Does Online Piracy make Computers Insecure? Evidence from Panel Data[1]

Rahul Telang

Carnegie Mellon University

Abstract

Using a panel data of more than 250 users for over a year, we examine the relation between user visits to content infringing sites and the number of malware found on their machines. A key aspect of our data is that the users are observed in the real world, at their homes, and the data is captured unobtrusively via background sensors. Thus the data provides an unbiased insight into how users navigate infringing sites and how it affects the health of their computers. We are able to classify user activities into various categories including visits to infringing sites. We then estimate a within-user model and find that when users spend more time on infringing sites in a given month, they are also more likely to download malware files on their machines in the same time period. In particular, we estimate that doubling the time spent on infringing sites leads to 20 percent increase in total malware files and 20 percent increase in malware files after removing potential adware. We also find no evidence that users who visit infringing sites more take more precautions. In particular, users visiting infringing sites are less (not more) likely to install an antivirus (AV) software.

[1] This material is based upon work supported by the NSA under contract# H9823014C0140. Rahul Telang also acknowledges generous support of IDEA (Initiative of Digital Entertainment Analytics) and help provided by Sarah Pearman, Jeremy Thomas, Yaswanth Jeganathan, Hyunjae Lee.

Introduction

The impact of online piracy on firm revenues and consumer behavior has been a topic of great debate for industry, policy makers and academics. Since the growth of Napster in late 90s, technology has made it easier for users to find, share, and consume copyrighted content for free. This has led to a variety of industry and policy efforts. Early part of the literature in management and economics focuses on how to measure the impact of piracy on firm revenues. Over the last decade or so, the literature has well documented that piracy has had an adverse impact on firm revenues and this has been true for both movies and music industry (see Smith and Telang 2014 for the review of this literature).

However, the other side of piracy debate is about consumer welfare. Even if piracy hurts industry revenues, it can be argued that consumers benefit from free consumption. Piracy also makes content available to a large number of users who otherwise would not consume that content legally. Piracy might also act a competitive force and keep prices low. In short, while piracy reduces producer profits, by increasing demand for content, the consumer surplus unambiguously increases. This has led to a policy debate on trade-off of anti-piracy regulations which restricts access to infringing content.

Much of the deleterious effects of piracy are understood to be in the form of hampering innovation. So, while piracy benefits consumers by making content available cheaply (in many cases for free), this benefit accrues only in the short run. In the long run, decreasing revenues for the industry would affect their ability to invest in content production and all else equal the quantity and the quality of content will decline. Thus, while piracy has a positive impact on demand, the supply side will be adversely affected in the long run. And, this will hurt consumers affecting the total welfare. But this potentially will not be seen in the data unless the adverse effects of piracy are very large and happen in a short period. The empirical research on the effects of piracy on supply is sparse and few available studies in recorded music industry find that the number of

new products has not declined even after the growth in piracy (Handke, 2012; Oberholzer-Gee & Strumpf, 2007; Waldfogel, 2012). But Telang and Waldfogel (2015) using the data on VCR entry in India find that large scale piracy did have a sharp negative impact on movie production in India.

There is one more important aspect of piracy which has not been as well studied. There is a cottage industry of websites hosting infringing content. These sites make it easy for users to search and find infringing content and download. Napster was the first such effort where users could search for songs of their liking and find a willing host to download from. The technology has evolved over a period of time with BitTorrent peer to peer technology becoming a popular way for users to share their content.[2] Various streaming websites now do the same task. Current controversy on stream ripping websites like YouTube-mp3.org points to the fast moving technology for sharing infringing content. A key effort of the content industry and regulators in many countries has been to either block or shut down popular sites which facilitate content search and download. The blocking of Pirate Bay, Megaupload, Kino (See Danaher et al 2016, Danaher and Smith 2014 and Peukert et al 2015) and many such popular websites show that much of the piracy is facilitated by these websites. Of course, there are many more sites where users access infringing content from. According to Google, in February 2016, Google received 75 million copyright takedown requests.[3]

For many of these sites, providing content for free is not an altruistic motive. These sites make money in variety of ways – predominantly from advertising. While most sites possibly make little or no money, some of these sites can attract large Internet traffic, and are attractive to advertisers as well. According to a study by Digital Citizens Alliance (Digital Citizen Alliance, 2015), the aggregate ad revenues for infringing sites is as high as $227 million in 2014. Many of these ads though can carry malware, adware and other questionable tracking software. When users click on the links, these software gets downloaded as well. Some other sites are purely malicious. They host infringing content

[2] See https://en.wikipedia.org/wiki/Legal_issues_with_BitTorrent
[3] https://www.theverge.com/2016/3/7/11172516/google-takedown-requests-75-million

to attract the users and then infect their machines with malware, Trojans and other tracking software which is used for variety of security breaches to turn into money (for example ransomware, or controlling user machine remotely as bots to launch attacks on other networks, send spams and so on). Thus there is a belief that while users may download content for free and enjoy those benefits in the short run, the possible security risk outweighs the short term benefits. So the widespread piracy is likely to not only hurt innovation in the long run, it will also lead to adverse security outcomes for users in the short run.

