UNITED STATES DISTRICT COURT
SOUTHERN DISTRICT OF FLORIDA

CASE NO. 11-20427-WILLIAMS/TURNOFF

DISNEY ENTERPRISES, INC.,
TWENTIETH CENTURY FOX FILM CORPORATION,
UNIVERSAL CITY STUDIOS PRODUCTIONS LLLP,
COLUMBIA PICTURES INDUSTRIES, INC., and
WARNER BROS. ENTERTAINMENT INC.,

Plaintiffs,

v.

HOTFILE CORP., ANTON TITOV, and
DOES 1-10.

Defendants.
_____/

HOTFILE CORP.,

Counterclaimant,

v.

WARNER BROS. ENTERTAINMENT INC.,

Counterdefendant.
_____/

DECLARATION OF DR. RICHARD WATERMAN IN SUPPORT OF PLAINTIFFS' MOTION FOR SUMMARY JUDGMENT AGAINST DEFENDANTS HOTFILE CORP. AND ANTON TITOV

PUBLIC REDACTED VERSION

I, Richard Waterman, hereby declare as follows:

1. I am an Adjunct Professor of Statistics at The Wharton School at the University of Pennsylvania, and the President and Co-Founder of Analytic Business Services, Inc., a consultancy focused on providing expert advice and opinions in the field of statistical analysis. I have been retained by the plaintiffs Disney Enterprises, Inc., Twentieth Century Fox Film Corporation, Universal City Studios Productions LLLP, Columbia Pictures Industries, Inc., and Warner Bros. Entertainment Inc. ("plaintiffs") to conduct an analysis and provide my conclusions regarding the level of infringing activity on www.hotfile.com ("Hotfile"). The statements made in this declaration are based on my personal knowledge or application of my specialized knowledge to facts or data of which I am aware. If called to testify, I would testify based on the best of my knowledge, information, and belief, as follows:

2. I received my Ph.D. in Statistics from the Pennsylvania State University in 1993. I have substantial experience designing and reviewing sampling protocols for various large organizations, such as the United States Postal Service, for whom I designed and analyzed a national multi-stage sample for the estimation of operational characteristics. I have designed sampling protocols involving various filesharing technologies, specifically BitTorrent, Gnutella and Usenet, as further explained below. I also have substantial experience in designing sampling protocols in the private sector, and have developed market research studies for numerous large corporate clients, which typically involve issues related to sampling. Further details of my professional history, including a list of publications I have authored during the last ten years, can be found on the resume attached as Exhibit A. Within the last four years, I have testified as an expert at trial or deposition in the following cases, as further outlined in Exhibit B: *Arista Records LLC, et al. v. Lime Group LLC, et al.* No. 06-Civ. 05936 (S.D.N.Y); *Columbia Pictures*

Industries, Inc. et al. v. Gary Fung, No. 06-CV-5578 (C.D. Cal.); and *Schappell v. GEICO Corporation*, No. 1333 S2001 (Pa. Commw Ct.). I have submitted expert reports in *Columbia Pictures Industries, Inc. et al. v. Gary Fung*, No. 06-CV-5578 (C.D. Cal.); *Arista Records LLC, et al. v. Usenet.com, Inc.*, No. 07-CV-08822 (S.D.N.Y.); *Schappell v. GEICO Corporation*, No. 1333 S2001 (Pa. Commw. Ct.); *Freedom Medical Supply, Inc. V. PMA Capital Ins. Co*., No. 003988 (Pa. Commw. Ct.); and *Blehm v. Albert Jacobs et al* 1:09-cv-02865-RPM, for the United States District Court of Colorado. I am being compensated for my services in this case at a rate of $450/hour ($550/hour for testimony).

3. I was asked by the plaintiffs to create a protocol for drawing a statistically reliable sample for a study analyzing the percentage of files downloaded daily that were identified as infringing from the website operated by the defendants, www.hotfile.com ("Hotfile"). To conduct that study, I designed a statistically valid sampling protocol that drew a sample of 1,750 files from Hotfile for further analysis to determine whether they were infringing. Based on the classifications of those files, which was done by Scott A. Zebrak, I was able to calculate the level of infringing daily download activity on Hotfile.

