NBER WORKING PAPER SERIES

THE NATURE AND INCIDENCE OF SOFTWARE PIRACY: EVIDENCE FROM WINDOWS

Susan Athey
Scott Stern

Working Paper 19755
http://www.nber.org/papers/w19755

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
December 2013

This research was conducted while both researchers were Consulting Researchers to Microsoft Research. Support was also provided by the Toulouse Network for Information Technology. This paper represents the research of the authors and does not reflect the views of any institution or organization. This paper has benefited greatly from seminar comments at the NBER Economics of Digitization Conference, Microsoft Research, the MIT Microeconomics at Sloan Conference, and by Ashish Arora, Shane Greenstein, Markus Mobius, and Pierre Azoulay. Exceptional research assistance was provided by Bryan Callaway and Ishita Chordia. Susan Athey has a long-term consulting relationship with Microsoft Corporation. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

At least one co-author has disclosed a financial relationship of potential relevance for this research. Further information is available online at http://www.nber.org/papers/w19755.ack

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

The Nature and Incidence of Software Piracy: Evidence from Windows
Susan Athey and Scott Stern
NBER Working Paper No. 19755
December 2013
JEL No. L86,O34

ABSTRACT

This paper evaluates the nature, relative incidence and drivers of software piracy. In contrast to prior studies, we analyze data that allows us to measure piracy for a specific product – Windows 7 – which was associated with a significant level of private sector investment. Using anonymized telemetry data, we are able to characterize the ways in which piracy occurs, the relative incidence of piracy across different economic and institutional environments, and the impact of enforcement efforts on choices to install pirated versus paid software. We find that: (a) the vast majority of "retail piracy" can be attributed to a small number of widely distributed "hacks" that are available through the Internet, (b) the incidence of piracy varies significantly with the microeconomic and institutional environment, and (c) software piracy primarily focuses on the most "advanced" version of Windows (Windows Ultimate). After controlling for a small number of measures of institutional quality and broadband infrastructure, one important candidate driver of piracy – GDP per capita – has no significant impact on the observed piracy rate, while the innovation orientation of an economy is associated with a lower rate of piracy. Finally, we are able to evaluate how piracy changes in response to country-specific anti-piracy enforcement efforts against specific peer-to-peer websites; overall, we find no systematic evidence that such enforcement efforts have had an impact on the incidence of software piracy.

Susan Athey
Graduate School of Business
Stanford University
655 Knight Way
Stanford, CA 94305
and NBER
athey@stanford.edu

Scott Stern
MIT Sloan School of Management
100 Main Street, E62-476
Cambridge, MA 02142
and NBER
sstern@mit.edu

I. Introduction

In the summer of 2009, Microsoft planned to release a new version of its flagship operating system, Windows 7. Relative to Windows Vista, Windows 7 offered significant improvements for consumers, including "driver support to multitouch groundwork for the future, from better battery management to the most easy-to-use interface Microsoft has ever had" (CNET, 2009a). The redesign of the core operating system, as well as the development of bundled applications and features, represented a significant investment on the part of Microsoft, with approximately 2500 developers, testers and program managers engaged on the project for multiple years. Perhaps more than any other Microsoft product before it, Windows 7 was designed with a global market in mind (Microsoft, 2009). Microsoft explicitly included a large number of features aimed at serving this global market, including the Multilingual User Interface included in Windows Ultimate and creating a low-priced version, Windows Home Basic, which was targeted specifically at emerging markets.

However, just weeks after the release of the final version of the software and individualized product "keys" to OEMs, a number of websites reported that one of the original equipment manufacturer master product keys issued to Lenovo had been hacked and released onto the Internet (CNET, 2009b). Websites quickly assembled step-by-step instructions into how to gain access to a pre-release, pirated version of Windows 7, and developed tools and protocols that allowed users to install an essentially complete version of Windows 7 Ultimate in a small number of transparent steps. While Microsoft chose to discontinue the leaked product key for OEM installation (they issued a new key for legitimate use by Lenovo), users were allowed to activate Windows 7 with the leaked key. In addition, though they did receive a modest functionality downgrade, users of the leaked Lenovo key were able to receive regularized product support and updates for their system. Microsoft argues that this approach ensures that they can "protect users from becoming unknowing victims, because customers who use pirated software are at greater risk of being exposed to malware as well as identity theft." (CNET, 2009a). Over the course of 2009, a number of additional leaked keys and methods for pirating Windows 7 appeared on the Internet, and, by 2012, there were a large number of country-specific unauthorized Windows installation web pages, often tailored to specific languages or countries.

By and large, most discussions of digital piracy – the use of the Internet to enable the unauthorized (and unpaid) replication of digital media including music, movie, and software – are based on specific instances of piracy, discussions of specific file-sharing websites (such as the Pirate Bay), or are closely tied to specific advocacy efforts. As emphasized by a recent National Academies study, the policy debate over piracy and the appropriate level of copyright enforcement is hampered by the lack of direct empirical evidence about the prevalence of piracy or the impact of enforcement efforts (Merrill and Raduchel, 2013). This empirical vacuum is particularly important insofar as appropriate policy over piracy requires the consideration of both benefits and costs of particular policies. For example, the case for aggressive enforcement against piracy is strongest when piracy results from simple lack of enforcement (or the absence of a legal framework for enforcing software copyright), while the argument for piracy tolerance is strongest when the primary impact of piracy is to provide access to low-income consumers whose alternative is non-consumption. The development of appropriate policy, therefore, depends on an empirical assessment of the form that piracy takes in key settings.

This paper addresses this need by undertaking a systematic empirical examination of the nature, relative incidence, and drivers of software piracy. We focus specifically on a product – Windows 7 – which was unambiguously associated with a significant level of private sector investment by a private sector company. The key to our approach is the use of a novel type of data that allows us to undertake a direct observational approach to the measurement of piracy. Specifically, we take advantage of telemetry data that is generated passively by users during the process of Windows Automatic Update (WAU) and is maintained in an anonymized fashion by Microsoft. For machines in a given geographic area, we are able to observe the product license keys that were used to initially authenticate Windows, as well as machine characteristics (such as the model and manufacturer). We are able to use these data to construct a conservative definition of piracy, and then calculate the rate of piracy for a specific geographic region.[1] The primary focus of our empirical analysis is then to assess how the rate and nature of that piracy varies across different economic, institutional and technological environments.

[1] In constructing a novel and direct observational measure of piracy, our work complements but also offers an alternative to the small prior literature on software piracy that has used a more indirect measure of piracy that *infers* the rate of piracy from the "gap" between the stock of sales / licenses allocated to a particular region/segment and audits of the "software load" for typical devices for users within that region/segment (Business Software Alliance, 2011).

We document a range of novel findings. First, we characterize the nature of "simple" software piracy. While software piracy has of course always existed, our examination of Windows 7 suggests that the global diffusion of broadband and peer-to-peer systems such as Pirate Bay has given rise to a distinctive "type" of software piracy: the potential for global re-use of individual product keys, with sophisticated and active user communities that develop easy-to-follow instructions and protocols. While the use of peer-to-peer networking sites has been associated for more than a decade with piracy for "smaller" products such as music or video, there is now a relatively direct way that any broadband user can access a fully functional version of Windows for free through the Internet. In particular, we document that a very small number of abused product keys are responsible for the vast bulk of all observed piracy, and that the vast majority of piracy is associated with the most advanced version of Windows (Windows Ultimate). This finding suggests that one proposed type of anti-piracy initiative – offering a "barebones" version at a greatly reduced price – may be of limited value, since such efforts will have no direct impact on the availability of a fully featured version of Windows for free (and may be considered a poor substitute). As well, we are able to detect a distinctive industrial organization to piracy: piracy rates are much higher for machines where the OEM does not install Windows during the production process, and the rate of piracy is much lower for machines produced by leading OEMs.

Third, we are able to evaluate how software piracy varies across different economic, institutional and technology environments. In addition to traditional economic measures such as GDP per capita (and more nuanced measures such as the level of income inequality), we also gather data characterizing the overall quality of the institutional environment (e.g., using measures such as the World Bank Rule of Law Index or the Foundational Competitiveness Index (Delgado, et al, 2011)), the ability of individuals within a country to take advantage of broadband, and the innovation orientation of a country. Our results suggest that the level of piracy is closely associated with the institutional and infrastructure environment of a country. For example, piracy is negatively associated with a number of traditional measures of the strength of legal and political institutions, is increasing in the accessibility and speed of broadband connections (faster broadband reduces the time required for pirating), and is declining in the innovation intensity of an economy. Most importantly, after controlling for a small number of measures for institutional quality and broadband infrastructure, the most natural

candidate driver of piracy – GDP per capita – has no significant impact on the observed piracy rate. In other words, while the pairwise correlation between piracy and GDP per capita is strongly negative, there is no direct effect from GDP per capita. Poorer countries tend to have weaker institutional environments (Hall and Jones, 1997, among many others), and it the environment rather than income per se which seems to be correlated with the observed level of piracy. Importantly, this finding stands in contrast to prior research, which has not effectively disentangled the role of institutions from the role of income per se.

Finally, we take advantage of time-series variation in our data to directly investigate the impact of the most notable anti-piracy enforcement efforts on the contemporaneous rate of Windows 7 piracy. Specifically, during the course of our 2011 and 2012 sample period, a number of individual countries imposed bans on the Pirate Bay website, the single largest source of pirated digital media on the Internet. Though such policy interventions are endogenous (the bans arise in response to broad concerns about piracy), the precise timing of the intervention is reasonably independent of Windows 7 piracy in particular, and so it is instructive to examine how a change in the level of enforcement against piracy impacts the rate of Windows 7 software piracy. Over a range of different anti-piracy enforcement efforts, we find no evidence for the impact of enforcement efforts on observed piracy rates. Overall, this paper offers the first large-scale observational study of software piracy. Our analysis highlights the value of emerging forms of passively created data such as the Windows telemetry data, and also the role of both institutions and infrastructure in shaping the overall level of piracy.

II. The Economics of Software Piracy

The economics of piracy and the role of intellectual property in software is a long-debated topic (Landes and Posner, 1989; Merrill and Raduchel, 2013; Danaher, Smith and Telang, 2013). Like other forms of intellectual property such as patents, the copyright system has the objective of enhancing incentives for creative work and technological innovation by discouraging precise copying of expression, and is a particularly important form of intellectual property for software. In the case of global software products such as Windows, uneven copyright enforcement across different countries can result in a reduction in incentives to innovation and a distortion in the level of country-specific investment (e.g., companies may limit investment in language and character support in countries with high rates of piracy). The impact

on regional investment would be of particular concern if the underlying driver of variation in piracy was the result of simple differences in legal institutions (such as the strength and respect for property rights) rather than the result of income differences (in which case there might also be a low willingness-to-pay for such value-added services). Piracy also has the potential to impose direct incremental costs on both software producers and purchasers of valid and updated software by facilitating the diffusion of viruses and other forms of malware. Because of the potential for a negative externality from the diffusion of pirated software, many software companies (including Microsoft) provide security updates (and some number of functionality updates) for pirated software. More generally, because software production is characterized by high fixed costs and near-zero replication costs, piracy redistributes the burden of funding the fixed costs of production onto a smaller share of the user population.

Interestingly, the main argument against strict copyright enforcement is also grounded in the structure of production costs. With near-zero costs of replication, enhancing access to a broader user base (whether or not they are paying or not) increases the social return of software (even as it limits private incentives to incur the initial sunk costs). This argument is particularly salient to the extent that there is a limited impact of piracy (or the level of copyright enforcement) on the level of creative expression or innovation (Waldfogel, 2011). However, many of the most widely diffused software products are produced by profit-oriented firms in which product development is the single most important component of overall costs.

It is also possible that the main impact of piracy arises not simply from enhancing access but facilitating implicit price discrimination (Meurer, 1997; Gopal and Sanders, 1998). If there is a strong negative relationship between price sensitivity and willingness to incur the "costs" of piracy (e.g., time, potential for functionality downgrades), then tolerance of piracy may facilitate a segmentation of the market, in which suppliers charge the monopoly price to the price-insensitive segment, and allow the price-sensitive segment to incur a higher level of transaction costs or a lower level of product quality. This argument is reinforced when the underlying product also exhibits significant network effects, so that even the price-insensitive consumers benefit from more widespread diffusion (Conner and Rumelt, 1991; Oz and Thisse, 1999). Importantly, the role that piracy plays in facilitating price discrimination depends on whether the segmentation that results between pirates and paying users reflects the type of consumer

heterogeneity emphasized in these models. For example, the price discrimination rationale is more pertinent to the extent that piracy is concentrated among low willingness-to-pay consumers (e.g., consumers with a low level of income).

