**Power Laws in altmetrics: An empirical analysis**

Sumit Kumar Banshala, Solanki Guptab, Hiran H Lathabaic, Vivek Kumar Singhd,[[1]](#footnote-1),

asumitbanshal06@gmail.com

Department of Computer Science & Engineering, Daffodil International University, Dhaka-1207. (Bangladesh)

asolankigupta2@gmail.com, chiranhl007@gmail.com, dvivek@bhu.ac.in,

Department of Computer Science, Banaras Hindu University, Varanasi-221005. (India)

Abstract

Power laws are a characteristic distribution found in both natural as well as in man-made systems. Previous studies have shown that citations to scientific articles follow a power law, i.e., the number of papers having a certain level of citation x are proportional to x raised to some negative power. However, the distributional character of altmetrics (such as reads, likes, mentions, etc.) has not been studied in much detail, particularly with respect to existence of power law behaviours. This article, therefore, attempts to do an empirical analysis of altmetric mention data of a large set of scholarly articles to see if they exhibit power law. The individual and the composite data series of ‘mentions’ on the various platforms are fit to a power law distribution, and the parameters and goodness of fit are determined, both using least squares regression as well as the Maximum Likelihood Estimate (MLE) approach. We also explore the fit of the mention data to other distribution families like the Log-normal and exponential distributions. Results obtained confirm the existence of power law behaviour in social media mentions to scholarly articles. The Log-normal distribution also looks plausible but is not found to be statistically significant, and the exponential distribution does not show a good fit. Major implications of power law in altmetrics are given and interesting research questions are posed in pursuit of enhancing the reliability of altmetrics for research evaluation purposes.

**Keywords:** Altmetrics, Exponential Distribution, Log-normal distribution, Power Laws, Scientometrics, Social Media Mentions.

# 1. Introduction

A power law is simply expressed by the fact that if the probability distribution of measuring a quantity x varies as the inverse power of the same variable raised to a value k, p(x) ~ x-α, then x is said to follow a power law with exponent -α (Newman, 2005). Power laws were first noticed a century ago by Pareto in the context of income distribution (Pareto, 1896). Pareto’s findings suggested the 80-20 rule, which illustrated that 80% of the world’s wealth is accumulated by 20% of the world population. In a generalized form, it states that if an observed phenomenon follows 80-20 rule, then 80% of total sampled occurrences correspond to 20% of the observed samples. This occurrence has been found prevalent in various domain and observed to be scale free (Crawford, 2020). Power Laws are abound in nature as well as in the artificially created technology systems, most recently the Internet (Adamic, 2000; Faloutsos et al., 1999). Basically, any heavily tailed distribution could follow or model into the power law. This specific feature was observed in different disciplinary areas at different periods of time, over the last century by different people, and early power laws go by different names like the Bradford law (Bradford, 1934), Zipf’s law (Zipf, 1949), Pareto’s law (Chen & Leimkuhler, 1986; Pareto, 1896). All these types of distribution models are said to imitate the Lotka’s Law (Lotka, 1926). Power law distributions involve various signature characteristics such as - if the full range is represented on the axes, the shape of the curve would be a perfect L. Also, if same distribution is plotted on log-log scale, then the curve is always linear in nature (Adamic, 2000). Further, the power law distribution is the only distribution that holds the scale–free property, i.e., the shape of distribution remains unchanged whatever be the scale at which it is observed (Newman, 2005).

In the domain of scholarly articles, citations have been considered as a widely known parameter of impact assessment for a long time. Several previous studies have shown that Power laws are observed in the distribution of citations to scientific papers (Ruiz-Castillo, 2012; Brzezinski, 2015; Redner, 1998, 2005; Thelwall & Sud, 2016). With the evolution of social media platforms and their increased use in scholarly publishing domain in the recent times, a new and quick event based impact measure has emerged, known as alternative metrics or ‘altmetrics’ (Priem et al., 2010; Priem & Hemminger, 2010). Altmetrics comprise of different kinds of events about scholarly articles, happening in social media platforms and academic social networks. Researchers are actively using platforms like Twitter and Facebook, as well as academic social networks like Academia, Mendeley, and ResearchGate, to share and disseminate their research results. Motivated by this, the researchers working on quantitative science studies are now studying several aspects of altmetrics ranging from country-level altmetric studies (Banshal et al., 2018, 2019a; Wang et al., 2016a) to discipline-specific coverage studies (Bar-ilan, 2014; Chen et al., 2015; Hammarfelt, 2014; Htoo & Na, 2017; Sotudeh et al., 2015; Vogl et al., 2018). Some other studies attempted to analyse disciplinary variations in altmetrics (Banshal et al., 2019b; Costas et al., 2015a; Holmberg and Thelwall, 2014; Mohammadi and Thelwall, 2013; Ortega, 2015; Thelwall and Kousha, 2017; Zahedi et al., 2014).

Some recent studies on altmetrics have pointed that altmetrics correlate with citations in different degrees (Banshal et al., 2021; Costas et al., 2015b; Eysenbach, 2011; Haustein et al., 2014; Peoples et al., 2016; Shema et al., 2014; Thelwall, 2018; Thelwall & Nevill, 2018). Given that the distribution of citations has been observed to follow Power Laws, it would be very interesting to analyse if such patterns also exist for altmetrics. However, the altmetric data distributions are not analysed in detail with respect to existence of power laws. In fact, there are very few studies that tried to analyse the distributional characteristics of altmetrics. Costas et al. (2016) studied distribution patterns of citations and altmetrics (Mendeley reads, tweets, blogs) and found that twitter and blog mentions are extremely skewed distributions. They observed that it was largely caused by a large share of publications without any mentions on these platforms. Though some other studies have indicated that altmetrics are skewed in nature (Thelwall & Nevill, 2018; Thelwall & Wilson, 2016), not much is known about the various statistical properties of altmetrics. Given the knowledge that altmetric distributions are skewed, it becomes more important to analyse and see if they follow Power Laws, and/ or fit to some other data distributions. This paper, therefore, attempts to do a systematic empirical analysis of existence of Power Laws in altmetrics and understand its type/ nature. Towards this end, we first survey some existing related studies on Power Laws and thereafter present the key research questions explored in our work.

# 2,. Related Work

Power Law properties are ubiquitous in various natural and man-made systems in the sense that Power laws are often seen to appear in large, interconnected and self-organising systems. It involves properties like its scale free nature which appears to be universal (Crawford, 2020). Because of this it has applications in various domains like Physics (e.g. sandpile avalanches), Economics, Linguistics, Biology (e.g. species extinction), Finance, Information and Computer Science, Geology, Social Science, Astronomy etc (Clauset et al., 2009; Newman, 2005). A lot of studies have been done regarding the emergence of Power Laws in diverse fields. Power Laws help in explaining various phenomena of Science (Mitzenmacher, 2004), Economics (Gabaix, 2009, 2016), Financial Time series (Bouchaud, 2001), and also network topologies (Faloutsos et al., 1999). The statistics for number of visitors to sites of the World Wide Web was measured and the distribution of visitors per site observed, and it approximated a Power Law (Adamic, 2000; Adamic & Huberman, 2001). There is a long list available as a proof of occurrence of Power Laws in various disciplines. Some of them mentioned in Newman (2005) are- word frequency count in texts, citations of scientific papers, web hits, copies of books sold, telephone calls, magnitude of earthquakes, diameter of moon craters, intensity of solar flares, intensity of wars, wealth of the richest people, frequencies of family names, population of cities and so on. Beyond the scale free properties, another probable reason of the ubiquitous existence power law might be the existence of exponential growth in any observed domain (Reed & Hughes, 2002).

The first mention about the existence of Power Laws in the scholarly articles was made by Price (1965) for the highly cited articles. Later on, he recommended ‘cummulative advantage’ mechanism which also can cause Power Law distribution (Price, 1976). Redner (1998) analyzed the citation distribution of scientiﬁc publications and found that the asymptotic tail of the citation distribution appears to be described by a Power Law, with exponent equals to 3. In another work, Redner analysed articles published over a century in the journal of physical review and found the tailed characteristics of citations (Redner, 2005). The exponent factor of highly cited papers and relatively lesser cited papers has been found to vary as studied by (Peterson et al., 2010). They analysed two types of mechanisms in citations namely ‘direct’ and ‘indirect’ mechanism. In this approach, three different set of scientific articles were examined. The existence of Power Laws in citations are relatively universally found in the context of research discipline (Radicchi et al., 2008). In a more comprehensive and elaborate empircal approach, Brzezinski (2015) detected power-law behaviour in the citation distribution. In this work, scientific papers published between 1998 and 2002 were drawn from Scopus and analysed. They found that the power-law hypothesis is not satisfied for around half of the Scopus fields of science.

