UNITED STATES DISTRICT COURT

DISTRICT OF MARYLAND

WIKIMEDIA FOUNDATION; NATIONAL ASSOCIATION OF CRIMINAL DEFENSE LAWYERS; HUMAN RIGHTS WATCH; AMNESTY INTERNATIONAL USA; PEN AMERICAN CENTER; GLOBAL FUND FOR WOMEN; THE NATION MAGAZINE; THE RUTHERFORD INSTITUTE; and WASHINGTON OFFICE ON LATIN AMERICA,

Plaintiffs,

v.

NATIONAL SECURITY AGENCY / CENTRAL SECURITY SERVICE; ADM. MICHAEL S. ROGERS, in his official capacity as Director of the National Security Agency and Chief of the Central Security Service; OFFICE OF THE DIRECTOR OF NATIONAL INTELLIGENCE; JAMES R. CLAPPER, in his official capacity as Director of National Intelligence; DEPARTMENT OF JUSTICE; and ERIC H. HOLDER, in his official capacity as Attorney General of the United States,

Defendants.

Hon. T. S. Ellis III

No. 15-cv-00662-TSE

DECLARATION OF DR. ALAN SALZBERG

I, Dr. Alan Salzberg, do hereby state and declare as follows:

Introduction

1. I am the Principal (and owner) of Salt Hill Statistical Consulting. My work includes statistical sampling, analysis, and review for government and industry. On several occasions, I have written expert statistical reports or testified as a statistical expert, both in court and in depositions.
2. I received a Ph.D. in Statistics from the University of Pennsylvania, where I also received a B.S. in Economics. I have taught courses in statistics and quantitative methods at the University of Pennsylvania and American University and have published several statistics papers in peer-reviewed journals. I am also the co-inventor on a U.S. Patent (#6,636,585) for a statistical process design to test the systems of telecommunications companies. A copy of my resume is attached as an appendix to this report.
3. My current and recent work includes: statistical sampling and analysis of financial records on behalf of the United States Geological Survey; statistical review of the sampling and estimation methodology used to audit Medicaid providers in New York State on behalf of the New York State Office of Medicaid Inspector General; analysis of failure rates and survival modeling regarding the chances of catastrophic failure of an undersea oil field on behalf of a major construction company; statistical sampling and analysis, including regression modeling and survival analysis, on behalf of the U.S. Department of Labor; statistical modeling and prediction related to determining the number of prescriptions filled for a variety of pharmaceutical products in separate projects for a pharmaceutical company and for an industry data provider; review and testing of telecommunications data and statistical methods on behalf of public service commissions (including statistical sampling).

4. The purpose of this declaration is to assess the claim and accompanying probability calculation found in paragraph 58 of Wikimedia's First Amended Complaint, which states that "the odds of the government copying and reviewing at least one of the Plaintiffs' communications in a one-year period would be greater than 99.9999999999%." Compl. ¶ 58. This declaration reviews that statistical claim, explaining the necessary assumptions under which it would be correct and incorrect. The declaration first summarizes my findings and then provides details of the analysis that led to these conclusions.

Summary of Findings

5. The Plaintiffs give no statistical foundation in the Complaint for three assumptions[1] necessary to the calculation in paragraph 58 of the Complaint. These assumptions are:
    a. There is a 0.00000001% chance that the NSA copies and reviews any one communication.
    b. The chance of copying and review for each communication is the same; and
    c. The fact that one communication was or was not copied and reviewed does not affect the chances of the copy and review of any other communication.
6. As I explain below, each of these assumptions are unsupported by any statistical foundation in the Complaint. The assumptions are nevertheless necessary to support the

[1] Plaintiffs also assume that their collective number of international communications per year is more than one trillion. This appears to be based, in large part, on "88 billion HTTP or HTTPS requests" to/from Wikimedia websites, cited in paragraph 88, for May 2015. This number is presumably multiplied by 12 months to arrive at one trillion per year, *see* Compl. ¶ 88. The unstated assumptions regarding these requests are that all twelve months over the last year maintained the same number of communications and that any HTTP request is a "communication."

calculations made in the Complaint, and Plaintiffs' calculation would be invalid if any one of these assumptions is not correct.