Both the benefits and costs to piracy may be evolving over time with the increasing diffusion of the Internet and broadband connectivity. During the era of desktop computing, software piracy required physical access to at least one copy of the software media (such as a disk or CD), the bulk of piracy involved a limited degree of informal sharing among end users, and so the level of piracy was likely to have been roughly proportional with the level of commercial sales (Peace, Galletta, and Thong, 2003). However, the Internet has has significantly increased the potential for digital piracy, since a single digital copy can now in principle be shared among an almost limitless number of users (and there is no requirement that pirates had any prior or subsequent social or professional contact with each other). Internet-enabled piracy has likely to have increased over the last decade with the diffusion of broadband and the rise of download speeds. Since the mid 2000s, there has been a very significant increase in the diffusion of broadband to mainstream consumers, in the United States and abroad (Greenstein and Prince, 2008), which has reduced the cost of large-scale software piracy. For example, a pirated version of Windows 7 requires downloading an ~ 10 GB file; it is likely that the extent and nature of piracy are qualitatively different when download times are at most a few hours as opposed to a few days. With the rise of the Internet and ubiquitous broadband connections, the potential for software piracy for large software product has become divorced from local sales of physical media.[2]

Despite the potential growing importance of software piracy, and the development of a rapidly emerging and even abundant literature examining the incidence of piracy and the role of copyright enforcement on digital mass media entertainment goods, such as music, movies, and books (Olberhozer-Gee and Strumpf, 2010; Merrill and Raduchel, 2013; Danaher, Smith, and Telang, 2013), Systematic empirical research on software piracy is at an early stage. Nearly all

[2] The rise of the Internet and broadband has also reshaped the interaction between users and software producers. During the desktop era, a software product was essentially static, and users received only limited updates or software fixes. With the rise of the Internet and broadband, software authorization and distribution is routinely achieved through an online connection, and users receive regular security and functionality updates to their software.

prior studies of software piracy depend on a single data source, the Business Software Alliance. The BSA measure is calculated based on an indirect auditing methodology (see BSA, 2011, for a more complete discussion of the BSA methodology). In particular, the BSA undertakes of the "software load" for typical devices within a particular region (broken down by particular types of software), and then compares the level of installed software with observed shipments and payments to software suppliers through authorized channels. In other words, the BSA *infers* the rate of piracy as the "residual" between the level of measured software and paid software in a given country and for particular software segment. Taken at face value, the BSA data suggests that software piracy is a highly significant phenomena; the BSA estimates that the annual "lost sales" due to piracy are worth more than $60 billion USD as of 2011, and that the rate of software piracy is well above 50% of all software in many regions around the world, including Latin America, Asia and Eastern Europe (North America registers the lowest level of piracy as a region). Though the BSA methodology for inferring piracy is imperfect, this approach has the advantage of offering a consistent measurement of piracy across countries, software product segments, and over time. However, as it is an inherently indirect measure, such data cannot be utilized for the types of observational studies that have sharpened our understanding of piracy in the context of areas such as music and movies.

A small literature exploits the BSA data to evaluate the extent of software piracy and the relationship between software piracy and the economic, institutional and technology environment. The most common focus of this literature is to examine the relationship between piracy and the level of economic development (Burke, 1996; Marron and Steele, 2000; Silva and Ramello, 2000; Gopal and Sanders, 1998, 2000). Over time, this literature has been extended to also include more nuanced measures of the institutional environment and the level of technology infrastructure (such as Banerjee, Khalid, and Strum, 2005; Bezmen and Depken, 2006). For example, Goel and Nelson (2009) focus on a broad cross-sectional examination of the determinants of the BSA piracy rate, including not only GDP per capita, but also measures of institutional "quality" such as the Heritage Foundation Property Rights and Economic Freedom Index. As well, Goel and Nelson include a number of measures of technology infrastructure; among other findings, they find that countries with higher prices for telephone service have a lower rate of piracy (i.e., reduced telecommunications access limits piracy). Finally, this literature suggests that measures of variation within the population, such as income inequality,

may also promote piracy; with a higher level of income inequality, the monopoly price for paying customers will be sufficiently high that a higher share of individuals will select into incurring the transactions costs associated with piracy (Andres, 2006).

Overall, our understanding of software piracy is still in a relatively embryonic state. On the one hand, similar to other debates about intellectual property enforcement, theory provides little concrete guidance about optimal policy in the absence of direct empirical evidence. The need for empirical evidence is particularly important given the likelihood that the nature and extent of piracy is changing as the result of the global diffusion of broadband infrastructure. At the same time, the extant empirical literature usefully highlights a number of broad correlations in the data, but has been limited by reliance on an indirect measure of piracy and a loose connection to the theoretical literature.

Three key issues stand out. First, while the prior literature emphasizes both the role of GDP per capita as well as the role of the institutional environment in shaping piracy, the policy debate suggests that it is important to disentangle the relative role of each, For example, if the primary driver of piracy is poverty (i.e., a negative association with GDP per capita), then the case for aggressive anti-piracy enforcement efforts is limited, as piracy is likely serving to simply enhance access to software but is not likely to be a source of significant lost sales. In contrast, if piracy is the result of a low-quality institutional environment, then any observed correlation with GDP per capita may be spurious; instead, the lack of strong legal and property rights institutions may be contributing to a low level of economic development as well as a high level of piracy. In that case, anti-piracy enforcement actions may have a salutary effect by directly enhancing the institutional quality and property rights environment of a given location. Second, the global diffusion of broadband may have changed the nature of piracy. To the extent that piracy is facilitated by broadband diffusion, the rate of piracy should be higher for countries and regions where broadband infrastructure is more prevalent (e.g., where there are higher access speeds and/or lower prices for broadband service). To the extent that changes in "frictions" like the cost of downloading have a non-trivial effect on piracy, it suggests that there are a fair number of individuals "at the margin" between pirating and not pirating, and that piracy can be influenced through institutional changes or frictions imposed by regulation or product design features that make piracy more challenging. Finally, existing studies have not been able to

isolate the impact of anti-piracy enforcement efforts on software piracy. Consistent with recent studies of enforcement efforts in music and movies, an observational study of software piracy alongside shifts over time in the level of enforcement may be able to offer direct evidence about the efficacy of such efforts in restricting the unauthorized distribution of software.

III. The Nature of Software Piracy: A Window onto Windows Piracy

In our initial investigation of software piracy, we found relatively little systematic information within the research literature about how software piracy actually works as a phenomena: how does one actually pirate a piece of software? How hard is piracy, and how does that depend on the type of software that one seeks to pirate, and the type of telecommunications infrastructure that you have access to? How does pirated software actually work (i.e., are there significant restrictions in terms of functionality or updates)? What are the main "routes" to piracy?

The Organization of Windows 7 Distribution Channels

To understand the nature of digital software piracy (and how we will measure piracy with our dataset), we first describe how users are able to receive, authenticate, and validate a legitimate copy of Windows, focusing in particular on the practices associated with individual copies of Windows 7. We then examine the nature of software piracy within that environment.

To authenticate a valid version of Windows requires a Product License Key, a code that allows Microsoft to confirm that the specific copy of Windows that is being installed on a given machine reflects the license that has been paid for that machine. Product License Keys are acquired as part of the process of acquiring Windows software, which occurs through three primary distribution channels: the OEM channel, the Retail channel, and Volume Licensing program.

The OEM Channel. By far, the most common (legal) way to acquire a copy of Windows is through an OEM. OEMs install Windows as part of the process of building and distributing computers, and the vast majority of OEM-built computers include a copy of Windows. To facilitate the authentication of Windows licenses, each OEM receives a number of specialized Product License Keys (referred to as OEM SLP keys) which they can use during this OEM-

installed process. In other words, while OEM SLP keys may be used multiple times, legal use of these keys can only occur on machines that (a) are from that specific OEM and (b) for machines where Windows was pre-installed. Users with OEM-installed Windows have the option to enroll in Windows Automatic Update, which provides security and functionality updates over time.

The Retail Channel. A second channel to legally acquire Windows is through a retail store (which can either be an online store or bricks and mortar establishment). The retail channel primarily serves two types of customers: users who are upgrading their version of Windows (e.g., from Windows XP or Vista), and users who purchased a "naked" machine (i.e., a computer that did not have a pre-installed operating system). Each Retail copy comes with a unique Retail Product Key, which is valid for use for a limited number of installations (usually 10). Retail product keys should therefore be observed only a very small number of times. Users with a Retail Key have the option to enroll in Windows Automatic Update, which provides security and functionality updates over time.

The Enterprise Channel. The final way to acquire Windows is through a contractual arrangement between an organization and Microsoft. For large institutional customers (particularly those that want to pre-install other software for employees), Microsoft maintains a direct customer relationship with the user organization, and issues that organization a Volume License Key Server, which allows the organization to create a specific number of copies of Windows for the organization. While each Volume License Key is unique, most Windows Enterprise customers receive updates through the servers and IT infrastructure of their organization, rather than being enrolled directly in programs such as Windows Automatic Update.

In each distribution channel, each legal Windows user undergoes a process of authenticating their copy of Windows. In the case of OEM-installed Windows or the Retail channel, that authentication process occurs directly with Microsoft. In the case of Windows Enterprise, that authentication occurs via the server system that is established as part of the contract between Microsoft and the volume license customer.

The Routes to Windows Piracy

We define software piracy as the "unauthorized use or reproduction of copyrighted software" (American Heritage Dictionary, 2000). While software piracy has always been an inherent element of software distribution (and has often closely been associated with hacker culture), the nature of piracy changes over time, and reflects the particular ways in which users are able to access software without authorization or payment. There seem to be three primary "routes" to piracy of a mass-market large-format software product such as Windows: Local Product Key Abuse, and Distributed Product Key Abuse, and Intentional Hacking.

Local Product Key Abuse: Since the development of software with imperfect version copying, individual users have occasionally engaged in the unauthorized "local" replication of software from a single legal version. Indeed, the ability to replicate a single copy of Windows across multiple computers is explicitly recognized in the Windows retail licensing contract, which allows users up to 10 authorized replications. Abuse of that license can involve significant replication of the software among social or business networks, or deployment within an organization well beyond the level which is specified in a retail license or which is reported through a volume license key server. Most users who engage in local product key abuse will continue to anticipate receiving software updates from the software vendor. A useful observation is that, when a certain limited number of copies is to be expected (e.g., less than 100), the seller can simply set a price to reflect the scalability of each piece of software once it is deployed in the field.

Sophisticated Hacking. A second route to piracy involves far more active involvement and engagement on the part of users, and involves an explicit attempt to "hack" software in order to disable any authentication and validation protocols that are built into the software. Though this does not seem to be the primary type of piracy that occurs in the context of a mainstream software product such as Windows, it is nonetheless the case that the ability to measure such piracy (particularly using the type of passively generated data that is at the heart of our empirical work) is extremely difficult.

Distributed Product Key Abuse. The third route to piracy is arguably the most "novel" and follows the evolution of piracy for smaller-sized digital products such as music or even movies. In distributed peer-to-peer unauthorized sharing, users access a software copy of Windows through a peer-to-peer torrent site such as the Pirate Bay (an ~ 10GB file), and then

separately download a valid / usable product license key from the Internet. Users then misrepresent that the key was obtained through legal means during the authentication and validation process. In our preliminary investigation of this more novel type of piracy, we found the "ecosystem" for peer-to-peer sharing to be very well-developed, with a significant level of focus in online forums and sites on pirating a few quite specific keys. To get a sense of how piracy occurs, and the role of globally distributed abused product key in that process, it is useful to consider a small number of "dossiers" that we developed for a select number of such keys:

The Lenovo Key ("Lenny"). Approximately three months prior to the commercial launch of Windows 7, Microsoft issued a limited number of OEM System Lock Pre-Installation (SLP) keys to leading OEMs such as Lenovo, Dell, HP, and Asus. Issuing these key allowed these OEMs to begin their preparations to pre-install Windows on machines for the retail market. Within several days of the release of these keys to OEMs, the Lenovo key for Windows 7 Ultimate was released onto the Internet (REFS). This widely reported leak led Microsoft to issue a separate key to Lenovo for the same product (i.e., so that all "legitimate" Lenovo computers would have a different product key than the key that was available on the Internet. As well, Microsoft imposed a functionality downgrade on users who authenticated Windows 7 with the Lenovo key; a message would appear every 30 minutes informing the user that their product key was invalid, and the desktop would be defaulted to an unchangeable black background. Within a few weeks (and still well before the commercial introduction of Windows 7), a number of websites had been established that provided step-by-step instructions about how to download a clean "image" of Windows 7 from a site such as the Pirate Bay or Morpheus and how to not only authenticate Windows with "Lenny" (the Lenovo product key) but also how to disable the limited functionality losses that Microsoft imposed on users that authenticated with the Lenovo key (Reddit, 2013; My Digital Life, 2013).[3] It is useful to emphasize that the Lenovo key leak allowed unauthorized users to gain access to a fully functional version of Windows 7 prior to its launch date and also receive functional and security updates on a regular basis. As of April, 2013, Google reports more than 127,000 hits for a search on the product key associated with

[3] This latter reference is but one of many making claims such as the following: "This is the loader application that's used by millions of people worldwide, well known for passing Microsoft's WAT (Windows Activation Technologies) and is arguably the safest Windows activation exploit ever created. The application itself injects a SLIC (System Licensed Internal Code) into your system before Windows boots; this is what fools Windows into thinking it's genuine." (My Digital Life, 2013). That post is associated with more than 7000 "thank yous" from users.

