Robust Methodologies for Modeling Web Click Distributions

Kamal Ali, Yahoo!
701 First Avenue
Sunnyvale, CA USA
kamal3@yahoo.com

Mark Scarr, Yahoo!
701 First Avenue
Sunnyvale, CA USA
mscarr@yahoo-inc.com

ABSTRACT

Metrics such as click counts are vital to online businesses but their measurement has been problematic due to inclusion of high variance robot traffic. We posit that by applying statistical methods more rigorous than have been employed to date that we can build a robust model of the distribution of clicks following which we can set probabilistically sound thresholds to address outliers and robots. Prior research in this domain has used inappropriate statistical methodology to model distributions and current industrial practice eschews this research for conservative ad-hoc click-level thresholds. Prevailing belief is that such distributions are scale-free power law distributions but using more rigorous statistical methods we find the best description of the data is instead provided by a scale-sensitive Zipf-Mandelbrot mixture distribution. Our results are based on ten datasets from various verticals in the Yahoo domain. Since mixture models can overfit the data we take care to use the BIC log-likelihood method which penalizes overly complex models. Using a mixture model in the web activity domain makes sense because there are likely multiple classes of users. In particular, we have noticed that there is a significantly large set of "users" that visit the Yahoo portal exactly once a day. We surmise these may be robots testing internet connectivity by pinging the Yahoo main website.

Backing up our quantitative analysis is graphical analysis in which empirical distributions are plotted against theoretical distributions in log-log space using robust cumulative distribution plots. This methodology has two advantages: plotting in log-log space allows one to visually differentiate the various exponential distributions and secondly, cumulative plots are much more robust to outliers. We plan to use the results of this work for applications for robot removal from web metrics business intelligence systems.

Categories and Subject Descriptors

G.3 [**Mathematics of Computing**]: Probability and Statistics

General Terms

Measurement, Theory

WWW 2007, May 8–12, 2007, Banff, Alberta, Canada.
ACM 978-1-59593-654-7/07/0005.

1. INTRODUCTION

Modern businesses rely on accurate counts of web page-views and clicks to calculate growth rates and market share. Per-user page-views in the millions (per month) and other suspicious statistics lead to the belief that a significant amount of traffic originates from robots. Failure to remove robots can mislead businesses about their growth metrics. Furthermore, we see higher temporal (month to month or day to day) variance from the robot population so failure to remove them also degrades statistical power [20] in comparing fielded systems to their beta counterparts.

In order to remove robots on a more principled basis, we need to have a better characterization of the distribution of click behavior. One method of removing robots is to identify them with outliers and remove outliers. Outlier removal using distributional methods proceeds by fitting a model to the observed distribution and then selecting a tail probability (say 0.1%) to use as a definition of an outlier. Using the model, we can then translate that probability into a statistically founded threshold of clicks and remove all "users" that exceed that threshold. Currently, businesses use very conservative ad-hoc thresholds for robot removal. Knowing the distribution more precisely would allow them to be more aggressive in removing robots and thus produce more accurate and stable metrics for business.

Most previous work in web traffic distribution modeling has been done for network caching and relay applications [7, 3] - little to none has been done for web analytics. The work that has been done did not appear to test a wide variety of distribution families and some authors used continuous distributions [9, 1], which are not appropriate for discrete (click and pageview) distributions. Finally, the work appears not to have used rigorous statistical methodology. For instance, many authors simply plot the observed distribution in log-log space (log of frequency of x versus $log(x)$) and then proceed to fit a straight line, using Pearson's correlation coefficient [16] to measure goodness of fit. This is completely incorrect because typically the points corresponding to low pageview and click counts represent millions of users and other points may only represent a single user. Thus, at the very least, a weighted regression is called for. Even so, outliers can have large impact on regression fitting methods and no indication was given that robust regression methods were used. Instead of regression, we use the maximum likelihood method (MLE) which is much more robust.

In our methodology, we try to fit a large number of distribution families to the data using the robust maximum-likelihood estimation [16] (MLE) method. Then we evalu-

ate the goodness of fit using the (log) likelihood of the data given the fitted distribution. Since some models have more parameters that can overfit the data, we use the Bayesian Information Criterion (BIC, [17]) correction to log-likelihood. Essentially this measure requires that more complex models "pay for their complexity" by providing a better fit to the data. In the case where we have nested models the likelihood-ratio test [16] is also used to see what other distributions are statistically indistinguishable from the winning one - thereby forming a set of winners. Although inappropriate for count data we also use two continuous distributions (Inverse Gaussian [9], log-normal [10]) that have been used in prior work so as to compare their scores to the theoretically correct discrete distributions.

Graphical methods (plotting the data) are a great aid in statistics because in two dimensions the human eye is great at pattern detection. However, in distribution modeling where some points have drastically different weight than others, the plots can be quite misleading. Nevertheless, properties such as curvature can be discerned from plots and this is important for this paper since if the observed distribution has a curved form in log-log space, it favors scale-sensitive models over scale-free power law distributions.

The rest of the paper is organized as follows: section 2 summarizes previous work in modeling distributions for the web and other data sets with various discrete distributions (Zipfian, Poisson, Negative Binomial) and simple mixture models. Section 3 gives density functions for the candidate distributions we use in this paper. Section 4 spells out our approach for parameter estimation and model comparison, section 5 describes the ten datasets and section 6 presents results for the various fitted distributions.

2. PREVIOUS WORK

Power-laws, Zipf distributions and Pareto distributions have become somthing of a fad recently, being very popular in explaining all manner of data (city sizes, galaxy sizes [10], words: [23], incomes [21]). We begin by carefully distinguishing these. The term "power law" is inappropriately general for our application since it goes well beyond distributions to describe any functional relationship between y and x where $y = ax^k$, hence we will not use this term. Zipf distributions are *discrete* distributions over a finite set $(1...N)$ with probability mass function $f(x; s, N) = \Omega(N)\frac{1}{x^s}$ where $\Omega(N) = 1/\sum_{i=1}^{N}\frac{1}{i^s}$ is a normalizing constant. For $s > 1$ this distribution is normalizable even if $N = \infty$ and becomes the Zeta distribution [19]. Some authors restrict the term "Zipf distribution" for the case $s = 1$ and call other cases "Zipf-like" or "Zipfian". Zipf distributions ($s = 1$) such as word distributions, have the scale-free property:

> Scale-free: A distribution f is scale free if *for all values of* x, the probability of $2x$ is half the probability of x. If a distribution is not scale-free, we will term it as *scale-sensitive*.

Applied to word rankings this means that for all ranks, the probability of seeing a word at that rank is twice the probability of seeing a word with twice the rank (Zipf's law [23]). Pareto [21] distributions are *continuous* analogues of the Zipf, with density given by $f(x; k) = k\frac{1}{x^{k+1}}$.