7. Moreover, even if the calculation were correct that it is highly probable that at least one communication of one of the nine Plaintiffs' were copied and reviewed, it does not indicate that *each* of the nine Plaintiffs' communications were copied and reviewed. In fact, these chances could be far smaller, as I explain below.

8. Based on my analysis below, it is not statistically inconsistent for the NSA to have reviewed a very large number of communications but still have reviewed none of the Plaintiffs' communications.

Detailed Findings

9. Paragraph 58 of the Complaint performs a calculation regarding "the odds of the government copying and reviewing at least one of the Plaintiffs' communications." Compl. ¶ 58. The calculation puts those chances at greater than 99.9999999999%, a number that for all practical purposes is 100%. However, the calculation of these chances requires a number of assumptions.

10. The calculation is based on a statistical probability distribution called the binomial distribution. This distribution requires the assumptions that the number of items (communications) is known, that the chances of copying and reviewing are known and the same for each communication (statistically, this is called "identically distributed"), and that that the copying or reviewing of one communication has no effect on the chances

of copying and reviewing any other communication (statistically, this is called "independence of observations").[2]

11. The first assumption, that the chances of copying and reviewing any one communication is known and is equal to 0.00000001% is set forth specifically in paragraph 58, but no statistical foundation is provided for it in the Complaint. If that assumption is incorrect, the calculation changes as a direct result. For instance, if the chance of copying and reviewing any one communication is equal to 0.00000000001% instead of 0.00000001%, the chances that at least one Plaintiff communication is copied and reviewed falls to 10%, even assuming the total number of Plaintiff communications is equal to more than one trillion. Further, if the chance of copying and reviewing any one communication is equal to 0.000000000001%, the chances that at least one of Plaintiffs' communications is reviewed falls to 1%. In this way, the validity of this assumption can drastically affect the conclusion set forth in paragraph 58 of the Complaint.

12. When Plaintiffs' assumptions are applied in determining the chances that at least one communication for a particular Plaintiff was copied and reviewed, the chances fall, because whatever the totality of Plaintiffs' communications are, each particular Plaintiff will have less than that total. So even if the calculation were correct that it is highly probable that at least one of the nine Plaintiffs' communications were copied and reviewed, the chances that any particular Plaintiff's communications were copied and reviewed depends (at least in part) on the total number of communications for that Plaintiff and is lower than the percentage chance set forth in paragraph 58.

[2] For my calculations, I used the R language function pbinom. The same can also be accomplished using the Poisson distribution in a situation in which typical calculators cannot precisely perform the calculation.

13. In order to perform the exact calculation, we would need to know the total number of communications for that particular Plaintiff. For example, if Plaintiff The Rutherford Institute of Charlottesville, Virginia, had one million communications each year, then the chances that at least one of that Plaintiff's communications (as opposed to Plaintiff Wikimedia's communications) would be copied and reviewed would be only about 1 in 10,000 (0.01%); this calculation assumes, of course, that the chance of copying and reviewing any one particular Plaintiff's communication remains the same as stated in paragraph 58 (0.00000001%).
14. The two implicit assumptions of "independence" and "identically distributed" (often grouped together and called "iid") are also critical to the calculation. The iid assumptions mean that the chances of copying and reviewing are the same for all communications and that the chances of any one item being copied and reviewed does not vary based on whether any other item is copied and reviewed. Thus, the assumption means communications from anywhere in the world all have equal chances of being copied and reviewed, such that the chance of copying and reviewing of a communication by someone in Iran is the same as the chance of copying and reviewing a communication by someone in Ireland. Furthermore, these assumptions also mean that if a communication sent from Iran from a particular computer at a certain time was copied and reviewed, the chances that a communication sent from that same computer one second later has no more or less chance of being copied than the original 0.00000001%.
15. Any clustering of the copying and reviewing of communications, whether by country or some other criteria, would mean that some groups would have different chances of being copied than some other groups and that the fact that a particular communication in one

group is reviewed or copied means other communications in that group are more likely to be copied.

