Power Law Distributions in Information Retrieval

Casper Petersen, University of Copenhagen, Denmark
Jakob Grue Simonsen, University of Copenhagen, Denmark
Christina Lioma, University of Copenhagen, Denmark

Several properties of information retrieval (IR) data, such as query frequency or document length, are widely considered to be approximately distributed as a power law. This common assumption aims to focus on specific characteristics of the empirical probability distribution of such data, e.g. its scale-free nature or its long/fat tail. This assumption however may not be always true. Motivated by recent work in the statistical treatment of power law claims, we investigate two research questions: (1) To what extent do power law approximations hold for term frequency, document length, query frequency, query length, citation frequency and syntactic unigram frequency? (2) What is the computational cost of replacing ad hoc power law approximations with more accurate distribution fitting? We study 23 TREC and 5 non-TREC datasets and compare the fit of power laws to 15 other standard probability distributions. We find that query frequency and 5 out of 24 term frequency distributions are best approximated by a power law. All remaining properties are better approximated by the Inverse Gaussian, Generalized Extreme Value, Negative Binomial or Yule distribution. We also find the overhead of replacing power law approximations by more informed distribution fitting to be negligible, with potential gains to IR tasks like index compression or test collection generation for IR evaluation.

Categories and Subject Descriptors: H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

Additional Key Words and Phrases: Statistical model selection, Power laws

ACM Reference Format:
Casper Petersen, Jakob Grue Simonsen, Christina Lioma, 2014. Power Law Distributions in Information Retrieval *ACM Trans. Inf. Syst.* 9, 4, Article 39 (September 2014), 35 pages.
DOI: http://dx.doi.org/10.1145/0000000.0000000

1. INTRODUCTION

Several properties of the documents and queries used in information retrieval (IR) are widely considered to be approximately distributed as a *power law* [Arampatzis and Kamps 2008; Baeza-Yates et al. 2007; Chau et al. 2009; Chaudhuri et al. 2007]. A power law is a relationship between two quantities x and y, such that $y \sim Cx^{-\alpha}$, where α is a non-zero real number and C is a normalisation constant.[1] A well known example of a power law is *Zipf's law* [Zipf 1935], which states that the probability of the j^{th} most frequent term in an English text of N terms is $(1/H_N) \cdot j^{-1}$, where H_N is the N^{th} harmonic number, i.e. the sum of reciprocals of the N first positive integers $H_N = \sum_{n=1}^{N} \frac{1}{n}$. Based on Zipf's findings, Luhn postulated that term salience is relative to its frequency rank, such that as term frequency decreases, term salience increases [Luhn 1958]. These observations by Zipf and Luhn lie at the basis of seminal IR research

[1] C is used as shorthand for $1/\zeta(\alpha, \mathrm{x_{min}})$ – see Table 6

Author's addresses: Casper Petersen and Jakob Grue Simonsen and Christina Lioma, Computer Science Department, University of Copenhagen, Denmark.

DOI: http://dx.doi.org/10.1145/0000000.0000000

in term weighting [Spärck-Jones 1972], extensions of which are still used today. An assumption that we still make is that several properties of the data used in IR (terms, queries, documents) are approximately distributed as a power law.

Power law assumptions extend beyond IR to the relationship of variables in natural and social phenomena, e.g. the number of criminal acts perpetrated by individuals [Cook et al. 2004] or the number of substrates in biochemical reactions within metabolic networks of yeast [Jeong et al. 2000] (see Clauset et al. [2007] or Newman [2005] for more examples). As with Zipf's law, many of these examples assert that the *probability* of something occurring is in a power law relationship with the *number* of something (e.g., the number of substrates in metabolic networks, number of crimes committed, number of bytes in a file, number of terms in a search query), and is hence an assertion about probability distributions. However, severe criticism has been raised against mainstream methods for identifying power laws in data [Clauset et al. 2007], mostly on methodological grounds that (a) alternative hypotheses are rarely tested (i.e., other distributions could fit the data just as well or better) and (b) probing for power laws is frequently done using ad hoc techniques. Indeed, studies testing alternative hypotheses show that several claims about power laws do not hold; e.g., word frequency distributions may often, depending on text genre or language, be more precisely approximated by other standard probability distributions [Baayen 2001, Chap. 1] and city sizes are more adequately described by a log-normal [Eeckhout 2004].

Motivated by this, we ask two research questions: (1) To what extent do power law approximations hold for term frequency, document length, query frequency, query length, citation frequency and syntactic unigram frequency? (2) What is the computational cost of replacing ad hoc power law approximations with more accurate distribution fitting? Our treatment of the second research question is short compared to our first research question, but important as efficiency is a key issue of IR systems.

Section 2 presents our motivation, and Section 3 discusses related work. To answer our first research question, in Section 4, we study 28 datasets and fit, to each of the above six properties, *both* a power law and 15 alternative standard distributions. We use relative goodness-of-fit testing (akin to Clauset et al. [2007]) to select the distribution that best fits each property. To our knowledge no such systematic study has been published in IR in the last decade. We find that the power law is the best-fitting discrete model for 5 out of 24 term frequency distributions and for query frequency distributions. All remaining data properties (document length, query length, citation frequency and syntactic unigram frequency) are better approximated by the Inverse Gaussian, Generalized Extreme Value, Negative Binomial or Yule. To answer our second research question, in Section 5, we measure the time required to fit each model. We find that the overhead of using more accurate distribution fitting is negligible, compared to ad hoc techniques. We discuss the relevance of our findings to IR tasks in Section 6 and future research directions in Section 7.

2. MOTIVATION

We study the validity of power law approximations for IR properties that have been associated with power laws in the past:

Term frequency. The probability that a term occurs $n = 1, 2, \ldots$ times in a dataset, is approximately power law distributed according to Chau et al. [2009], Chaudhuri et al. [2007] and Momtazi and Klakow [2010].

Query frequency. The probability that a query occurs $n = 1, 2, \ldots$ times in a query log/stream[2], is approximately power law distributed according to Baeza-Yates and Tiberi [2007], Baeza-Yates et al. [2007] and Ding et al. [2011].

Query length. The probability that a query has $n = 1, 2, \ldots$ terms, is approximately power law distributed according to Clements et al. [2010], though only the tail is power law distributed according to Arampatzis and Kamps [2008].

Document length. The probability that a document has $n = 1, 2, \ldots$ terms, is approximately power law distributed according to Sato and Nakagawa [2010] and Srivastava et al. [2013].

Citation frequency. The probability that a paper is cited $n = 1, 2, \ldots$ times, is approximately power law distributed according to Redner [1998].

Syntactic unigram frequency. The probability that a syntactic unigram (a term and its parts of speech) occurs $n = 1, 2 \ldots$ times in a dataset, is approximately power law distributed according to Lioma [2007].

Our study is motivated by examples in various fields where reliance on the wrong distribution has had serious negative impact. For instance the 100 billion USD hedge fund Long Term Capital Management blowup was the result of an event ten standard deviations away from the mean in a Gaussian distribution, which should take place once per lifetime of the universe [Lowenstein 2000]. However, the assessment of the potential for overall losses was based on a Gaussian distribution, which severely underestimates the chance of a cataclysmic event [Buchanan 2004]. Similarly, Babbel et al. [2009] give examples on how the consideration of non-Gaussian distributions could have changed U.S. court case outcomes from employment discrimination to criminal prosecution.

Within IR, the literature is replete with examples of power laws fitted to empirical data (Section 3.1). However, there may be *practical* implications of using the best-fitting model instead of assuming a power law. Knowing the (approximately) best distribution is important for predicting the probability of some outcome in statistical inference, as using the "right" model will reduce uncertainty and lead to better predictions. Mao and Lu [2013] find that click trends of PubMed articles can be substantially better predicted for new articles when modelled using a log-normal instead of a power law. Further IR tasks relying on distributional assumptions might also benefit from using the most accurate model. For instance, when ranking with language models, Momtazi and Klakow [2010] use a Pitman-Yor process to assign probabilities to unseen words. As a Pitman-Yor process generates power law distributed probabilities of words, if it turns out that other distributions may fit this data better, using the best-fitting distribution may be beneficial. In index compression, Zipf's law is often used to describe the distribution of term frequencies, and can be used with γ-coding to estimate the size of an inverted index [Manning et al. 2008]. However, if the term frequency distribution is *not* approximately Zipfian, the estimate of the size of the index can be more than double the size of the actual index [Manning et al. 2008]. Using the best-fitting statistical model should provide a more accurate estimate on the index size that will aid in making more informed storage decisions. In test collection generation, queries and documents are often sampled from collections so that they approximately follow the *distributions* of real life queries and collections. Hence, knowing the right distribution may produce synthetic test collections that better approximate real life test collections both in terms of frequency *and* length of the queries/documents.

Finally, consider a search engine that allocates bandwidth based on query volume. A higher query volume requires more bandwidth. To ensure high throughput and fast

[2]For example, the probability that a query occurs $n = 1, 2, \ldots n$ times in a query log/stream means that we count how many times each query occurs and feed this vector/list of counts into our implementation

response times, whenever the free bandwidth drops below a threshold, more bandwidth must be allocated to prevent throttling users. However, as bandwidth costs money, allocation should only happen when necessary. Fig. 1 shows a theorised situation where the query volume is assumed to be power law distributed (solid line) when in fact the query volume is (empirically) distributed by a Negative Binomial (dashed line). The horizontal line shows the threshold where more bandwidth must be allocated. The x-axis shows the number of queries/minute (query volume) and the y-axis the free bandwidth percentage. We see two distinct areas where the power law assumption

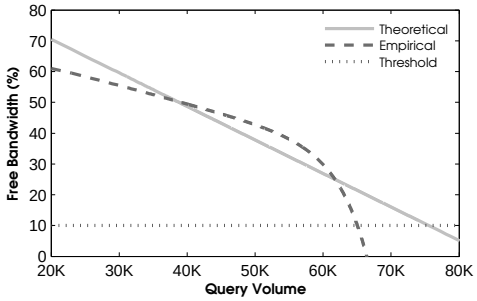


Fig. 1: Theorised (solid) and empirical (dashed) distributions of query volume of a search engine. The horizontal line is the threshold at which additional bandwidth must be allocated to ensure that end-users are not throttled.

departs from the empirical: From 20K to approximately 37K (and from 62K to 80K), the power law assumption indicates, incorrectly, that more bandwidth is available compared to the empirical distribution. In contrast, from 40K to approximately 62K, the power law assumption indicates that less bandwidth is available than the empirical distribution. The important scenario is from 62K to 80K where the empirical distribution indicates that bandwidth must be allocated at approximately 65K queries, whereas the power law assumption indicates that the query volume should be approximately 75K before bandwidth allocation becomes necessary. This difference suggests that users will experience throttling if the wrong distribution is assumed.

3. RELATED WORK

We review previous work on power law distributions in IR (Section 3.1), applications of power laws to IR (Section 3.2), and methods for detecting power laws (Section 3.3). Tables 1 through 5 summarise our literature review.

3.1. Power Law Distributions in IR

Distributions of IR data properties that have been reported to approximate a power law include: the frequency of terms (both single and co-occurring), not only in documents [Chaudhuri et al. 2007; Momtazi and Klakow 2010], but also in query logs [Chau et al. 2009; Baeza-Yates and Saint-Jean 2003; Baeza-Yates et al. 2007; Ding et al. 2011; Tong et al. 2013]; the frequency of term n-grams in various languages, for instance English [Egghe 2000; Quan Ha et al. 2003], Chinese [Dahui et al. 2005; Quan Ha et al. 2003] and Hungarian [Dominich and Kiezer 2005];[3] the frequency of syntactic n-grams [Lioma 2007]; and the number of documents per category in large Web taxonomies [Liu

[3]Egghe [2000] also investigates Flemish, Chinese and and Greek term n-grams.

et al. 2005].[4] In image IR, power laws have been found to approximate: the number of tags and comments assigned to images in a large Flickr dataset [Bolettieri et al. 2009]; the degree distribution of "image-similarity" networks for supporting image IR through browsing [Heesch and Rüger 2004]; and the number of faces in images [Baeza-Yates et al. 2004]. In music IR, tag vocabulary size is reported to grow according to a power law [Levy and Sandler 2009]. Moreover, node indegree distributions in music recommendation networks are reported to have a power law decay [Cano et al. 2006]. Similarly, the number of songs shared per user in a U.S. peer-to-peer (P2P) network is reported to also follow a power law [Koenigstein et al. 2010].

Web properties[5] that have been reported to be power law distributed include: tag distributions in collaborative tagging and social bookmarking systems [Bischoff et al. 2008; Halpin et al. 2007; Hotho et al. 2006; Peters and Stock 2010]; collaborative recommendation [Ye et al. 2011]; microblogging [Galuba et al. 2010]; information acquisition [Gatterbauer 2011]; intensities of webuser activity [Zhou et al. 2008]; session lengths [Xue et al. 2004]; number of replies made to weblogs and webposts [Mishne and Glance 2006]; degree distribution of on-line social networks [Garcia et al. 2013; Kwak et al. 2010]; number of comments per article for ranking on-line news [Tatar et al. 2014]; PageRank scores [Becchetti and Castillo 2006; Benczur et al. 2005]; weblinks and connected components [Broder et al. 2000; Soboroff 2002]; and click graphs [Cao et al. 2008; Craswell and Szummer 2007]. For instance, Bischoff et al. [2008] report that tag distributions from three different social tagging websites follow a power law, and Soboroff [2002] finds that the WT10g dataset is similar to the Web insofar as exhibiting power law degree distributions.

The process by which power laws emerge varies from, e.g. the choice of dampening factor in the PageRank computation [Becchetti and Castillo 2006], to models of behavioural psychology, e.g., *preferential attachment*: new objects tend to attach to popular objects [Barabási et al. 1999]. The latter has been used to claim the origins of power laws in Chinese: a preferential selection mechanism of words is used when new phrases are introduced to explain "emergent novelty" [Dahui et al. 2005]. Preferential attachment has, however, been questioned as a suitable mechanism for explaining such self-reinforcing behaviour [Huberman and Adamic 1999; Pennock et al. 2002].

3.2. Applications of Power Laws in IR

Power law assumptions about query frequency in search logs have been used repeatedly. For example, Asthana et al. [2011] use this assumption to improve expected accuracy for the top-1 document retrieved for higher values of the power law scale factor (α), in unstructured P2P search using non-uniform document replication strategies. Using the same assumption, Ding et al. [2011] derive an upper bound on the fraction of frequent queries received by a search engine in any interval of time. They find this fraction to be below current cache hit-rates, suggesting that query loads must increase. Hagen et al. [2010] multiply n-gram counts of disjoint segments of queries with a "power law factor" to obtain competitive query segmentation results. In retrieval effectiveness prediction, Azzopardi [2009] generates queries to test if retrieval precision is approximately power law distributed, but finds little support for the hypothesis.

Power law approximations of term (co-)occurrence may be useful for: (1) indexing [Arampatzis and Kamps 2008; Chaudhuri et al. 2007], based on the idea that match list size follows Heap's law: as the corpus grows, increasingly fewer new words are introduced thus limiting the number of new keyword-combinations, (2) ranking, based on the idea that language modelling smoothing can be adapted per term frequency

[4]Yang et al. [2003] further use the power law to bind the complexity of classifiers in category distributions.
[5]Many of the properties discussed so far were measured with Web data; see Tables 1 through 5.

[Kawamae 2014; Momtazi and Klakow 2010], and (3) for caching, using a least-recently-used policy that considers not only temporal and spatial locality of terms, but also term correlations [Tong et al. 2013].

Power laws have also been used to approximate the distribution of prior evidence in language modelling, e.g. to derive popularity scores for websites [Hauff and Azzopardi 2005], and as indicators of prior probability of document relevance based on (i) citations [Meij and de Rijke 2007] and (ii) link evidence [Kamps and Koolen 2008]. Power law approximations of part-of-speech n-grams [Lioma 2007] have also improved retrieval precision with pruned indices [Lioma and Ounis 2007], term weighting [Lioma and van Rijsbergen 2008], and query reformulation [Lioma and Ounis 2008]. Sigurbjörnsson and van Zwol [2008] use the power law distribution of tag frequencies in Flickr to define effective "promotion" functions for tag recommendation, while Peters and Stock [2010] suggest using only the most frequent tags for tag recommendation in collaborative tagging systems to boost recommendation precision.