There is some prominent industry studies that quantities the risk carried by these sites. A report by Digital Citizens Alliance conducted by RiskIQ (Digital Citizen Alliance, 2015) probes 800 sites dedicated to infringing movies and TV shows and find that 1 out of every 3 such sites have malware. Consumers are 28 times more likely to get malware from these sites and many times the malware is delivered by simply visiting the sites, even without clicking a link. However, the data collection was done by examining the sites and did involve actual downloading of any files or clicking on ads. In reality, users may adopt various strategies to avoid malware. The study also followed specific protocol (how many pages to visit on a site, for example) which may not mimic the actual user behavior. Finally, since the study did not involve downloading or clicking the links, there was no delivery of actual malware. So the results should be interpreted as potential risk from visiting the sites as opposed to actual intrusion.

Some other work tries to correlate rate of piracy in a country with the malware infection rates. For example, a report from Business Software Alliance (http://global.bsa.org/internetreport2009/2009internetpiracyreport.pdf) suggests that counties which have high software piracy rates also have high malware infection rate.

The challenge in this space has always been to assemble a dataset of real user behavior over time to establish a relation between their navigation to infringing pages and the risk of security intrusion. In Wondracek et al (2010), authors examine the riskiness of online adult sites. Author crawl the adult websites to measure the riskiness of these

sites and how might they can compromise user security. Authors also host couple of such sites on their own, to understand how easy it is to compromise user machines. The study describes the economic models and deceptive strategies of adult websites in attracting visitors and possible attack vectors for users visiting these sites. However, no real user data was explored.

We are aware of only two studies (Canali, Bilge and Balzarotti 2014 and Ovelgonne et al 2016) which have examined real user behavior with risky outcomes. Both studies examine Symantec data (available via Worldwide Information Network Environment - WINE). The first study correlated users browsing behavior over 3 month period to classify them into risky outcomes. However, risk is measured as visits to possibly malicious sites. The study identifies some features (like hours spent on web or the number of sites visited) and correlate them with whether users are more (less) likely to visit a malicious site. In short, the paper does not actually measure any intrusion attempts or increases probability of finding malicious software on user machine as a result of user browsing. The paper also, does not focus on infringing sites, which is the key focus of our analysis. The second study (Ovelgonne et al 2016) does measure possible intrusion attempts on user machines as measured by Symantec anti-virus, but the study does not correlate it with users' web browsing. They only look at various files downloaded by users (software, gaming and so on) called binaries and whether these downloads correlate with malicious intrusions detected by Symantec AV.

While both these studies have large user base (thousands of users), they suffer from obvious selection (only those users who have Symantec AV and are willing to share are in the study) and having no information about users themselves. These studies also do not study the infringing behavior per se.

We believe our study overcomes this significant gap in the literature – what effect does online piracy have on computer security? Using a panel of users through IRB approved SBO (Security Behavior Observatory) from Carnegie Mellon University, we bring in a

unique dataset to answer this question. While we describe the data in more detail in the next section, we highlight unique features that are particularly salient for our analysis.

1. User behavior is observed in the real world setting. We collect user data (for example which sites they are going, or what they download) using sensors that work in the background without intrusion. We are also able to observe the same user for a period of time (months in most instances). Unlike Symantec sample, our panel of users do not need to have an anti-virus installed a pre-condition for being in the sample.
2. Since we observe users over a period of time, we are able to create a panel dataset which overcomes limitations of cross-section data. This allows us to make stronger causal statements than we can make with mere cross-section data. Our data also extends to a longer period of time allowing us to control for short term noise affecting the results and make more robust claims on user behavior.
3. We capture comprehensive information about users' web activity (like web browsing, emails or games, or downloads and so on). We also collect data on what programs are installed or what files are downloaded by users. We also have data on users' demographics. With a comprehensive panel data in place, we can tease out the effect of infringement carefully and robustly.

To summarize, we are able to observe user behavior of about 250 users over a period of time (1 year) to analyze how visits to infringing sites affect the health of their computer after controlling for their other activities. We find that doubling the time spent on infringing sites leads to 20 percent increase the count of malware found on user machines. Our data allows us to separate malware into adware and more malicious malware like Trojans, Virus and so on. The results remain unchanged when ignore adware and only include more serious malware. We also find no evidence that users who visit infringing sites more take more precautions. In particular, users visiting

infringing sites are less (not more) likely to install an antivirus (AV) software. We show that our results are robust to many different specifications and measurements.

The paper is organized as follows. In section 2, we provide details on data collection and summary. In section 3, we present the econometric model and in section 4 we provide results and discussion. We conclude in section 5.

2. Data Collection and SBO

The details of Security Behavior Observatory can be found in Forget et al (2014 and 2016). SBO was developed to understand user behavior *in setu* and how it translates to the observed security of their machines. Technically, the SBO consists of a set of "sensors" monitoring various aspects of participants' computers to provide a comprehensive overview of user activity that regularly reports (encrypted) measurements to our secure server. Our monitoring provides us with the opportunity to characterize which user actions led to insecure computing states. We can also interact with our participants via surveys and interviews to get insights into their behaviors. [4]

Participants are recruited from a telephone service that calls individuals in Pittsburgh, PA. We recruit SBO participants from a university service that telephones individuals to notify them about ongoing experiments in Pittsburgh, Pennsylvania. Potential participants are contacted to complete a brief pre-enrollment survey to ensure they are over 18 and own a Windows Vista, 7, 8, or 10 personal computer. A member of our research team then calls participants to walk them through the following tasks while they are in front of their computers:

1. Read and complete a consent form, which clearly informs participants that the researchers may collect data on all activity on their computer, except personal file contents, e-mails sent or received, contents of documents on Google Docs, and bank card numbers.