Lenny, and both the Windows 7 software image and the Lenovo product key are widely available through sites such as the Pirate Bay.

The Dell Key ("Sarah"). Though the Lenovo key received the highest level of media and online attention (likely because it seemed to be the "first" leaked OEM key associated with Windows 7), the Dell OEM SLP key for Windows 7 was also released onto the Internet within weeks after its transmission to Dell, and months before the commercial introduction of Windows 7. Similar to Lenny, a large number of websites were established providing step-by-step instructions about how to download an image of Windows that would work with the Sarah product key, and instructions about how to use the leaked product key, and disable the minor functionality downgrades that Microsoft imposed on users with this key. In contrast to the Lenovo key, the Dell key was never discontinued for use by Dell itself; as a result, there are literally millions of legitimate copies of Windows 7 that employ this key. However, by design, this key should never be observed on a non-Dell machine, or even a Dell machine that was shipped "naked" from the factory (i.e., a Dell computer that was shipped without a pre-installed operating system). For computers that validate with this key, a simple (and conservative) test of piracy is an observation with Sarah as the product key on a non-Dell machine or a Dell machine that was shipped "naked" (a characteristic also observable in our telemetry data).

The Toshiba Key ("Billy"). Not all OEM SLP Windows keys are associated with a high level of piracy. For example, the Toshiba Windows Home Premium Key is associated with a much lower level of piracy. This key was not released onto the Internet until just after the commercial launch of Windows 7 (October, 2009), and there are many fewer Google or Bing hits associated with this product key (less than 10% of the number of hits associated with the Lenovo and Dell keys described above). In other words, while this version of Windows could be pirated at a much more intensive level if other copies (including all copies of Windows Ultimate) were unavailable, the Windows piracy community seems to focus their primary attention on a small number of keys, with a significant focus on leading OEM SLP Ultimate keys.

Overall, these short dossiers of the primary ways in which retail Windows 7 software piracy has actually been realized offers some insight into the nature of Windows piracy as a a phenomena, and guidance as to the relative effectiveness of different types of enforcement actions either by government or by Microsoft. First, while discussions of software piracy that

predates widespread broadband access emphasizes the relatively local nature of software piracy (e.g., sharing of physical media by friends and neighbors, instantiating excess copies of a volume license beyond what is reported to a vendor such as Microsoft), Windows 7 seems to have been associated with a high level of digital piracy associated with a small number of digital point sources. In our empirical work, we will explicitly examine how "concentrated" piracy is in terms of the number of product keys that are associated with the vast bulk of piracy. Second, the globally distributed nature of the ways to access a pirated version of Windows suggests that it may be difficult to meaningfully impact the piracy rate simply by targeting a small number of websites or even product keys. Based on the voluminous material and documentation publicly available on the internet (and reachable through traditional search engines!), it is likely that small changes in the "supply" of pirated software might have little impact on the realized level of piracy.

IV. Data

The remainder of this paper undertakes a systematic empirical examination of the nature and incidence of software piracy. Specifically, we take advantage of a novel dataset that allows us to observe statistics related to a large sample of machines that install Windows on a global basis which receive regular security and functionality updates from Microsoft. Though these data have important limitations (which we discuss below), they offer the opportunity to undertake a direct observational study of software, and in particular the ability to identify whether machines in a given region are employing a valid or pirated version of Windows. We combine this regional measure of piracy with measures of other attributes of machines as well as regional variables describing the institutional, economic, and technology environment to evaluate the nature and relative incidence of piracy.

Windows 7 Telemetry Data

Our estimates of the piracy rates of Windows 7 are computed by drawing on a dataset that captures information about machines (including "hashed" data providing their regional location) that enroll in a voluntary security update program known as Windows Automatic Updates (WAU). When a machine enrolls in the program, a low-level telemetry control, formally known as Windows Activation Technologies (WAT), is installed, which performs periodic validations of the machine's Windows 7 license. During each of these validations, which occur every 90 days by default, data is passively generated about a machine's current hardware, operating system configuration and basic geographic information. This information is transmitted to Microsoft and maintained in a "hashed" manner consistent with the privacy protocols established by Microsoft.[4] More than 400 million individual machines transmitted telemetry information to Microsoft during 2011 and 2012, the period of our sample.

We make use of a research dataset consisting of an anonymized sample of 10 million machines, where, for a given machine, the dataset includes the history of validation attempts for that machine over time. For each of these validation episodes, the dataset includes information on the broad geo-location of the machine at the time of validation,[5] the product key used to

[4] During the validation process, no personal information that could be used to identify an individual user is collected. For more details, see http://www.microsoft.com/privacy/default.aspx .

[5] The geographical location of a machine during its WAT validation attempt is constructed based on the Internet Protocol (IP) address that was used to establish a connection with Microsoft in order to undergo validation. In order to preserve anonymity, only the city and country from which the IP address originates is recorded in our dataset.

activate Windows 7, the version of Windows 7 installed, and a set of machine characteristics, including the manufacturer (OEM) and the machine model, the PC architecture, and whether an OEM installed a version of Windows during the manufacturing process.

Though the Windows telemetry data offers a unique data source for observing software in the field, users face a choice about whether to enroll in the WAU program. Self-selection into Windows Automatic Update engenders two distinct challenges for our data. First, Windows Enterprise customers and others that employ Volume Licensing contracts with Microsoft primarily opt out of WAU and instead manage updating Windows through their own IT departments (a process which allows them, for example, to also include organization-specific updates as well). While we do observe a small number of machines that report a volume license key, we exclude this population entirely from our analysis in order to condition the analysis on users who attempt to validate with either an OEM SLP or Retail Product Key license.[6] In that sense, our empirical analysis can be interpreted as an examination of piracy by individual users and organizations without any direct contract with Microsoft. Second, users with pirated versions of Windows may be less likely to enroll in the Automatic Update program. As such, conditional on being within the sample of users who validate with an OEM SLP or Retail Product key, we are likely estimating a lower bound on the rate of piracy within the entire population of machines.

Defining Piracy

Using the system information recorded by the Windows Telemetry data, we are able to check for the presence of key indicators that provide unambiguous evidence for Piracy. We take a conservative approach to defining each of these in order to ensure that our overall definition of piracy captures only machines consistently possessing what we believe are unambiguous indicators of piracy. Consistent with the discussion in Section III, we therefore identify a machine as non-compliant for a given validation check if it meets one or more of the following criteria:

- For those validating with an OEM Key:

[6] There are number of reasons for doing this, most notably the fact that due to the highly idiosyncratic nature of VL agreements, it is extremely difficult to determine what constitutes an abused VL product key.

    a. Machines Associated with Known Leaked and/or Abused Keys in *and* in which there is a mismatch between the OEM associated with the key and the OEM of the machine.
    b. Machines with an unambiguous mismatch between the Product Key and Other Machine-Level Characteristics
- For those validating with a Retail Key:
    a. Known Leaked and/or Abused Retail Product Keys with more than 100 observed copies within the machine-level WAT population dataset

This definition captures the key cases that we highlighted in Section III. For example, all machines that validate with the "leaked" Lenovo product key, Lenny, will be included in this definition, since this a known leaked key which should not be matched with any machine. As well, this captures all uses of the Dell OEM SLP key (Sarah) in which validation is attempted on a non-Dell machine or a Dell machine that also reports having been shipped "naked" from the factory (in both of these cases, there would be no legal way to receive a valid version of the Dell OEM SLP key). Similarly, any machine that was exclusively designed for Windows XP or Vista (so that no OEM key for Windows 7 was ever legally installed on that machine) but reports an OEM SLP Windows 7 key will be measured as an instance of non-compliancy.

Because we observe the full history of validation attempts for any given machine (though the data for each machine is anonymized beyond its broad regional location), we are able to define piracy as the persistence of non-compliance across all validation attempts by a given machine. In other words, if a machine originally uses a non-compliant version of Windows but then re-authorizes with a valid license key, we define that machine as being in "compliance" in terms of our overall definition. We therefore define a machine as a pirate if, for each of the validation attempts associated with that machine, it satisfies one of the unambiguous non-compliant criteria stated above.

We then construct our key measure, PIRACY RATE, as the aggregation of piracy across the machines within the sample within a given region divided by the number of machines we observe within that region (see Table 1). Overall, weighted by the number of machines per country, the overall piracy rate is just over 25%; if each of the 95 countries in our sample is treated as a separate observation, the average country-level piracy rate is just under 40%.

Machine Characteristics

We are able to observe and then aggregate a number of additional characteristics of machines within a given country. While the decision of whether to pirate Windows and other hardware and vendor choices is clearly endogenous, we nonetheless believe that it is informative to understand what types of machines tend to be associated with pirated software (or not). Specifically, we define four measures that we believe usefully characterize key machine attributes and in which it is useful to compare how the rate of piracy varies depending on machine characteristics:

- FRONTIER MODEL: An indicator equal to 1 for machine models that were exclusively built following the launch of Windows 7.
- LEADING MANUFACTURER: An indicator for whether a machine was produced by one of the leading 20 OEMs, as determined by their market share within the telemetry population.
- FRONTIER ARCHITECTURE: An indicator equal to one for machines with a 64-bit CPU instruction set (also known as an x86-64 processor). Approximately 63% of the machines in our sample are equipped with an x86-64 processor.
- WINDOWS HOME PREMIUM / PROFESSIONAL / ULTIMATE. An indicator for whether the installed version of Windows 7 on a machine is Windows Home Premium, Professional, or Ultimate, respectively.

Economic, Institutional and Infrastructure Variables

Once we have classified each of the machines in our sample, we construct a measure of the incidence of piracy for each of the 95 countries in our sample, which we then incorporate into a dataset of country-level economic, institutional and technology infrastructure variables. Our data on country-level characteristics can be classified into three broad categories: (i) economic and demographic factors; (ii) institutional quality and (iii) technology and innovative capacity. The variable names, definitions, and means and standard deviations are in Table 1. For our basic economic and demographic measures, such as GDP per Capita, the current rate of inflation, and Population, we use standard data from the IMF (for the most current year (2012 in nearly all

cases)). The Gini coefficient for each country is drawn from the CIA World Factbook, and a measure of the lending interest rate is drawn from the Economist Intelligence Unit.

We then incorporate four different measures of overall "institutional quality" of a country. Our first measure, Foundational Competitiveness, is drawn from Delgado et al (2012), who develop a multi-attribute measure that captures a wide range of factors that contribute to the baseline quality of the microeconomic environment, as well as the quality of social and political institutions within a given country. Foundational competitiveness incorporates a wide range of prior research findings on the long-term drivers of country-level institutional quality, and reflects differences across countries in their institutional environment in a way that is distinguishable from simply the observed level of GDP per capita (Delgado, et al, 2012). We also include two additional contemporary measures of institutional quality, including the Rule of Law measure developed as a part of the World Bank Doing Business Indicators (Kaufmann, Kraay, and Mastruzzi, 2008), and a Property Rights Index developed by the Heritage Foundation. Finally, building on Acemoglu, et al (2001), we use Settler Mortality (as measured in the early 19th century) as a proxy for the historical origins of long-term institutional quality. Environments where European Settler Mortality was low led to more investment in setting up more inclusive institutions, resulting in an historical path leading to more favorable institutions over time. We will therefore be able to examine how the historical conditions giving rise to institutions in a given location impacts the rate of piracy today. It is important to note that all of these measures are highly correlated with each other, and our objective is not to discriminate among them in terms of their impact on piracy. Instead, we will evaluate how each of these measures relates to piracy, and in particular consider whether their inclusion reduces the relative salience of contemporary economic measures such as GDP per capita or the Gini coefficient.

Finally, we use a number of measures of the technological and innovative capacity of a country. In terms of telecommunications infrastructure, we use two different measures (from the International Telecommunications Union) of broadband infrastructure, including Broadband Speed and Broadband Monthly Rate. We also investigate alternative measures of the information technology and internet infrastructure, including the percentage of households with a computer and the percentage of the population with access to an Internet connection. Interestingly, for the purpose of evaluating the incidence of operating systems piracy, we believe

that measures associated with broadband connectivity are likely to be particularly important, since low-cost and rapid broadband connection would be required for downloading the large files that are required for Windows 7 piracy. Finally, though we experimented with a wide range of measures, we use the number of USPTO-filed patents per capita as our measure of the innovation orientation of an economy (other measures lead to similar findings).

V. Empirical Results

Our analysis proceeds in several steps. First, we examine some broad patterns in the data, highlighting both the nature and distribution of Windows 7 piracy around the world. We then examine the impact of the economic and institutional environment on piracy, both looking at cross-country comparisons, and a more detailed examinations of cities within and across countries. We also briefly consider how the rate of piracy varies with particular populations of machines and computer characteristics, in order to surface some of the potential mechanisms that are underlying differences in piracy across different environments and among individuals within a given environment. Finally, we undertake a preliminary exercise to assess the causal short-term impact of the primary anti-piracy enforcement effort – the blocking of websites such as Pirate Bay – on observed piracy rates in our data.