4. In its general design, this study was similar to other statistical studies that I have designed and implemented to measure the level of infringing activity on online services. I designed these studies in connection with the following cases: *Arista Records LLC, et al. v. Lime Group LLC, et al.* No. 06-Civ. 05936 (S.D.N.Y) ("*Limewire*"); *Columbia Pictures Industries, Inc. et al. v. Gary Fung*, No. 06-CV-5578 (C.D. Cal.) ("*Fung*"); and *Arista Records LLC v. Usenet.com, Inc.*, No. 07 Civ. 8822 (S.D.N.Y.) ("*Usenet.com*"). In each case, I designed a protocol to select a sample of content files available on and/or downloaded from different online networks via services offered by the defendants in those cases (respectively, Limewire,

the BitTorrent websites Isohunt and Torrentbox, and Usenet.com). In each case, the sample was obtained and then analyzed by a separate copyright analyst, an expert who classified whether each of the files was "highly likely infringing" to distribute. And in each case, I relied on those classifications and calculated an estimate of the level of infringement on the service. I understand that my conclusions were accepted by the reviewing Court in each of those cases.

5. While each of the previous studies I designed examined different sets of data to measure the overall level of infringement, in this case, I was able to examine a set of data that was particularly helpful in estimating the overall level of infringing download activity on Hotfile: data that showed the number of recorded "daily downloads" of files in a particular day, which is used by Hotfile in its formula to calculate payments to Hotfile's "Affiliates." That data summarized the information that would be available in a "log file" of the same recorded downloads – a download-by-download log of users downloading files from Hotfile. It is the most useful data available to analyze the actual use of a site or service to distribute content. In many instances (for example, in the *Usenet.com* case) this kind of download data simply has not been available to a outside party. Indeed, this information is not available to an outside observer of Hotfile, and I understand that it was obtained in discovery in this litigation.

Overview.

6. I devised a methodology, described in more detail below, to allow for a scientifically reliable and representative sample of files to be selected from the population of interest, in order to calculate the level of infringing download activity on Hotfile. After the sample had been drawn and the content obtained if available, the sample files were reviewed by a copyright analyst, Mr. Scott Zebrak. Mr. Zebrak reported his conclusions as to the infringement status of each of those files, and I understand that Mr. Zebrak is submitting a

declaration also describing the process he followed and his analysis. For the determination of the copyright infringement status of each file in the sample, I relied on the work and conclusions of Mr. Zebrak. Relying on a third party expert's classifications to reach particular conclusions that are integrated into a statistical analysis is a common, accepted, and oftentimes necessary practice in designing statistical studies. In relying on a copyright analyst to make conclusions about the infringement status of the sample files, I have following an identical protocol as in the analyses I conducted in *Limewire*, *Fung*, and *Usenet.com*. In each of those cases, the copyright analyst reviewed the sample files to determine whether the files were "highly likely infringing," among other classifications.

7. Based on Mr. Zebrak's conclusions regarding the infringement status of the content files in the sample, I performed statistical analyses to derive the results for the infringement study. In general, in reaching my opinions and conclusions, I relied upon my specialized knowledge, education, and experience as applied to the facts and data discussed below, as well as data about downloads from Hotfile produced by defendants, and the work and conclusions of Mr. Zebrak.

8. Based upon my review of the most recent data provided by Mr. Zebrak, approximately 90.2% of all daily downloads of files on Hotfile were downloads of infringing or highly likely infringing content; approximately 5.3% of the downloads of files per day on Hotfile were downloads of non-infringing files; and the remaining approximately 4.5% of the downloads of files per day on Hotfile were downloads of files whose copyright status could not be reliably determined in the time allowed. Additionally, 0.5% of the works were identified as being likely illegal to distribute. These are included in my numbers above in the "non-infringing" category,

making the infringement analysis here conservative. This analysis was based on data showing downloads of files that was provided by Hotfile.

Identification of the Population of Interest.