Previous work can be classified according to the types of distributions that were tried and also more subtly, the meaning of the x axis: in some work, the x axis refers to the value of a random variate (e.g. number of clicks) whereas in others it refers to a *rank* of that variable. Table 1 summarizes previous work with respect to this classification. Laherrere and Sornette [10] take yet another approach, modeling rank as a function of the random variate. The Inverse-Gaussian [9], Log-Normal [10] and Weibull [10] distributions are continuous distributions yet they are being used to model discrete data.

The paper of Huberman *et al.* [9] claims to have discovered a "strong law of surfing": that the distribution of clicks is distributed as Inverse Gaussian (IG). However, we believe this claim to be too strong. In particular, for our datasets, we have found other distributions to offer statistically significantly better fit to the data than the Inverse Gaussian. We doubt there is a single distribution that will fit all kinds of web surfing let alone constitute a "law of surfing". Huberman also claims the Inverse Gaussian has theoretical motivation and argues that the utility of a web surfer mirrors that of economic options whose prices are known to follow an Inverse-Gaussian distribution. They also show that the page-view distribution of URLs is fitted by an Inverse-Gaussian. Their approach seems not to have tried a lot of distributions, let alone discrete ones - it seems only that they have considered the Inverse Gaussian because of its theoretical underpinning and then proceeded to see if it gives a good enough fit to the data.

Huberman's work is the only other we are aware of that points out the curvature of the distribution in log-log space thus deprecating the scale-free power law distributions. Previous authors did not rule out better fits by scale-sensitive curved forms in log-log space: they only demonstrated that they got a good-enough fit by a line in log-log space and thus concluded the distribution must be power-law.

Laherrere and Sornette justify their choice of Weibull because *"tails of pdfs of products of a finite number of random variables is generically a stretched exponential"* [6]. However, it appears to us that there is a serious problem with their methodology. They evalute goodness of fit in log-log space using Pearson's correlation coefficient which is incorrect since the plotted points with lower rank represent many more points than those with higher ranks.

Finally, a note on previous work in mixture and zero-adjusted models. User web behavior can be naturally partitioned based on the presence or absence of a click, in other words zero click *vs.* non-zero click behavior. The zero click class might be thought of as primarily containing robot traffic, and the non-zero click class principally "human" traffic. At the very least, if we see that a model fits the data well except at the zero point, we can conclude that the excess of users with zero clicks may be due to a secondary phenomenon: that of robots. This is the approach of mixture models in which the zero class is modeled by a (single-valued degenerate) distribution and the positive clicks are modeled by a discrete positive-valued distribution. Such models are called zero-altered or zero-inflated and have been used in numerous applications to model data with an excess of zeroes or where the zero-class of the underlying process has special meaning (number of defects [11], number of dental cavities [14], and number of car crashes [12]). In all of these domains, zero has a special meaning in that it is usually the default scenario indicating lack of an accident or problem.

Entity	Measure	Best-fit distribution	X-axis	Y-axis	Author
Word	Frequency	Zipf $k = 1$	Rank	Frequency	Zipf [23]
City	Population	Weibull	Population	Rank	Laherrere, Sornette [10]
URL	Page-views	Zipf-like, various k	Page-views	Frequency	Breslau [3]
Search Queries	Frequency	Weibull	Rank	Frequency	Abdullah [1]
URL	Page-views	Inverse Gaussian	Page-views	Frequency	Huberman *et al.* [9]
Session	Clicks	Inverse Gaussian	Clicks	Frequency	Huberman *et al.* [9]
Session	Clicks	ZAZM: Zero-altered Zipf-Mandelbrot	Clicks	Frequency	Scarr, Ali (this paper)
User	Page-views	ZAZM	Page-views	Frequency	Scarr, Ali (this paper)

Table 1: Prior work in distribution modeling can be partitioned into those modeling the random variate x versus those modeling the rank of x. The web papers differ subtly in that some are modeling clicks per user or clicks per session or searches per query.

3. DISTRIBUTION THEORY

Since this paper is concerned with modeling count data, in the form of user clicks, we limit our attention to discrete distributions; in particular the Negative Binomial [2], Zipf [23], Zipf-Mandelbrot [13] (of which Zipf is a special case), Logarithmic Series [22] and Yule-Simon [18]. The Poisson distribution which best fits data when its mean and variance are equal is not used due to the fact our click data (table 2) have variance-to-mean ratios far in excess of 1. The Inverse Gaussian [9], Weibull [10] and Log-Normal [10] distributions, which are continuous, are only included for comparison as they have been used to model web click behavior in the past.

In addition a class of simple mixture models are also considered, motivated by the fact that zero-click users may behave differently to non-zero click users. A number of "users" are actually robots that ping the Yahoo site every day to test if they are connected to the internet. This is a different generative phenomenon than that of regular human search and hence justifies using a mixture model. Zero-altered or zero-inflated models [8, 11, 14] for count data are mixture models that assume with probability p the only possible observation is 0 and with probability $1-p$ a discrete random variable is observed.

The zero-altered model mixes a degenerate distribution with point mass of 1 at zero with a truncated count distribution for example truncated Poisson, truncated Negative Binomial or any discrete distribution bounded below by 1. On the other hand the zero-inflated model accounts for some of the zeros through the non-degenerate distribution (*e.g.* Poisson, Negative Binomial, *etc...*) and some through the degenerate (zero) distribution.

We now present formulae for the probability mass and density functions of the various distributions discussed above, by considering a random variable X.

Zero-Altered (ZA):

The probability mass or density function for a zero-altered mixture model is defined as:

$$f_X(x;p,\boldsymbol{\theta}) = \begin{cases} 0 & x < 0 \\ p & x = 0 \\ (1-p)f_Y(x;\boldsymbol{\theta}) & x > 0 \end{cases} \quad (1)$$

where $0 \le p \le 1$, $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$ is a vector of parameters and $f_Y(x;\boldsymbol{\theta})$ is any valid probability distribution on $[1, \infty)$. From the above, the random variable X takes the value zero with probability p and values greater than zero with probabilities $(1-p)f_Y(x;\boldsymbol{\theta})$.

Zero-Inflated (ZI):

The probability mass or density function for a zero-inflated mixture model is defined as:

$$f_X(x;p,\boldsymbol{\theta}) = \begin{cases} 0 & x < 0 \\ p + (1-p)f_Y(0;\boldsymbol{\theta}) & x = 0 \\ (1-p)f_Y(x;\boldsymbol{\theta}) & x > 0 \end{cases} \quad (2)$$

where $0 \le p \le 1$, $\boldsymbol{\theta}$ is a vector of parameters and $f_Y(x;\boldsymbol{\theta})$ is any valid probability distribution on $[0, \infty)$.