[4] The following text is borrowed from Forget et al paper (2016) page 99.

2. Provide the names and e-mail addresses of others users of the computer to be instrumented, so we may obtain their consent.
3. Download and install the SBO data collection software.
4. Complete an initial demographics questionnaire.

Once all the computers users have consented and we begin receiving data, we send participants a $30 Amazon.com gift card. Participants are then paid $10 per month their computers continue transmitting data to our server. Data transmission occurs in the background, requiring no user action. We encourage and promptly respond to questions about the study via phone or e-mail. We assert that maintaining the confidentiality of their data is our primary concern. Participants may withdraw from the SBO at any time. If we unexpectedly stop receiving data from a machine, we con- tact the participant to attempt to resolve the issue.

Data collection architecture.

The SBO relies on a client- server architecture with several client-side sensors collecting different types of data from participants' machines (Forget et al 2014). Examples of collected data include processes, installed software, web browsing behavior, network packet headers, wireless network connections, Windows event logs, Windows registry data, and Windows update data. The SBO data collection architecture is implemented with multiple technologies: Java, C#, C++, Javascript, SQL, Python, PHP, WiX, and command-line batch scripts.

The SBO architecture provides security and confidentiality of participants' data as follows. All communication between users' machines and our collection server is authenticated and encrypted using unique client-server key pairs. The server only accepts connections from authenticated machines on one specific port. Finally, the data collection server is not used for analysis. Instead, a data analysis server retrieves participants' data from the collection server for long-term storage. The data analysis

server is only accessible from within our institution's network. All data analysis must be performed on the server. No collected data is authorized for transfer from the data analysis server.

For the analysis in this paper, we need access to users' web navigation and file system to identify malwares. We also have data on whether users have an anti-virus installed on their machine and whether the AV detects any intrusion. Finally, we are also able to observe some security precautions user take (like clearing the cookies on their browser or navigating the Internet using anonymous mode).

Classifying Web Navigation Data

We first classify user visits to various sites based on the URLs they visited. This would allow us to control for other possibly websites who may be responsible for potential malware. In particular, we want to isolate infringing sites visits from visits to other potential sites and applications that may also lead to intrusions.

Infringing Sites:

To classify infringing sites we use Google transparency report and a few other lists[5] [6]. If users visit any of the domain listed in these reports, we classify them as visits to an infringing site. It is not always easy to find when a session starts and ends because users might open a browser window and but might not actually be browsing. To avoid over estimating the time spent on a site, we only count the time if we see a user action on that site every 15 minutes. In short, a session is assumed to last for 15 minutes unless we see user taking an action which will extend the session for next 15 minutes. We use the same algorithm for visits to all sites including infringing sites.

[5] https://ustr.gov/sites/default/files/USTR-2015-Out-of-Cycle-Review-Notorious-Markets-Final.pdf

[6] http://www.mpaa.org/wp-content/uploads/2015/10/MPAA_Notorious_Markets_2015-Final1.pdf

The visits to other websites are classified into following categories. The classification was derived from Alexa. Thus for each user, every time they start their browser, we know how long they were active on each of the following websites including for the infringing sites.

Music	Movies
TV	Banking
Gambling	Gaming
shopping	Emails
Social Networking sites	Adult websites

Number of movie downloads and total downloads:

By collecting filesystem metadata, we identified files downloaded from the web by users through web browsers, torrent applications, or any other means. Using a combination of file extension, filename, and file path heuristics we determined which files were media or video files.

Malware Classification

We used *virustotal.com* to identify known malicious files and programs on user machines. This service aggregates malware scanning results from over 50 anti-virus (AIV) products and provided us with a measure of the number of anti-virus products identifying each file or program as malicious. Since for each potential malware signature, we get different number of AIVs classifying it as malware, we weight the number of malwares by the proportion of AIVs which identify it as malware. To keep the analysis tractable, we assign a weight of 0.3 to a malware when 30-50% of malware identify it as malware. We assign a weight of 0.6 when 50-70% AIV identify it as malware and a weight of 1 when more than 70% AIVs identify a signature as malware. Thus, to

count the number of malware on a user machine on a given day, each malware is assigned a weight and then we take an expectation and round it to the nearest integer. [7]

Malware Severity Classification

We can further classify the malware based on the severity. While we do not know directly how severe a malware is, we want to ensure that adware, which is a very common way for websites to display ads to user, are not counted as malware. While it is possible that some of the adware can be malicious, a large number of adware is for the purpose of delivering ads to end users and while undesirable to end users, they may not be malicious. In our data, we can identify each malware signature as potentially adware or not. While classification of even adware is also probabilistic, in our analysis we classify a malware as an adware if more than 30% of AIVs classify it as an adware. We will provide a separate analysis for total malware count and malware count without adware.

We should note that our measures of malware is probably an undercount of the actual number of malware files found on user machines since *virustotal* is not able to identify all malware signatures.

Presence of Antivirus (AV)

We used scans of the installed software on each user's computer to extract details on any anti-virus product they used. We matched this information with the publicly available information on each product to classify each anti-virus product.

