# Pirates keep downloading even when it is free: evidence from a randomized experiment *

Miguel Godinho de Matos[1,a], Pedro Ferreira[2,3,b], and Michael D. Smith[3,b]

[1]Católica-Lisbon, School of Business and Economics
[2]Department Engineering and Public Policy, Carnegie Mellon University
[3]Heinz College, Carnegie Mellon University
[a]*miguel.godinhomatos@clsbe.lisboa.ucp.pt*
[b]{*pedrof,mds*}*@cmu.edu*

July 2, 2015

**Abstract**

*We partner with a major multinational telecommunications provider to analyze the effect of channel competition on the consumption of media of digital pirates. For a period of 45 consecutive days we activated, for free, a set of new TV channels dedicated to movies and TV shows to a group of randomly selected households. We use this exogenous increase in the TV content available to these households to study whether they reduce their consumption of Internet traffic in exchange for additional TV time. Using an Intention To Treat framework of analysis and instrumental variables derived from our randomized schedule, we find evidence of a strong substitution between TV and the Internet. We estimate that households that watched the new channels increased their daily TV viewership time by 4.9% and reduced their daily downloaded and uploaded traffic by 7.0%. Using information from Bit Torrent logs that allow us to identify part of the population of pirate households, we show this treatment also affected the pirate population. We show that torrent usage intensity decreased but that the probability of using Bit Torrent did not change. We also show that the population of pirates who are heavy users of Internet traffic are much less affected by our treatment. Finally, we show that the effect of thi treatment is short lived and does not extend beyond the experimental period.*

*This work has been partially supported by the Motion Pictures Association of America Through the IDEA Center at Carnegie Mellon University, by the iLab and the Living Analytics Research Center at the Heinz College. We also thank our Industrial Partner for their support.

# 1 Introduction

Since the introduction of Napster in 1999 and its close down in 2001, digital piracy has been subject to substantial research efforts in the fields of economics, marketing and management of information systems (Danaher et al., 2013). Two questions that often emerge are whether (1) firms suffer from illegal digital copies of copyrighted content and (2) the availability of such copies increases societal welfare (Peitz and Waelbroeck, 2006).

Theory proposes that when one consumer's willingness to pay for digital products is independent of other consumers, consumers are perfectly informed about product characteristics and there are no means of indirect appropriation of rents, then firms suffer from illegal copies of copyrighted content (Peitz and Waelbroeck, 2006). However, firms may profit from piracy when any of these three conditions is relaxed (Peitz and Waelbroeck, 2006) (e.g. if consumers pirate as a means of sampling as suggested in Chellappa and Shivendu (2005) or when pirated products exhibit network effects such as in Reavis Conner and Rumelt (1991)). The impact of piracy on societal welfare is harder todetermine and can go either way in any scenario (Peitz and Waelbroeck, 2006). For example, if pirates value content below market prices, piracy creates consumer surplus without ever affecting firm revenues in the short run (Rob and Waldfogel, 2007).

Most research papers on digital piracy quantify the magnitude of the displacement of digital content sales that is caused by piracy. Studies vary in terms of their focus (music vs. motion picture industry) as well as in the type of data and methods that they use (quasi experiments, survey based or observational) (Danaher et al., 2013). An initial stream of literature focused on the displacement of music sales by illegal downloads. Examples of prominent studies include those using survey data at individual level such as (Rob and Waldfogel, 2006; Zentner, 2006; Waldfogel, 2010) or those using city-level cross sectional data such as (Liebowitz, 2008). Except for (Oberholzer-Gee and Strumpf, 2007) these studies find that illegal music file-sharing is harmful for music sales. Oberholzer-Gee and Strumpf (2007) relate piracy downloads by album, collected from OpenNap servers, to the sales of the same album in the US. They use instrumental variables to address the endogeneity concerns, namely unobservables that could derive both album sales and illegal downloads. They do

not find evidence of a harmful effect of downloads on sales. However, the exogeneity of their instrument has been highly criticized (Rob and Waldfogel, 2006; Liebowitz, 2008).

The development of the BitTorrent protocol in 2003 generalized the downloads of movies and of TV shows. Following this trend there was a wave of papers analyzing the impact of illegal downloads of movies on the motion pictures industry revenue: Bounie et al. (2006); Rob and Waldfogel (2007); Bai and Waldfogel (2012) use survey data and find that file-sharing is harmful for the motion pictures industry. A similar conclusion is obtained by (Danaher et al., 2010; Danaher and Smith, 2014) using data from natural experiments. In this line of research, Smith and Telang (2009) are an exception because they show that the availability of the pirated copies at a given point in time has no impact on subsequent DVD sales. They reach this result exploring the fact that TV movie broadcasts induce both legal DVD sales as well as illegal downloads and they use these broadcasts as an exogenous shock to demand. However, as the authors acknowledge, their results do not look at the impact of piracy in the early stages of a movie's life cycle where piracy may have a significant negative impact on sales.

Other works in digital piracy focus on studying deterrence. Examples include looking at the impact of legal litigation on piracy. Researchers find that increasing threats decreases piracy but that the availability of illegal content remains substantial (Bhattacharjee et al., 2007). Additionally, some attention has also been directed at understanding the protection of digital content via Digital Rights Management (DRM). For example, (Sinha et al., 2010) show that DRM may actually increase piracy because it decreases the usability of digital content. This, in turn, reduces the consumers' willingness to pay for content.

More recently, the continuous development of digitization and the bundling of fixed and mobile communication services with high speed Internet, TV service, Video-on-Demand (VoD), Digital Video Recording (DVR) and streaming of TV over IP directly to devices such as laptops and smart phones (Bronnenberg et al., 2010) researchers started exploring the role of channel competition as a potential way to displace piracy Danaher et al. (2010). The rationale stems from the increase in content variety that consumers can enjoy (Resnick and Varian, 1997) and, more importantly, from the observation that a significant part of the new content that is available can be consumed for free using new technologies such

as digital video recording (VCR) or time shifted TV (TSTV)(Wilbur, 2008). Yet, so far, there are almost no studies that analyze the impact of direct competition from new digital distribution channels on digital piracy. The only example known to us is Danaher et al. (2010) where the authors conclude that removing TV content from iTunes led to a large increase in piracy of the same content, which hints at substitution between channels (in this case, between Internet downloads and the iTunes store).

Our paper contributes to advance our understanding of how channel competition interacts with piracy in the context of the motion pictures industry. We partner with a major multinational telecommunications provider to analyze the effect of channel competition on the consumption of media by digital pirates. For a period of 45 consecutive days we activated, for free, a set of new TV channels (hereinafter called the cinema pack) dedicated to movies and TV shows to a group of randomly selected consumers. We use this exogenous increase in the TV content availability to these households to study whether they reduce their consumption of Internet traffic in exchange for additional TV time. Using an Intention To Treat (ITT) framework of analysis and instrumental variables derived from our randomized schedule, we find evidence of a strong substitution between the TV and the Internet. We estimate that households that watched the new channels increased their daily TV viewership time by 4.9% and reduced their amount of daily downloaded and uploaded traffic by 7.0%. Using information from Bit Torrent logs that allow us to identify part of the population of pirate households, we show this treatment also affected the pirate population. We show that torrent usage intensity decreased, but that the probability of using Bit Torrent did not change. We also show that the population of pirates who are heavy users of Internet are much less affected by our treatment. Finally, we show that the effect of treatment is short lived and does not extend beyond the experimental period, namely once households are asked to pay for the new TV channels.

Our paper is organized into seven sections. In section 2 we describe our industrial partner and the datasets that we obtained from it. In section 3 we present our experimental design in detail. Section 4 shows detailed descriptive statistics that provide an model free overview of our results. Sections 5 and 6 present the main results of our paper. Section 7 discusses implications, future work and concludes.

# 2 Context and Dataset

Our industrial partner, hereinafter called TELCO, is a large multi-national telecommunications firm with core operations in Europe. It is the market leader of Pay-TV services in the country we analyze serving more than 1 million households. TELCO offers high speed Internet, TV, Video-on-Demand (VoD), Digital Video Recording (DVR), online TV and Automated Cloud Recording (ACR).

## 2.1 The TV service

TELCO's basic TV service includes 100 channels and free access to a VoD store from which clients can rent movies more than 2000 at any time. In addition, TELCO sells packages of channels including documentary, music, sports and cinema that households can subscribe on a-la carte model. The cinema pack includes 10 channels – 8 with movies and 2 with TV shows. This package can be purchased for 13 USD/month and it is the core focus of our study. The cinema pack broadcasts recent movies and TV shows including popular titles such as Netflix's House of Cards, Suits and Fargo. More importantly, in what pertains to TV shows, the cinema pack broadcasts content that has been aired only a couple of days ago in the US.

The channels included in the cinema pack allow for Automated Cloud Recording (ACR) free of charge. ACR stores in the cloud without user intervention up to one week of the content that aired in a subset of the TV channels offered by TELCO. In practice, this means that at any time, the TELCO's consumers of PayTV can choose to stream, in real time, on their TV screen, any program that aired in the past week in any of the channels supporting ACR. Therefore, these consumers have access to a variety of entertainment content that is orders of magnitude larger than what is used to be any time before.

## 2.2 The Internet service

TELCO offers both fixed and mobile Internet to its clients. In this paper we focus exclusively on the fixed service for which we have data. Fixed Internet access offers downstream speeds ranging between 3 Mbps and 300 Mbps. Upload speeds vary between 1 Mbps and 10 Mbps.

There are no restrictions to the amount of traffic that each household can use per month.

Overall, in TELCO's network, 35% of all Internet traffic (downloads + uploads) pertains to web browsing, 19% to Internet video and 37% to protocols related to peer-to-peer file sharing. The remaining 9% are dispersed across several protocols including video-gaming and voice-over-ip. From the identifiable traffic using peer-to-peer protocols Bitorrent streams account for 20% of the overall download traffic and 68% of upstream traffic. Peer-to-peer file sharing traffic originates in a small fraction of the consumers (less than 10%) that make a disproportionate use of network resources.

## 2.3 Our Dataset

We obtained an anonymized monthly panel of all client accounts for the months of April and May 2014 and then for the months of December 2014, January and February 2015. For each month, and each client account, the dataset contains a snapshot of all the products subscribed at the end of the month as well as variables that summarize the consumption of each service during the month. Specifically, the dataset contains information on bundle subscriptions, TV use (e.g., hours watched, channels viewed), download and upload consumption and a log of identifiable torrent streams.

We used the data for the months of April and May to set up our experiment. Then the analysis in this paper focuses on the period ranging from December 2014 to February 2015. We do not have access to data between May and December 2014. Moreover, we used this time to setup our pilot study in a way that would ensure confidentiality and anonimity. TELCO collects only aggregate Internet usage statistics. It does not hold data at the household level on links accessed. We overcome this limitation by relying on a third party provider of torrent traffic that regularly monitors the most popular active swarms available online and keeps a log of all IP addresses that are seen sharing content. It also registers identifiers for the content that is shared. These IP addresses could then be matched to the range of IP addresses that TELCO provides to the households it serves. This matching was carried out by a fourth entity that provides data escrow services. This ensured that no one would be able to revert the record encryption keys matching client accounts to their illegal online activities nor to track who the original data providers were. We did not disclose to TELCO

the entity providing the torrent logs and we did not disclose to the torrent data provider the source of user level data.

While we monitored closely all data collection processes at TELCO, we had no control over the data collection process of the torrent streams. The data provider ensures that the torrent logs it gave us are a representative sample of torrent streams. Yet, they do not account for the overall torrent traffic that traverses the Internet. Therefore, the IP addresses found sharing content by the data provider are most likely the most active Bit Torrent users but not necessarily a representative sample of the population using Bit Torrent worldwide. The practical consequence for our study is that we can know for sure that a household is doing piracy if we have a stream in our logs that matches its IP address, but a household that we cannot match to our torrent logs may still be sharing copyrighted content.

# 3 Experimental Design and Setup

In Randomized Control Trials (RCTs) one creates groups of statistical units of interest, in our case TELCO's clients, that are similar with regards to baseline characteristics, both observed and unobserved, that may influence the outcome of treatment (Rubin, 1974). With simple randomization, the researcher makes a single randomly assigned list of subjects before the experiment. Some subjects are assigned the treatment condition — in our case treated households receive the **cinema pack** for free during 45 consecutive days — and the remaining subjects are assigned to the control state — in our case control households are left alone. The objective of this setup is to causally attribute differences in variables of interest after treatment — in our case TV time, downloads, uploads and torrent activity — to the effect of treatment — in our case having the cinema pack (Kernan et al., 1999).

Simple randomization can fail if it creates groups of treated households that are unbalanced for critical characteristics that are known or suspected to affect treatment outcomes (Kernan et al., 1999). Furthermore, simple randomization can also fail if the analysis of interest is at the level of a particular subgroup of the population that is rare because of insufficient statistical power (Assmann et al., 2000). In our case, our goal is to study the impact of channel competition on the pirate's consumption of TV and Internet. Therefore,

we needed to find a large enough group of pirate households. According to the torrent logs that we had access to for the months of April and May, known pirates are a rare population. They account for approximately only one tenth of the households that can receive the **cinema pack**. Furthermore, we believe that the willingness to use the cinema pack is related to unobserved household characteristics that determine the household's piracy activity. These features of our setup favor an experimental study planed with a stratified sample rather than a simple randomization schedule.

Stratified randomization is a two-step procedure in which first clients are divided into strata that are likely to influence the effectiveness of treatment only then and within each stratum clients are assigned a treatment condition according to distinct randomization schedules (Simon, 1979). Stratification has two major advantages over simple random sampling: (1) it allows for obtaining estimates of the effect of treatment for each stratum with some desired statistical power; and (2) it enhances precision when it is based on variables related to the effect of the treatment (Levy and Lemeshow, 2013). In the next sections we describe our criteria for stratification as well as the details of the experimental procedure.

## 3.1 Stratifying the Sample

The initial setup phase occurred in the month of November 2014 using household data from the months of April and May 2014. Let $H$ denote the set of TELCO's households and $h$ a particular household. Using the monthly user panel for April-May 2014, we develop a model $M()$ that given a vector of household characteristics $X_h$ determines the likelihood that $h$ engages in illegal downloads through BitTorrent: $M(X_h) = p_h$. Model $M$ was developed using state-of-the-art data mining algorithms for classification tasks. The details of the algorithm used and tested, training parameters, cross validation and performance evaluations are presented in the appendix A.

The purpose of model $M$ was to stratify TELCO's clients into homogeneous groups regarding the set of observable characteristics that were associated with piracy behavior. As stated earlier our hypothesis was that illegal downloading activity would be correlated with the likelihood of our treatment having any effect. Taking this into consideration we stratified the population using the confusion matrix for the output of the machine learning

algorithm that partitions the user space in four strata: $G01$ - clients that model $M$ predicts as non pirates that we could not match to the torrent log (April-May 2014); $G02$ - clients that model $M$ predicts as pirates, but that we could not match the torrent log; $G03$ - clients predicted as non pirates by $M$, but that we could match to the torrent log; and $G04$ - clients that $M$ predicted as pirates that we could match to the torrent log.

Table 1 shows the output of applying model $M()$ to all TELCO's households that had both TV (with a set-top-box) and Internet service. It shows the size of each strata that we considered in our analysis.

Table 1: Experimental groups of households as classified by our algorithm (details in Appendix).

| Group Name | Treatment Assignment | | |
|---|---|---|---|
| | Confirmed Pirate | Predicted Pirate | Overall Population |
| G01 | No | No | 355,420 |
| G02 | No | Yes | 34,169 |
| G03 | Yes | No | 20,831 |
| G04 | Yes | Yes | 20,702 |

We use a machine learning algorithm to stratify our data rather than observable household characteristics because (1) we have no clear theory as to what household observables would be good determinants of piracy behavior; and (2) we want to keep the number of stratum to a minimum to allow for a large enough sample size in each stratum. As detailed in appendix A, the most influential variables that $M$ uses to predict confirmed piracy behavior are total Internet download and upload traffic and Internet service tenure. Consequently, table 1 partitions households in terms of their intensity of Internet usage and separates households that are actually pirates from those that have a pattern of Internet usage consistent (inconsistent) with piracy.