Negative Binomial (NB) (discrete): $X \sim NB(k, \mu)$, the probability mass function is:

$$f(x;k,\mu) = \frac{\Gamma(x+k)}{\Gamma(k)\Gamma(x+1)}\left(\frac{k}{\mu+k}\right)^k\left(1-\frac{k}{\mu+k}\right)^x,\ x \ge 0 \quad (3)$$

where $k > 0$ and $\mu \ge 0$. To fit a 'zero-altered" Negative Binomial model (ZANB) the probability mass function of a zero-truncated Negative Binomial distribution is required, this is defined as:

$$f_0(x;k,\mu) = \frac{f(x;k,\mu)}{1 - f(0;k,\mu)},\ x > 0 \quad (4)$$

From equations (1, 3, 4) the Zero-Altered Negative Binomial $ZANB(p, k, \mu)$ probability mass function is:

$$f(x;p,k,\mu) = \begin{cases} 0 & x < 0 \\ p & x = 0 \\ (1-p)\frac{\Gamma(x+k)(\frac{k}{\mu+k})^k(1-\frac{k}{\mu+k})^x}{\Gamma(k)\Gamma(x+1)(1-(\frac{k}{\mu+k})^k)} & x > 0 \end{cases} \quad (5)$$

Zipf (Z) (discrete): $X \sim Z(s, N)$, the probability mass function is:

$$f(x;s,N) = \Omega(N)\frac{1}{x^s},\ x \in (0, N] \quad (6)$$

where $s > 0$, N is finite and $\Omega(N) = 1/\sum_{i=1}^{N}\frac{1}{i^s}$ is a normalising constant.

Zipf-Mandelbrot (ZM) (discrete): $X \sim ZM(s, q, N)$, the probability mass function is:

$$f(x;s,q,N) = \Omega(q,N)\frac{1}{(x+q)^s},\ x \in [0, N] \quad (7)$$

where $s > 0$, $q \ge 0$, N is finite and $\Omega(q,N) = 1/\sum_{i=1}^{N}\frac{1}{(i+q)^s}$ is a normalising constant. Clearly, setting $q = 0$ yields the Zipf distribution in equation (6) so non-zero q indicates curvature in log-log space.

Logarithmic-Series (LS) (continuous): $X \sim LS(k)$, the probability mass function is:

$$f(x;k) = \frac{-1}{\log(1-k)}\frac{k^x}{x},\ x>0,\ k\in(0,1) \quad (8)$$

Yule-Simon (YS) (discrete): $X \sim YS(k)$, the probability mass function is:

$$f(x;\rho) = \rho B(x,\rho+1),\ x>0,\ \rho>0 \quad (9)$$

where $B(,)$ is the Beta function.

Inverse-Gaussian (IG) (continuous): $X \sim IG(\lambda,\mu)$, the probability density function is:

$$f(x;\lambda,\mu) = \sqrt{\frac{\lambda}{2\pi x^3}}\exp(\frac{-\lambda(x-\mu)^2}{2\mu^2 x}),\ x\geq 0 \quad (10)$$

with , $\lambda,\mu>0$.

Log-Normal (LN) (continuous): $X \sim LN(\mu,\sigma)$, the probability density function is:

$$f(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}}\exp\left\{-\frac{(\log(x)-\mu)^2}{2\sigma^2}\right\},\ x\in\mathbb{R} \quad (11)$$

with $\mu\in\mathbb{R}$ and $\sigma>0$. The Log-Normal distribution is the probability distribution of any random variable whose logarithm possesses a Normal or Gaussian distribution. In other words if $Z\sim N(\mu,\sigma^2)$ is a Normally distributed random variable then $X=e^Z$ has a Log-Normal distribution.

Weibull (W) (continuous): $X \sim W(\alpha,\beta)$, the probability density function is:

$$f(x;\alpha,\beta) = \frac{\alpha}{\beta}(\frac{x}{\beta})^{\alpha-1}\exp(-(\frac{x}{\beta})^\alpha),\ x\geq 0,\ \alpha,\beta>0 \quad (12)$$

Zero-altered or zero-inflated models can be constructed using equations (1, 2) and any of the probability mass or density functions above. Since we are particularly interested in the zero class, zero-altered models only are used for those distributions where $x>0$, such as Zipf, Logarithmic series *etc...*

4. METHODOLOGY

We now describe the method used to fit the distributions proposed in section 3 to our observed data. In addition we also discuss how to compare different fitted models. Maximum likelihood estimation *e.g.* [16, 5], a standard statistical modeling technique, is used to fit the models. It possesses a number of desirable properties and is a widely used parameter estimation tool. Once various models have been fitted they can be compared using the Bayesian Information Criterion (BIC) [17] and graphical methods such as Cumulative Distribution plots.

4.1 Maximum Likelihood Estimation

Suppose we have an independant and identically distributed (*i.i.d.*) sample $\boldsymbol{X}=(X_1,X_2,\ldots,X_n)$ with joint probability density or mass function $f(\boldsymbol{x}|\boldsymbol{\theta})$, where the data $\boldsymbol{x}=(x_1,x_2,\ldots,x_n)$ and the parameter vector $\boldsymbol{\theta}=(\theta_1,\theta_2,\ldots,\theta_k)$. Given observed values $X_i=x_i$, $i=1,\ldots,n$ the *likelihood* is:

$$L(\boldsymbol{\theta}|\boldsymbol{x}) = \prod_{i=1}^{n} f(x_i|\boldsymbol{\theta}) \quad (13)$$

In the case of discrete data the *likelihood* measures the probability of observing the given data as a function of $\boldsymbol{\theta}$. The maximum likelihood estimate (MLE) $\hat{\boldsymbol{\theta}}$, is the value of $\boldsymbol{\theta}$ that maximises the likelihood *i.e.* makes the observed data "most likely". In practice it is usually easier to equivalently maximize the *log-likelihood*:

$$l(\boldsymbol{\theta}|\boldsymbol{x}) = \sum_{i=1}^{n}\log\{f(x_i|\boldsymbol{\theta})\} \quad (14)$$

As an example, consider an *i.i.d.* sample from a Poisson distribution with parameter λ, then:

$$f(x_i|\lambda) = \frac{\lambda^{x_i}e^{-\lambda}}{x_i!}\ \forall i=1,\ldots,n \quad (15)$$

From equation (13) the *likelihood* is:

$$L(\lambda|\boldsymbol{x}) = \prod_{i=1}^{n}\frac{\lambda^{x_i}e^{-\lambda}}{x_i!} \quad (16)$$

Since it is easier to work with the *log-likelihood*, from equation (14) we have:

$$l(\lambda|\boldsymbol{x}) = \log(\lambda)\sum_{i=1}^{n}x_i - n\lambda - \sum_{i=1}^{n}\log(x_i!) \quad (17)$$

Differentiating the log-likelihood in equation (17) with respect to λ and setting to zero gives:

$$l'(\lambda) = \frac{1}{\lambda}\sum_{i=1}^{n}x_i - n = 0 \quad (18)$$

Expressing equation (18) in terms of λ gives the familiar MLE $\hat{\lambda}=\overline{x}$.

In the above example, we maximized the *log-likelihood* with respect to a single parameter λ. More generally, the model may contain several parameters, in which case we compute partial derivatives and set each in turn to zero. Depending on the particular distribution used in the *likelihood*, a closed form solution may or may not exist. For the Poisson distribution the MLE of λ has an algebraic solution that is the sample mean of the data. In cases where there is no closed form solution an iterative method is employed using a modification of the Broyden, Fletcher, Goldfarb and Shanno quasi-Newton algorithm [4] within the statistical software package R [15].