Detecting intrusions via AV.

Additionally, our software regularly scraped the log files of several of the most common anti-virus products found on user machines including Kaspersky, Norton, Webroot,

[7] We try some combinations of weights and our results are quite robust.

Malwarebytes, and Windows Defender. This data allowed us to find instances where the user's anti-virus product successfully identified and cleaned malware from the computer.

Presence of Adblockers

From our Google Chrome browser extension, we identified extensions installed and enabled on the user's computer that provided ad-blocking capabilities. This included products such as Ad Block Plus.

2.1 Summary statistics:

We have data for about 312 users who were in our sample in 2016. The starting period for our sample is January 2016 and the end period is December 2016. But not every user is observed for all months (some may have joined the panel in the middle of the year while other may have quit during the year. We also drop the month if user had no activity recorded). We require users to have at least two months of data before we include them in our sample. We also have demographic information about users and in some cases have missing data. This leads to a few users being dropped from our final sample. After cleaning up, we have 253 users whose data is used. All data is aggregated at the monthly level unless otherwise noted.

Infringement is defined as the time spent by a user on infringing websites. So for the analysis going forward, we will treat it a continuous variable (number of minutes spent on infringing websites in a given month). We apply the same strategy and calculate the time spent on other websites as well. For description purpose, we first split the sample into users who never have visited any infringing site (infringe Time = 0) during the period we observe them and users who have spent non-zero time on these sites. We provide a summary statistics across these two segments. All numbers are monthly averages.

79 users in our sample never visited or spent any time on infringing sites while 174 of them spent non-zero times on these sites for the duration of their data. All variables with suffix "T" are continuous measures and reflect the time spent on the activity (for example musicT is the amount of time spent by the user on legitimate music websites). AntiVir (is 1 if users have installed any anti-virus software) is dummy variables. MovieDC is the number of movies downloaded by end users and DownloadDC are the numbers of other downloads (document, images and so on) and ClearBC is the number of times a user clears cookies in her browser. The demographic variables are self-explanatory. MalT is the number of total malware found on user machine during the study period and MalT_noad is the number of malware without the adware.

Table 1: Summary Statistics

	infringer 0			infringer 1		
VARIABLES	N	mean	sd	N	mean	sd
totalT	79	559	595	174	1869	2363
infringeT	79	0	0	174	97	349
incognitoT	79	15	81	174	83	388
musicT	79	3	9	174	24	66
movieT	79	2	6	174	16	47
tvT	79	78	173	174	298	550
bankingT	79	19	54	174	53	117
gamblT	79	1	3	174	30	261
gameT	79	2	5	174	36	247
shopT	79	50	98	174	190	415
emailT	79	42	116	174	179	548
socialT	79	116	253	174	590	1104
adultT	79	1	7	174	53	512
clearBrC	79	0.1	0.5	174	2	14.5
AntiVir	79	0.7	0.5	174	0.6	0.5
movieDC	79	0.2	0.8	174	0.7	1.9
DownloadC	79	53	77	174	98	106
MalT	79	1.4	2.6	174	1.5	3.1
MalT_noad	79	0.7	1.4	174	1.2	2.5
adblocker_F	79	0.2	0.4	174	0.4	0.5
age	79	42.4	17.8	174	34.8	15.5
income	71	50,528	46216	157	46274	47981
student	79	0.2	0.4	174	0.3	0.5
edu	79	0.7	0.5	174	0.6	0.5
race	71	0.6	0.5	157	0.6	0.5

It is immediate that users who are infringing are also users who generally spend a lot more time on the Internet. People who infringe spent about 97 minutes a month on infringing sites and more than 1800 minutes (30 hours) a month of total time on Internet. People who do not infringe spend significant less time on the Internet, 559 minutes (approximately 10 hours). This is an important reason that cross section comparisons are fraught with selection and would lead to suspect results. As we will explain, we model a within-user model to avoid this problem. Users who infringe are spending more time on most other activities as well (social network, TV shows and even adult websites). They also download more movies and download more in general. There are no significant demographic differences between two samples. Students are slight more likely to infringe but the education and race is similar. Infringers though tend to be slightly younger (35 year vs 42 years).

In terms of computer security, average number of malware files found on user machine is about 1.5 for users who infringe and 1.4 for those who do not. Recall that we do not count malware signatures with fewer than 30% of AIVs classify as malware. [8] Once we remove the potential adware files, the number of malware count decline to about 0.7 for non-infringers and about 1.4 for infringers. Surprisingly, there is no difference in the presence of anti-virus software use across two samples. One would have thought that users who infringe more are more likely to be careful and take precaution (like have anti-virus software installed). Both samples have about 60% installation rate. Non-infringers though are lot less likely to use incognito more or clear their browser cookies.

We also provide a correlation table below.

