ISSI 2023 Proceedings, v1, pp 139–154
DOI: 10.5281/zenodo.8305637 https://cns-iu.github.io/workshops/2023-07-02_issi/

# On the quality of altmetric data: An exploratory analysis using Benford's Law

Solanki Gupta[1], Vivek Kumar Singh[2], Sumit Kumar Banshal[3]

[1]*solankigupta2@gmail.com,* [2]*vivek@bhu.ac.in*
Department of Computer Science, Banaras Hindu University, Varanasi-221005 (India)

[3]*sumitbanshal06@gmail.com*
Department of Computer Science & Engineering, Alliance University, Bengaluru (India)

## Abstract
Altmetrics, or alternative metrics, have gained a lot of popularity during last few years. It is now being explored for several purposes ranging from impact measure to research evaluation. Researchers are even exploring use of altmetrics as a proxy for societal impact of research. However, at the same time there are several studies which have cautioned about use of altmetrics on account of quality and reliability of altmetric data. This study proposes a framework to evaluate the quality of altmetric data by applying Benford Law. A large sized altmetric data sample is considered and the fits with Benford's Law are computed. Results for fit on first and second leading digit of altmetric data show adherence to Benford's Law and hence provides evidence towards authenticity of altmetric data considered. Statistical tests confirm the findings. The study suggests that Benford's Law fit can be used as a method to test the quality of altmetric data. Further, the use of a large-sized data sample is likely to reduce the impact of manipulations in altmetric data. Relevant implications of the research are discussed.

**Keywords:** Altmetrics, Altmetric data quality, Benford Distribution, Benford's Law, Social media mentions, Scientometrics.

## Introduction
Altmetrics is a web-based metric that focuses on scholarly influence measured via online tools and environments. A variety of sources are included in altmetrics, such as blog mentions, citations, Wikipedia articles, tweets, Facebook updates, recommendation services, readership data on social reference managers and bookmarking websites. Altmetrics accrue very quickly (Mohammadi & Thelwall, 2014; Bornmann, 2014b; Haustein et al., 2014a; Priem et al., 2011) and offer access to a wider range of viewpoints from audiences including experts, college students, the government, and the general public (Bornmann, 2014b). Some researchers are even exploring it as a proxy measure for societal impact of research (Tahamtan & Bornmann, 2020; Thelwall, 2021; Bornmann, 2014a; Garcovich & Adobes Martin, 2020). Altmetric data is also used to address how research has a broader influence on policy, including effects on clinical practices, technical applications, education, health policy, and other areas (Haustein et al., 2014a; Haustein et al., 2014b). Nowadays, use of altmetrics for funding, recruiting, tenure, and promotion decisions is also being explored (Lin, 2012; Thelwall et al., 2015). Influenced by this, many researchers actively promote themselves and their work on different social media platforms. Sometimes use of bots is also seen in such dissemination. Altmetric data are collected from various platforms usually through some public APIs or a web-crawling or scraping method.

The trend towards use of altmetrics for various purposes has also resulted in debate about its quality and reliability. Some researchers argue that altmetrics should be avoided (Cheung, 2013) or should only be used in combination with more traditional metrics to quantify scholarly impact (Bornmann, 2014a) because of issues with data quality and validity (Haustein, 2014). It has also been argued that altmetrics can be distorted and manipulated, for example, by creating several user accounts or deploying bots to increase metrics (Cheung, 2013). According to a

recent study on the impact of bots on tweet counts, up to 9% of tweets are found to be generated by bots (Haustein et al., 2016). Altmetrics also suffers from accidental manipulation as the research articles may be discussed on the social web for negative purposes, such as to criticise them (Shema, Bar-Ilan & Thelwall, 2012), to accuse the authors of fraud, to discuss papers that have been retracted (Marcus & Oransky, 2011), for irrelevant purposes like spam or automated mentions (for example, a journal tweeting all of its articles when they are published), or simply because they have interesting or fascinating titles. These manipulations are linked with the altmetric gaming (Roemer & Borchardt, 2015; Strielkowski & Chigisheva, 2018). Despite these issues, altmetrics is still gaining popularity and credibility amongst the broader research community, including being explored for research evaluation purposes. However, the quality of altmetric data is of paramount importance for establishing trust in its use.

This article presents a new approach to test the quality of altmetric data by applying "Benford's Law" (Kössler, Lenz & Wang, 2019) on the altmetric data. Although this law is in existence for quite some time and have been extensively researched, it has not been explored for evaluating the quality of altmetric data. It is considered an interesting tool/ method for data quality assessment and to detect the anomalies in the data. These anomalies can be due to various reasons including due to data being manipulated/ gamed. In a given data set, Benford's Law can identify the probability of very probable or highly unlikely frequencies of numbers. The probabilities are calculated using mathematical logarithms for individual digits in large data sets of randomly generated numbers. Benford's Law can be used to catch that these numbers are purposely manipulated or not. To verify the data quality, empirical distribution of data is compared with Benford distribution and then the fits are visualized for anomaly detection. It can help in answering a question of the form "what is the quality and authenticity of altmetric data?" Therefore, it presents a highly suitable choice for evaluating the quality of altmetric data and there lies the novelty of this work.

**Benford's Law**

Benford's Law is an observation about the frequency distribution of leading or first digits and called as "Law of first digit" or "phenomenon of Significant digits" or "The Law of Anomalous Numbers**".** It is also referred as the Newcomb-Benford Law since physicist F. Benford reportedly made a new discovery of the law in 1938 (Benford, 1938) after it was first reported by polymath Newcomb in 1881 (Newcomb, 1881). It describes a fixed probability distribution for the leading digits in natural and social processes.

This probability distribution for first digit is described as –

$$\mathrm{P(d)} = \log_{10}(\mathrm{d}+1) - \log_{10}(\mathrm{d}) = \log_{10}(1+\frac{1}{d})$$

The first digit described by Benford's Law satisfies a uniform logarithmic distribution. The expected digit distribution is given in Table 1. The first digit distribution ranges from 1 to 9 whereas the second digit distribution takes the values from 0 to 9.