Groups $G02$ and $G04$ are households that use Internet intensively. Households in $G02$ register 3.7 GB of daily downloads and 2.3 GB of daily uploads on average. Households in $G04$ register a daily average of 4.4 GB of downloads and 3.0 GB of uploads on average. Households in $G01$ and $G03$ are more modest users of Internet. Households in $G01$ download

1.2 GB per day and upload 270 Mbytes per day while households in $G03$ download 2.1GB per day and upload approximately 800 MB of content per day (all statistics are at the average). Appendix B contains further details on the descriptive statistics of the households in each strata in the period preceding the experiment.

## 3.2 Minimum Sample Size

In the absence of priors for the potential magnitude of the effect of treatment we determine the minimum sample size for each stratum assuming that after treatment, on average, households would see their preferred TV show using the TV rather than downloading it illegally from the web. An effect less than this is also arguably uninteresting from an economic point of view. Taking this into consideration, we planed our sample to identify changes of 450 MB of downloads per day and 15 minutes of TV usage per day with a confidence level of 95% and power of 80%. Note that 450 MB of downloads corresponds to approximately 15 minutes of youtube video at 1080p. According to Netflix 450 MB is also the average size of a TV show. Table 2 summarizes our sample size calculations. We need at least 3057 households in each treatment condition to identify our hypothesized effect of treatment in the most demanding sub-population ($G04$). We follow the standard procedure (Kernan et al., 1999) and allocate similarly the sample across strata given that our major goal is to test hypotheses across strata, in particular compare pirate households against non pirates. After negotiations with TELCO we random sample 9,000 households in each stratum. Treatment assignment within each stratum followed a simple randomized scheduled such that half of TELCO's clients in each stratum were assigned the treatment condition and the other half were assigned the control condition.

## 3.3 The experiment

The experiment consisted in awarding access to the cinema pack for a period of 45 consecutive days free of charge to TELCO households as described in the previous section. The activation of the cinema pack was done automatically and did not require client intervention. At the time of activation, each TELCO client listed for treatment received an e-mail together with

Table 2: Minimum sample sizes required to identify changes in 450MB of downloads per day and 15 minutes of TV time per day with a 95% confidence level and 80% power.

| Measure | Stratum | Historical daily sd | Min. Sample Size |
|---|---|---|---|
| Downloads | G01 | aprox 2,000 MB / day (Apr-May) | 311 |
| | G02 | aprox 3,000 MB / day (Apr-May) | 698 |
| | G03 | aprox 2,400 MB / day (Apr-May) | 447 |
| | G04 | aprox 6,300 MB / day (Apr-May) | 3,057 |
| TV Time | G01 | aprox 2.4 hours / day (Apr-May) | 1,329 |
| | G02 | aprox 2.6 hours / day (Apr-May) | 1,698 |
| | G03 | aprox 2.9 hours / day (Apr-May) | 2,113 |
| | G04 | aprox 2.5 hours / day (Apr-May) | 1,570 |

a text message sent to their mobile device that announced the activation of the cinema pack as a Christmas gift. Clients were told that the gift would extend through the entire month of January. After the experimental period, clients in the treatment condition also received a follow up e-mail and a text message notifying them of the end of the free sampling period and inviting them to subscribe the service for a fee.

For technical reasons, the activation of the cinema pack took place over 3 days – 15th to the 18th of December. The order of activation assignment was such that all clients in the same stratum were activated at the same time. The order of stratum activation was randomized and the time of activation was logged for each household. During the experiment, for each and every household, we monitored daily TV usage and daily aggregated Internet traffic (both download and upload).

## 3.4 Unexpected Reductions in Sample Size

At the time of randomization we did not have information about whether households had opted out from marketing campaigns. We did not know if households were included in other marketing trials that would preclude them from participating in our campaign. We were unaware of the validity of client contact information registered in the TELCO system and

between the time of listing and the start of the experiment, households could have subscribed to the cinema pack by themselves. These issues led us to remove 3,270 households from the initial list of 36,000 households listed to be part of this experiment.

More importantly, TELCO has two types of clients with set-top-boxes allowing for the automated activation of the cinema pack: up-to-date and legacy. They provide slightly different TV interfaces and use different monitoring tools. In fact, we are unable to confidently track the TV and Internet usage of legacy set-top-box, thus we exclude an additional 7,590 households from our analysis in this paper. We also needed to excluded the 1,549 clients that churned during the experiment or that upgraded (downgraded) from (to) the legacy set-top-box software because if they had been randomly assigned the cinema pack this assignment would have been automatically canceled for business reasons out of our control.

Finally, we exclude an additional 2,125 households for whom we did not have one day of TV and Internet usage information in each of the three periods of our experimental study: pre-experiment (2014-12-01 - 2014-12-13), experiment (2014-12-19 - 2015-02-02) and post-experiment (2015-02-07 - 2015-02-28). These were households that had their Internet modem and set-top-box turned off during the experiment for whom we cannot reliably measure the effect of treatment.

We keep 21,466 households in our study and note that all reasons for sample attrition are orthogonal to our random assignment. Therefore, all results reported in this paper have a causal interpretation. The caveat is that we need to acknowledge that they may not be representative of the population that we initially plan to use in our experiment but rather from a subset that excludes all legacy clients, clients opting out from of marketing campaigns, all churners and all clients that had their set-top-box and Internet modem turned off during the our experiment.

Table 3 summarizes the sample of households used in our experiment. Sections 4.1, 5.1 and Appendix B show extensively that treatment assignment was unrelated with being included in this sample. Finally we note that despite the trimming described above the sample size is still large enough to identify the hypothesized magnitude of the effect of treatment in each stratum with a confidence level of 95% and a power of 80%.

Table 3: Details of Random assignment of households to different treatment cond

| Group Name | Condition | | Sample Size |
|---|---|---|---|
| | Confirmed Pirate | Predicted Pirate | (T = Treated) |
| G01 | No | No | 6,107 (T = 3,077) |
| G02 | No | Yes | 5,134 (T = 2,508) |
| G03 | Yes | No | 4,307 (T = 2,153) |
| G04 | Yes | Yes | 5,918 (T = 2,963) |

# 4 Descriptive Statistics and Model Free Evidence

## 4.1 Balance Statistics

Table 4 shows that despite trimming our randomized scheduled achieved a good balance in key observed user characteristics across treatment and control groups.

Table 4: Overall balance statistics for pre-trearment variables for the period ranging from December 1-13.

| | tx.mn | tx.sd | ct.mn | ct.sd | std.eff.sz | stat | p | ks | ks.pval |
|---|---|---|---|---|---|---|---|---|---|
| Population Weights | 13.288 | 15.990 | 13.081 | 15.877 | 0.013 | 0.950 | 0.342 | 0.006 | 1 |
| Pirate Score | 0.502 | 0.290 | 0.508 | 0.288 | -0.020 | -1.489 | 0.137 | 0.013 | 1 |
| Tenure (TV months) | 93.812 | 57.908 | 93.249 | 57.365 | 0.010 | 0.714 | 0.475 | 0.011 | 1 |
| Tenure (Internet months) | 67.168 | 35.271 | 67.204 | 35.203 | -0.001 | -0.075 | 0.940 | 0.008 | 1 |
| Tenure (Telephone months) | 55.478 | 19.534 | 55.612 | 19.446 | -0.007 | -0.502 | 0.616 | 0.008 | 1 |
| Active Contract | 0.808 | 0.394 | 0.807 | 0.395 | 0.003 | 0.237 | 0.813 | 0.001 | 1 |
| Download (MB per day) | 2,905.670 | 4,274.043 | 2,945.733 | 4,659.159 | -0.009 | -0.655 | 0.512 | 0.018 | 1 |
| Upload (MB per day) | 1,626.739 | 4,377.533 | 1,629.845 | 4,195.579 | -0.001 | -0.053 | 0.958 | 0.017 | 1 |
| TV Channels zapped | 11.762 | 8.295 | 11.773 | 8.104 | -0.001 | -0.096 | 0.923 | 0.012 | 1 |
| TV Time - New Chan (Hours per day) | 0.062 | 0.272 | 0.066 | 0.285 | -0.014 | -1.023 | 0.306 | 0.008 | 1 |
| TV Time - All Chans (Hours per day) | 4.383 | 2.496 | 4.426 | 2.486 | -0.017 | -1.248 | 0.212 | 0.012 | 1 |
| VoD streams | 0.970 | 1.114 | 0.973 | 1.103 | -0.003 | -0.225 | 0.822 | 0.014 | 1 |
| VoD TV time shifts | 0.022 | 0.172 | 0.025 | 0.204 | -0.017 | -1.275 | 0.202 | 0.007 | 1 |
| VoD title page view | 0.086 | 0.406 | 0.092 | 0.397 | -0.015 | -1.064 | 0.287 | 0.009 | 1 |
| Number of days | 9.980 | 3.538 | 10.043 | 3.489 | -0.018 | -1.304 | 0.192 | 0.009 | 1 |

For each covariate, balance is assessed using t-test for the difference in means between treatment and control and the Kolmogorov-Smirnov statistic (KS) showing that both samples (treated and control) were drawn randomly from the same distribution (Conover and Conover, 1980; McCaffrey et al., 2013). In all cases, we cannot reject the null hypothesis that the treatment and control groups are statistically equivalent at 95% confidence level

both in terms of means and distributions. Additional balance statistics within each stratum are provided in Appendix B. These results show that balance is achieved both overall and within each stratum.

## 4.2 TV and Internet Usage

Figure 1 previews the effect of treatment assignment on TV time watched by each household (first panel), on the total amount of Internet usage measured by the sum of downloads and uploads per day (second panel) and on the ratio of TV time watched per MB of Internet consumed converted to equivalent Youtube minutes.[1]. We use the logarithm of the measures of interest due to the large variance of Internet usage, in particular, in what upload traffic is concerned. This figure suggests that after treatment households increased their TV usage and decreased their Internet traffic. The figure also suggests that the effect of treatment ceased as soon as the experiment ended.

Figures 2 and 3 show the use of the cinema pack across our sample. Figure 2 focus on the average number of minutes per household per day spent watching the cinema pack. Figure 3 shows the total TV usage time besides the time spent watching the cinema pack. This figure suggests that there was substitution in TV usage towards the new set of channels that ceased as soon as the experiment ended.

## 4.3 Torrent Logs During the Experiment

Figure 4 decomposes the torrent logs that we match to the households in our sample. This figure shows that over 60% of the streams pertain to content that could not be identified. Videos account for 28% of the time and were the second most common stream type. Audio, software and other content types (e.g. text) were much less frequently found and together amount to the remaining 12% of the streams.

Figure 5 details the video content shared in the torrents that we match to the households in our sample. TV show episodes account for 38% of the streams and movie account for 35% of the streams. The remaining 27% of the streams pertain to adult content and other types

[1]Each youtube minute at 1080p corresponds to 30 MB of Internet traffic

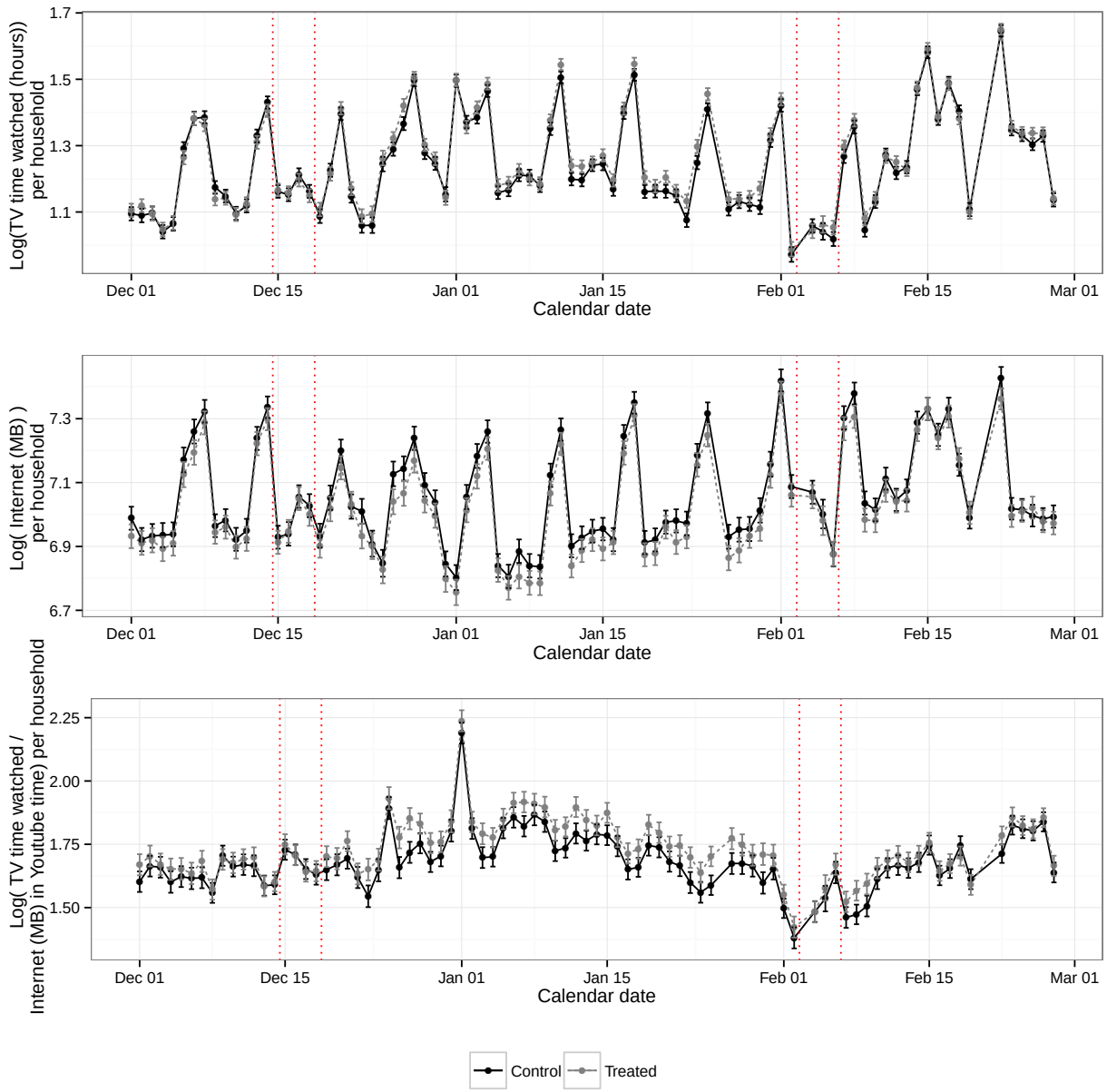



Figure 1: Summary of Internet and TV usage before, during and after the experiment. The marks in the graph show the 95% confidence intervals.

of videos.