[8] The number of malware found is larger when we include those files (closer to 18). Again, our results are robust to these measurements.

	totalT	infringeT	musicT	movieT	movieDC	tvT	shopT	socialT	emailT	bankingT	adultT	gamblT	gameT	movieDC	DownloadC
totalT	1.00														
infringeT	0.36	1.00													
musicT	0.21	0.15	1.00												
movieT	0.23	0.26	0.06	1.00											
movieDC	0.00	0.01	0.00	0.01	1.00										
tvT	0.55	0.34	0.14	0.19	0.01	1.00									
shopT	0.54	0.16	0.11	0.21	0.00	0.20	1.00								
socialT	0.63	0.50	0.15	0.26	0.00	0.48	0.36	1.00							
emailT	0.36	0.20	0.04	0.06	0.01	0.08	0.33	0.29	1.00						
bankingT	0.38	0.11	0.05	0.08	0.00	0.14	0.29	0.21	0.37	1.00					
adultT	0.33	0.11	0.01	0.02	0.00	0.07	0.15	0.03	0.03	0.06	1.00				
gamblT	0.21	0.03	0.18	0.01	0.00	0.01	0.13	0.00	0.02	0.02	0.01	1.00			
gameT	0.14	0.06	0.08	0.03	0.00	0.07	0.06	0.06	0.06	0.01	0.00	0.04	1.00		
movieDC	0.00	0.01	0.00	0.01	1.00	0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.00	1.00	
DownloadC	0.04	0.00	0.03	-0.01	0.08	0.04	0.00	0.01	0.01	0.01	-0.01	-0.01	0.00	0.08	1.00

Not surprisingly, many variables are correlated with total time but none of them of the correlations are too large. Total time is consists of some other activities as well that are not listed in the earlier section.

Our identification though relies on intensity of infringement and how it affects outcome. So the summary statistics which compares no infringement with some infringement does not paint the full picture of how infringement might affect computer health. In particular, our identification relies on month to month changes in infringing intensity and how it affect the probability and count of malware. We now formally present the model and analyze our data.

3. Model and Results:

Our identification strategy is to look at variation in infringement intensity over time within a user and whether it affects incidences of malware infection. We also add a large number of possible controls to account for the fact that other potential risky

behavior (visiting adult websites for example) may be responsible for malware incidences. Thus use of panel data along with detailed controls should allow us to tease out the effect of infringement intensity on user outcomes.

Since the dependent variable (number of new malware found in a month) is a continuous and count variable, we try two different specifications. In the first specification, we estimate a linear fixed effect model. We first take a log of most of the independent variables since there is a large variance across users and over time and then estimate the following model.

$$MalWare_{it} = \alpha_i + \delta_t + \beta_1 Log(TotalT_{it}) + \beta_2 Log(\inf ringeT_{it}) + \beta Control_{it} + \varepsilon_{it}$$

α and δ capture user and time fixed effects (we use monthly dummies). TotalT is the total time spent on computer and infringeT is the time spent on infringing sites in a given month *t*. Thus β_2 is the main estimate of interest for us which estimates the effect of infringement time at time *t* on number of malware found at time *t*. Controls are all other variables (as listed in Table 1) that may possibly have an effect on the malware incidence. Since this is a Linear-Log specification, β_2 is interpreted as having a linear β_2 increase in the number of malware when there is 100 percent change in the intensity of infringement. In other words, 100% increase in time spent (doubling the time) on infringing sites is associated with β_2 increase in number of malware files found. This interpretation is true for other variables as well.

This is a classic within-user regression that controls for any user specific preferences that we cannot control. In short, a user may be prone to risk taking which we cannot fully capture in our controls. But the user specific dummy should account for such user preferences. Given very detailed controls we have, we believe this model allows us to make causal claims on the effect of infringing on computer health. Notice that in our model, controlling for other factors, if a user spends more (less) time on infringing sites, β_2 will estimate its effect on malware count. Users can be quite heterogeneous in their

preferences for infringing. A within user model accounts for such cross-section heterogeneity.

In case, we are worried that our results are driven by time spent on infringing sites and few large observations maybe driving the results, we also use infringe dummy instead of amount of time spent. So if a user visits an infringing site in a given month (no matter how many times), the infringe_dummy = 1 and 0 otherwise. In this specification, the amount of infringing time is irrelevant.[9] We report these results in the appendix and find that results remain unchanged.

However, since the malware incidence are count data with lots of zeros, a more effective strategy would be use count model (either Poisson or Negative Binomial regression). The limitations of these models is that they will drop all observations when malware count are zero for entire period (which is quite common in our data) and infringement time is zero as well. The fixed effect models also does not allow us to use user demographic data. So we use random effect models for Negative Binomial regression.[10] The negative binomial model can be written as:

$$f(y_{it} \mid \lambda_{it}, \theta_i) = \frac{\Gamma(\lambda_{it} + y_{it})}{\Gamma(\lambda_{it})\Gamma(y_{it} + 1)} \left(\frac{\theta_i}{1 + \theta_i}\right)^{y_{it}} \left(\frac{1}{1 + \theta_i}\right)^{\lambda_{it}}$$

Where y_{it} is the number of malware incidences, θ_i are individual fixed effects and λ_{it} are function of control variables (X_{it}).

The interpretation of parameters in the NBD model difference than the linear model. A one percent increase in infringing intensity is associated with an approximately β_2

[9] This also avoids the worry that 15 minutes time per session may introduce any measurement error.
[10] The fixed effect models on count data are estimated via conditional likelihood and are the same as linear fixed effects (see https://statisticalhorizons.com/fe-nbreg)

percent increase in the incidences of malware count.[11] As we will show below that both models directionally perform similarly.