16. The iid assumptions are sweeping but are nonetheless necessary for the calculation to be correct. In order to account for or remove them, we would need to know the specific chances that are appropriate to apply to Plaintiffs' communications and the exact nature of how the Plaintiffs' communications are clustered.

17. By way of illustration of the iid assumptions, consider a statistical survey that selects people at random from some population. Such a survey has a selection method that is iid—selection of one person provides no information on whether another person is sampled and the chances of any one person being selected are the same. Only careful attendance to the mechanics of the survey—delineation of all possible respondents and statistically random sampling of a set of them—can ensure that the survey is truly random and that the iid property holds.

18. A statistically haphazard survey will generally be far from random. Consider a survey, even a very large one, where someone stands on a street corner and questions passers-by. This survey is certainly haphazard in design, and it is equally certainly not random. For example, even if it is known that on a random day 10% of people in the U.S. carry umbrellas, a survey done in Phoenix on a sunny summer day is unlikely to yield any people with umbrellas while one done in Seattle on a rainy winter day is likely to yield many. The assumptions the Plaintiffs use would say that if 1,000 are surveyed, then there

is a greater than a 99.9999999999% chance someone surveyed will be carrying an umbrella without regard to whether the survey was in Seattle or Phoenix.[3]

19. Likewise, even a very large operation of copying and reviewing communications may completely miss some communication while copying and reviewing nearly 100% of others. To be accurate, the Plaintiffs' calculation requires that the copying and review of communications be like a good statistical survey in that the selection for copying and reviewing is random. But Plaintiffs' assertions about how the process works—through the copying of "*certain* high-capacity cables, switches, and routers" (Compl. ¶ 49)—would mean, if accurate, that the process is, in statistical terms, haphazard like the survey

[3] These two assumptions are also critical to the accuracy of percentage chances in the scenario of an hourly forecast of rain. Suppose that the chances of rain any morning hour between 9 and noon are 0.6, or 60%. If this is the case and the chances are iid in each of the three hours 9am to 10am, 10am to 11am and 11am to noon, then the chances of *no* rain in any hour is 40% (100% - 60%), or 0.4, to the power of 3, which equals 6.4%. If the identically distributed assumption is violated the chances of rain could average 60% but would be different each hour. Thus, the chances in the first hour could be 100% and the chance in each of the next two hours could be 40%, a combination which still produces an average of 60%. However, the chance of no rain is 0 rather than 6.4% since during the first hour the chance of rain is 100%. And if the independence assumption is violated, the chances may be correct at 60% per hour, but, if it is not raining the first hour, it may be very unlikely it will rain in either of the other two. In this case, the chance of no rain would be 40% for the first hour, but 0% in the second and third hour if it does not rain in the first hour (and 100% in the second and third hour if it does rain the first hour). This would be the case if a storm that lasts three hours may or may not hit the area. If it does hit the area, it will begin between 9am and 10am and continue through noon. In this case, the chances of no rain between nine and noon are 40%. This example also shows that without the iid assumption, which allows the chances for each time period to be treated independently and all chances to be assumed to be the same, the calculation of the chances can be far different than with the assumption.

with the umbrellas.[4] Therefore, the statistical assumptions Plaintiffs have made in paragraph 58 are inconsistent with how they say this copying and reviewing process works. Using my earlier example of the umbrella survey, Plaintiffs have calculated the chances of any one person carrying an umbrella by using a studiously random statistical model to determine how many people are carrying an umbrella without regard for whether the survey itself occurred in Phoenix on a summer day or on a rainy day in Seattle in the winter. In terms of the umbrella survey, however, Plaintiffs have tried to apply that statistically random model to a statistically haphazard survey that occurs in certain cities at a certain time of the year.