In generating test collections, Cantone et al. [2009] develop a finite-state model that produces text having a Zipfian character distribution. Their model however does not succeed in producing text similar to natural English, French, German and Italian. Ferrer-i Cancho and Elvevåg [2010] suggest, however, that term frequencies in random texts, i.e. texts formed by selecting arbitrary symbols from a vocabulary and placing them in consecutive order, are not power law distributed. Similarly, in synthesising representative web-document sets to support accurate investigation into link-based retrieval, Gurrin and Smeaton [2004] require that the in-degree distribution of links is approximately power law distributed.

Power laws have also been reported in the context of searching P2P networks [Adamic et al. 2001; Jin et al. 2006; Sripanidkulchai et al. 2003]. Adamic et al. [2001] study efficient searching in power law degree distributed networks using random walks. As random walks gravitate towards high-degree nodes, this results in reduced search costs, which is important as the network grows. Jin et al. [2006] propose a semantic overlay model of power law P2P networks, where each node connects to semantically equivalent ones and find 60–150% improvement in recall over the Interest-based shortcut [Sripanidkulchai et al. 2003] model and a Gnutella network.[6]

3.3. Methods for Detecting Power Laws

We find four main methods for detecting power laws in our survey: (1) *graphical* methods, (2) *straight-line approximations*, (3) *generative* methods, and (4) statistical tests for model selection that use *relative goodness-of-fit* estimation.

3.3.1. Graphical Methods. Detecting power laws is most frequently done using *graphical* methods [Bauke 2007], i.e. a visual representation of data, such as a histogram. Recall that a power law is a relationship between two quantities x and y such that $y \sim Cx^{-\alpha}$, where C is a normalisation constant. Taking the logarithm gives $\log y \sim -\alpha \log x + C$, that is, x should approximately follow a straight line when plotted on double logarithmic axes. Checking for a power law is therefore straightforward [Clauset et al. 2007]:

Step 1: Measure the quantity x
Step 2: Construct x's distribution (e.g. a histogram of frequencies)
Step 3: Plot the distribution on double-logarithmic axes
Step 4: Check if the distribution approximates a straight line

These four steps constitute an *ad hoc* method for detecting power laws. In our literature survey we found that Steps 2 and 4 are where graphical methods differ.

[6]The authors specify Gnutella version 0.4 but give no further details on the dataset/model.

In constructing the distribution of x in Step 2, x is typically binned. We found direct, logarithmic and partial logarithmic binning to be most commonly used. In *direct* binning, each unique element in x is a bin. The bin holds the count of that element, divided by the cardinality of x to ensure a probability distribution. However, as direct binning produces a "noisy" tail (due to data sparsity) in datasets where many values occur infrequently, *logarithmic* binning uses bins whose widths w increase exponentially. To ensure the result is a probability distribution, each bin is divided by the cardinality of x times the width of the bin. While this may reveal trends not visible in the power law tail [Milojević 2010], information from the plot is lost as ranges of data values are reduced to a single point (i.e. the count in that bin). To avoid smoothing data that are not sparse, Milojević [2010] proposes using *partial* logarithmic binning where only data above some threshold where data are deemed sparse are logarithmically binned, which requires setting an ad hoc threshold. Another "binning" approach is to use Pareto quantile-quantile or QQ-plots [Cirillo 2013], where data are plotted (on the x-axis) against standard quantiles of a Pareto distribution[7] (on the y-axis). If data are Pareto distributed[8], they will approximately lie on the line $x = y$, i.e. a straight line.

A final graphical approach for detecting a power law is the complementary cumulative distribution function (CCDF) [Chakrabarti and Faloutsos 2006; Clauset et al. 2007] of the data, i.e. $1 - F(x)$, where $F(x)$ is the cumulative distribution function[9] of the data x. The CCDF is a monotonically non-increasing function describing the probability that a real-valued stochastic variable X will have a value larger than or equal to $x \in X$.

Direct binning, logarithmic binning, Pareto QQ and CCDF plots are shown in Fig. 2a–2d, using the document length distribution of ClueWeb09 cat. B. for the 10% least spam-like documents (using the Fusion spam rankings by Cormack et al. [2011]). We see that each binning technique produces a different visual representation of the data. For example, one may be more inclined to claim a straight line approximation when using a Pareto QQ-plot (Fig. 2c) than direct binning (Fig. 2a). Using these graphical methods to assess if data follows a straight line, without further statistical testing, is therefore error-prone. Despite being *ad hoc* approaches to data analysis, graphical techniques are the most popular in our literature survey (Tables 1 through 5).

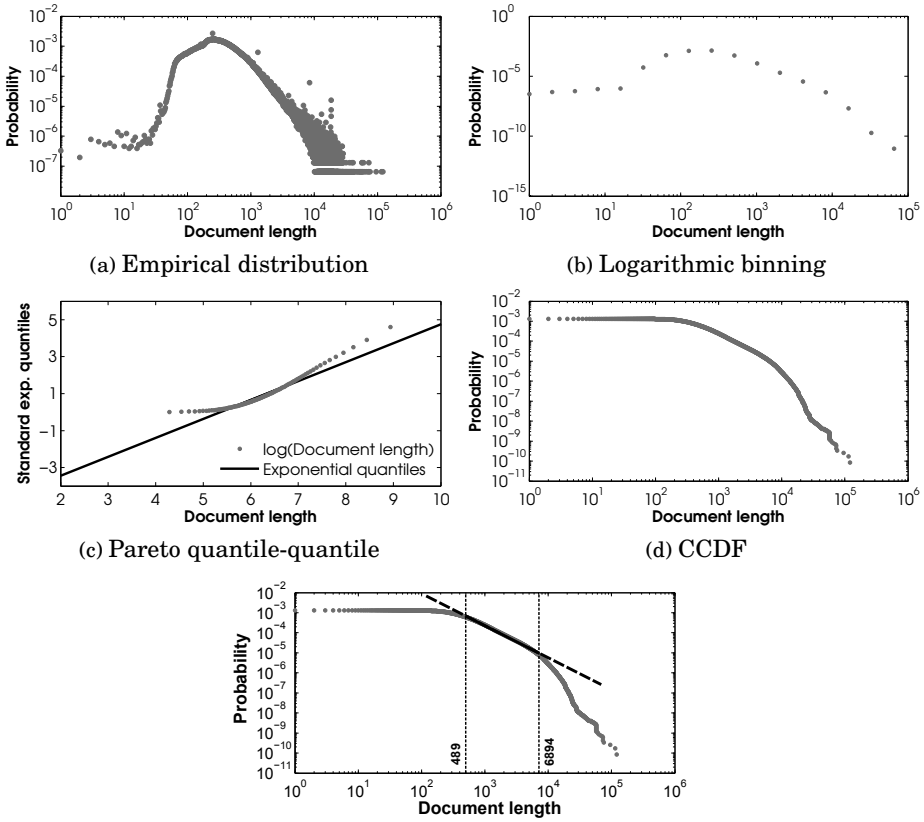
3.3.2. Straight Line Approximation. After the distribution of x is determined, it must be assessed if it follows a power law, that is, if it can be approximated by a straight line. The most common approach in our survey was visual inspection of the distribution on double-logarithmic axes. In some cases, a straight line was fitted to the distribution using linear least squares (LLS) [Milojević 2010] to illustrate how well the data (or parts of it) "fit" a power law, but rarely was the fit quantified using, for example, the coefficient of determination (R^2) [Benczur et al. 2005] or Spearman's ρ [Gabaix 2009; Medina et al. 2000; Palmer and Steffan 2000; Perc 2010].

If data are found to be power law distributed, the power law exponent α can be estimated using different approaches (see e.g. Chakrabarti and Faloutsos [2006]), such as maximum likelihood [Box and Cox 1964], Hill estimation [Hill 1975] and its bias-reducing versions [Feuerverger and Hall 1999], nonparametric estimators [Crovella and Taqqu 1999], slope of regression line (using LLS and C as the y-intercept), and minimisation of the Kolmogorov-Smirnov distance [Clauset et al. 2007]. Of these, the slope of the regression line was the most common in our review.

[7]Often the theoretical quantiles of an exponential distribution are used, as a log-transformed Pareto random variable is exponentially distributed [Beirlant et al. 2006].

[8]A Pareto distribution is nearly identical to a Zipfian. As Newman [2005] writes: "Zipf made his plots with x on the horizontal axis and $P(x)$ on the vertical. Pareto did it the other way around".

[9]The cumulative distribution function $F(x)$ is given by $F(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{x_i \leq x}$, where $1_{x_i \leq x}$ is an indicator function emitting 1 if $x_i \leq x$ and 0 otherwise.



(a) Empirical distribution

(b) Logarithmic binning

(c) Pareto quantile-quantile

(d) CCDF

(e) CCDF (central part fitted)

Fig. 2: Common graphical methods. Data are document lengths (number of terms in a document) of the 10% least spam documents of ClueWeb09 cat. B. on double logarithmic axes (base 2). (2a) Direct binning (empirical distribution). (2b) Logarithmic binning. (2c) Pareto QQ-plot to standard exponential quantiles. (2d) Complementary cumulative distribution function (CCDF) $1 - F(x)$, where $F(x)$ is the cumulative distribution function. (2e) CCDF with a power law fitted to the "central part" of the data.

3.3.3. Generative Models. When empirical data are not available (such as when defining a prior distribution as in Kawamae [2014]), generative processes can generate data distributed according to a power law. These processes are typically variations of preferential attachment [Mitzenmacher 2004], where new nodes in a network attach to existing nodes with a probability proportional to their degree. The Pitman-Yor [Pitman and Yor 1997] and Simon process [Simon 1955] are examples of such generative processes, though other generative models producing power law distributions have been proposed by Mandelbrot [1953] for rank-frequency distributions of words (though not without criticism [Miller 1957]), and multiplicative models (such as the double Pareto [Reed 2003]) for modelling income [Mitzenmacher 2004]. Arampatzis and Kamps [2009], for example, generate natural language queries whose lengths are drawn from a fitted Poisson-Zipf distribution and Volkovich et al. [2007] use the theory of regular variation [Meerschaert and Scheffler 2001] to show that PageRank and indegree are approximately power law distributed with the same exponent for three different Web datasets.

3.3.4. Model Selection and Relative Goodness-of-Fit (GOF) Estimation. Clauset et al. [2007] compare power law claims to alternative models as follows: they first use standard maximum likelihood estimation (MLE) to fit power laws to data, and then statistical tests to compare the power law fit to *alternative* models. Fitting various distributions to data, and selecting the "best" is known as *model selection*. A standard method is to (a) use MLE to find the parameters of each model and (b) use a statistical significance test to identify which model minimises some, typically information-theoretic, divergence measure such as the Kullback-Leibler (KL) divergence, to an unknown "true" model.[10] Alternatives to MLE are sometimes used to minimise some error measure, e.g., the least-squares distance, but usually do not compare to alternative models [Burnham and Anderson 2002; Corder and Foreman 2009].

A well-known KL-divergence-based test for model selection is the Akaike Information Criterion (AIC) [Akaike 1974]. Alternatives exist, e.g. the likelihood ratio-based Vuong's Closeness Test [Vuong 1989]. Generally, there are three main approaches to model selection using optimisation of some selection criteria [Burnham and Anderson 2002]: (i) use criteria that estimate KL divergence, e.g., AIC and Vuong's test; (ii) use criteria that estimate the (dimension of the) underlying "true" model, such as the Bayesian Information Criterion [Schwarz 1978], the Minimum Description Length principle [Grünwald 2007], and variations of means-squares methods, notably Mallows' C_p [Mallows 1973]; and (iii) ad hoc methods such as least-squares ranking. These selection criteria concern the *relative* GOF of several models to data, i.e., they can only tell if a model fits better than others; they do not concern *absolute* GOF, that is, if a model fits *well*. Standard tests exist for testing if a sample is drawn from a distribution (i.e., if the model fits well), e.g., the χ^2 and G tests, as well as tests specialised to specific distributions [Burnham and Anderson 2002; Lehmann and Romano 2006]. Most absolute GOF tests posit a null hypothesis that data are drawn from a model, and thus cannot reject the null hypothesis at small sample sizes, but sometimes aggressively reject the null hypothesis at large sample sizes (and thus sometimes are "badness-of-fit" tests [Roberts and Pashler 2000; Schunn and Wallach 2005]).

Table 1: Overview of related work using *graphical methods*. * exponent/coefficient reported.

POWER LAW FOUND USING GRAPHICAL METHODS		
Publication	*Dataset*	*Distribution*
Adamic et al. [2001]	AT&T call graph Gnutella	Frequency of nodes with outdegree=k* Frequency of nodes with k links*
Baeza-Yates and Saint-Jean [2003]	Query log	# Documents containing a term* Frequency of single-word queries*
Baeza-Yates et al. [2007]	Yahoo! UK Web log Yahoo! UK query log	Frequency of queries* Frequency of query terms* Frequency of document terms*
Bischoff et al. [2008]	last.fm, del.icio.us, Flickr last.fm and lyricsdownload.com Stanford WebBase	Frequency of tags Frequency of tags in track lyrics Frequency of anchor text
Bolettieri et al. [2009]	Flickr subset	Frequency of tags w.r.t. # images they are found in
Chaudhuri et al. [2007]	MSN news AQUAINT Movies (Polarity) & products	Frequency of top 32K terms Frequency of top 26K terms Sizes of multi-keyword indexes Frequency of top 32K terms Frequency of top 90K term n-grams (n=[2,3])
Dahui et al. [2005]	Ancient Chinese characters	Frequency of characters
Ding et al. [2011]	Excite query log	Frequency of queries Frequency of term n-grams (n=[2,3])
Egghe [2000]	Dewey [Heaps 1978, p. 200]	Frequency of English characters

Continued on next page

[10] A "true" but unknown model is assumed to quantify the information loss of selecting the "wrong" model.

Table 1 – continued from previous page

	Author report Chinese characters Beckman [1999] Books [Yannakoudakis et al. 1990]	Frequency of Flemish characters Frequency of top k={100,300} characters Frequency of Greek characters
Heesch and Rüger [2004]	Image graph	Frequency of nodes with indegree k (in a range)
Hotho et al. [2006]	del.icio.us	Frequency of tags Frequency of tagged URLs Frequency of users tagging
Kamps and Koolen [2008]	INEX 2006	# websites with indegree=k
Koenigstein et al. [2010]	Gnutella crawl (song similarity graph)	# Songs shared by users (cutoff) # Users sharing a song Frequency of degrees
Krause et al. [2008]	del.icio.us MSN click log AOL click log	Frequency of session degree Frequency of clicked URLs degree Frequency of queries/tags degree
Kwak et al. [2010]	Twitter	# of followers (in a range)* # Users participating in retweets
Levy and Sandler [2009]	last.fm & MyStrands	Tag vocabulary growth (Heaps' law)*
Lioma [2007]	Associated Press TREC Disk 4 & 5 WT2g and WT10g	Part-of-speech n-grams (n=$[1, \ldots, 100]$)
Meij and de Rijke [2007]	TREC Genomics 2004 & 2006	# Received citations
Ripeanu and Foster [2002]	Gnutella crawl	# Nodes with k links
Xue et al. [2004]	MSN Query log	Query session distribution
Ye et al. [2011]	Foursquare & Whrrl	Probability that users visit nearby points-of-interest
Zhang and Lesser [2006]	Scale-free network generated using Palmer and Steffan [2000]	Frequency of outdegrees
Zhou et al. [2008]	del.icio.us	Frequency of users with k tags

Table 2: Overview of related work using a *Pitman-Yor process*. * exponent/coefficient reported.

POWER LAW FOUND USING PITMAN-YOR PROCESS		
Publication	*Dataset*	*Distribution*
Kawamae [2014]	Words (Pitman-Yor process)	Frequency of terms
Momtazi and Klakow [2010]	TREC disks 4 & 5	Frequency of terms

Table 3: Overview of related work using *model selection*. * exponent/coefficient reported.

