On the *Google*-Fame of Scientists and Other Populations

James P. Bagrow* and Daniel ben-Avraham*

**Department of Physics, Clarkson University, Potsdam NY 13699-5820*

Abstract. We study the fame distribution of scientists and other social groups as measured by the number of *Google* hits garnered by individuals in the population. Past studies have found that the fame distribution decays either in power-law [1] or exponential [2] fashion, depending on whether individuals in the social group in question enjoy true fame or not. In our present study we examine critically *Google* counts as well as the methods of data analysis. While the previous findings are corroborated in our present study, we find that, in most situations, the data available does not allow for sharp conclusions.

1. INTRODUCTION

The concept of Fame within a population has critical social and economic impact. Recently, the idea of using the number of *hits* returned from a search of a person's name on *Google* as a means of quantifying that person's fame has been explored [1, 2]. A seminal paper explored the fame of a unique population, that of World War I "ace" pilots [1], and found, among other things, a power-law decay in the tail of the distribution. More recent work [2] has applied this to a population of scientists who have published on the *cond-mat* e-print archive[1]. The tail of their fame distribution was best fit by an exponential. On the other hand, the fame of other populations was found to follow a power-law decay. The difference was attributed to the fact that scientists habitually use the World Wide Web as a professional means of communication and cite each other on the web in relation to their published work.

Google's goal as a service is to provide accurate search results to its users. For the purposes of determining a subject's fame, what is most relevant is not having accurate results listed first, as it is for most users, but to have an accurate *count* of those results. Unfortunately, *Google* does not provide enough accuracy, and there are several reasons for this [3].

Google acknowledges that the hits count given is an estimate, but does not elaborate on the accuracy of this estimation nor reveal how it is calculated. It seems reasonable to assume that very small counts are more accurate than larger ones. This means that the error is largest in the tail of the fame distribution, and it is this region that is of most interest. In addition, the tail of the distribution is more likely to contain results that are over-counted, further compounding the error.

In [1], over-counting was prevented by verifying each hit by hand, a time-consuming procedure that limited the sample size. At the time of this writing, *Google* only returns the first 1000 hits, so it is impossible to verify the accuracy of any results beyond that number, and one must trust in *Google's* estimation. Even manual verification is limited.

The previous searches in [1, 2] used a search lexicon including the boolean `OR` operator. We have since found out that Google returns incorrect hit counts when `OR` is used [3]. For a simple illustration, a search for `cars OR automobiles` returns 80.5 million hits (at the time of this writing) while searches for `cars` and `automobiles` return 94.2 million and 8.82 million hits, respectively, violating basic set theory. Thus, the previous work must be reproduced using a better lexicon. In the current work, all our searches avoid the problematic `OR` operator. See Table 1.

Despite these issues, *Google* still provides an excellent tool for research. It is the simplest means of getting the most information available and it commands a very large sample space. For example, the work in [4] uses hit counts to "teach" the semantic meaning of words to software — a central problem in Artificial Intelligence. Related words such as 'painter' and 'artist' will have many more joint occurrences than disparate words, such as 'plumber' and 'artist', leading to higher hit counts. Their work confirms that *Google* yields reasonable results when avoiding the `OR` and using the `AND` operator only.

Google has been generous enough to open their search interface to allow tools to be created that can perform

[1] `http://arxiv.org/archive/cond-mat`

Google searches automatically[2]. We have used this to eliminate the laborious task of entering single queries and recording the hits count. Larger populations can be searched much more quickly using an automatic tool. For the present work, we used a script that performed the searches from a web server. An easier way still is with the open-source *PyGoogle* package[3], which integrates the *Google* search interface with the Python programming language.

In both [1] and [2], fits were performed by binning the search results with exponentially-sized bins and then fitting to the binned data using least-squares. A better technique than binning, when working with sparse data, is examining cumulative distributions [5], as we do in the present work. Also, it has been shown that there are problems with using least-squares fits to logarithmic plots [6]. One problem is that the log operation magnifies the error in the tail. Least-squares fitting assumes that errors for each data point are uniform and will not properly weigh the noisier tail. In this work, we use a more robust technique, that of Maximum Likelihood, to achieve less biased fits. See section 2.

All distributions studied here exhibit a power-law tail, although for many the tail covers a very narrow range. For the scientists populations, we observe a power-law tail only in the top 12% of the data. Most of the data for scientists is best fit by an exponential, just as found in the previous study [2]. In contrast, other populations do not fit to an exponential over any sizable range. See section 4.