20. In conclusion, the chances calculated in paragraph 58 of the Complaint depend on assumptions for which no statistical basis is provided in the Complaint. If any of these assumptions are incorrect—and Plaintiffs' description of the process of copying and review suggest that these assumptions are incorrect—then the chances of one of Plaintiffs' communications being copied and reviewed could be far less than 100%.

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

DATE: August 4, 2015

Alan Salzberg
Digitally signed by Alan Salzberg
DN: cn=Alan Salzberg, o=Salt Hill Statistical Consulting, ou, email=salzberg@salthillstatistics.com, c=US
Date: 2015.08.04 08:13:28 -04'00'

ALAN SALZBERG

[4] Such a method of copying and reviewing, if the NSA does in fact use that method, may mean that Plaintiffs' communications have no chance of being copied, as would be the case if Plaintiffs' communications do not happen to go through the copied cables, switches, and routers.

ALAN J. SALZBERG, PH.D.
salzberg@salthillstatistics.com
646-461-6153

EXPERIENCE

Salt Hill Statistical Consulting, Founder and Principal, 2000-present
Founder and Principal of a statistical consulting company (formerly Quantitative Analysis). The firm is skilled at presenting complex ideas to non-experts. Capabilities include development and implementation of statistical techniques as well as critical review and audit of existing statistical estimates, samples, and models. The company's clients are law firms, government, and private corporations and have included: United States Department of Labor; Pfizer; Barnes & Thornburg; Honeywell; K&L Gates; City of New York

Summit Consulting, Teaming Partner, 2009-present
Consult on multiple engagements with economic consulting firm on large-scale government projects. Served as a Director at the firm in 2014.

Analysis & Inference, Inc., CEO, 1991-1995 and 2008-2013
Led a statistical consulting company that provides consulting services to corporations, law firms, and government.

KPMG LLP, Practice Leader, Quantitative Analysis Group – New York, 1996-2000
Established and led the New York office of KPMG's Quantitative Analysis Group. Built a consulting practice with annual revenues of $4 million.

Morgan Stanley, Associate, 1988-1990, 1995-1996
Performed statistical modeling and software design.

EDUCATION

Ph.D., Statistics, Wharton School, University of Pennsylvania, 1995
M.A., Statistics, Wharton School, University of Pennsylvania, 1992
B.S., Economics (concentration in Economics and Finance), *cum laude*, Wharton School, University of Pennsylvania, 1988

ENGAGEMENTS

- Served as a statistical consultant in the development of dynamic models for residential property valuation across the United States in order to determine whether certain residential mortgage-backed securities (RMBS) were fairly valued. Made use of statistical and econometric techniques including regression modeling, statistical sampling, bootstrapping, and bias adjustment.

- On behalf of a Fortune 100 company, evaluated models that estimated the potential liability in more than 10,000 asbestos settlements. In addition, reviewed the likely bias and other issues with

a model that predicted the "propensity to sue" for future claims. Wrote two expert reports concerning findings and testified as a statistical expert regarding those findings.

- On behalf of the New York State Office of Medicaid Inspector General, reviewed the sampling and estimation methodology used to audit Medicaid providers in New York State. Reviewed and critiqued specific methodologies in ongoing matters, and provided recommendations for improving the statistical audit process.

- In a series of matters on behalf of the law department for a major city, created and analyzed a massive real estate database, modeled market and sales values, and wrote expert reports to determine potential biases of alternative methods of valuing commercial real estate. Determined the validity of assumptions about lease lengths, turnover rates, and other issues affecting rents and property values. Testified as a statistical expert in one of these matters.

- On behalf of the United States Department of Labor, acted as the principal investigator on a study of industry compliance with certain labor laws. Developed and pulled a statistical sample for evaluation. Performed survival analysis to better understand how long certain industry investigations would last and the likely outcomes of such investigations.