**Table 1: Expected digit distribution specified by Benford's Law**

| Digits | First digit (%) | Second digit (%) |
|---|---|---|
| 0 | | 11.9 |
| 1 | 30.1 | 11.3 |
| 2 | 17.6 | 10.8 |
| 3 | 12.4 | 10.4 |

| | | |
|---|---|---|
| 4 | 9.6 | 10.0 |
| 5 | 7.9 | 9.6 |
| 6 | 6.7 | 9.3 |
| 7 | 5.8 | 9.0 |
| 8 | 5.1 | 8.7 |
| 9 | 4.5 | 8.5 |

Before applying this law to any empirical data, that dataset should fulfil the following criteria (Tošić & Vičič, 2021):

1. The numbers need to be random and not assigned with no imposed minimums or maximums.
2. The number should cover several orders of magnitude (e.g., data with plenty of values in the hundreds, thousands, tens of thousands, etc.).
3. Data is right skewed i.e.; the distribution has a long right-tail rather than being symmetric.
4. Dataset should preferably cover at least 1000 samples (though Benford's law has been shown to hold true for datasets containing as few as 50 numbers).

Benford's Law provides a method which is useful to quantify the data quality. If a data set does not follow Benford's Law, it implies that either data was not generated in a totally random manner, or it used a restricted or manipulated set of numbers as a potential leading digit. The implications of testing the Benford distribution have been well established (Hill, 1995a, 1995b). Furthermore, the possibility of turning Benford's Law into a real tool able to detect the falsification and the frauds in several domain such as election fraud detection (Nigrini, 2012; Pericchi & Torres, 2011; Mebane, 2011), trade fraud detection (Cerioli et al., 2019), image manipulation detection (Iorliam et al., 2014), accounting fraud detection (Durtschi, Hillison & Pacini, 2004), scientific fraud detection (Horton, Kumar & Wood, 2020) etc. has been explored. Beyond fraud detection tool, the usefulness of Benford's Law to ascertain overall data quality of microeconomic data (Gonzalez-Garcia & Pastor, 2009) has also been seen. Benford's Law has been demonstrated as simple and objective tool in data quality evaluation for cancer registry data (Crocetti & Randi,2016), survey dataset (Kaiser, 2019), CIC data (Huang, Niu & Yang, 2020) etc. It has also been used to assess the data quality of Covid 19 dataset (Idrovo, & Manrique-Hernández, 2020; Lee, Han & Jeong, 2020; Silva, & Figueiredo Filho, 2021; Natashekara, 2022).

Recently, this law has been substantially used to validate bibliometric indexes including the number of articles, number of articles per researchers, number of journal categories, number of countries, impact factors, half-life, and immediacy score etc. (Alves, Yanasse & Sohma, 2014; 2016; Tošić & Vičič, 2021). The distribution of citations has also been found to follow Benford's Law (Campanario & Coslado, 2011; Mir, 2016). Since the altmetrics is known to correlate with citations in different degrees (Banshal et al.,2021; Costas, Zahedi & Wouters, 2015; Peoples et al., 2016; Thelwall, 2018; Thelwall & Nevil, 2018), it would be quite interesting to explore whether similar patterns exist for altmetrics data too.

Benford's Law may be explored in the domain of altmetric data for following additional reasons as well:

- Altmetric mentions are randomly generated.
- It includes data which have multiple order of magnitude.
- The process of generating the Altmetric mentions follow power laws (Banshal et al., 2022).

It is claimed that Benford's Law tends to be most accurate when values are distributed across multiple orders of magnitude, especially if the process generating the numbers is described by a power law (Pietronero et al., 2001; Golbeck, 2015). Hence the results of this study should be more accurate and helpful to gain more insights about altmetric mentions.

**Related work**

Altmetrics is a significantly researched domain nowadays, with a variety of studies. The altmetric events across various platforms like blog, Twitter, Mendeley, CiteULike etc., are being explored for a variety of reasons including to understand the relationship of altmetrics with citations (Herrmannova, Stahl & Patton, 2018; Peters et al., 2016; Shema, Bar-IIan & Thelwall, 2014; Sotudeh, Mazarei & Mirzabeigi, 2015; Thelwall, 2018; Thelwall et al., 2013; Thelwall & Nevill, 2018; Zahedi, Costas & Wouters, 2014; Banshal, Singh & Muhuri, 2021; Ortega, 2016; Snijder, 2016). Some of these studies have shown that the presence of an article in social media platforms enhances the probability of the article getting cited. The existence of power laws is altmetrics has also been confirmed by a recent study (Banshal et al., 2021; 2022) just as it has been found for citation distribution (Brzezinski, 2015). Some researchers are even exploring it as a proxy measure for societal impact of research (Tahamtan & Bornmann, 2020; Thelwall, 2021; Bornmann, 2014a; Garcovich & Adobes Martin, 2020). The use of altmetrics for funding, recruiting, tenure, and promotion decisions has also been explored (Lin, 2012; Thelwall et al., 2015).