Figures 6 and 7 speak to the quality of our model based stratification approach, in particular, when we recall that we can only prove that a household engages in piracy when we observe it in our torrent logs. Figure 6 shows the fraction of households that we match to the torrent logs during the months of December and January across strata. Likewise

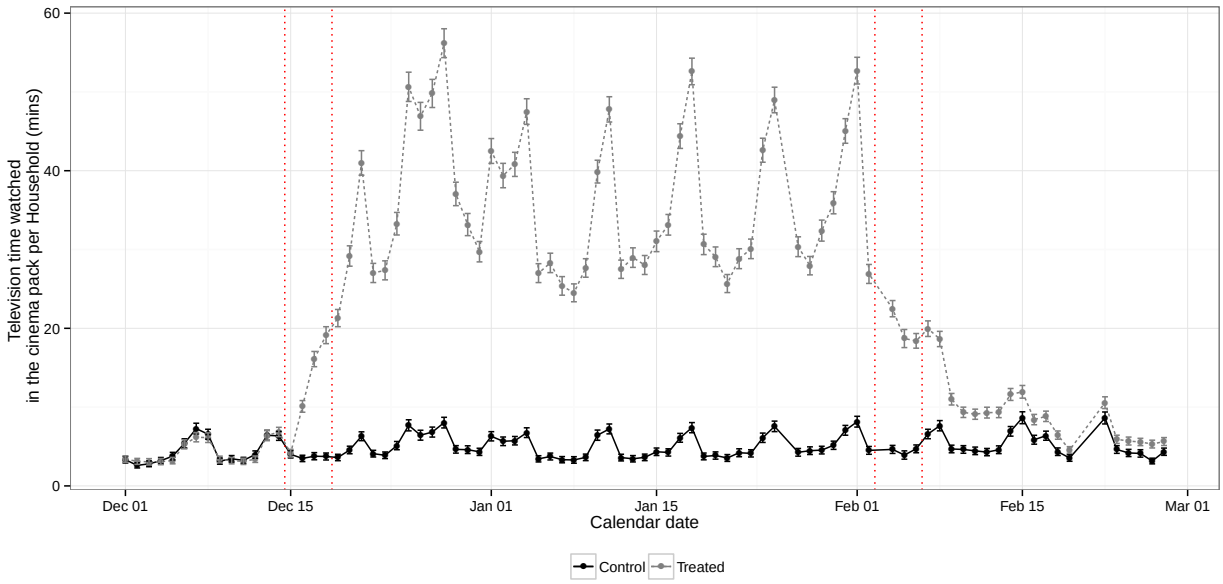



Figure 2: Average TV time watched in channels included in the cinema pack.

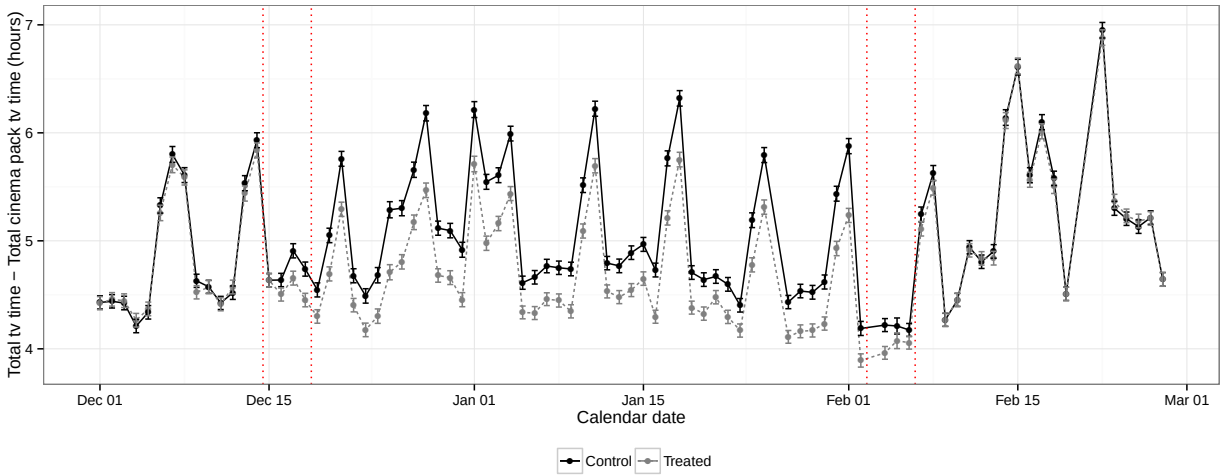



Figure 3: Average TV time watched over all channels except those included in the cinema pack.

for Figure 7 on a daily basis. We confirm that the vast majority of users in $G03$ and $G04$ are pirates. Also, households in $G04$ are more active pirates than households in $G03$. The proportion of users that we match in the daily torrent logs is twice as large for the former.

We also see that only a small fraction of households in group $G01$ used Bit Torrent during

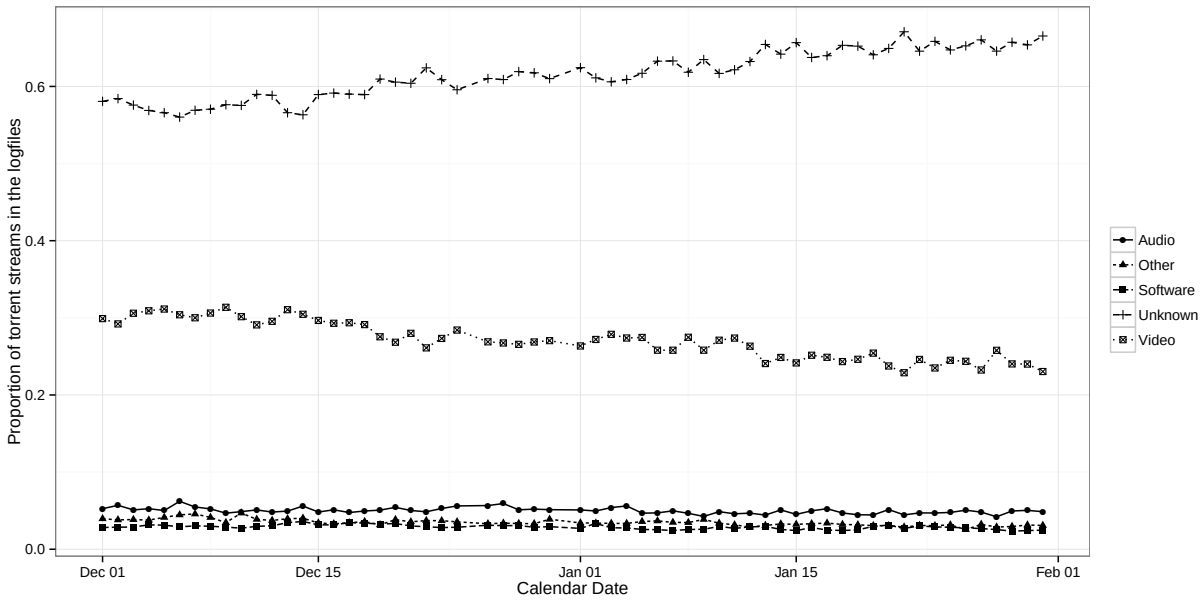



Figure 4: Torrent streams by content type during the period of analysis.

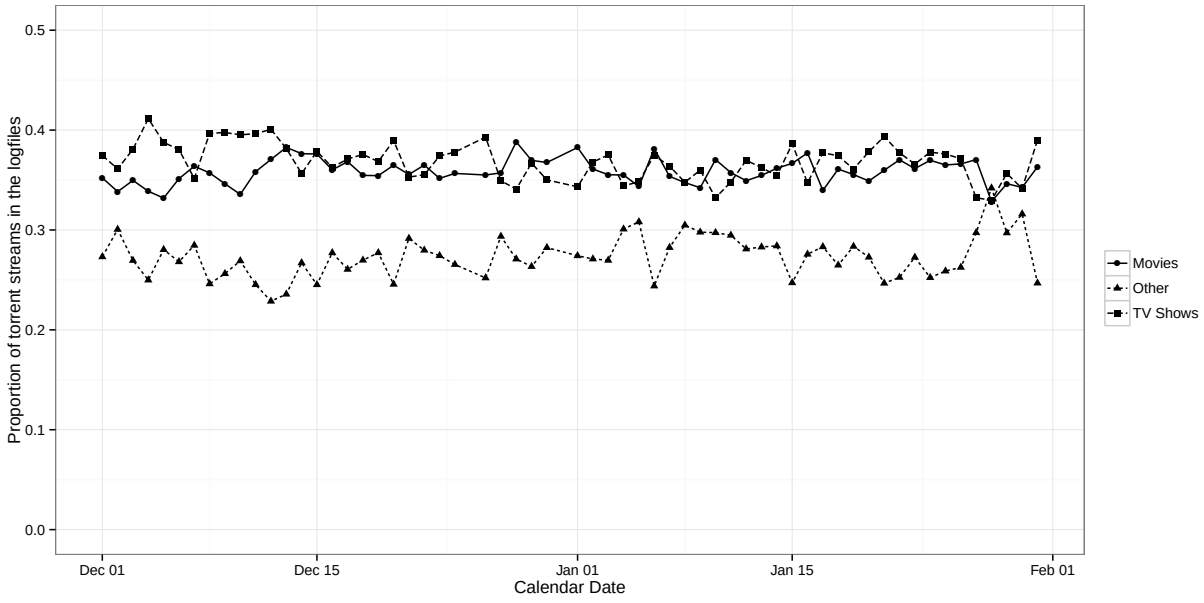



Figure 5: Video torrent streams shared by sub-type during the period of analysis.

the experiment. Yet, we find a significant number of households in $G02$ sharing files using Bit Torrent. Recall that our model classified households in these groups as pirates though we could gather of pirate activity at the time of the experimental setup. Given the amount of Internet traffic that households in $G02$ consume and the difficulty in locating their IP

addresses in the torrent logs, we suspect that they may be more sophisticated Internet users who are able to better camouflage their online (illegal) activities.

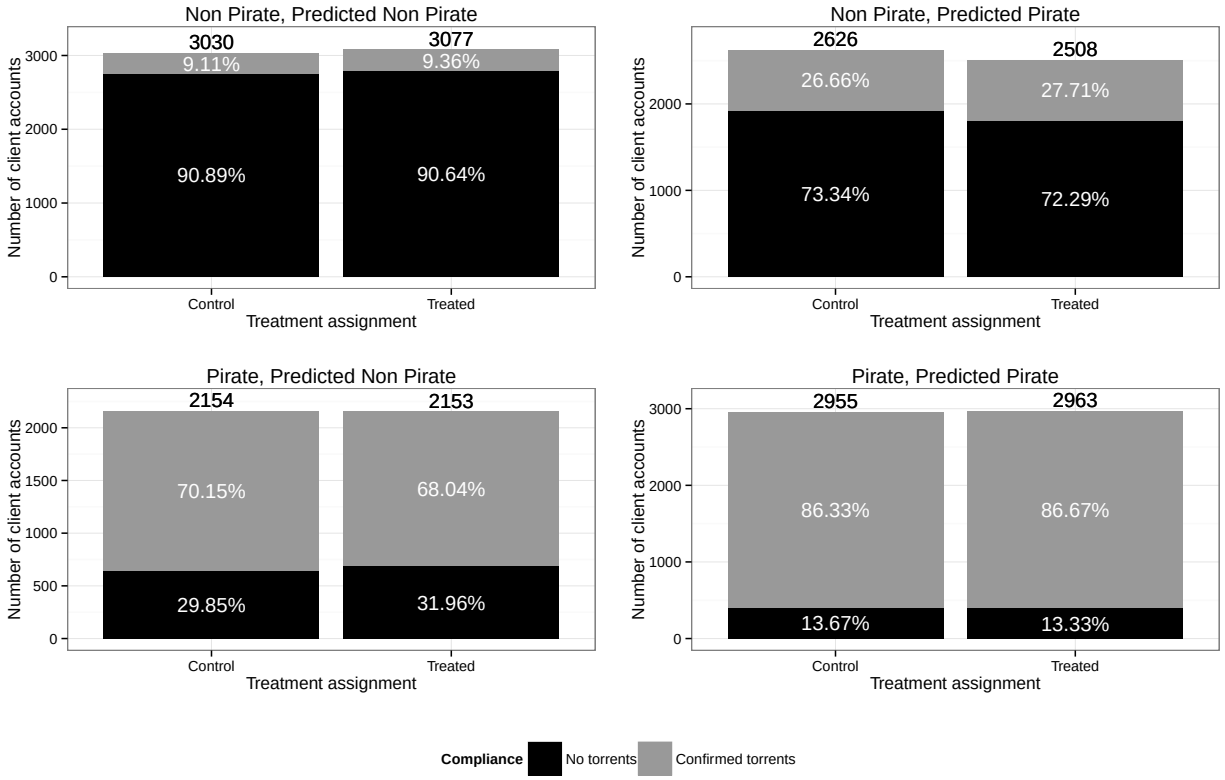



Figure 6: Households matched to Bit Torrent logs in each stratum over the entire period of analysis.

Note, however, that we cannot use the torrent logs from December and January to re-classify households because this period of time coincides with the implementation of our experiment and our treatment could have changed the behavior of households. Furthermore, if we moved households across strata we would have no good way to calculate their new sampling weights. Therefore, going forward, we keep the dataset as originally planed. The novelty is that we know that some households use Bit Torrent to engage in piracy that we have not flagged before as doing so. This does not invalidate our analysis in anyway but highlights a limitation of our study in that we can only identify pirates after they show up in our logs. Therefore, our classification of pirates is conservative in the absolute number of households that we find pirating and aggressive in the sense that it likely captures well the

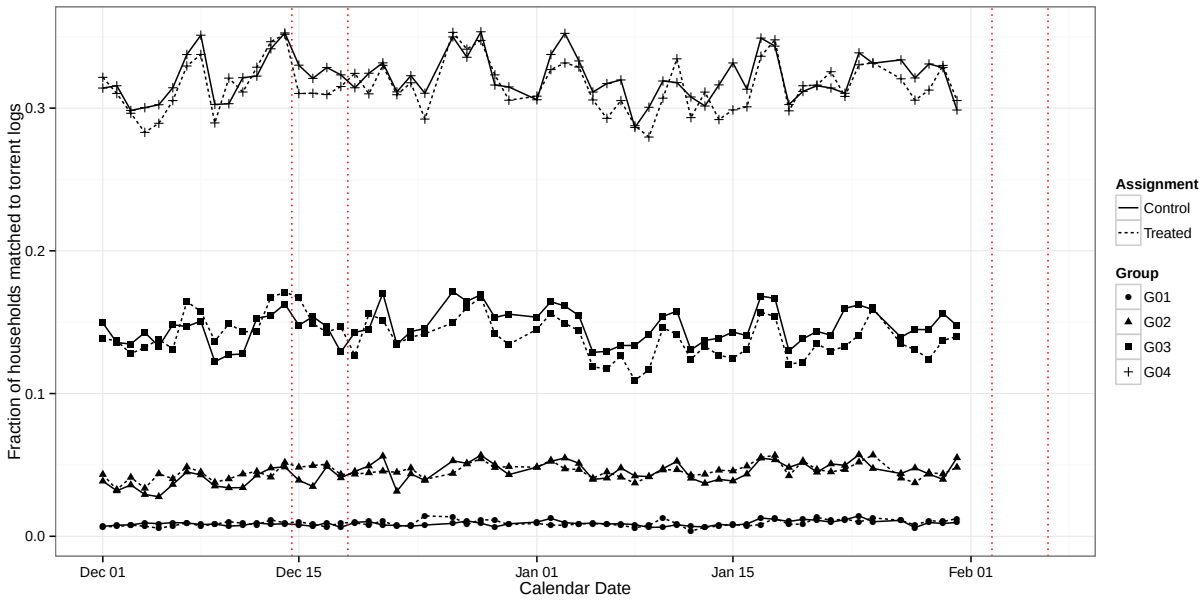



Figure 7: Fraction of households matched to Bit Torrent logs in each stratum each day over the period of analysis.

most active pirates. Finally, we note that our analyses focus on households in $G03$ and $G04$, which we knew to be pirates at the time of the experimental setup. Figures 8 and 9 show the fraction of households in $G03$ and $G04$ (confirmed pirates as of May 2014) that download movies (excluding adult content) and TV show episodes, respectively. These figures provide preliminary evidence that pirate households do seem to have reduced their probability of sharing movies through Bit Torrent (even if they might have changed the intensity of their sharing activities) but that they might have reduced their likelihood of sharing TV shows.

