EJW
ECON JOURNAL WATCH
Scholarly Comments on Academic Economics

ECON JOURNAL WATCH 13(3)
September 2016: 373–396

Why the Oberholzer-Gee/ Strumpf Article on File Sharing Is Not Credible

Stan J. Liebowitz[1]

LINK TO ABSTRACT

The most influential article on the effect of piracy (file sharing) on the sound recording industry appeared as the lead article in the *Journal of Political Economy* in 2007.[2] Its authors, Professors Felix Oberholzer-Gee and Koleman Strumpf—referred to here as OS—performed a regression analysis using data from 2002 and concluded that piracy had no impact on record sales, even though the birth of file sharing coincided with what in hindsight can be described as a financial near-collapse of the sound recording industry (worldwide sales down 60–70 percent in real dollars). The financial decline, along with the lack of any other empirically plausible cause of the decline, has led almost all current industry participants and most researchers to conclude that piracy has been the primary if not the complete cause of the decline,[3] although many industry critics, who are well represented in the academy, still cling to the notion that piracy was not a significant cause of the decline and often cite the OS article as support for that view.

1. University of Texas at Dallas, Richardson, TX 75080. The author wishes to thank the Center for the Economic Analysis of Property Rights, at the University of Texas at Dallas, for financial support, and Jonathan Lee for access to his data. No other entity or person has paid for this research. The author often consults on media topics and has been paid to testify as an expert witness both for and against the sound recording industry.

2. According to Google Scholar, Oberholzer-Gee and Strumpf (2007) has received more citations than any other paper on piracy in the economics literature of which I am aware. According to Web of Science (in May 2016), it was the fifth-most cited paper that the *Journal of Political Economy* published in 2007.

3. Most of the decline took place before 2008, when there was no credible alternative explanation for the decline. For a summary of the empirical studies on this topic see Liebowitz (2016a), which concludes that the average academic study found that piracy was responsible for the entire decline in industry sales through the time periods studied in those articles.

The OS analysis was notable for using data from an actual pirate server in its main analysis, although it also included several secondary "quasi experiments" using other data (OS 2007, 33–38). This comment focuses on issues related to the primary OS data and analysis. A straightforward replication of OS's work is not possible because OS have not made their data from the pirate server public, in spite of having promised to do so.[4] I should also note that I have elsewhere replicated their secondary tests using publicly available data, and those replications found results very different than those reported by OS (see Liebowitz 2016b).[5]

My examination here investigates the small bits of their main data that they have made public—summary statistics, estimated coefficients—and also investigates the logic behind their analysis. I conclude that irregularities in their data, conflicts between their evidence and their hypotheses, and a flawed research design prevent the OS analysis from being an informative research work.

A brief word on the method used by OS will be useful. OS obtained download data from a very small (.01 percent market share) centralized file-sharing system, covering a 17-week period at the end of 2002 (that is, September through December). OS extracted the American downloads and matched them to songs from a sample of generally successful music albums over the 17 weeks. Extracting the information from server logs and matching them to music albums required ingenuity and effort on OS's part.

Comparing the weekly sales of albums to the number of unauthorized downloads of the songs on those albums suffers from a well-known simultaneity problem because the popularity of a song influences both its sales and downloads. In an attempt to overcome this problem, OS chose a dazzlingly recondite instrument for downloads—the number of German pre-university students on school vacation each week during the 17-week period of their analysis. The variable representing the number of German pre-university students on school vacation is referred to in this paper by the abbreviation NGSV, which stands for Number of German Schoolkids on Vacation. After performing their regressions, OS concluded, in contrast to most of the literature, that piracy had no impact on record sales.

4. At a public forum in 2004, Strumpf stated an intention to make the data available "as soon as the legal environment quiets down" (search the text at this **link** for "soon"). Four years after that promise, OS claimed to have signed a non-disclosure agreement, although when reporters asked to see the agreement, OS were unwilling to provide it (see Häring 2008; Glenn 2008).

5. In 2007 and 2010 I submitted to the JPE comments regarding different aspects of the OS paper (and placed them on SSRN). Both comments were rejected, in my opinion incorrectly. As the research community has increasingly shown an interest in replication, evidenced in the growth of sites such as Retraction Watch and PubPeer, and with the addition of some new results, I thought this might be a propitious time to resubmit these comments to outlets that have expressed a positive view of replication. The current paper contains material in common with the 2010 comment, and a second paper, Liebowitz (2016b), contains material related to the 2007 comment.

The reasoning behind the use of the NGSV instrument for downloads is not difficult to follow, but the logic requires several assumptions, some of which are never tested by OS. First, German K–12 students are presumed to provide a large number of files to American pirates. OS provide minimal support for this claim, instead providing evidence that the entire German population provides a relatively large number of files to American pirates. Second, OS assert that German students are more likely to keep their computers attached to file-sharing networks during school vacations than on days when they attend school, but OS have never provided any evidence in support of the assumption and it remains untested. If the home computers of German students instead were continuously hooked up to file-sharing services, German school vacations could not have had any impact on American pirates.

With these two assumptions, OS argue that German school holidays increase the computer usage of German kids who are also pirates, increasing the availability of their music files to American pirates. This extra availability of German school-kids' files is supposed to greatly enhance the ease of American piracy, and therefore its amount, so that if American piracy had an impact on American record sales this impact would be found by econometrically examining the influence of German school holidays (instrumenting for American downloads) on American record sales. If German school holidays (working through American downloads) do not affect American record sales (as OS found), OS would (and did) conclude that American piracy does not affect American record sales. That, in a nutshell, is the OS methodology.

The plan of my critique proceeds as follows. First, I demonstrate that the OS measurement of piracy—derived from their never-released dataset—appears to be of dubious quality since the aggregated weekly numbers vary by implausibly large amounts not found in other measures of piracy and are inconsistent with consumer behavior in related markets. Second, the average value of NGSV (German K–12 students on vacation) reported by OS is shown to be mismeasured by a factor of four, making its use in the later econometrics highly suspicious. Relatedly, the coefficient on NGSV in their first-stage regression is shown to be too large to possibly be correct: Its size implies that American piracy is effectively dominated by German school holidays, which is a rather farfetched proposition. Then, I demonstrate that the aggregate relationship between German school holidays and American downloading (as measured by OS) has the opposite sign of the one hypothesized by OS and supposedly supported by their implausibly large first-stage regression results.

After pointing out these questionable results, I examine OS's chosen method. A detailed factual analysis of the impact of German school holidays on German files available to Americans leads to the conclusion that the extra files

available to Americans from German school holidays made up less than two-tenths of one percent of all files available to Americans. This result means that it is essentially impossible for the impact of German school holidays to rise above the background noise in any regression analysis of American piracy.