POWER LAW FOUND USING MODEL SELECTION		
Publication	*Dataset*	*Distribution*
Arampatzis and Kamps [2008]	TREC Million Query 2007 TREC Terabyte 2005 and 2006 AOL [Pass et al. 2006]	Query length (see [Arampatzis and Kamps 2009])*
Azzopardi [2009]	Associated Press Wall Street Journal AQUAINT	Vocabulary distribution*
Garcia et al. [2013]	FaceBook, Friendster, Livejournal, Myspace, Orkut	Frequency of nodes with k indegree*
Gatterbauer [2011]	Web crawl Delicious	Frequency of domains with $\geq k$ inlinks* Frequency of tags repeated k times*
Tatar et al. [2014]	20minutes	# of comments*

Table 4: Overview of related work using *linear regression*. * exponent/coefficient reported.

POWER LAW FOUND USING LINEAR REGRESSION		
Publication	*Dataset*	*Distribution*
Baeza-Yates and Tiberi [2007]	Yahoo! query log	Frequency of queries* Frequency of queries with no clicks* Frequency of URLs with k clicks* Size of connected components* Frequency of nodes with k indegree*
Becchetti and Castillo [2006]	.GR Web crawl	PageRank scores*
Broder et al. [2000]	Web crawl (1999)	# Pages with indegree=k* # Pages with outdegree=k* # Connected components of size n*
Cano et al. [2006]	MSN Entertainment Amazon	Frequency of nodes with k indegree* Frequency of nodes with k in/outdegree*
Cantone et al. [2009]	Characters (generated by own model) Terms (generated by own model)	Frequency of characters* Frequency of terms in natural language texts
Cao et al. [2008]	Click-through bipartite graph	Frequency of edges with weight=w* # of query nodes (in a range) # URLs clicked for a query*
Chau et al. [2009]	Timway query log (Chinese)	Frequency of term n-grams (n=$[1, \ldots, 6]$)*
Galuba et al. [2010]	15M tweets	# Tweets & URLs posted by users* # Tweets & users mentioning a URL* # Weakly connected components and their sizes*
Gurrin and Smeaton [2004]	SPIRIT subset Random sample	# websites with k outdegree* # websites with k indegree*
Halpin et al. [2007]	del.icio.us (Popular section) del.icio.us (Random section)	Frequency of tags relative to position* Frequency of tags relative to position*
Liu et al. [2013]	TREC07p & TanCorp	Frequency of term n-grams (n=$[1, \ldots, 4]$)*
Mishne and Glance [2006]	Blogpulse [Glance et al. 2004]	# Comments per Weblog/post*
Quan Ha et al. [2003]	Wall Street Journal TREC dataset (Chinese) Mandarin News (Chinese)	Frequency of term n-grams*(n=[1...5]) Frequency of compound terms * (a compound term are 2+ characters) Frequency of compound term n-grams*(n=[1...5])
Redner [1998]	ISI citation	# of citations (in a range)*
Sigurbjörnsson and van Zwol [2008]	Flickr	Frequency of tags* # tags per photo*
Soboroff [2002]	WT10G	# Pages with k indegree* # Pages with k outdegree* # Connected components of size n*
Tong et al. [2013]	Query log	Frequency of terms* Frequency of co-occurring terms*

Table 5: Overview of related work with *no evidence*. * exponent/coefficient reported.

NO EVIDENCE		
Publication	*Dataset*	*Distribution*
Asthana et al. [2011]	—	Frequency of queries (from [Baeza-Yates et al. 2007])
Babbar et al. [2014]	Directory Mozilla	# Documents per category (from [Liu et al. 2005; Yang et al. 2003])
Baeza-Yates et al. [2004]	Images from .CL domain	# Images with area of size k (pixels) # Human faces in images*(for home page images)
Benczur et al. [2005]	Web crawl (31.1M pages)	PageRank scores
Craswell and Szummer [2007]	Query log	Frequency of URLs per query Frequency of queries per URL # Repetitions of query-URL pairs
Hagen et al. [2010]	Google n-gram dataset	Query segmentation scores
Hauff and Azzopardi [2005]	WT2g	Popularity scores
Jin et al. [2006]	Power law graph	Unknown*
Lioma and Ounis [2007]	WT10g	Part-of-speech blocks
Liu et al. [2005]	Yahoo! Directory	# Documents per category
Peters and Stock [2010]	—	Frequency of "power" tags (system not evaluated)

Ramasubramanian and Sirer [2004]	—	Frequency of queries [Breslau et al. 1999; Jung et al. 2002]
Sripanidkulchai et al. [2003]	Gnutella connectivity graphs 2001	Frequency of nodes with degree=k^*

4. TO WHAT EXTENT DO POWER LAW APPROXIMATIONS HOLD?

To answer this research question, we follow Clauset et al. [2007] because their method (i) is a systematic approach for fitting statistical models to data that (ii) allows for model selection by comparing power law fits to those of alternative models, and that (iii) has been used in IR before [Arampatzis and Kamps 2008; Azzopardi 2009]. Next, we describe our methodology (Section 4.1), experimental setup (Section 4.2), and findings (Section 4.3).

4.1. Methodology

To find which of the 16 models best fits our data, we first use MLE to estimate the best-fitting parameters of each model, and then statistical tests to see which model fits best. The box at the end of this section summarises our methodology.

4.1.1. Step 1. Maximum Likelihood Fitting. A (parametric) *model* is a set $\mathcal{N} = \{f(\cdot|\theta) : \theta \in \Theta\}$ of probability density functions[11] (PDF), where θ is the vector of parameters of the model, ranging over the possible values in Θ. While the word "model" is standard in statistical model fitting, note that each model is a set of probability distributions. "Fitting a model" means in this context "fitting a probability distribution to the data". For example, the model consisting of all Gaussian distributions $\mathcal{N}$ is:

$$\begin{aligned}\mathcal{N} &= \{f(x|\theta) : \theta \in \Theta\} \\ &= \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} : \mu \in \mathbb{R},\ \sigma^2 \in \mathbb{R}_+,\ x \in \mathbb{R} \right\}\end{aligned} \tag{1}$$

Given specific values for θ, the corresponding PDF will show that some data are more probable than other data. The task of MLE is to find the PDF that *most likely* produced the data. The *log-likelihood* function $\mathcal{L}$ is a function of the *parameters* of $\mathcal{N}$ given the data that summarise the amount of "evidence" in the data (i.e. how likely the data are) given specific values of θ. Assuming $x = \{x_1, \ldots, x_n\}$ is an i.i.d. sample, $\mathcal{L}$, for $\mathcal{N}$, is:

$$\mathcal{L}(\theta|x) = \sum_{i=1}^{n} \ln f(x_i|\theta) = -n\log(\sigma^2) - n\log(\sqrt{2\pi}) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 \tag{2}$$

and we seek to maximise the average log-likelihood, that is, obtain $\hat{\theta}$:

$$\hat{\theta} \in \left\{ \arg\max_{\theta\in\Theta} \frac{\mathcal{L}(\theta|x)}{n} \right\} \tag{3}$$

For each dataset x and for each candidate model $\mathcal{N} = \{f(x|\theta) : \theta \in \Theta\}$, we compute $\hat{\theta}$ to obtain the distribution $f(x|\hat{\theta})$. This distribution can be seen as the model $\mathcal{N}$ fitted to the data using $\hat{\theta}$ as the "best" parameters.

4.1.2. Step 2. Comparison of Models. Methods for finding the model that fits the data best include Bayesian Information Criteria [Schwarz 1978], the *J* test [Davidson and MacKinnon 1981], minimum description length [Grünwald 2007] and the Clarke test [Clarke 2007]. We use Vuong's Closeness test [Vuong 1989] as per Clauset et al. but any other test could have been used. Vuong's test compares two models by their likelihood

[11] Or probability mass functions if the model is defined only for discrete data.

which leads to probabilistic conclusions about their relative fits. We also include the AIC [Burnham and Anderson 2002] (with finite sample size correction) as it has been argued that it has advantages over likelihood ratio tests (like Vuong's test) [Posada and Buckley 2004]. Both tests are based on KL divergence and will select the same model in the limit of large sample sizes, namely the model that minimises the information loss w.r.t. the unknown "true" model; however, they differ as Vuong's test is probabilistic, hence intuitively gives "confidence" in its prediction of the superiority of one model over another. We include both tests, as different predictions from the two tests for some dataset will indicate that the sample may not be large enough to rule out one model.

Vuong's Test. Vuong's Closeness Test, one of the least controversial tests for non-nested model testing [Clarke 2003], compares a pair of models $M_1 = \{f_1(\cdot|\theta_1) : \theta_1 \in \Theta_1\}$ and $M_2 = \{f_2(\cdot|\theta_2) : \theta_2 \in \Theta_2\}$, by computing:

$$Z_V = \frac{\sum_{i=1}^{n} \ln f(x_i|\hat{\theta}_1) - \sum_{i=1}^{n} \ln f_2(x_i|\hat{\theta}_2) + c}{\sqrt{n\hat{\omega}^2}} \tag{4}$$

where

$$\hat{\omega}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\ln\frac{f_1(x_i|\hat{\theta}_1)}{f_2(x_i|\hat{\theta}_2)}\right)^2 - \left(\frac{1}{n}\sum_{i=1}^{n}\ln\frac{f_1(x_i|\hat{\theta}_1)}{f_2(x_i|\hat{\theta}_2)}\right)^2 \tag{5}$$

and c is a correction that penalises models with many parameters when applied to small samples (the correction is given at the end of this section). The numerator in Eqn. 4 (excluding c) is the difference in log-likelihood ratios of the maximised likelihood functions and $\hat{\omega}^2$ is the sample variance of the log-likelihood ratios.

Vuong's Closeness Test rejects the null hypothesis that M_1 and M_2 are equivalent, in favour of the hypothesis that M_1 is better than M_2 at significance level p, if Z_V exceeds the $(1-p)$-quantile of the standard Gaussian. Roughly, p can be interpreted similarly to standard significance levels of significance tests, i.e., very low values of p (e.g., $p < .001$) may be interpreted as strong evidence that M_1 is a better fitting model than M_2.

When using Vuong's Closeness Test, we only consider two of the three possible cases (as do Clauset et al. [2007]): nested[12] and non-nested. The last case, overlapping distributions[13], is not included because none of the models we consider are overlapping.

Akaike Information Criterion. In the Akaike Information Criterion (AIC) the following is computed for a model Y with $\hat{\theta}$ and c:

$$Z_A = 2\left(|\hat{\theta}| - \sum_{i=1}^{n}\ln f(x_i|\hat{\theta})\right) + c \tag{6}$$

where c is again correcting for finite sample sizes that penalise models with many parameters when applied to small samples. Small Z_A values indicate "better fits", hence models with many parameters (i.e., large values of $|\hat{\theta}|$) are penalised, as are models where the sample size is not many times larger than $|\hat{\theta}|$. While the AIC test cannot be used as a GOF criterion for single models, it can be used for comparison of multiple different models on the same samples [Burnham and Anderson 2002], the favoured model being the one with the smallest Z_A value.

Correcting for Finite Sample Sizes. Both Vuong's test and the AIC test are derived under the condition that the sample size is infinite. When the sample size is finite, as in the case of IR data, both tests must be corrected. For both tests, this correction

[12]Vuong's test for nested models should asymptotically be χ^2 distributed [Vuong 1989].
[13]Let F and G be two models. F and G are overlapping if $F \cap G \neq \emptyset$ and $F \subsetneq G$ and $G \subsetneq F$.

is based on penalising a model according to its number of (free) parameters, i.e. its lack of parsimony. For Vuong's test, as we test many models with different numbers of parameters, we use the correction based on the Schwartz information criterion [Schwarz 1978] suggested by Vuong [1989], namely $c = -(|\hat{\theta}_1| - |\hat{\theta}_2|) \log n)/2$, which vanishes in the limit of large n as $\log n \in o(\sqrt{n})$. For the AIC test, we use the standard correction, appropriately called "AIC with correction for finite sample size" (AICc) [Burnham and Anderson 2002; Hurvich and Tsai 1989], namely $c = (2|\hat{\theta}|(|\hat{\theta}| + 1))/(n - |\hat{\theta}| - 1)$.

4.1.3. Step 3. Model Selection. A GOF test of a statistical model summarises how well the model fits the data [Maydeu-Olivares and Garca-Forero 2010]. Two kinds of GOF tests are *absolute* and *relative*. Absolute tests summarise the discrepancy between a statistical model and the data. Relative tests summarise the discrepancy between two statistical models. As our aim is to investigate if models other than power laws fit IR data properties better, we use *relative* GOF. In our experiments we do *not* claim that certain data properties are distributed by certain models, but rather suggest that certain models better fit the data than others. Our choice of GOF is different than that of Clauset et al. [2007] who use a specific Monte Carlo GOF test of the null hypothesis that data follows a power law: in their approach, data are fitted to a power law and a large number of synthetic datasets are subsequently generated and their fit compared to the original fit using an appropriate distance function. If the synthetic datasets frequently have a greater distance than the original data have to their own model, the fit is poor. However, we note that this test suffers from similar problems as many ordinary GOF tests [Roberts and Pashler 2000; Schunn and Wallach 2005]: it only *rejects* models, and will do so as sample sizes become large. This is especially damaging for IR datasets as they tend to be large. Other options for assessing the absolute GOF of a model is Pearson's statistic or ordinary likelihood ratio tests, but these tend not to yield accurate p-values [Maydeu-Olivares and Garca-Forero 2010].

Methodology Outline

(1) **Model estimation step**: Estimate the parameters of the models.
(2) **Model comparison step**: Compare each pair of models using Vuong's likelihood ratio (LR) test [Vuong 1989].
(3) **Model selection step**: Choose as best-fitting the model that, depending on the sign of Vuong's test, "wins" the most comparisons with all other models.

4.2. Experimental Setup

We study the datasets listed in Table 7 (indexed with Indri 5.7[14], without stemming or stop word removal – for a description of the datasets see Appendix N) and the data properties shown in Table 8. We study various distributions in these datasets using 16 statistical models. Many standard statistical models exist (see e.g. Forbes et al. [2011] and Johnson et al. [2002]), including specialised models, such as the double Pareto [Reed and Jorgensen 2004], or mixture models, e.g. Poisson-power law (Arampatzis and Kamps [2008]). While such specialised models may approximate data better than any single model, they typically require estimating a transition point where one model starts approximating the data "better" than another. Similarly, a model can be fitted only to part of the data (for power-laws typically the "central" or "tail" part, see Fig. 2(e)), requiring two such transition points to demarcate which part of the data is being fitted. Finding such transition points in a principled manner is non-trivial, as estimating them by, say MLE, adds one new parameter to each model. In addition, the method used to

[14]http://www.lemurproject.org/indri/.

find such transition points may aggressively throw away data. As an example, we apply the method of Clauset et al. to the Excite query log, the iSearch and the Congressional Record (CR) datasets. In the Excite query log, each data point is the length (number of terms) of a query; in the iSearch dataset, each data point is the length (number of terms) of each document and in the CR dataset, each data point is the frequency of a given term in the collection. Clauset's method proceeds to discard 99.99% of the data points in the Excite query, (ii) 96.97% of the data points in the iSearch dataset and (iii) 99.78% of the data points in the CR dataset. Similar examples are found in [Garcia et al. 2013]. Furthermore, the running time of their method is quadratic in the difference between the largest and smallest value in each dataset making it infeasible for datasets with outliers.

To avoid such complications, we use only standard statistical models commonly found in statistical software like MATLAB, R or OCTAVE with support in either the non-negative or the positive integers (for discrete models), or the positive reals (for continuous models) and hence eschew finding transition points in the data – that is, we consider *only* fits of models to the full range of each data property. We use the 16 models listed in Table 6 (see Appendix A for a more detailed version). We do not use the Binomial, Uniform and Bernoulli distributions as (i) the Binomial is well-approximated by a Gaussian for large sample sizes, (ii) the Uniform distribution assumes every observation has equal probability, which is highly unlikely for IR data properties, and (iii) the Bernoulli distribution only has binary outcomes.