## 4.4 Treatment Compliance

Ideally units allocated to each treatment condition would comply with the assignment. However, in practice, subjects may choose not to do so either because they obtain treatment by other means or because they do not take the treatment awarded (Frangakis and Rubin, 1999; Hollis and Campbell, 1999; Gupta, 2011). In situations where there are deviations from the scheduled randomized assignment, comparing the treatment and the control group will measure only the effect of the Intention To Treat (ITT) (Frangakis and Rubin, 1999;

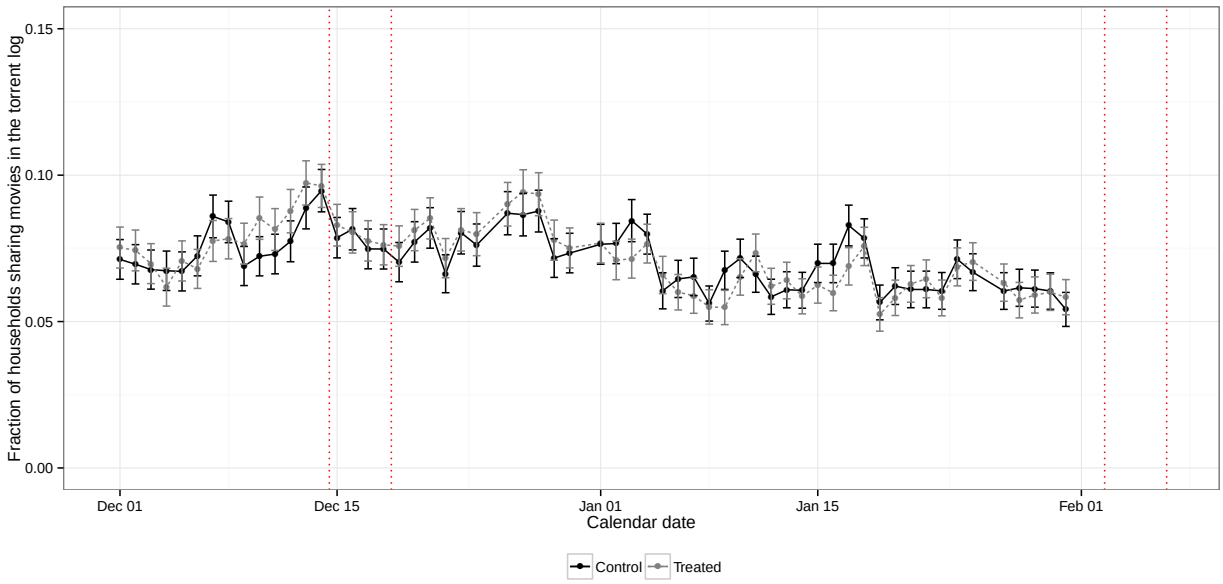



Figure 8: Fraction of households in groups $G03$ and $G04$ matched to movie torrent log streams (excludes adult content).

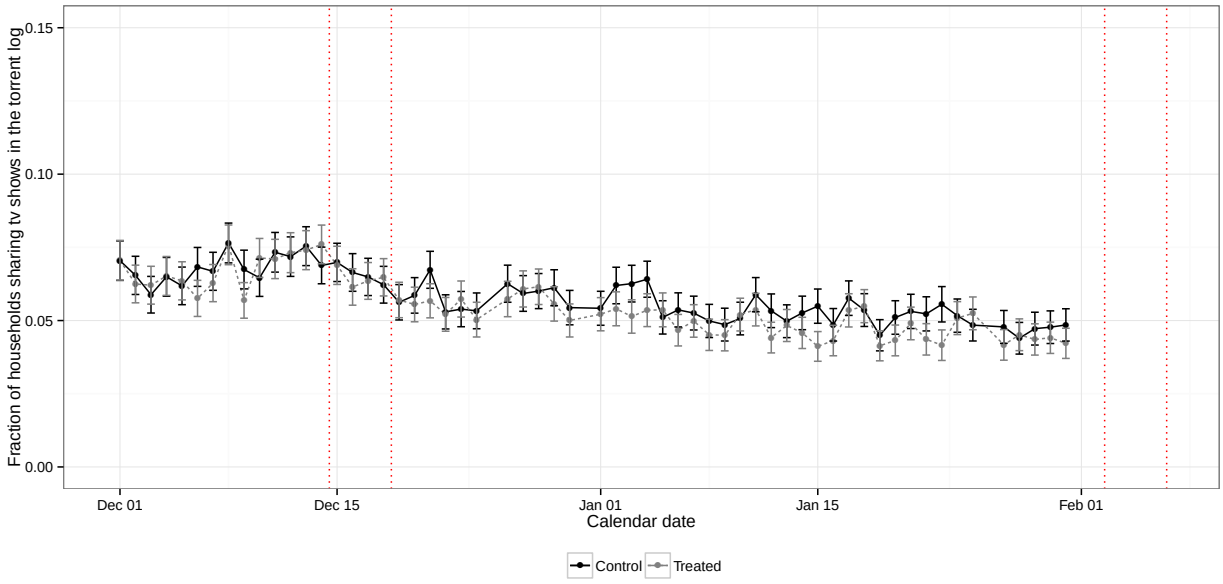



Figure 9: Fraction of households in groups $G03$ and $G04$ matched to TV shows torrent log streams.

Hollis and Campbell, 1999), which provides a lower bound for the Average Treatment Effect (ATE) (Angrist et al., 1996).

In our experimental context the deviations from the scheduled randomized assignment of treatment were significant. A number of households that were given the cinema pack did not use it. Furthermore, some households who were not awarded the cinema pack purchased it independently of our assignment. We follow (Angrist et al., 1996) to account for these deviations and report the ITT effect and the Local Average Treatment Effect (LATE) where we use treatment assignment as an instrumental variable for treatment compliance. The latter is the average effect over the subpopulation of households '*whose behavior can be modified by randomized assignment (compliers)*"(Angrist, 1991).

We quantify compliance by looking at the fraction of households that received the cinema pack that used it for more than 90 minutes at least once during the period of the experiment[2]. In addition, each day we measure the fraction of households that used the cinema pack for more than 90 minutes during that day. Figures 10 and 11 summarize the compliance levels overall and within each stratum over the entire experimental period. Figure 12 summarizes the daily compliance levels over all strata. These figures show that there is non-compliance across both the treatment and the control groups.

# 5 The Effect of the Cinema Pack on TV and Internet consumption

We analyze the effect of treatment assignment on the total usage of TV measured in hours/day and on the total amount of download and upload traffic measured in MB/day. We estimate the effect across all households included in the experiment and we split the analysis in three distinct experimental periods: the pre-treatment phase (2014-12-01 - 2014-12-13), the treatment phase (2014-12-19 - 2015-02-28) and the post treatment phase (2015-02-08 - 2015-02-28). The main goal of this section is to test whether the cinema pack had an effect on the overall consumption of media at household level. We will then look at household piracy behavior in section 6.

[2]The majority of movies range from 90 to 120 minutes. Our results are robust to changes in this definition. We tested 20, 30, 60, 90, 120 minutes and the results are qualitatively the same. They are available upon request.

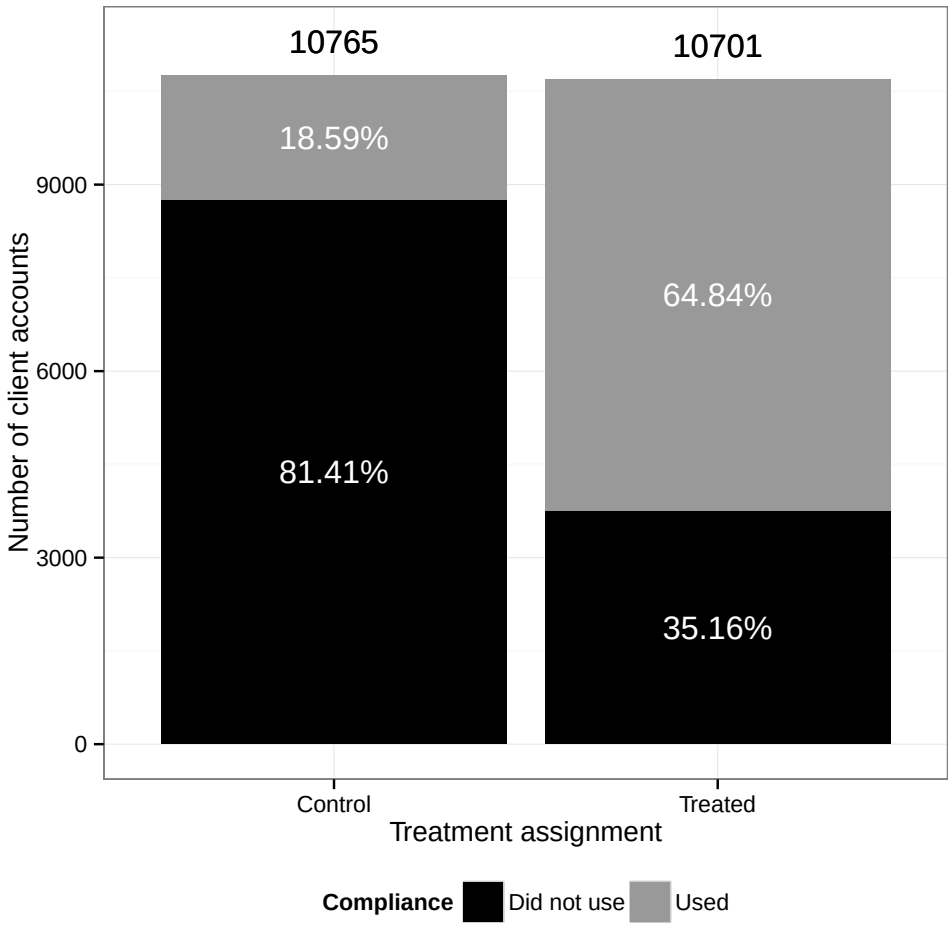



Figure 10: Compliance with treatment over the entire treatment period.

## 5.1 The Pre-Treatment Effect

Table 5 tests the effect of treatment assignment in the period prior to the experiment. The purpose of this regression is to show that before the activation of the cinema pack both the treatment and the control groups had the same level of TV and Internet usage both overall and within each stratum. For each of the three regressions presented in this table, we test the null hypothesis that treatment had no effect, that is, the coefficient on treated and the coefficients on treated interacted with stratum are jointly zero: $Treated = 0, Treated{*}G02 = 0, Treated * G03 = 0, Treated * G04 = 0$. The p-values for these tests are 0.867, 0.419 and 0.641, respectively for columns (1), (2) and (3). Therefore, in all cases we cannot reject the null hypothesis that the treatment and control groups have the same level of TV usage and

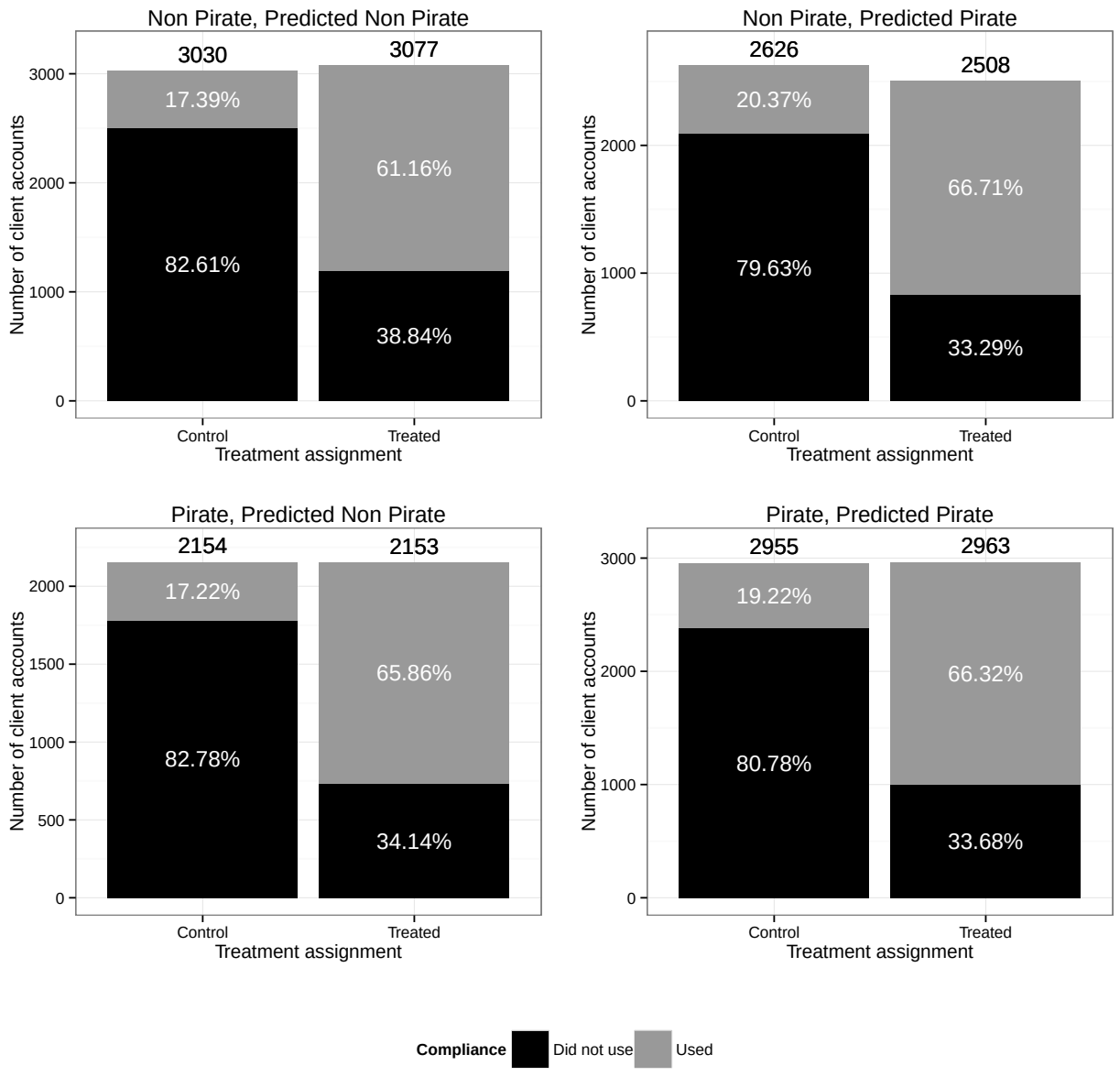



Figure 11: Compliance with treatment in each stratum.

of Internet usage both in terms of upload and download traffic. This analysis adds to the descriptive statistics provided in subsection 4.1 and shows balance between treatment and control even after trimming the sample as described in section 3.4. This table also shows that the baseline usage of TV is similar across strata but that Internet usage, both uploads and downloads, vary widely across strata, with households in $G02$ and $G04$ doing the most use of the Internet.

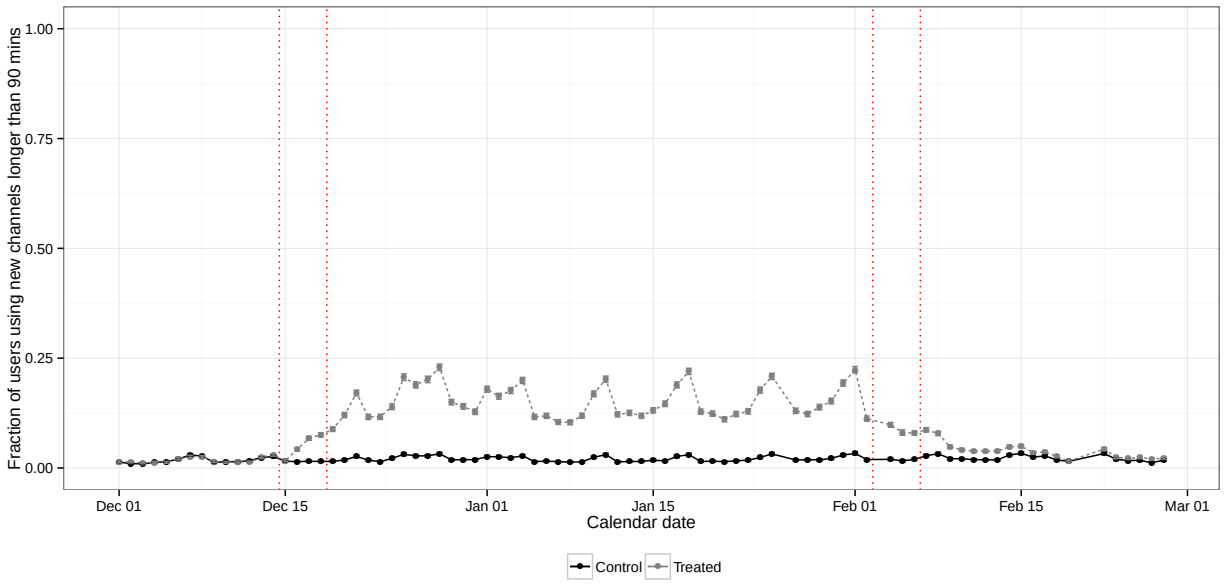



Figure 12: Compliance with treatment over time.