The implausibility of OS's download data

The novelty of the OS article lies in its authors' access to a dataset that consists of downloads from actual pirate servers during the final 17 weeks of 2002. The use of a dataset based on actual downloads, everything else equal, is preferable to using proxies for downloads, such as Internet usage or downloads reported by respondents to surveys. But actual pirate downloads suffer from a strong simultaneity bias with album sales, and using albums as the unit of analysis, as OS do, brings its own potential pitfalls.[6]

Nevertheless, it is clear that for OS's project to have any validity, their download data must be an accurate representation of actual U.S. piracy behavior. OS (2007, 7) acknowledge this point by stating that "an important question is whether our sample [of pirate downloads] is representative of data on all P2P networks," and they send the reader to an appendix of an earlier version of their paper for more details, where they had stated:

> Our inferences about the effect of file sharing on record sales would be invalid if we had an unrepresentative sample of downloads. …
>
> [W]e considered whether our most popular downloads were also common in other file sharing networks. To do this, we compared the top ten downloads each week in our data with the concurrent list from http://www.bigchampagne.com. BigChampagne generates their own weekly top lists…. Over our seventeen week sample period, two-thirds of our top ten downloads also appear in the BigChampagne top ten list.[7]

6. Using albums as the unit of analysis creates a potential fallacy of composition. The impact of downloading on individual albums may be very different than the impact of downloading on the entire industry, particularly if downloading provides publicity about individual albums that affects their relative market shares. Extra downloading of an album that, say, puts it on top-100 download charts, might increase the market share (and possibly the actual sales) of the album even if downloads as a whole are negatively related to album sales as a whole. If so, an analysis based on albums could provide very misleading results even if all the data and analyses were pristine (see Hammond 2016 for a more complete discussion).

7. The quotation comes from Appendix A of the June 2005 version of their paper (**link**). OS use several approaches to verify the similarity of their download data to the rest of the piracy universe, but mainly examine whether their rankings of downloaded albums is similar to other rankings of pirated songs, not whether the weekly changes in downloads are similar to those in other data sources. In one instance they

Although there is both a time series component and a cross section component in their panel data, OS's attempt to check on the reasonableness of their download data appears to be largely limited to the cross section element (across albums) as indicated by their focus on the similar list of top albums between the various piracy networks. Their results, however, are also dependent on the time series component of the data.

In an early version of their paper, OS provided the aggregate weekly number of American downloads from the servers.[8] These weekly downloads were aggregated from the 260,889 American pirate downloads that OS culled from their full set of worldwide downloads over the 17-week period, from their target servers.[9]

Before examining OS's download data, it is useful to think a little about the likely characteristics of weekly American pirate downloading, based upon what we know about consumption of entertainment products. In other words, what kind of weekly trends in downloading might we expect, at a national level, over these 17 end-of-year weeks?

By way of analogy, weekly television viewing, the leading entertainment activity of Americans, is fairly constant throughout the year, perhaps deviating 5 to 10 percent above or below its mean during the year, and generally being lowest in the summer.[10] Aggregate music radio audiences generally have very little change during the year.[11] Given these general facts, if a researcher were to create a dataset purporting to represent national statistics, but with large weekly fluctuations in radio listening or television viewing, say several hundred percent or more, we would have every right to be suspicious of the data.

Alternatively, music piracy activity might be thought to be particularly related to sales of music albums. Record sales are basically flat during the first nine or ten months of the year. The variability within those months appears to be greater than is the case for television and radio, however.[12] A large increase in record sales

---

compare *availability* of songs, not downloads, by week for a limited number of weeks, which cannot be used to verify their download data.

8. These data appear in Table 3, "Downloads and Matched Songs," in the first public version of the OS paper, dated March 2004 (**link**).

9. OS then try to match these American pirate downloads to the songs in a sample of 680 albums, chosen because they were successful enough to make it onto a Nielsen SoundScan chart of leading albums (by genre) over the final 17 weeks of 2002. The process of matching downloads to albums reduces the number of pirate downloads to 47,709, eliminating more than 80 percent of the downloads (see OS 2007, 8–9). OS's main results are based on these 680 albums over the 17-week interval, providing about 10,000 album-week observations (the quantity of downloads and sales of a given album in a given week).

10. Monthly television ratings are usually lowest in the summer. See the information available **here**.

11. See, e.g., the figure "Listening Patterns" in Arbitron's "Radio Today" report for 2004 (**link**), which shows virtually identical time spent listening by season in 2002–2003.

12. During 2003–2005, ignoring the months November and December, average monthly changes were 16 percent, with the highest individual monthly change being 32 percent, according to Nielsen Soundscan.

regularly occurs during the Christmas holiday season, where sales increase to about double the normal level in the weeks before the holiday.[13] Even in OS's sample of album sales during these 17 weeks, the largest weekly change if you exclude the run-up to Christmas is 11 percent, and the largest weekly change including all weeks is 41 percent.

With this as a backdrop, what should we expect for pirate downloading during the last 17 weeks of the year? It is difficult to come up with reasons for important weekly trends during the first half of this period. It is likely to depend on the number of new hit songs, which is often random, although major albums are often released during the Christmas selling season. But perhaps downloads might be expected to decline during the Christmas season since music albums are often holiday gifts, and many potential gift recipients are not pleased to receive a present consisting of pirated music. All in all, there seems little reason to expect major weekly changes in quantity of pirate downloads during most of these 17 weeks.

Does the weekly OS data conform to these general expectations? Absolutely not. We find amazingly large weekly variations, unlike other media/entertainment trends. Figure 1 provides the details.

Figure 1. OS American pirate downloads

Data source: Table 3 in March 2004 working paper version of "The Effect of File Sharing on Record Sales: An Empirical Analysis" (**link**).