Although not discussed in detail here it is also possible to compute variances based on the Fisher Information [16] and hence confidence intervals for the maximum likelihood estimates described above.

4.2 Comparing Models

The *log-likelihood* in equation (14) can be computed for the various parametric models of interest and used as a basis for model comparison. This makes sense as the model that has the largest *log-likelihood* is considered to be the most "likely" given the observed data. However since we are investigating models with differing numbers of parameters, rather than comparing *log-likelihoods* directly we use the Bayesian Information Criterion (BIC) [17]. The BIC penalizes models with more parameters so that to "win" in

Property	Session-level data	User-level data
X-axis	Clicks/session	Clicks/user
Number records	1,411,298	22,913,854
Mean	2.78	2.29
Median	1	0
Variance	22.94798	75.84359
Variance to Mean Ratio	8.24	34.03
Minimum #Clicks	0	0
Maximum #Clicks	499	5,428

Table 2: Summary of our data sets.

a BIC-sense, the extra parameter needs to justify its addition with a commensurate increase in log-likelihood.

The Bayesian Information Criterion (BIC) [17] is defined as:

$$BIC = -2l(\boldsymbol{\theta}|\boldsymbol{x}) + k\log(n) \tag{19}$$

where $l(\boldsymbol{\theta}|\boldsymbol{x})$ is the maximized *log-likelihood*, n is the number of observations and $k = |\boldsymbol{\theta}|$ is the number of model parameters. We wish to minimize the BIC with respect to the estimated model parameters. As can be seen from equation (19) the BIC attaches a penalty to the addition of extra parameters, forcing it to prefer lower order models especially for large n.

4.3 Graphical Methods

In the related work section we argued why linear fitting in log-log space where one axis denotes frequency is poor methodology. In this section we describe the benefits and pitfalls of the Cumulative-Probability (Cum-Prob) plot.

Cum-Prob plots: A Cumulative-Probability (cum-prob) plot is related to a QQ plot. To compute the x coordinate of the i'th point, the area under the histogram (*e.g.* of clicks) up to and including the i'th distinct value is divided by the full area under the histogram. In other words, x values are empirical CDF (Cumulative Distribution Function) values. The y values are obtained by first doing a MLE best-fit but this time referring to the theoretical CDF. Let c_i be the i'th distinct click value over the set $\mathcal{C}$ of all possible click values, then CDF is a function taking random click values to the zero-one set: $CDF : \mathcal{C} \rightarrow [0,1]$. Thus both x and y values are normed to the [0,1] interval.

5. DESCRIPTION OF DATASETS

We show results for our two primary datasets: their basic parameters are in table 2. The first dataset is for our main web search engine: a random sample of sessions collected from a week's worth of data. A session is terminated by the standard 30 minutes of inactivity. The second dataset is a random sample of per *user* clicks, integrated over one month of activity for that user on another Yahoo website. Thus dataset 2 consists of a mixture of sessions.

Figures 1 and 2 show histograms of the data. We have zoomed in to the top 10 click values since the rest of the histogram has very low values. Note that for the per-session web search dataset, one-click occurs more frequently than any other value and in particular, more than zero clicks. A zero-click session is possible since a session can have pageviews but zero clicks. If one were to use an exponential discrete model, one might expect the zero probability to be higher

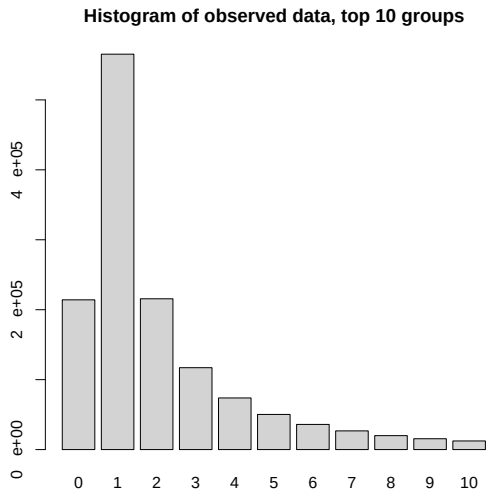

Figure 1: Sessionized data: frequency of web-search clicks per-session (top 10 values).

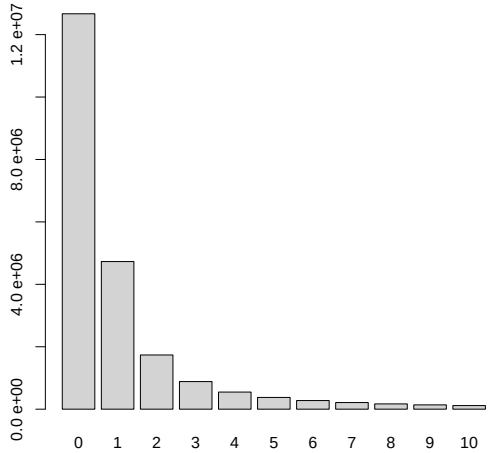

Figure 2: User-level data: frequency of Yahoo website clicks per-user (top 10 values).

than the probability for one click. Thus the data indicate immediately the need to try for a mixture model with a special component at zero - so-called Zero-Adjusted mixture models. For the per-user month-level dataset (figure 2), zero clicks are more common than one click. Note that the user-level data is from another website (not web search) so it is not as if multiple sessions from figure 1 are being included in figure 2.

Ten other datasets from various other Yahoo verticals[1] were also analyzed to see if our results are generalizable. In order to get verticals that span the gamut of user behavior

[1]Verticals are websites such as mail.yahoo, travel.yahoo etc.

we sorted all the websites at Yahoo by volume of traffic and then randomly picked a website in each decile of the sorted list.

6. RESULTS

We begin by examining both datasets in log-log space (figures 3 and 4). Log-log space brings out differences in the various exponential distribution families that are just not apparent in usual histograms because all such distributions have long tails. The other advantage of log-log space is that curvature can be easily visually spotted and this usually indicates the data is not from a scale-free distribution (scale-free has linear form in log-log space).

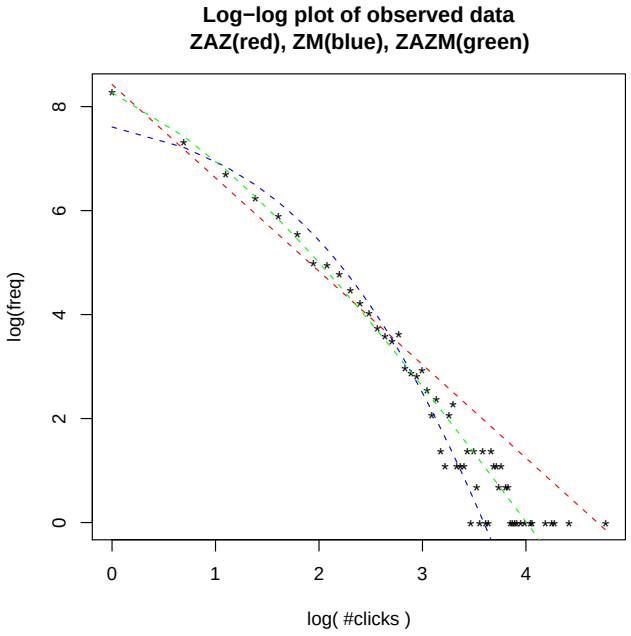

Figure 3: Per-session clicks for search engine results exhibit curvature in log-log space. Observations are dark asterisks; fitted models are curves or lines.