9. The first step in devising the sampling protocol was to define the relevant "population of interest." The objective of the Hotfile study was to analyze the daily percentage of downloads of files from Hotfile that were infringing.

10. In order to make statistical inferences about this population, I decided to analyze the data contained in a file called "dailydownload" that records downloads of files on a daily basis. I understand that Hotfile uses the download records in "dailydownload" in its formula to determine payments to its "Affiliates," who are compensated in part based on the number of downloads of their uploaded files. Thus, by analyzing this file, I was able to analyze the level of infringement of the downloads that Hotfile's Affiliate program directly incentivizes.

11. In this case, my study examined the number of daily downloads of content files from January 2011, the month before this litigation was brought. My decision to examine January 2011 was made prior to selecting any of the files in the sample. There were strong reasons for choosing January 2011 as the period of time from which to draw the sample. I understand that, prior to this litigation, Hotfile did not always preserve content files that were removed from Hotfile (for example, because they were subject to an infringement takedown notice.) My reasonable expectation is that Hotfile, in general, would be more likely to be in possession of actual content files available on Hotfile in more recent periods of time. Having the actual content file to review would provide more information for the copyright infringement analysis conducted by Mr. Zebrak. If we drew a sample of downloads from January 2011, we would be more likely to obtain the associated files from Hotfile for those downloads than if we

drew a sample of downloads from previous months. Review of Hotfile data confirms that expectation. That data shows that cohorts[1] of files uploaded in earlier months generally were less likely to be available from Hotfile after this litigation than cohorts of files uploaded in months closer to January 2011. For example, only 15-20% (at most) of files uploaded shortly after the launch of Hotfile were still available as of the initiation of litigation, whereas up to 80% of the more recent files were available. By choosing a more recent month of daily download data, the study had access to more information about the content files downloaded by Hotfile users (*i.e.*, the content file itself to review, in addition to metadata about the file).

12. There were also strong reasons for choosing a month prior to when the litigation was brought. I understand that, in mid-February 2011, after this case was filed, Hotfile publicly announced that it was changing its "repeat infringer" policy and it terminated substantial numbers of users. I expect that this kind of change would affect Hotfile user download behavior. Indeed, I have examined data about Hotfile's revenue, number of uploads, number of downloads, and available files[2] over time. All of this data shows relatively steady increase in growth through January 2011, then a distinct "break point" occurring in February 2011, the month in which this litigation began, when the levels of revenue, number of uploads, number of downloads, and the available files dropped suddenly.

13. While defendants did not produce actual download log data for the period before February 2011, they did produce the "dailydownload" table for the period prior to the litigation. The "dailydownload" data efficiently summarizes all the necessary information that would be found in a log file to enable an infringement analysis of the recorded downloads. My

[1] A cohort of files is defined as a sequential set of one million file uploads. My understanding is that Hotfile generally numbers its uploads sequentially.

[2] Available files is defined as the cumulative sum of the difference between uploaded and deleted files.

understanding is that this table identifies files that were downloaded in a specific day (represented in the "uploadid" field), the date of download (represented in the "date" field), and the number of "premium" and "free" downloads of the files (represented in the "premium" and "free" fields). Further, my understanding is that "premium" and "free" downloads are downloads by different kinds of users: those who have purchased Hotfile Premium subscriptions, and those that have not, respectively. Adding the two together gives the number of recorded downloads per day for the file on the indicated date. Thus, the "dailydownload" data contains a summary of information of recorded downloads by file for any particular day, which Hotfile uses in conjunction with its Affiliate program.

14. I also understand that defendants have implemented business rules such that "dailydownload" does not measure all downloads from Hotfile, but only records those downloads used in the compensation formula for Affiliates. However, I believe that the "dailydownload" table is in fact very useful to analyze because it is an operational data set showing the download activity Hotfile considers when determining how to pay Affiliates – the users that Hotfile encourages to supply content to the site and promote its download. An analysis of the "dailydownload" data shows what kind of download activity Hotfile is incentivizing through its Affiliate program, which is a central part of its business model.

Sample Selection.