IR datasets typically only contain discrete data, and their properties are well-defined for the set of non-negative integers. We use both discrete and continuous models, because (i) this gives a greater range of models with which to study the data properties, and (ii) the continuous models can in general be discretised using approximations or continuity corrections. We report both the overall best-fitting model (which may be continuous) *and* the best-fitting discrete model.

Six of the above 16 models have parameters that can be interpreted as modelling certain behaviour (for the rest, the parameters simply dictate the "shape" of the distribution). These six models are: (1) *Exponential.* The μ parameter describes the waiting time between Poisson distributed events. Events are Poisson distributed if a number of these occur in a fixed period of time and if they occur with a known average rate l and independently of the time passed since the last event. (2) *Gamma* models the time for a events to occur, given that the events occur randomly in a Poisson process with average waiting time between events equal to b. (3) *Log-normal* arises as the *product* of any i.i.d. positive random variables. The (log-transformed) parameters have the same meaning as for a Gaussian. (4) *Negative Binomial.* If $r \notin \mathbb{N}^+$, an interpretation based on Bernoulli trials cannot be given [Bean 2001]. If $r \in \mathbb{R}^+$, r controls the amount of excess correlation in the data or the deviation of the Negative Binomial from the Poisson distribution [Gregoriou 2009; Hilbe 2011]. (5) *Power law.* The exponent α indicates how steep the power law is. The steeper the power law, the more skewed the distribution. This has implications in several areas. In caching for example, a small α results in weaker temporal locality and worse cacheability [Jin and Bestavros 2000][15]. (6). *Weibull* is extensively used to model life-data (quantities having a limited life-span) where a shape parameter $a<1$ indicates "infant-mortality" (that is, a large number of whatever is being observed fails/dies early on in its life). An $a=1$ indicates a constant rate of failure (that is, if the object of interest has survived so far there is a constant, but small, chance of failure/death) and an $a>1$ indicates that failure/death increases as the object gets older mandated on it having survived past the $a=1$ stage.

[15]Kunegis and Preusse [2012], however, argue that higher values of α make for a more equal distribution based on Lorenz curves and the Gini coefficient.

DISCRETE MODELS			
Model	Probability mass function (PMF)	Conditions	Nested in:
Geometric (Geo)	$f(x\|p)=(1-p)^x p$	$x\in\mathbb{N}_0$, $0<p\leq 1$	NBin
Negative Binomial (NBin)	$f(x\|r,p)=\binom{r+x-1}{x}p^r(1-p)^x$	$x\in\mathbb{N}_0, r,p\in\mathbb{R}^+$, $p\in(0,1)$	
Poisson (Poiss)	$f(x\|\lambda)=\frac{\lambda^x}{x!}\exp(-\lambda)$	$x\in\mathbb{N}_0$, $\lambda\in\mathbb{R}^+$	
Power law (P-law)	$f(x\|\alpha,x_{\min})=\frac{x^{-\alpha}}{\zeta(\alpha,x_{\min})}$	$x,x_{\min}\in\mathbb{N},\alpha\in\mathbb{R}^{>1}$	
Yule-Simon (Yule)	$f(x\|p)=p\beta(x,p+1)$	$x\in\mathbb{N}$, $p\in\mathbb{R}^+$	
CONTINUOUS MODELS			
Model	Probability distribution function (PDF)	Conditions	Nested in:
Exponential (Exp)	$f(x\|\mu)=\frac{1}{\mu}\exp\left(\frac{-x}{\mu}\right)$	$x\in\mathbb{R}_0^+$, $\lambda\in\mathbb{R}^+$	Gamma,Wbl,GP
Gamma (Gamma)	$f(x\|a,b)=\frac{x^{a-1}\exp\left(\frac{-x}{b}\right)}{b^a\Gamma(a)}$	$x\in\mathbb{R}_0^+$, $a,b\in\mathbb{R}^+$	
Gaussian (Gauss)	$f(x\|\mu,\sigma^2)=\frac{\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}{\sqrt{2\pi\sigma^2}}$	$x\in\mathbb{R}$, $\mu\in\mathbb{R}$, $\sigma^2\in\mathbb{R}^+$	
Generalized Extreme Value (GEV)	$f(x\|k,\mu,\sigma)=\left(\frac{1}{\sigma}\right)\exp\left(-\left(1+k\frac{(x-\mu)}{\sigma}\right)^{-\frac{1}{k}}\right)\times \left(1+k\frac{(x-\mu)}{\sigma}\right)^{\left(-1-\frac{1}{k}\right)}$	$x\in[-\infty;\mu-\sigma/k]$ $\sigma\in\mathbb{R}^+,\mu,k\in\mathbb{R}$	
Generalized Pareto (GP)	$f(x\|k,\sigma,\theta)=\left(\frac{1}{\sigma}\right)\left(1+k\frac{(x-\theta)}{\sigma}\right)^{-1-\frac{1}{k}}$	$x\geq\theta : k>0$ $\theta\leq x\leq\theta-\frac{\sigma}{k} : k<0$	
Inverse Gaussian (IGauss)	$f(x\|\mu,\lambda)=\left(\frac{\lambda}{2\pi x^3}\right)^{\frac{1}{2}}\exp\left(\frac{-\lambda(x-\mu)^2}{2\mu^2 x}\right)$	$x\in\mathbb{R}^+,\mu,\theta\in\mathbb{R}^+$	
Logistic (Log)	$f(x\|\mu,\sigma)=\frac{\exp\left(\frac{x-\mu}{\sigma}\right)}{\sigma\left(1+\exp\left(\frac{(x-\mu)}{\sigma}\right)\right)^2}$	$x\in\mathbb{R}$, $\mu\in\mathbb{R}$, $\sigma\in\mathbb{R}^+$	
Log-normal (Logn)	$f(x\|\mu,\sigma^2)=\frac{\exp\left(\frac{-(\ln x-\mu)^2}{2\sigma^2}\right)}{x\sigma\sqrt{2\pi}}$	$x,\sigma^2\in\mathbb{R}^+$, $\mu\in\mathbb{R}$	
Nakagami (Naka)	$f(x\|\mu,\omega)=2\left(\frac{\mu}{\omega}\right)^\mu\frac{x^{2\mu-1}}{\Gamma(\mu)}\exp\left(\frac{-\mu x^2}{\omega}\right)$	$x\in\mathbb{R}^+$, $\mu,\omega\in\mathbb{R}^+$	
Rayleigh (Rayl)	$f(x\|b)=\frac{x}{b^2}\exp\left(\frac{-x^2}{2b^2}\right)$	$x\in\mathbb{R}_0^+$, $b\in\mathbb{R}^+$	Wbl
Weibull (Wbl)	$f(x\|a,b)=ba^{-b}x^{b-1}\exp\left(-\left(\frac{x}{a}\right)^b\right)$	$x\in\mathbb{R}_0^+$, $a,b\in\mathbb{R}^+$	

Table 6: Our 16 statistical models identified by name, PDF and conditions. Notation: γ: Incomplete gamma function, ϕ: Gaussian, erf: Error function, ζ: Hurwitz zeta function.

4.3. Experimental Findings

We now present the results separately for each of the six data properties.

4.3.1. Term Frequency. Table 9 shows the results for the term frequency distribution. All plots and tables pertaining to this property are found in Appendix F (KS-distance plots), Appendix E (Vuong tables) and Appendix M (distribution plots). For all datasets, the best overall fit is the continuous Generalized Extreme Value (GEV). By the Fisher-Tippett-Gnedenko Theorem, the (normalised) extrema of any sequence of i.i.d. variables are GEV distributed, hence GEV is often used to describe extreme or rare events [Shukla et al. 2010]. A rare event corresponds here to a rare term that is likely to appear in the tail of the term frequency distributions. As the number of rare terms increases, the tail becomes sparser, and it is this data sparsity the GEV can approximate better than the other models. The best-fitting *discrete* model is, for most (19/24) datasets, Yule. For large values of x [Drucker 2007, chap. 1]: $f(x|k)\propto\frac{1}{x^{k+1}}$, that is, the tail of Yule's model is a realisation of Zipf's law [Simon 1955]. That Yule's model –and not a power law– is the best-fitting discrete model suggests a preferential attachment mechanism, i.e. that "popular" words are likely to "attach" to new words. The best-fitting discrete model for the Congressional Record, Federal Register, iSearch and ClueWeb data is the power law. Vuong's test shows that the results of the pair-wise comparison of models are statistically significant at the $p<.05$ level. Different best-fitting models are found by Vuong's test and the AICc for six datasets, indicating that some discrepancy in the data remains unexplained. The different best-fitting models all have AICc values very close to each other. E.g., for the Financial Times, Vuong's test finds Yule's model the

Collection	Datasets	Abbr		Size
TREC datasets				
			# documents	
TIPSTER	Wall Street Journal (1987–1992)	wsj	173,252	686M
	Federal Register (1988–1989)	fr	45,820	485M
	Associated Press (1988–1990)	ap	242,918	995M
	Department of Energy abstracts	doe	226,087	299M
	Computer Select disks (1989–1992)	ziff	293,121	919M
	San Jose Mercury News (1991)	sjm	90,257	365M
	U.S. Patents (1983-1991)	patents	6,711	216M
Total			1,078,166	3,965M
			# documents	
TREC 4 & 5	Financial Times Limited (1991–1994)	ft	210,158	766M
	Congressional Record of the 103rd Congress (1993)	cr	27,922	277M
	Federal Register (1994)	fr94	55,630	315M
	Foreign Broadcast Information Service (1996)	fbis	130,471	576M
	Los Angeles Times (1989–1990)	latimes	131,896	596M
Total			556,077	2,530M
			# documents	
AQUAINT	New York Times (1999–2000)	nyt	314,452	2,113M
	Associated Press (1999–2000)	apw	239,576	939M
	Xinhua News Agency (1996–2000)	xie	479,433	975M
Total			1,033,461	4,027M
			# documents	
AQUAINT-2	Agence France Presse (2004–2006)	apf_eng	296,967	889M
	Central News Agency (Taiwan) (2004–2006)	cna_eng	19,782	50M
	Xinhua News Agency (2004–2006)	xin_eng	170,228	360M
	Los Angeles Times-Washington Post (2004–2006)	ltw_eng	65,713	384M
	New York Times (2004–2006)	nyt_eng	159,400	1,100M
	Associated Press (2004–2006)	apw_eng	194,687	666M
Total			906,777	3,449G
			# documents	
ClueWeb cat. B.	Websites in English (2009)	CW09	49,2M	169G
	Websites in English (2012)	CW12	52,3M	180G
Total			101,5M	349G
Non-TREC datasets				
			# documents	
iSearch [2010]	Physics articles, metadata and book recods	is_full	453,254	4,179G
			# citations	
	Citations	is_cit	3,768,410	102M
Total			4,221,664	4,281G
			# queries	
Excite	Commercial query log (1999)	excite	2,5M	115M
Microsoft	Commercial query log (2006)	MSN	14,9M	1,1G
			# n-gram	
Google books [2013]	Syntactic unigrams	books	13,6M	3,6G

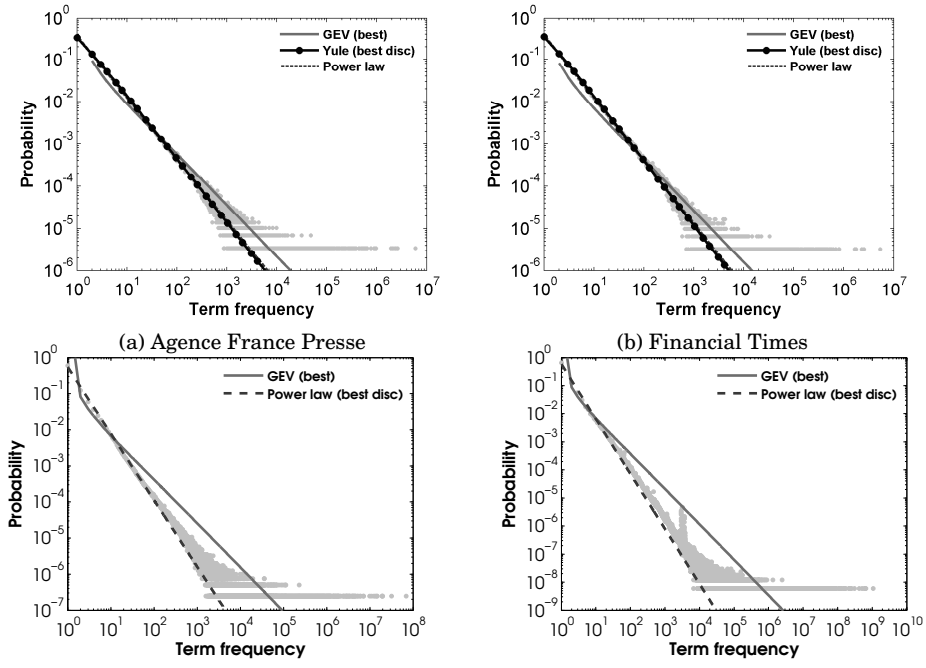
Table 7: Datasets used. For more detailed descriptions see Appendix N

best-fitting, where AICc finds the power law to be best-fitting. In Fig. 3b, we see that distinguishing between the GEV, Yule or power law is difficult, which may explain why Vuong's test and the AICc disagree. The difference in AICc values for the power law and Yule fit is around 500 indicating a near identical fit.[16]

[16]The difference in AICc values between Yule and the third-best-fitting model is over $700{,}000$.

Collections	Datasets	Term frequency	Document length	Query frequency	Query length	Citation frequency	Syn. n-gram frequency
TIPSTER	All datasets (see Table 7)	✓	✓				
TREC 4 & 5	All datasets (see Table 7)	✓	✓				
AQUAINT	All datasets (see Table 7)	✓	✓				
AQUAINT-2	All datasets (see Table 7)	✓	✓				
ClueWeb cat. B.	Websites in English (2009)	✓	✓				
	Websites in English (2012)	✓	✓				
iSearch [2010]	Full-text physics articles, metadata and book records	✓	✓				
	Citations					✓	
MSN	Commercial query log			✓	✓		
Excite	Commercial query log			✓	✓		
Google books	Syntactic unigrams						✓

Table 8: Data properties studied for each dataset.



(a) Agence France Presse (b) Financial Times

(c) iSearch (d) ClueWeb12 cat. B.

Fig. 3: Term frequency distribution for different datasets. We show the best-fitting (i) overall model (solid), (ii) discrete model (dot-solid) and (iii) power law (dashed). (i) & (ii) are the best-fits according to Vuong's test. The power law is plotted as reference.