The figures show the distribution of data, a non-mixture Zipf-Mandelbrot, a mixture Zipf-Mandelbrot and a mixture Zipf - we cannot fit a non-mixture Zipf since it is undefined for $x = 0$. The figures also show a "linear fit" regression line. This line is obtained by doing linear regression on the *plotted* points. Note how it provides a poor fit for points near $\log(\text{clicks}) = 0$ which are just the points that represent millions of users (or sessions). This is the fundamental problem with doing line-fitting of the plotted points: all points are equally treated whereas in reality some points correspond to millions of users and others may correspond to a single user.

Although visual examination of the data seems to indicate curvature, a more rigorous confirmation would be to fit models assessed by BIC values (figures 5 and 6). If the winning model has, by theory, a curved form in log-log space, then it would confirm the visual observation. The figures show the smallest (best fitting) BIC values consistently come from the zero-altered mixture models, with zero-altered Zipf-Mandelbrot (ZAZM) being the best fit for our data.

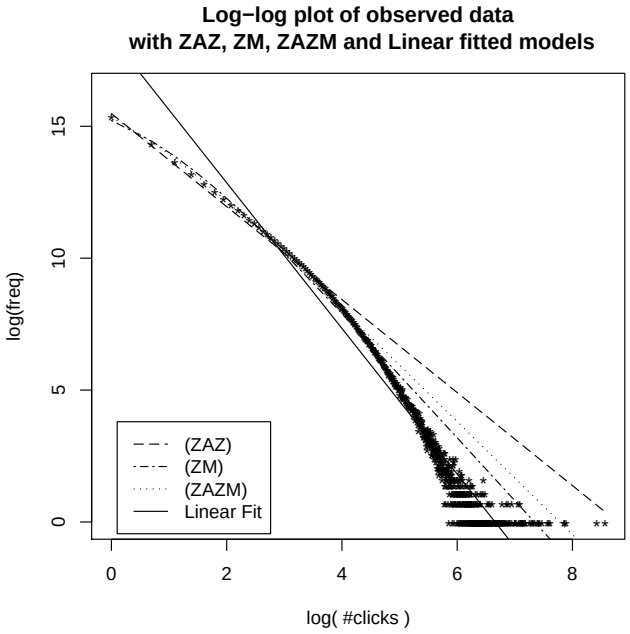

Figure 4: Per-user clicks (integrated over 1 month) for another large Yahoo website exhibit curvature in log-log space.

The main three results of our paper are supported by these results:

- **Curvature:** In log-log space our data is curved as indicated by the fact that the best fitting distribution, Zipf-Mandelbrot, by theory has a curved form in log-log space. This is also visually confirmed. Another way to quantify the amount of curvature is to examine the confidence interval around the MLE (maximum likelihood estimate) of the q parameter in the Zipf-Mandelbrot distribution. For web search, we obtained $q = 2.0 \pm 0.01$ and for the per-user dataset we obtained $q = 0.9 \pm 0.002$. Thus in both cases, $q = 0$ (which would indicate no curvature) is emphatically excluded. Curved forms in log-log space do not have the scale-free property of pure Zipfian (power-law) distributions - they have a natural, distinguished scale. (*e.g.* [10]).
- **ZAZM:** The particular model form with best BIC fit is the ZAZM (Zero-Adjusted Zipf-Mandelbrot) model for both datasets.
- **Zero-Altered Mixture:** Zero-altered mixture models do better than their non-mixture counterparts, irrespective of the distribution type. This is consistent with our hypothesis that a different generative component (robots) are major contributors to the zero-click observations.

6.1 Results on ten other websites

Next, we examine the ten other datasets (table 3). First note the (Log-Likelihood) LL column: this shows that the best fit for each website was obtained by the Zero-Altered Zipf-Mandelbrot mixture - not any other mixture. Second,

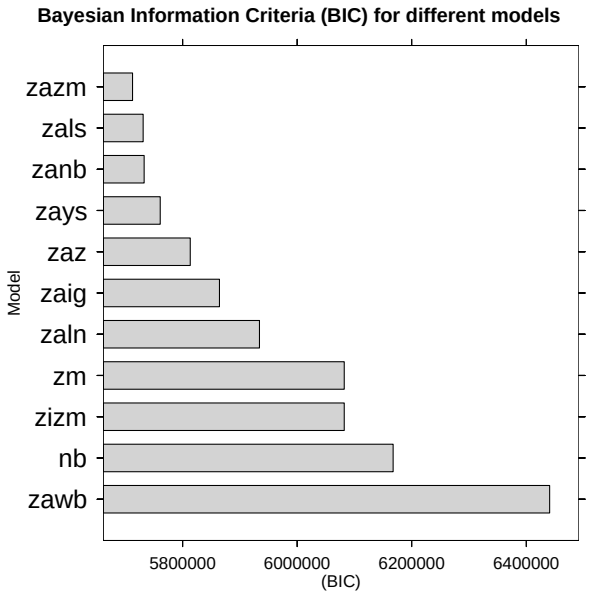

Figure 5: Sessionized web data BIC fit values (low values are better): Zero-alteration makes a big difference and within that class, the Zipf-Mandelbrot is the best.

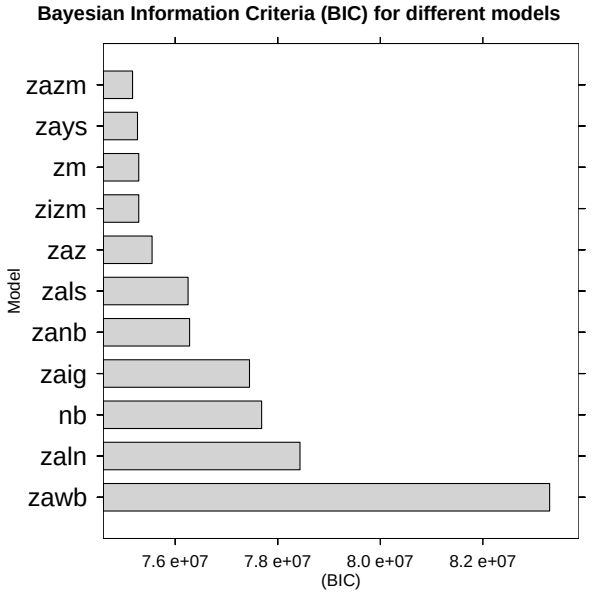

Figure 6: User-level BIC values (lower is better): Zero-altered models (ZA..) do better than zero-inflated (ZI..) or unadjusted (e.g. ZM). Within the zero-adjusted set, Zipf-Mandelbrot does the best.

note that for many of these minor websites, BIC indicates the addition of a parameter in going from Zipf to Zipf-Mandelbrot is not "worth it". In other words, some of these websites *do* have a linear form in log-log space. However, even in these cases, the best *fit* per se (measured by LL) is provided by ZAZM. And since LL measures pure fit (BIC is an adjusted measure of fit) one would never lose (fitting-wise) by using ZAZM (this is clear when one recalls that the ZAZ model is just a special case of the ZAZM model).