2. MAXIMUM LIKELIHOOD FITTING

A better technique to determine the parameter(s) of a probability distribution from sampled data is that of Maximum Likelihood. The results are more robust in terms of error-weighing. This is a very common technique and is covered in many statistics and regression texts.

To briefly illustrate how maximum likelihood works, let us derive the *Maximum Likelihood Estimator* (MLE) for λ, the parameter for an exponential probability distribution:

$$P(x) = \lambda e^{-\lambda x} \tag{1}$$

The goal of Maximum Likelihood is to find the *most likely* λ given the existing data. For this, we start with the probability of the experiment given λ, assuming independence of the data points:

$$P(x_1, x_2, \ldots x_N | \lambda) = \prod_{i=1}^{N} \lambda e^{-\lambda x_i} \tag{2}$$

$$= \lambda^N \exp\left(-\lambda \sum_{i=1}^{N} x_i\right) \tag{3}$$

where x_i is the unbinned data gathered from the experiment. This function is the total probability of all measurements occurring in the experiment. From this, we define the *likelihood function*, using Bayes' Theorem:

$$l(\lambda | x_1, x_2, \ldots x_N) = P(x_1, x_2, \ldots x_N | \lambda) \frac{P(\lambda)}{P(x_1, x_2, \ldots x_N)} \tag{4}$$

This is the *likelihood* of λ given the experimental data. Assuming $P(x_1, x_2, \ldots x_N) = 1$ (the experiment has already occurred) and $P(\lambda)$ is uniformly distributed (all λ's are equally likely), then $l(\lambda|x) \propto P(x|\lambda)$. To find the most likely λ, we must find the maximum of this function with respect to the parameter λ. To simplify the calculation, we will instead maximize the log-likelihood function, L, which is equivalent:

$$L = \ln(l) = N \ln \lambda - \lambda \sum_{i=1}^{N} x_i \tag{5}$$

$$\frac{dL}{d\lambda} = \frac{N}{\lambda} - \sum_{i=1}^{N} x_i = 0 \tag{6}$$

$$\lambda = N \Big/ \left(\sum_{i=1}^{N} x_i\right) \tag{7}$$

This is just the inverse of the mean, exactly as expected for an exponential distribution. We need not account for the proportionality between $l(\lambda|x)$ and $P(x|\lambda)$ because we only used the derivative of L

The other probability distribution we are concerned with is the power-law distribution:

$$P(x) \propto x^{-\gamma} \tag{8}$$

For the MLE of γ, we reproduce the derivation given in [5]. The first step is to normalize Eqn. (8) for the given data points:

$$P(x) = Cx^{-\gamma} = \frac{\gamma - 1}{x_{\min}} \left(\frac{x}{x_{\min}}\right)^{-\gamma} \tag{9}$$

where C is a constant of proportionality and $x_{\min}$ is the smallest data point from the given sample. This is then used to get the probability of the experiment:

$$P(x_1, x_2, \ldots x_N | \gamma) = \prod_{i=1}^{N} P(x_i) = \prod_{i=1}^{N} \frac{\gamma - 1}{x_{\min}} \left(\frac{x_i}{x_{\min}}\right)^{-\gamma} \tag{10}$$

[2] http://code.google.com

[3] http://pygoogle.sourceforge.net

This is proportional to $l(\gamma|x)$ as before:

$$l(\gamma|x_1,x_2,...x_N) = \prod_{i=1}^{N} \frac{\gamma-1}{x_{\min}} \left(\frac{x_i}{x_{\min}}\right)^{-\gamma} \tag{11}$$

Again, we work with $L = \ln l$, which is equivalent for finding the most likely γ. Then:

$$L = \sum_{i=1}^{N} \left(\ln(\gamma-1) - \ln x_{\min} - \gamma \ln \frac{x_i}{x_{\min}}\right) \tag{12}$$

$$= N\ln(\gamma-1) - N\ln x_{\min} - \gamma \sum_{i=1}^{N} \ln \frac{x_i}{x_{\min}} \tag{13}$$