Fig. 3 shows four term frequency distributions with the GEV, Yule and/or power law. These distributions resemble a cone: a narrow stem (for the most frequent terms) that fans out as terms become rarer. The best-fitting models approximate the distributions similarly; the Yule and power law, for example, are nearly indistinguishable as hinted by the small difference in AICc (see above), and while both models approximate the distribution below the $\sim$100 most frequent terms, they both fail to model the tail. The GEV underestimates the probability of the $\sim$100 most frequent terms, but approximates

Collections	Dataset	Discrete	Continuous
TIPSTER	Wall Street Journal	Yule (p=1.506)	Generalized Extreme Value (k=4.198,σ=0.819,μ=1.195)
	Federal Register	Yule (p=1.563)	Generalized Extreme Value (k=4.318,σ=0.79,μ=1.183)
	Associated Press	Yule (p=1.529)	Generalized Extreme Value (k=4.142,σ=0.6999,μ=1.169)
	Department of Energy	Yule (p=1.646)	Generalized Extreme Value (k=4.221,σ=0.7295,μ=1.173)
	Computer Select disks‡	Yule (p=1.597)	Generalized Extreme Value (k=3.819,σ=1.853,μ=1.484)
	San Jose Mercury News	Yule (p=1.515)	Generalized Extreme Value (k=4.229,σ=0.7439,μ=1.176)
	U.S. Patents‡	Yule (p=1.554)	Generalized Extreme Value (k=4.211,σ=0.7171,μ=1.17)
TREC 4 & 5	Financial Times Limited‡	Yule (p=1.536)	Generalized Extreme Value (k=4.139,σ=0.6858,μ=1.166)
	Congressional Record of the 103rd Congress	Power law (α=1.635)	Generalized Extreme Value (k=3.825,σ=0.5372,μ=1.14)
	Federal Register	Power law (α=1.631)	Generalized Extreme Value (k=4.117,σ=0.677,μ=1.164)
	Foreign Broadcast Information Service‡	Yule (p=1.604)	Generalized Extreme Value (k=3.796,σ=1.759,μ=1.462)
	Los Angeles Times	Yule (p=1.509)	Generalized Extreme Value (k=3.75,σ=1.216,μ=1.323)
AQUAINT	New York Times	Yule (p=1.498)	Generalized Extreme Value (k=3.936,σ=2.866,μ=1.727)
	Associated Press	Yule (p=1.475)	Generalized Extreme Value (k=3.991,σ=3.381,μ=1.846)
	Xinhua News Agency‡	Yule (p=1.618)	Generalized Extreme Value (k=4.193,σ=0.7286,μ=1.174)
AQUAINT-2	Agence France Presse	Yule (p=1.502)	Generalized Extreme Value (k=3.994,σ=3.296,μ=1.834)
	Central News Agency (Taiwan)	Yule (p=1.496)	Generalized Extreme Value (k=3.845,σ=2.585,μ=1.671)
	Xinhua News Agency	Yule (p=1.55)	Generalized Extreme Value (k=3.788,σ=1.622,μ=1.427)
	L.A. Times-Washington Post	Yule (p=1.503)	Generalized Extreme Value (k=3.941,σ=3.119,μ=1.79)
	New York Times‡	Yule (p=1.524)	Generalized Extreme Value (k=4.215,σ=0.7283,μ=1.173)
	Associated Press	Yule (p=1.503)	Generalized Extreme Value (k=3.832,σ=2.562,μ=1.667)
iSearch [2010]	Full-text physics articles, metadata and book records	Power law (α=1.844)	Generalized Extreme Value (k=4.126,σ=0.6834,μ=1.165)
ClueWeb cat. B.	Web pages in English (2009)	Power law (α=1.817)	Generalized Extreme Value (k=3.664,σ=0.4585,μ=1.125)
	Web pages in English (2012)	Power law (α=1.958)	Generalized Extreme Value (k=3.648,σ=0.442,μ=1.121)

Table 9: Best-fitting discrete and continuous models for term frequencies. ‡ indicates where Vuong's test and the AICc found different best-fitting models.

rarer terms better than both the Yule or power law. Fig. 3 also shows that as the number of unique terms (x-axis) increases, the GEV increasingly deviates from a straight line to approximate the extreme/rare events.

Overall, the distributions confirm what is already known: very few words have a very high frequency and many words have a very low frequency. This happens because most highly frequent words belong to a *closed grammatical class* (e.g., determiners, prepositions): no new words of that type can be produced in language (i.e. no new ways of producing equivalents to `the` or `of`), hence their frequency is boosted by their repeated use. Conversely, the numerous but infrequent words belong to an *open grammatical*

class (e.g. nouns, verbs): new words of that type can be produced in language (e.g. `to tweet`, `a googler`) analogously to the new meanings that emerge. Hence, the frequency of these types of words is "diluted" by varied usage, leading to a long tail. Collectively, the findings from the term frequency distribution in iSearch, ClueWeb09 cat. B., ClueWeb12 cat. B. Federal Register and the Congressional Records agree with Zipf's original findings. Note that we considered the probability distribution for term frequency where $P(X{=}n)$ means "what is the probability that an arbitrary term occurs n times in the dataset", whereas in Zipf's original findings $P_{\text{Zipf}}(Y{=}j)$ means "what is the probability that an arbitrary term is the j^{th} most frequent term", or equivalently "how many times does the j^{th} most frequent term occur in the dataset". For large datasets, we can obtain a Zipf-type power law from the probability distribution on terms [Bookstein 1990] as follows: If the probability $P(X{=}n)$ that an arbitrary term occurs satisfies $P(X{=}n){=}Cn^{-\alpha}$, then

$$P_{\text{Zipf}}(Y{=}j){=}f_{\max}\left(\frac{(\alpha{-}1)j}{f_{\max}}{+}1\right)^{-\frac{1}{(\alpha-1)}} \tag{7}$$

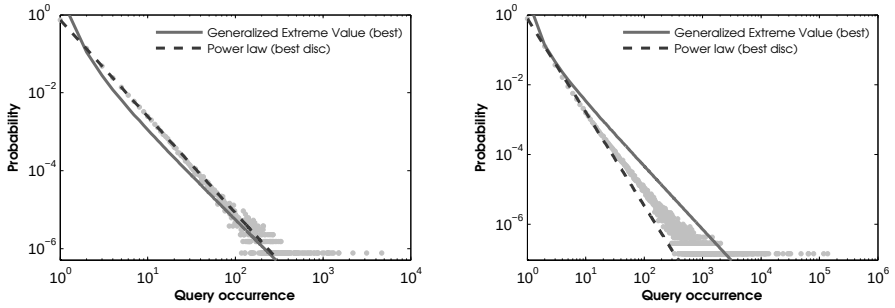
where $f_{\max}$ is the probability of the most frequent term (i.e., $j{=}1$) in the dataset.

4.3.2. Query Frequency. Table 10 shows that the GEV and power law are the best-fitting overall and discrete models to the distributions of query frequencies in the Excite and MSN datasets according to both Vuong's test and the AICc. All plots and tables pertaining to this property are found in Appendix J (KS-distance plots), Appendix I (Vuong tables) and Appendix M (distribution plots). Fig. 4 shows the query frequency distributions from both datasets. For the Excite dataset (Fig. 4a), the power law approximates the distribution of query frequencies well until the query occurrence is approximately 10^2. Above 10^2 data sparsity gives rise to a tail that the power law cannot approximate. The GEV fits the data similarly, but underestimates the probability of query occurrences in the interval 10^0 to 10^2. Similarly to the power law, the GEV cannot approximate the tail. For the MSN dataset (Fig. 4b), the power law approximates the query frequency distribution until approximately $10^{1.5}$, after which the probability of query occurrence is consistently underestimated. In contrast, the GEV overestimates the probability of query occurrence in the same interval, but unlike the power law, it does approximate the tail. That the GEV can approximate the tail for the MSN and not the Excite dataset, is likely caused by the former having more "extreme" outliers and an overall higher density of infrequent query occurrences. Our results cooperate the findings of e.g. Baeza-Yates and Saint-Jean [2003], Baeza-Yates and Tiberi [2007] and Baeza-Yates et al. [2007], and also suggest that the upper bound derived in Ding et al. [2011] is reasonable (Section 3.2).

Collections	Dataset	Discrete	Continuous
Query frequency	Excite	Power law (α=2.474)	Generalized Extreme Value (k=0.815,σ=0.152,μ=1.076)
	MSN	Power law (α=2.613)	Generalized Extreme Value (k=1.241,σ=0.205,μ=1.106)

Table 10: Best-fitting discrete and continuous models for query frequency.

4.3.3. Query Length. Table 11 shows that the Negative Binomial and Inverse Gaussian are the best-fitting discrete and overall models to the distributions of query lengths in the Excite and MSN data according to both Vuong's test and the AICc. All plots and tables pertaining to this property are found in Appendix L (KS-distance plots), Appendix K (Vuong tables) and Appendix M (distribution plots).



(a) Excite (b) MSN

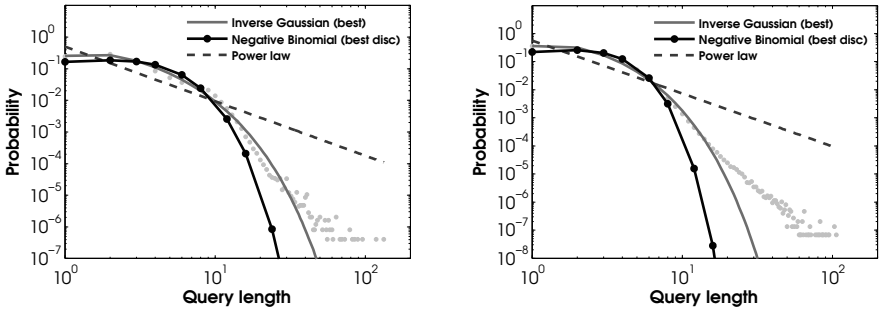
Fig. 4: Query frequency distribution for the Excite and MSN logs. We show the best-fitting (i) overall model (solid), (ii) discrete model (dot-solid) and (iii) power law (dashed). The first two models are the best-fitting according to Vuong's test. The power law is plotted as reference.

Fig. 5 shows the distribution of query lengths from both datasets. We see that the distributions in both datasets are poorly approximated by a power law across the entire data range, whereas the distributions of smaller query lengths are well-approximated by the Negative Binomial and Inverse Gaussian. For the Excite data, the Negative Binomial approximates queries of length $n \lesssim 10$ well, whereas the Inverse Gaussian can well-approximate query lengths up to $n \approx 30$. Neither model provides a good approximation to the tail (queries above $n \approx 40$). For the MSN data, both models approximate query lengths up to $n \approx 10$ (the inverse Gaussian up to $n \approx 20$) well, but neither model provides a good approximation to the tail. Our results agree with Arampatzis and Kamps [2008] who found that a mixed model of the power law and Poisson gave better fits to the tail, albeit at the cost of calculating a cutoff point where one should switch between the two distributions. For empirical distributions with several modes, such as the iSearch document length distribution (shown in Fig. 6c), several such points may require estimation (see Section 4.2).

Collections	Dataset	Discrete	Continuous
Query length	Excite	Negative Binomial (r=4.238,p=0.556)	Inverse Gaussian (μ=3.249,λ=5.274)
	MSN	Negative Binomial (r=43.671,p=0.948)	Inverse Gaussian (μ=2.401,λ=5.598)

Table 11: Best-fitting discrete and continuous models for query lengths.

4.3.4. Document Length. Table 12 shows the best-fitting overall and the best-fitting *discrete* model according to Vuong's test. All plots and tables pertaining to this property are found in Appendix D (KS-distance plots), Appendix C (Vuong tables) and Appendix M (distribution plots). For eight datasets, the best-fitting models selected by Vuong's test and the AICc were different. Similarly to the term frequencies, the disagreement is between models whose AICc values are relatively close to each other (or nested in which case we favour the most parsimonious model). For example, for the Associated Press dataset, the best fit is the Rayleigh model according to Vuong's test and the Nakagami



(a) Excite (b) MSN

Fig. 5: Distribution of query lengths for the Excite and MSN datasets. We show the best-fitting (i) overall model (solid), (ii) discrete model (dot-solid) and (iii) power law (dashed). The first two models are the best-fitting according to Vuong's test. The power law is plotted as reference.

model according to the AICc. The difference in AICc values is $\approx$17, suggesting that both models fit the data equally well.[17]

Gamma was the best-fitting continuous model overall in most cases. The Gamma distribution is used to model the time required for a events to occur, given that the events occur randomly in a Poisson process with a mean time between events of β. For document length, this means that if we know that a document of length n occurs in an incoming stream of documents, on average, every m^{th} seconds, the Gamma distribution models the number of seconds before the next document of length n appears in the stream. For datasets where document lengths are *not* well fitted by the Gamma, the log-normal or GEV provide the best fit for most datasets. From the set of discrete models, the Negative Binomial or Geometric are the best fits for all datasets.

Fig. 6 shows the distribution of document length for four datasets. These distributions deviate substantially from a straight line. All datasets have one or more local maxima before the distribution drops off sharply. Before the peak, the probability of a document of length n increases with larger n. The iSearch dataset has three clear local maxima due to its heavily curated nature and to the standardised document lengths of scientific publication environments: iSearch is a collection of full length articles, their metadata and short bibliographic records of books in physics. These three types of data correspond to the three peaks in Fig. 6c. The U.S. Patents dataset (Fig. 6d) is the smallest of the datasets (6711 documents) and shows two distinct local maxima at $n\approx100$ and $n\approx3000$. We attribute this bi-modality to the small number of documents combined with the substantial differences in lengths (90% of the documents are above length n=1519).

For all datasets, the power law fails to approximate the general shape of the empirical distribution. For the Xinhua dataset (Fig. 6a), both the Gamma and Negative Binomial approximate the distribution for the same range of data, differing only in their approximation for short ($n\lesssim30$) documents where both models overestimate the probability of document lengths. In the tail of the distribution ($n\gtrsim900$), both models are visually indistinguishable and both underestimate the probability of larger documents.

For ClueWeb12 cat. B., document lengths in the range $30\lesssim n\lesssim900$ are well approximated by both the best-fitting overall model – the Generalized Pareto – and the best discrete model – the Negative Binomial. Neither model fits document lengths above

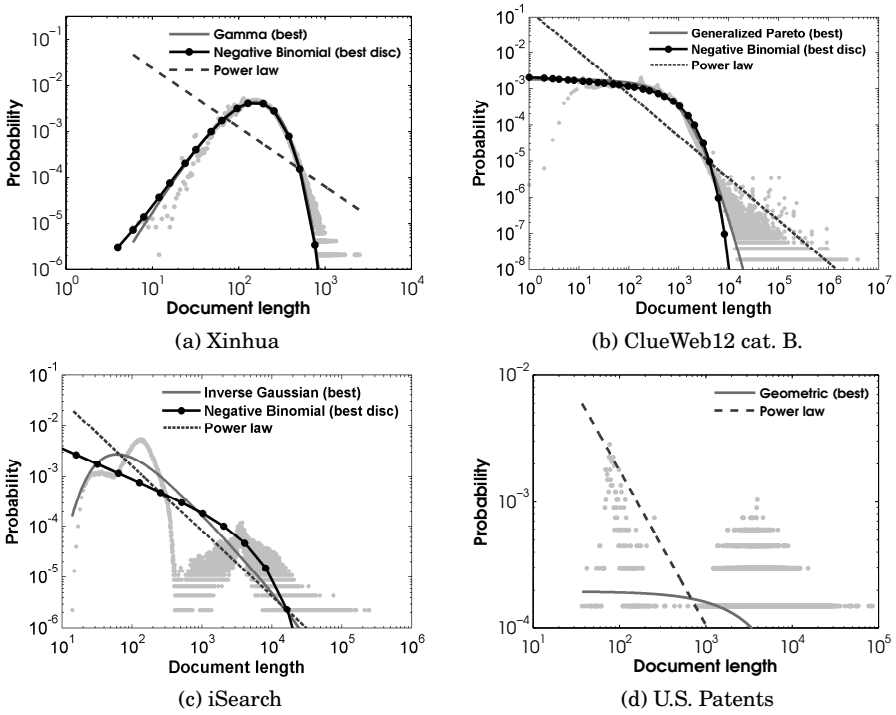
[17]An AICc value of 17 is substantially smaller than 500 found for term frequencies. However, these numbers are *relative* as the number of terms is substantially larger than the number of documents.