A more robust and quantitative analysis of statistical distributions was performed by Thelwall (2016c) by considering data from 26 Scopus subject areas for seven years including 911,971 journal articles. The study considered three different models-the hooked Power law model, the Truncated Power Law model, and the Discretised log normal model. It was explored whether the discretised log-normal and hooked power law distributions were plausible for citation data. It also tested if there were too many uncited articles, and zero inflated variants of the discretised log-normal and hooked power law distributions (Thelwall, 2016b). The study examined the best options for modelling and regression, and tried to detect which distribution best fit citation data by including and excluding uncited articles (Thelwall, 2016a). It was concluded that the hooked power law and discretised log-normal distribution were the best options for complete citation data. Though several studies (Costas et al., 2015a; Haustein et al., 2014; Eysenbach, 2011; Mohammadi et al., 2016; Thelwall, 2018; Thelwall & Nevill, 2018) have found positive correlations between citations and altmetrics, but distribution patterns of altmetric data are not well explored.

The only major studies on distribution analysis of altmetric data include analysis of online events- web usage counts (Wang et al., 2016b), readership (Thelwall & Wilson, 2016) and download patterns (Duan & Xiong, 2017). Wang et al. (2016b) investigated the Web of Science usage count to understand the distribution and the relation with citation counts. They analysed data for about 12,000 articles from five journals of information science domain. The ‘usage count’ for each article was analysed for fit with Power Law distribution and positive evidence of Power Law behaviours was observed for usage counts. The study by Thelwall & Wilson (2016) analysed about 332,000 scientific publications in 45 subfields of medical sciences domain drawn from Scopus database. The study observed the relationship between Mendeley reads and citations and examined the distribution patterns of reads by exploring fits to either log-normal or hooked power law. They concluded that, hooked power law is a better fit in case of citation data, but for Mendeley reads, it varied across 45 sub-fields. Duan & Xiong (2017) analysed the download patterns from the Chinese Library & Information Sciences using two-step cluster analysis, outlining the distribution for the same. The authors collected data from eleven chinese journals and grouped them into clusters based on the type of downloads. They concluded that the majority of sets followed power function in the case of absolute downloads. Costas et al. (2016) also studied distribution patterns of citations and altmetrics (Mendeley reads, tweets, blogs) and found that twitter and blog mentions are extremely skewed distributions. They observed that it was largely caused by a large share of publications without any mentions on these platforms. There are, however, no previous studies on examining the distribution patterns of mentions in the popular social media platforms like Twitter, Facebook, Blogs etc., particularly with respect to existence of power law behaviours. This research gap motivated us to explore if altmetric mentions exhibit power laws.

**3. Objectives/ Research Questions**

The article attempts to answer following key research questions:

1. Do altmetrics indicators (mentions and Altmetric Attention Score) follow Power Laws?
2. Will other distributions like exponential and log-normal distributions, truncated power law and stretched exponential be plausible for altmetrics?

In order to answer the first question above, the following question has to be addressed: *Does the power-law model fit well to the empirically found distributions related to altmetrics?* In order to answer the second question, the altmetric data has to be fit to other distributions like- log-normal and exponential. The following sections present the various relevant definitions and explain the outline of the methodological approach.

**4. Definitions**

In this section first we define the distributions that are used in this study.

Let us consider that “” indicates the quantity of interest, i.e., the variable whose distribution we want to study. Formally, it is described that a random variableobeys a particular law if it is drawn from a probability distribution . For example, obeys a power law if it is drawn from a probability distribution

|  |  |  |
| --- | --- | --- |
|  |  | (1) |

where “α” is a constant parameter of the distribution known as the exponent or scaling parameter (Clauset et al., 2009).

Some of the common distributions of the form , where represents the functional form and “” the normalization constant, are given in **table 1**. A brief definition of the major distributions are given below.

**Table 1. Statistical Definition of the power-law and other common distributions.**

|  |  |
| --- | --- |
| **Distributions Name** |  |
| Power Law |  |
| Truncated Power Law |  |
| Exponential | λ>0, x>0 |
| Stretched Exponential |  |
| Log-normal |  |

***Exponential Distribution:*** A continuous random variable X is said to follow exponential distribution if for any positive value λ, it has the probability density function (pdf) defined as given in table 1 (fourth row), where λ is known as parameter of exponential Distribution.

***Stretched Exponential Distribution*:** The stretched exponential function is defined as given in table 1 (fifth row). It is obtained by inserting a fractional [power law](https://en.wikipedia.org/wiki/Power_law) into the [exponential function](https://en.wikipedia.org/wiki/Exponential_function). In general applications, *x* lies between 0 and +∞. With *β* = 1, the usual exponential function is recovered. With a stretching exponent *β* between 0 and 1, the graph of log *f* versus x is characteristically stretched, hence it is named as stretched exponential. It is also Known as “Weibull Distribution”.

# *Log-normal Distribution:* When the natural logarithm of a dataset under consideration follows a normal distribution, it can be assumed to be log-normally distributed and such a distribution is termed as ‘log-normal distribution’. It is a continuous probability distribution. As natural logarithm cannot be calculated for zero or negative numbers, only positive values are possible for such distributions. The log-normal distribution is expressed as given in table 1 (sixth row). It has two standard parameters, the mean (µ) and standard deviation (σ).

#### **Truncated Power Law:** It is expressed as given in table 1 (third row). It is also Known as **“**Power law with exponential cut-off”. A power law with an exponential cut off is obtained when a power law is multiplied by an exponential function. α and are its two parameters.

{\displaystyle f(x)\propto x^{-\alpha }e^{-\beta x}.}

**5. Outline of the methodological approach**

The detailed approach followed in this article is summarized in the flow chart shown in **fig. 1**.

For obtaining primary evidence of the existence of power law (research question 1), the standard strategy for revealing a power law is to draw a size-frequency plot, i.e., to plot the number of papers N(k) with a given level of mentions k. In other words, we use the fact that a histogram of a quantity with a power law distribution appears as a straight line when plotted on a logarithmic scale (Newman, 2005). Therefore, the mentions were plotted using log-log plot and the parameters of fit calculated to have primary evidence of a power-law. The plots were made for altmetric attention score (AAS) and also mention data for Twitter, Facebook, News and Blog platforms, and the observed values of power-law exponents were computed.

Although least square method is quite popular in estimation of power law but there are several problems involved in this approach because these estimates of the slope are subject to systemic and potentially large errors. Hence, it is necessary to verify exponent values with other approaches also. One of the most common approaches for such estimations is Maximum-Likelihood-Estimation approach (MLE).

**Figure. 1: Systematic procedure for testing existence of power laws**

**Maximum Likelihood Estimation (MLE) for scaling exponent**

As we know that power laws are undefined for x=0, hence there must be some minimum value which is to be defined. This minimum could be a theoretical minimum, a noise threshold, or the minimum value observed in the data.

In order to calculate exponent using MLE, we have to first determine the value of that makes the probability distributions of measured data and the best fit power law model as similar as possible above. Next question arises that from what minimal value the scaling relationship of the power law begins. In order to address this, we follow the approach of (Clauset et al., 2009). They calculated the value of by creating a power law fit starting from each unique value in the dataset, then selecting the one that results in the minimal Kolmogorov-Smirnov (KS) distance, D, between the data and the fit. Here D is described as

|  |  |  |
| --- | --- | --- |
|  | x>=xmin | (2) |

Here S(x) is the CDF of the data for the observations with value at least , and P(x) is the CDF for the power-law model that best fits the data in the region x ≥ . Our estimate ˆ is then the value ofthat minimizes D.

Assuming that our data are drawn from a distribution that follows a power law exactly for *x*≥ , we use maximum likelihood estimators (MLEs) of the scaling parameter. The value of exponent is described by the equation

|  |  |  |
| --- | --- | --- |
|  |  | (3) |

Note: For detailed derivation refer (Clauset et al., 2009; Newman, 2005)

The method described above provides a reliable way to test whether a given data is plausibly drawn from a power-law distribution. In practice, these estimators give quite good results. However, the results don't describe the complete story. Even if our data provides good fit for a power law, it is still possible that another distribution, such as an exponential or a log-normal, might have a better fit. We can eliminate this possibility by using a goodness-of-fit test. We can simply calculate a *p*-value for a fit to the competing distributions and compare it with power law.

