UNITED STATES DISTRICT COURT
SOUTHERN DISTRICT OF NEW YORK

ARISTA RECORDS LLC; ATLANTIC RECORDING CORPORATION; BMG MUSIC; CAPITOL RECORDS, INC.; ELEKTRA ENTERTAINMENT GROUP INC.; INTERSCOPE RECORDS; LAFACE RECORDS LLC; MOTOWN RECORD COMPANY, L.P.; PRIORITY RECORDS LLC; SONY BMG MUSIC ENTERTAINMENT; UMG RECORDINGS, INC.; VIRGIN RECORDS AMERICA, INC.; and WARNER BROS. RECORDS INC., Plaintiffs, v. LIME GROUP LLC; LIME WIRE LLC; MARK GORTON; and GREG BILDSON, and M.J.G. LIME WIRE FAMILY LIMITED PARTNERSHIP Defendants.	CIVIL ACTION NO. 06 CV. 5936 (GEL)

DECLARATION OF MATTHEW G. MERCURIO, PH.D. IN SUPPORT OF DEFENDANTS' RESPONSES TO PLAINTIFFS' MOTION FOR PARTIAL SUMMARY JUDGMENT

I, Matthew G. Mercurio, Ph.D., hereby declare as follows:

1. My name is Matthew G. Mercurio, Ph.D. I reside in San Anselmo, California. I am over eighteen years of age, of sound mind, and in all ways qualified and competent to make this Declaration. I have personal knowledge of the facts contained in this Declaration and they are true and correct.

2. I am a director in the Forensic and Litigation Consulting practice of FTI Consulting. I hold a BA degree in economics and mathematics from Boston University and masters and Ph.D. degrees in economics from Princeton University. I specialize in applied

statistical analysis and econometric modeling, and the application of those skills to public and private litigation, as well as other legal matters. My particular areas of expertise include survey design and sampling analysis, analysis of both cross-sectional and time-series data, and limited dependent and latent variable analysis. I have substantial experience in designing and implementing as well as reviewing sampling protocols in a wide variety of matters. A true and correct copy of my CV, including all publications I have authored during the last ten years as well as all cases in which I have provided deposition or testimony is attached as Exhibit A. The documents I have reviewed or considered in reaching my conclusions are listed in Exhibit B. A true and correct copy of the report I authored in this case is attached as Exhibit C.

3. Statistical sampling is a widely accepted and well understood method for drawing inferences from a subset of a population which can then be reliably extrapolated to the whole. Sampling is generally used when it is impractical, expensive, or impossible to analyze a population as a whole. The subject of sample design is concerned with how to select the part of a population which is to be included in the sample in a way which follows proper statistical practice to ensure correct and reliable results.

4. The first step in designing a sample is to define a population in terms of units of analysis. The *population* is the entire group of units of interest; the group about which the researcher wishes to draw conclusions. In this case, the population of interest is the set of all files made available by users of the LimeWire client.

5. Once the population of interest is identified, the next step is the construction of a list of all of the sampling units, commonly referred to as the *sampling frame*. In simple terms the sampling frame represents a list of the units of the population which are available for sampling.

6. There are many circumstances in which the sampling frame will not correspond exactly to the population. For example, consider a situation in which a researcher wishes to conduct a survey of individuals in a particular region to explore opinions on various modes of public transportation. To execute the survey, the researcher may choose to conduct a phone survey by selecting names from the telephone book. Therefore, the phone book represents the sampling frame in this example. However, the phone book is not a complete list of the population of interest – in the first place, it only represents households of interest rather than all persons. Furthermore, it does not include those persons who use only a cell phone or those who do not even have telephone service. Individuals may also have moved in or out of the area since the phone book was last printed, and other individuals may elect not to be listed in the phone book. Therefore, the results of such a sample can only be generalized to the universe of people whose names appear in the telephone book, which bears a complicated and uncertain relationship to the actual population which was originally desired. In any case, the results of such a survey cannot be reliably generalized to all the people in the region of interest.

7. In general the sampling frame might not represent all the available units in the population because it may be incomplete, partly illegible, or it may contain an unknown amount of duplicate values. In such cases it must be clearly explained how the survey will be used in making inferences about the population to assure the representativeness of the sample. The sampling frame must be consistent with the objectives and adequately reflect the universe of interest. If there are deficiencies in the sampling frame, an adjustment should be made to compensate for them. Then the sample, the set of units selected from the population for the study, can be drawn using the appropriate sampling methodology considering the context of the problem and the objective of interest.