We estimate the model for both with and without including adware counts. Thus in our first specification, the dependent variable is total number of malware and in the second specification, the dependent variable is the number of malware without adware. In each specification, we estimate both the linear fixed effect model and the random effect negative binomial model.

3.1 Results

3.1.1 Infringement and Total Malware

We first report results on the effect of Infringement on total Malware count. We report the fixed effect model and Negative Binomial model side by side. In the linear fixed effect model, the dependent variable is linear total count of malware in a given month for a user. In the NBD specification, the dependent variable is logged value of malware count.

[11] A more correct interpretation would (exp(β_2)-1)% increase in incidences of malware but for the smaller values of β_2, *they are approximately the same.*

Table 2: Estimates for Total Malware Outcomes

VARIABLES	(4) MalT_noad	(1) MalT
LNinfringeT	***0.054****	***0.190*****
	(0.027)	(0.059)
LNtotalT	0.059	0.117
	(0.083)	(0.123)
LNmusicT	-0.026	-0.119*
	(0.026)	(0.066)
LNmovieT	0.008	0.017
	(0.025)	(0.068)
LNtvT	-0.012	-0.004
	(0.023)	(0.054)
LNbankingT	0.040	-0.048
	(0.031)	(0.054)
LNgamblT	-0.070	-0.045
	(0.127)	(0.087)
LNgameT	-0.001	-0.071
	(0.030)	(0.065)
LNsocialT	-0.060	-0.080
	(0.043)	(0.057)
LNshopT	0.009	0.054
	(0.025)	(0.059)
LNemailT	0.025	0.014
	(0.027)	(0.046)
movieDC	-0.058*	-0.140
	(0.034)	(0.181)
DownloadC	-0.002	0.003
	(0.002)	(0.003)
LNadultT	0.031	-0.032
	(0.051)	(0.068)
age		-0.003
		(0.007)
race		-0.261
		(0.236)
student		-0.534*
		(0.292)
edu		-0.298
		(0.209)
Constant		1.751
		(0.527)
Observations	1,537	1,302
R-squared	0.135	
Number of Users	253	228

In column 1, we report the fixed effect estimates and in column2, we report the NBD estimates. The standard errors are reported in the parenthesis below the estimate. Note our model is very detailed and allows for all controls (we do not report estimates on monthly dummies and individual dummies to avoid clutter). With these detailed controls, we may not expect any significance. But it is very interesting to note that the most significant variable is InfringeT and this is true in both fixed effect and in NBD model. In both models, the parameter is estimated with high precision ($p<0.05$ in fixed effect model, and $p<0.001$ in NBD model). This is despite that we have many zeros in the dependent variable. Our estimates suggest that in given month, more time spent on infringing sites leads to a higher number of malware files found on user machine in that month. For the Fixed effect model, controlling for all factors, doubling the amount of time spent (a 100% increase) on infringing sites increases the number of malware count by almost 0.05 units. While this number may look small, the mean number of malware count on a user machine is 0.24 per month. So a 0.05 increase translates to a 20 percent increase in malware count due to infringing alone. The estimate on NBD model is similar, doubling the amount of time spent (a 100% increase) on infringing sites leads to about 20 percent increase ($e^{0.189} – 1$) in malware count. In short, our model estimate remains the same despite the change in specification.

The only other variable of significant is the number of more legal movie downloads are associated with fewer malware counts (fixed effect model) though somewhat surprisingly more visits to adult sites is not associated with more malware (after control for other activities). Notice that we lose a few observations in the NBD model since we are able to include demographic variables and they are not consistently available for all participants in our panel.

It is clear from these results that visits to infringing sites leads to an increase in number of total malware files on users' machines and this results is both economically and statistically significantly.

One may worry that these results may be driven by adware (which may or may not be as malicious). So we re-run our analysis now removing the potential adware from our malware count. Thus our dependent variable is the number of malware count without the adware. We follow the same strategy as before and estimate both the linear fixed effects and NBD model.

Table 3: Estimates for Malware Outcomes without Adware

VARIABLES	(2) MalT (without adware) Fixed effects	(4) MalT (without adware) NBD
LNinfringeT	***0.034****	***0.198*****
	(0.021)	(0.063)
LNtotalT	0.026	0.136
	(0.058)	(0.133)
LNmusicT	-0.026	-0.150**
	(0.019)	(0.072)
LNmovieT	0.006	0.045
	(0.019)	(0.072)
LNtvT	-0.007	-0.006
	(0.013)	(0.058)
LNbankingT	0.033	-0.039
	(0.023)	(0.058)
LNgamblT	-0.075	-0.088
	(0.108)	(0.093)
LNgameT	-0.002	-0.054
	(0.021)	(0.068)
LNsocialT	-0.039	-0.101*
	(0.032)	(0.060)
LNshopT	0.006	0.065
	(0.019)	(0.065)
LNemailT	0.013	0.005
	(0.020)	(0.049)
movieDC	-0.045*	-0.103
	(0.023)	(0.177)
DownloadC	0.001	0.005*
	(0.001)	(0.003)
LNadultT	0.040	0.003
	(0.041)	(0.070)

age		-0.004
		(0.008)
race		-0.173
		(0.257)
student		-0.576*
		(0.319)
edu		-0.304
		(0.228)
Constant	1.281***	0.501
	(0.381)	(0.802)
Observations	1,537	1,302
Number of Users	253	228
Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1		