Though altmetrics is becoming popular for a variety of reasons, it is also being questioned on account of data quality, manipulation and gaming. The altmetric comprises of a wide variety of data drawn from heterogenous sources and hence the data heterogeneity is a major feature of altmetric data. This has been seen as one of the possible reasons for data quality issues (Haustein, 2016). The sparsity of data is another concern for scientists while evaluating altmetrics (Priem, Piwowar & Hemminger, 2012). There are different ways to handle a bunch of zeros in Altmetrics data (Thelwall & Nevill, 2018), however, the generation of data might be problematic as the some recent trends or eye-catchy titles might have drawn a lot of attention across social media platforms (Haustein et al., 2014b). Moreover, the quality of image or pictures might be another reason to attract more transactions across online platforms (Finch, O'Hanlon & Dudley, 2017). For obvious reasons due to the nature of social media platforms, the data can be manipulated even in the case of specialized platforms like Mendeley (Mohammadi & Thelwall, 2019). Additionally, the presence of bots can do the harm of illegal manipulation of this kind of data (Haustein et al., 2016). The behavioural patterns of publishers can also influence the social media transaction based on their integration of tools (Karmakar, Banshal & Singh, 2020). As the altmetric mentions attracts the eyes of even funding agencies to assess young researchers (Thelwall et al., 2015), there would be more problems of artificially inflating the altmetric data. The commercial nature of these kind of event data makes the data more prone to manipulations (Adie & Roe, 2013). The altmetric data is believed to be more susceptible towards the 'Gaming' for several obvious reasons (Roemer & Borchardt, 2015). Therefore, it is important that a method may be developed to test the altmetric data quality.

Benford's Law offers an effective mechanism for testing data quality (Gonzalez-Garcia & Pastor, 2009; Crocetti & Randi, 2016; Huang, Niu & Yang, 2020) and data reliability (Kaiser, 2019). Benford's Law has been used in diverse domains. For example, in Economics it is utilized for identifying the regularities and discrepancies of credit default swaps (Ausloos, Castellano & Cerqueti, 2016), and in Finance it is utilized as a decision support device for financial investments and helpful to recognize financial risks (Cerqueti, Maggi & Riccioni,2022) etc. Riccioni & Cerqueti, (2018) used Benford's Law for analysis of the statistical regularity of financial data and assess about the reliability of stock exchange data of

several countries, specifically prices and volume of stock. It is claimed that Benford's Law tends to be most accurate when values are distributed across multiple orders of magnitude, especially if the process generating the numbers is described by a power law (Pietronero et al., 2001; Golbeck, 2015). In case of altmetrics, these conditions are met and hence Benford's Law can be applied in the domain of altmetrics. This article thus attempts to bridge the research gap about assessment of quality of altmetric data by applying Benford's Law for the purpose.

**Data & methodology**

The altmetric data for evaluation is downloaded from altmetric aggregator Altmetric.com. The data for the whole year of 2021 was downloaded. The download was done in the month of Sep. 2022. A total of 1,787,976 items are found with different altmetric mentions. The downloaded altmetric data included 47 columns, out of which 18 columns are indicators based on social media activities of the scholarly publications. These indicators are 'Altmetric Attention Score', 'News mentions', 'Blog mentions', 'Policy mentions', 'Patent mentions', 'Twitter mentions', 'Peer review mentions', 'Weibo mentions', 'Facebook mentions', 'Wikipedia mentions', 'Google+ mentions', 'LinkedIn mentions', 'Reddit mentions', 'Pinterest mentions', 'F1000 mentions', 'Q&A mentions', 'Video mentions', 'Syllabi mentions', and 'Number of Mendeley readers'. The field of 'Number of Dimensions citations' was also downloaded. The downloaded data was analysed on each of these fields to find out which fields satisfy the criteria for applying Benford's Law. It was found that only 'Altmetric Attention Score', 'Twitter mentions', 'Number of Mendeley readers' and 'Number of Dimensions citations' had enough data values (Criterion 1), and the data values comprise several orders of magnitude (Criterion 2). The other altmetric data fields don't have enough data. Therefore, the analysis was performed only on the following altmetric data fields (as they satisfy the given criteria): 'Altmetric Attention Score', 'Twitter mentions', and 'Number of Mendeley readers', along with 'Number of Dimensions citations' as an additional field. Further in order to avoid analytic distortion, we evaluated summaries of the mean, standard deviation, skewness, and Kurtosis for the selected altmetric fields (Table 2).

It can be observed that for all the four fields, the mean values are quite higher, and all distributions appear positively skewed. The tailedness of a distribution using kurtosis is also measured to distinguish the nature of distribution relative to a normal distribution. It is observed that all the mentions have high kurtosis, that led to flat tail distribution rather than normal distribution. Thus, these statistics shows that the data of the selected fields is suitable for studying Benford's Law.

**Table 2. Descriptive Statistics for Altmetric mentions data**

| Mentions | Mean | Standard Deviation | Skewness | Kurtosis |
|---|---|---|---|---|
| **AAS** | 25.15 | 188.94 | 66.96 | 8325.25 |
| **Twitter Mentions** | 26 | 410.55 | 138.62 | 31957.56 |
| **Number of Mendeley Readers** | 19.81 | 19.81 | 103.99 | 32472.30 |
| **Number of Dimensions citations** | 4.8 | 30.23 | 280.63 | 135693.26 |

Computer programs were written in Python to assess the distribution of the retrieved first and second digits of each mention. The data for the field of 'Number of Dimensions Citations' is also analysed. For each of these mentions, the frequency of occurrence of the extracted digits is plotted. For the first digit and the second leading digit, this empirical distribution is compared with the distribution predicted by Benford's Law. As will be seen later, the Benford distribution fits well with the altmetric mentions. The statistical goodness of fit tests are used to further

confirm the fit of the empirical data. Some of the well-known statistics used for such purpose are Pearson's Chi squared and Kolmogorov-Smirnov D.

The Chi squared test has an excess power problem because as the number of observations increases (it is believed to reach 5000 records and above (Nigrini, 2012)), it becomes more susceptible to insignificant spikes, which leads to the conclusion that the data does not comply (Golbeck, 2015; Tošić & Vičič, 2021). Since each mention in the dataset includes more than 5000 entries, the Pearson's Chi squared test is not appropriate to determine the goodness of fit for this study.