Decile	Best parsimonious fit (BIC)	Best fit (LL)	Sample size
1	ZAZM	ZAZM	22,913,854
2	ZAZ	ZAZM	239,520
3	ZAZM	ZAZM	160,137
4	ZAZ	ZAZM	21,033
6	ZAZ	ZAZM	6,461
7	ZAZ	ZAZM	6,724
9	ZAZ	ZAZM	294

Table 3: Best fitting models for Yahoo websites chosen to span all Yahoo websites in terms of user volume. Even though deciles 5, 8 and 10 had plenty of data, they had too few *distinct* values for number of clicks to allow fitting of parameterized models.

6.2 Graphical Results

Having established that, at least for our datasets, the Zipf-Mandelbrot (and its Zero-Altered counterpart) offer a very good fit[2], we now examine supporting graphical methods: the Quantile-Quantile (QQ) plot and Cumulative-Probability (Cum-Prob) plot.

Figure 7 shows the cum-prob plot for the winning ZAZM model on the web per-session data. The cumulative distribution plots are less vulnerable to outliers because they plot areas rather than values of the random variates themselves. An outlier has a large value but typically does not constitute a large percentage of the overall mass (clicks summed across all users). Hence its effect is reduced in the cum-prob plot.

We can re-examine graphically our comparison between the best BIC scoring distribution (ZAZM - Figure 7) and one that was not competitive (Inverse Gaussian - Figure 8). One can clearly see the ZAZM provides a good fit between theory and observation whereas the Inverse Gaussian touted by [9] as providing a strong law of surfing does not hold for Yahoo web-search data.

Figures 9 through 14 provide graphical confirmation of the LL/BIC results. ZAZM is the best fit with ZAZ being the runner-up and the other doing much less well. The inappropriate continuous distributions (Inverse Gaussian, Weibull and Log-Normal) are included here only because they have been used in prior literature [9, 10, 1], which did not investigate Zipf-Mandelbrot or mixtures. Negative-Binomial (NB) is included because clicks are discrete counts and NB is often used for over-dispersed (variance-to-mean ratio > 1) data. Log-Normal is shown because it is the first distribution people think of when they think of skewed data (and it also has a curved form in log-log space). However, it is not discrete and it did not perform nearly as well as the Zipf-Mandelbrot.

[2]Note we can never say for sure that the data *originate* from a particular form of distribution, only that of the distributions we have tried, such and such a distribution provides the best fit.

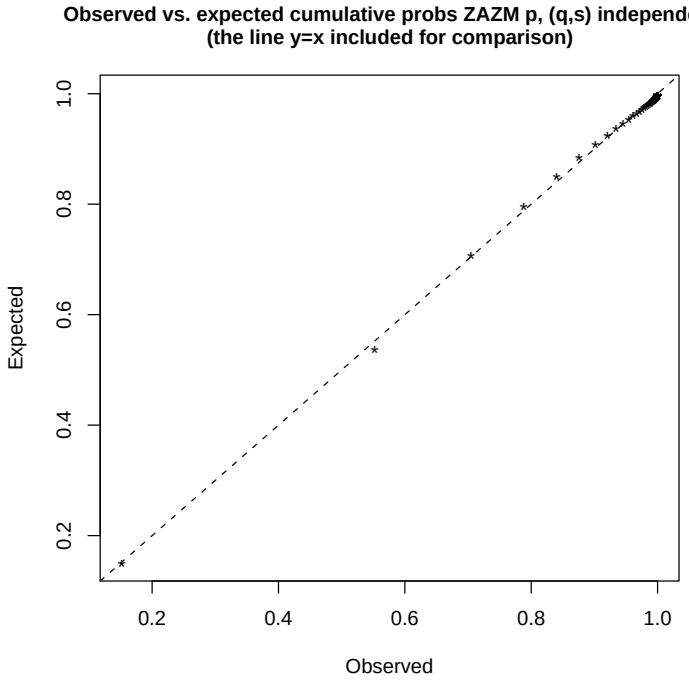

Figure 7: Web session data: Cumulative distribution plot. Winning model: Zero-Altered Zipf-Mandelbrot.

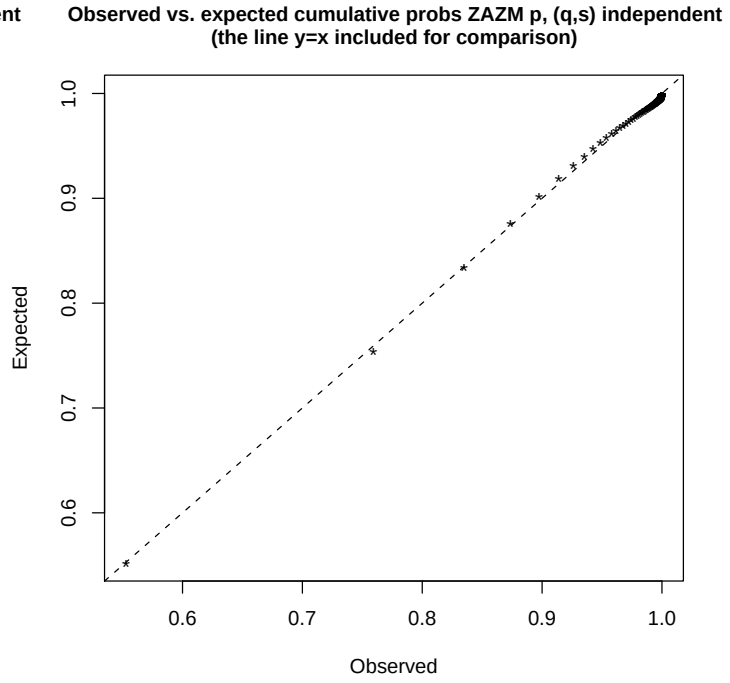

Figure 9: Second dataset: Cumulative distribution plot. Winning model: Zero-Altered Zipf-Mandelbrot.