The Nature and Incidence of Piracy

We begin in Figure 1, where we consider how piracy is promulgated, focusing on the incidence of individual product keys within the population of pirated machines. The results are striking. Consistent with our qualitative discussion in Section 3, the vast bulk of observable piracy is associated with a relatively small number of product keys. The top "5" keys each account for more than 10% of observed piracy in our data, and more than 90% of piracy is accounted for by the top 12 product keys. At least in part, this extreme concentration is consistent with the idea that global piracy is associated with user communities that provide easy-to-follow instructions associated with individual product keys, and so there is similarity across users in their precise "route" to piracy. Of course, it is likely that enforcement efforts that focused on individual keys would likely simply shift potential pirates to other potential keys (and websites would spring up to facilitate that process).

We continue our descriptive overview in Figures 2A and 2B, where we break out the rate of piracy by the type of OEM and the type of machine. In Figure 2B, we examine how the rate of piracy varies by whether the machine is associated with a leading OEM or not. On the one hand, the rate of piracy is much higher among machines that are shipped from "fringe" rather than leading OEMs. However, this is only a small share of the entire sample (less than 20%). In other words, while the incidence of piracy is much lower among machines produced by leading OEMs, the bulk of piracy is nonetheless associated with machines from leading OEMs. Similarly, Figure 2B describes how the rate of piracy varies depending on whether a particular computer model was introduced before or after the debut of Windows 7. Just over a third of observed copies of Windows 7 are associated with machines that were produced prior to the introduction of Windows 7, and so are likely machines that are "upgrading" from Windows Vista (or an even earlier version of Windows). Interestingly, the rate of piracy for machines associated with clear upgrades is only slightly higher than for machines produced after the introduction of Windows 7 (29 versus 22 percent).

Finally, there is striking variation across the *version* of Windows installed. While the piracy rate associated with Home Premium is quite modest, more than 70% of all piracy is of Windows Ultimate, and, amazingly, nearly 70% of all observed copies of Windows Ultimate are pirated (Figure 3). Windows piracy is associated with machines that, by and large, are not produced by leading manufacturers, and, conditional on choosing to pirate, users choose to install the most advanced version of Windows software.

The Economic, Institutional, and Technological Determinants of Software Piracy

We now turn to a more systematic examination of how the rate of piracy varies by region (the drivers of which are the main focus of our regression analysis in the next section). Figures 4A and 4B highlight a very wide range of variation across regions and countries. While the observed piracy rate in Japan is less than 3%, Latin America registers an average piracy rate of 50%. Interestingly, after Japan, there is a group of advanced English-speaking countries – the United States, Canada, Australia and the United Kingdom – which register the lowest rates of piracy across the globe. At the other extreme is Georgia, where the observed piracy rate reaches nearly 80%. Perhaps more importantly, a number of large emerging countries such as Russia, Brazil and China each are recorded at nearly 60%. Finally, it is useful to note that a number of

reasonably "wealthy" countries (e.g., South Korea, Taiwan, and Israel) boast piracy rates between 30 and 40%. As emphasized in Figure 5 (which simply plots the country-level piracy rate versus GDP per capita), there is a negative but noisy relationship between piracy and overall prosperity.

These broad correlations provide the foundation for the more systematic examination we begin in Table 2. We begin with (2-1) a simple regression which documents the relationship illustrated in Figure 5 -- there is a negative correlation between piracy and GDP per capita. However, as we discussed in Section 3, the relationship between piracy and GDP per capita is subtle: is this relationship driven by the fact that poor countries tend to have poor institutional environments (and so are likely to engage in more piracy) or does this reflect differences in opportunity cost or price sensitivity? To disentangle these effects, we include a simple set of measures associated with country-level institutional quality. In (2-2), we include Foundational Competitiveness (Delgado, et al, 2012) as an overall index that aggregates many different facets of the institutional environment, and in (2-3) we focus on a more straightforward (but perhaps more blunt) measure of institutional quality, the World Bank Rule of Law Index. In both cases, the coefficient on GDP per capita declines by half, and is only marginally significantly different from zero.

We investigate this further in (2-4) where we include a small number of additional controls for the quality of the telecom infrastructure and the degree of innovation orientation of the economy. On the one hand, consistent with earlier studies emphasizing the importance of the telecommunications infrastructure in piracy (Goel and Nelson, 2009), piracy is declining in the price of broadband access, and (not significantly) increasing in average broadband speed. The relative significant of these two coefficients depends on the precise specification (and they are always jointly significanatly different from zero); the overall pattern suggests that piracy is sensitive to the ability to download and manage large files, consistent with the hypothesis that broadband downloads of pirated content is a primary channel through Windows piracy occurs. As well, the piracy rate is declining in patents per capita -- the rate of piracy by consumers and businesses is lower in countries with a higher rate and orientation towards innovation. Finally, computer purchasing and procurement is a capital good; in countries with a higher lending rate, the observed rate of piracy is higher (and this result is robust to the use of the real rather than

nominal lending rate as well). Perhaps most importantly, once controlling for these direct effects on the piracy rate (and across a wide variety of specifications including only a subset or variant of these types of measures), GDP per capita is both small and insignificant.

Rather than income per se, the results from Table 2 provide suggestive evidence that piracy rates are driven by the institutional and technological attributes of a given country, including most importantly whether they have institutions that support property rights and innovation. Poorer countries tend to have weaker institutional environments (Hall and Jones, 1998, among many others), and it is the environment rather than income per se which seems to be correlated with the observed level of piracy. We explore the robustness of this core finding in Table 3 where we examine several alternative ways of capturing the baseline institutional environment of a country, and evaluate the impact of GDP per capita on piracy once such measures are included. In (3-1) and (3-2), we simply replace the Foundational Competitiveness Index with the World Bank Rule of Law measure and the Heritage Foundation Property Rights Index, respectively. In both cases, the broad pattern of results remains the same, and the coefficient on GDP per capita remains very small and statistically insignificant. In (3-3) and (3-4), we extend this analysis by focusing on the subset of countries highlighted in the important work of Acemoglu, et al (2001). Acemoglu, et al argue that the colonial origins of individual countries have had a long-term impact on institutional quality, and they specifically highlight a measure of settler mortality (from the mid-1800s) as a proxy for the "deep" origins of contemporary institutional quality. We build on this idea by directly including their measure of settler mortality. Though the sample size is much reduced (we are left with only 43 country-level observations), the overall pattern of results is maintained, and there is some (noisy) evidence that Settler Mortality itself is positively associated with piracy (i.e., since a high level of settler mortality is associated with long-term weakness in the institutional environment); most notably, in both (3-3) and (3-4), the coefficient on GDP per capita remains small and insignificant.

We further explore these ideas by looking at a few case studies, examples of city pairs that share roughly the same income level but are located in countries with wide variation in their institutional environment. Drawing on city-specific GDP per capita data from the Brookings Global Metro Monitor Project, we identify four city-pairs with similar income levels but wide

variation in measured institutional quality (as measured by the World Bank Rule of Law Index). The results are striking. While Johannesburg and Beijing have roughly the same GDP per capita, the piracy rate in Beijing is recorded to be more than twice as high as Johannesburg (a similar comparison can be made between Sheghzen, China, and Berlin, Germany). A particularly striking example can be drawn between Moscow, Russia, and Sydney Australia, where relatively modest differences in "prosperity" cannot explain a nearly four-fold difference in the observed piracy rate. While these suggestive examples are simply meant to reinforce our more systematic regression findings, we believe that this approach – where one exploits variation within and across countries in both GDP and institutions through regional analyses – offers a promising approach going forward in terms of evaluating the drivers of piracy in a more nuanced way.

Figure 6 sharpens this analysis by plotting the actual piracy rate versus the predicted piracy rate (as estimated from (2-4)). Several notable countries with high piracy rates and intense public attention on the issue (such as China and Brazil) have an observed piracy rate only slightly above that which would be predicted by their "fundamentals." The leading English-speaking countries and Japan have low piracy rates, but those are even lower than predicted by the model. Finally, it is useful to highlight some of the most notable outliers: New Zealand registers a piracy rate far below that which would be predicted by observable factors, and South Korea realizes a level of piracy well above that which would be predicted by observables. Overall, our results suggest that the wide variation of piracy observed across countries reflects a combination of systematic and idiosyncratic factors.

Finally, in Table 5, we examine a number of other potential drivers of piracy that have been discussed in the prior literature. For example, in (5-1) and (5-2), we include measures of population and population density, while in (5-3), we include a measure of country-level income inequality. The inclusion of these measures does not have a material effect on our earlier findings, and are estimated to have a small and insignificant impact. Similar patterns are observed when we include a measure of overall Internet penetration, or a measure of inflation. While the small size of our country-level dataset precludes use from drawing firm conclusions about the relative importance of these additional factors, our overall pattern of results suggests that software piracy is closely associated with fundamental features of the institutional and

technological environment, rather than being primarily driven by measures of income or income inequality.

The Relationship Between Piracy and Machine Characteristics

While the primary focus of the analysis in this paper has been on the impact of the broader economic, institutional and technological environment on country-level piracy, it is also useful to explore the composition of piracy *within* a country, and specifically examine the relationship between piracy and other elements of the machines that users are purchasing and/or upgrading. To do so, we re-organize our dataset to capture the level of piracy within a given country for a certain "type" of machine (e.g., the rate of piracy for computers that are produced by a leading OEM after Windows 7 was introduced. We are therefore able to examine how the rate of piracy varies among different populations of machines; as well, we control for country-level differences in the overall rate of piracy by including country-level fixed effects in our specifications. We weight the regressions so that each country is weighted equally, but we weight each "machine type" within a country according to its share within the country-level population. The results are presented in Table 6. First, consistent with the global averages we presented in Figures 2A and 2B, (6-1) and (6-2) document that the rate of piracy is much higher for machines that are produced by "fringe" manufacturers or assemblers, and is modestly higher among machines that are unambiguously receiving an "upgrade" (i.e., from Windows Vista) as the machine was not produced after Windows 7 was launched. Perhaps more interestingly, there is a very strong interaction effect between these two machine characteristics. Essentially, the highest rate of piracy is observed among "older" machines (i.e. not Windows 7 models) that are produced by "fringe" manufacturers or assemblers. This core pattern of interaction is robust to the inclusion or exclusion of a variety of controls, including a control for whether the machine is has frontier hardware (i.e., a 64-bit versus 32-bit microprocessor) and also if one accounts for the precise version of Windows which is installed. As well, consistent with our earlier descriptive statistics, it is useful to note that the rate of piracy is much higher for machines with Windows Pro and Windows Ultimate; given the global availability of all versions of Windows, it is not surprising that pirates choose to install the highest level of software available.

The Impact of Anti-Piracy Enforcement Efforts on Software Piracy

Finally, we take advantage of time-series variation in our data to directly investigate the impact of the most notable anti-piracy enforcement efforts on the contemporaneous rate of Windows 7 piracy. Specifically, during the course of our 2011 and 2012 sample period, a number of individual countries imposed bans on the Pirate Bay website, the single largest source of pirated digital media on the Internet. Though such policy interventions are broadly endogenous (the bans arise in response to broad concerns about piracy), the precise timing of the intervention is arguably independent of changes over time in Windows 7 piracy in particular, and so it is instructive to examine how a change in the level of enforcement against piracy impacts the rate of Windows 7 software piracy.

We examine three interventions: the ban of Pirate Bay by the United Kingdom in June, 2012, by India in May, 2012, and by Finland in May, 2011. For each country, we define a "control group" of peer countries that can be used as a comparison both in terms of the pre-intervention level of piracy as well as having enough geographic / cultural similarity that any unobserved shocks are likely common to both the treatment and control countries. For the United Kingdom, the control group is composed of France and Ireland; for India, we include both geographically proximate countries such as Bangladesh and Pakistan, as well as the other BRIC countries; and for Finland, we use the remainder of Scandinavia. For each country and for each month before and after the intervention, we calculate the rate of piracy among machines that are first observed within the telemetry data for that month. As such, we are able to track the rate of "new" pirates within each country over time. If restrictions on the Pirate Bay were salient for software piracy, we should observe a decline in the rate of new piracy for those countries impacted by the restriction (relative to the trend in the control countries), at least on a temporary basis. Figure 7 present the results. Across all three interventions, there does not seem to be a meaningful decline in the rate of piracy after the Pirate Bay restriction, either on an absolute basis or relative to the trend followed by the control countries. We were unable to find a quantitatively or statistically significant difference that resulted from these interventions. This "non-finding" suggests that, at least for operating system piracy, the main focus on supply-side enforcement effects may be having a relatively small impact; there may simply be too many alternative sources of pirated Windows, and the pirate user community may be sufficiently pervasive so as to provide potential pirates with new routes to piracy in the face of supply-side enforcement efforts.

VI. Conclusions

The primary contribution of this paper has been to conduct the first large-scale observational study of software piracy. By construction, this is an exploratory exercise, and even our most robust empirical findings are limited to considering the specific domain of piracy of Windows 7. With that said, we have established a number of novel findings that should be of interest to researchers in digitization and piracy going forward.