8. I have reviewed Dr. Waterman's report in this case and I attended his deposition. I understand that Plaintiffs retained Dr. Waterman in this case in an attempt to develop a protocol to determine the authorization status of files made available for download via the LimeWire client and the frequency of download requests for those files. Dr. Waterman describes the exercise as a three phase procedure. Because there is no centralized listing of the files made available for download on the Gnutella network, the first phase of the protocol designed by Dr. Waterman was to develop a library of files (and associated file hashes) to serve as a sampling frame. Dr. Waterman refers to this database as the "Master Hash Library." Phases two and three of Dr. Waterman's protocol represent two sampling exercises.

9. In order for the sample drawn by Dr. Waterman to be valid, the sampling frame from which it was drawn must be a valid representation of the entire population at issue. Because there exists no valid sampling frame corresponding to the population of interest in this matter (all files made available for download using the LimeWire client), Dr. Waterman must, as a preliminary step, create a sampling frame to represent that population. In my opinion there are several issues which render the sampling frame constructed by Dr. Waterman inappropriate for the purpose of representing the population at issue. In short, as detailed below, the master hash library developed by Dr. Waterman does not provide a comprehensive or even representative image of all the possible files available for download on the LimeWire client at any point in time. The inadequacy of this sampling frame must to be taken into consideration in evaluating Dr. Waterman's conclusions and the reliability of his inferences about the population.

10. Dr. Waterman utilizes the "What's New?" search feature of the LimeWire client as an entry point into creating the master hash library. The actual files returned by the "What's New?" search are immaterial as far as the database is concerned – the search is only utilized as a

method for locating a selection of hosts. Once those hosts are found, the entire contents of their directories are consumed into the master hash library database. Dr. Waterman states in his report that "The client was directed to initiate a search request using the "What's New?" function with the default settings enabled." Dr. Waterman's protocol then directed the user to issue a "browse host" command on a randomly selected result from the "What's New?" search, with each result weighted by source count. The file hashes for all files in that user's shared directory were collected in a database. The collected file hashes from the execution of these steps over a one-week period form the master hash library upon which Dr. Waterman's report relies. In a subsequent step, Dr. Waterman's protocol initiated an attempt to download a random selection of these files.

11. Because the search results from the "What's New?" search form the basis of the selection of users whose search folders will subsequently be examined to create the master hash library, the results returned by this search are the most critical element in assessing the representativeness of the sampling frame developed using Dr. Waterman's protocol. As discussed in greater detail below, the sampling frame is biased and does not represent the population, and consequently, the results derived from Dr. Waterman's sample are statistically unreliable.

12. The most simple and straightforward method of sampling from a population is referred to as *simple random sampling*. When using simple random sampling, each item in the sampling frame has a known and identical probability of being drawn into the sample, and that probability is simply one divided by the total number of elements in the sampling frame. For example, Dr. Waterman uses a simple random sample to select elements from his sampling frame (the master hash library). He generates a list of 10,000 uniformly distributed random

numbers taking on a value from 1 to 6,908,689 (what he says is the size of the master hash library), with each value being equally likely. The sample of hashes drawn represents a simple random sample from the sampling frame.

13. While simple random sampling is straightforward to understand and apply, it is impractical in certain situations. For example, assume you wanted to conduct a survey of mass transit ridership in New York City, NY, and it was decided that a face-to-face interview with each survey respondent was necessary. Assume further that it was decided that a sample size of 400 was needed to generate sufficient precision for the results of the survey. A simple random sample of 400 individuals would be an exceedingly time consuming process, necessitating 400 separate trips across the five boroughs to complete the survey. However, you could divide the population into clusters of city blocks or high rise buildings, then select a small number of these, say 20, at which you would then sample 20 residents. This reduces the number of trips from 400 (one trip for each interviewee in the simple random sampling case) to just 20 trips using the latter method (one trip for each building or city block). This technique, referred to as *cluster sampling*, is an example of nonprobability sampling, in that each observation in the sample does not have an equal probability of being sampled (indeed, for clusters not selected the probability than one of the elements in that cluster will be selected is zero).