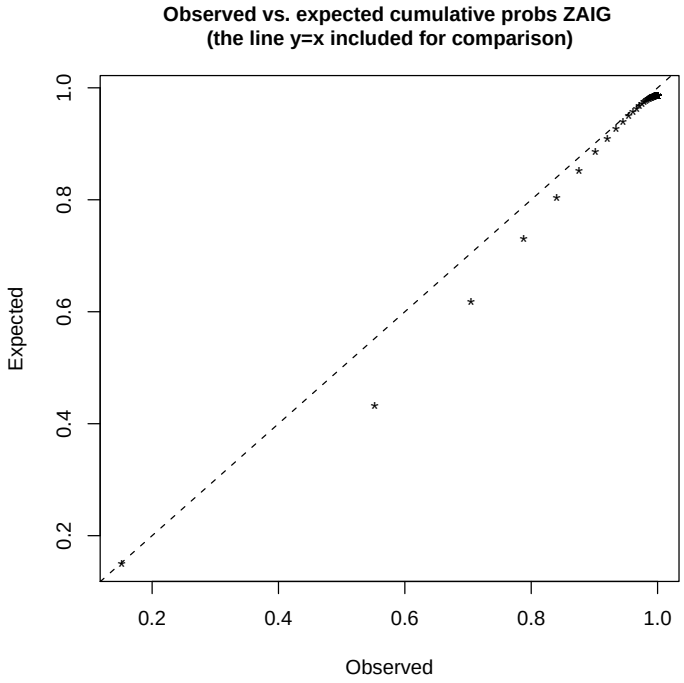

Figure 8: Web session data: Cumulative distribution plot. Losing model from strong law of surfing: Zero-Altered Inverse Gaussian.

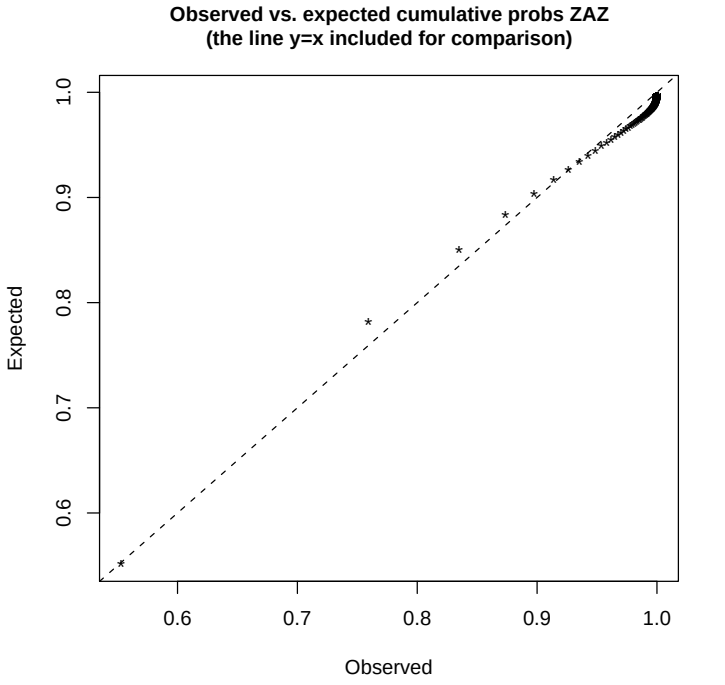

Figure 10: Second dataset: Drop the Mandelbrot correction (Zero-Altered Zipf).

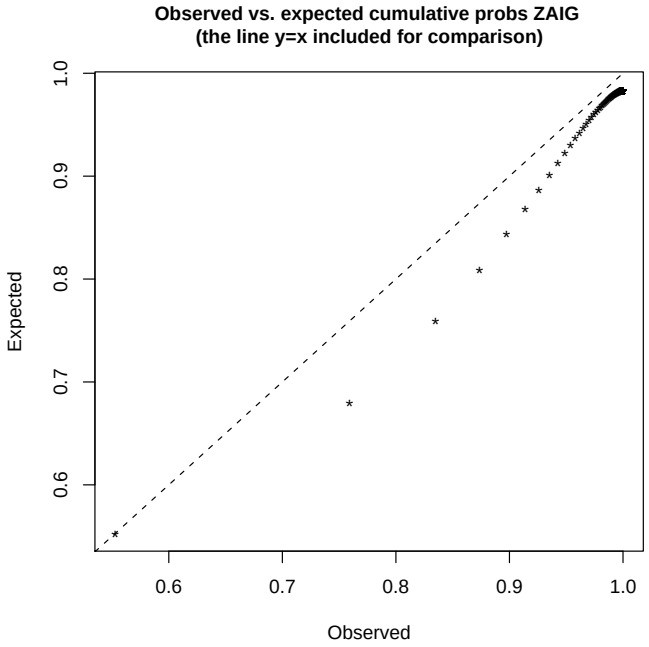

Figure 11: Second dataset: A losing model: inappropriate continuous Zero-Altered Inverse Gaussian. Inverse Gaussian was used in Huberman's strong law of surfing.

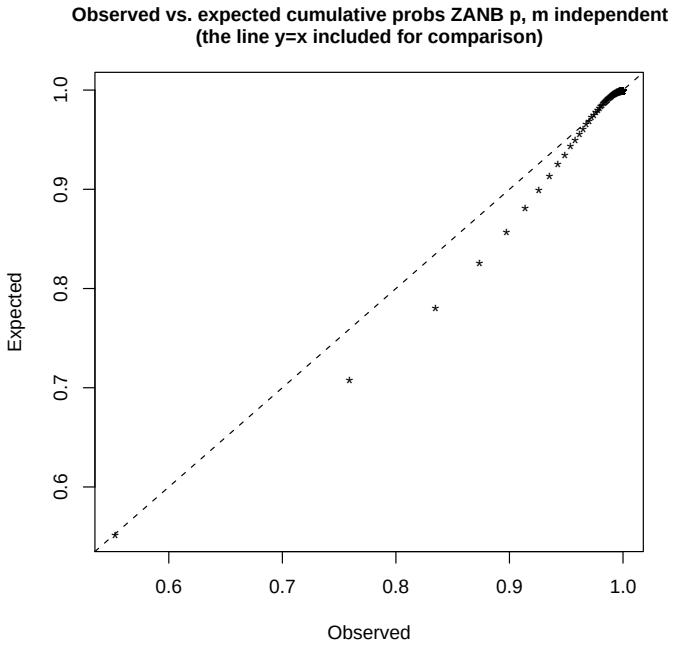

Figure 12: Second dataset: A losing discrete model with over-dispersion (Zero-Altered Negative Binomial). Negative-Binomial is a reasonable guess since it is discrete and has long tail.

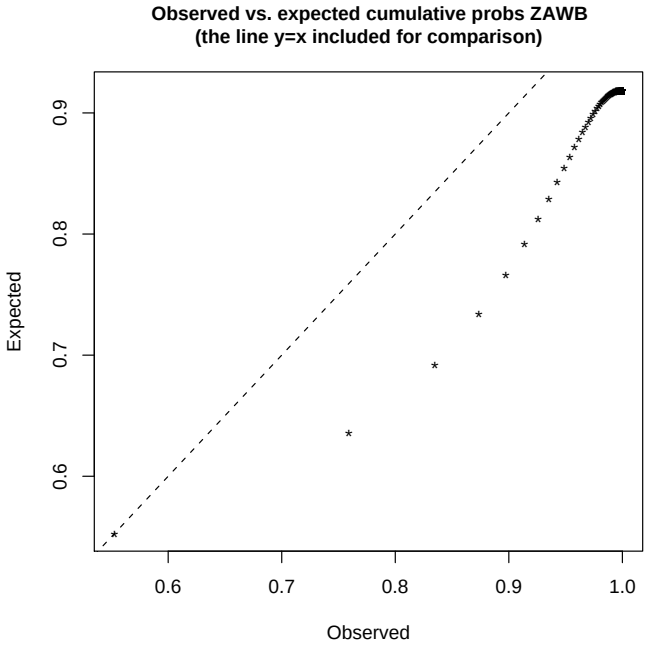

Figure 13: Second dataset: A losing continuous model: Zero-Altered Weibull. Weibull was used by Laherrere *et al.*.