The alternative one i.e., Kolmogorov–Smirnov statistic for goodness of fit is applied next. The Kolmogorov-Smirnov statistic measures the difference between the reference distribution's cumulative distribution function and the sample's empirical distribution function and is described as follows-

$$D_{max} = \sup_{x \epsilon R} |F_n(x) - F_o(x)|$$

To apply this test, a hypothesis testing approach is followed. A null hypothesis (H0) which considers that both empirical and theoretical (Benford) distributions are same is developed. A statistical hypothesis test is a technique for determining if the available data are sufficient to support a specific hypothesis. One can make probabilistic claims regarding population parameters through hypothesis testing. Here, the level of significance (α) is chosen as 0.05 or 95% of confidence interval for the assessment of results of test statistics and interpretation of hypothesis.

**Results**

Since real data doesn't always follow an exact distribution, most studies use graphical representations to highlight any unusual patterns in the data. Therefore, first the plots are presented to show the fit for the empirical data. The first digit distribution for all selected indicators (including Dimensions citations) can be found in Figure 1 - 4. For all the figures, the empirical distribution is presented along with the expected Benford distribution. It is observed that the indicators follow the Benford distribution quite well.

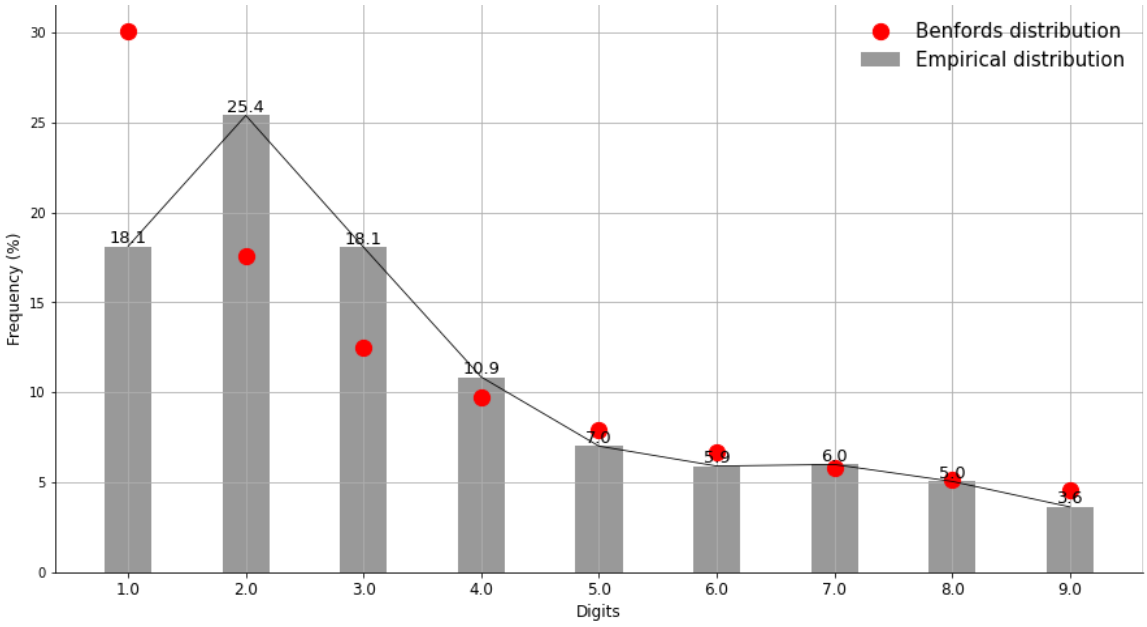


**Figure 1: First digit distribution of Altmetric Attention Score**

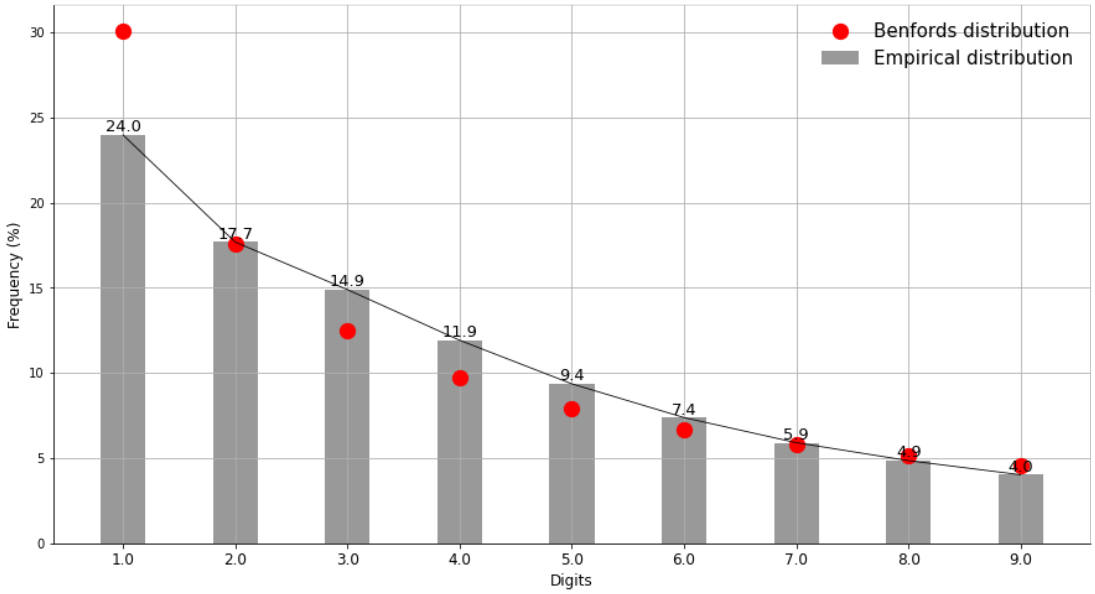


**Figure 2: First digit distribution of Twitter mentions**

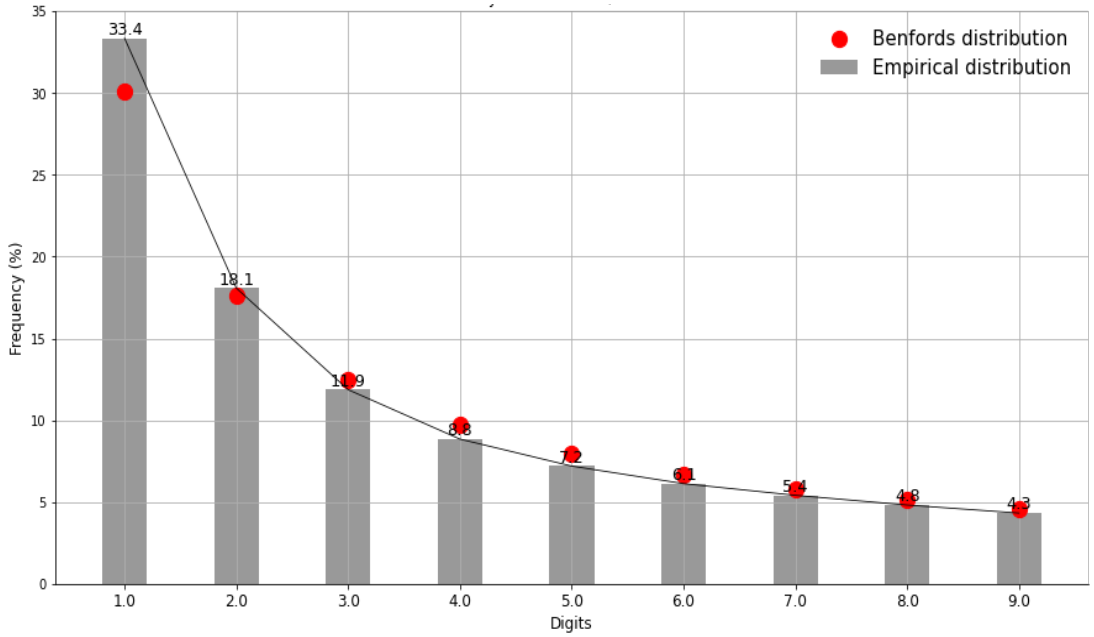


**Figure 3: First digit distribution of Number of Mendeley readers**

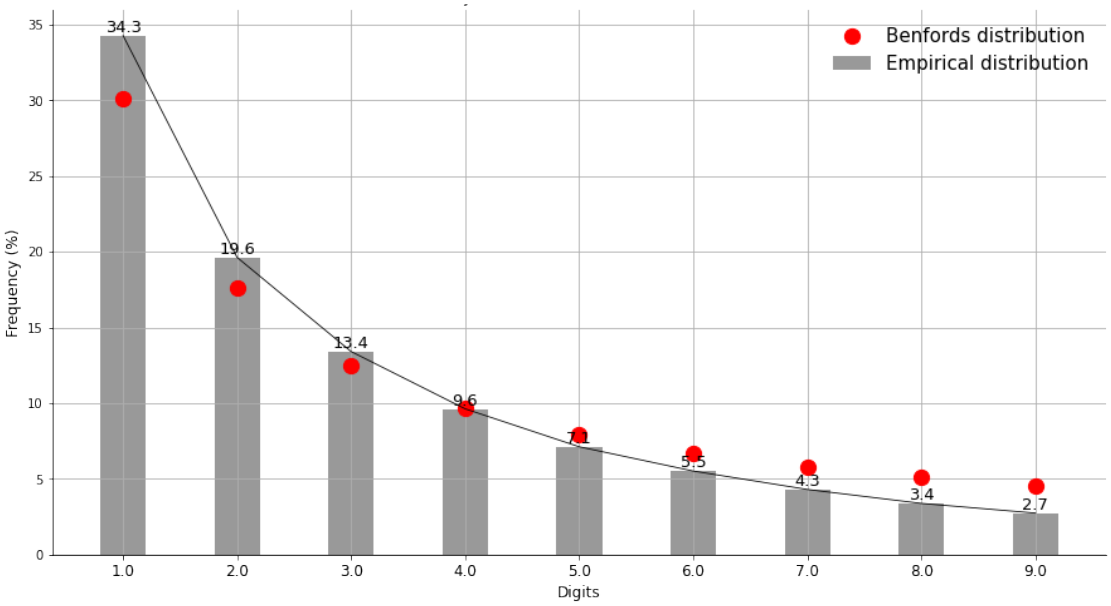


**Figure 4: First digit distribution of Number of Dimension citations**

To further evaluate about the agreement of Benford's Law with altmetric mentions, this law is extended for second digits as well. The plots are accordingly drawn for the second leading digit. It is observed that the second digit distribution also follows Benford distribution, although this distribution is considerably flatter (Figure 5 to 8).