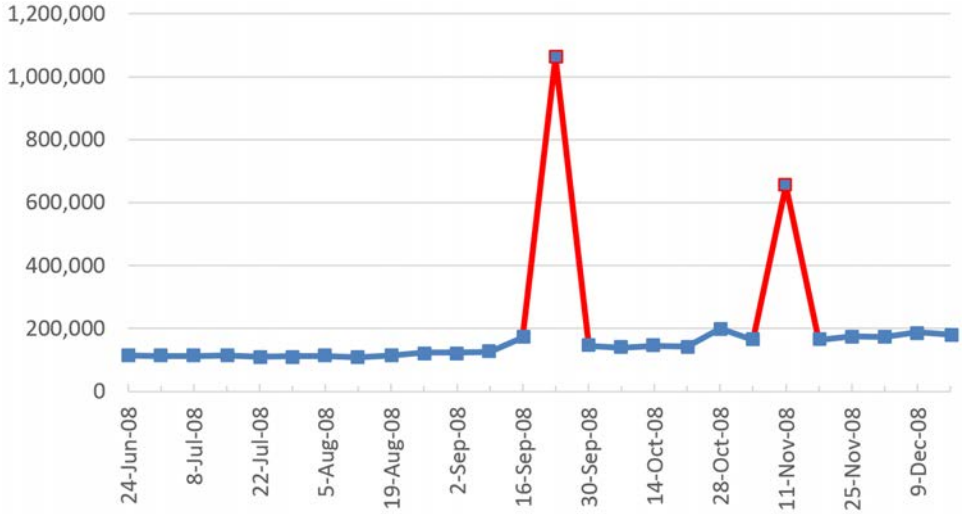
Relative to other weeks in the sample, the number of downloads is extremely low during the first two weeks. In fact, downloads in the second week are less

13. It is well known that sound-recording sales peak at Christmas in the U.S., and Nielsen SoundScan monthly aggregated data show this result as well.

than nine percent of the average weekly number for the entire period. From the second week to the fourth week, downloads jump 12-fold. From the ninth to the eleventh week downloads jump over six-fold. From the second to the eleventh week downloads jump 22-fold. These are remarkably large jumps that seem unreasonable for an activity undertaken throughout the country by tens of millions of Americans.

The OS piracy data not only have suspicious jumps, but they also do not match the variation in other measures of piracy. For example, Jonathan Lee (2016), using weekly download music album data from a pirate system for the last half of 2008, finds that "file sharing activity is fairly constant." Figure 2 represents the fairly smooth downloading activity on the system he was following, during 2008, but I draw attention, using the color red, to two peaks in the data that occurred because the rules of the network changed in a way that artificially enhanced downloading during those weeks.[14]

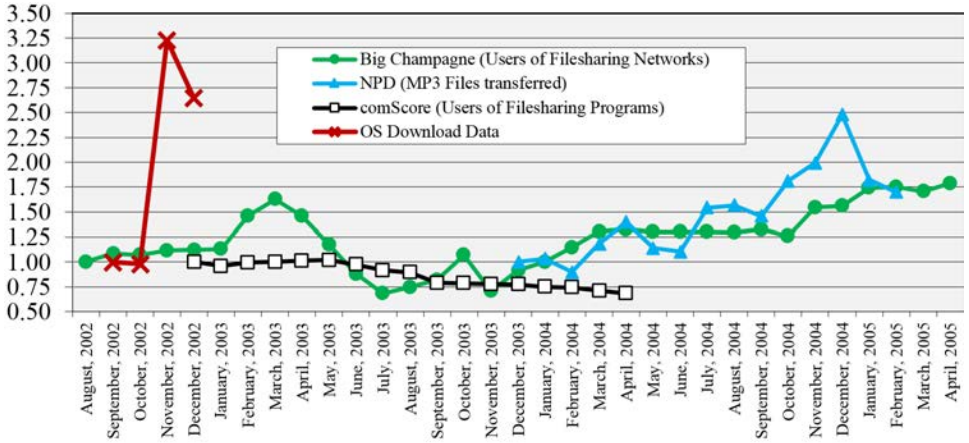
Figure 2. Lee's weekly pirate downloads



14. Mr. Lee generously offered to provide his data upon request, and I thank him for that and for providing a detailed explanation of their construction. The red peaks reflect changes in the rules of the system that Lee was following. During certain weeks 'freeleeches' were allowed where downloaders could download as much as they wanted without having to follow the uploading policy that this pirate server imposed on its users and without reducing their 'buffers' or 'ratios.' In those weeks when 'freeleeches' were allowed, there were temporary spikes in downloading that Lee believes were due to these changes in the tracker's rules and which provided the exogenous instruments used in his analysis. Even these spikes (of 300 percent and 500 percent), which reflect only the temporary conditions on this one pirate system, are not as large as some of the weekly changes found in Figure 1.

We can also compare the OS data to monthly measurements of file sharing, as done in Figure 3 (a modified version of Figure 2 in Liebowitz 2006). Each of the three data sources has its first month normalized to 1, and then the later months are all normalized relative to the first month. The OS data, for the four months of its existence, are in red.

Figure 3. Various measurements of file sharing



The data source that is contemporaneous with the OS data is BigChampagne (mentioned by OS in the quote above) which measures the number of pirates who are online. Data from comScore measures the number of pirates. Neither comScore nor BigChampagne display anything like the extreme variability found in the OS data. In fact, the monthly change in the OS data, from October to November, is more than 50 times as large as what the BigChampagne data show for the same month and about seven times as large (in percentage terms) as the largest month-to-month change in three years of BigChampagne data. The changes from November to December are not even the same sign.

Data from NPD, which started measuring the number of pirated mp3 file transfers in late 2003, have their largest absolute monthly change (from December 2004 to January 2005) of 43 percent.[15] By way of comparison, the largest monthly change in OS data is more than five times as great. The OS dataset, therefore, appears *sui generis*, even though, as OS state, the dataset needs to be representative of all American downloading if it is to be the basis for useful results. With the entire OS research enterprise being built upon the downloading data, this finding should be a cause for concern.

15. Using the lower value as the base year.

The OS download data in the charts above are taken from OS's tables listing the *full* American download numbers. In other words, these are the American download numbers *prior* to OS's attempt to match American downloads to American sound recordings. Thus these highly dubious numbers are not due to the elimination of many observations that occurred in OS's matching process, nor to any quirks from their particular selection of matched albums.

The mismeasured and astonishingly influential German school holidays

Another canary in the OS coal mine—remember, their data set is not publicly available—is their average value of NGSV, the number of German schoolkids on vacation. This is the instrument that OS have chosen to break the strong simultaneity that exists between the pirating of songs and sales of those songs.

In table 5 of OS's article, it is reported that the mean value of NGSV is 9.855 (million). Meanwhile, the maximum value, occurring during Christmas, when *all* German schoolkids are on vacation (as shown in OS's Figure 1), is 12.491. The straightforward implication is that during this 17-week period the average share of students on vacation is 79 percent!

This school system sounds like the realization of a lazy student's daydreams. What kind of school system must this be where students spend most of their time on vacation? Can Germany really be so lax with regard to requiring K–12 students to attend school? These numbers are worth a second look.