There are several existing methods which can directly compare two distributions against each other. In this section, we describe one of such method, the likelihood ratio test.

**Goodness-of-fit test (Log likelihood test) as supporting evidence**

Log likelihood test is a methodology to assess the evidence regarding two simple statistical hypotheses**.** The basic idea behind log likelihood ratio test is to calculate the ratio of the likelihoods between the two candidate distributions, or equivalently the logarithm R of the ratio, which is positive or negative depending on which distribution is better. In case of the event of a tie, the value is zero. Ratio value will be positive if data is more likely towards the first candidate distribution and negative if data is more likely towards the second candidate. However, whether we can trust the sign of the ratio or not, can be tested using Voung’s test (Voung, 1989). More details of the test are presented later in the paper.

Not all the distributions are capable to fit the power law data. Since it is previously shown that power law appears in tails too (Newman, 2005), therefore, we tested heavily tailed distributions in this paper. Some popular heavily tailed distributions are - Exponential distribution, Lognormal distribution, Weibull distribution etc.

Thus, we try to compare power law with exponential and log-normal distributions. In some real-world observations, though power law model seems to provide best fit to the data, power law behaviour is exhibited in the tail region only, not with the whole data. In such cases, such data is considered to follow truncated power law or power law with exponential cut off. Therefore, the need to check whether power law model fits at least as much as that of truncated power law

Apart from these, stretched exponential (Weibull) model is also known to produce heavy-tailed distributions. Thus, it is required to test the plausibility of whether stretched exponential model offer better fit to the data than the exponential distribution. The cases of power law and truncated power law comparison and exponential and stretched exponential comparison can be termed as a nested distribution log likelihood test.

**Log-likelihood test for nested Distributions**

Here the term “nested” means that one family of distributions is a subset of the other. When distributions are nested it is always considered that the larger family of distributions will provide a fit at least as good as the smaller, since every member of the smaller family is also a member of the larger. In this case, a slightly modified likelihood ratio test is needed to properly distinguish between such models (Jeff Alstott, 2014). The reliability of the sign of log-likelihood ratio can be determined by Voung’s test as in the case of non-nested distributions. But in case of nested we test the sign of log-likelihood on the basis of p values obtained directly.

As described above, for this analysis, we have to compare two family of distributions described as ‘power law and truncated power law’ and ‘exponential and stretched exponential’. Now, we are in a position to describe the details of data used for our study and the results of our analysis using the systematic procedure.

## **6. Data**

For this study, we obtained data from Altmetric.com. First, we take the publication records for the year 2016 from Web of Science as reference and then identify the articles that do not have a DOI. Such articles are excluded. The articles that have a DOI are then looked up in Altmetric.com for social media mention data. Only those records which are found tracked by Altmetric.com constitute the sample data for analysis.

There were a total of 2,528,868 publication records in Web of Science for the year 2016. Out of these, only 1,785,149 publication records had a DOI. A look up was then made for the articles with DOI in Altmetric.com for social media mention data. Altmetric.com was found to have data for 902,990. For the remaining articles, Altmetric.com did not have any tracked data. However, this does not mean that such articles can be treated as having ‘0’ mentions. Its only that Altmetric.com has no tracked social media data for such articles. In fact, some other altmetric aggregators (such as PlumX) may have social media data available for more articles. Therefore, one cannot say that such articles did not get any social media attention, and hence cannot be assumed to have ‘0’ mentions or otherwise. Thus, our data sample for analysis comprised of 902,990 articles, that are found tracked by Altmetric.com. Out of these 902,990 articles, not all articles have a non-zero mention for all social media platforms. Many articles included in the data sample have a ‘0’ mention in one or other platforms.

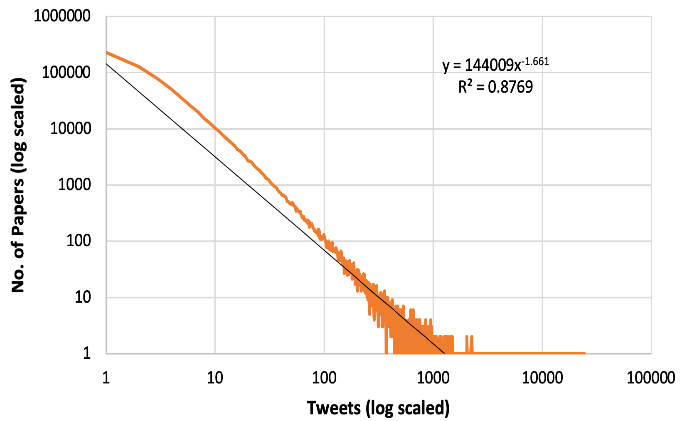
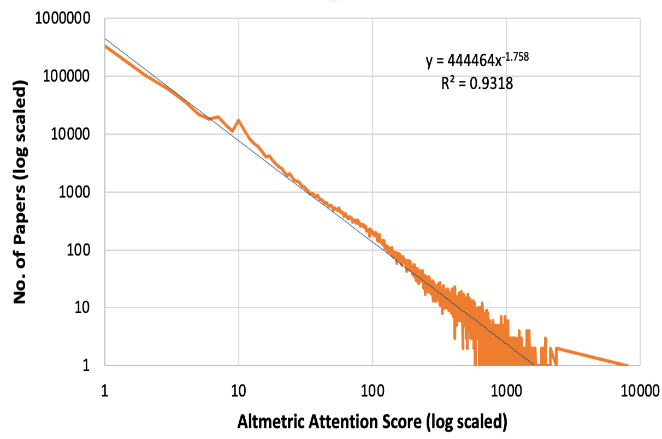
The altmetric data was collected from the Altmetric.com aggregator through an automated DOI-based lookup process. The Altmetric.com aggregator accumulates social and online mentions of research output for more than 18 different social and online platforms (such as Twitter, Facebook, Blog, News, Mendeley etc.). The data obtained from Altmetric.com had 46 fields, including DOI, title, Twitter mentions, Facebook mentions, News mentions, altmetric attention score, OA Status, Subjects (FoR), publication date, URI, etc. We have used the altmetric attention score (AAS) and mention data for Twitter, Facebook, News and Blog platforms for the analysis.

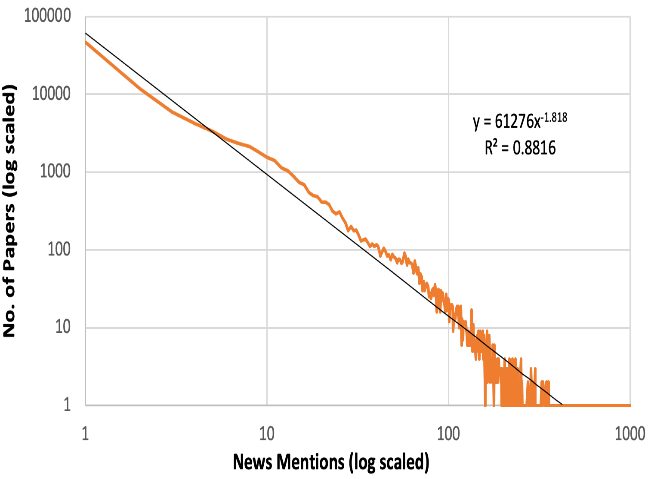
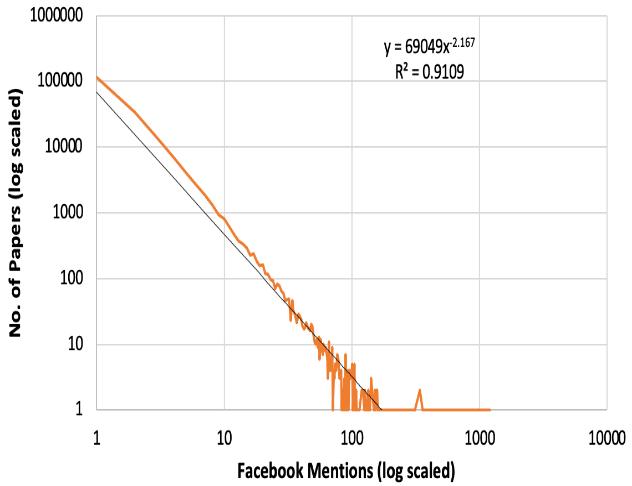
**7. Results**