The results are similar to the earlier results except that estimate in the linear fixed effect case is less precise (significant at $p<0.1$). The numbers are still economically significant. The estimate of 0.034 translates to a 20 percent increase in malware count (the mean monthly counts are about 0.17 malware per month per user) if users double the amount time they spent on infringing sites. The NBD model also projects a 21 percent increase in malware count due infringing. In short, whether we include total malware count or malware count without adware, we find that time spent on infringing sites increases the malware count by almost 20 percent. And, this is true in both specifications. Thus our results seem quite robust to how we count malware or what particular specification we use (also see appendix where we use a dummy variable for measuring infringement)

Similar to last specification, none of the other variables are significant expect for movie downloads. We find small evidence that user specific demographic variables play a small role but most are imprecisely estimated.

Broadly, we find strong evidence that visiting infringing sites is more likely to lead to malware files on users' machines.

4. Are Infringers more Careful?

We now explore whether infringers are more computer savvy and careful. The earlier analysis suggests that, infringers even if they are taking any precautions, are more likely to find malware files on their computer. We now examine whether infringers are more likely to take precautions. One can argue that users who infringe maybe more web savvy and hence be able to avoid malware despite visiting infringing sites and downloading content.

In particular, we measure users' decision to install antivirus (AV) software. We check if users who infringe more are more likely to have Anti-Virus software (AVs) installed. Based on the scan of installed programs, we are able to measure if users have any of the popular AVs installed on their machines. This is an important distinction from work done with the Symantec WINE data. Those studies could only look at the data where users already had a Symantec AV installed.

Ideally we would like to run the panel data analysis like we did earlier for examining AV installation decision. However, in most cases, the decision to install AV is not dynamic (i.e. users do not install or uninstall it on a month-to-month basis). So we do a cross section analysis. We calculate the mean usage statistics for all variables in Table 1 and regress against the AV installation or not. We test if intensity of infringement affects with decision to have AV or not. The model we test is:

$$Logit(AV_i) = \beta_0 + \beta_1 Log(InfringeT_i) + \beta_2 Control_i + \beta_3 Demographic_i + \varepsilon_i$$

Where *i* indexes a user and AV is a binary variable which takes value 1 for users who have installed AV and 0 otherwise. InfringeT is the average monthly minutes spent on infringing web sites and control variables are the other control variables used in

previous analysis. β_2 thus is a vector of estimated parameters. We also use the demographic controls. Our variable of interest is β_1 which estimates how the intensity of infringement affects the decision to install AV or not. In particular, we test if users who infringe more are also more likely to have an AV installed.

The following Table report the estimates.

Table 5: Estimates on AV installation

VARIABLES	(1) AntiVir_F
LNinfringeT	***-0.122***
	(0.124)
LNtotalT	0.556*
	(0.300)
musicT	-0.004
	(0.002)
LNmovieT	-0.012
	(0.147)
LNtvT	-0.057
	(0.145)
LNbankingT	0.205*
	(0.120)
LNgamblT	-0.311**
	(0.157)
LNgameT	0.0841
	(0.129)
LNsocialT	0.121
	(0.135)
LNshopT	-0.351**
	(0.146)
LNemailT	-0.117
	(0.105)
movieDC	-0.146*
	(0.080)
LNadultT	-0.122
	(0.126)
Age	0.032**
	(0.015)
Edu	-0.913**
	(0.355)
income	1.27e-06
	(3.38e-06)
student	-1.351***
	(0.414)
Constant	-1.979

	(1.228)
Observations	228

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

While the estimate on totalT is positive and significant, estimate on infringeT is negative though insignificant. So users who browse more overall, are more likely to have AV installed but users who spend more time on infringing websites are not more likely to install an AV. If anything, there is a negative relation between AV and time spent on infringing. This suggests that users who infringe are not any more web savvy or even careful when they visit infringing sites. This could also be due to the fact that users may not perceive visiting infringing sites as a risky decision. This collaborates with our earlier results that infringing users are more likely to see malware files on their machine because they are not any more likely to take precautions or AVs are not particularly effective.

Estimates on demographic variables are also interesting. Students and more educated users are more risk taking and less likely to have an AV while older users are potentially more likely to have AV installed (an estimate of -0.91 on education translates to about a 21% decrease in AIV installation rate for users who have college education and more)

We also explore some other metrics which potentially measure users' security and privacy precautions. For example, we have data on clearing cookies (which may make it harder for advertisers for target an IP address) and installing adblockers. However, this data is limited to few browsers (for example Chrome). So we lose some data points when analyzing these metrics and hence do not report them.

5. Conclusion and Discussion

Policy discussions on piracy and copyrights focus on the short term and long term trade-offs for consumers and producers. On the one hand, consumers benefit in the short run due to cheaper and widespread access of content because of piracy. But on the other

hand, consumers suffer in the long run if piracy hurts creativity and innovation. Over time, piracy may reduce the quality and quantity of content produced. However, piracy may also impose some significant short term costs to users. Many piracy sites may be unsafe for users to navigate. They may be cause for malware, Trojans and other insecurities. When users visit infringing sites, these insecure software may also download on user machines and inflict economic and other potential opportunity costs to users. In many instances, users may be unaware of these threats.