14. Thus, contrary to simple random sampling, where single subjects are selected from the population, in cluster sampling the subjects are selected in naturally occurring groups or clusters. The clusters themselves are referred to as the primary sampling units or PSUs, while the actual items of interest within the clusters are referred to as the secondary sampling units, or SSUs. This approach overcomes the constraints of costs and time associated with a highly geographically dispersed population.

15. The obvious drawback to cluster sampling should be apparent: The residents within each block or high rise building may have very similar lifestyles, and thus may have very similar attitudes towards public transportation, whereas in the simple random sampling case, each one of the 400 interviewees lives in a different area of the city. In the most extreme case, if each member of the cluster is identical in their attitude towards public transportation, then cluster sampling will only provide as much information as a simple random sample of 20 citizens, rather than 400. Because of this limitation, the formulas used to generate basic statistics such as the average and the confidence interval are totally different for cluster sampling than for simple random sampling. "One of the biggest mistakes made by researchers using surveys is to analyze a cluster sample as if it were a simple random sample."[1]

16. While the selection of file hashes from the master hash library in Dr. Waterman's protocol is based on simple random sampling, the original selection of files into the master hash library is based on cluster sampling, not simple random sampling. As discussed above, the "What's New?" search feature is used to select a sample of hosts (or primary sampling units). The "browse host" feature is then used to collect the file hashes for all files in that host's search directory (the secondary sampling units) into the master hash library.

17. First, the use of cluster sampling in the creation of the master hash library produces a sampling frame which is not representative of the universe of files made available by LimeWire users. Although it is not possible to do so, suppose one were able to draw a completely random sample of 10,000 files made available by LimeWire users at a point in time. Given the large number of users at any given time relative to the number of files shared by any one user, it is highly likely that these 10,000 files would be drawn from 10,000 separate hosts, and would reflect the tremendous diversity of files shared by these various users. Now suppose

[1] Lohr, Sharon L. Sampling: Design and Analysis. Pacific Grove, CA: Duxbury Press, 1999, p. 133.

instead one draws a sample of 100 users and takes 100 files from their respective search directories. This would also produce a total of 10,000 files, however this selection is likely to be far less diverse than the simple random sample. Those who prefer Rap or Classical music or modern art or scientific papers are likely to have significant numbers of those files in their search directories, and thus there will be large concentrations or clusters of those types of files. In the same way, Dr. Waterman's decision to select all of the files made available by a random selection of users results in a very different master hash library than would result from a truly random sample of all files made available. As discussed above, in cluster sampling not all elements of the population (in this case, file hashes) has an equal probability of being in the sample. For those hosts not selected by the "What's New?" search, their files have a zero probability of being selected.

18. Second, Dr. Waterman further states that in selecting hosts whose search directories were subsequently added to the master has library, each result from the "What's New?" search was weighted by its source count. This weighting imposes further bias because it only selects hosts (primary sampling units) which have popular files (*i.e.*, files made available by a large number of users) in their directories. Remember that the actual results of the "What's New" search are irrelevant in terms of the files actually selected for the master hash library – the results of the search are used to identify *hosts*. That is, the higher the source count the more likely the user who made that file available is to be selected. Using our example from above, now the selection of 100 hosts is no longer random but skewed towards hosts which have popular files available in their search directories. As such, collecting files only from hosts with popular files in their shared directories does not provide a representative image of all the files made available by users of the LimeWire client.

19. Because of the impracticality of actually downloading all seven million of the files identified for the creation of the master hash library, Dr. Waterman's protocol did not direct that the files in question be downloaded directly from the LimeWire client after the "browse host" command was issued. The attempt to download the files was made later, after a random subset of the master hash library was selected, using the magnet links that were downloaded when the file hashes were originally identified. According to information I learned from Sam Berlin of Lime Wire, Magnet links can only be used to download files directly from users who are not behind a firewall, however, and such users represent approximately only 30% of all users. In other words, the master hash library was constructed by collecting the search directories of both firewalled and non-firewalled hosts. Later, when the attempt was made to download a random sample of files from those hosts, if a selected item could not be downloaded, it was skipped and an attempt was made to download the next item on the list.

20. In this case, because magnet links do not work for hosts behind a firewall, most of the download requests initiated at this stage of Dr. Waterman's protocol would have failed, with requests for rarer files on the Gnutella network (those files made available by fewer users) more likely to fail. Because non-firewalled hosts are not a representative sample of all hosts using the LimeWire client, Dr. Waterman's master hash library is not a representative sample of all files made available by LimeWire users. In other words, the use of magnet links in Dr. Waterman's protocol increases the chances of having more unauthorized files in the master hash library, yielding a sampling frame that is not representative.