First, our research underscores the global nature of software piracy, and the role of large-scale global sharing of software and piracy protocols. Relative to the pre-Internet era where piracy may indeed have been pervasive but its diffusion was local (almost by definition), the diffusion of the Internet, the widespread availability of broadband, and the rise of user communities that specifically provide guidance about how and what to pirate have changed the nature of contemporary software piracy.

Second, though the type of data that we use are novel, the bulk of our analysis builds on a small but important literature that has linked the rate of piracy to the economic, institutional and technological environment. At one level, our findings using observational data are broadly consistent with that prior literature; however, our analysis has allowed us to clarify a key empirical distinction: at least in the context that we examined, it is the quality of the institutional environment, rather than income per se, which is more closely linked with piracy. This finding is particularly salient, since a key argument against copyright enforcement depends on income-based price discrimination. Clarifying the distinction between the quality of institutions and income can be seen in a particularly sharp way by comparing cities that have similar income levels but are located in countries with different institutional environments. Though we only undertake a small number of comparisons of this type, our exploratory work looking at cities suggests a future direction of research that can sharpen our identification argument: do cities that are at different levels of income but share the same institutions behave more similarly than cities with the same level of income but with different institutions? Finally, our observational data allows us to directly assess the impact of the most high-profile enforcement efforts against piracy – the choices by individual countries to restrict access to the Pirate Bay over the last several years. Over a number of different "experiments," and examining a number of alternative control groups, we are not able to identify a meaningful impact of these enforcement efforts on

the observed rate of Windows 7 piracy. While such enforcement efforts may be having a meaningful effect on other types of piracy (e.g., movies or music), supply-side enforcement initiatives have not yet meaningfully deterred large-scale operating systems piracy.

More generally, our analysis highlights the potential value of exploiting new types of data that passively capture user behavior in a direct way. By observing the actual choices that users make about what types of software to install (and where and in conjunction with what types of machine configurations), our analysis offers new insight into the both the nature and incidence of software piracy. By and large, our results are consistent with prior measures such as those produced by the Business Software Alliance that suggest that the rate of software piracy is a large and meaningful economic phenomena. Our results suggest that those earlier findings are not simply the result of the BSA methodology but reflect the underlying phenomena. This is particularly important since the rate of piracy is extremely low in the United States, and so claims about piracy are often met with some skepticism. Our direct observational approach not only reinforces those earlier findings, but has allowed us to document both the nature and drivers of piracy in a way that may be instructive for policy and practice going forward.

VII. References

Acemoglu, D., Johnson, S. and Robinson, J. 2001. "The Colonial Origins of Comparative Development: An Empirical Investigation." *The American Economic Review*. 91(5): 1369-1401

American Heritage Dictionary, 4th Edition. 2000. "Software Piracy." Houghton Mifflin.

Andres, A. R. 2006. "Software Piracy and Income Inequality." *Applied Economics Letters*. 13: 101–105.

Banerjee, D., Khalid, A. M., and Strum, J. E. 2005. "Socio-economic Development and Software Piracy: An Empirical Assessment." *Applied Economics*. 37: 2091–2097.

Bezmen, T. L., and Depken, C. A. 2006. "Influences on Software Piracy: Evidence from Various United States." *Economics Letters*. 90: 356–361.

Brookings. 2012. "Slowdown, Recovery, and Interdependence." Report by Emilia Istrate and Carey Anne Nadeu. http://www.brookings.edu/~/media/Research/Files/Reports/2012/11/30%20global%20metro%20monitor/30%20global%20monitor%20appendixb.pdf.

Burke, A. E. 1996. How Effective are International Copy- right Conventions in the Music Industry?" *Journal of Cultural Economics*. 20: 51–66.

Business Software Alliance. 2011. "Ninth Annual BSA Global Software 2011 Piracy Study." http://globalstudy.bsa.org/2011/.

CNET. 2009. "Microsoft Windows 7." Online Professional Review. http://reviews.cnet.com/windows/microsoft-windows-7-professional/4505-3672_7-33704140-2.html.

CNET. 2009b. "Microsoft Acknowledges Windows 7 Activation Leak." News by Dong Ngo. http://news.cnet.com/8301-10805_3-10300857-75.html.

Conner, K. R. and Rumelt, R. P. 1991. "Software Piracy: an Analysis of Protection Strategies." *Management Science*. 37(2): 125–37.

Danaher, B., Smith, M.D., Telang, R and Chen, S. 2012. "The Effect of Graduated Response Anti-Piracy Laws on Music Sales: Evidence from an Event Study in France." *Journal of Industrial Economics*. Forthcoming.

Danaher, B., Smith, M.D. and Telang, R. 2013. "Piracy and Copyright Enforcement Mechanisms." *Innovation Policy and the Economy*. Forthcoming

Delgado, M., Ketels, C., Porter, E and Stern, S. 2012. “The Determinants of National Competitiveness.” National Bureau of Economic Research Working Paper No. 18249.

Goel, Rajeev and Nelson, M. 2009. “Determinants of Software Piracy: Economics, Institutions, and Technology.” *Journal of Technology Transfer*. 34(6):637-658.

Gopal, R. D. and Sanders, G. L. 1998. “International Software Piracy: Analysis of Key Issues and Impacts.” *Information Systems Research.* 9(4): 380–97.

Gopal, R. D. and Sanders, G. L. 2000. Global Software Piracy: You Can’t Get Blood Out of a Turnip.” *Communications of the ACM*. 43(9)” 82–9.

Greenstein, S., & Prince, J. 2006. The Diffusion of the Internet and the Geography of the Digital Divide in the United States. NBER Working Paper No. 12182.

Kaufmann, D., Kraay, A., and Mastruzzi, M. 2009. “Governance Matters VIII: Aggregate and Individual Governance Indicators, 1996-2008.” World Bank Policy Research Working Paper No. 4978.

Hall, Robert E., and Charles I. Jones. 1997. "Levels of Economic Activity across Countries." *American Economic Review*. 87(2): 173-77.

Lakhani, K.R. and Hippel, E.von. 2003. “How Open Source Software Works: “Free” User-to-user Assistance.” *Research Policy.* 32: 923-943.

Landes, W.M. and Posner, R.A. 1989. “An Economic Analysis of Copyright Law.” *Journal of Legal Studies.* 18: 325-366.

Lerner, J. and Tirole, J. 2005. “The Economics of Technology Sharing: Open Source and Beyond.” *Journal of Economic Perspectives.* 19(2): 99-120/

Marron, D. B. and Steel, D. G. 2000. “Which Countries Protect Intellectual Property? The Case of Software Piracy.” *Economic Inquiry*. 38: 159–74.

Merrill, S. and Raduchel, W. 2013. *Copyright in the Digital Era: Building Evidence for Policy.* National Academic Press. Washington, D.C.

Meurer, M.J. 1997. “Price Discrimination, Personal Use and Piracy: Copyright Protection of Digital Works.” *Buffalo Law Review.* https://ssrn.com/abstract=49097.

Microsoft. 2009. “Announcing the Windows 7 Upgrade Option Program & Windows 7 Pricing-Bring on GA!” Windows 7 Blog by Brandon LeBlanc. http://blogs.windows.com/windows/archive/b/windows7/archive/2009/06/25/announcing-the-windows-7-upgrade-option-program-amp-windows-7-pricing-bring-on-ga.aspx.

MyDigitalLife. 2013. “Windows Loader: Current Release Information.” Forum. http://forums.mydigitallife.info/threads/24901-Windows-Loader-Current-release-information.

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.

Oberholzer- Gee, F. and Strumpf, K. 2010. File Sharing and Copyright. *Innovation Policy and the Economy* 10.

Oz, S. and Thisse, J. F. 1999. “A strategic approach to software protection.” *Journal of Economics and Management Strategy*. 8(2): 163–90.

Peace, A. G., Galletta, D. F. and Thong, J. Y. L. 2003. “Software piracy in the workplace: A model and empirical test.” *Journal of Management Information Systems*. 20(1): 153-177.

Peitz, M. and Waelbroeck, P. 2006. “Piracy of Digital Products: A Critical Review of the Theoretical Literature.” *Information Economics and Policy*. 28(4): 449-476.

Reddit. 2013. “Is Anyone Using a Pirated Copy of Windows 7 or 8?” Reddit Thread. http://www.reddit.com/r/Piracy/comments/1baus9/is_anyone_using_a_pirated_copy_of_windows_7_or_8/.

Silva, F. and Ramello, G. B. 2000. “Sound Recording Market: the Ambiguous Case of Copyright and Piracy.” *Industrial and Corporate Change*. 9: 415–42.

Waldfogel, J. 2011. “Bye, Bye, Miss American Pie? The Supply of New Recorded Music since Napster.” NBER Working Paper No. 15882.

Table 1: Summary Statistics[7]

		Country Level N = 95	Country Level-Weighted by Machines N = 95
Dependent Variable			
Piracy Rate	Share of non-compliant (ie- pirated) machines	.38	.25
Machine Characteristics			
Frontier Model	Windows 7 ready model	.57 (.09)	.63 (.48)
Leading Manufacturer	Indicator for whether machine is produced by one of 20 top manufacturers (by market share)	.75 (.12)	.80 (.40)
Frontier Architecture	Indicator for 64-bit CPU architecture	.50 (.16)	.63 (.48)
Windows Ultimate	Indicator for whether machine is Windows Ultimate	.40 (.20)	.25 (.43)
Windows Professional	Indicator for whether machine is Windows Professional	.19 (.06)	.18 (.38)
Windows Home Premium	Indicator for whether machine is Windows Home Premium	.41 (.19)	.58 (.49)
Economic, Institutional, and Demographic Indicators			
GDP Per Capita	GDP Per Capita (IMF)	22215.42 (17898.45)	32498.37 (15628.68)
Foundational Competitiveness Index	Competitiveness Index score (Delgado et. Al., 2012)	.22 (.78)	.55 (.76)
WB Rule of Law	World Bank Rule of Law Index	.36 (.97)	.75 (.99)
Settler Mortality	European Settler Mortality (Acemoglu et. Al., 2001)	111.96 (298.43)	39.12 (107.39)
Property Rights	Heritage Foundation Property Rights Index	53.66 (24.99)	63.47 (26.64)
Gini Coefficient	Gini Coefficient for income inequality (CIA World Factbook, 2008)	38.3 (10.18)	39.57 (8.58)
Lending Rate	Lending Interest Rate (EIU)	8.83 (5.93)	7.38 (7.72)
Inflation	Annual (%) change in CPI (IMF)	4.63 (3.66)	3.53 (5.77)
Population (in millions)	Total Population (IMF)	61.32 (189.35)	167.52 (219.43)
Population Density	People per Sq. KM (WDI)	301.25 (1021.36)	188.29 (696.51)
Measures of Innovative & Technological Capacity			
Patents Per Capita	USPTO-filed patents per one million inhabitants (USPTO)	34.5 (70.09)	118.37 (116.96)
Broadband Speed	Wired broadband speed per 100 Mbit/Sec (ICT/ITU)	4.96 (12.84)	6.73 (16.98)
Broadband Monthly Rate	Wired broadband monthly subscription charge (USD) (ICT/ITU)	24.15 (11.47)	24.12 (10.07)
Computer	(%) of households with a computer (ICT/ITU)	54.07 (27.30)	66.94 (22.08)
Internet	Internet users (%) of population (ICT/ITU)	52.21	65.72

[7] With exception to the CIA Factbook's Gini coefficient, which was computed in 2008, we take the average of all indicators over our sample period (2011-2012), unless otherwise indicated.

(23.99) (20.74)

TABLE 2

SOFTWARE PIRACY AND THE ECONOMIC, INSTITUTIONAL, AND INFRASTRUCTURE ENVIRONMENT

	Windows 7 Piracy Rate			
	1	2	3	4
Ln GDP Per Capita	**-.151**	**-.082**	**-.061**	-.026
	(.014)	**(.021)**	**(.018)**	(.02)
Competition Index		**-.096**		***-.039***
		(.017)		***(.019)***
WB Rule of Law			**-.097**	
			(.014)	
Ln Patents Per Capita*				**-.023**
				(.007)
Ln Broadband Download Speed				.008
				(.009)
Ln Broadband Monthly Rate				**-.087**
				(.026)
Lending Rate				**.005**
				(.001)
Observations	**95**	**95**	**95**	**95**
R-Squared	**.599**	**.674**	**.708**	**.762**

Notes: Bold, Bold-Italic, and Italic numbers refer to coefficients significant at 1%, 5% and 10% levels. Robust standard errors in parentheses.