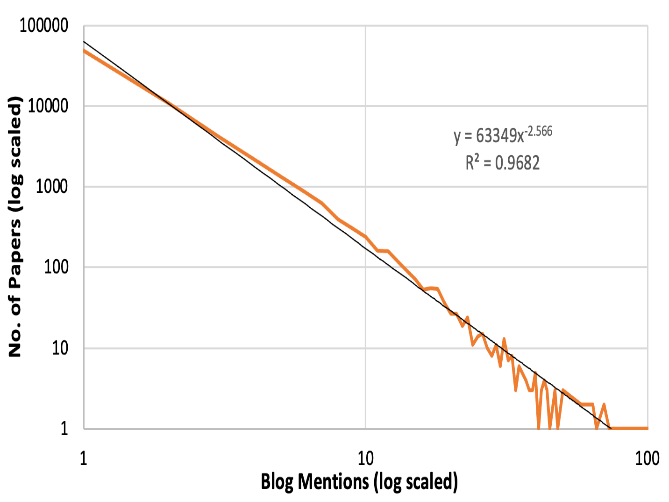
In order to understand the distribution of the data, the four prominent platforms namely, Facebook, Twitter, Blogs, News in Altmetric.com data along with the composite score-Altmetric attention score or AAS (a combined index created out of all the other indicators)- have been explored. The mentions were plotted using log-log plot and the parameters of fit were calculated to obtain primary evidence about the plausibility of a power-law. We find that the altmetrics data when plotted on log-log plot show a reasonably good linear fit. **Table 2** shows the value of different parameters and the **figure 2** shows the log-log plot for AAS and various other mentions. Please note that all Logarithm plots are on base 10.

**Table 2. Parameters of Fit to a power law for altmetric data**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Data Source** | **Data points** | **C** | **α** | **R2** |
| *AAS* | *757784* | *444464* | *1.758* | *0.9318* |
| *Twitter* | *700985* | *144009* | *1.661* | *0.8769* |
| *Facebook* | *185243* | *69049* | *2.167* | *0.9109* |
| *News* | *98499* | *61276* | *1.818* | *0.8816* |
| *Blog* | *70387* | *63349* | *2.614* | *0.9682* |

****

****

****

**Fig. 2 Log-log scale plot for altmetric attention score (AAS) and mentions data from four platforms**

Although we get high value of R2 (**Table 2**) but we can’t use it for hypotheses testing because in the terminology of statistical theory, the value of R2 has very little power as a method of hypothesis test since the probability of successfully detecting a violation of the power-law assumption is low (Clauset et al., 2009).Thus, the R2 value and the log-log plots indicate for the existence of power law behaviour, which we need to confirm further using MLE approach.

In order to try the Maximum likelihood estimation (MLE) approach on the dataset, we used the code based on approaches proposed by (Jeff Alstott, 2014)[[2]](#footnote-2), and estimated the model fit and calculated the values of (using eq.2) and the exponent α (using eq. 3). The values of observed parameters are shown in the **table 3**. From a perusal of the values observed, it can be said that MLE approach also confirms the power law behaviour (as discussed further in the ‘Discussion’ section).

**Table 3: Power law exponent and minimum exponent values are calculated using MLE approach**

|  |  |  |
| --- | --- | --- |
| **Mentions** |  | **Exponent α** |
| Twitter mentions | 3.0 | 1.486 |
| Facebook mentions | 3.0 | 1.404 |
| Blog mentions | 118.0 | 1.394 |
| News mentions | 86.0 | 1.575 |

According to Newman (2015), for a power law distribution with exponent α, all existing moments can be computed as:

Where *m* is the *m*th moment. *m*=1 indicates the first moment or mean, represented as or.*m*=2 indicates the second moment or mean square, represented as **,** which in turn can give an idea about variance or standard deviation.

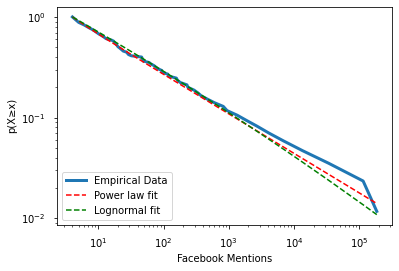
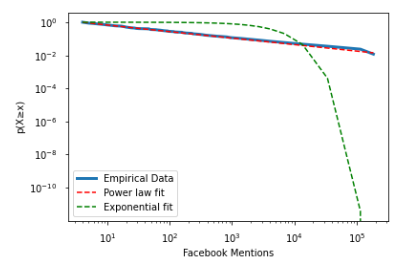
This means, become infinite for distributions with α ≤ 2. Similarly, become infinite for distributions with α ≤ 3 and so do variance or standard deviation.

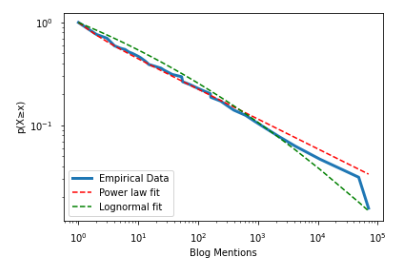
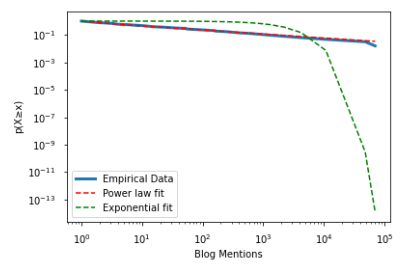
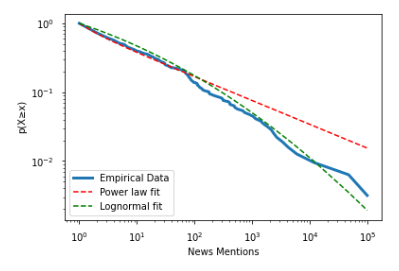
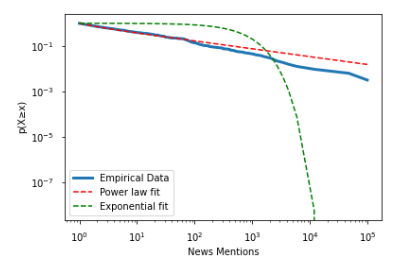
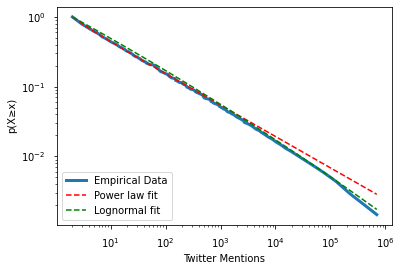
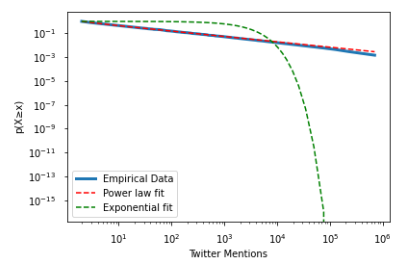
Practically, for a dataset (of finite size) with α ≤ 2 (and 3), there exists a mean which is dominated by the largest sample in the dataset. This mean appears to be finite because of the finite size of dataset which in turn ensures the finite value of largest sample in the dataset. However, with more samples drawn, the mean exhibit divergence or increase without bound under the influence of the then larger sample value. This means such distributions lack a characteristic mean or representative value. Similar is the case of variance and standard deviation.

In our case, since all the four datasets are found to be of α ≤ 2 (and 3), a characteristic value of mention score cannot be used as representative score for the datasets and a bound cannot be fixed for variation/deviation of values from a representative score even if we attempt to fix the representative score.

**Comparison of models (non-nested distributions)**

Next, in order to assess the plausibility of altmetric data for other distributions, we compared our power law fits with other non-nested distributions like Log-normal and exponential. For this we followed a common approach of visualization of power law distribution: we plot the cumulative distribution function on log-log scales, and again look for straight line behaviour. Here, we excluded altmetric attention score (AAS) as it is a composite of many indicators. Plots for mentions data from different platforms comparing empirical data fit along with power law and exponential and lognormal are shown in **fig. 3**.





**Figure 3. Figure indicates the complementary cumulative distribution function, and fitted** **power law and exponential/lognormal distributions** **for each social media mention. Left side plots compare power law fit with exponential fit and right side plots compare power law fit with log-normal fit. The blue line curve shows empirical fit of data.**

In above plots, p(X>=x) indicates the complementary cumulative distribution function (CCDF) This is also known as the survival function. These plots are designed using approach proposed in (Jeff Alstott, 2014). These plots are helpful in visualizing the fit of data. From **fig. 3** (left side plots), we observe that power-law provides better fit to the data as compared to exponential fit for the different social media mentions. From right side plots, it is observed that- (i) for twitter and news mentions lognormal seems to offer slightly better fit than power law (ii) whereas in case of facebook and blog mentions it is not clear which model is better fit for the data. Hence, we need to verify these mixed observations with some other standard procedures like Goodness-of-fit test etc. to arrive at a conclusion.