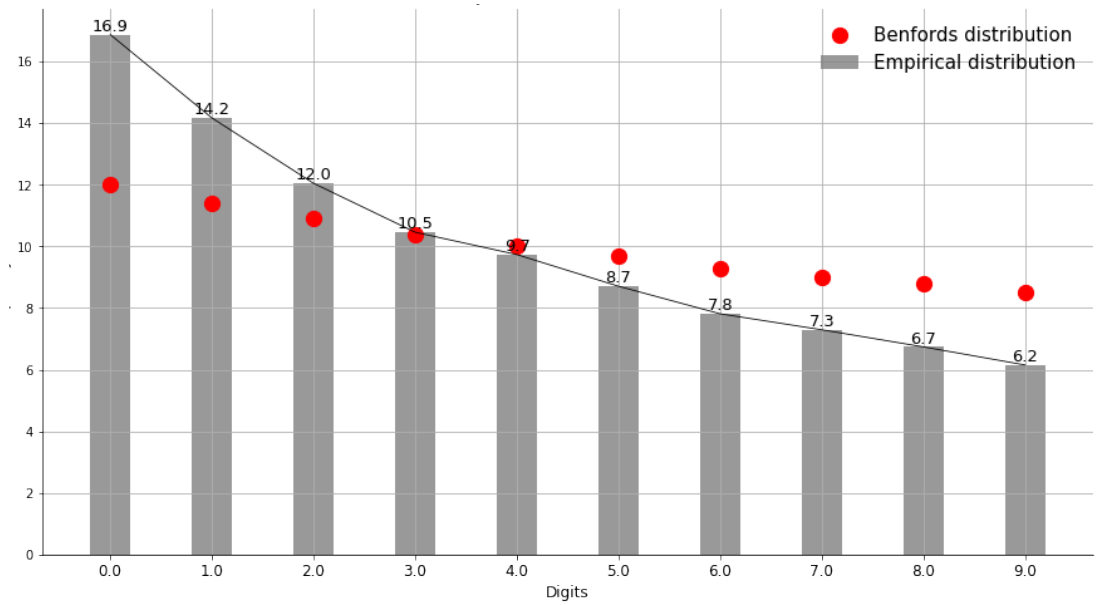


**Figure 5: Second digit distribution of Altmetric Attention Score**

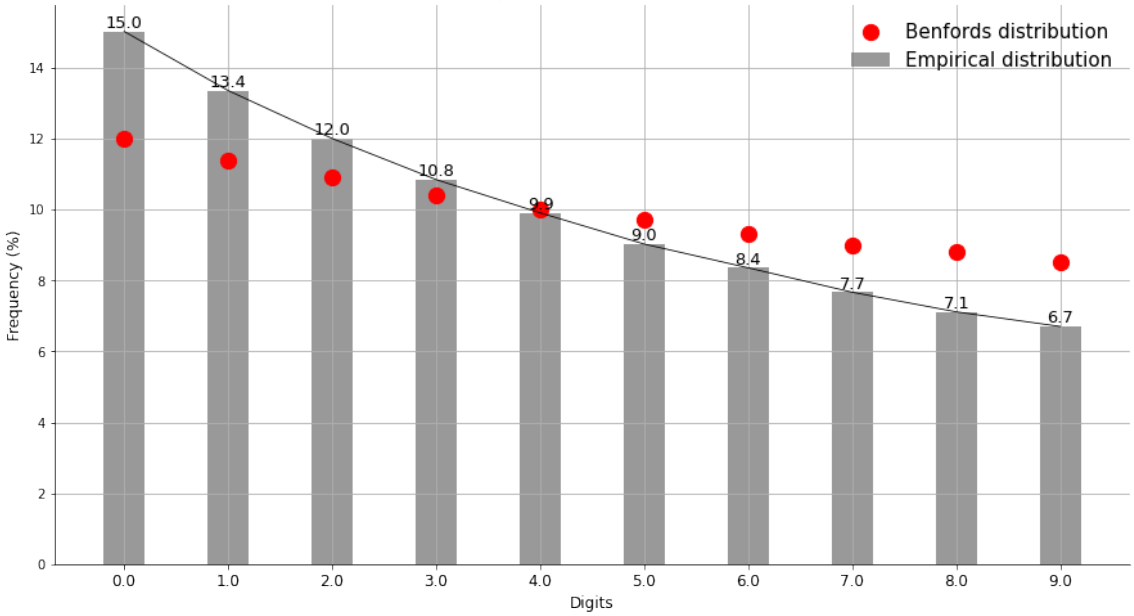


**Figure 6: Second digit distribution of Twitter mentions**

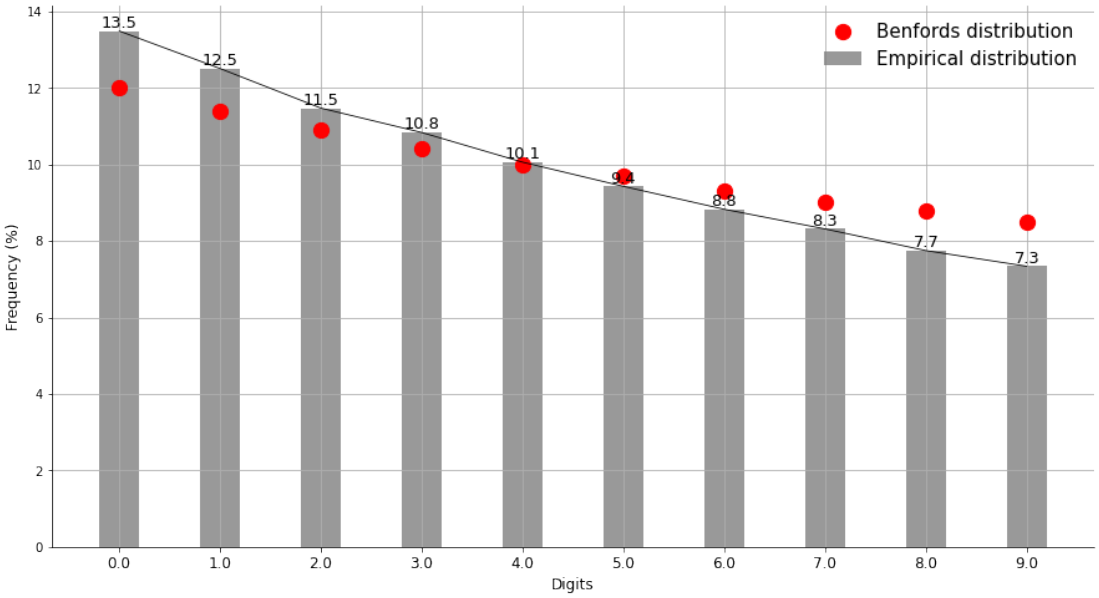


**Figure 7: Second digit distribution of Number of Mendeley readers**

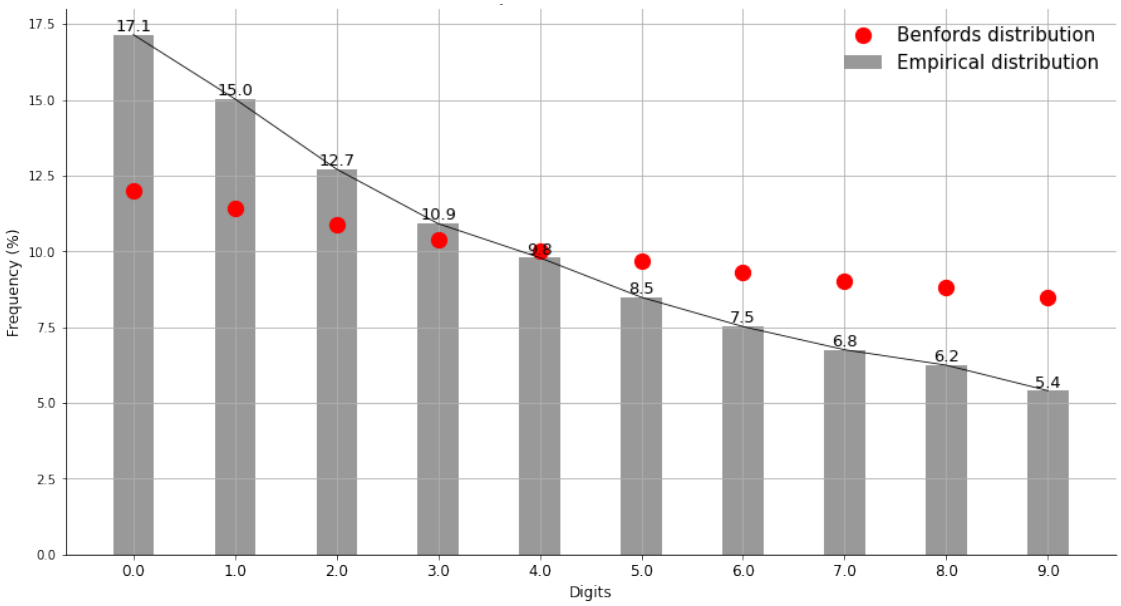


**Figure 8: Second digit distribution of Number of Dimension citations**

There are two type of discrepancies that may occur for the digits- undershoot or overshoot with the expected values. The spikes above Benford's Law line (i.e., overshoot) are the numbers of interest as they reflect the cases that involve some kind of manipulation. It is observed that there is significant overshoot appearing for digits 0, 1, 2 and 3 in second digit distribution of all altmetric mentions. However, a consistent overshoot for all the altmetric mentions in case of first digits is not observed. For example, overshoot appeared for digit 2, 3 and 4 for altmetric attention score; in digit 3, 4 and 5 for twitter; in digit 1 for number of Mendeley readers; in 1, 2 and 3 for number of dimensions citations. However, results with a digit lower than the probability of occurrence are usually ignored.

To further quantify how well these empirical distributions adhere to Benford's Law, the hypothesis testing approach is used. There are several tests through which one can verify the hypothesis but here the two tailed KS statistics is applied for the assessment of the hypothesis. The null hypothesis (H0) is tested with the level of significance as 0.05. This means that the null hypothesis is rejected in favour of the alternative hypothesis if the p-values are less than 0.05. The results obtained from the KS test are summarized in the table 3 and 4.

**Table 3: ks test results for first significant digits**

| Data mentions | AAS mentions | twitter mentions | no. of Mendeley readers | no. of Dimension citations |
|---|---|---|---|---|
| **p value** | 0.98 | 1.0 | 1.0 | 0.73 |
| **ks statistics** | 0.22 | 0.11 | 0.11 | 0.33 |