* Ln Patents Per Capita is defined as Ln(1 + Patents Per Capita)

TABLE 3

ALTERNATIVE MEASURE OF INSTITUTIONAL QUALITY

	Windows 7 Piracy Rate			
	1	2	3	4
Ln GDP Per Capita	-.015	-.018	-.019	-.006
	(.018)	(.017)	(.039)	(.039)
WB Rule of Law	**-.074**			
	(.021)			
Ln Patents Per Capita*	-.013	***-.016***	-.022	-.023
	(.009)	***(.008)***	(.014)	(.01)
Ln Broadband Download Speed	.013	.014	4e-4	.005
	(.009)	(.009)	(.015)	(.014)
Ln Broadband Monthly Rate	-.087	**-.076**	***-.106***	**-.122**
	(.025)	**(.024)**	***(.045)***	**(.045)**
Lending Rate	**.005**	**.005**		**.005**
	(.001)	**(.001)**		**(.002)**
Prop Rights		**-.003**		
		(.001)		
Ln Settler Mortality			*.045*	.04
			(.026)	(.027)
Observations	**95**	**93**	**43**	**43**
R-Squared	**.786**	**.79**	**.664**	**.694**

Notes: Bold, Bold-Italic, and Italic numbers refer to coefficients significant at 1%, 5% and 10% levels. Robust standard errors in parentheses.

* Ln Patents Per Capita is defined as Ln(1 + Patents Per Capita)

TABLE 4

City Pair Comparisons

Rule of Law and GDP Per Capita Comparisons by Piracy Rate					
Pair	**City**	**Country**	**GDP Per Capita (Thousands $, PPP rates)**	**Rule of Law Index (WB)**	**Piracy Rate**
1	Johannesburg	South Africa	17.4	0.10	0.24
1	Beijing	China	20.3	-0.45	0.55
2	Kuala Lumpur	Malaysia	23.9	0.51	0.29
2	Sao Paulo	Brazil	23.7	0.013	0.55
3	Moscow	Russia	44.8	-0.78	0.56
3	Sydney	Australia	45.4	1.77	0.15
4	Shengzhen	China	28	-0.45	0.44
4	Berlin	Germany	33.3	1. 69	0.24

TABLE 5

OTHER POTENTIAL DRIVERS OF PIRACY

	Windows 7 Piracy Rate				
	1	2	3	4	5
Ln GDP Per Capita	-.025	-.028	-.026	-.012	-.025
	(.02)	(.02)	(.026)	(.024)	(.02)
Competition Index	***-.039***	*-.034*	***-.052***	-.032	*-.036*
	(.019)	*(.02)*	***(.022)***	(.021)	*(.019)*
Ln Broadband Download Speed	.008	.01	.009	.01	.008
	(.009)	(.009)	(.01)	(.009)	(.009)
Ln Broadband Monthly Rate	**-.087**	**-.094**	**-.09**	**-.087**	**-.087**
	(.026)	**(.027)**	**(.027)**	**(.025)**	**(.026)**
Ln Patents Per Capita*	**-.023**	**-.024**	*-.02*	**-.022**	**-.023**
	(.007)	**(.007)**	*(.01)*	**(.007)**	**(.007)**
Lending Rate	**.005**	**.005**	**.005**	**.005**	**.005**
	(.001)	**(.001)**	**(.002)**	**(.001)**	**(.002)**
Ln Population	3e-4				
	(.007)				
Ln Population Density		-.009			
		(.006)			
Gini Coefficient			2e-4		
			(.001)		
Internet				-.001	
				(.001)	
Inflation					.002
					(.003)
Observations	**95**	**95**	**85**	**95**	**95**
R-Squared	**.762**	**.767**	**.769**	**.764**	**.762**

Notes: Bold, Bold-Italic, and Italic numbers refer to coefficients significant at 1%, 5% and 10% levels. Robust standard errors in parentheses.

* Ln Patents Per Capita is defined as Ln(1 + Patents Per Capita

TABLE 6

PIRACY AND MACHINE CHARACTERISTICS

	Windows 7 Piracy Rate			
	1	2	3	4
OEM Leading Manufacturer	**-.391**		**-.274**	**-.234**
	(.011)		**(.011)**	**(.009)**
Windows 7 Model		**-.227**	.003	-.005
		(.008)	(.011)	(.008)
OEM Leading * Win 7 Model			**-.180**	**-.047**
			(.013)	**(.009)**
Frontier Architecture				**-.094**
				(.005)
Windows Professional				**.048**
				(.008)
Windows Ultimate				**.426**
				(.07)
Observations	**4518**	**4518**	**4518**	**4518**
R-Squared	**.487**	**.360**	**.534**	**.862**

Notes: Bold, Bold-Italic, and Italic numbers refer to coefficients significant at 1%, 5% and 10% levels. Robust standard errors in parentheses.

* Ln Patents Per Capita is defined as Ln(1 + Patents Per Capita)

FIGURE 1

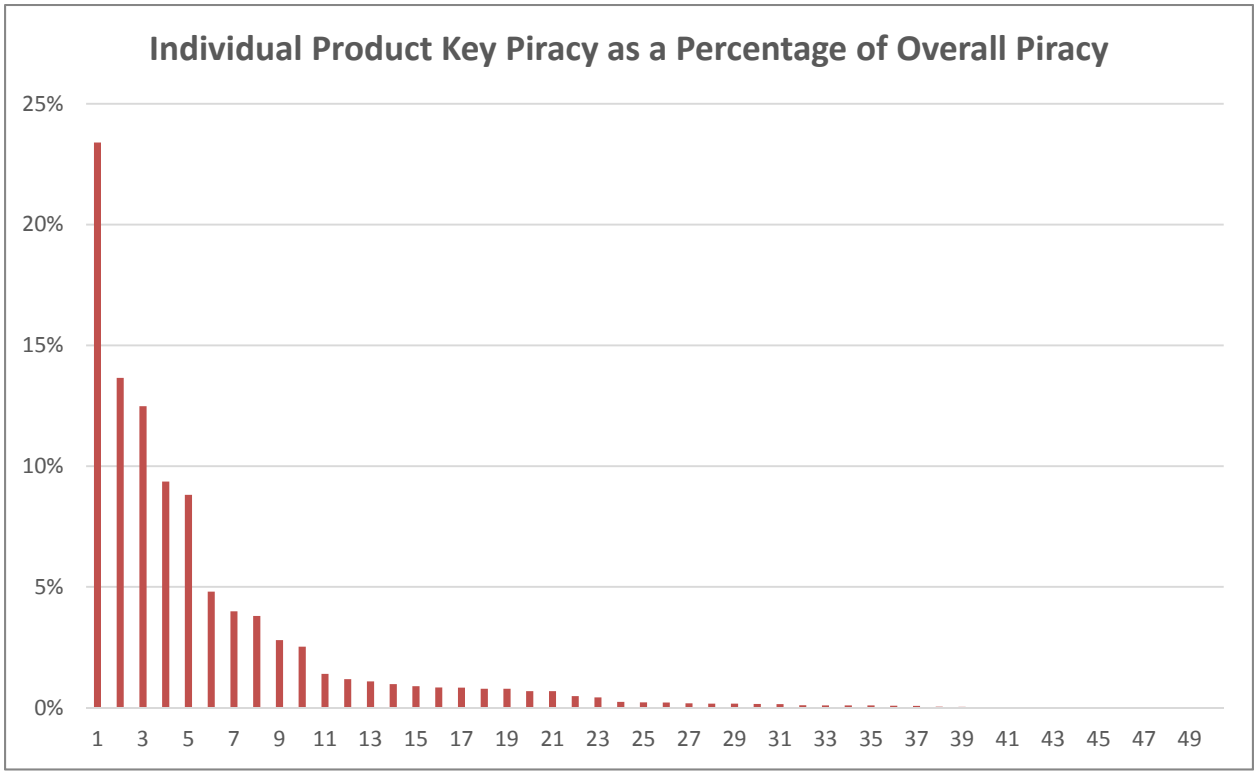
Individual Product Key Piracy as a Percentage of Overall Piracy
25%
20%
15%
10%
5%
0%
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49

FIGURE 2A

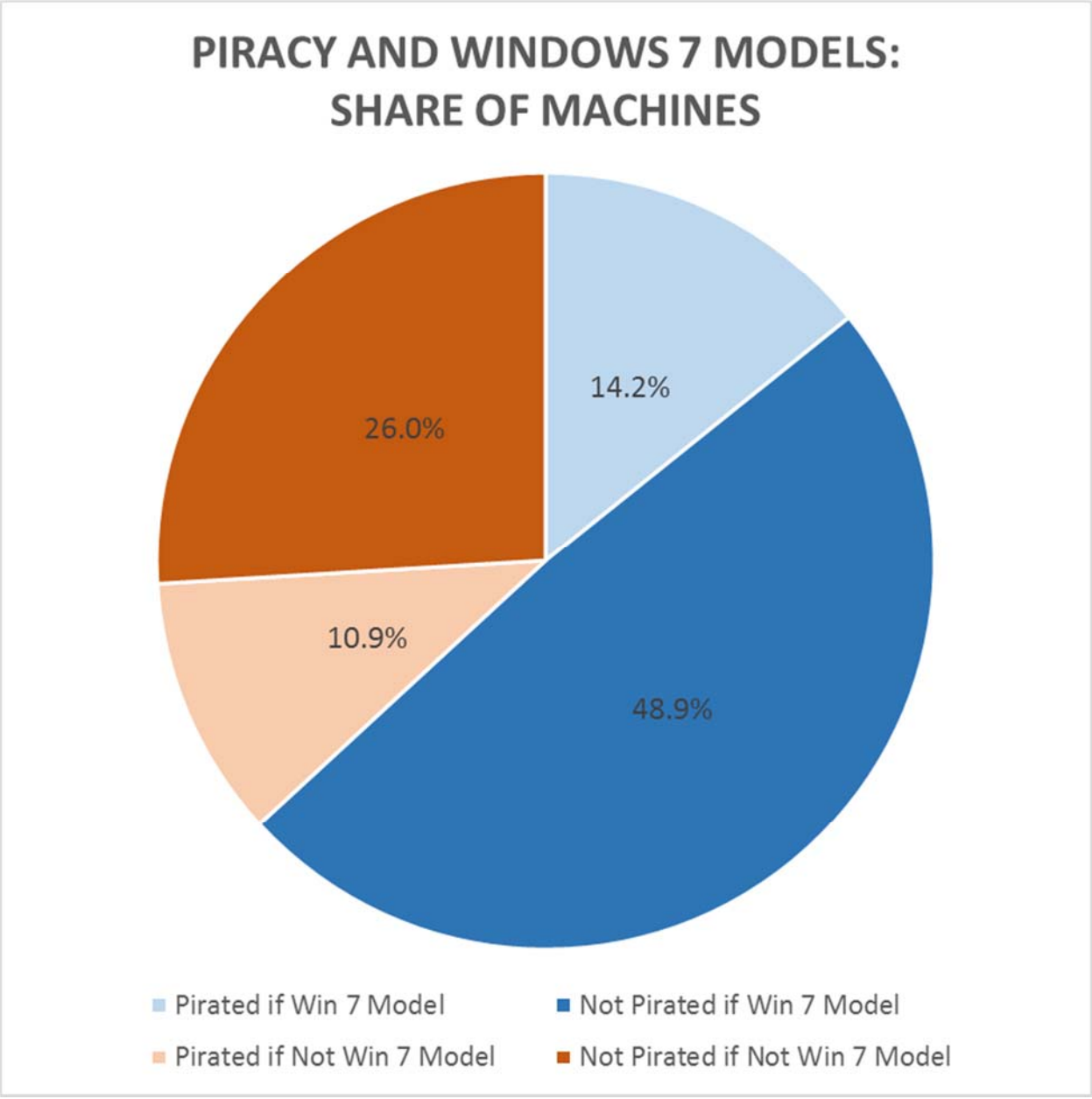
PIRACY AND WINDOWS 7 MODELS:
SHARE OF MACHINES
14.2%
48.9%
10.9%
26.0%
Pirated if Win 7 Model
Not Pirated if Win 7 Model
Pirated if Not Win 7 Model
Not Pirated if Not Win 7 Model