It may be noted that, comparing the power-law fit of our data with all competing distributions would not be possible. Indeed, as is usually the case with data fitting, it will almost always be possible to find a class of distributions that fits the data better than the power law if we define a family of curves with a sufficiently large number of parameters. Fitting the statistical distribution of data should therefore be approached using a combination of statistical techniques that constitutes a reasonable model for the data. Statistical tests can be used to rule out specific hypotheses.

Hence, we perform log-likelihood ratio test to analyse the goodness of fit and try to analyse which distribution provides better fit. For this purpose, in each case we drew n - independent values from each distribution and estimated the value of for each set of values, then calculated the likelihood ratio for the data above and the corresponding p-value. This procedure was repeated 1000 times to assess sampling fluctuations.

As discussed in Goodness-of-fit section, the sign obtained with log-likelihood ratio is indicative of the more plausible model that might have generated the data. But the sign of the log-likelihood ratio alone, will not definitively indicate which model suits better because it is subject to statistical fluctuation. If its true value, meaning that its expected value over many independent data sets drawn from the same distribution is close to zero, then the fluctuations could change the sign of the ratio and hence the results of the test cannot be trusted. In order to make a firm choice between distributions we need a log-likelihood ratio that is sufficiently positive or negative that it could not plausibly be the result of a chance fluctuation from a true result that is close to zero (Clauset et al., 2009).

Hence, to make a judgment about whether the observed value of R is sufficiently far from zero, we have to observe the size of the expected fluctuations. It means that we need to know the standard deviation σ on R and convert log-likelihood ratio to normalized log-likelihood ratio. The normalized log-likelihood ratio normalizes R by its standard deviation,

|  |  |  |
| --- | --- | --- |
|  | Normalized log likelihood ratio=R/ (σ/n1/2) | (4) |

The normalized log-likelihood ratio is what is directly used to calculate p and assess the plausibility of sign associated with normalized log-likelihood ratio. This we can estimate using a method proposed by Vuong (1989) dubbed as Vuong’s test. This method gives a p-value that tells us whether the observed sign of R is statistically significant. If this p-value is small (say, p < 0.01), then it is unlikely that the observed sign is a chance result of fluctuations and the sign is a reliable indicator of which model is the better fit to the data. If p is large, on the other hand, the sign is not reliable and the test does not favour either model over the other. It is one of the advantages of this approach that it can tell us not only which of two hypotheses is favoured, but also when the data are insufficient to favour either of them (Clauset et al., 2009).

The **table 4** shows the calculations for non-nested distributions using code described by (Jeff Alstott, 2014) which includes log-likelihood ratio, normalized log-likelihood ratio and p-values (or probability values). Positive values of log-likelihood ratio indicates that power law is favoured over the alternative. A negative value on the other hand shows that alternative distribution is more feasible than power law. For non- nested alternatives, we have to calculate normalized log-likelihood ratio using eq. 4 & its corresponding p-values. This measure is useful for applying Voung’s test. For further details refer discussion section.

**Table 4: Tests of power-law behaviour in the social media mentions using log-likelihood ratios**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Mentions | Log-normal | | | Exponential | | |
| R  (log-likelihood) | Normalized  log-likelihood | p value | R  (log-likelihood) | Normalized log-likelihood | p value |
| Twitter mentions | -0.738 | -1.144 | 0.252 | 2149.502 | 7.866 | 3.634e-15 |
| Facebook mentions | -0.011 | -0.104 | 0.916 | 354.705 | 7.901 | 2.757e-15 |
| Blog mentions | -0.014 | -0.124 | 0.9005 | 39.702 | 4.507 | 6.547e-06 |
| News mentions | -0.097 | -0.284 | 0.776 | 73.720 | 3.735 | 0.0001 |

**Nested Distributions**

In above sub section we compared power law with log-normal and exponential distributions where we took both distributions as a non-nested distribution, i.e., we compare likelihood of two distributions that are not related. Next, we have to compare the likelihoods of distributions that are nested versions of each other like ‘power law and truncated power law’ and ‘exponential and stretched exponential’. The main difference between nested and non-nested approach is in how the different factors are related to one another. Nested approach is basically multi model approach. In a nested design, the levels of one factor only make sense within the levels of another factor.

As mentioned above, in case of nested alternative, log-likelihood ratio signs and p values are used to test the results. Here again, we calculated likelihood ratio and corresponding p-values using code described by Jeff Alstott (2014). Values obtained are arranged in **table 5**. Here, the log-likelihood ratio values are obtained with negative sign, which indicates that it favours the second distribution. However, these observations of log-likelihood tests are to be seen with associated p values. The p-values for exponential and stretched exponential case is zero which shows there is no significance. Thus, from a perusal of the p-values, we can say that the stretched exponential distribution, that looked plausible, is actually not confirmed, as it is not significant. Whereas in case of power law and truncated power law, p values are much better as compared to exponential and stretched exponential. All these values are helpful is assesing the details regarding these nested distributions. For further details, please refer discussion section.

**Table 5: Tests of nested distribution in the social media mentions using log-likelihood ratios**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Mentions | Power law & truncated power law | | Exponential & Stretched exponential | |
| Likelihood ratio | P value | Likelihood ratio | P value |
| Twitter mentions | -0.724 | 0.228 | -2142.958 | 0.000 |
| Facebook mentions | -0.535 | 0.300 | -353.696 | 0.000 |
| Blog mentions | -0.197 | 0.529 | -39.657 | 0.000 |
| News mentions | -0.396 | 0.373 | -73.743 | 0.000 |

# 8. Discussion

The existence of power law characteristics have been widely explored for diverse data from many disciplinary domains, including citations, which are the most prevalent and useful impact indicator of the scholarly articles (Ruiz-Castillo, 2012; Brzezinski, 2015; Peterson et al., 2010; Radicchi, Fortunato, & Castellano, 2008; Thelwall, 2016c, 2016a). Given that the altmetric mentions are found to correlate with citations (Thelwall & Nevill, 2018; Banshal et al., 2021; Shema et al., 2014; Thelwall, 2018) and that they are observed to be skewed in nature (Thelwall & Nevill, 2018; Thelwall & Wilson, 2016; Costas et al., 2016), it would be an interesting exercise to explore if altmetric mentions follow power laws. Yet, this property had not been put to the test so far. Our approach has been to investigate the plausibility of the existence of power law behaviours in altmetric mentions. The plots in the log-log space indicate existence of power laws in different kinds of altmetric mentions as well as the composite altmetric score.

***Power Law in Altmetrics***

The first step in scrutinizing the existence of power laws in altmetric data is the log-log plot for mentions across several social media platforms- Facebook, Twitter, etc. and the composite score built from all individual scores. Altmetric.com also collects data from other platforms not mentioned here, which we have not included in the study due to the sparseness of data. The plots in **Figure 1** show an approximate straight line fit to the data plotted on a log-log plot. This is the first characteristic of power laws and thus first indication of the presence of power laws in altmetric mentions. These plots also specify the parameters of fit. These values help in analysing the nature of the plots and also determine where the majority of the distribution of data lies. The low (<=2) value of the power-law scaling exponent reflects an infinite (or divergent) mean, whereas mean is finite/converges for exponent>2. If α<2, it means that most of the mentions lies in tail of the distribution. Further, the median exists for exponent α>1 and also, variance exists if α>3. Thus, in the present case, it can be suspected that, for altmetric attention score, Facebook and blog, most of the mentions lies in the tail of the distribution.

In the **table 2**, value of R2 indicates the variance explained by the model, high value of R2 is necessary for acceptance of the power law distribution (Clauset et al., 2009). But as mentioned before, it is only necessary and not a sufficient condition for assessment. Hence, we can go for other evidence in order to assess the power law behaviour in altmetrics mentions. Thus, these parameters of fit provide evidence of existence and degree of fit, but do not prove that the power law is the sole model for the observed data. We use MLE approach for further confirmation.