## 5.2 The Effect of Treatment

Tables 6 and 7 repeat the analysis in the previous section for TV and Internet use but now on a daily basis during the experimental period. Table 6 reports the effect of randomized treatment assignment on consumer behavior (i.e. the ITT effect). Column (1) measures the effect of the cinema pack on average household daily television time. On average, across all strata, households assigned to the treatment condition increased their TV usage by 2.3% ($p-value < 0.01$) when compared to households assigned to the control condition. In columns (2) and (3) we report the ITT effect in downloads and uploads. On average, download activity decreased by 3.4% ($p - value < 0.1$) and we cannot reject the null hypothesis that uploads remained unchanged. Given the high variability associated with both download and upload traffic we have less statistical power to identify changes in Internet usage than to identify changes in TV usage.

Table 7 reports the LATE of the cinema pack on TV usage and Internet consumption. The relationship between the instrument and the endogenous variable is presented in table 8. The latter table shows that treatment assignment is not a weak instrument the actual consumption of the cinema pack both overall and for each stratum. As expected, the former

Table 5: Pre-experimental Analysis.

| | *Dependent variable:* | | |
|---|---|---|---|
| | Log(TV Time) | Log(Downloads) | Log(Uploads) |
| | (1) | (2) | (3) |
| Treated | −0.009 | −0.032 | −0.016 |
| | (0.016) | (0.046) | (0.040) |
| Treated * G02 | 0.015 | 0.068 | 0.053 |
| | (0.025) | (0.060) | (0.065) |
| Treated * G03 | −0.008 | −0.037 | −0.027 |
| | (0.026) | (0.065) | (0.063) |
| Treated * G04 | 0.004 | 0.016 | −0.034 |
| | (0.024) | (0.057) | (0.063) |
| G02 | −0.025 | 1.439*** | 2.061*** |
| | (0.018) | (0.042) | (0.045) |
| G03 | −0.001 | 0.829*** | 0.985*** |
| | (0.019) | (0.046) | (0.044) |
| G04 | −0.025 | 1.667*** | 2.538*** |
| | (0.017) | (0.040) | (0.044) |
| Constant | 1.116*** | 5.689*** | 3.426*** |
| | (0.014) | (0.034) | (0.030) |
| Day Dummies | Yes | Yes | Yes |
| Observations | 232,602 | 232,602 | 232,602 |
| $R^2$ | 0.014 | 0.125 | 0.194 |
| Adjusted $R^2$ | 0.014 | 0.125 | 0.194 |
| F Statistic | 159.969*** | 1,654.865*** | 2,800.861*** |
| *Note:* | | | *p<0.1; **p<0.05; ***p<0.01 |
| | | | Cluster Robust Standard Errors in () |

table provides a much stronger effect than the ITT effect on both TV and Internet use. Column (1) shows that on average, across all strata, households that watched the channels in the cinema pack for more than 90 minutes at least once during the experiment increased their daily TV usage by 4.9% ($p - value < 0.01$). Column (2) shows that daily downloads reduced by an average of 7.0% ($p - value < 0.01$) and column (3) shows that daily uploads reduced by 7.1% ($p - value < 0.01$).

Table 6: Intention To Treat (ITT) Effect.

| | *Dependent variable:* | | |
|---|---|---|---|
| | Log(TV Time) | Log(Downloads) | Log(Uploads) |
| | (1) | (2) | (3) |
| Treated | $0.023^{***}$ | $-0.034^{*}$ | $-0.034$ |
| | (0.008) | (0.020) | (0.022) |
| G02 | $-0.005$ | $1.450^{***}$ | $1.993^{***}$ |
| | (0.011) | (0.029) | (0.030) |
| G03 | $-0.010$ | $0.821^{***}$ | $0.974^{***}$ |
| | (0.011) | (0.031) | (0.030) |
| G04 | $-0.020^{*}$ | $1.671^{***}$ | $2.453^{***}$ |
| | (0.011) | (0.027) | (0.029) |
| Constant | $1.095^{***}$ | $5.681^{***}$ | $3.344^{***}$ |
| | (0.012) | (0.027) | (0.025) |
| Day Dummies | Yes | Yes | Yes |
| Observations | 783,787 | 783,787 | 783,787 |
| $R^2$ | 0.015 | 0.119 | 0.180 |
| Adjusted $R^2$ | 0.015 | 0.119 | 0.180 |
| F Statistic | $249.337^{***}$ | $2{,}154.314^{***}$ | $3{,}506.093^{***}$ |
| *Note:* | | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |
| | | | Cluster Robust Standard Errors in () |

## 5.3 The Post-Treatment Effect

We look at whether the treatment had any effect beyond the experimental period during which access to the cinema pack was free on charge. We are particularly interested in knowing whether clients that were given a free sample were more likely to subscribe the cinema pack after the sampling period. Table 9 presents results from a Linear Probability Model (LPM) for the likelihood of subscribing the cinema pack in the month after the experiment. Over all strata we cannot reject the null hypothesis that treatment assignment did not change the likelihood of subscribing the cinema pack after the experiment[3].

With the subscription rate unchanged in response to treatment assignment it was ex-

[3]We use LPM because the covariates of interest included in this model are discrete. In such a case, LPM and associated estimation techniques like two stage least squares (2SLS), are as appropriate for limited dependent variables (binary outcomes) as for other types of dependent variables (Angrist, 2001). In any case, results using logit, probit and probit with instrumental variables yield the same outcomes and are available upon request.

Table 7: Local Average Treatment Effect (LATE).

| | *Dependent variable:* | | |
|---|---|---|---|
| | Log(TV Time) | Log(Downloads) | Log(Uploads) |
| | (1) | (2) | (3) |
| Used | $0.049^{***}$ | $-0.070^{***}$ | $-0.072^{***}$ |
| | (0.008) | (0.020) | (0.022) |
| G02 | $-0.007$ | $1.453^{***}$ | $1.996^{***}$ |
| | (0.011) | (0.029) | (0.030) |
| G03 | $-0.012$ | $0.823^{***}$ | $0.975^{***}$ |
| | (0.011) | (0.031) | (0.030) |
| G04 | $-0.022^{**}$ | $1.674^{***}$ | $2.456^{***}$ |
| | (0.011) | (0.027) | (0.029) |
| Constant | $1.086^{***}$ | $5.693^{***}$ | $3.357^{***}$ |
| | (0.011) | (0.026) | (0.025) |
| Day Dummies | Yes | Yes | Yes |
| Observations | 783,787 | 783,787 | 783,787 |
| $R^2$ | 0.019 | 0.118 | 0.179 |
| Adjusted $R^2$ | 0.019 | 0.118 | 0.179 |
| F Statistic | $301.256^{***}$ | $2{,}135.583^{***}$ | $3{,}497.357^{***}$ |
| *Note:* | | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |
| | | | Cluster Robust Standard Errors in () |

tremely unlikely that we would observe any sustainable change in TV or Internet consumption after the deactivation of the cinema pack offer. Table 10 confirms that as soon as the treatment was removed the behavior of the treated households converged to that of the control group. This highlights that offering the cinema pack for free for a period of 45 days was not effective at inducing long term changes in behavior.

# 6 The Effect of the Cinema Pack on Piracy

In this section we focus on characterizing the impact of the cinema pack offer on piracy and on the behavior of pirates towards media consumption. A major difficulty in our study is to separate the population of pirates from the population of lawful users of the Internet. We determine that a household is pirating if that household is identified in our BitTorrent logs as having downloaded copyrighted content. However, there are several mechanisms that pirates can use to avoid detection. For example, they can use (1) virtual private networks to

Table 8: First stage regression for IV estimation.

| | *Dependent variable:* | | | | |
|---|---|---|---|---|---|
| | Used | | Used * G02 | Used * G03 | Used * G04 |
| | (1) | (2) | (3) | (4) | (5) |
| Treated | $0.482^{***}$ | $0.455^{***}$ | 0.00004 | 0.00004 | 0.00005 |
| | (0.006) | (0.011) | (0.00003) | (0.00002) | (0.00003) |
| Treated * G02 | | $0.030^{*}$ | $0.485^{***}$ | $-0.00002$ | $-0.00003$ |
| | | (0.017) | (0.012) | (0.00003) | (0.00005) |
| Treated * G03 | | $0.052^{***}$ | 0.00001 | $0.507^{***}$ | 0.00001 |
| | | (0.017) | (0.00004) | (0.013) | (0.00005) |
| Treated * G04 | | $0.032^{**}$ | $-0.00002$ | $-0.00002$ | $0.487^{***}$ |
| | | (0.016) | (0.00004) | (0.00003) | (0.011) |
| G02 | $0.046^{***}$ | $0.031^{***}$ | $0.213^{***}$ | 0.00000 | $-0.00000$ |
| | (0.008) | (0.011) | (0.008) | (0.00002) | (0.00003) |
| G03 | $0.024^{***}$ | $-0.002$ | 0.00001 | $0.179^{***}$ | 0.00001 |
| | (0.009) | (0.011) | (0.00003) | (0.009) | (0.00003) |
| G04 | $0.040^{***}$ | $0.024^{**}$ | $-0.00003$ | $-0.00003$ | $0.206^{***}$ |
| | (0.008) | (0.011) | (0.00003) | (0.00002) | (0.008) |
| Constant | $0.179^{***}$ | $0.192^{***}$ | $0.004^{***}$ | $0.001^{**}$ | $0.002^{***}$ |
| | (0.006) | (0.007) | (0.001) | (0.001) | (0.001) |
| Day Dummies | Yes | Yes | Yes | Yes | Yes |
| Observations | 783,787 | 783,787 | 783,787 | 783,787 | 783,787 |
| $R^2$ | 0.253 | 0.253 | 0.533 | 0.544 | 0.527 |
| Adjusted $R^2$ | 0.253 | 0.253 | 0.533 | 0.544 | 0.527 |
| F Statistic | $5{,}419.750^{***}$ | $5{,}116.415^{***}$ | $17{,}177.190^{***}$ | $17{,}954.480^{***}$ | $16{,}772.690^{***}$ |

*Note:* $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01
Cluster Robust Standard Errors in ()

masquerade their true IP address; (2) encrypted streams that evade deep packet inspection; (3) services other than Bit Torrent to share and obtain copyrighted material. In practice, there might be pirates in the population of households that we classify as non-pirates.

For example, in our sample, stratum $G02$ is particularly interesting. On average, in this stratum households download over 3.7 GB/day and upload over 2.3 GB/day. However, we do not find them listed in the Torrent logs that we have access to during the months of April and May 2014. It is possible that these are more sophisticated pirates that are able to masquerade their illegal activities online but, simply put, we do not know if this is true. In the analysis below, we study only households in groups $G03$ and $G04$, who we know for

Table 9: Results for the subscription of the cinema pack after the experiment.

| | *Dependent variable:* | |
|---|---|---|
| | Flg Cinema Pack | |
| | OLS | IV |
| | (1) | (2) |
| Treated | −0.002 | |
| | (0.004) | |
| Used | | −0.004 |
| | | (0.009) |
| G02 | 0.021*** | 0.021*** |
| | (0.005) | (0.005) |
| G03 | 0.009 | 0.009 |
| | (0.006) | (0.006) |
| G04 | 0.018*** | 0.018*** |
| | (0.005) | (0.005) |
| Constant | 0.082*** | 0.083*** |
| | (0.004) | (0.005) |
| Observations | 21,466 | 21,466 |
| $R^2$ | 0.001 | −0.002 |
| Adjusted $R^2$ | 0.001 | −0.002 |
| Residual Std. Error | 0.290 | 0.290 |
| F Statistic | 4.564*** | |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 | |
| | Robust Standard Errors in () | |

sure are pirates (that is, we have definite proof that they did at least one illegal Bit Torrent download in the months of April through May 2014), due to the uncertainty surrounding households in strata $G01$ and $G02$. $G04$ contains pirate households that are heavy Internet users while $G03$ contains pirate households that consume less Internet.

## 6.1 The Pirates' Consumption of TV and Internet

Table 11 reports the ITT effect of the cinema pack on TV time, downloads and uploads. Table 12 reports the corresponding LATE. The former table suggests that households in $G03$ increase TV consumption and decrease both their download and upload activity. However, the ITT cannot reject the null hypothesis that households in $G04$ (pirates that are heavy Internet users) were not affected by treatment assignment. In fact, in columns (4) through (6) we cannot reject the null hypothesis that $Treated{+}Treated{*}G04 = 0$. The corresponding F tests and p-values are $F = 2.466(p - value = 0.116)$, $F = 0.961(p - value = 0.327)$ and $F = 1.368(p - value = 0.242)$ respectively.

Table 10: Impact of treatment assignment after the experimental period.

| | *Dependent variable:* | | |
|---|---|---|---|
| | Log(TV Time) | Log(Downloads) | Log(Uploads) |
| | (1) | (2) | (3) |
| Treated | 0.008 | −0.009 | −0.008 |
| | (0.008) | (0.020) | (0.023) |
| G02 | −0.020* | 1.396*** | 1.973*** |
| | (0.012) | (0.029) | (0.032) |
| G03 | −0.021* | 0.750*** | 0.890*** |
| | (0.012) | (0.031) | (0.030) |
| G04 | −0.034*** | 1.589*** | 2.409*** |
| | (0.011) | (0.027) | (0.031) |
| Constant | 1.381*** | 6.125*** | 3.833*** |
| | (0.011) | (0.026) | (0.025) |
| Day Dummies | Yes | Yes | Yes |
| Observations | 330,900 | 330,900 | 330,900 |
| $R^2$ | 0.090 | 0.120 | 0.186 |
| Adjusted $R^2$ | 0.090 | 0.120 | 0.186 |
| F Statistic | 1,421.788*** | 1,957.675*** | 3,280.016*** |
| *Note:* | | | *p<0.1; **p<0.05; ***p<0.01 |
| | | | Cluster Robust Standard Errors in () |

As before, the LATE estimates are stronger across both groups of households. Table 12 shows that treated households in $G03$ increase TV consumption by 6.8% and decrease their download and upload activity in 21.2% and 16.1%, respectively. In this case we strongly reject the null hypothesis that $Treated + Treated * G04 = 0$: $F = 10.723 (p - value = 0.001)$. In fact, column (4) in this table shows that treated households in $G03$ and $G04$ were equally affected by the cinema pack with respect to their increase TV viewership. Column (5) shows that households in $G04$ reduced their download activity less than households in $G03$. Still, we reject the null hypothesis that $Treated + Treated * G04 = 0$: $F = 4.049 (p - value = 0.0442)$. Therefore, both groups of households reduce their download activity due to the cinema pack but non heavy users of Internet were most affected. Finally column (6) in this table shows that households in $G03$ reduced upload traffic by 16.1% but in this case we cannot reject the null hypothesis that households in $G04$ did not change their upload traffic: $F = 1.947 (p = 0.163)$ for $Treated + Treated * G04 = 0$. Therefore, pirates who are heavy

Internet users keep on uploading content even after being given the cinema pack for free.

