

REPORT OF DR. INGRAM OLKIN

BACKGROUND AND QUALIFICATIONS

I am currently a Professor of Statistics and Education at Stanford University. My academic background is a Bachelor's degree in mathematics from the College of the City of New York in 1947, a Master's degree in mathematical statistics from Columbia University in 1949, and a Doctorate in theoretical statistics from the University of North Carolina, Chapel Hill. From 1951 to 1960 I served on the faculties of Michigan State University and the University of Minnesota, and from 1961 to the present, on the faculty of Stanford University.

I have been Chair of the Department of Statistics at Stanford University and the University of Minnesota. I am the recipient of a Guggenheim Fellowship; an Alexander von Humboldt Senior Research Fellowship; a Lady Davis Fellowship at Hebrew University; and an Overseas Fellow, Churchill College, Cambridge. I have received an Honorary Doctor of Science degree from DeMontfort University, as well as the Wilks Medal from the American Statistical Association. I have received a Lifetime Achievement Award from the American Psychological Association, and have been elected a Fellow of several statistical societies. I have served on numerous governmental panels and committees including the National Science Foundation, National Institutes of Health, National Center for Education Statistics, Bureau of the Census, and Research Council of the National Academy of Sciences.

My field of statistical expertise is in statistical inference, statistical models in the social and behavioral sciences, multivariate statistical analysis, and methods for combining information. I have written on sampling methods and have taught courses in

this area. I have authored or edited numerous books and over 150 scientific papers. My curriculum vitae and bibliography, which includes a listing of all my published papers, is attached as Appendix A.

I have not given court testimony during the past four years. I am being compensated at my standard hourly rate of \$400.

SCOPE OF WORK

I was asked by counsel for the plaintiffs in the lawsuits against NAPSTER (Index Nos. C99-05183 and C00-0074 [MHP]) to design and implement a sampling protocol that would allow one to draw fair conclusions about the activities occurring on the NAPSTER system regarding, and to provide an opinion on the following two subjects:

- (a) the percent of NAPSTER users that make copyrighted works available to others for unauthorized exchange using the NAPSTER system (the "User Project"); and
- (b) the frequency with which NAPSTER users actually trade copyrighted works without permission (the "Download Project").

SUMMARY OF CONCLUSIONS

1. User Project

Based on the sampling that I directed I have concluded that at any given time, virtually all users of NAPSTER are making copyrighted files available for downloading by others without the permission of the copyright owners.

2. Download Project

Based on the sampling that I directed I have concluded that a very high percentage of files actually exchanged by NAPSTER's users are copyrighted works that are likely being traded without permission.

DISCUSSION OF SAMPLING PROTOCOL

In order to create any sampling procedure, the sample should be representative of the population and not give undue weight to an "item." An item might be a time of day, an individual, and so on.

A second aspect is the determination of the size of the sample to be taken in order to provide with 95% confidence a margin of error that is no greater than 3%, which is a standard used, for example, in opinion polls. To accomplish this accuracy, standard statistical methodology yields a required sample size of 1067 items. A larger sample size would provide an even smaller margin of error, and accordingly, to be conservative, I used the figure of 1150 items.

The basis for the sampling procedure that I used is what is called "simple random sampling." In such a sampling procedure, each item has the same opportunity to be chosen as does every other item. Thus, for example, the choice of a random hour during a 24-hour day is obtained by labeling tickets from 1 to 24 and drawing one ticket from an urn. This is the mode of a lottery. However, it is more efficient to have a computer provide the draws instead of actually drawing tickets from an urn, and there currently exist a number of algorithms that provide for simple random sampling. Accordingly, all sampling of the data taken from NAPSTER was performed using an algorithm.

After we generated the two populations of material (User and Download), the files were sent to other persons for a copyright analysis. Specifically I understand that the information was sent to persons at the Recording Industry Association of America ("RIAA") and persons at the Harry Fox Agency ("HFA") to determine copyright ownership of the songs set forth in the data.

I have drawn my conclusions based upon the information obtained as a result of that analysis, and I did not independently perform such analysis. For purposes of my conclusions I assume reasonable accuracy of that analysis.

The actual sampling was designed in two parts: (1) for the User Project, a sampling procedure that took place independently of NAPSTER's participation, and (2) for the Download Project a sampling procedure of data that NAPSTER captured at times that I designated.

1. Sampling Protocol for the User Project

The sample was taken once every hour for four days for a total period of 96 hours. This permitted us to generate a very large database of millions of files and many thousands of users.