15. I designed a protocol to draw a statistically reliable sample from the daily download data in January 2011 as follows: In order to analyze the number of downloads per day in that month, I decided to select different random days in January 2011, and take a sample of downloads from each of those days. I designed the protocol to randomly select five weekdays

and two weekend days. This eliminated any potential bias in choosing days within that month to analyze.

16. In the first step of the protocol, I randomly selected five weekdays and two weekend days, by consecutively assigning each weekday in January 2011 a number and consecutively assigning each weekend day in January 2011 a number. I then used a standard random number generator to generate a separate list of numbers for the set of weekdays and the set of weekend days. This is a standard and universally accepted means to generate a simple random sample. The days selected by this process were January 5, 11, 20, 21, and 24 (weekdays) and January 1 and 30 (weekend days).

17. Overall, the "dailydownload" table shows 145,691,820 downloads of files from Hotfile in the month of January 2011. On each date selected, the "dailydownload" table shows the number of recorded downloads of files per day. The combination of the "free" and "premium" downloads per day for the selected days were as follows:

Date	Download Count
2011-Jan-01	4,180,329
2011-Jan-05	4,677,811
2011-Jan-11	4,568,087
2011-Jan-20	4,496,274
2011-Jan-21	4,631,944
2011-Jan-24	4,738,937
2011-Jan-30	5,125,537

18. Within each selected day, the sample frame was obtained by taking the dailydownload data and expanding the record of each file to capture the total number of recorded

downloads of that file on that day. For example, if a file was downloaded 5 times in a day, the record would be expanded to reflect five separate downloads of that file. This method permits simple random sampling of the complete set of recorded downloads of all files in a day. The sample size was selected to obtain a 95% confidence interval with a margin of error of plus or minus 5%. (Because of the consistency of daily download infringement proportions, the final margin of error of the study was considerably smaller.) This allows for a high level of confidence that the results of the study reflect the percentage of infringing downloads per day for any randomly selected day in the entire population, together with a high level of precision. To target this level of precision, I concluded that the Hotfile sample size should be 1,750 (250 per day), which is also consistent with sample sizes in other similar online infringement studies conducted in other cases.

19. I used "simple random sampling" to draw the sample within each day. "Simple random sampling" is a universally accepted statistical methodology in which each item has the same opportunity to be chosen as any other item. In this case, each download of a file in a particular day had the same chance to be chosen as any other download of any file within that day. For each day, I used a standard random number generator to generate a list of numbers to select the downloads that constitute the sample. This too is a standard and universally accepted means to generate a simple random sample.

20. I am attaching herewith as Exhibit C the download instructions that implement the sampling protocol I have described. The protocol provides that files in the sample of 1,750 files will be replaced under only two limited circumstances. First, if the file appeared by its metadata to contain child or other illegal pornography, it was not included in the sample. Second, if the content file was corrupt, inoperable, or unplayable/undisplayable, for reasons other than being

password-protected or encrypted, it was not included in the sample. In each of those cases, the files were replaced in the sample by another randomly selected file according to the protocol.

21. Mr. Zebrak provided an analysis showing his conclusions as to which of the 1,750 sample files analyzed were determined to be either confirmed or highly likely copyright infringing, with the result broken down by download date. He also provided information as to which files he classified as non-infringing, those "unknowable" files as to which no determination could be made, and "illegal" files that did not appear to be copyright infringing but that Mr. Zebrak concluded were likely illegal to distribute for other reasons. The infringement determinations of each download by day are itemized in the attached Exhibit D.

Conclusions.

22. Based upon my review of the data provided by Mr. Zebrak, using scientifically valid and accepted statistical calculations, I am able to conclude that, on a daily basis, approximately 90.2% of all downloads of files on Hotfile were downloads of infringing or highly likely infringing content; approximately 5.3% of the downloads of files per day on Hotfile were downloads of non-infringing or highly likely non-infringing files; and the remaining approximately 4.5% of the downloads of files per day on Hotfile were downloads of files whose copyright status could not be reliably determined in the time allowed. Additionally, 0.5% of the works were identified as being likely illegal to distribute. These are included in my numbers above in the "non-infringing" category. These conclusions unequivocally apply to January 2011, the month of data from which the sample was drawn. Additionally, as I explain below, these numbers are probative of the levels of infringement in prior months.