FIGURE 2B

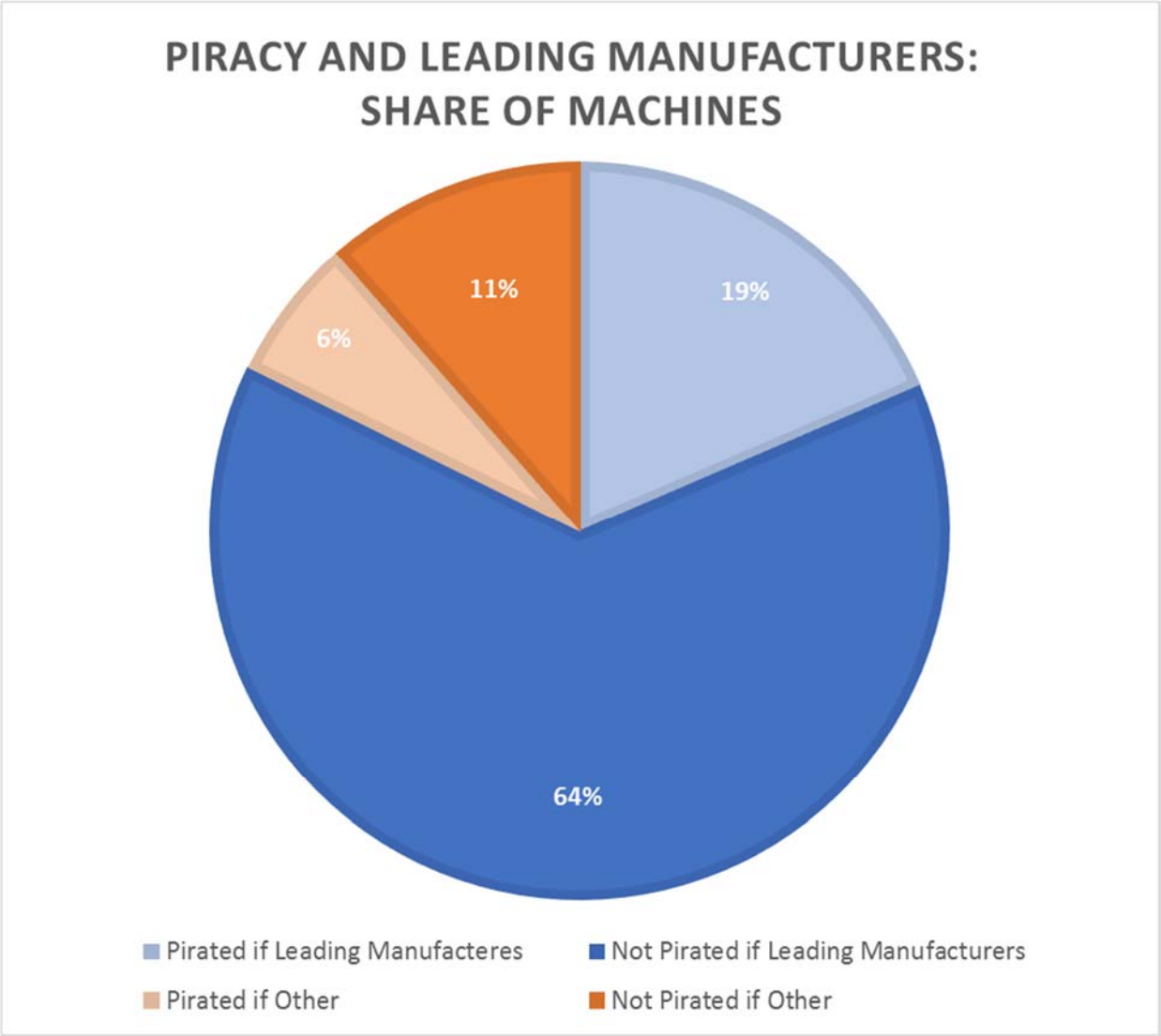
PIRACY AND LEADING MANUFACTURERS:
SHARE OF MACHINES
19%
64%
6%
11%
Pirated if Leading Manufacteres
Not Pirated if Leading Manufacturers
Pirated if Other
Not Pirated if Other

FIGURE 3

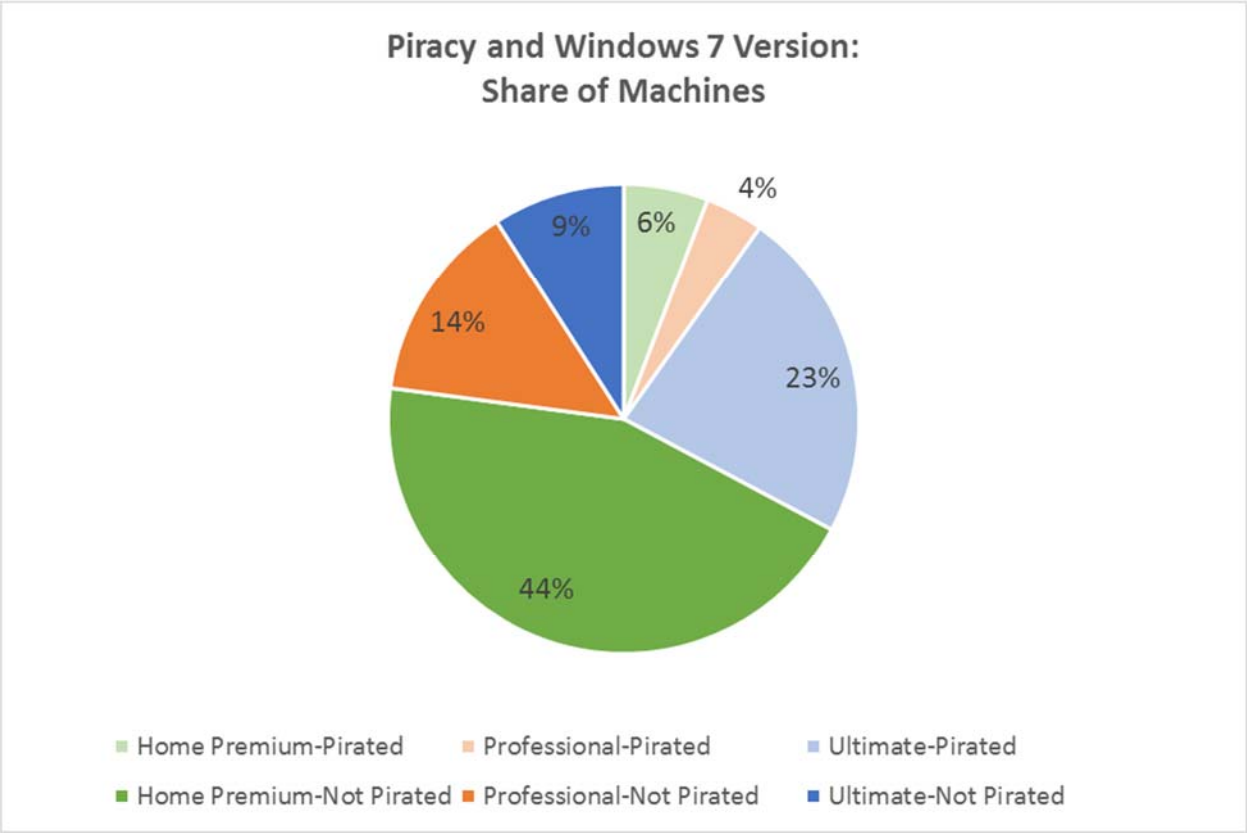
Piracy and Windows 7 Version:
Share of Machines
4%
6%
9%
14%
23%
44%
Home Premium-Pirated
Professional-Pirated
Ultimate-Pirated
Home Premium-Not Pirated
Professional-Not Pirated
Ultimate-Not Pirated

FIGURE 4A

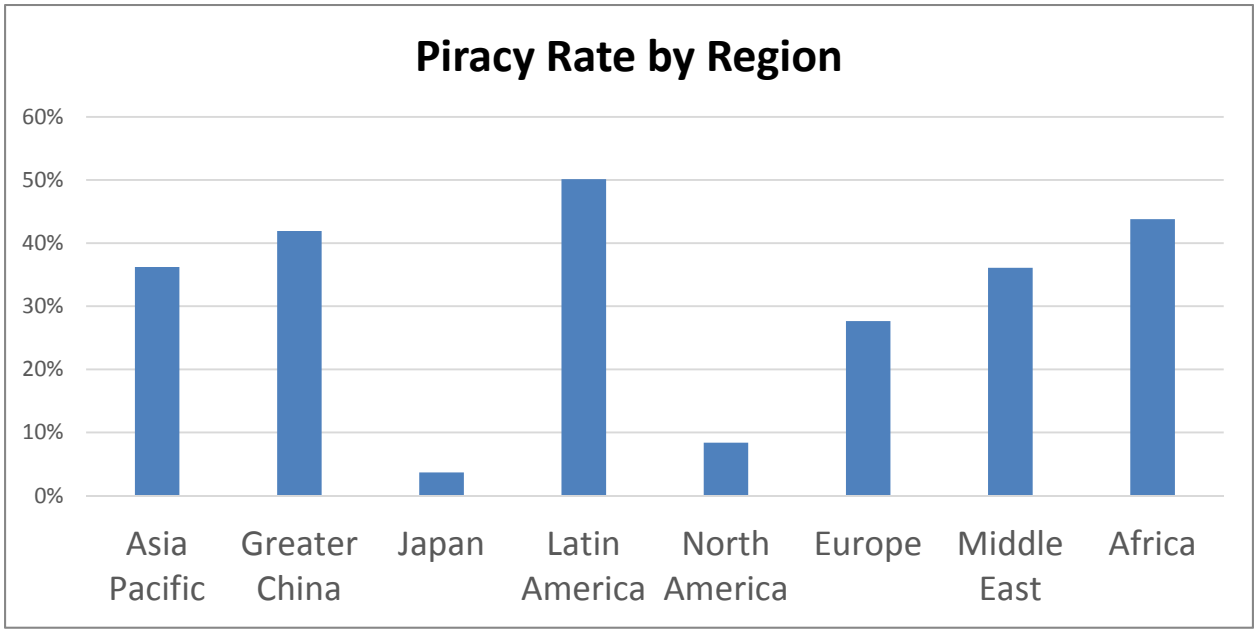
Piracy Rate by Region
60%
50%
40%
30%
20%
10%
0%
Asia Pacific
Greater China
Japan
Latin America
North America
Europe
Middle East
Africa

FIGURE 4B

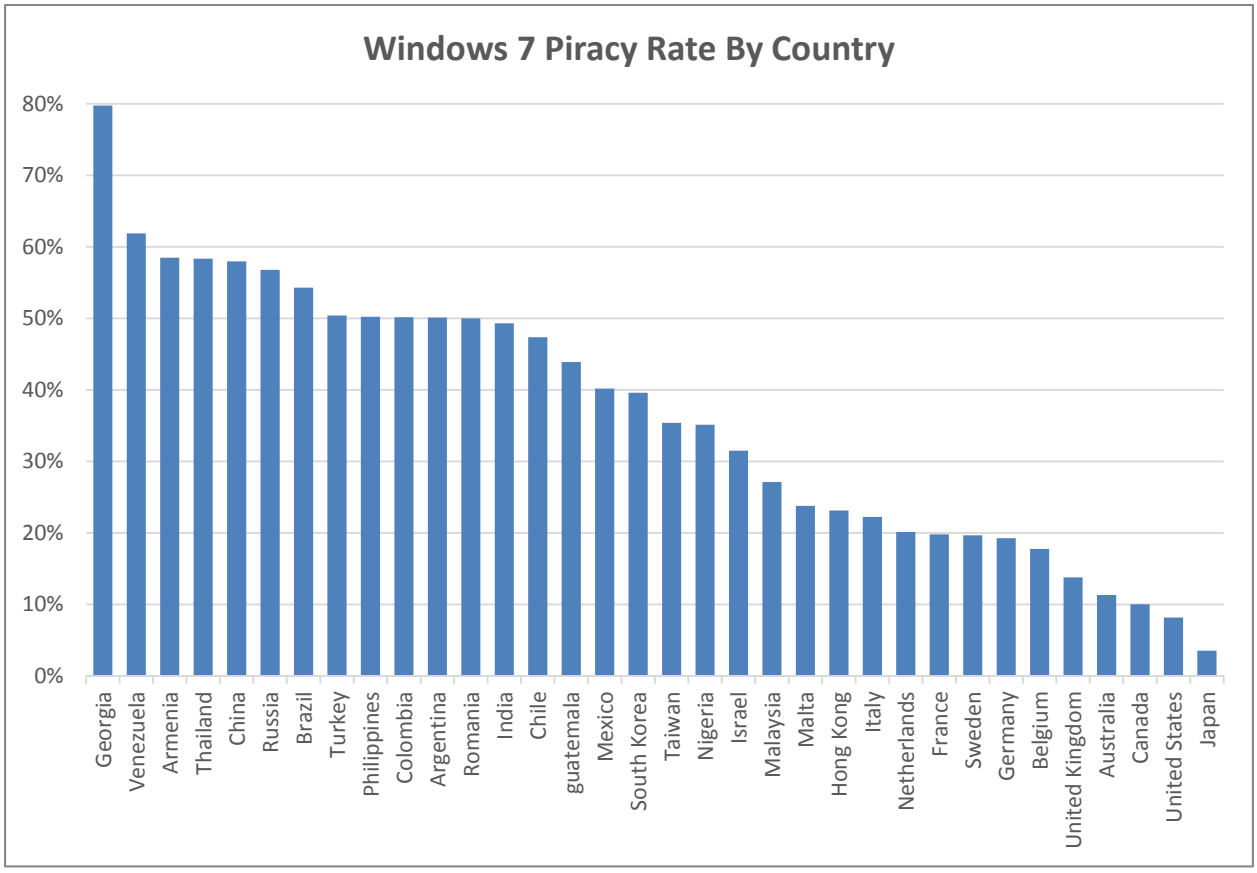
Windows 7 Piracy Rate By Country
80%
70%
60%
50%
40%
30%
20%
10%
0%
Georgia
Venezuela
Armenia
Thailand
China
Russia
Brazil
Turkey
Philippines
Colombia
Argentina
Romania
India
Chile
guatemala
Mexico
South Korea
Taiwan
Nigeria
Israel
Malaysia
Malta
Hong Kong
Italy
Netherlands
France
Sweden
Germany
Belgium
United Kingdom
Australia
Canada
United States
Japan