The MLE for γ can then be found:

$$\frac{dL}{d\gamma} = \frac{N}{\gamma-1} - \sum_{i=1}^{N} \ln \frac{x_i}{x_{\min}} = 0 \tag{14}$$

$$\gamma = 1 + N/\left(\sum_{i=1}^{N} \ln \frac{x_i}{x_{\min}}\right) \tag{15}$$

Maximum Likelihood derives an estimator for a distribution's parameter(s), regardless of whether the sampled data truly does come from such a distribution. Hence, one needs a way to test how well the estimator matches the sample. For our purposes, the Kolmogorov-Smirnov (KS) Test works quite well [6]. This test compares the cumulative distribution function (CDF) of the hypothesized probability distribution to the empirical CDF of the sampled data. The test statistic is:

$$K = \sup_x |F(x) - S(x)|, \tag{16}$$

where $F(x)$ is the hypothesized CDF and S(x) is the empirical CDF. K is then compared with a critical value (for the given significance level) which can be found in a table or generated by software. MATLAB's Statistics Toolbox has a built-in KS-Test function, `kstest()`.

3. POPULATIONS

We have been able to greatly expand upon the number of searches performed compared to previous work. In addition, due to the problems with the `OR` operator, we have performed multiple searches of the same population using progressively inclusive lexicons. Here we describe the populations studied.

Scientists: Two populations of scientists were used in this study. The smaller one (of size 449) is the same population used in [2]. The larger population (of size 1625) is a list of authors who have published recently on cond-mat and was harvested using arXiv's OAI XML feed[4].

Aces: The population of 1851 aces contains the 393 German aces studied in [1] as well as all the listed aces of other nationalities[5].

Actors: The actors population contains 778 actors who were born on the second or third of each month between the years 1950 and 1955, as collected from the archives of the Internet Movie Database[6]. These selection criteria were chosen to insure a mostly uniform sample and to give all the chosen subjects roughly the same career length.

Villains: The villains population was gathered from a user-contributed list of antagonists from fictional media[7]. This list contains both fictitious characters and real people who have appeared in fictional works. Since this list was generated by users, the characters must already enjoy a substantial level of popularity.

Programmers: Similar to the villains population, this population was collected from a user-contributed list of famous programmers[8]; people who have made a large contribution to computing, the Internet, etc., such as Tim Berners-Lee, who invented the World Wide Web, and Bill Gates, a co-founder of Microsoft. As with the villains, it seems safe to assume that this population is "famous".

Clarkson Students: The students population was chosen from Clarkson University's student directory. It consists of all students (undergraduate and graduate) whose last name contains the letter "e". This criterion was chosen simply to make it easy to harvest a large collection of names from the online student directory. We assume this is a "non-famous" population, in that the students are too young to have amassed any real fame.

Runners: This population was used previously [2]. The original searches used the erroneous `OR` operator and are here reproduced without it.

4. RESULTS AND ANALYSIS

Table 1 contains the power-law exponents and search lexicons for the populations studied. Many of the power-law exponents are ≈ 2, as first predicted in [1]. All populations display a power-law tail, regardless of whether they

[4] See `www.openarchives.org`

[5] See `www.theaerodrome.com`

[6] `www.imdb.com`

[7] See `en.wikipedia.org/wiki/List_of_villains`

[8] See `en.wikipedia.org/wiki/List_of_programmers`

are “famous” or not. It should be pointed out, however, that for some populations the range fitting a power-law is extremely narrow, casting doubt on this interpretation. In those cases, an exponential distribution may fit as well. Most of the scientists distributions fit an exponential over much of the “non-tail”. See Table 2. Clarkson students, another population assumed to be non-famous, does not fit to an exponential over such a range. This is further evidence that the exponential distribution for scientists stems from their use of the World Wide Web as a professional means for disseminating research, rather than related to fame.

The power-law exponent tends to increase as the restriction due to the lexicon increases. This is expected because a more restrictive search will make high hit counts less frequent, increasing the slope of the tail. Figure 1 contains *rank / frequency* plots for several populations to illustrate this effect. The plots are proportional to the empirical CDF, $P(X > x)$. Note that individual searches which return zero hits are not shown, changing the maximum rank between lexicons. This is most evident in the Students population: the third lexicon is very restrictive and many students garnered zero hits.

The proposed model in [1] was shown to have a power-law exponent that approaches 2 asymptotically, from *above*, as the number of relevant web pages citing the population in question increases over time. The population changes in size in Table 1 are due to progressively restrictive lexicons and do not pertain to the same phenomenon. On the other hand, we are unable to account for the many instances of power-law exponents smaller than 2 observed, as any reasonable extension of the theory in [1] yields powers $\gamma \geq 2$.