- For major pharmaceutical company, analyzed company and external marketing data to determine reliability and potential biases in using external data sources. Analyzed physician-specific data for a period of 36 months concerning product marketing to approximately 1 million prescription drug subscribers.

- In complex litigation matter involving an undersea oil field, analyzed data from several years of inspections and repairs to determine likelihood of a catastrophic failure that would result in a major oil spill. Used survival analysis to determine the likelihood of such an event for different inspection and repair cycles.

- On behalf of several state public service commissions, directed data analysis and statistical design in a series of tests of Bell South, Verizon, SBC-Ameritech, and Qwest. Beginning in 1998, developed software and procedures for calculating performance metrics and evaluating the competitive environment. Testified before several state public service commissions, including New York, Virginia, Florida, Michigan, and Colorado.

- Using social security and insurance company data, developed two probability-based models in order to match unclaimed assets with the individual owners of those assets. The models were successfully implemented at our client, a financial services company, and used to assist state agencies in locating unclaimed assets.

- For hedge fund, performing an ongoing series of projects related to pricing risk and return of various investment options. Using standard and proprietary statistical techniques and software, developing models to select appropriate investment funds according to risk and term of investment.

77 Water Street, 7th Floor • New York, NY 10005 • 646 461 6153 • salzberg@salthillstatistics.com

www.salthillstatistics.com

- For large direct market publisher, improved customer response modeling while reducing the costs of test marketing. Overall test marketing was reduced by combining data for various market segments. This method also increased the precision of the scores assigned to customers concerning their propensities to purchase individual books. These improvements were expected to lead to cost savings and revenue improvement totaling about $1 million annually.

- Modeled television audience ratings to determine the Public Broadcasting System's share of cable royalty distributions. Used statistical methods to determine a reliable estimate of PBS's cable royalty share. The estimate resulted in a multi-million dollar decision in favor of the Public Broadcasting System by the Cable Royalty Tribunal.

- Lead statistician in the design and implementation of a sample of all personal property and equipment on behalf of the United States Internal Revenue Service. The population of interest involved more than one million items contained in over 1,000 buildings. The sample design, implementation, and resulting estimates and projections were subject to intense scrutiny by the United States General Accounting Office.

- For the United States Department of Justice, designed and implemented a sample to estimate the number of immigrants improperly granted citizenship. The sample was designed to provide precision of plus or minus less than 1%, for a population of more than 1 million immigrants. The work was the focus of intense congressional scrutiny and received substantial review in the media.

- On behalf of Fortune 100 company, created statistical models to determine the probabilities and likely severities of accidents for different employee and accident types. This project resulted in recommended annual savings of $3 million.

- On behalf of the Arava Institute of Environmental Studies, advised on design and sampling methodology for a broad-based survey of environmental education in middle and high schools. More than 7,000 students were surveyed in a sample that was stratified by size of town, income level, and other socio-economic variables. Performed weighted statistical analysis to project survey results to the population. Presented results before Israeli Congressional committee in July 2007.

- For the United States Customs Service (Department of Homeland Security), assisted with sampling of financial statement information. Designed and wrote sampling plans, helped implement the plans, and created spreadsheet calculator to analyze results. In an earlier engagement, evaluated the credibility of statistical sampling and analysis used to track and categorize imports, for the Office of Inspector General. Suggested improved methods of sampling and implementation.

- Designed and implemented several studies of stock basis in corporate mergers. One universe comprised over 100 million shares and more than 20,000 shareholders, yet the sample design resulted in a highly precise estimate using data for fewer than 1,000 shareholders.

RESEARCH

77 Water Street, 7th Floor • New York, NY 10005 • 646 461 6153 • salzberg@salthillstatistics.com

www.salthillstatistics.com

An excerpt from my "What are the chances?" blog appears in Lundsford, Andrea L. and Ruszkiewicz, John, *Everything's an Argument, 6th Edition, 2012* and Lundsford, Andrea L., Ruszkiewicz, John, and Walters, Keith, *Everything's an Argument with Readings, 6th Edition, 2012*.