Collections	Dataset	Discrete	Continuous
TIPSTER	Wall Street Journal‡	Geometric (p:0.00223)	Inverse Gaussian (λ:258,μ:447.5)
	Federal Register	Negative Binomial (r:0.5986,p:0.0004)	Generalized Extreme Value (k:0.7565,σ:330.6,μ:324)
	Associated Press‡	Negative Binomial (r:3.137,p:0.0067)	Rayleigh (b:370)
	Department of Energy‡	Negative Binomial (r:3.989,p:0.03)	Rayleigh (b:99.24)
	Computer Select disks‡	Negative Binomial (r:0.7491,p:0.002)	Log-normal (μ:5.098,σ^2:1.221)
	San Jose Mercury News	Geometric (p:0.002442)	Exponential (μ:408.5)
	U.S. Patents‡	Geometric (p:0.0002)	—
TREC 4 & 5	Financial Times Limited	Negative Binomial (r:1.469,p:0.0035)	Gamma (a:1.462,b:273.4)
	Congressional Record of the 103rd Congress	Negative Binomial (r:0.4669,p:0.0003)	Generalized Extreme Value (k:0.8658,σ:219.1,μ:215.6)
	Federal Register	Negative Binomial (r:2.404,p:0.0035)	Log-normal (μ:6.287,σ^2:0.6704)
	Foreign Broadcast Information Service‡	Geometric (p:0.002)	Generalized Extreme Value (k:0.5833,σ:164.7,μ:214)
	Los Angeles Times‡	Geometric (p:0.002)	Gamma (a:1.035,b:485.1)
AQUAINT	New York Times‡	Negative Binomial (r:2.017,p:0.0025)	Nakagami (μ:0.6925,w:88712)
	Associated Press	Negative Binomial (r:2.626,p:0.0061)	—
	Xinhua News Agency	Negative Binomial (r:4.146,p:0.019)	Gamma (a:4.055,b:51.09)
AQUAINT-2	Agence France Presse	Negative Binomial (r:2.039,p:0.006)	Gamma (a:2.021,b:166.1)
	Central News Agency (Taiwan)	Negative Binomial (r:2.688,p:0.01)	Gamma (a:2.653,b:97.19)
	Xinhua News Agency	Negative Binomial (r:2.945,p:0.013)	Gamma (a:2.897,b:76.55)
	L.A. Times-Washington Post	Negative Binomial (r:1.715,p:0.0025)	Nakagami (μ:0.5993,w:70213)
	New York Times	Negative Binomial (r:3.189,p:0.004)	—
	Associated Press	Negative Binomial (r:2.437,p:0.006)	Gamma (a:2.416,b:160.3)
iSearch [2010]	Full-text physics articles, metadata and book records	Negative Binomial (r:0.3981,p:0.0001)	Inverse Gaussian (λ:183.2,μ:2261)
ClueWeb cat. B.	Web pages in English (2009)	Negative Binomial (r:1.14,p:0.0014)	Log-normal (μ:6.183,σ^2:0.9606)
	Web pages in English (2012)	Negative Binomial (r:0.8921,p:0.0011)	Generalized Pareto (k:0.2121,σ:558.6,θ:0)

Table 12: Best-fitting discrete and continuous models for document lengths. '–' in the Continuous column means the best-fitting model is the one listed in the Discrete column. ‡ indicate datasets where Vuong's test and the AICc found different best-fitting models.

$n{\approx}1000$ however, and both models overestimate the probability of small documents. We make a similar observation with the U.S. Patents dataset where the best-fitting Geometric distribution (and power law) fail to model any range of the data. Together, the iSearch and U.S. Patents datasets suggest that *before* fitting any standard statistical distribution to data, one should look at its shape.

4.3.5. Citation Frequency & Syntactic Unigrams. Table 13 shows the best-fitting overall and discrete models for the citations and syntactic unigrams datasets according to both Vuong's test and the AICc. All plots and tables pertaining to this property are found in Appendix H (KS-distance plots), Appendix G (Vuong tables) and Appendix M



(a) Xinhua (b) ClueWeb12 cat. B.

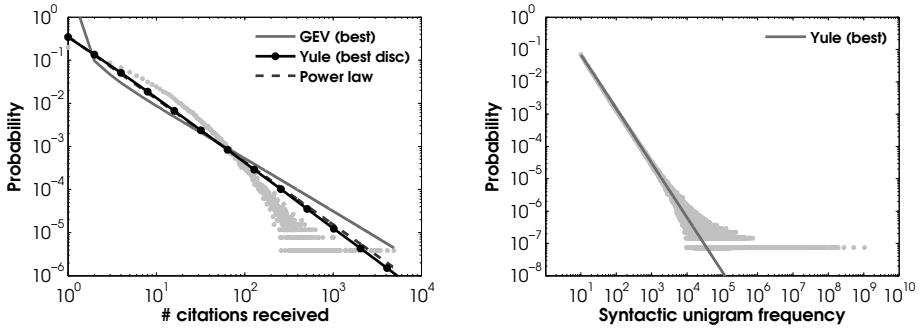
(c) iSearch (d) U.S. Patents

Fig. 6: Distribution of document lengths for different datasets. We show the best-fitting (i) overall model (solid), (ii) discrete model (dot-solid) and (iii) power law (dashed). The first two models are the best-fitting according to Vuong's test. The power law is plotted as reference.

Collections	Datasets	Discrete	Continuous
iSearch	Citations	Yule (p=1.518)	Generalized Extreme Value (k=3.761,σ=2.098,μ=1.556)
Google books	Syntactic unigrams	Yule (p=1.683)	—

Table 13: Best-fitting discrete and continuous models for citations and syntactic unigrams. For syntactic unigrams, the Yule model was the best-fitting model overall.

(distribution plots). The discrete model that best approximates the citation frequency distribution is Yule (Fig. 7a), which means that the tail of the distribution is a realisation of Zipf's law. However, the Yule distribution deviates from a pure power law; in our case, the parameter p=1.518 means that the probability of a paper receiving many citations decreases faster than a power law would suggest. The GEV – the best-fitting model – and Yule both fail to approximate the general shape of the distribution. Fig. 7b shows the distribution of syntactic unigrams in the Google Books dataset. This distribution visually resembles the cone-shapes of the term frequency distributions (Section 4.3.1), and is best fitted by the discrete Yule, though the model fails to approximate the tail of the distribution. Both the GEV and power law also fit the distribution well according to both Vuong's test and AICc.



(a) iSearch citations (b) Syntactic unigrams

Fig. 7: Distribution of citations in iSearch and syntactic unigrams in Google books. We show the best-fitting (i) overall model (solid), (ii) discrete model (dot-solid) and (iii) power law (dashed). The first two models are the best-fitting according to Vuong's test. The power law is plotted as reference.

4.4. Failure Analysis: Model Misestimation and Fits

To assess the deviation of the models we fitted from the data, we use the Kolmogorov-Smirnov (KS) distance (Eqn. 9) that measures the maximum difference between the cumulative distribution of the data and the fitted model, and hence may be interpreted as the maximum *mis*estimate that the fitted model does when predicting the probability $P(X \leq n)$ of the data having *at most* n occurrences of the property. Let the empirical cumulative distribution function (ECDF) of a sample $x=\{x_1,...,x_n\}$ be given by

$$F_{\text{emp}}(x)=\frac{1}{n}\sum_{i=1}^{n}1_{x_i \leq x} \tag{8}$$

where $1_{x_i \leq x}$ is an indicator function emitting 1 if $x_i \leq x$ and 0 otherwise. Let F be the cumulative distribution function (CDF) of $f(\cdot|\hat{\theta})$, that is, $F(x)$ is the probability that the underlying stochastic variable X has value *at most* x. The KS distance between the ECDF and the CDF of the fitted model with MLE $\hat{\theta}$ is then

$$\mathrm{D}_{\{\mathrm{x}_1,\ldots,\mathrm{x}_\mathrm{n}\}}=\sup_x |F_{\text{emp}}(x)-F(x)| \tag{9}$$

where $\sup$ is the supremum and x ranges over the domain of definition of both F_{emp} and F. If F is the CDF (i.e., the cumulative mass function) of a discrete model with support exactly on the positive integers, the KS distance is [Arnold and Emerson 2011]:

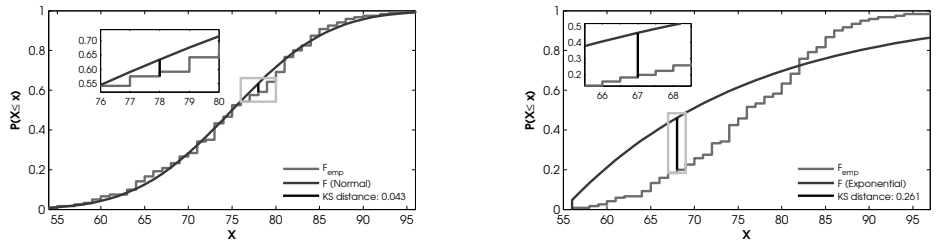
$$\mathrm{D}_{\{\mathrm{x}_1,\ldots,\mathrm{x}_\mathrm{n}\}}=\max_{1\leq i\leq n}\left(|F(x_i)-F_{\text{emp}}(x_i)|,\, |F(x_i-1)-F_{\text{emp}}(x_{i-1})|\right) \tag{10}$$

where we set $F(x_{-1})=0$. If F is the CDF of a continuous model with support on the positive reals, then

$$\mathrm{D}_{\{\mathrm{x}_1,\ldots,\mathrm{x}_\mathrm{n}\}}=\max_{1\leq i\leq n}\left(\max(|F(x_i)-F_{\text{emp}}(x_i)|,\, |F(x_i)-F_{\text{emp}}(x_{i-1})|)\right) \tag{11}$$

where we set $|F(x_1)-F_{\text{emp}}(x_{-1})|=0$. As the KS-distance of the data and a fitted model is a number between 0 and 1, it can be interpreted as the maximum difference in probability between the cumulative probability distribution of the fitted model and the data: roughly, if the fitted distribution "predicts" the percentage of data below some threshold x, the KS-distance is the maximum deviation (for all x) of this prediction from the actual data. As an example, Fig. 8 shows the ECDF of a random set of 120 twitter messages (their lengths) and two MLE fitted models (Gaussian and Exponential). The

Gaussian distribution (Fig. 8a) qualitatively approximates the data well evidenced by a KS-distance of 0.043. In contrast, the Exponential (Fig. 8b) does not provide as good an approximation with a KS-distance of 0.261.



(a) Gaussian (b) Exponential

Fig. 8: Examples of KS-distances for the lengths of 120 randomly selected twitter messages (step function). MLE fitted models are shown as smooth curves. Insets zoom in on the KS-distance (vertical bar). (8a) An MLE fitted normal distribution with KS-distance$=0.043$. (8b) An MLE fitted Exponential distribution with KS-distance$=0.261$.

KS-distances for all datasets and models are found in Appendix O. We now summarise the results. The GEV, Generalized Pareto and Yule typically have the lowest KS-distance of all models. These findings generally support our results from term frequencies where these models were also the best fits according to both Vuong's test and the AICc. However, as the KS-distance does not penalise models for having many parameters, the two former models (which both have three parameters) may have an advantage. We observe that in some cases the model with the smallest KS-distance is not the same as the one found using statistical analysis. E.g., while the power law was the best fit as per AICc and Vuong's test for the term frequency distribution of ClueWeb09 cat. B., the KS distance is 0.441. This indicates that for some n, the probability that an arbitrary term has frequency at most n is misestimated by the power law by more than 44%, in contrast to, e.g., the Yule that at most misestimates by 12.4%. One feature of the KS distance is that very rare events contribute little to the cumulative probability mass. Thus, if the KS distance is small, the *absolute* misestimation Δp of the probability for rare events will be small, but the *relative* misestimation $\Delta p/p$ of a rare event may be very large. Thus, a model may have very low KS distance, and seemingly fit very well for frequent events, but fail to adequately predict rare events – a phenomenon avoided by using MLE where the best-fitting models may have larger KS distance in this case. E.g., for the iSearch citations (Fig. 9), where the most frequent event is that a given paper is not cited very much, the best-fitting models are GEV (continuous) and Yule (discrete). Both models fail to fit the "shape" of the data, but do not make very large misestimations for very rare events; in contrast, the Weibull model has the lowest KS distance to the data of all models, and has a better fit to the shape –for papers that are not cited much– but the relative error in probability becomes enormous for highly cited papers.

4.5. Summary findings

Our results (Section 4.3 and Section 4.4) show that power law models describe the data better than other standard distributions for five term and two query frequency datasets. In all remaining datasets, other distributions such as the Yule, Generalized Extreme Value or Negative Binomial approximate the empirical distribution of term frequencies, document lengths, query lengths, citation frequency and syntactic unigram frequencies

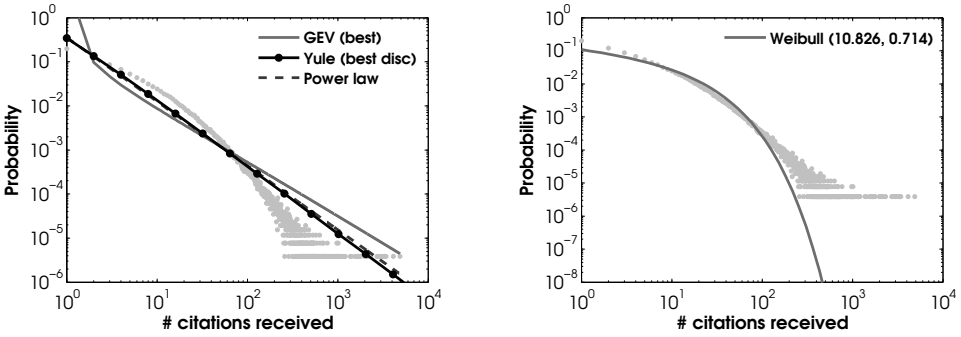


Fig. 9: Citations with PDF/PMF of the best-fitting models. Fig. 9a is identical to Fig. 7a. Fig. 9b: Citations with the Weibull model, which captures the shape of the distribution for little-cited papers, but misestimates the probability for highly cited papers. The Weibull PDF continues downwards: while the frequency of the rightmost data point is $e^{-12,46}$, Weibull predicts it to be $e^{-83.26}$; if drawn, this would be halfway down the page.

better than a power law. We also found (Section 4.4) that the fitted Generalized Extreme Value, Generalized Pareto and Yule models tend to have the smallest deviations from these empirical distributions.

5. COMPUTATIONAL COST AND CONSEQUENCES TO OPERATIONAL IR SYSTEMS

In our first research question, we investigated the extent to which six different data properties of IR datasets were approximately power law distributed using a statistically principled approach. We now look at computational costs of this approach. Specifically, we measure the *elapsed* time required to fit each model from Table 6 to each data property per dataset in Table 7. The results will indicate if more accurate distribution fitting techniques, like the one we use, can be efficient alternatives to the graphical methods commonly used.

To answer our second research question, we fit each data property to all 16 models three times and report the median time. We use MATLAB's `tic` timer to measure the elapsed time in seconds. All experiments are done on an unladen server with a single Intel Core i7 CPU with 32GB memory running Ubuntu Linux 13.10 and MATLAB R2013B on a 256GB OCZ Agility 3 SSD.

All results can be found in Appendix B, and here we give a summary. We find that each model can generally be fitted to a dataset property in, *at most*, a few seconds. For several models, this is expected, as many of the them have closed form solutions for parameter estimation, but also models with no closed form solution (such as the Negative Binomial and the power law) are fitted in seconds. The one exception is the GEV, which consistently takes substantially longer time than all other models. This is because we use a different optimiser for the MLE, which requires explicit calculation of the gradients. We use an optimiser from OCTAVE, as MATLABs default causes the GEV to converge to a *boundary point* in parameter space. When at a boundary point, the asymptotic normality of the MLE is violated and likelihood ratio tests must be corrected [Kopylev 2012]. Aside from using faster machinery, faster optimisers exist, e.g. in R, that can drastically reduce the fitting time for GEV. While this increase in time typically means only a few minutes instead of seconds, for larger datasets (such as ClueWeb 12 cat. B.), fitting the GEV takes close to 14 hours – roughly double the time of ClueWeb09 cat. B. likely caused by the doubling in number of terms ($\approx 168M$ for ClueWeb12 and $\approx 94M$ ClueWeb09), despite both datasets having approximately the same number of documents. All remaining models also take longer time to estimate

as the size of the dataset grows. For most datasets this is only a few seconds, bar the largest datasets (ClueWeb09/ClueWeb12 cat. B. and Google books) where model fitting takes anywhere from a second (using the Geometric Model) to 14 hours.

In summary, we find that methods such as ours for accurate statistical model fitting can take, on average, several minutes for the type of large datasets used in IR. However, as the size of the datasets grows, some models, in particular the GEV, can take several hours to fit. This can be remedied by using optimised solvers and faster hardware.

6. DISCUSSION

We now discuss each data property in turn, highlighting areas in IR where our findings support existing research/knowledge. This does not invalidate claims made in previous research, but may encourage researchers to consider distributions other than the power law for their purpose.