Table 11: Effect of Intention To Treat (ITT) of the cinema pack on pirates.

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | Log(TV Time) | Log(Downloads) | Log(Uploads) | Log(TV Time) | Log(Downloads) | Log(Uploads) |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Treated | $0.029^{**}$ | $-0.063^{**}$ | $-0.065^{**}$ | $0.035^{*}$ | $-0.107^{**}$ | $-0.081^{*}$ |
| | (0.012) | (0.026) | (0.033) | (0.018) | (0.043) | (0.046) |
| Treated * G04 | | | | −0.010 | 0.076 | 0.029 |
| | | | | (0.024) | (0.054) | (0.065) |
| G04 | −0.009 | $0.850^{***}$ | $1.479^{***}$ | −0.004 | $0.811^{***}$ | $1.465^{***}$ |
| | (0.012) | (0.027) | (0.032) | (0.017) | (0.037) | (0.045) |
| Constant | $1.081^{***}$ | $6.503^{***}$ | $4.298^{***}$ | $1.078^{***}$ | $6.525^{***}$ | $4.306^{***}$ |
| | (0.016) | (0.030) | (0.034) | (0.017) | (0.034) | (0.037) |
| Day Dummies | Yes | Yes | Yes | Yes | Yes | Yes |
| Observations | 372,080 | 372,080 | 372,080 | 372,080 | 372,080 | 372,080 |
| $R^2$ | 0.015 | 0.061 | 0.102 | 0.015 | 0.061 | 0.102 |
| Adjusted $R^2$ | 0.015 | 0.061 | 0.102 | 0.015 | 0.061 | 0.102 |
| F Statistic | $120.503^{***}$ | $514.622^{***}$ | $902.618^{***}$ | $118.034^{***}$ | $504.763^{***}$ | $883.896^{***}$ |

*Note:* $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01
Cluster Robust Standard Errors in ()

Overall, these results show that increasing the content available on the TV deviates attention from the Internet even across the population of pirates. The treatment works best for pirates that are non heavy users of internet and does not dissuade pirates who are heavy Internet users from uploading as much content as they did before. Table 13 reinforces this idea and provides an estimate for the elasticity of substitution between time spent watching TV and MB of Internet traffic used. To obtain this estimate we instrument the log of TV time with the randomized treatment indicator. Note that Table 11 shows that treatment assignment determines TV usage. This information together with the randomized nature of the assignment makes treatment assignment a good instrument to measure the effect of TV time on downloads. Table 13 shows that when TV time increases by 1% Internet consumption reduces on average by 3.1% and by 1.9% for heavy Internet users and these effects are statistically different. A similar albeit smaller result in magnitude is obtained for upload traffic.

Table 12: Local Average Treatment Effect (LATE) of the cinema pack on pirates.

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | Log(TV Time) | Log(Downloads) | Log(Uploads) | Log(TV Time) | Log(Downloads) | Log(Uploads) |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Used | $0.058^{***}$ | $-0.128^{***}$ | $-0.131^{***}$ | $0.068^{***}$ | $-0.212^{***}$ | $-0.161^{***}$ |
| | (0.012) | (0.026) | (0.033) | (0.018) | (0.044) | (0.046) |
| Used * G04 | | | | $-0.018$ | $0.147^{***}$ | 0.052 |
| | | | | (0.023) | (0.054) | (0.064) |
| G04 | $-0.010$ | $0.852^{***}$ | $1.481^{***}$ | $-0.002$ | $0.787^{***}$ | $1.458^{***}$ |
| | (0.012) | (0.027) | (0.032) | (0.016) | (0.036) | (0.043) |
| Constant | $1.069^{***}$ | $6.528^{***}$ | $4.323^{***}$ | $1.065^{***}$ | $6.564^{***}$ | $4.336^{***}$ |
| | (0.016) | (0.029) | (0.033) | (0.017) | (0.033) | (0.036) |
| Day Dummies | Yes | Yes | Yes | Yes | Yes | Yes |
| Observations | 372,080 | 372,080 | 372,080 | 372,080 | 372,080 | 372,080 |
| $R^2$ | 0.019 | 0.060 | 0.102 | 0.019 | 0.059 | 0.102 |
| Adjusted $R^2$ | 0.019 | 0.060 | 0.102 | 0.019 | 0.059 | 0.102 |
| F Statistic | $148.988^{***}$ | $503.199^{***}$ | $896.827^{***}$ | $145.206^{***}$ | $488.094^{***}$ | $877.401^{***}$ |

*Note:* $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01
Cluster Robust Standard Errors in ()

## 6.2 The Pirates' Use of BiTorrent

The previous section provides strong evidence of substitution between TV usage and Internet consumption. These results are obtained both over all users and over the pirate population. We now study whether the reduction in Internet traffic can be linked to a reduction in piracy activity. To answer this question we build a cross section with households in $G03$ and $G04$ (confirmed pirates prior to the experiment) and we define our dependent variable as binary outcome equal to one if a household appears in our torrent logs during the experiment and zero otherwise. We use a cross section rather than a daily panel because we want to minimize the potential of measurement error associated with the dependent variable. Our concern is not bias in the estimation of the effect of treatment, because assignment was randomized. However, we are concerned that the uncertainty associated to knowing for sure whether a household was sharing files through Bit Torrent will increase our standard errors (Wooldridge, 2010).

Table 14 presents the results of the effect of treatment assignment on the likelihood of

Table 13: Substitution of TV by Internet usage.

| | *Dependent variable:* | | | |
|---|---|---|---|---|
| | Log(Downloads) IV | Log(Uploads) IV | Log(Downloads) IV | Log(Uploads) IV |
| | (1) | (2) | (3) | (4) |
| Log(TV Time) | $-2.211^{***}$ | $-2.259^{***}$ | $-3.088^{***}$ | $-2.346^{***}$ |
| | (0.027) | (0.027) | (0.056) | (0.043) |
| Log(TV Time) * G04 | | | $1.822^{***}$ | $0.182^{***}$ |
| | | | (0.060) | (0.055) |
| G04 | $0.829^{***}$ | $1.458^{***}$ | $-1.454^{***}$ | $1.230^{***}$ |
| | (0.039) | (0.042) | (0.089) | (0.084) |
| Constant | $8.892^{***}$ | $6.739^{***}$ | $10.010^{***}$ | $6.850^{***}$ |
| | (0.054) | (0.056) | (0.086) | (0.071) |
| Day Dummies | Yes | Yes | Yes | Yes |
| Observations | 372,080 | 372,080 | 372,080 | 372,080 |
| $R^2$ | 0.001 | 0.003 | 0.001 | 0.003 |
| Adjusted $R^2$ | 0.001 | 0.003 | 0.001 | 0.003 |
| F Statistic | −5,059.524 | −4,110.529 | −4,958.194 | −3,994.675 |
| *Note:* | | | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |
| | | | | Cluster Robust Standard Errors in () |

torrent activity in the period preceding treatment. Likewise for Table 15 for the experimental period. The even columns in the latter table report the LATE. We use LPM rather than logistic or probit regression models because all explanatory variables are binary[4]. The former table shows that before the experiment there is no evidence that treatment and control groups behaved differently with respect to the probability of using Bit Torrent. This was true for households in $G03$ and $G04$. The latter table shows that during the experiment we cannot reject the null hypothesis that the overall likelihood of engaging in torrent traffic did not change in response to treatment. The same is true for torrent traffic that is specifically related to movies (excluding adult content). However, for households in $G03$, torrent traffic pertaining to TV shows was less likely during the experiment. As before, households in $G04$ are less affected. In fact, we cannot reject the null hypothesis that households in $G04$ continued to use Bit Torrent to share content related to TV shows. The joint significance test for $Treated + Treated \times G04 = 0$ yields $F = 1.961 (p - value = 0.1614)$ and for

[4]In this case there is no advantage from using a binary outcome model over the linear probability. In fact, logit and probit models yield the exact same results as discussed in (Angrist, 2001)

$Used + Used \times G04 = 0$ it yields $F = 1.959 (p - value = .162)$.

Table 14: Effect of treatment assignment on torrent use before the experiment.

| | *Dependent variable:* | | |
|---|---|---|---|
| | Flg. Torrent | Flg. Movie Torrent | Flg. TVShow Torrent |
| | *OLS* | *OLS* | *OLS* |
| | (1) | (2) | (3) |
| Treated | −0.011 | 0.016 | −0.007 |
| | (0.015) | (0.012) | (0.011) |
| Treated*G04 | 0.025 | −0.005 | 0.015 |
| | (0.019) | (0.018) | (0.016) |
| G04 | 0.219*** | 0.191*** | 0.164*** |
| | (0.014) | (0.012) | (0.012) |
| Constant | 0.468*** | 0.192*** | 0.155*** |
| | (0.011) | (0.008) | (0.008) |
| Observations | 10,224 | 10,224 | 10,224 |
| $R^2$ | 0.055 | 0.041 | 0.038 |
| Adjusted $R^2$ | 0.054 | 0.041 | 0.038 |
| Residual Std. Error | 0.477 | 0.453 | 0.425 |
| F Statistic | 196.695*** | 145.386*** | 136.081*** |
| *Note:* | | | *p<0.1; **p<0.05; ***p<0.01 |
| | | | Robust Standard Errors in () |

Table 16 reports results using the logarithm of the average number of torrent streams matched to households on a daily basis. These results confirm our previous findings and show that even if the probability of doing at least one torrent download in the experimental period did not change, the overall intensity of torrent download reduced for households in $G03$. This reduction was particularly relevant for content related to TV shows.

Overall, our results suggest that pirates do not stop pirating even when more content becomes available conveniently, without schedule limitations and for free, as in the cinema pack. We also show evidence that there is substitution between TV and Internet usage and that some of the substitution originates in the reduction of torrent downloads. Still, overall, pirate households in the treatment and control groups keep using Bit Torrent.

Finally, the fact that recent TV shows are the main driver of the reduction in torrent intensity speaks to the importance of content recency on the pirates' activity. As described in section 2.1 the cinema pack included TV channels that broadcast both movies and TV shows. But in what pertains to TV shows, the content available has been typically aired only a couple of days ago in the US. Movies made available through the cinema pack channels

Table 15: Effect of treatment assignment on torrent use during the experiment.

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | Flg. Torrent | | Flg. Movie Torrent | | Flg. TVShow Torrent | |
| | *OLS* | *instrumental variable* | *OLS* | *instrumental variable* | *OLS* | *instrumental variable* |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Treated | −0.022 | | −0.022 | | −0.033*** | |
| | (0.015) | | (0.014) | | (0.013) | |
| Treated*G04 | 0.023 | | 0.021 | | 0.051*** | |
| | (0.018) | | (0.019) | | (0.018) | |
| Used | | −0.045 | | −0.045 | | −0.069*** |
| | | (0.030) | | (0.029) | | (0.026) |
| Used * G04 | | 0.047 | | 0.043 | | 0.107*** |
| | | (0.037) | | (0.040) | | (0.038) |
| G04 | 0.181*** | 0.173*** | 0.187*** | 0.180*** | 0.153*** | 0.134*** |
| | (0.012) | (0.018) | (0.014) | (0.020) | (0.013) | (0.018) |
| Constant | 0.649*** | 0.657*** | 0.339*** | 0.346*** | 0.238*** | 0.250*** |
| | (0.010) | (0.014) | (0.010) | (0.014) | (0.009) | (0.013) |
| Observations | 10,224 | 10,224 | 10,224 | 10,224 | 10,224 | 10,224 |
| $R^2$ | 0.048 | 0.045 | 0.039 | 0.037 | 0.036 | 0.034 |
| Adjusted $R^2$ | 0.048 | 0.045 | 0.038 | 0.037 | 0.036 | 0.034 |
| Residual Std. Error | 0.423 | 0.424 | 0.487 | 0.487 | 0.460 | 0.460 |
| F Statistic | 172.934*** | | 137.357*** | | 128.375*** | |

*Note:* *p<0.1; **p<0.05; ***p<0.01
Robust Standard Errors in ()

are usually within the initial DVD release window (thus about 1-1.5 years old).

## 6.3 Piracy at the Content Level

(Danaher and Smith, 2014) concluded that removing NBC content from the iTunes store in December 2007 produced a large increase in piracy levels of that same content. These results hinted at a substitution effect between Internet downloads and the iTunes store and suggested that legal distribution channels of content could dissuade users from engaging in piracy. In the current paper, we show that when households are given more TV content for free, they substitute away from the Internet towards the TV where they can obtain similar content. The novelty in our study is that we also show that overall pirates do not shutdown their torrent activity. Our results show that pirates reduce the number of downloads of content that is related to what they have available for free, but they continue downloading other content that they would otherwise have to pay for. This is a gloomier perspective of the impact of channel competition on piracy. Channel competition shifts the focus of the

Table 16: Effect of treatment assignment on number of torrent streams matched to households included in the experiment.

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | log(Torrent) | | log(Movie Torrent) | | log(TVShow Torrent) | |
| | *OLS* | *instrumental variable* | *OLS* | *instrumental variable* | *OLS* | *instrumental variable* |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Treated | $-0.030^{*}$ | | $-0.005$ | | $-0.013^{**}$ | |
| | (0.016) | | (0.005) | | (0.006) | |
| Treated*G04 | $0.046^{*}$ | | 0.011 | | 0.016 | |
| | (0.027) | | (0.009) | | (0.010) | |
| Used | | $-0.062^{*}$ | | $-0.011$ | | $-0.027^{**}$ |
| | | (0.033) | | (0.010) | | (0.012) |
| Used * G04 | | $0.095^{*}$ | | 0.023 | | 0.034 |
| | | (0.058) | | (0.020) | | (0.021) |
| G04 | $0.389^{***}$ | $0.372^{***}$ | $0.080^{***}$ | $0.075^{***}$ | $0.061^{***}$ | $0.055^{***}$ |
| | (0.019) | (0.028) | (0.006) | (0.009) | (0.007) | (0.010) |
| Constant | $0.325^{***}$ | $0.336^{***}$ | $0.055^{***}$ | $0.057^{***}$ | $0.055^{***}$ | $0.060^{***}$ |
| | (0.012) | (0.017) | (0.003) | (0.005) | (0.005) | (0.006) |
| Observations | 10,224 | 10,224 | 10,224 | 10,224 | 10,224 | 10,224 |
| $R^2$ | 0.072 | 0.071 | 0.027 | 0.026 | 0.016 | 0.016 |
| Adjusted $R^2$ | 0.071 | 0.070 | 0.027 | 0.026 | 0.016 | 0.015 |
| Residual Std. Error | 0.732 | 0.733 | 0.254 | 0.254 | 0.270 | 0.270 |
| F Statistic | $263.178^{***}$ | | $93.942^{***}$ | | $55.805^{***}$ | |

*Note:* $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01
Robust Standard Errors in ()
Constant 1 added to all torrent measures

pirates' attention but does not dissuade them from pirating.

We also note that the context of the present paper is distinct from the one studied in Danaher and Smith (2014). Not only the markets are different but Danaher and Smith (2014) studied the removal of content from a online digital store, while we study the introduction of free TV content in households. More importantly we note that Danaher and Smith (2014) provide results from an analysis at the content level while we provide results at the household level, which we believe provides a richer picture of the behavior of pirates. Still, we find that our results are in line with those reported in Danaher and Smith (2014). To see this, we aggregate our data at the content level. We focus on TV shows that are broadcasted in the cinema pack and that were downloaded by the population of pirate households in our sample prior to the experimental period. We have one observation per pirate stratum, treatment assignment and content ID and we aggregate torrent streams in the periods before and

during the experiment. This new dataset allows us to run a differences in differences model similar to the one presented in (Danaher and Smith, 2014). Furthermore, in our case this model measures the causal effect of treatment due to our randomized schedule of treatment assignment. Therefore, we use these data to measure, at content level, the magnitude of the reduction in download streams of the content that the cinema pack made available to treated households.