Of course, the German school system is not a vacation fantasyland. My calculations, based on publicly available German school data, indicate that the average share of students on vacation in those 17 weeks was actually 18 percent, not 79 percent.[16] My calculations are based on reports, by week, listing which German states have school vacations, and how many students are in those states. The reader can examine the raw weekly vacation values in my Figure 4, found in the next section.

Since there are only 17 weeks, there can be only 17 weekly observations for NGSV. But because OS use a pooled cross-section time-series analysis, the average value of NGSV in OS's Table 5 is based on 10,093 observations. Each album-week observation is supposed to include the value of NGSV for that week, along with other variables. Note that a fully balanced panel of 17 weeks and 680 albums

16. Interestingly, the mean value in OS's Table 5 is also inconsistent with the weekly values shown in OS's Figure 1. The NGSV shares in their Figure 1, although very close to mine, are still a little too high compared to my estimates.

would provide 11,560 observations. Could missing album-weeks in the OS sample have caused the measured mean of NGSV to take on its very strange value, due to an unbalanced panel? The answer is 'no.' There is only one week where NGSV is greater than OS's 79 percent average and that is the week of Christmas, when 100 percent of students are on vacation. It is mathematically impossible for 1467 missing observations to have raised the actual 18 percent average value to a level anywhere near OS's reported 79 percent. The most that would be possible would be to raise the average NGSV to 21 percent.

It appears, therefore, that there is something very amiss with the OS NGSV numbers that were used in their regressions. With the NGSV values used in the regressions being too large, on average by a factor of four, there is no telling what the impact might be on the regression estimates since we have no idea how these measurements correlate with the correct values. Add this to the fact that the measurement of American downloads is suspicious, and perhaps it is not such a surprise that the coefficient relating the two seems so unreasonable, as I now describe.

In the first stage of their instrumented regressions, OS regress NGSV, along with a small number of other variables, on their measure of American pirate downloads. The results, reported in their Table 7, show a significant *positive* relationship between NGSV and American downloads. OS state:

> The first-stage estimates imply that a one-standard-deviation increase in the number of children on vacation [that is, NGSV] boosts [American] weekly album downloads by slightly more than one-half of their mean, an effect that is statistically significant and economically meaningful. (OS 2007, 23)

OS draw the correct inferences from their statistics, although their use of the term "economically meaningful" is somewhat too modest. It is important to note the full implication of this remarkably large coefficient on NGSV. A one-standard-deviation change in NGSV represents a mere 29 percent of German schoolkids.[17] When this 29 percent share of German students goes on vacation, the large positive coefficient implies an increase in American downloading of 54 percent from its mean (and in Model 5 of their Table 7, the coefficient is three times as large!).

17. From OS's Table 5, the standard deviation of NGSV is 3.6 million, and the total number of students (the maximum value, which is also the total because in week 17 all students are on vacation) is reported to be 12.491 million, meaning that the standard deviation change in NGSV is equivalent to 29 percent of German schoolkids. I should note that OS are including all German K–12 students, which greatly exaggerates the number of German students who are likely to be engaged in music piracy. I discuss this in more detail in footnote 30.

But this positive coefficient also implies, given the summary statistics in OS's Table 5 (including their incorrect average value of NGSV), that when NGSV drops from its average value to zero, American downloads drop by an amount that is 150 percent of the average number of downloads. Since we are dealing with a drop of 150 percent when 100 percent is the most that downloads could actually drop, zero downloading would be an implication (unless other factors were increasing downloads by more than 50 percent). Thus American downloading would be expected to drop to zero in 7 of the 17 weeks of the study, when all German students are in school. That result seems counter to all understanding of actual American piracy (including OS's own dubious data), none of which finds American piracy frequently dropping to zero.

If the correct average share of NGSV were used, instead of the grossly inflated average reported by OS, weeks where all the German students were in school would no longer bring American downloading to a complete halt, but large drops would still be implied. The difference between all students being at school or being home would cause American piracy to change by a factor of four.[18] In other words, American downloading would be almost 300 percent higher on days when all German students are at home compared to days when all German students are at school. American piracy should also soar by several hundred percent in late July and early August (weeks not in the OS dataset), when most German students are on summer vacation. Yet in their first 'quasi-experiment' elsewhere in their paper, using other data, OS claim to have demonstrated the very opposite empirical result—that American piracy *falls* in the summer even though German students are on vacation.[19]

Although it is difficult to believe that American downloaders could be so strongly impacted by NGSV as implied in the OS regressions, it is even more difficult when we consider the choices available to American pirates. The OS results imply that American pirates do not bother replacing the no-longer-available pirate files of vacationing German students with the other 99 percent of pirate files available from those pirates who are not German students.[20]

Before closing this section, I want to note a possible suggestion that the implications of the strong coefficient in the first-stage regression, which surely

18. Given the first-stage coefficient on NGSV (.671), when the share of NGSV changes from zero to 100 percent, American downloads (in OS's sample) change by 8.38 units per album-week. The average number of downloads (from their Table 5) is 4.36. Starting from a mean of 18 percent for NGSV, I add 82 percent of the 8.38 variation to the mean for when NGSV is 1 and subtract 18 percent of 8.38 for when NGSV is 0.
19. The claim is found on page 36 of their article. Using BigChampagne data, I have demonstrated (Liebowitz 2016b) that OS's putative empirical confirmation of this claim is incorrect.
20. I will demonstrate below that German K–12 student vacations influence considerably less than 1 percent of worldwide pirate files accessed by Americans.

seem unreasonable, might be an artifact of the download data being heavily skewed toward the more successful records, just as record sales are heavily skewed.[21] Contrary to this suggestion, skewness does not seem likely to be responsible for these extreme results.

For example, we could reduce the skewness by eliminating the large number of observations (album-weeks) with zero or very few downloads, and the extreme results will still hold, albeit they will be somewhat less extreme. First, imagine removing two-thirds of the observations with the lowest downloads, which seems likely to remove all the observations with zero downloads. Although it is likely that the regression coefficient on NGSV would rise, let's be conservative and assume that it remains constant. In this case, the average number of downloads rises to 13.1.[22] But the implied difference in American downloads between when all German kids are in school and when none are in school is still a very large 72 percent.[23] Even removing 95 percent of the observations leads to a difference in American downloads of 39 percent.[24]

In these highly restricted samples, predicted changes in American piracy due to German school vacations are still large enough that they should be able to swamp other economic forces causing short-term changes in American piracy. In order to believe OS's results, it must be the case that American piracy is the tail being wagged by the German school holiday dog. It should be even more difficult to accept OS's powerful positive relationship between NGSV and American piracy, based on individual albums, when the overall weekly changes in American piracy are shown to be *negatively* related to NGSV, which is the subject of the next section.