6.1. Data Properties

Table 14 summarises our findings for our first research question, which show that for most datasets, the data properties we studied, except term frequency and query frequency, are *not* (approximately) power law distributed. Term frequency distributions have long been claimed to be power law distributed. However, we find an approximate power law for only 5 of 24 datasets (iSearch, ClueWeb09 cat. B., ClueWeb12 cat. B, Congressional Record, Federal Register). For these datasets, computations assuming that Zipf's law holds, such as computation of the space requirements for an inverted index compressed with any standard encoding such as γ encoding [Manning et al. 2008, Ch. 5], will be more accurate than if other models, for example the Poisson model, are used. All remaining term frequency distributions are better approximated by the discrete Yule and continuous GEV. As the tail of the Yule model is a realisation of Zipf's law, our findings lend support to those of e.g. Chau et al. [2009], Chaudhuri et al. [2007], Momtazi and Klakow [2010], but because the term frequency distribution is sparse in the tail (examples in Fig. 3) the Yule fails to approximate rare terms; something the GEV is better at, which likely explains why it was the better overall model.

The distributions of document lengths often contained one or more clear local maxima with a non-symmetric spread of data around them (Fig. 6). This makes the power law a poor choice for approximating such distributions. Our results show that multi-modal models such as the Negative Binomial, Gamma, GEV or log-normal are consistently better-fitting models, making our findings at odds with Sato and Nakagawa [2010] and Srivastava et al. [2013].

Term Frequency	Document Length	Query Frequency	Query Length	Citations Received	Syntactic Unigrams
5/24	0/24	2/2	0/2	0/1	0/1

Table 14: Number of approximate power laws found for each data property per dataset.

Similarly to Baeza-Yates et al. [2007] and Ding et al. [2011], we find query frequencies approximated by a power law (Fig. 4). This indicates that the increase in expected accuracy for top-1 documents obtained by Asthana et al. [2011] is reasonable and that the upper bound of query frequency received by a search engine derived by Ding et al. [2011] may be correct.

The distributions of query lengths from both query logs show that neither are well-approximated by a power law (Fig. 5) and instead, the Inverse Gaussian and Negative Binomial are the best-fits. This agrees with Arampatzis and Kamps [2008].

The citation frequency distribution, like the term frequency distributions, is best fitted by the GEV and Yule (Fig. 7a), where the GEV better approximates highly cited papers. The Yule suggests that the distribution is the result of a preferential attachment: highly cited articles will continue to be cited. This agrees with Meij and de Rijke [2007] who found a preferential attachment prior of received citations to outperform a "scale-free" prior. The GEV has, to the best of our knowledge, not been used for this purpose.

The Yule is the overall best-fitting model for the syntactic unigrams; this is in contrast to our findings of term frequencies where the GEV was the best fit. One explanation is that the tail of the distribution is not sufficiently "heavy" for the GEV to be favoured.[18] Comparing to ClueWeb12 cat. B. for example, the number of terms with frequency $x{>}1.000.000$ in the syntactic unigrams dataset is less than half. Another explanation is the difference in domain (websites for ClueWeb12 cat. B. and UK books for syntactic unigrams). Finally, the preferential attachment process implied by the Yule model could mean that existing words are frequently used to describe "new" words entering the vocabulary.

Cost of Obtaining Better Fits. The time required to fit most models was a few seconds, which increased to a few minutes as the size of the datasets grew. However, some combinations of model and dataset took substantially longer time, i.e. fitting the GEV to the ClueWeb datasets took several hours as a different optimiser was used. Faster and more powerful computers will reduce the time required to fit models to large-scale datasets (up to 400% using a dedicated dual-CPU Xeon server in our experiments).

Discretisation. Several data properties were best-fitted by a continuous model, despite the data being discrete. Thus the best-fitting continuous models must be discretised, which carries the risk of potential loss of statistical accuracy [Chen and Pollino 2012]. Commonly used discretisation techniques (such as equal-width or equal-frequency intervals – see e.g. Hammond and Bickel [2013] or Muhlenbach and Rakotomalala [2005]) may generate inappropriate division of the data as variations within intervals are ignored, and care must be taken if statistical properties of the original fitted distribution, such as its moments, are to be preserved.

7. CONCLUSIONS

Many properties of IR data are thought to be approximately distributed according to power laws. Motivated by recent work in the statistical treatment of power law claims, we investigated two research questions: (1) To what extent do power law approximations hold for term frequency, document length, query frequency, query length, citation frequency, and syntactic unigram frequency? (2) What is the computational cost of replacing empirical power law approximations with more accurate distribution fitting?

To answer our first research question, we systematically compared the fit of the power law to the fits of 15 other standard probability distributions on 28 datasets of different sizes (from 7,000 to 52M data points). We found that a power law best approximated the distribution of query frequency and the distribution of 5 out of 24 term frequency datasets. All remaining term frequency distributions and data properties (document length, query length, citation frequency and syntactic unigram frequency), were better approximated by the Inverse Gaussian, Generalized Extreme Value, Negative Binomial or Yule. To answer our second research question, we measured the time required to fit each model to each data property three times. We found that the median time taken for most models is a few seconds and, overall, the time taken is low enough for the approach to be usable.

[18]As Spierdijk and Voorneveld [2009] state in their investigation of the application of Yule's distribution to model gold record sales by U.S. artists: "... the Yule distribution captures stardom, but not superstardom".

Our findings show that researchers should be wary of making, or repeating, claims of approximate power laws. As most models can be accurately fitted to most datasets in a few seconds, we suggest that such methods be used instead of ad hoc techniques. Using accurate fitting techniques may benefit IR tasks such as predicting the probability of unseen words in language modelling, index compression, or test collection generation.

In the future, we intend to work towards discretising the best-fitting continuous models we found, as they fitted the data better than discrete models in most cases, and to investigate practical applications of our findings to IR tasks where knowing the correct distribution can impact performance, for instance with social media data.

REFERENCES

Lada A Adamic, Rajan M Lukose, Amit R Puniyani, and Bernardo A Huberman. 2001. Search in power-law networks. *Physical review E* 64, 4 (2001), 046135.

Hirotugu Akaike. 1974. A new look at the statistical model identification. *IEEE Trans. Automat. Control* 19, 6 (1974), 716–723.

Avi Arampatzis and Jaap Kamps. 2008. A study of query length. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 811–812.

Avi Arampatzis and Jaap Kamps. 2009. A signal-to-noise approach to score normalization. In *Proceedings of the 18th ACM conference on Information and knowledge management*. ACM, 797–806.

Taylor B Arnold and John W Emerson. 2011. Nonparametric goodness-of-fit tests for discrete null distributions. *The R Journal* 3, 2 (2011), 34–39.

Harshvardhan Asthana, Ruoxun Fu, and Ingemar J Cox. 2011. On the feasibility of unstructured peer-to-peer information retrieval. In *Advances in Information Retrieval Theory*. Springer, 125–138.

Leif Azzopardi. 2009. Query Side Evaluation: An Empirical Analysis of Effectiveness and Effort. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '09)*. ACM, 556–563.

Harald Baayen. 2001. *Word frequency distributions*. Springer.

Rohit Babbar, Ioannis Partalas, Eric Gaussier, and Massih-reza Amini. 2014. Re-ranking Approach to Classification in Large-scale Power-law Distributed Category Systems. In *Proceedings of the 37th International ACM SIGIR Conference on Research (SIGIR 2014)*. ACM, 1059–1062.

David F Babbel, Vincent J Strickler, and Ricki S Dolan. 2009. Statistical String Theory for Courts: If the Data Don't Fit. *Legal Technology Risk Management* 4 (2009), 1.

Ricardo Baeza-Yates, Aristides Gionis, Flavio Junqueira, Vanessa Murdock, Vassilis Plachouras, and Fabrizio Silvestri. 2007. The impact of caching on search engines. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 183–190.

Ricardo Baeza-Yates, Javier Ruiz-del Solar, Rodrigo Verschae, Carlos Castillo, and Carlos Hurtado. 2004. Content-based image retrieval and characterization on specific web collections. In *Image and Video Retrieval*. Springer, 189–198.

Ricardo Baeza-Yates and Felipe Saint-Jean. 2003. A three level search engine index based in query log distribution. In *String Processing and Information Retrieval*. Springer, 56–65.

Ricardo Baeza-Yates and Alessandro Tiberi. 2007. Extracting semantic relations from query logs. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 76–85.

Albert-László Barabási, Réka Albert, and Hawoong Jeong. 1999. Mean-field theory for scale-free random networks. *Physica A: Statistical Mechanics and its Applications* 272, 1 (1999), 173–187.

Heiko Bauke. 2007. Parameter estimation for power-law distributions by maximum likelihood methods. *The European Physical Journal B-Condensed Matter and Complex Systems* 58, 2 (2007), 167–173.

Michael A. Bean. 2001. *Probability: the science of uncertainty with applications to investments, insurance, and engineering*. Vol. 6. American Mathematical Soc.

Luca Becchetti and Carlos Castillo. 2006. The distribution of PageRank follows a power-law only for particular values of the damping factor. In *Proceedings of the 15th international conference on World Wide Web*. ACM, 941–942.

Casper Beckman. 1999. Chinese character frequencies. http://casper.beckman.uiuc.edu/~c-tsai4/chinese/charfreq.html. (1999). No longer available.

Jan Beirlant, Yuri Goegebeur, Johan Segers, and Jozef Teugels. 2006. *Statistics of extremes: theory and applications*. John Wiley & Sons.

Andras A Benczur, Karoly Csalogany, Tamas Sarlos, and Mate Uher. 2005. SpamRank–Fully Automatic Link Spam Detection Work in progress. In *Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web*.

Kerstin Bischoff, Claudiu S Firan, Wolfgang Nejdl, and Raluca Paiu. 2008. Can all tags be used for search?. In *Proceedings of the 17th ACM conference on Information and knowledge management*. ACM, 193–202.

Paolo Bolettieri, Andrea Esuli, Fabrizio Falchi, Claudio Lucchese, Raffaele Perego, Tommaso Piccioli, and Fausto Rabitti. 2009. CoPhIR: a test collection for content-based image retrieval. *arXiv preprint arXiv:0905.4627* (2009).

Abraham Bookstein. 1990. Informetric distributions, part I: Unified overview. *American Society for Information Science* 41, 5 (1990), 368–375.

George EP Box and David R Cox. 1964. An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)* (1964), 211–252.

Lee Breslau, Pei Cao, Li Fan, Graham Phillips, and Scott Shenker. 1999. Web caching and Zipf-like distributions: Evidence and implications. In *INFOCOM'99. Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, Vol. 1. IEEE, 126–134.

Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. 2000. Graph structure in the web. *Computer networks* 33, 1 (2000), 309–320.

Mark Buchanan. 2004. Power Laws & the New Science of Complexity Management. *Strategy+ Business* 34 (2004), 1–8.

Kenneth P Burnham and David R Anderson. 2002. *Model selection and multimodel inference: a practical information-theoretic approach*. Springer.

Pedro Cano, Oscar Celma, Markus Koppenberger, and Javier M Buldu. 2006. Topology of music recommendation networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science* 16, 1 (2006), 013107.

Domenico Cantone, Salvatore Cristofaro, Simone Faro, and Emanuele Giaquinta. 2009. Finite State Models for the Generation of Large Corpora of Natural Language Texts. In *Proceedings of the 7th International Workshop on Finite-state Methods and Natural Language Processing*, Vol. 191. IOS Press, 175.

Huanhuan Cao, Daxin Jiang, Jian Pei, Qi He, Zhen Liao, Enhong Chen, and Hang Li. 2008. Context-aware query suggestion by mining click-through and session data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 875–883.

Deepayan Chakrabarti and Christos Faloutsos. 2006. Graph mining: Laws, generators, and algorithms. *ACM Computing Surveys (CSUR)* 38, 1 (2006), 2.

Michael Chau, Yan Lu, Xiao Fang, and Christopher C Yang. 2009. Characteristics of character usage in Chinese Web searching. *Information Processing & Management* 45, 1 (2009), 115–130.

Surajit Chaudhuri, Kenneth Church, Arnd Christian König, and Liying Sui. 2007. Heavy-tailed distributions and multi-keyword queries. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 663–670.

Serena H Chen and Carmel A Pollino. 2012. Good practice in Bayesian network modelling. *Environmental Modelling & Software* 37 (2012), 134–145.

Pasquale Cirillo. 2013. Are your data really Pareto distributed? *Physica A: Statistical Mechanics and its Applications* 392, 23 (2013), 5947–5962.

Kevin A Clarke. 2003. Nonparametric model discrimination in international relations. *Journal of Conflict Resolution* 47, 1 (2003), 72–93.

Kevin A Clarke. 2007. A simple distribution-free test for nonnested model selection. *Political Analysis* 15, 3 (2007), 347–363.

Aaron Clauset, Cosma R Shalizi, and Mark EJ Newman. 2007. Power-law distributions in empirical data. *SIAM review* 51, 4 (2007), 661–703.

Maarten Clements, Arjen P de Vries, and Marcel JT Reinders. 2010. The influence of personalization on tag query length in social media search. *Information Processing & Management* 46, 4 (2010), 403–412.

Will Cook, Paul Ormerod, and Ellie Cooper. 2004. Scaling behaviour in the number of criminal acts committed by individuals. *Journal of Statistical Mechanics: Theory and Experiment* 2004, 07 (2004), P07003.

Gregory W Corder and Dale I Foreman. 2009. *Nonparametric statistics for non-statisticians: a step-by-step approach*. John Wiley & Sons.

Gordon V Cormack, Mark D Smucker, and Charles LA Clarke. 2011. Efficient and effective spam filtering and re-ranking for large web datasets. *Information retrieval* 14, 5 (2011), 441–465.

Nick Craswell and Martin Szummer. 2007. Random walks on the click graph. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 239–246.

Mark E Crovella and Murad S Taqqu. 1999. Estimating the heavy tail index from scaling properties. *Methodology and computing in applied probability* 1, 1 (1999), 55–79.

Wang Dahui, Li Menghui, and Di Zengru. 2005. True reason for Zipf's law in language. *Physica A: Statistical Mechanics and its Applications* 358, 2 (2005), 545–550.

Russell Davidson and James G MacKinnon. 1981. Several tests for model specification in the presence of alternative hypotheses. *Econometrica: Journal of the Econometric Society* (1981), 781–793.

Shuai Ding, Josh Attenberg, Ricardo Baeza-Yates, and Torsten Suel. 2011. Batch query processing for web search engines. In *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, 137–146.

Sandor Dominich and Tamas Kiezer. 2005. Zipfs Law, Small World and Hungarian Language. *Alkalmazott Nyelvtudomány* 1, 2 (2005), 5–24. In Hungarian.

Joshua Drucker. 2007. *Regional dominance and industrial success: a productivity-based analysis*. ProQuest.

Jan Eeckhout. 2004. Gibrat's law for (all) cities. *American Economic Review* (2004), 1429–1451.

Leo Egghe. 2000. The distribution of N-grams. *Scientometrics* 47, 2 (2000), 237–252.

Ramon Ferrer-i Cancho and Brita Elvevåg. 2010. Random texts do not exhibit the real Zipf's law-like rank distribution. *PLoS One* 5, 3 (2010), e9411.

Andrey Feuerverger and Peter Hall. 1999. Estimating a tail exponent by modelling departure from a Pareto distribution. *The Annals of Statistics* 27, 2 (1999), 760–781.

Catherine Forbes, Merran Evans, Nicholas Hastings, and Brian Peacock. 2011. *Statistical distributions*. John Wiley & Sons.

Xavier Gabaix. 2009. Power Laws in Economics and Finance. *Annual Review of Economics* 1 (2009), 255–93.

Wojciech Galuba, Karl Aberer, Dipanjan Chakraborty, Zoran Despotovic, and Wolfgang Kellerer. 2010. Outtweeting the Twitterers - Predicting Information Cascades in Microblogs. In *Proceedings of the 3rd conference on Online social networks*.

David Garcia, Pavlin Mavrodiev, and Frank Schweitzer. 2013. Social resilience in online communities: The autopsy of friendster. In *Proceedings of the first ACM conference on Online social networks*. ACM, 39–50.