In this paper, we collect a unique dataset of more than 250 users and analyze whether frequent visits to infringing sites lead to poor computer health and higher possibility of downloading malware files. As part of Carnegie Mellon's SBO project, we are able to monitor detailed user behavior and measure how much time users spend on infringing sites. We can control for visits to other websites which may confound our analysis. Our sensors are also able to identify malwares files downloaded on user machines. We are able to measure some other security metrics as well. Thus we are able to analyze whether more time spent on infringing sites lead to more downloading of malware files. The panel nature of our data allows to make causal inferences. Thus we identify the effect by looking at changes in users visit frequency to infringing sites over time and whether it causes changes in number of malware files found on user machines. The panel data is critical for our analysis because cross section comparison of users will lead to significant selection issues. We also able to control for variety of other navigational behavior. We believe our data and analysis is first of its kind to answer this question.

We find that more visits to infringing sites does lead to more number of malware files being downloaded on user machines. In particular doubling the amount of time spent on infringing sites cause a 20 percent increase in malware count. Even after we classify malware files into adware and remove them from analysis, our results still suggest that there is a 20 percent increase in malware count due to visits to infringing sites. These results are robust to various controls and specifications. We also find that users who visit infringing sites do not take any more precautions than other users. In particular, we find no evidence that such users are more likely to install anti-virus software. If

anything, we find that infringing users are more risk taking. We find that students and young users are less likely to install AV.

While our data and analysis is unique, it is not without limitations. Our malware classification can be improved. Currently we are able to remove adware from analysis but we do not have a strong measure on which malware are more (less) severe. We also identify malware using public sources like virustotal.com. This potentially is undercounting the actual number of malware. Future work can tally the malware signatures with large AV providers to get a more precise number. Currently, we only measure the amount of time spent on infringing sites but do not measure actual user behavior on the sites. In the future work, we can classify user actions more precisely to gain deeper insights into how user actions lead to malware downloads. Ideally, we would like to explore use of instrumental variables to control for some potential endogeneity.

Despite the limitations, we believe our paper provides a deeper understanding into how infringing sites can cause proliferation of malware. We hope our work will be a springboard for similar empirical work.

Bibliography

Aguiar, L., Claussen, J., and Peukert, C. *Online Copyright Enforcement, Consumer Behavior, and Market Structure.* Working paper. Copenhagen Business School, Copenhagen, Denmark, 2015; http://ssrn.com/abstract=2604197

Canali D, Leyla Bilge, Davide Balzarotti. 2014. On the Effectiveness of Risk Prediction Based on Users Browsing Behavior, ASIA CCS'14.

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

Danaher, B., Smith, M.D., and Telang, R. Piracy and copyright enforcement mechanisms. Chapter 2 in *Innovation Policy and the Economy, Volume 14*, J. Lerner and S. Stern, Eds. National Bureau of Economic Research, University of Chicago Press, Chicago, IL, 2014, 31–67.

Digital Bait- How Content Theft Sites and Malware are Exploited by CyberCriminals to hack into Internet Users' Computers and Personal Data, December 2015, http://www.digitalcitizensalliance.org/news/press-releases-2015/digital-bait-internet-users-at-high-risk-of-malware-from-content-theft-70-million-underground-market/

Handke, Christian, 2012. "Digital copying and the supply of sound recordings," Information Economics and Policy, Elsevier, vol. 24(1), pages 15-29.

Oberholzer, F., K. Strumpf. 2007. The Effect of File Sharing on Record Sales. An Empirical Analysis. *Journal of Political Economy,* **115**(1) 1-42.

Ovelgonne Michael, Tudor Dumitras, Aditya Prakash, V. S. Subrahmanian, Benjamin Wang (2016). Understanding the Relationship between Human Behavior and Susceptibility to Cyber-Attacks: A Data-Driven Approach. Working Paper.

Smith D Michael, R Telang (2014). Assessing the Academic Literature Regarding the Impact of Media Piracy on Sales. Working paper, ssrn.

Waldfogel Joel 2014. "Digitization and the Quality of New Media Products: The Case of Music," in Economics of Digitization, University of Chicago Press, Avi Goldfarb, Shane Greenstein, and Catherine Tucker, eds.

Telang R, J Waldfogel 2016. Piracy and New Product Creation: A Bollywood Story. Working Paper. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2478755

Wondracek G, T Holz, C Platzer (2010). Is the Internet for Porn? An Insight Into the Online Adult Industry. Downloaded from https://www.researchgate.net/publication/242581158

Appendix

Results using Infringe_dummy instead of Infringe Time.

VARIABLES	(2) MalT
infringe_dummy	0.259**
	(0.125)
LNtotalT	0.087
	(0.099)
LNmusicT	-0.033
	(0.031)
LNmovieT	0.016
	(0.028)
LNtvT	-0.032
	(0.028)
LNbankingT	0.045
	(0.036)
LNgamblT	-0.082
	(0.145)
LNgameT	0.001
	(0.033)
LNsocialT	-0.070
	(0.050)
LNshopT	0.005
	(0.029)
LNemailT	0.031
	(0.031)
movieDC	-0.067**
	(0.033)
LNadultT	0.047
	(0.057)
Constant	2.146***
	(0.720)
Observations	1,302
R-squared	0.155
Number of users	228

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1