5. CONCLUSIONS

A purely visual inspection of plots such as those in Figure 1 may lead one to conclude that a search is exponentially or power-law distributed, but this is misleading and subjective. The eye will overweigh the number of data points in the tail, due to the logarithmic axes. Objective hypothesis tests such as the KS-test must be used.

In addition to problems with hit estimation, `OR`, etc., the choice of a lexicon has noticeable impact. In the rank / frequency plot for the aces population in Figure 1, the second search shows a much cleaner tail, though again this region contains less than 6% of the aces. All of these factors make it difficult to test theories. For the size of populations involved, *Google* hits have too much “noise” to accurately distinguish distributions.

ACKNOWLEDGMENTS

We are grateful to Eduard Vives for suggesting the analysis of cumulative plots by the method of maximum likelihood and to the Universidad de Granada for facilitating discussions leading to this work. DbA thanks NSF grant PHY-0140094 for partial support of this work.

REFERENCES

1. M.V. Simkin and V.P. Roychowdhury, “Theory of Aces: Fame by chance or merit?" (preprint, arxiv.org/abs/cond-mat/0310049, 2003).
2. J.P. Bagrow, H.D. Rozenfeld, E.M. Bollt, and D. ben-Avraham, “How Famous is a Scientist? – Famous to Those Who Know Us." cond-mat/0404515, Europhys. Lett., **67**, (4) 511-516 (2004).
3. G.R. Notess, “Google Inconsistencies.” `http://www.searchengineshowdown.com/features/google/ir`
4. R. Cilibrasi, and P.M.B Vitanyi. “Automatic Meaning Discovery Using Google." (preprint, arxiv.org/abs/cs.CL/0412098, 2004).
5. M.E.J. Newman, “Power laws, Pareto distributions, and Zipf’s law.” *Contemporary Physics* in press (2004). cond-mat/0412004.
6. M.L. Goldstein, S.A. Morris, and G.G. Yen. “Problems with Fitting to the Power-Law Distribution." (preprint, arxiv.org/abs/cond-mat/0402322, 2004).

TABLE 1. MLE Power-Law Fits to Search Results. All fits pass the KS-test ($\alpha = 0.05$).

Population (Size)	Search	γ	Fitting Range*	Lexicon
Scientists (449)	1	1.82	Top 99	<name>
	2	2.18	" "	<name> physics
	3	2.29	" "	<name> statistical physics
	4	2.69	" "	<name> statistical physics condensed
Scientists (1625)	1	2.02	Top 105	<name>
	2	1.77	Top 240	<name> physics
	3	2.08	Top 150	<name> statistical physics
	4	2.00	Top 210	<name> statistical physics condensed
Aces (1851)	1	2.74	Top 99	<name> WWI
	2	3.62	" "	<name> WWI ace
Actors (778)	1	1.88	Top 120	<name>
	2	2.04	" "	<name> movie
	3	2.10	" "	<name> movie actor
Villains (421)	1	1.57	Top 99	<name>
	2	1.86	" "	<name> villain
	3	2.03	" "	<name> villain evil
Programmers (148)	1	1.88	Top 59	<name>
	2	2.16	" "	<name> programmer
	3	2.03	" "	<name> computer
	4	2.43	" "	<name> computer programmer
Clarkson Students (1533)	1	1.74	Top 119	<name>
	2	1.99	" "	<name> clarkson
	3	2.57	" "	<name> "clarkson university"
Runners (222)	1	1.71	Top 99	<name>
	2	1.92	" "	<name> olympics

* For example, Top 99 means that the fit was applied to only the 99 highest-ranked searches. Note that some sub-samples constitute less than 10 percent of the population and that the tail contains the noisiest, and therefore least reliable, data.

TABLE 2. MLE Exponential Fits to Search Results. All fits pass the KS-test ($\alpha = 0.05$) except for Scientists (1625) Search 4.