Table 17: Download of TV shows before and during the experiment.

| | *Dependent variable:* | |
|---|---|---|
| | Log(Number of Streams) | |
| | OLS | FE |
| | (1) | (2) |
| Treated | 0.215 | |
| | (0.197) | |
| Treated * During | $-0.522^{*}$ | $-0.522^{**}$ |
| | (0.289) | (0.225) |
| During | $-0.041$ | $-0.041$ |
| | (0.199) | (0.150) |
| G04 | $1.471^{***}$ | |
| | (0.145) | |
| Constant | $1.886^{***}$ | |
| | (0.319) | |
| IMDb ID | Yes | No |
| Observations | 128 | 128 |
| $R^2$ | 0.895 | 0.164 |
| Adjusted $R^2$ | 0.876 | 0.080 |
| Residual Std. Error | 0.752 | |
| F Statistic | $48.437^{***}$ | $6.096^{***}$ |
| *Note:* | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 | |
| | Robust standard errors in () | |

Table 17 reports our results. Column (1) reports the differences in differences for the amount of torrent streams pertaining to TV shows provided by the cinema pack during the period of our study. Column (2) provides estimates using the within panel data estimator for the same effect. Column (1) also shows that there is no reason to believe that households in the treatment and control groups downloaded different amounts of episodes of these TV shows episodes prior to the experiment. However, both column (1) and (2) show that, on average, households in the treatment group were logged in 52% fewer streams downloading/uploading from the set of TV shows available in the cinema pack offer.These results suggest a dramatic reduction in piracy activity in response to similar content being made

available free of charge and conveniently on the household TV. However, pirate households do not stop pirating. Rather, they pirate other content.

# 7 Conclusion

We partner with a major multinational telecommunications provider to analyze the effect of channel competition on the consumption of media by digital pirates. For a period of 45 days we activated for free a set of new TV channels dedicated to movies and TV shows, called the cinema pack, to a stratified sample of randomly selected consumers. Consumer strata were determined resorting to a state-of-the-art machine learning algorithm used to partition the space of households into four groups: (1) clients predicted as non pirates and confirmed non pirates; (2) clients predicted as pirates and confirmed not pirates; (3) clients predicted as non pirates and confirmed pirates and (4) clients predicted and confirmed pirates. A household is confirmed pirate when its IP address appears in Bit Torrent logs. Our stratified designed allowed us to study the impact of TV consumption on Internet use for a known population of pirates as well as for the general population of households of our industrial partner with similar statistical power.

We estimate that overall, households watching the new channels exogenously awarded as a gift, increased daily TV consumption by 4.9%, measured in terms of minutes watched, and reduced their daily downloaded and uploaded activities by 7.0%, measured in terms of amount of traffic exchanged with the Internet. In particular, we show that the sub-population of confirmed pirates was also affected by our treatment. On average, across the pirate population, households increased their TV consumption by 5.8% and reduced their download activity by 12.8%. A detailed analysis of pirates reveals that households that are heavy Internet users increase their TV consumption after treatment as much as the remaining population of households, but reduced their download traffic about half as much as non heavy Internet users.

In addition, we find no change in the probability of households using Bit Torrent during the experimental period. We match households in our sample to a dataset of Bit Torrent streams provided by a third party to reach this conclusion. However, we find evidence of a

reduction in the overall number of streams that were logged for households in the treatment group and we find that such reduction is driven by fewer streams related to TV shows. Nevertheless, these changes affect only the pirate population of non heavy users of Internet. We also show that the number of Bit Torrent streams in our torrent logs associated to TV shows offered as part of the cinema pack reduced by as much as 52% across the households in the treatment group when compared to the the households in the control group. For this analysis, we aggregate our logs of torrent streams at content level so that we can focus on TV shows that are included in the cinema pack and that were downloaded by our sample of households before the experiment started. These results speak to the wide range of content that pirates obtain from Bit Torrent and to the unlikely success of channel competition as a deterrent for piracy as a whole.

Our results suggest that content producers can benefit from channel competition because once their content is available in more platforms it will be less likely obtained from illegal sources. However, for policy makers our results suggest that channel competition is unlikely to stop piracy. As content availability increases pirates reallocate their portfolio of downloads and uploads but not seem to stop sharing content with their peers. Finally, we also show that while giving away content for free reduces piracy temporarily, this effect ceasses once users are again asked to pay for the content. In our setting the behavior of treated households converged to that of control households in the post-experiment period. Unlike the music industry where free sampling of songs has been argued to promote album sales, in our case we do not find evidence that the cinema pack sold more after the sampling period.

# References


Angrist, J. (1991). Instrumental variables estimation of average treatment effects in econometrics and epidemiology.

Angrist, J. D. (2001). Estimation of limited dependent variable models with dummy endogenous regressors. *Journal of business & economic statistics*, 19(1).

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455.

Assmann, S. F., Pocock, S. J., Enos, L. E., and Kasten, L. E. (2000). Subgroup analysis and other (mis) uses of baseline data in clinical trials. *The Lancet*, 355(9209):1064–1069.

Bai, J. and Waldfogel, J. (2012). Movie piracy and sales displacement in two samples of chinese consumers. *Information Economics and Policy*, 24(34):187–196.

Bhattacharjee, S., Gopal, R. D., Lertwachara, K., Marsden, J. R., and Telang, R. (2007). The effect of digital sharing technologies on music markets: A survival analysis of albums on ranking charts. *Management Science*, 53(9):1359–1374.

Bounie, D., Bourreau, M., Waelbroeck, P., and others (2006). Piracy and the demand for films: Analysis of piracy behavior in french universities. *Review of Economic Research on Copyright Issues*, 3(2):15–27.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Bronnenberg, B. J., Dube, J.-P., and Mela, C. F. (2010). Do digital video recorders influence sales? *Journal of Marketing Research*, 47(6):998–1010.

Chellappa, R. K. and Shivendu, S. (2005). Managing piracy: Pricing and sampling strategies for digital experience goods in vertically segmented markets. *Information Systems Research*, 16(4):400–417.

Conover, W. J. and Conover, W. (1980). Practical nonparametric statistics.

Danaher, B., Dhanasobhon, S., Smith, M. D., and Telang, R. (2010). Converting pirates without cannibalizing purchasers: The impact of digital distribution on physical sales and internet piracy. *Marketing Science*, 29(6):1138–1151.

Danaher, B. and Smith, M. D. (2014). Gone in 60 seconds: The impact of the megaupload shutdown on movie sales. *International Journal of Industrial Organization*, 33:1–8.

Danaher, B., Smith, M. D., and Telang, R. (2013). Piracy and copyright enforcement mechanisms. Working Paper 19150, National Bureau of Economic Research.

Frangakis, C. E. and Rubin, D. B. (1999). Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes. *Biometrika*, 86(2):365–379.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Gupta, S. K. (2011). Intention-to-treat concept: A review. *Perspectives in Clinical Research*, 2(3):109–112.

Hearst, M. A., Dumais, S. T., Osman, E., Platt, J., and Scholkopf, B. (1998). Support vector machines. *Intelligent Systems and their Applications, IEEE*, 13(4):18–28.

Hollis, S. and Campbell, F. (1999). What is meant by intention to treat analysis? survey of published randomised controlled trials. *BMJ*, 319(7211):670–674.

Kernan, W. N., Viscoli, C. M., Makuch, R. W., Brass, L. M., and Horwitz, R. I. (1999). Stratified randomization for clinical trials. *Journal of clinical epidemiology*, 52(1):19–26.

Kuhn, M. (2008). Building predictive models in r using the caret package. *Journal of Statistical Software*, 28(5):1–26.

Levy, P. S. and Lemeshow, S. (2013). *Sampling of populations: methods and applications*. John Wiley & Sons.

Liebowitz, S. J. (2008). Research note-testing file sharing's impact on music album sales in cities. *Management Science*, 54(4):852–859.

McCaffrey, D. F., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramchand, R., and Burgette, L. F. (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in medicine*, 32(19):3388–3414.

Oberholzer-Gee, F. and Strumpf, K. (2007). The effect of file sharing on record sales: An empirical analysis. *Journal of political economy*, 115(1):1–42.

Peitz, M. and Waelbroeck, P. (2006). Piracy of digital products: A critical review of the theoretical literature. *Information Economics and Policy*, 18(4):449–476.

Reavis Conner, K. and Rumelt, R. P. (1991). Software piracy: an analysis of protection strategies. *Management Science*, 37(2):125–139.

Resnick, P. and Varian, H. R. (1997). Recommender systems. *Communications of the ACM*, 40(3):5658.

Rob, R. and Waldfogel, J. (2006). Piracy on the high cs: Music downloading, sales displacement, and social welfare in a sample of college students. *Journal of Law and Economics*, 49(1):29–62.

Rob, R. and Waldfogel, J. (2007). Piracy on the silver screen. *The Journal of Industrial Economics*, 55(3):379–395.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

Simon, R. (1979). Restricted randomization designs in clinical trials. *Biometrics*, pages 503–512.

Sinha, R. K., Machado, F. S., and Sellman, C. (2010). Don't think twice, it's all right: music piracy and pricing in a drm-free environment. *Journal of Marketing*, 74(2):40–54.

Smith, M. D. and Telang, R. (2009). Competing with free: the impact of movie broadcasts on DVD sales and internet piracy 1. *MIS Quarterly*, 33(2):321338.

Suykens, J. A. and Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural processing letters*, 9(3):293–300.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

Waldfogel, J. (2010). Music file sharing and sales displacement in the itunes era. *Information Economics and Policy*, 22(4):306–314.

Wilbur, K. C. (2008). How the digital video recorder (dvr) changes traditional television advertising. *Journal of Advertising*, 37(1):143–149.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. The MIT Press, second edition edition.

Zentner, A. (2006). Measuring the effect of file sharing on music purchases. *Journal of Law and Economics*, 49(1):63–90.

# A Model to predict piracy

We use the pre-experimental dataset that covers the period April - May 2015 to build a classification algorithm that allows for stratifying our household sample according to observable features that correlate with piracy — in this case, Bit Torrent downloads. We use the Caret framework to train and evaluate the performance of different machine learning algorithms (Kuhn, 2008) on their ability to predict if a household will be present in the torrent logs that we had access to. All algorithms that we used were trained and tested using 5 fold cross validation repeated 10 times. The outcome of this analysis is depicted in Figure 13. The Figure shows that most models fit our data well. In particular, the area under the curve (AUC) which measures the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one is near or above 80% across the algorithms that we use. This is the usual rule of thumb criteria to assume that a model is a good for predictive purposes.

Variable selection is an integral part of gradient boosted model trees (GBM) (Friedman, 2001), random forests (RFOREST) (Breiman, 2001) and classification and regression trees algorithm (RPART) (Breiman et al., 1984). For Support Vector Machines with radial Kernel (SVM) (Hearst et al., 1998; Suykens and Vandewalle, 1999) and for the Logit model, feature selection is a separate step. To train these models we selected features to include in the fit process using LASSO (Tibshirani, 1996). Taking into consideration that GBM had better scores in every performance metric we opted to use this algorithm as the basis of our stratification strategy. Using the output of the trained GBM model, the population classifier was constructed such that households with GBM scores above 50% were marked as pirates, while households with GBM scores below this threshold were marked non-pirates. Figure 14 plots the ROC curve of our trained model. The black dot identifies the classifier that we used to stratify households.

After classifying households we look for whether they show up in the torrent logs. This allows for creating the four strata described in section 3.1. Households that were found in the torrent logs were confirmed pirates. Otherwise they were marked as non pirates. Finally, Figure 15 summarizes the features that GBM uses to classify households as pirates and non

pirates. This figure shows that upload traffic is the main determinant followed by how long ago households subscribed Internet service and by whether they are legacy or up-to-date.

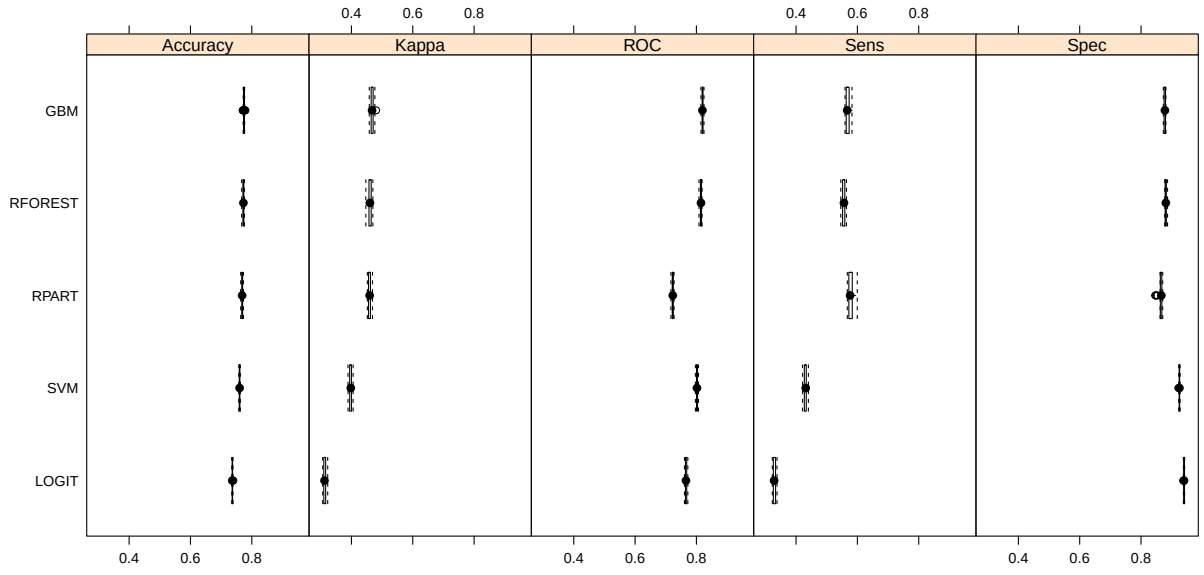



Figure 13: Machine learning algorithms tested.