Wolfgang Gatterbauer. 2011. Rules of Thumb for Information Acquisition from Large and Redundant Data. In *Advances in Information Retrieval - 33rd European Conference on IR Research, ECIR 2011, Dublin, Ireland, April 18-21, 2011*. 479–490.

Natalie Glance, Matthew Hurst, and Takashi Tomokiyo. 2004. Blogpulse: Automated trend discovery for weblogs. In *WWW 2004 workshop on the weblogging ecosystem: Aggregation, analysis and dynamics*, Vol. 2004. ACM.

Yoav Goldberg and Jon Orwant. 2013. *A Dataset of Syntactic-Ngrams over Time from a Very Large Corpus of English Books*. Technical Report. Google. http://commondatastorage.googleapis.com/books/syntactic-ngrams/index.html.

Greg N Gregoriou. 2009. *Operational Risk Toward Basel III: Best Practices and Issues in Modeling, Management, and Regulation*. Vol. 481. John Wiley & Sons.

Peter Grünwald. 2007. *The minimum description length principle*. MIT press.

Cathal Gurrin and Alan F Smeaton. 2004. Replicating web structure in small-scale test collections. *Information retrieval* 7, 3-4 (2004), 239–263.

Matthias Hagen, Martin Potthast, Benno Stein, and Christof Braeutigam. 2010. The power of naive query segmentation. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 797–798.

Harry Halpin, Valentin Robu, and Hana Shepherd. 2007. The complex dynamics of collaborative tagging. In *Proceedings of the 16th international conference on World Wide Web*. ACM, 211–220.

Robert K Hammond and James E Bickel. 2013. Reexamining discrete approximations to continuous distributions. *Decision Analysis* 10, 1 (2013), 6–25.

Claudia Hauff and Leif Azzopardi. 2005. Age dependent document priors in link structure analysis. In *Advances in Information Retrieval*. Springer, 552–554.

Harold S Heaps. 1978. *Information Retrieval: Computational and Theoretical Aspects*. Academic Press, Inc., Orlando, FL, USA.

Daniel Heesch and Stefan Rüger. 2004. NN^k networks for content-based image retrieval. In *Advances in Information Retrieval*. Springer, 253–266.

Joseph Hilbe. 2011. *Negative binomial regression*. Cambridge University Press.

Bruce M Hill. 1975. A simple general approach to inference about the tail of a distribution. *The Annals of Statistics* 3, 5 (1975), 1163–1174.

Andreas Hotho, Robert Jäschke, Christoph Schmitz, and Gerd Stumme. 2006. Information Retrieval in Folksonomies: Search and Ranking. *The Semantic Web: Research and Applications* (2006), 411–426.

Bernardo A Huberman and Lada A Adamic. 1999. Evolutionary dynamics of the world wide web. *arXiv preprint cond-mat/9901071* (1999).

Clifford M Hurvich and Chih-Ling Tsai. 1989. Regression and time series model selection in small samples. *Biometrika* 76, 2 (1989), 297–307.

Hawoong Jeong, Bálint Tombor, Réka Albert, Zoltan N Oltvai, and Albert-László Barabási. 2000. The large-scale organization of metabolic networks. *Nature* 407, 6804 (2000), 651–654.

Hai Jin, Xiaomin Ning, and Hanhua Chen. 2006. Efficient search for peer-to-peer information retrieval using semantic small world. In *Proceedings of the 15th international conference on World Wide Web*. ACM, 1003–1004.

Shudong Jin and Azer Bestavros. 2000. Sources and characteristics of Web temporal locality. In *Modeling, Analysis and Simulation of Computer and Telecommunication Systems, 2000. Proceedings. 8th International Symposium on*. IEEE, 28–35.

Norman L Johnson, Samuel Kotz, and Narayanaswamy Balakrishnan. 2002. *Continuous Multivariate Distributions, volume 1, Models and Applications*. Vol. 59. New York: John Wiley & Sons.

Jaeyeon Jung, Emil Sit, Hari Balakrishnan, and Robert Morris. 2002. DNS Performance and the Effectiveness of Caching. *IEEE/ACM Transactions on Networking* 10, 5 (2002), 589–603.

Jaap Kamps and Marijn Koolen. 2008. The importance of link evidence in Wikipedia. In *Advances in Information Retrieval*. Springer, 270–282.

Noriaki Kawamae. 2014. Supervised N-gram Topic Model. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining (Web Search and Data Mining '14)*. 473–482.

Noam Koenigstein, Yuval Shavitt, Ela Weinsberg, and Udi Weinsberg. 2010. On the Applicability of Peer-to-peer Data in Music Information Retrieval Research. In *International Society for Music Information Retrieval*. 273–278.

Leonid Kopylev. 2012. Constrained parameters in applications: Review of issues and approaches. *International Scholarly Research Notices* 2012 (2012).

Beate Krause, Robert Jäschke, Andreas Hotho, and Gerd Stumme. 2008. Logsonomy-social information retrieval with logdata. In *Proceedings of the nineteenth ACM conference on Hypertext and hypermedia*. ACM, 157–166.

Jérôme Kunegis and Julia Preusse. 2012. Fairness on the web: Alternatives to the power law. In *Proceedings of the 4th Annual ACM Web Science Conference*. ACM, 175–184.

Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is Twitter, a social network or a news media?. In *Proceedings of the 19th international conference on World wide web*. ACM, 591–600.

Erich L. Lehmann and Joseph P Romano. 2006. *Testing statistical hypotheses*. Springer.

Mark Levy and Mark Sandler. 2009. Music information retrieval using social tags and audio. *IEEE Transactions on Multimedia* 11, 3 (2009), 383–395.

Christina Lioma. 2007. *Part of Speech n-Grams for Information Retrieval*. Ph.D. Dissertation. University of Glasgow.

Christina Lioma and Iadh Ounis. 2007. Light Syntactically-based Index Pruning for Information Retrieval. In *Advances in Information Retrieval, 29th European Conference on IR Research, ECIR 2007, Rome, Italy, April 2-5, 2007, Proceedings*. 88–100.

Christina Lioma and Iadh Ounis. 2008. A Syntactically-based Query Reformulation Technique for Information Retrieval. *Information Processing & Management* 44 (2008), 143–162.

Christina Lioma and Cornelis Joost van Rijsbergen. 2008. Part of speech n-grams and information retrieval. *Revue française de linguistique appliquée* 13, 1 (2008), 9–22.

Tie-Yan Liu, Yiming Yang, Hao Wan, Hua-Jun Zeng, Zheng Chen, and Wei-Ying Ma. 2005. Support vector machines classification with a very large-scale taxonomy. *ACM Knowledge Discovery and Data Mining: Explorations Newsletter* 7, 1 (2005), 36–43.

Wuying Liu, Lin Wang, and Mianzhu Yi. 2013. Power Law for Text Categorization. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*. Springer, 131–143.

Roger Lowenstein. 2000. *When genius failed: The rise and fall of Long-Term Capital Management*. Random House Trade Paperbacks.

Hans P Luhn. 1958. The automatic creation of literature abstracts. *IBM Journal of research and development* 2, 2 (1958), 159–165.

Marianne Lykke, Birger Larsen, Haakon Lund, and Peter Ingwersen. 2010. Developing a test collection for the evaluation of integrated search. In *ECIR'10*. 627–630.

Colin L Mallows. 1973. Some comments on C_P. *Technometrics* 15, 4 (1973), 661–675.

Benoit Mandelbrot. 1953. An informational theory of the statistical structure of language. *Communication theory* 84 (1953).

Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*. Vol. 1. Cambridge university press Cambridge.

Yuqing Mao and Zhiyong Lu. 2013. Predicting clicks of PubMed articles. In *AMIA Annual Symposium Proceedings*, Vol. 2013. American Medical Informatics Association, 947.

Alberto Maydeu-Olivares and Carlos Garca-Forero. 2010. Goodness-of-Fit Testing. In *International Encyclopedia of Education* (3 ed.), Baker E. Peterson, P. and B. McGaw (Eds.). Elsevier, 190–196.

Alberto Medina, Ibrahim Matta, and John Byers. 2000. On the origin of power laws in Internet topologies. *ACM SIGCOMM Computer Communication rReview* 30, 2 (2000), 18–28.

Mark M Meerschaert and Hans-Peter Scheffler. 2001. *Limit distributions for sums of independent random vectors: Heavy tails in theory and practice*. Vol. 321. John Wiley & Sons.

Edgar Meij and Maarten de Rijke. 2007. Using Prior Information Derived from Citations in Literature Search. In *Recherche d'Information et ses Applications*.

George A Miller. 1957. Some effects of intermittent silence. *The American Journal of Psychology* (1957), 311–314.

Staša Milojević. 2010. Power law distributions in information science: Making the case for logarithmic binning. *Journal of the American Society for Information Science and Technology* 61, 12 (2010), 2417–2425.

Gilad Mishne and Natalie Glance. 2006. Leave a reply: An analysis of weblog comments. In *Third annual workshop on the Weblogging ecosystem*.

Michael Mitzenmacher. 2004. A brief history of generative models for power law and lognormal distributions. *Internet mathematics* 1, 2 (2004), 226–251.

Saeedeh Momtazi and Dietrich Klakow. 2010. Hierarchical Pitman-Yor language model for information retrieval. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 793–794.

Fabrice Muhlenbach and Ricco Rakotomalala. 2005. Discretization of continuous attributes. *Encyclopedia of Data Warehousing and Mining* 1 (2005), 397–402.

Mark EJ Newman. 2005. Power laws, Pareto distributions and Zipf's law. *Contemporary physics* 46, 5 (2005), 323–351.

Christopher R Palmer and Greg Steffan. 2000. Generating network topologies that obey power laws. In *Global Telecommunications Conference, 2000. GLOBECOM'00. IEEE*, Vol. 1. IEEE, 434–438.

Greg Pass, Abdur Chowdhury, and Cayley Torgeson. 2006. A Picture of Search. In *Proceedings of the 1st International Conference on Scalable Information Systems (InfoScale '06)*. Article 1.

David M Pennock, Gary William Flake, Steve Lawrence, Eric J Glover, and Clyde L Giles. 2002. Winners don't take all: Characterizing the competition for links on the web. *Proceedings of the national academy of sciences* 99, 8 (2002), 5207–5211.

Matjaž Perc. 2010. Zipfs law and log-normal distributions in measures of scientific output across fields and institutions: 40 years of Slovenias research as an example. *Journal of Informetrics* 4, 3 (2010), 358–364.

Isabella Peters and Wolfgang G Stock. 2010. "Power tags" in information retrieval. *Library Hi Tech* 28, 1 (2010), 81–93.

Jim Pitman and Marc Yor. 1997. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability* (1997), 855–900.

David Posada and Thomas R Buckley. 2004. Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. *Systematic biology* 53, 5 (2004), 793–808.

Le Quan Ha, Ji Ming, and Francis Jack Smith. 2003. Extension of Zipfs law to word and character n-grams for English and Chinese. In *Journal of Computational Linguistics and Chinese Language Processing*. Citeseer.

Venugopalan Ramasubramanian and Emin Gün Sirer. 2004. Beehive: Exploiting power law query distributions for O (1) lookup performance in peer to peer overlays. In *Symposium on Networked Systems Design and Implementation, San Francisco CA*.

Sidney Redner. 1998. How popular is your paper? An empirical study of the citation distribution. *The European Physical Journal B-Condensed Matter and Complex Systems* 4, 2 (1998), 131–134.

William J Reed. 2003. The Pareto law of incomes: an explanation and an extension. *Physica A: Statistical Mechanics and its Applications* 319 (2003), 469–486.

William J Reed and Murray Jorgensen. 2004. The double Pareto-lognormal distributiona new parametric model for size distributions. *Communications in Statistics-Theory and Methods* 33, 8 (2004), 1733–1753.

Matei Ripeanu and Ian T Foster. 2002. Mapping the Gnutella Network: Macroscopic Properties of Large-Scale Peer-to-Peer Systems. In *IPTPS*. 85–93.

Seth Roberts and Harold Pashler. 2000. How persuasive is a good fit? A comment on theory testing. *Psychological review* 107, 2 (2000), 358.

Issei Sato and Hiroshi Nakagawa. 2010. Topic models with power-law using Pitman-Yor process. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 673–682.

Christian D Schunn and Dieter Wallach. 2005. Evaluating goodness-of-fit in comparison of models to data. University of Saarland Press, Saarbrueken, 115–154.

Gideon Schwarz. 1978. Estimating the Dimension of a Model. *The Annals of Statistics* 6, 2 (1978), 461–464.

Ripunjai K Shukla, Mohan Trivedi, and Manoj Kumar. 2010. On the proficient use of GEV distribution: a case study of subtropical monsoon region in India. *Annals of Computer Science Series* 8, 1 (2010).

Börkur Sigurbjörnsson and Roelof van Zwol. 2008. Flickr tag recommendation based on collective knowledge. In *Proceedings of the 17th international conference on World Wide Web*. ACM, 327–336.

Herbert A Simon. 1955. On a class of skew distribution functions. *Biometrika* (1955), 425–440.

Ian Soboroff. 2002. Does wt10g look like the web?. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 423–424.

Karen Spärck-Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation* 28, 1 (1972), 11–21.

Laura Spierdijk and Mark Voorneveld. 2009. Superstars without talent? The Yule distribution controversy. *The Review of Economics and Statistics* 91, 3 (2009), 648–652.

Kunwadee Sripanidkulchai, Bruce Maggs, and Hui Zhang. 2003. Efficient content location using interest-based locality in peer-to-peer systems. In *INFOCOM 2003. Twenty-Second Annual Joint Conference of the IEEE Computer and Communications. IEEE Societies*, Vol. 3. IEEE, 2166–2176.

Nitish Srivastava, Ruslan Salakhutdinov, and Geoffrey E Hinton. 2013. Modeling Documents with Deep Boltzmann Machines. In *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 616–625.

Alexandru Tatar, Panayotis Antoniadis, Marcelo D De Amorim, and Serge Fdida. 2014. From popularity prediction to ranking online news. *Social Network Analysis and Mining* 4, 1 (2014), 1–12.

Jiancong Tong, Gang Wang, Douglas S Stones, Shizhao Sun, Xiaoguang Liu, and Fan Zhang. 2013. Exploiting query term correlation for list caching in web search engines. In *Proceedings of the 22nd ACM international conference on information & knowledge management*. ACM, 1817–1820.

Yana Volkovich, Nelly Litvak, and Debora Donato. 2007. Determining factors behind the PageRank log-log plot. In *Algorithms and Models for the Web-Graph*. Springer, 108–123.

Quang H Vuong. 1989. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica: Journal of the Econometric Society* (1989), 307–333.

Gui-Rong Xue, Hua-Jun Zeng, Zheng Chen, Yong Yu, Wei-Ying Ma, WenSi Xi, and WeiGuo Fan. 2004. Optimizing Web Search Using Web Click-through Data. In *Proceedings of the 13th ACM International Conference on Information and Knowledge Management (CIKM '04)*. ACM, 118–126.

Yiming Yang, Jian Zhang, and Bryan Kisiel. 2003. A scalability analysis of classifiers in text categorization. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 96–103.

Emmanuel J Yannakoudakis, Ioannis Tsomokos, and Paul J Hutton. 1990. n-Grams and their implication to natural language understanding. *Pattern Recognition* 23, 5 (1990), 509–528.

Mao Ye, Peifeng Yin, Wang-Chien Lee, and Dik-Lun Lee. 2011. Exploiting geographical influence for collaborative point-of-interest recommendation. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. ACM, 325–334.

Haizheng Zhang and Victor Lesser. 2006. Multi-agent based peer-to-peer information retrieval systems with concurrent search sessions. In *Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems*. ACM, 305–312.

Ding Zhou, Jiang Bian, Shuyi Zheng, Hongyuan Zha, and Clyde L Giles. 2008. Exploring social annotations for information retrieval. In *Proceedings of the 17th international conference on World Wide Web*. ACM, 715–724.

George K Zipf. 1935. *The psycho-biology of language*. Houghton, Mifflin.