23. Using standard and universally accepted statistical methods to calculate a margin of error at a 95% confidence level yields a margin of error for this study of approximately 1.3%. This indicates a high level of reliability.

24. In my professional opinion, the sampling procedures used in the Hotfile study are based on standard and universally accepted statistical methods, and provide a scientifically valid sample from which we can reliably estimate the incidents of copyright infringement through the Hotfile website during the time period from which the sample was drawn.

25. Further, as I explained above, the population I sampled consisted of daily downloads from January 2011, immediately prior to the filing of the complaint. As a result, my study provided unequivocal evidence about the level of infringement in that month. It also allows conclusions to be drawn about months prior to January 2011, when analyzed in conjunction with other data. As a leading statistical treatise writer notes, "Judgment about the extent to which these conclusions [about the sampled population] will also apply to the target population must depend on other sources of information," and it is helpful to consider any "supplementary information that can be gathered about the nature of the differences between the sample and target population."[3] In designing studies in the real world, it is not uncommon for a sample to be limited to a subset of the target population, and then to attempt to draw conclusions about a broader population, based on consideration of additional data.

26. I believe the level of infringement on Hotfile in January 2011 is indicative of infringement on Hotfile in previous months, for reasons I explain below. As I explained in my deposition in this case, I would expect that the level of infringement in months prior to January 2011 would be similar if Hotfile had been relatively stable during that time, and there had been

[3] William G. Cochran, Sampling Techniques (3rd ed., 1977), at page 5.

no watershed events that I would have expected to change user behavior and the level of infringement on the site. I have since reviewed data regarding the development of the Hotfile site over time. As reflected in Exhibit E, it appears that Hotfile grew at a relatively stable rate from its launch in February 2009 through January 2011, as revenues, numbers of uploads and downloads, and numbers of live files on the site had relatively stable growth. During this time, I understand that Hotfile maintained a relatively consistent business model, making money by charging users for "Premium" subscriptions, and paying its Affiliate users to upload content that was widely downloaded. I have also reviewed information regarding the changes that Hotfile claims to have made to address infringing activity at various points in time, including by reviewing transcripts of the deposition of Anton Titov, which were taken only after my first deposition in this case. While I have no information about the effectiveness of any of Hotfile's steps in reducing infringement, by intention, none of these would be expected to increase the amount of infringement on Hotfile. Further, as reflected in the data, each of the steps taken prior to February 2011 appears to have had little effect on Hotfile's consistent pattern of growth.

27. Based on the stability in Hotfile's growth from the beginning of its operation, the consistency of Hotfile's business model over that time, and the fact that changes in Hotfile's operation purportedly addressing infringing activity during that time were unlikely to result in increased infringement, I would not anticipate that the level of infringing daily downloads on Hotfile for periods prior to January 2011 would be materially lower than the level of infringing daily downloads in January 2011. Indeed, it would be reasonable to expect that, if Hotfile's steps to purportedly address infringement had any effect, the incidence of infringing downloads from Hotfile in January 2011 would be lower than in previous months, making the conclusions from January likely conservative regarding overall levels of infringement.

28. I am aware that defendants' expert Dr. Daniel Levy has characterized Hotfile's growth rate as being somehow unusual, but I disagree with that characterization and note that the level of relatively quick, consistent growth that Hotfile experienced up until February 2011 is consistent with patterns of growth for many successful online businesses.[4] In his analysis, Dr. Levy confuses changes in the size of a company's user base with changes in the behavior of those users. Hotfile's steady growth rate suggests that its user base was growing, and does not itself suggest that users' likelihood to engage in infringing activities changed over time. In contrast, the sudden drop-off in growth in February 2011 does suggest that Hotfile underwent a dramatic change at that time that would make it difficult to draw conclusions about levels of infringing activity at that time based on pre-February 2011 data.