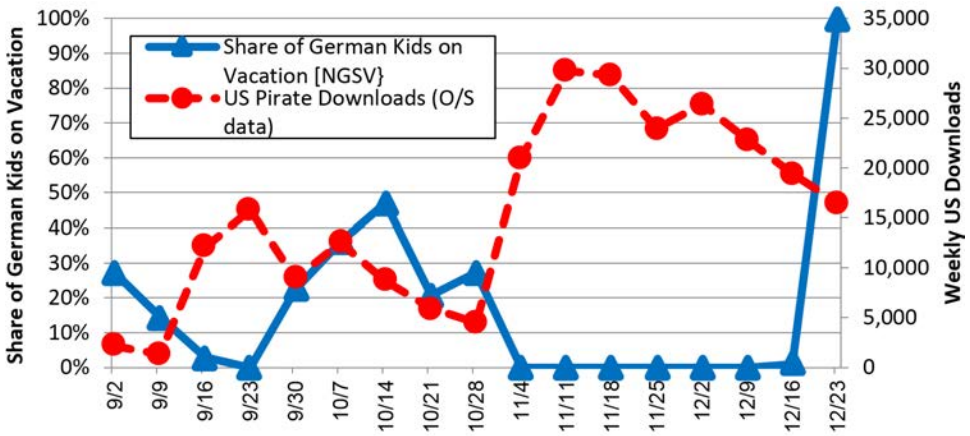
OS's aggregate data conflict with their hypothesis and results

The hypothesis underlying OS's analysis is that when German kids go on vacation they run their home computers for more hours per day, increasing the available files to American pirate downloaders. This implies a *positive* association

21. This skewness is why, when OS generalized their results to the entire industry, they should have given greater weight to more successful albums than to less successful albums. Blackburn (2006), in his analysis of similar data, spends considerable effort trying to avoid duplicating OS's oversight.
22. Assuming that all the removed observations have zero downloads implies that the average of the remaining one-third of the observations have an average that is three times as high as the original average.
23. For this calculation, as well as the next, I use the same procedure as discussed in footnote 18.
24. In a response to the editor of the *Journal of Political Economy*, OS stated that the 95th percentile observation had an average download value of 23.

between NGSV and American downloads, and an enormous positive relationship is what they found in the first-stage regressions, as just discussed. But let's bring this hypothesis to the weekly data, which are represented in Figure 4.

Figure 4. NGSV and U.S. downloads, 2002



The German students on vacation (NGSV) are represented by the blue line and triangles, with the weekly values on the left axis. In early September, some students are still at the end of their summer vacations, with autumn vacations taking place in late September, October, and very early November, after which school vacations come to an end (until Christmas).

The red dashed lines with circles illustrate the number of U.S. downloads in the OS dataset, with the values on the right hand axis (this is just Figure 1 again). It is fairly obvious that the greatest amount U.S. piracy occurred in November and December, although that is also when NGSV was at its lowest point. Weeks when German students are on vacation are not associated with greater amounts of U.S. downloading.

Instead, the data provide a *negative* correlation: −0.71 (significant at the one-percent level) between NGSV and American downloads if the Christmas week is excluded, or −0.37 (not significant) if the Christmas week is included. Further, even including the polynomial time trend used by OS in their panel regressions, there is a negative relationship between these variables (significant at the one-percent level for the full sample). These negative correlations are the very opposite of the premise underlying OS's use of NGSV, and the opposite sign of their coefficient on NGSV in their first-stage regressions.

We must ask whether it is mathematically possible, short of an error, for NGSV and aggregate U.S. piracy to be negatively related while at the same time NGSV and the piracy of individual albums are positively related. The answer

appears to be 'no,' because there does not appear to be any fallacy-of-composition issue here. The weekly change in aggregate piracy (downloads) is merely the sum of the weekly change in piracy of individual albums, and since there is no cross sectional variation in NGSV, the sign of the aggregate relationship should be the same as the summed signs of the disaggregated relationships (weighted by downloads). It appears, therefore, that either there is an error in the OS data or else in OS giving equal weight to all albums in their regressions.[25] These dueling signs are also not due to different controls, because some of these first-stage regressions control for no other variables.[26]

Note the importance of this result. The reasoning that OS offer for using NGSV as an instrument is that high levels of NGSV are supposed to lead to high levels of American piracy, and the relationship is supposedly strong enough so as to measurably affect American record sales (if piracy reduces sales). Yet, over the 17 weeks of aggregate OS data, the result is the exact opposite—NGSV is negatively related to the quantity of American downloading.

The Princess and the Pea, or why the plan of OS's paper could not have worked

Exclusive of the problems already discussed, NGSV could only be a practical instrument for American downloading if its impact were large enough to rise above the background noise. The results of OS—specifically the large coefficient and small standard error on NGSV in the first stage—imply that the impacts of NGSV on American piracy rise well beyond the background noise. In this section, I demonstrate that the actual number of files that vacationing German pirates make available to Americans is extremely small, far too small, indeed, to have any measurable impact on American piracy. The analysis is based on simple logic, arithmetic, and, once informed by facts, common sense, thereby avoiding the complexity of econometric issues.

---

25. Because the aggregate measure of piracy implicitly weights albums with many downloads far more than albums with few downloads, it is possible, if the many albums with low downloads have those downloads increased when NGSV increases, that unweighted regression results such as those reported by OS might provide misleading results about the overall relationship between NGSV and downloads. The problem with OS giving equal weight to all albums in their regressions was discussed in footnote 21. This weighted/unweighted explanation is different than a fallacy-of-composition effect, discussed in footnote 6.

26. In some first-stage regressions (see model III, Table 7 of the December 2004 version of their paper (**link**)), OS control only for album fixed effects and a polynomial time trend. In Table 7 of their published paper, Model 2 also includes the U.S. MTV rank, but the coefficient on NGSV is essentially unchanged (.671 versus .670) with the inclusion of this variable (OS 2007, 24).

I refer, in the heading to this section, to Hans Christian Andersen's fairy tale "The Princess and the Pea." For those who don't remember the story, a prince wanted to marry a princess, and one way to recognize a real princess was whether she could (magically) feel a minuscule pea under dozens of mattresses. I mention this because the variation in NGSV that might impact American downloaders is extremely small, rather like the pea in the fairy tale.