**Table 4: ks test results for second significant digits**

| Data mentions | AAS mentions | twitter mentions | no. of mendeley readers | no. of dimension citations |
|---|---|---|---|---|
| **p value** | 0.4 | 0.4 | 0.3 | 0.5 |
| **ks statistics** | 0.417 | 0.417 | 0.78 | 0.167 |

From tables 3 and 4, it is observed that all the p-values are higher than the level of significance i.e., 0.05. This suggests that alternative hypothesis may be accepted, i.e., the empirical altmetric mentions provide a good fit for Benford's Law for both, the first and the second leading digits. However, this fit is more accurate for first digits rather than second since p-values are closer to 1 in case of first digit distribution. Thus, the observations, further confirmed by statistical tests, show that the altmetric mentions follow Benford's Law.

In order to further evaluate the strength of this statistical claim, the parameters of both the data distributions i.e., Benford distribution (also referred as expected distribution) and empirical data distribution are compared. The absolute difference between mean and variance values for the different altmetric fields in the two distributions is computed and shown in table 5 (first significant digit) and table 6 (second significant digit).

It is observed that the absolute difference between the observed mean (for altmetric mentions in the empirical distribution) and the expected mean (i.e., Benford distribution) is very small. For the second digit, the absolute difference between mean values is very small and lies between 0.24 to 0.58. However, this difference is almost negligible for the first digit distribution since all vales are less than 0.16, except for the no. of Dimension citations. This indicates that the means of both the data distributions almost coincide for most of the altmetric mentions, which further confirms that the distribution observed from the empirical data is similar to the expected distribution provided by the Benford's Law. This indicates that the altmetric data considered follows Benford's Law.