***Power law exponent and MLE Approach***

From **table 3**, we observed that for twitter and Facebook mentions theis low i.e., we observe power law behaviour in almost full range, whereas in case of blog and news mentions, the power law applies only in the tail of the distribution because is much greater. Hence, we conclude that Blog and News mentions have a power law tail whereas Twitter and Facebook follow a power law behaviour in almost full range. Also, as the exponent α<2 for all the four platforms, most of the mentions lies in the tail of the distribution.

***Power law distribution and its comparison with Log-normal and Exponential***

From **table 4**, we observed sign of normalized log-likelihood is positive with very high value when we compare power law with exponential, and also that the p values are significantly small. This indicates that the normalized log-likelihood value is significantly different from zero, confirming that the power law is better fit for social media data as compared to exponential distribution. On the other hand, the sign of normalized log-likelihood is negative with very low value (close to zero) when we compare power law with log-normal. Negative sign can imply the possibility of better fit by log-normal, but smaller value normalized log-likelihood hints that power law too offer almost the same fit. In such cases, p values can confirm whether log-likelihood have a clear upper hand or not. Here the p values are larger than 0.01, indicating that the sign is not reliable and the test does not favour either model over the other (Vuong, 1989). Thus, we cannot confirm the plausibility of log-normal distribution.

***Nested distribution comparison on Altmetric data***

From **table 5**, we obseve that log-likelihood ratio is negative which shows that second distribution does provide better fit. This always holds as described by assumptions. But when we observe p values, in the case of power law and truncated power law, since p values indicates significance, the plausibility that truncated power law can be better fit than power law holds. In case of exponential vs stretched exponential, since p-values indicates no significance, there is no strong indication of either exponential fits better than stretched exponential and vice versa. Since, stretched exponential is the distribution of our interest as it can produce heavy tailed distributions like power law, and as there is not enough indication that it is better than exponential (to which it is related), the conclusion we can have now is that power laws do exist in altmetrics and most possibly altmetric data can exhibit power law at least from a cut-off point . Further, explorations with more datasets (when they evolve out of sparsity) are required to confirm whether other possible candidate distributions like log-normal and exponential can represent the altmetric data distribution more suitably. Same holds for variants of power laws like hooked power law.

So by the systematic procedure, it can be concluded that altmetrics does exhibit power law at least from a cut-off point. Now, we discuss some of the possible reasons for the exhibition of power law and implications of power law in altmetrics.

***Social media preferences and Power Law in altmetrics***

One of the major well-studied reasons for power law can be the ‘preferential attachment’ prevalent in some of the real-world systems.  In the case of altmetrics, the ‘preferential attachment’ manifests in the form of preferential sharing, mentioning, and other activities in social media and other platforms. This phenomenon is similar to citation preferences in the case of traditional indicators. The preferences in social media activity can be affected largely by the influence of the authors of scholarly publications. Also, aspects like discussion of trending/populist themes in the publications, catchy titles or the combination of both etc., can provide a leverage for such preferences. Though both of these can happen in the case of citations too, unlike altmetrics, the existence of gate-keeping mechanisms in the form of peer review processes (despite its limitations) restrict their effect to a somewhat smaller scale. Thus, whether the preferential attachment is a mostly self-organized preferential attachment or is it an influenced preferential attachment matters most.  Self-organization is the spontaneous emergence of order in several systems achieved as global order based on local interactions. These are usually not under the influence of any external agents or aspects, but triggers from an internal event. In case of altmetrics, self-organization happens if the entry of a newly published (possibly meritorious and impactful) paper (event) in scientific literature can trigger local interactions when someone notices it and shares or mentions it on social media which in turn propagates in an evolutionary fashion. However, if the social, personal and professional influence of the author(s) is the factor that triggers and propagates the local interactions (the process of sharing something that came to our notice when others share) rather than the merit of the scholarly work is not, the local interactions (i.e., sharing mentioning, etc.) cannot be regarded as spontaneous but under the influence of an external aspect namely the apparent influence of the author(s). Here, the influence of the authors acts as an aspect that triggers and propagates the interaction rather than the influence of the paper's merit and hence it cannot be regarded as self-organized preferential attachment. In other words, the degree or level of self-organization in the preferential attachment behaviour in the underlying social media that generates altmetric scores determines the reliability and usability of altmetric as an early impact indicator of scholarly output. However, the degree or level of self-organization is difficult to determine.

***Implications of the Power Law in Altmetrics***

Of the above two aspects (i.e., influence of the author(s) and discussion of populist/trending themes) that influence the preferential attachment, the influence of authors in social media platforms can have the following implications. High skewness in reputation or altmetric recognition is a major characteristic, as in the case of citations. This means, that a greater majority or major fraction of published works might struggle to receive substantial attention in social media platforms from the scientific community and therefore might be having a minor fraction of the altmetric scores. While the building up of citation scores depend on so many factors including the peer review process, the accumulation of altmetric scores from the social media attention of articles seems to be more dependent on the diffusion of information in social networks. Therefore, publications from authors who enjoy central or information regulatory positions in such platforms can be expected to have a natural advantage in terms of visibility and hence more chance for them to gain more attention/recognition. This can happen in two ways. Owing to their influence or reach, their active interventions in popularisation of their own works or works that can give mileage to their works can gain more visibility that eventually be reflected as early impact captured at altmetric platforms. Secondly, by the activity of relatively less influential, but large number of other people who tend to follow the influential people or their activities. Thus, the quantity and strength of their direct and indirect connections in the form of friends, followers, etc., (including those authors who themselves are in regulatory positions) can attribute to the regulatory power of an author. Thus, altmetric platforms can have publications from at least four major classes of scientists:

1. LSR-LSMR: Scientists with low scientific reputation (LSR) and low social media reputation (LSMR)
2. LSR-HSMR: Scientists with low scientific reputation (LSR) and high social media reputation (HSMR)
3. HSR-HSMR: Scientists with high scientific reputation (HSR) and high social media reputation (HSMR)
4. HSR-LSMR: Scientists with high scientific reputation (HSR) and low social media reputation (LSMR)

This opens up many issues that need investigation. These are:

1. Are all the occupants of such positions really meritorious/competent authors (i.e., HSR-HSMR types) or are there at least some not-so-high contributing authors (LSR-HSMR types) with high networking skills capable of making undue advantage? In other words, are all the central or regulatory positions in social networks/social media platforms are a result of real reputation or is it just because of an ‘inflated reputation’ as discussed by Hall (2014)?
2. How to determine the immunity of a social media platform towards the inflated reputation persons (LSR-HSMRs)?
3. How to improve the immunity of social media platforms and save them from ‘inflated reputation crisis’ if there is a way to determine the immunity?

These concerns cannot be addressed until the following key issue is resolved- “How to determine whether an author’s reputation is inflated or genuine?” This issue is already being addressed by the scientific community. Hall (2014) outlines the danger of the emerging practice of bestowing high value on social media communication to such an extent that these can overshadow the metrics that reflects scientific value of publications like citation indices. He also introduced a measure of discrepancy (termed as the ‘Kardashian index’) between a scientist’s social media profile and publication record. This index seems to be capable of determining whether a scientist is ‘Kardashian’ (the term he used to indicate inflated reputation persons or LSR-HSMR type) or not. That means, whether the scientist is one with an overblown profile or not. For measurement of the immunity of platforms, the magnitude of presence of Kardashians or scientists with overblown profiles have to be explored.

Thus, the effectiveness of altmetric indicators is deeply related to the immunity of altmetric platforms towards the inflated reputations. If all or most of the platforms are immune or can be made immune to the inflated reputations, the exhibition of power law can be treated as a very positive phenomenon where ‘rich get richer’ will naturally mean ‘worthy get worthier’. However, the above-mentioned key issue related to inflated reputation seems to remain as an open problem for a while and once it is properly addressed, the revelation about the existence of power law in altmetrics might strengthen the conjecture about altmetrics/altmetric scores as a forerunner of citations. This might bring scientometric community much closer to the confirmation of predictive power of altmetrics, but not without a caveat. With the existence of power law, it is clear that major share of altmetric mentions is received by a smaller share of a few works only. Though this can be expected in the case of lognormal too, scale-invariance associated with power laws might make it more robust for predictive purposes, especially if there is self-organization in play. However, as there are no sufficient indications from existing literature about this, rigorous research is required to confirm this. In fact, this can be treated as an open problem which is related to an open problem of general interest- “in what situations the small differences between power laws and log-normal distributions manifest themselves in vastly different qualitative behavior, and in what cases a power law distribution can be suitably approximated by lognormal distribution”, stated by Mitzenmacher, (2004).