An example indicates the nature of the problem I am suggesting. Assume you know a large extended family in Germany that listens to American music and is heavily engaged in making pirated files of American music available on file-sharing networks. Assume also that you know when this family has its file-sharing software running and when it does not. Even if the extra files made available by this family, when its file-sharing software is running, make it technically easier for some dozens or even hundreds of Americans to download files, no one would believe that this small number of additional files could have a measurable impact on overall American downloading, or on the sales of sound recordings in America. That is because the impact of these extra files would be swamped by the behavior of millions of other pirates, in America and elsewhere. The noise from everyone else, which is in the vicinity of a million times larger, would overwhelm the signal from this particular family.

That is why OS emphasize how important German files are to American downloaders. OS wish to convince the reader that the impact on American piracy from German school vacations can make it through the overall noise and be detected in their regressions. But OS do not investigate the actual number of pirate files affected by NGSV. Instead, OS provide an estimate of the share of American pirate files supplied by *all Germans throughout this time period*. The analysis below demonstrates, however, that the number of files controlled by German student-pirates affected by school holidays is orders of magnitude less than the estimate suggested by OS.

OS's analysis, therefore, assumes that NGSV is a boulder, upon which any sleeping individuals would be acutely aware of their discomfort, even with many mattresses under them. But when NGSV is really pea-sized, OS's econometric apparatus would need to be like the princess, (magically) capable of detecting an extremely small object that is heavily insulated and far away.

The rest of this section examines the factors influencing the impact of NGSV on American downloaders. First, I demonstrate that German files are over-represented in the OS sample. Next, the portion of German pirate files controlled by German secondary-school students affected by German school holidays is shown to be quite small. Further, the typical school holiday in the OS data affects only a small portion of German students at any one time. Finally, the extra files made available by a student being on vacation is estimated to be only a portion of

the files normally made available. The net result of these multiple factors is that changes in NGSV should change the number of pirate files available to Americans by less than two-tenths of one percent, a vanishingly small impact on the number of files available to Americans. In order to demonstrate how robust this conclusion is, I try, whenever a choice is available, to take the option that is more favorable to the OS analysis.

Germans are overrepresented in the OS database

Based on their 2002 sample, OS claim that American pirate downloaders received 16.5 percent of their files from German file sharers, a value reproduced in the first column (second row) of my Table 1, which also contains related statistics about the German and American populations. The second column shows that OS report German pirates to account for more than 13 percent of all pirates in their dataset. Note as well, as shown in the bottom row, that OS report that Americans download 2.7 times as many files from other Americans as from Germans and that there are about 2.3 times as many American pirates as German pirates.

These are surprising measurements because in 2002 the U.S. had a population between 15 and 29 years of age that was four times as large as that of Germany,[27] a broadband penetration rate that was 1.7 times higher than in Germany (**link**), and a higher Internet penetration rate (**link**). If anything, we might expect the number of U.S. pirates to be more than four times as large as the number of German pirates and for American pirates to download from other Americans at least four times as frequently as they download from Germans (ignoring time zone and repertoire differences, which should only enhance this expectation).

TABLE 1. Statistics related to file-sharing activity in the United States and Germany

	(1) [OS dataset] Origin of Files Americans Downloaded*	(2) [OS dataset] Share of World File Sharers*	(3) 2003 Share of World File Sharers**	(4) 2002 Share of World Internet Users*
United States	45.1%	30.9%	55.4%	27.4%
Germany	16.5%	13.5%	10.2%	5.3%
Ratio U.S./Germany	2.7	2.3	5.4	5.2

*Source: OS (2007, Table 2). **Source: OECD (2004, 190).

That expectation is supported when we look at data based on entire populations instead of OS's tiny servers. The OECD (2004, 190) reports that American pirates were more than five times as numerous as those in Germany, as shown

27. Population statistics come from an OECD **database**.

in column 3.[28] And, in data cited by OS (2007, Table 2), the CIA found that in 2002 American Internet users were more than five times as numerous as those in Germany, as shown here in column 4.

What is the implication of the OS data likely having a large over-representation of Germans relative to Americans? It means that the importance of German files to American pirates would be considerably less than that reported by OS. To correctly gauge how many German files provided to American pirates are going to be affected by changes in NGSV, we would need to scale the share of Germans to a more realistic level. Given that the OS data seems to overweight Germans relative to Americans at about a 2:1 ratio, we can merely cut the estimated share of German files used by Americans in half. This implies that the full population of American file sharers, as opposed to those captured in OS's sample, are likely to download about 8.3 percent of their files from Germans, not 16.5 percent.

Schoolchildren as a fraction of German file sharers

Even 8.3 percent is a large enough number that it would not be surprising if changes in the file-sharing activity of those 8.3 percent could measurably influence the file-sharing activity of Americans. But how many of those 8.3 percent of German pirate files are actually controlled by German kids on school vacation? OS conduct no examination of the share of German files that can be attributed to German *students*, although that would seem to be a crucial factor in the analysis they propose.

Although I was unable to find estimates of the file-sharing population distributions for Germany, I was able to find such distributions for the U.S. and France. The U.S. and French numbers, although they measure slightly different relationships,[29] are fairly similar to each other and I presume also similar to those from Germany. Table 2 provides these data.

The numbers in Table 2 clearly indicate that secondary school students (ages 12–17), the only age group in this table that would be affected by secondary school holidays,[30] are not the main users of file-sharing networks. In the U.S., individuals

28. OECD's measure of file sharers is based on data from BigChampagne.

29. The U.S. numbers are based on the share of the *population* whereas the French numbers measure share of *Internet users*. The U.S. share of Internet users engaging in file sharing is higher than the numbers in Table 2, and similar to the French numbers. But the share of Americans using the Internet in 2002 was about double the share in France (OECD, 2004, Figure 4.1), so if compared in terms of the share of population, the U.S. likely would have larger shares in all age groups than France.

30. When OS provide statistics on the number of schoolchildren, they include all children in school. This grossly overstates the number of potential child pirates, since it is unreasonable to believe that five-year-olds, or seven-year-olds, or even most 10-year-olds were engaging in music piracy of songs of interest to American pirates. For one thing, only 20 percent of German primary school students (up to age 10) were

between 18 and 29 years of age are equally likely to use these networks, and there are more individuals in this age bracket than there are in the in the 12–17 bracket. It is also probable that the highest level of use in the U.S. would be among those aged 18–24, who are often in college or university, although the U.S. data are not broken down in that way. In France, individuals aged 18–24 are seen to have the most intense use, although the average member of the 18–39 age group is more likely to use file-sharing networks than are secondary students.