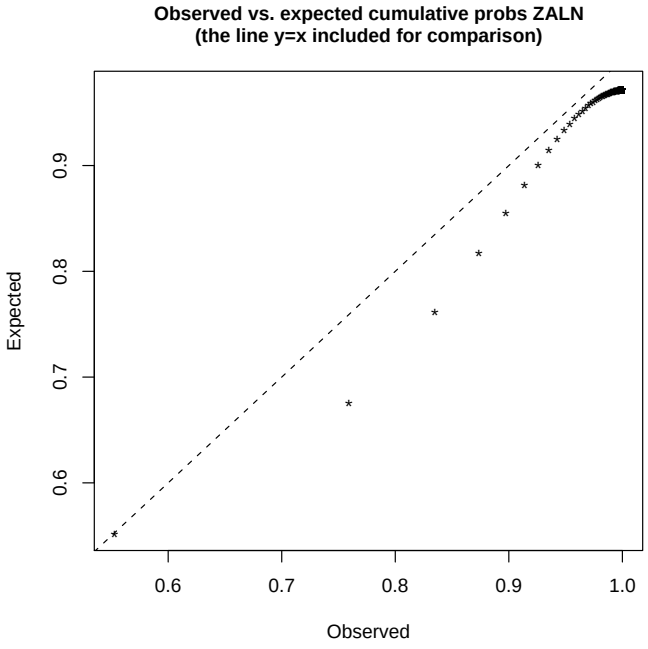

Figure 14: Second dataset: Another losing continuous model: Zero-Altered Log-normal. Log-normal is one of the simplest models that has curved form in log-log space but it is continuous.

7. CONCLUSIONS

Prevailing wisdom is that the distribution of web clicks and pageviews follows a scale-free power law distribution. However, we have found that a statistically significantly better description of the data is the *scale-sensitive* Zipf-Mandelbrot distribution and that mixtures thereof further enhances the fit. Previous analyses have three disadvantages: they have used a small set of candidate distributions, analyzed out-of-date user web behavior (circa 1998) and used questionable statistical methodologies. Although we cannot preclude that a better fitting distribution may not one day be found, we can say for sure that the *scale-sensitive* Zipf-Mandelbrot distribution provides a statistically significantly stronger fit to the data than the scale-free power-law or Zipf on a variety of verticals from the Yahoo domain. The distribution has a definite curved form in log-log space which in turn indicates it is not scale free.

Secondly, we have shown that better results are obtainable using a mixture model which treats the zero-class as special. This is warranted because the generative process of zero clicks might contain a significant proportion of robot "users" and thus would be different than the generative process for non-zero clicks (containing mostly human users). Since we have compared zero-adjusted mixture models to non-mixture models we have taken care to use the BIC log-likelihood scoring method since it makes some adjustments for varying complexity of the models.

Finally, we have argued that the practice of fitting plotted points in log-log space is incorrect methodology and is sensitive to outliers. We instead propose using the Cumulative-Probability plots which plot empirical cumulative distributions against theoretical cumulative distributions. We plan to use the thresholds resulting from these methods to set probabilistically founded threshold levels for removing outliers and robots and thus to enjoy more stable and accurate metrics.

8. REFERENCES

[1] G. Abdulla. *Analysis and Modeling of World Wide Web Traffic*. PhD thesis, Virginia Tech, 1998.

[2] A. Agresti. *Categorical Data Analysis*. Wiley Series in Probability and Statistics, 2002.

[3] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker. Web caching and zipf-like distributions: Evidence and implications. In *INFOCOM (1)*, pages 126–134, 1999.

[4] R. H. Byrd, J. Nocedal, and C. Zhu. A limited memory algorithm for bound constrained optimization. *Journal of Scientific Computing (SIAM)*, 16:1190–1208, 1995.

[5] G. Casella and R. L. Berger. *Statistical Inference*. Duxbury Press, 1990.

[6] U. Frisch and D. Sornette. Extreme deviation and applications. *J. Phys. I France 7*, 7:1155–1171, 1997.

[7] S. Glassman. A caching relay for the World Wide Web. *Computer Networks and ISDN Systems*, 27(2):165–173, 1994.

[8] D. C. Heilbron. Zero-altered and other regression models for count data with added zeroes. *Biometrics*, 36:531–547, 1994.

[9] B. A. Huberman, P. L. T. Pirolli, J. E. Pitkow, and R. M. Lukose. Strong regularities in world wide web surfing. *Science*, 280:95–97, 1998.

[10] J. Laherrere and D. Sornette. Stretched exponential distributions in nature and economy: "fat tails" with characteristic scales. *The European Physical Journal B*, 2:525, 1998.

[11] D. Lambert. Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34:1–14, 1992.

[12] D. Lord, S. P. Washington, and J. N. Ivan. Poisson, poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. *Accident Analysis and Prevention*, 37:35–46, 2005.

[13] B. Mandelbrot. An informational theory of the statistical structure of language. In W. Jackson, editor, *Communication Theory*. Betterworths, 1953.

[14] S. M. Mwalili, E. Lesaffre, and D. Declerck. The zero-inflated negative binomial regression model with correction for misclassification: An example in caries research. Technical Report TR0462, IAP Statistics Network, 2005.

[15] R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2004. ISBN 3-900051-00-3.

[16] J. A. Rice. *Mathematical Statistics and Data Analysis*. Wadsworth & Brooks/Cole, 1988.

[17] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464, 1978.

[18] H. A. Simon. On a class of skew distribution functions. *Biometrika*, 42:425–440, 1955.

[19] E. C. Titchmarsh. *The Theory of the Riemann Zeta Function, 2nd ed.* Oxford Science Publications, Clarendon Press, Oxford, 1986.

[20] D. G. Uitenbroek. *SISA Pairwise tests*. http://home.clara.net/ sisa/ pairwhlp.htm, 1997.

[21] D. von Seggern. *CRC Standard Curves and Surfaces*. CRC Press, 1993.

[22] J. R. Wilson. Logarithmic series distribution and its use in analyzing discrete data. In *Proceedings of the Survey Research Methods Section, American Statistical Association*, pages 275–280, 1988.

[23] G. K. Zipf. *Human Behaviour and the Principle of Least-Effort*. Addison-Wesley, Cambridge, MA, 1949.