29. Further, it is not uncommon for a business to have data from only certain periods of time. Therefore, statistical studies of operating businesses often must draw broader conclusions based on a subset of data that is available. I note that in the *Limewire*, *Fung*, and *Usenet.com* studies, for example, I had access to only a snapshot of information about site activity at a particular point in time. In *Limewire*, I analyzed a population of files that were available at a particular period of time in which the data about the files was collected. In *Fung*, I had access to a population of dot-torrent files that were available on the torrent websites from a particular period of time, as well as log data from one of the websites for a particular period of time. Similarly, in *Usenet.com*, I had access to a population of files on Usenet that were available at a particular period of time, and in fact Usenet services routinely delete files after a relatively short period of time. In each of these cases, I understand that information about

[4] See, for example, http://blogs.reuters.com/felix-salmon/2011/09/26/notes-on-groupon.

infringement occurring at a given period of time was credited by the reviewing courts in finding liability for infringement.

30. Finally, I understand that, as part of its business rules for recording downloads to compensate Affiliates, Hotfile records "dailydownload" counts by "free" (non-Premium) users located only in 54 specific countries. Thus, there are some downloads of files by "free" users in non-Affiliate countries in January 2011 that were not included in the population from which the statistical sample was drawn. This is not a criticism of whether the 90.2% figure is itself an accurate conclusion about the sampled population, but rather a suggestion that there is some set of downloads from Hotfile I have not yet analyzed in my study. However, as I explained above, I believe it is quite useful to examine the downloads that the Affiliate program was incentivizing. I note that in the *Usenet.com* case, for example, my study only looked at a subset of all available files, music files, that were infringed on Usenet.com. I understand that the court agreed that looking at that subset of content was useful in assessing the liability for infringement on the site.

31. In any event, there are not a substantial number of these downloads from "free" users in non-Affiliate countries. Based on the review of the data I have been provided, the vast majority – 84% – of the total "free" downloads from Hotfile are recorded in the "dailydownload" data. I have calculated this by comparing the "free" download counts of the sample files recorded in the "dailydownload" data with the "free" download counts in the "uploaddownload" data (which recorded a greater number of such downloads) for all time periods.[5]

[5] Hotfile's expert Dr. Levy has also suggested that some other categories of downloads may not be counted in the "dailydownload" totals because, for example, Hotfile suspects the downloading user may be cheating the Affiliate system. To the extent those downloads for January 2011 were nevertheless recorded in "uploaddownload," they are factored into the 84% figure above.

32. Further, based on the data and testimony I have reviewed, I have no reason to believe that the "free" downloads from non-Affiliate countries would have a materially different level of infringement than all other downloads from Hotfile. First, Hotfile's Anton Titov, who I understand to have been involved in the development and operation of Hotfile, testified that he had no reason to believe that the downloading patterns of users from the non-Affiliate countries were any different than that of users from Affiliate countries. Moreover, absent evidence to the contrary, that conclusion makes sense. Looking at the sample files recorded in the "dailydownload" table, my analysis of the unrecorded "free" downloads of those files shows the overall level of infringing downloads was not materially different.[6] That leads me to conclude that "free" users in non-Affiliate countries were downloading infringing files at a rate similar to all other Hotfile users.

33. In sum, my study here has determined that, on a daily basis, approximately 90.2% of the downloads from Hotfile are copyright infringing. I believe that this result is indicative of the level of infringement in months prior to January 2011. That level of infringement is remarkably consistent with the infringement levels found in other online infringement studies that have been credited by the courts.

[6] There is no systematic difference between the proportion of downloads not recorded in "dailydownload" as between infringing works and the remainder. In looking at the sample of 1,750 downloaded files, I found that on average 16% of downloads of infringing files by "free" users in non-Affiliate countries were not counted in "dailydownload," and on average 17% of the downloads of the remainder of files by "free" users in non-Affiliate countries were not counted in "dailydownload."

I declare under penalty of perjury that the foregoing is true and correct.

Executed on February 16th 2012, at Philadelphia, PA.

Dr. Richard Waterman