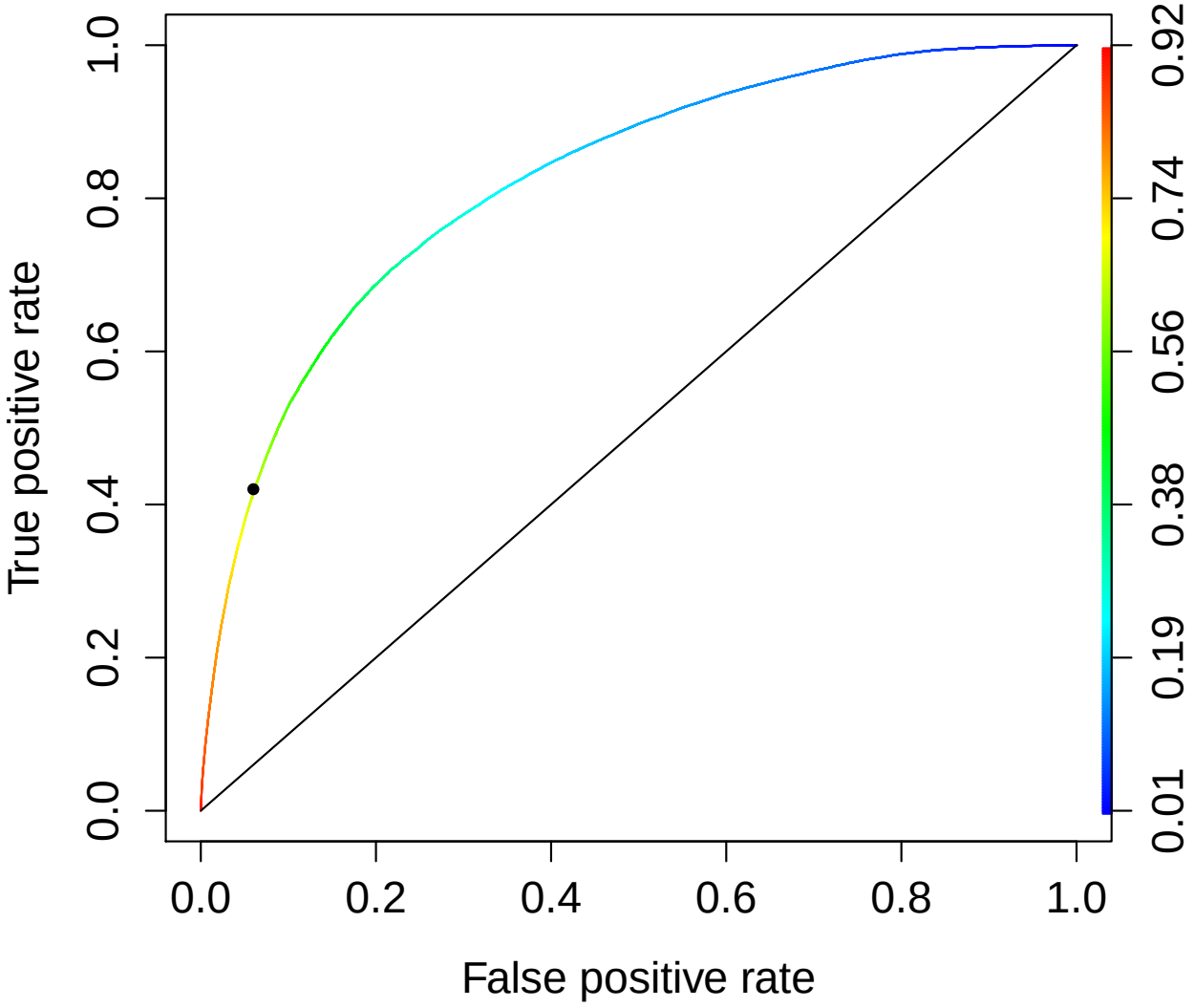



Figure 14: ROC chart for gradient boosted model trees algorithm.

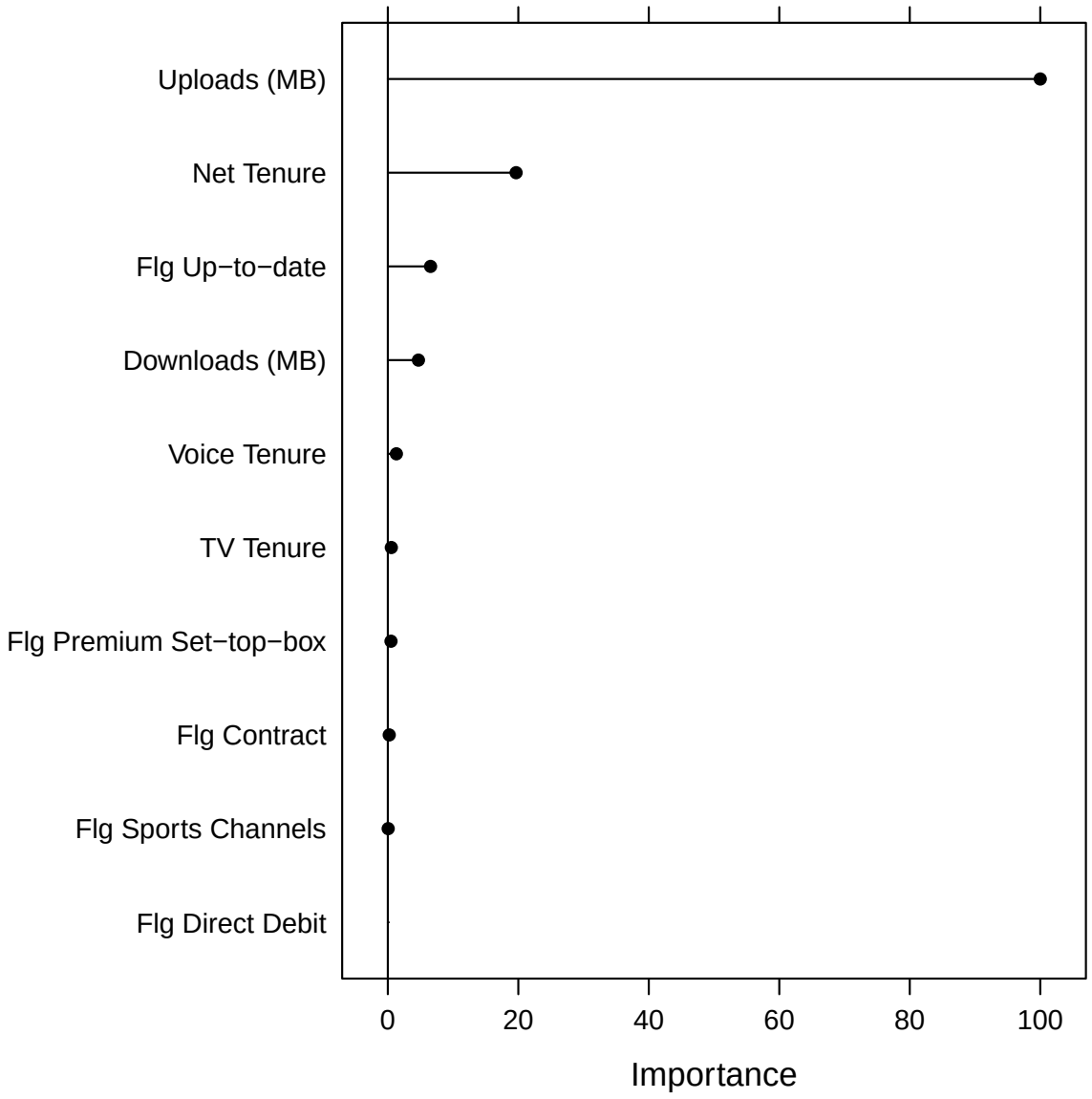



Figure 15: Variable importance plot for the gradient boosted model tree algorithm.

# B Stratum Balance

Table 18: $G01$ - Non pirate and predicted non pirate

| | tx.mn | tx.sd | ct.mn | ct.sd | std.eff.sz | stat | p | ks | ks.pval |
|---|---|---|---|---|---|---|---|---|---|
| Pirate Score | 0.181 | 0.135 | 0.183 | 0.135 | -0.013 | -0.492 | 0.623 | 0.015 | 1 |
| Tenure (TV months) | 86.878 | 60.930 | 87.588 | 60.957 | -0.012 | -0.454 | 0.650 | 0.015 | 1 |
| Tenure (Internet months) | 56.289 | 35.253 | 57.165 | 36.007 | -0.025 | -0.959 | 0.338 | 0.016 | 1 |
| Tenure (Telephone months) | 49.037 | 21.923 | 49.151 | 21.792 | -0.005 | -0.204 | 0.839 | 0.013 | 1 |
| Active Contract | 0.824 | 0.381 | 0.818 | 0.386 | 0.016 | 0.633 | 0.527 | 0.006 | 1 |
| Download (MB per day) | 1,217.236 | 1,913.677 | 1,242.850 | 2,124.813 | -0.013 | -0.494 | 0.621 | 0.028 | 1 |
| Upload (MB per day) | 269.733 | 1,531.998 | 229.026 | 862.270 | 0.033 | 1.280 | 0.200 | 0.021 | 1 |
| TV Channels zapped | 11.676 | 7.940 | 11.518 | 8.036 | 0.020 | 0.773 | 0.439 | 0.019 | 1 |
| TV Time - New Chan (Hours per day) | 0.055 | 0.256 | 0.056 | 0.240 | -0.003 | -0.125 | 0.901 | 0.014 | 1 |
| TV Time - All Chans (Hours per day) | 4.393 | 2.403 | 4.474 | 2.476 | -0.033 | -1.289 | 0.198 | 0.026 | 1 |
| VoD streams | 0.927 | 1.074 | 0.943 | 1.113 | -0.015 | -0.591 | 0.555 | 0.016 | 1 |
| VoD TV time shifts | 0.021 | 0.167 | 0.030 | 0.173 | -0.054 | -2.107 | 0.035 | 0.023 | 1 |
| VoD title page view | 0.092 | 0.393 | 0.115 | 0.447 | -0.055 | -2.136 | 0.033 | 0.021 | 1 |
| Number of days | 10.065 | 3.471 | 10.059 | 3.464 | 0.002 | 0.074 | 0.941 | 0.011 | 1 |

Table 19: $G02$ - Non pirate and predicted pirate

| | tx.mn | tx.sd | ct.mn | ct.sd | std.eff.sz | stat | p | ks | ks.pval |
|---|---|---|---|---|---|---|---|---|---|
| Pirate Score | 0.763 | 0.065 | 0.764 | 0.064 | -0.003 | -0.121 | 0.904 | 0.022 | 0.690 |
| Tenure (TV months) | 99.347 | 55.084 | 99.905 | 55.204 | -0.010 | -0.362 | 0.717 | 0.016 | 0.918 |
| Tenure (Internet months) | 75.900 | 34.593 | 75.829 | 33.590 | 0.002 | 0.074 | 0.941 | 0.022 | 0.660 |
| Tenure (Telephone months) | 60.441 | 17.468 | 60.789 | 17.241 | -0.020 | -0.717 | 0.474 | 0.016 | 0.912 |
| Active Contract | 0.808 | 0.394 | 0.809 | 0.393 | -0.004 | -0.151 | 0.880 | 0.002 | 1 |
| Download (MB per day) | 3,770.558 | 4,935.325 | 3,738.642 | 5,185.037 | 0.006 | 0.225 | 0.822 | 0.019 | 0.786 |
| Upload (MB per day) | 2,331.399 | 5,367.296 | 2,229.316 | 4,445.092 | 0.021 | 0.739 | 0.460 | 0.015 | 0.929 |
| TV Channels zapped | 11.809 | 8.427 | 12.094 | 8.339 | -0.034 | -1.213 | 0.225 | 0.031 | 0.432 |
| TV Time - New Chan (Hours per day) | 0.061 | 0.270 | 0.074 | 0.312 | -0.046 | -1.652 | 0.099 | 0.015 | 0.933 |
| TV Time - All Chans (Hours per day) | 4.402 | 2.601 | 4.390 | 2.429 | 0.005 | 0.181 | 0.857 | 0.030 | 0.460 |
| VoD streams | 0.983 | 1.112 | 0.950 | 1.085 | 0.030 | 1.064 | 0.288 | 0.031 | 0.443 |
| VoD TV time shifts | 0.024 | 0.203 | 0.025 | 0.246 | -0.001 | -0.033 | 0.973 | 0.004 | 1 |
| VoD title page view | 0.083 | 0.509 | 0.083 | 0.373 | 0.001 | 0.031 | 0.976 | 0.011 | 0.997 |
| Number of days | 9.931 | 3.581 | 10.043 | 3.466 | -0.032 | -1.128 | 0.259 | 0.020 | 0.762 |

Table 20: $G03$ - Pirate and predicted non pirate

| | tx.mn | tx.sd | ct.mn | ct.sd | std.eff.sz | stat | p | ks | ks.pval |
|---|---|---|---|---|---|---|---|---|---|
| Pirate Score | 0.304 | 0.151 | 0.311 | 0.150 | -0.044 | -1.456 | 0.145 | 0.033 | 0.186 |
| Tenure (TV months) | 93.136 | 57.657 | 91.215 | 55.605 | 0.034 | 1.111 | 0.267 | 0.026 | 0.451 |
| Tenure (Internet months) | 66.586 | 34.231 | 66.391 | 34.349 | 0.006 | 0.186 | 0.852 | 0.018 | 0.868 |
| Tenure (Telephone months) | 54.911 | 18.944 | 55.335 | 18.639 | -0.023 | -0.740 | 0.459 | 0.022 | 0.643 |
| Active Contract | 0.802 | 0.398 | 0.793 | 0.405 | 0.024 | 0.775 | 0.439 | 0.010 | 1 |
| Download (MB per day) | 2,193.011 | 3,156.323 | 2,212.563 | 3,544.914 | -0.006 | -0.191 | 0.849 | 0.032 | 0.223 |
| Upload (MB per day) | 806.465 | 2,448.565 | 722.401 | 2,329.261 | 0.035 | 1.153 | 0.249 | 0.035 | 0.134 |
| TV Channels zapped | 11.640 | 8.179 | 11.948 | 7.965 | -0.038 | -1.249 | 0.212 | 0.040 | 0.064 |
| TV Time - New Chan (Hours per day) | 0.063 | 0.280 | 0.067 | 0.311 | -0.014 | -0.452 | 0.651 | 0.020 | 0.766 |
| TV Time - All Chans (Hours per day) | 4.375 | 2.505 | 4.464 | 2.512 | -0.035 | -1.150 | 0.250 | 0.030 | 0.264 |
| VoD streams | 1.020 | 1.180 | 1.010 | 1.108 | 0.009 | 0.286 | 0.775 | 0.018 | 0.865 |
| VoD TV time shifts | 0.021 | 0.151 | 0.022 | 0.163 | -0.007 | -0.235 | 0.814 | 0.011 | 1 |
| VoD title page view | 0.081 | 0.341 | 0.088 | 0.396 | -0.017 | -0.553 | 0.580 | 0.011 | 0.999 |
| Number of days | 9.982 | 3.557 | 10.078 | 3.530 | -0.027 | -0.887 | 0.375 | 0.029 | 0.324 |

Table 21: $G04$ - Pirate and predicted pirate

| | tx.mn | tx.sd | ct.mn | ct.sd | std.eff.sz | stat | p | ks | ks.pval |
|---|---|---|---|---|---|---|---|---|---|
| Pirate Score | 0.757 | 0.080 | 0.757 | 0.082 | 0.001 | 0.036 | 0.971 | 0.023 | 1 |
| Tenure (TV months) | 96.826 | 56.475 | 94.636 | 56.055 | 0.039 | 1.495 | 0.135 | 0.026 | 1 |
| Tenure (Internet months) | 71.509 | 33.653 | 70.442 | 33.810 | 0.032 | 1.214 | 0.225 | 0.024 | 1 |
| Tenure (Telephone months) | 58.385 | 16.955 | 57.846 | 17.373 | 0.031 | 1.206 | 0.228 | 0.024 | 1 |
| Active Contract | 0.796 | 0.403 | 0.803 | 0.397 | -0.019 | -0.716 | 0.474 | 0.007 | 1 |
| Download (MB per day) | 4,446.922 | 5,292.187 | 4,522.641 | 5,917.408 | -0.013 | -0.518 | 0.605 | 0.030 | 1 |
| Upload (MB per day) | 3,037.274 | 5,766.615 | 3,195.712 | 6,039.995 | -0.027 | -1.030 | 0.303 | 0.024 | 1 |
| TV Channels zapped | 11.900 | 8.621 | 11.622 | 8.056 | 0.033 | 1.279 | 0.201 | 0.026 | 1 |
| TV Time - New Chan (Hours per day) | 0.070 | 0.284 | 0.069 | 0.284 | 0.005 | 0.211 | 0.833 | 0.011 | 1 |
| TV Time - All Chans (Hours per day) | 4.362 | 2.495 | 4.381 | 2.526 | -0.007 | -0.285 | 0.776 | 0.016 | 1 |
| VoD streams | 0.967 | 1.108 | 0.998 | 1.105 | -0.028 | -1.057 | 0.290 | 0.048 | 1 |
| VoD TV time shifts | 0.022 | 0.162 | 0.024 | 0.217 | -0.008 | -0.289 | 0.773 | 0.006 | 1 |
| VoD title page view | 0.085 | 0.361 | 0.079 | 0.361 | 0.018 | 0.677 | 0.499 | 0.008 | 1 |
| Number of days | 9.933 | 3.556 | 10.003 | 3.507 | -0.020 | -0.755 | 0.450 | 0.014 | 1 |