TABLE 2. File-sharing usage by age group

U.S. October 2002; % of population*		France June 2003, % of Internet users**	
12–17	41%***	12–17	31%
18–29	41%	18–24	47%
30–49	21%	25–39	31%
50–64	8%	40–59	22%
65+	3%	60+	11%

*Source: PewInternet.org spreadsheet (link); October 2002 results for Question ACT35 multiplied by Q6. **Source: OECD (2004, 195). ***Pew numbers usually only include those above 18 years of age; an OECD document (2004, 194) provides a 2001 Pew value for the percentage of file shares among of Internet users aged 12–17, that being 53 percent, which I adjust here (using Internet use rates for 18-29 year-olds) to the base of the overall population, not just Internet users.

It is not difficult to use these numbers to approximate the share of files controlled by those 12–17 years of age. Because it is likely the case that the age groups with the highest file-sharing usage are also those with the greatest intensity of file sharing, I assume that the intensity of piracy for an age group is proportional to its participation rate.[31] I also assume that the overall age distribution is uniform, though in Germany this would over-count younger Germans relative to the actual distribution and is beneficial to the OS story.[32] In the U.S., I exclude all pirates over the age of 50, expecting that they more than balance any pirates in Germany who are less than 12 years of age. In France I do the same thing for pirates over the age of 60.

With these assumptions, secondary students would control 18 percent of pirate files in in the U.S. and 13 percent of those in France. I will take an average of these two numbers to represent the share of files controlled by secondary school

taking English as a foreign language in 2002 (link), so their interest in English language songs was likely very limited.

31. Given that the French statistics are based on percentage of Internet users (not population), this tends to overemphasize the importance of old pirates in France because older age groups in France are less likely to use the Internet in the first place.

32. The German population is skewed toward middle-age individuals. For example, per OECD (link), there were 46 percent more Germans aged 30–39 than aged 10–19 in 2002.

students in Germany—15 percent. Approximately one out of seven pirated files downloaded from Germany by Americans, then, is likely to come from secondary school students.

One other very simple factor ignored by OS is that students go to school only on weekdays. Thus, if file sharing were uniform across the days of the week, school holidays would affect only five-sevenths of days when piracy might occur. Of course, since file sharing is likely to be higher during the weekend when everyone has more free time, this assumption of a uniform distribution overestimates the share of weekly files impacted by school holidays and is thus beneficial to the OS hypothesis.

In sum, school holidays affect only the 15 percent slice of the German file-sharing population representing secondary students, and it does so on only 71 percent of the days in a week, leading to a potential reduction in German files due to German school holidays of 10.6 percent, or about one-tenth of the pirate files of all Germans available to Americans in a week.

Because we had already calculated that all Germans provided about 8.3 percent of the files used by Americans on file-sharing networks, we can now more precisely say that over a period including weekdays and weekends, the pirate files controlled by German secondary students that are potentially influenced by school vacations are about 0.9 percent (8.3 percent times 10.6 percent) of the worldwide pirated files available to Americans.

How large is the typical supply 'shock' due to school holidays?

Figure 4 demonstrated that, for those weeks where some students were on school vacations, the share of kids on vacation was on the order of 20–40 percent, with an average of 25 percent for all weeks with non-zero vacations.[33] Since the vacation times are adjusted by the German government to avoid peak-load vacation travel problems, as noted by OS, it is not surprising that the share of students on vacation in any one week is not very large. But this also means that the 'shocks' brought about by the NGSV variable are only about one-fourth of what you would expect if all the students went on vacation at the same time.

If, as reported in the previous section, German students potentially at school only provide 0.9 percent of all pirated files, and if only one-fourth of the students

33. This calculation includes several weeks with very low vacation shares but excludes the Christmas week because that is a national holiday, not a school vacation in the sense that OS use them. Including zero vacation weeks drops the average share of students on vacation to 18 percent. If weeks with a share of students on vacation below three percent are dropped the average is 28 percent.

are affected in an average week when school vacations (as opposed to national holidays) take place, then the size of the typical German vacation shock on the availability of pirate files to Americans would be about one-fourth of 0.9 percent, or 0.23 percent.

How do school holidays affect files from German schoolkids?

When German secondary students experience school holidays, how does this change the number of files that they make available to American downloaders? After all, we are not really interested in the number of files controlled by German students. Instead, we are interested in the *change* in the number of those files made available to Americans when German students go on school vacations. That is the number that we would need to know in order to measure the influence of German vacations on American downloading.

OS do not delve into this subject—they merely assume that the German students' computers are on for longer periods of time because the students are not at school. They ignore the possibility that these kids might take a trip with their families during this period or that the computers might be left on all the time. The key assumption that OS make is that during school holidays German schoolkids have their computers on all day and that the kids turn them off when they go to school. Since, on a typical school day, kids are at school for fewer hours than they typically sleep each night, I suppose we should likewise figure that kids turned their computers off when they are sleeping.

Assume that the baseline day when schools are open looks like this: students get up to go to school at 7 a.m. without turning on their computers, and arrive home at 1 p.m., the typical time for school to end in Germany. The students then keep the computer on until bedtime, assumed to be 11 p.m. In this case, the computer would be on for 10 hours during school days.

Now take the case where the students are on vacation. Let's assume they sleep in until 9 a.m. and then keep the computers on until bedtime, which we assume is now 1 a.m. The computers would be on for 16 hours, as opposed to the original 10 hours, for a 60 percent increase.[34]

34. Moreover, I ignore consideration of when those extra download files from German kids are available to American pirates, to the benefit of the OS thesis because the fit is not a good one. Four of the six extra hours of German file availability occur during the German hours of 9 a.m. to 1 p.m., when German kids would otherwise be at school. But these hours correspond to 2 a.m. through 6 a.m. in the U.S. Central Time Zone. Virtually no American pirates are going to be awake at those late-night times. So four of the six extra hours are of virtually no value to American pirates.

A 60 percent change in the availability of German students' files seems like a fairly substantial number, and it is, but it is *less* than the number of files they normally make available to Americans. To arrive at the number of additional files made available to Americans due to German school vacations, we need to multiply 0.23 percent, the share of pirate files impacted by vacations, by 60 percent, the change in file availability during vacations. The result is 0.14 percent.