Despite these issues, for practical purposes, identification of authors of works that received major share of mentions can be viewed as a major step for determination of the representation of Kardashians or LSR-HSMR types, because from that set, with the aid of Kardashian index, we can determine Kardashians. Our work, which revealed the existence of power law in altmetrics can be regarded as a stride towards the first vital step, especially when the extent/level of self-organization in the underlying social media platforms that generates altmetric data cannot be easily determined.

# 9. Conclusion

Altmetrics or alternative metrics and citations are known to exhibit many similarities and differences, of which similarities convincingly outweigh the dissimilarities. Inspired by the striking parallelism induced by the similarities including skewed distributions, we investigated whether altmetrics follow power laws like citations. Though several studies have reported that altmetrics also shows skewed distribution like citations, apart from power law, there are some distribution models like exponential and lognormal that might tend to generate skewed distribution. Thus, a systematic analysis needs was needed to confirm which model fits altmetrics data in the best possible manner. Based on the empirical analysis, we report the occurrence of power laws in altmetrics, where our data on altmetrics constitute counts of mentions about scientific articles. The existence of power laws in altmetrics can further strengthen the parallelism between altmetrics and citations. This feat was not achievable with some existing studies that explored correlation between both wherein varying strength of correlation were observed. It is highly interesting to see more and more newer systems tend to follow the power law which is ubiquitous in many natural and man-made systems. Not only do the above categories follow power laws, but we find that the composite altmetrics index (AAS), created by combining individual altmetric indices, also follows a power law (table 2 and figure 1). Our study also draws supporting evidence for empirical results. We observe that blog and news mentions have a power law tail whereas Twitter and Facebook support power law with almost complete range (table 3).

Our second objective was to check whether other distributions like log-normal and Exponential distributions are plausible for the datasets under consideration. We observe that for our data, log-normal may be a plausible fit but exponential distribution is not a good fit (figure 3 and table 4). We also compared power law with truncated power law and observed that truncated power law provides better fit than power law (table 5). Thus, further explorations may be pursued to check the plausibility of other distributions too. Especially variants of power laws like hooked power laws are much recommended because of the high goodness of fit of power law and truncated power law for these datasets. Implications of power law in altmetrics is discussed and major issues that are to be addressed to enhance the reliability of altmetrics as a tool for research performance assessment are stated. As future endeavours, one can address the key issue itself, namely the ‘determination of inflated reputation’, by possibly exploring and building up on the work by Hall (2014) that can lead to solutions to other issues.

**10. Limitations of this research**

Like most of the research works that have an empirical analysis element, this work is also prone to many limitations. Major limitation of this study roots from the inclusion of papers with ‘DOI’ only for the analysis, which may have led to the exclusion of some publications that might have non-zero altmetric scores. However, since the exact number of such publications and scores of each publication are not known, whether this exclusion might have caused a selection bias is not clear. There are no studies regarding whether publications without DOI but having some other identifier (such as PubMed ID, ISBN, Handles, arXiv IDs, etc.), which is supported by Altmetric.com, can garner same level of visibility (that may lead to citations as well as altmetric scores) as that of the ones having DOIs. Without a clear knowledge related to at least one of the above two cases, the existence of ‘selection bias’ and how it affects the generalizability of our results cannot be determined. However, we feel that, it is important to issue a caution that our study may be prone to the ‘selection bias’. Apart from this, as discussed earlier we have considered altmetric data from ‘Altmetric.com’ with respect to four variables only. Extensive studies using more variables from different altmetric databases including PlumX can shed more light on whether our results are generalizable. These two limitations also offer room for further interesting research explorations.

# Acknowledgments

The authors would like to acknowledge the access to Altmetric.com data provided by Stacy Konkiel, Director of Research Relations at Digital Science.

# Funding Support

This work is supported by extramural research grant no: MTR/2020/000625 from Science and Engineering Research Board (SERB), India to the fourth author.

# References

Adamic, L. A. (2000). Zipf, power-laws, and pareto-a ranking tutorial. *Xerox Palo Alto Research Center, Palo Alto, CA, Http://Ginger. Hpl. Hp. Com/Shl/Papers/Ranking/Ranking. Html*.

Adamic, L. A., & Huberman, B. A. (2001). The Web’s hidden order. *Communications of the ACM*, *44*(9), 55–60.

Banshal, S. K., Singh, V. K., & Muhuri, P. K. (2021). Can altmetric mentions predict later citations? A test of validity on data from ResearchGate and three social media platforms. In *Online Information Review:* Vol. 45 No. 3, pp. 517-536. https://doi.org/10.1108/OIR-11-2019-0364

Banshal, S.K., Singh, V.K., Kaderye, G., Muhuri, P.K. and Sánchez, B.P. (2018), “An altmetric analysis of scholarly articles from India”, *Journal of Intelligent & Fuzzy Systems*, Vol. 34 No. 5, pp. 3111–3118.

Banshal, S.K., Singh, V.K., Muhuri, P.K. and Mayr, P. (2019a), “How much research output from India gets social media attention?”, *Current Science*, Vol. 117 No. 5, pp. 753–760.

Banshal, S.K., Singh, V.K., Muhuri, P.K. and Mayr, P. (2019b), “Disciplinary Variations in Altmetric Coverage of Scholarly Articles”, *In 17th INTERNATIONAL CONFERENCE ON SCIENTOMETRICS & INFORMETRICS (ISSI)*, pp. 1870–1881.

Bar-ilan, J. (2014), “Astrophysics publications on arXiv , Scopus and Mendeley : a case study”, *Scientometrics*, Vol. 100 No. 1, pp. 217–225.

Bouchaud, J.-P. (2001). Power laws in economics and finance: some ideas from physics. *Quantitative Finance*, *1*(1), 105–112. https://doi.org/10.1080/713665538

Bradford, S. C. (1934). Sources of information on specific subjects. *Engineering*, *137*, 85–86.

Brzezinski, M. (2015). Power laws in citation distributions: evidence from Scopus. *Scientometrics*, *103*(1), 213–228. https://doi.org/10.1007/s11192-014-1524-z

Chen, K., Tang, M., Wang, C. and Hsiang, J. (2015), “Exploring alternative metrics of scholarly performance in the social sciences and humanities in Taiwan”, *Scientometrics*, Vol. 102 No. 1, pp. 97–112.

Chen, Y., & Leimkuhler, F. F. (1986). A relationship between Lotka’s law, Bradford’s law, and Zipf’s law. *Journal of the American Society for Information Science*, *37*(5), 307–314.

Clauset, A., Shalizi, C. R., & Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Review*, *51*(4), 661–703. https://doi.org/10.1137/070710111

Costas, R., Zahedi, Z. and Wouters, P. (2015a). The thematic orientation of publications mentioned on social media: Large-scale disciplinary comparison of social media metrics with citations.*Aslib Journal of Information Management*, Vol. 67 No. 3, pp. 260–288.

Costas, R., Zahedi, Z., & Wouters, P. (2015b). Do “altmetrics” correlate with citations? Extensive comparison of altmetric indicators with citations from a multidisciplinary perspective. *Journal of the Association for Information Science and Technology*, *66*(10), 2003–2019. https://doi.org/10.1002/asi.23309

Costas, R., Haustein, S.; Zahedi, Z.; Larivière, V. (2016). Exploring paths for the normalization of altmetrics: Applying the Characteristic Scores and Scales. The 2016 Altmetrics Workshop, Bucharest, Romania, 27 September 2016. Available at: https://hdl.handle.net/1887/47068

Crawford, T. W. (2020). Scale, Analytical. In A. Kobayashi (Ed.), *International Encyclopedia of Human Geography (Second Edition)* (Second Edition, pp. 89–96). Elsevier. https://doi.org/https://doi.org/10.1016/B978-0-08-102295-5.10024-1

Duan, Y., & Xiong, Z. (2017). Download patterns of journal papers and their influencing factors. *Scientometrics*, *112*(3), 1761–1775. https://doi.org/10.1007/s11192-017-2456-1

Egghe, L. (2005). Power laws in the information production process: Lotkaian informetrics. Elsevier Ltd.

Egghe, L. (2009). Performance and its relation with productivity in Lotkaian systems. *Scientometrics*, 81(2), 567-585.

Eysenbach, G. (2011). Can tweets predict citations? Metrics of social impact based on Twitter and correlation with traditional metrics of scientific impact. *Journal of Medical Internet Research*, *13*(4). https://doi.org/10.2196/jmir.2012

Faloutsos, M., Faloutsos, P., & Faioutsos, C. (1999). On power-law relationships of the internet topology. *Computer Communication Review*, *29*(4), 251–261.