21. In computing the percentage of files in his sample which are "Confirmed Infringing (Record Company)" or "Highly Likely Infringing," Dr. Waterman excludes files that contain viruses, spam or spoof files, and pornography from his analysis. Furthermore, Dr.

Waterman excludes files for which Mr. German determined the authorization status as "unknowable." In principle, there is no reason to exclude such files. If the goal of the exercise conducted by Dr. Waterman is to determine the proportion of *all* content made available by LimeWire users that is infringing, then the exclusion of these files is not appropriate. In other words, because these files are likely to be authorized or noninfringing content, their exclusion serves only to reduce the denominator of his calculations and thus biases his estimate of the proportion of infringing content upward.

22. In addition, the meaning of the term "Highly Likely Infringing" used in Dr. Waterman's report (in reference to Mr. German's declaration) is not at all transparent. A significant proportion of the files labeled as "Highly Likely Infringing" are not audio or video files at all. While I do not claim to be an expert in computer file types or copyright status of particular files, my cursory review reveals that many of the files labeled as "Highly Likely Infringing" are in fact not unauthorized files at all, but various operating system files and small graphic images. In other words, Dr. Waterman's own protocol appears to indicate (despite the flaws outlined above) that a significant portion of files made available are in fact not infringing.

23. Even if the sample of files used in Dr. Waterman's analysis of download requests was representative of all files available, the requests for those specific files provide no insight into the population of download requests for all files made available by users of the LimeWire client at any given time. Consider the following: If a particular user's shared folder contains only unauthorized content, then 100% of the requests for downloads of those files will by definition be for unauthorized content. Similarly, if a particular user's shared folder contains only authorized content, then 100% of the requests for downloads of those files will by definition be for authorized content. Simply put, the underlying ratio of unauthorized to authorized content

in a particular set of files made available for download significantly affects the ratio of download requests for such content, completely irrespective of the ratio of such requests in the full population. Given a population which was, according to analysis by counsel for the Plaintiffs, 92.7% unauthorized content, it is not surprising that 98.8% of the requests for downloads were for unauthorized content; Indeed, it would be surprising if that were not the case. But this fact reveals nothing about the overall ratio of requests for downloads in the population if users of the LimeWire client.

24. In any event, as discussed above, the sample of files collected through Dr. Waterman's protocol does not represent a random sample of the universe of files made available by users of the LimeWire client. From a statistical perspective, it is virtually impossible to survey what users are searching for over the decentralized Gnutella network made up of millions of users. As such, Dr. Waterman's protocol is incapable of yielding meaningful results which can be reliably extrapolated to the total population of download requests.

25. Due to the structure of the LimeWire system, *i.e.*, the lack of any centralized listing of all available files and the fact that the list of available files changes depending on the various peers and ultrapeers who log on to and log off from the system, there is no representative master hash library available from which to sample and no conceivable way in which a valid statistical sample can be taken. Dr. Waterman attempts to create a representative sampling frame using his "master hash library" for the purposes of his report, but he ultimately falls short in that effort. Because the master hash library described in Dr. Waterman's report is assembled by collecting all of the files made available by a certain number of users rather than a truly random sample of all files made available by all users, it is not representative of the population. The weighting of the selection of hosts based on the popularity of the files being shared further

distorts the sampling frame. In addition, the sampling frame is constructed from collecting files from both firewalled and non-firewalled hosts, but the actual downloading of those files is attempted through magnet links, which cannot be used for firewalled user. As a result of these flaws, the sampling frame constructed by Dr. Waterman is biased and not representative of the population. Consequently, the sample drawn from the sampling frame is not representative, and the analyses performed by Dr. Waterman using the sample are flawed and statistically unreliable.

26. Since the sampling frame and hence the sample misrepresent the status of unauthorized files, the conclusions derived from the sample are neither reliable nor accurate. If the master hash library incorrectly contains a significant number of unauthorized files, then obviously the requests for downloads from those files will reflect the same inaccurate proportions. As such, in my professional opinion as a statistician, the results contained in Dr. Waterman's report should not be relied upon for any purpose.

I declare under penalty of perjury under the laws of the United States of America that the foregoing is true and correct and that this Declaration is executed in San Anselmo, California on September 10, 2008.

Matthew G. Mercurio, Ph.D.