"Law and Statistics of Combining Categories: Wal-Mart and Employment Discrimination Cases", with Albert J. Lee, *Proceedings of the 2010 Joint Statistical Meetings of the American Statistical Association*, 2010.

"Evaluating the Environmental Literacy of Israeli Elementary and High School Students," with Maya Negev, Gonen Sagy, and Alon Tal, *Journal of Environmental Education*, Winter 2008.

"Trends in Environmental Education in Israel," with Gonen Sagy, Maya Negev, Yaakov Garb, and Alon Tal, *Studies in Natural Resources and Environment*, Vol. 6, 2008. [In Hebrew]

"Results from a Representative Sample in the Israeli Educational System," with Gonen Sagy, Maya Negev, Yaakov Garb, and Alon Tal, *Studies in Natural Resources and Environment*, Vol. 6, 2008. [In Hebrew]

"Comment on Local model uncertainty and incomplete-data bias by Copas and Li," with Paul R. Rosenbaum, *Journal of the Royal Statistical Society, Series B*, 2005.

"Determining Air Exchange Rates in Schools Using Carbon Dioxide Monitoring", with D. Salzberg and C. Fiegley, presented at the *American Industrial Hygiene Conference and Expo*, 2004.

"The Modified Z versus the Permutation Test in Third Party Telecommunications Testing", *Proceedings of the 2001 Joint Statistical Meetings of the American Statistical Association*.

"Removable Selection Bias in Quasi-experiments," *The American Statistician*, May 1999.

"Skewed oligomers and origins of replication," with S. Salzberg, A. Kervalage, and J. Tomb, *Gene*, Volume 217, Issue 1-2 (1998), pp. 57-67.

"Selection Bias in Quasi-experiments," (Doctoral Thesis), 1995.

***Patent** (#6,636,585)* One of five inventors on a patent for statistical process design related to information systems testing.

PRESENTATIONS

- Panelist and Presenter of "Secrets to Effective Communication for Statistical Consultants,", Joint Statistical Meetings of the American Statistical Association, 2013, with Ghement, Isabella; Mangeot, Colleen; Rantou, Elana; Schuenemeyer, Jack; and Turner, Ralph.

- Lectured on "Statistics in Predictive Coding" as part of a one day seminar sponsored by the Cowen Group and Equivio in the area of e-discovery, 2012.

77 Water Street, 7th Floor • New York, NY 10005 • 646 461 6153 • salzberg@salthillstatistics.com

www.salthillstatistics.com

- Presented paper (with Albert Lee) entitled "Law and Statistics of Combining Categories: Wal-Mart and Employment Discrimination Cases" at the Joint Statistical Meetings of the American Statistical Association, 2010.

- Delivered presentation on census data from the New York City Housing and Vacancy Survey, before the New York City Rent Guidelines Board, 2007.

- Part of a team of five presenting results before an Israeli congressional committee regarding a nationwide public school survey, 2007.

- Served on panel and presented "The Modified Z versus the Permutation Test in Third Party Telecommunications Testing" at the Joint Statistical Meetings of the American Statistical Association, 2001.

- Delivered talk regarding "Skewed oligomers and origins of replication" at Hebrew University in Jerusalem, 1999.

PERSONAL

Married, with two daughters and a son.
Languages: English (native), Hebrew (conversational).
Member, Park Slope Food Coop.
Member, 39 Plaza Housing Corp (residential coop). Board member, 2012-2015.
Enjoy ultimate Frisbee, basketball, biking, hiking, running, tennis, chess, and bridge.

77 Water Street, 7th Floor • New York, NY 10005 • 646 461 6153 • salzberg@salthillstatistics.com
www.salthillstatistics.com