In order to use the NAPSTER search engine, a "word" is required. The term word is not well defined. It could be (1) the name of a performer, (2) the name of a song, (3) a meaningless set of letters, or (4) an actual word taken from the dictionary. We decided not to employ method (3), that is, meaningless letter combinations, on the assumption that this procedure would not replicate how NAPSTER users typically search. Methods (1) or (2), moreover, could generate some bias in that they would guide the result obtained, for example, known songs by known artists. Thus, for example, the name of a

popular performer would be expected to yield an overestimate of usage of that performer's works. In order to avoid such a bias we chose method (4) and randomly sampled words from an electronic dictionary.

Twenty-five of these randomly selected words were input into NAPSTER during each sampling session. To choose a word, a letter was chosen in proportion to its frequency in the English language. Given a letter, words were then chosen at random from the Random House Unabridged Electronic Dictionary.

Each search provided a list of users and file names. We used the results of these random searches to identify individual NAPSTER users. Then, the files that each user was offering through the NAPSTER system at that time were also captured.

The outcome of this procedure led to the collection of approximately 24,000,000 songs and approximately 28,000 users in the four-day period. It was now required to choose a subsample of 1150 users.

There are a variety of methods to choose a subsample from a larger sample. One such method is to list the items in the larger sample in all possible orders, thereby giving equal weight to every ordering of the items. Once this is done we choose a single ordering at random, and then choose the first number of elements for the subsample. This method is called a random permutation procedure, and is readily carried out by computer. To accomplish this I generated a random permutation from which the first 1150 users constituted the final subsample. Henceforth, for the purposes of this discussion, the word sample will actually refer to the subsample, when appropriate.

2. Sampling Protocol for the Download Project

I estimated that downloads from eight separate times would provide a statistically valid panorama of usage that would be representative of the general availability of files on the NAPSTER system. Two groups of days were chosen, one from weekdays (Monday, Tuesday, Wednesday, Thursday), and one from weekends (Friday, Saturday, Sunday). I then generated 16 time-day combinations from the weekday group, and 12 time-day combinations from the weekend group. Subsequently I randomly selected four times to be sampled from each group for a total of eight sample times. This allowed for a statistically valid sampling of the actual downloading occurring in NAPSTER's system.

Further, the sampling assured that the time-day choices were different. This resulted in the following list of eight time-day choices:

01:00 Thursday	02:00 Friday
02:00 Monday	10:00 Sunday
17:00 Monday	15:00 Friday
22:00 Wednesday	17:00 Saturday

For each of the eight times that a sample was run, we captured data from NAPSTER's servers for a five minute period. As a result of this data collection we obtained a population of approximately 574,185 files actually downloaded. We then generated a list of 1150 songs in a manner analogous to that described in the User Project to obtain our sample.

CONCLUSIONS AND FINDINGS

1. The User Project

For the User Project, the persons analyzing the songs were asked to review the files that each user had to offer to determine whether they could find two (2) entries that represent copyrighted songs offered for downloading without permission. We found this to be the case for all of the 1150 users selected.

2. The Download Project

For the Download Project, I have concluded the following based on my review of the data I have collected as set forth above:

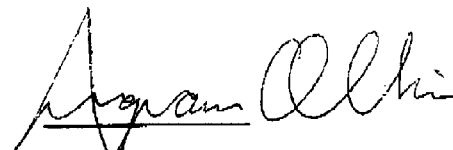
- (a) 1002 files, or 87.1 percent of the files in the sample, are songs that belong to or are administered by plaintiffs or other copyright holders and were selected for downloading using the NAPSTER system, without permission.
- (b) 37 files, or 3.2 percent of the files in the sample, are songs that are likely copyrighted and being traded without the authority of the rights holder, based on the preliminary analysis of RIAA/HFA, but sufficient information was not available to allow for definitive confirmation in the available time.
- (c) 3 files, or less than 0.26 percent of the files in the sample, are songs which are apparently being made available on NAPSTER without objection from the rights holders.
- (d) 108 files, or 9.4 percent of the files in the sample, did not present enough data for conclusion regarding copyright ownership in the available time.

The sampling procedures adhered to, in my opinion, indicate that the sampled files and users are representative of the NAPSTER population as a whole. Therefore, I am able to conclude that virtually all NAPSTER users offer copyrighted materials for distribution on NAPSTER without authority, and that copyright infringement appears to be a central and dominant part of the NAPSTER system.

I declare that the foregoing represents my opinions and conclusions under penalty
of perjury.

Stanford California

June 12, 2000



Ingrid Olkin, Ph.D.