In perspective: NGSV's impact on American file sharing

Returning full circle, we began with OS's claim that, in their sample, German files represented 16.5 percent of pirate files downloaded by Americans. Evidence from sources measuring the entire market, however, indicated that Germans were overrepresented in OS's sample by a factor of 2, leaving 8.3 percent as a more reasonable estimate of the share of American pirate downloads taken from German sources. We then examined the likely share of German pirate files controlled by German secondary school students (15 percent), and also noted that weekends are not affected by German school holidays. Taking account of these two factors led to a conclusion that the pirate files of German schoolkids that could be affected by German school holidays represented slightly less than 1 percent of the files taken by American pirates. And then, since we are only interested in the *extra* files available to Americans due to German secondary school vacations, it was necessary to examine the share of German students at any one time affected by school vacations. The share of German students impacted by the average vacation was about 25 percent, so the files that Americans might download from German schoolkids that could be influenced by German students on vacation was reduced to slightly less than one quarter of one percent. The final consideration was measuring the presumed increase in availability of files when German students are on school vacation compared to when they are not on vacation. We concluded that this increase could be in the vicinity of 60 percent. Since the total files that could be affected by German vacations were, again, slightly less than 0.25 percent, a 60 percent increase means the bottom-line increase in pirate-file availability to Americans caused by a typical German school holiday would be 0.14 percent. And that calculation incorporates several assumptions that would inflate the value.

Let's try to put this resulting percentage, roughly one-seventh of one percent, into more intuitive contexts.

First, OS claim that the average download time for a music file in their sample, counted from the download request to the completed download, was 1,496 seconds, or about 25 minutes.[35] If the extra German files from the German holi-

35. Calculated from OS's (2007) Table 6, sum of items in the last row, for the first three columns.

days increased the speed with which Americans could find and download songs in proportion to the increased quantity of file availability, the savings in time would be about 2 seconds out of the original 1,496 seconds. It seems inconceivable that American downloaders could even perceive a change this small without the careful use of a stopwatch. Surely, common sense tells us that if the additional files from German school vacations cannot even be noticed by American downloaders, then the change cannot seriously affect Americans' downloading behavior. Nor does it seem possible that the change could rise above the background statistical noise, nor impact American record sales.

Alternatively, we can view this 0.14 percent as a share of the 320 million Americans, a share which would be 445,000. This is equivalent to a small-sized city, such as the Shreveport-Bossier City MSA. Do we really believe that if a tornado shut down Shreveport for a week, this event would be detectable in weekly national U.S. record sales figures where the *smallest* weekly change over these seventeen weeks, using OS's data, was more than twenty-six times as large? Or could we believe that piracy activities of a city of this size could *dominate* the piracy behavior of the entire U.S.? Could we even believe that it could have a large enough impact on American downloading to be measurable, when more than 99.8 percent of the other pirates throughout the world are making their files available to Americans?

Finally, there is one last point to be made here. OS find that NGSV strongly reduced the time it took for Americans to download files, just as they found it strongly influenced American downloads. But their measure of download time is not an independent data source—it comes from their dataset, viz., the same publicly unavailable dataset that contains their data on download counts. Whatever is causing the erroneous reported relationship between NGSV and the number of American downloads is most likely also causing an erroneous reported relationship between NGSV and the completion time of American downloads.

Conclusion

This paper has investigated Professors Oberholzer-Gee and Strumpf's influential article on the impact of piracy on record sales. By closely examining their results, the snippets of data they have made public, and the logic behind their analysis, it was possible to discover numerous problems with their published results.

First, their data purportedly measuring American downloading, upon which their entire edifice stands, demonstrates a level of weekly variability that is inconsistent with other measures of American downloading and also is at variance with

reasonable expectations about the likely weekly variance. If the downloading data are defective, then the rest of their main empirical work would be invalid.

Second, their claimed results imply that German schoolkids are almost always on vacation, which is factually incorrect, and their measured impact of German schoolkid piracy on American piracy is unbelievably large, implying that American piracy is virtually controlled by German school vacations.

Third, although their panel regression results imply a powerful positive impact of German schoolkids on American downloading, their aggregate weekly data reveal a strongly negative relationship between American downloading and German schoolkids on vacation, an inconsistency that seems to imply some sort of error.

Fourth, and most importantly, their chosen instrument, the number of German schoolkids on vacation, is shown, though a careful analysis of raw data, to be of a vanishingly small size. The minuscule size of their key instrument implies that it would be impossible to measure its impact on American pirates or American record sales, given normal background noise.

Based upon this analysis, and additional issues that I have raised elsewhere (Liebowitz 2016b), I believe that the OS article and its conclusions can be deemed unreliable. Because OS have not fulfilled their promise to make their data available, it is not possible to know the exact causes of their inconsistent and unbelievable results.

Appendix

Underlying data for Figures 1–4 can be downloaded **here (.xlsx)**.

References

Blackburn, David. 2006. The Heterogeneous Effects of Copying: The Case of Recorded Music. Working paper. **Link**

Glenn, David. 2008. Dispute Over the Economics of File Sharing Intensifies. *Chronicle of Higher Education*, July 17. **Link**

Hammond, Robert G. 2016. The Fallacy of Composition and Disruption in the Music Industry. In *Business Innovation and Disruption in the Music Industry*, eds. Patrik Wikström and Robert DeFillippi, 73–94. Cheltenham, UK: Edward Elgar.

Häring, Nobert. 2008. Der Download-Krieg der Ökonomen. *Handelsblatt*, March 4. **Link**

Lee, Jonathan. 2016. Purchase, Pirate, Publicize: The Effect of File Sharing on Album Sales. *Queen's Economics Department Working Paper* 1354, Queen's University (Kingston, Canada). **Link**

Liebowitz, Stan J. 2006. File Sharing: Creative Destruction or Just Plain Destruction? *Journal of Law and Economics* 49(1): 1–28.

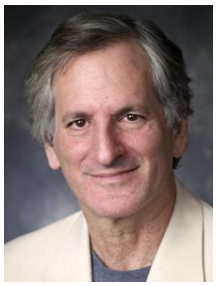
Liebowitz, Stan J. 2016a. How Much of the Decline in Sound Recording Sales Is Due to File-Sharing? *Journal of Cultural Economics* 40(1): 13–28.

Liebowitz, Stan J. 2016b. Replicating Four 'Quasi-Experiments' and Three Facts from Oberholzer-Gee/Strumpf's Piracy Article. Working paper. **Link**

Oberholzer-Gee, Felix, and Koleman Strumpf. 2007. The Effect of File Sharing on Record Sales: An Empirical Analysis. *Journal of Political Economy* 115(1): 1–42.

Organisation for Economic Co-operation and Development (OECD). 2004. *OECD Information Technology Outlook 2004*. Paris: Organisation for Economic Co-operation and Development. **Link**

About the Author



Stan Liebowitz is the Ashbel Smith Professor of Managerial Economics at the University of Texas at Dallas. He has authored many articles examining the influence of new technology on intellectual property, beginning with a study on photocopying's impact, for the Canadian government, and continuing with digital copying. He has also written numerous articles about network effects and a few articles about measuring research performance. Most of his research is available on the SSRN. His email is liebowit@utdallas.edu.

Go to archive of Comments section
Go to September 2016 issue