Gabaix, X. (2009). Power laws in economics and finance. *Annu. Rev. Econ.*, *1*(1), 255–294.

Gabaix, X. (2016). Power laws in economics: An introduction. *Journal of Economic Perspectives*, *30*(1), 185–206.

Hall, N. (2014). The Kardashian index: a measure of discrepant social media profile for scientists. *Genome biology*, 15(7), 1-3.

Hammarfelt, B. (2014), “Using altmetrics for assessing research impact in the humanities”, *Scientometrics*, Vol. 101 No. 2, pp. 1419–1430.

Haustein, S., Peters, I., Sugimoto, C. R., Thelwall, M., & Larivière, V. (2014). Tweeting biomedicine: An analysis of tweets and citations in the biomedical literature. *Journal of the Association for Information Science and Technology*, *65*(4), 656–669.

Holmberg, K. and Thelwall, M. (2014), “Disciplinary differences in Twitter scholarly communication”, *Scientometrics*, Vol. 101 No. 2, pp. 1027–1042.

Htoo, T.H.H. and Na, J.-C. (2017), “Disciplinary Differences in Altmetrics for Social Sciences”, *Online Information Review*, Vol. 41 No. 2, pp. 235–251.

Jeff Alstott, E. B. (2014). powerlaw: A Python Package for Analysis of Heavy-Tailed Distributions. PLoS ONE , 9 (1).

Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences*, *16*(12), 317–323.

Mitzenmacher, M. (2004). A brief history of generative models for power law and log-normal distributions. *Internet Mathematics*, *1*(2), 226–251. https://doi.org/10.1080/15427951.2004.10129088

Mohammadi, E., Thelwall, M., & Kousha, K. (2016). Can Mendeley bookmarks reflect readership? A survey of user motivations. *Journal of the Association for Information Science and Technology*, *67*(5), 1198-1209

Mohammadi, E. and Thelwall, M. (2013), “Assessing non-standard article impact using F1000 labels”, *Scientometrics*, Vol. 97 No. 2, pp. 383–395.

Newman, M. (2015). *Networks*. Oxford University Press. ISBN 978-0-19-920665-0.

Newman, M. E. J. (2005). Power laws, Pareto distributions and Zipf’s law. *Contemporary Physics*, *46*(5), 323–351. https://doi.org/10.1080/00107510500052444

Ortega, J.L. (2015), “Disciplinary differences in the use of academic social networking sites”, *Online Information Review*, Vol. 39 No. 4, pp. 520–536.

Pareto, V. (1896). *Coursd’économie politique: professé à l’Universi̧té de Lausanne* (Vol. 1). F. Rouge.

Peoples, B. K., Midway, S. R., Sackett, D., Lynch, A., & Cooney, P. B. (2016). Twitter predicts citation rates of ecological research. *PLoS ONE*, *11*(11), 1–11. https://doi.org/10.1371/journal.pone.0166570

Peterson, G. J., Pressé, S., & Dill, K. A. (2010). Nonuniversal power law scaling in the probability distribution of scientific citations. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(37), 16023–16027. https://doi.org/10.1073/pnas.1010757107

Price, D. J. D. S. (1965). Networks of scientific papers. *Science*, 510–515.

Price, D. J. D. S. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, *27*(5), 292–306.

Priem, J., & Hemminger, B. H. (2010). Scientometrics 2.0: New metrics of scholarly impact on the social Web. *First Monday*, *15*(7), 1–19.

Priem, J., Taraborelli, D., Groth, P., & Neylon, C. (2010). *altmetrics: a manifesto*. 1–77. http://altmetrics.org/manifesto/

Radicchi, F., Fortunato, S., & Castellano, C. (2008). Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences*, *105*(45), 17268–17272.

Redner, S. (1998). Rapid Note How popular is your paper? An empirical study of the citation distribution. *THE EUROPEAN PHYSICAL JOURNAL B*, *4*(2), 131–134.

Redner, S. (2005). Citation statistics from 110 years of Physical Review. *Physics Today*, *58*(6), 49–54. https://doi.org/10.1063/1.1996475

**﻿**Reed, W. J., & Hughes, B. D. (2002). From gene families and genera to incomes and internet file sizes: Why power laws are so common in nature. *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*, *66*(6), 4. https://doi.org/10.1103/PhysRevE.66.067103

Ruiz-Castillo, J. (2012). The evaluation of citation distributions. SERIEs 3, 291–310. https://doi.org/10.1007/s13209-011-0074-3

Shema, H., Bar-IIan, J., & Thelwall, M. (2014). Do Blog Citations Correlate With a Higher Number of Future Citations? Research Blogs as a Potential Source for Alternative Metrics. *JOURNAL OF THE ASSOCIATION FOR INFORMATION SCIENCE AND TECHNOLOGY*, *65*(5), 1018–1027. https://doi.org/10.1002/asi

Sotudeh, H., Mazarei, Z. and Mirzabeigi, M. (2015), “CiteULike bookmarks are correlated to citations at journal and author levels in library and information science”, *Scientometrics*, Springer Netherlands, Vol. 105 No. 3, pp. 2237–2248.

Thelwall, M. (2016a). Are the discretised log-normal and hooked power law distributions plausible for citation data? *Journal of Informetrics*, *10*(2), 454–470. https://doi.org/10.1016/j.joi.2016.03.001

Thelwall, M. (2016b). Are there too many uncited articles? Zero inflated variants of the discretised log-normal and hooked power law distributions. *Journal of Informetrics*, *10*(2), 622–633. https://doi.org/10.1016/j.joi.2016.04.014

Thelwall, M. (2016c). The discretised log-normal and hooked power law distributions for complete citation data: Best options for modelling and regression. *Journal of Informetrics*, *10*(2), 336–346. https://doi.org/10.1016/j.joi.2015.12.007

Thelwall, M. (2018). Early Mendeley readers correlate with later citation counts. *Scientometrics*, *115*(3), 1231–1240. https://doi.org/10.1007/s11192-018-2715-9

Thelwall, M. and Kousha, K. (2017), “ResearchGate articles: Age, discipline, audience size, and impact”, *Journal of the Association for Information Science and Technology*, Vol. 68 No. 2, pp. 468–479.

Thelwall, M., & Sud, P. (2016). Mendeley Readership Counts: An Investigation of Temporal and Disciplinary Differences. *Journal of the Association for Information Science and Technology*, *67*(12), 3036–3050. https://doi.org/10.1002/asi

Thelwall, M., & Wilson, P. (2016). Mendeley Readership Altmetrics for Medical Articles: An Analysis of 45 Fields. *Journal of the Association for Information Science and Technology*, *67*(8), 1962–1972. https://doi.org/10.1002/asi

Thelwall, M., & Nevill, T. (2018). Could scientists use Altmetric.com scores to predict longer term citation counts? *Journal of Informetrics*, *12*(1), 237–248.

Vogl, S., Scherndl, T. and Ku, A. (2018), “# Psychology : a bibliometric analysis of psychological literature in the online media”, *Scientometrics*, Vol. 115 No. 3, pp. 1253–1269.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica , 307–333.

Wang, X., Fang, Z., Li, Q. and Guo, X. (2016a), “The Poor Altmetric Performance of Publications Authored by Researchers in Mainland China”, *Frontiers in Research Metrics and Analytics*, Vol. 1, p. 8.

Wang, X., Fang, Z., & Sun, X. (2016b). Usage patterns of scholarly articles on Web of Science: a study on Web of Science usage count. *Scientometrics*, *109*(2), 917–926. https://doi.org/10.1007/s11192-016-2093-0

Zahedi, Z., Costas, R. and Wouters, P. (2014), “How well developed are altmetrics? A cross-disciplinary analysis of the presence of ‘alternative metrics’ in scientific publications”, *Scientometrics*, Vol. 101 No. 2, pp. 1491–1513.

Zipf, G. K. (1949). Human behaviour and the principle of least-effort. Cambridge MA edn. Reading: Addison-Wesley.

1. Corresponding author. Email: [vivek@bhu.ac.in](mailto:vivek@bhu.ac.in) [↑](#footnote-ref-1)
2. [powerlaw/powerlaw.py at master · jeffalstott/powerlaw · GitHub](https://github.com/jeffalstott/powerlaw/blob/master/powerlaw.py) [↑](#footnote-ref-2)