FIGURE 5

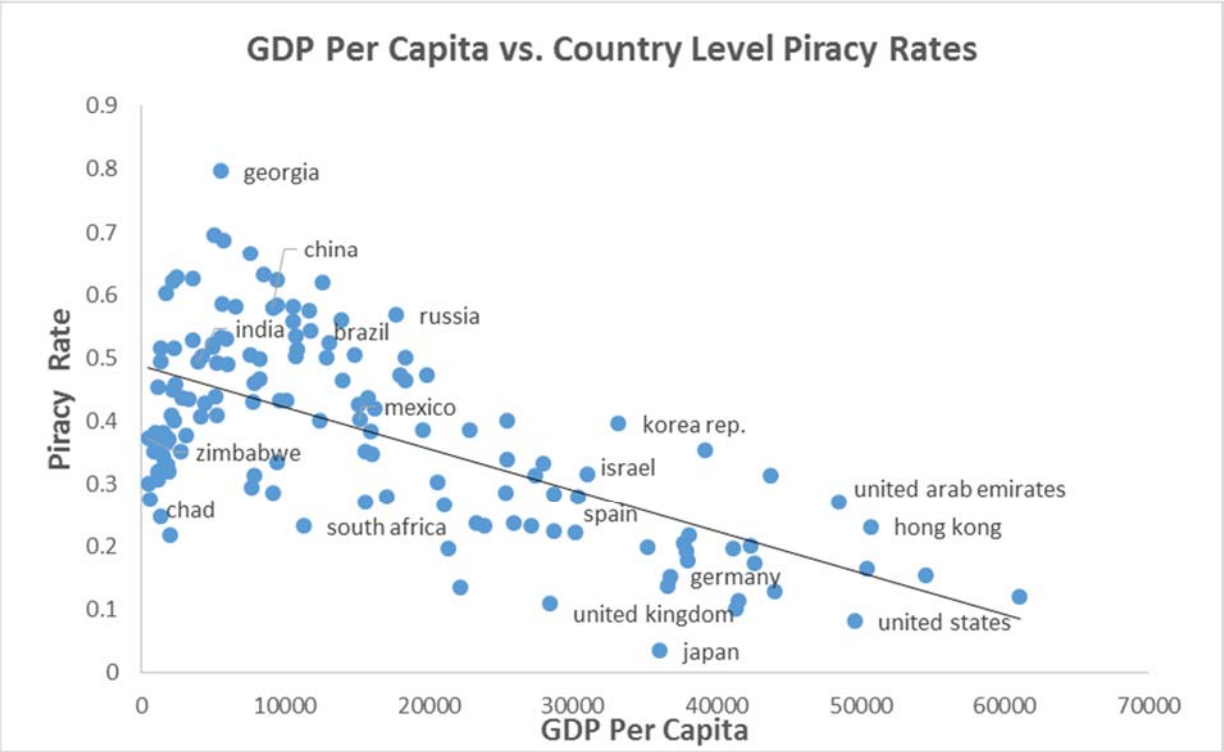
GDP Per Capita vs. Country Level Piracy Rates
Piracy Rate
GDP Per Capita
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
10000
20000
30000
40000
50000
60000
70000
georgia
china
india
brazil
russia
mexico
korea rep.
zimbabwe
israel
united arab emirates
chad
spain
hong kong
south africa
germany
united kingdom
united states
japan

FIGURE 6

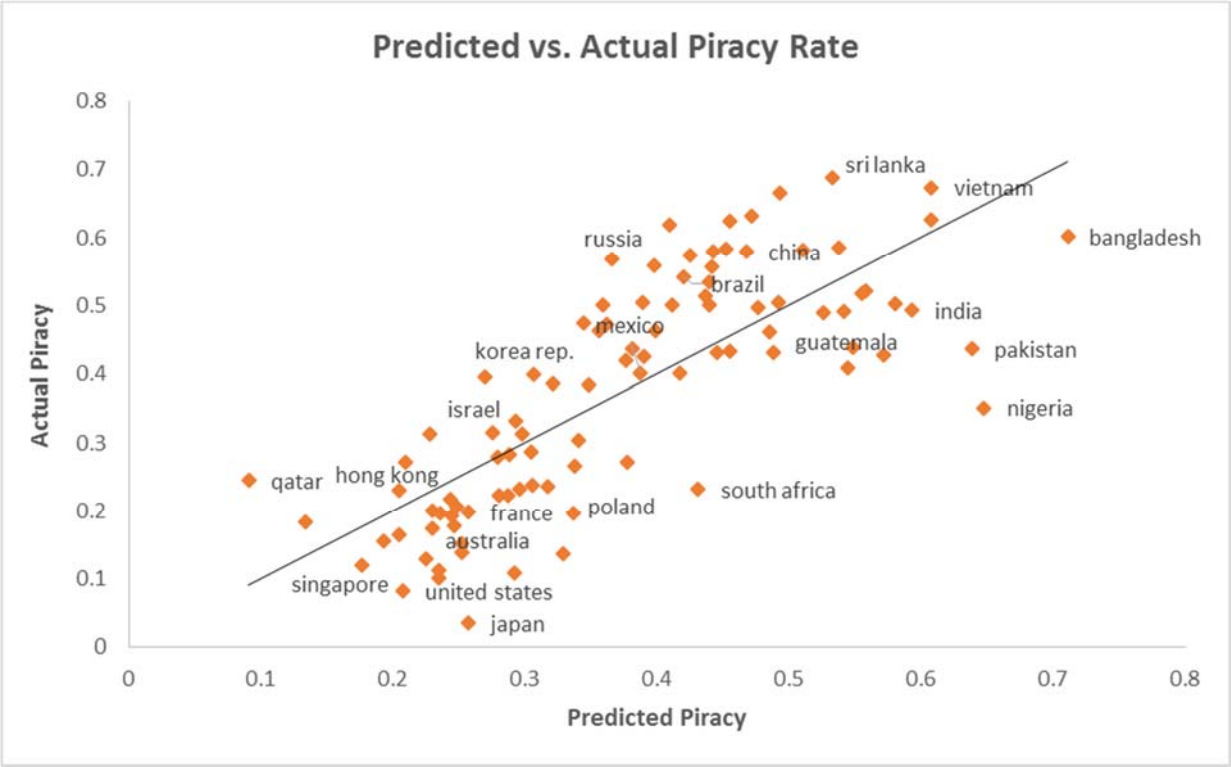
Predicted vs. Actual Piracy Rate
Actual Piracy
Predicted Piracy
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
sri lanka
vietnam
bangladesh
russia
china
brazil
india
mexico
guatemala
pakistan
korea rep.
nigeria
israel
qatar
hong kong
south africa
france
poland
australia
singapore
united states
japan

FIGURE 7A

UK: PIRACY RATE

EFFECTIVE BAN DATE: JUNE 2012

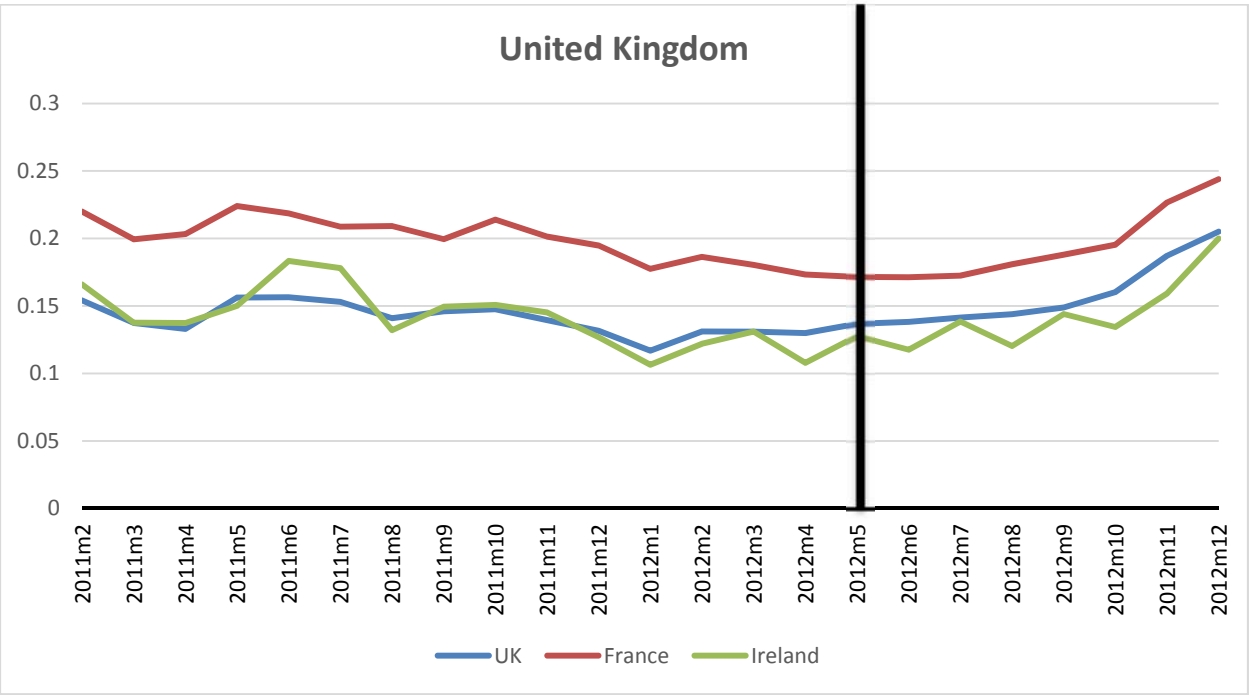

FIGURE 7B

INDIA: PIRACY RATE

EFFECTIVE BAN DATE: MAY 2012

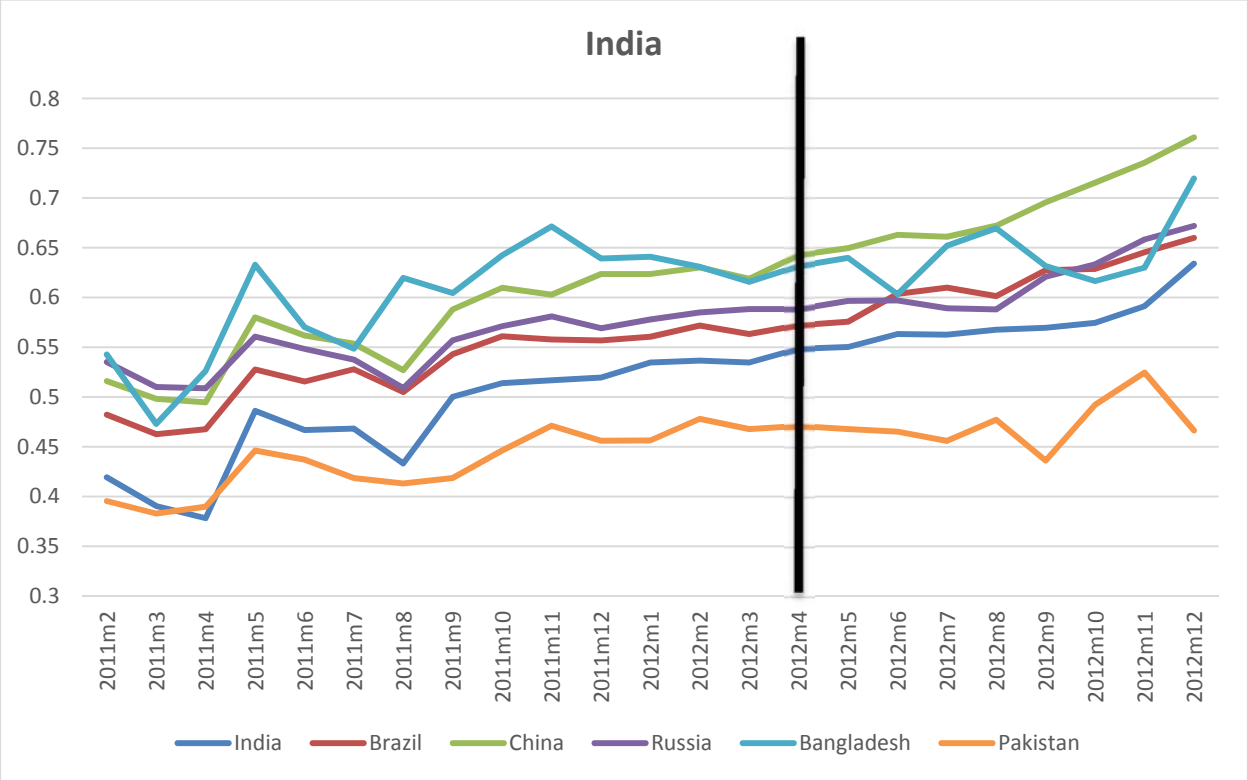

FIGURE 7C

FINLAND: PIRACY RATE

EFFECTIVE BAN DATE: NOVEMBER 2011