Population (Size)	Search	λ^{-1}	K	CV^*	Fitting Range
Scientists (449)	1	1040	0.0877	0.0908	230 - 449
	2	591	0.0663	0.0706	85 - 449
	3	385	0.0650	0.0688	65 - 449
	4	134	0.0602	0.0644	10 - 449
Scientists (1625)	1	468	0.0495	0.0505	910 - 1625
	2	343	0.0385	0.0395	455 - 1625
	3	161	0.0397	0.0416	570 - 1625
	4	75	0.0423	0.0369	280 - 1625

* The critical value that is compared to K. A distribution passes the KS-test when $K < CV$.

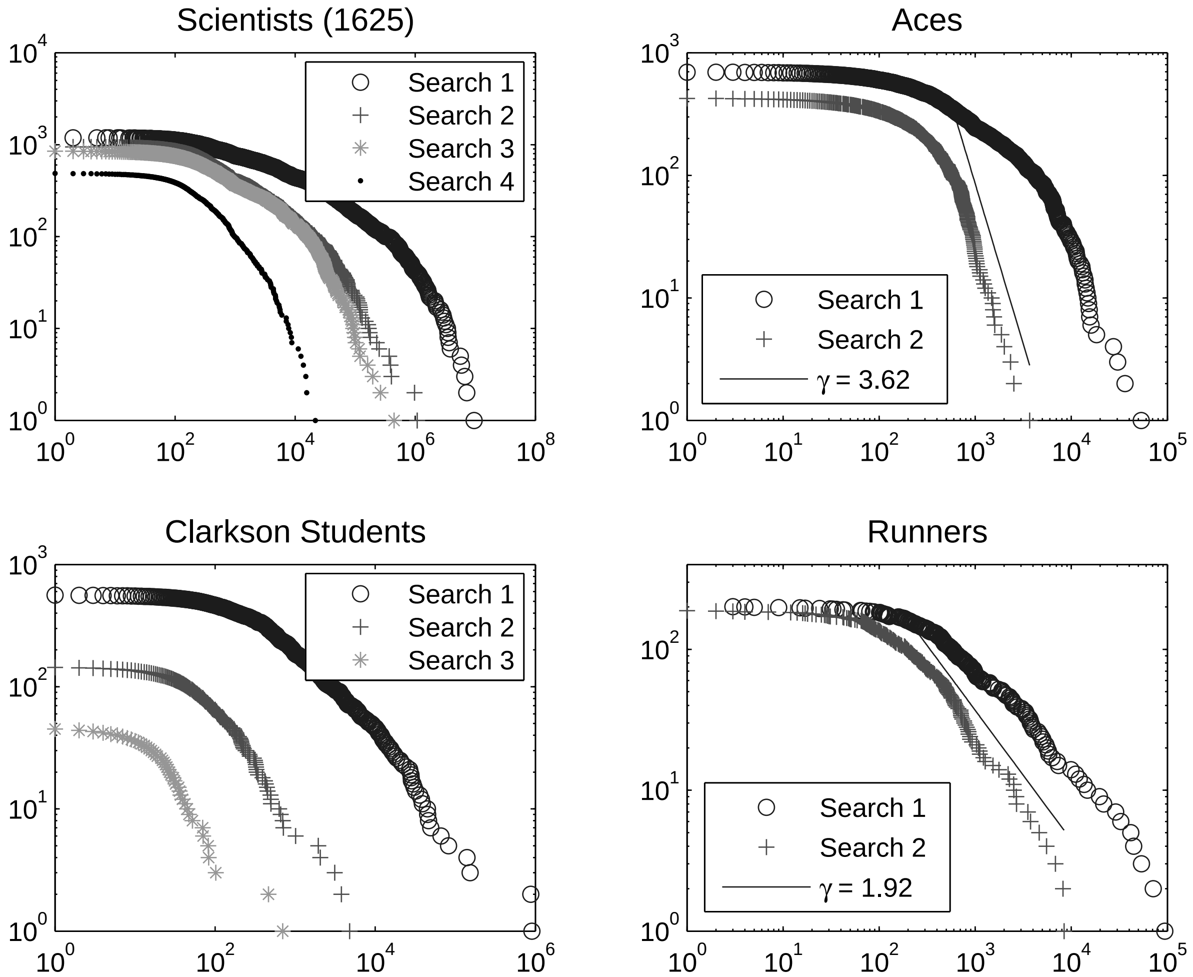

FIGURE 1. Rank / Frequency plots for several populations. The horizontal axis is the number of *Google* hits and the vertical axis is the rank of the (sorted) data points. Note that straight lines (offset for clarity) will have slope $-\gamma+1$.