**Table 5: Parameter assessment for first significant digit distribution**

| | ***Statistic/parameter*** | ***Mean*** | ***Absolute Difference*** | ***Variance*** | ***Absolute difference*** |
|---|---|---|---|---|---|
| **Expected data distribution** | | 3.43 | - | 6.04 | - |
| **Empirical data distribution** | AAS | 3.51 | 0.08 | 5.13 | 0.91 |
| | Twitter mentions | 3.59 | 0.16 | 5.55 | 0.54 |
| | No. of Mendeley readers | 3.28 | 0.15 | 5.97 | 0.07 |
| | No. of Dimension citations | 3.02 | 0.41 | 4.95 | 1.09 |

**Table 6: Parameter assessment for second significant digit distribution**

| | ***Statistic/parameter*** | ***Mean*** | ***Absolute Difference*** | ***Variance*** | ***Absolute difference*** |
|---|---|---|---|---|---|
| **Expected data distribution** | | 4.17 | - | 8.21 | - |
| **Observed** | AAS | 3.59 | 0.58 | 8.07 | 0.14 |

| | | | | |
|---|---|---|---|---|
| Twitter mentions | 3.75 | 0.42 | 8.06 | 0.15 |
| No. of Mendeley readers | 3.93 | 0.24 | 8.09 | 0.12 |
| No. of Dimension citations | 3.45 | 0.72 | 7.75 | 0.46 |

**Conclusion**

This study explored the application of Benford's Law for assessment of quality of altmetric data. For the evaluation of altmetric data quality, the leading digits of altmetric indicators such as 'Altmetric Attention Score', 'Twitter mentions', and the 'Number of Mendeley readers' are evaluated for fit with Benford's distribution. To the best of our knowledge there are no previous studies on exploring the utility of Benford's Law for the assessment of quality of altmetric data, and there lies the novelty of this work. The results show that the distribution of first digit of all the four mentions agree with Benford distribution (figure 1-4). The distribution is extended and found to be fit for the second digit as well (figure 5-8). To further confirm these empirical data fits, KS test statistics is used. A tabular summary of observed results shown in table 3 and 4 confirm an agreement with Benford's Law with significance level as 0.05. The conformity is stronger for the first leading digit as compared to the second digit. The absolute difference between means of observed and expected distributions are also computed and the values indicate that the two distributions are quite close with almost coinciding means. Thus, the study confirms that the given altmetric mention data, along with no. of Dimensions citations, follow Benford's Law.

Benford's Law has been used worldwide as a first step in the detection of non-conformance or data fraud of numerical data. It has helped in examining data quality and check data abnormalities to see if data is being manipulated. A recent study also showed that Benford's Law can be used effectively to perform integrity checks on large databases (Morales, Porporato & Epelbaum, 2022). In addition, it is used to evaluate the quality of the information (Kaiser, 2019). The present study utilized Benford's Law with the same motive, i.e., to assess the quality of altmetric data. Though in this case data fraud in altmetric mentions is not anticipated but altmetric data is yet known to be prone for manipulations by various means. Thus, a test of conformity of altmetric data with Benford's Law may provide an indication of manipulation or gaming of altmetrics. The violation to the law could be a sign that points towards need for further investigation into the data quality.

In the present study, conformity of altmetric mentions as well as citations is seen with Benford's Law, which may be a significant evidence towards quality of altmetric data. However, one must keep in mind that this study used a significantly large sized data sample (altmetric data for whole of the year 2021). It is indeed a fact that altmetric data are prone to manipulations and artificial inflation of scores by various means is seen in many cases. However, the large sample size used in this case and the observed fit for Benford distribution suggest that manipulations may not be as severe in the altmetric data sample considered. Thus, few inferences could be made from this. One is that use of a large sized sample of altmetric data is likely to reduce the effect/ impact of manipulations. Second, fit to Benford's Law can be used as a test for authenticity of a given sample of altmetric data before using it for other purposes. Another aspect is whether this approach of using fit to Benford's Law would also be able to detect manipulation of altmetric data by gaming through bots. Though, we are not completely sure, but it is expected that such manipulations will result in a situation where the empirical data may not fit Benford's Law and therefore, this approach may be able to detect such artificial inflation

of counts through gaming by using bots. However, more research and evidence are needed to establish this.

This study demonstrates the applicability of Benford's Law on four indicators namely 'Altmetric Attention Score', 'Twitter mentions', 'Number of Mendeley readers' and also the 'Number of Dimensions citations'. Distribution for other altmetric mentions could not be tested for adherence to Benford's Law due to insufficient amount of data. This study could therefore be further extended with a larger data sample which has sufficient data points for other types of altmetric mentions as well. Thus, these findings are an excellent starting point for future in-depth research on application of Benford's Law on altmetric data that may also focus on finding explanations and reasoning for the observed patterns.

## Declarations

**Funding and/or Conflicts of interests/Competing interests:** The authors declare that the manuscript complies with the ethical standards and there is no conflict of interests whatsoever. Further, this work is supported by the following extramural research grants: (a) HPE Aruba Centre for Research in Information Systems at BHU (Project Code: M-22-69 of BHU), and (b) SERB MATRICS project "Quantitative Analysis of Bibliometric & Altmetric Data of Scientific Articles" (MTR/2020/000625).

## References

Adie, E., & Roe, W. (2013). Altmetric: enriching scholarly content with article-level discussion and metrics. Learned Publishing, 26(1), 11-17.

Alves, A. D., Yanasse, H. H., & Soma, N. Y. (2014). Benford's law and articles of scientific journals: Comparison of JCR® and Scopus data. *Scientometrics*, *98*(1), 173-184.

Alves, A. D., Yanasse, H. H., & Soma, N. Y. (2016). An analysis of bibliometric indicators to JCR according to Benford's law. *Scientometrics*, *107*(3), 1489-1499.

Ausloos, M., Castellano, R., & Cerqueti, R. (2016). Regularities and discrepancies of credit default swaps: a data science approach through Benford's law. Chaos, Solitons & Fractals, 90, 8-17.

Banshal, S. K., Basu, A., Singh, V. K., Gupta, S., & Muhuri, P. K. (2021). Do'altmetric mentions' follow Power Laws? Evidence from social media mention data in Altmetric. com. 18th International Conference on Scientometrics and Informetrics (ISSI), 81–93.

Banshal, S. K., Gupta, S., Lathabai, H. H., & Singh, V. K. (2022). Power Laws in altmetrics: An empirical analysis. Journal of Informetrics, 16(3), 101309. https://doi.org/10.1016/j.joi.2022.101309

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. Online Information Review, 45(3), 517–536. https://doi.org/10.1108/OIR-11-2019-0364

Benford, F. (1938). The law of anomalous numbers. *Proceedings of the American philosophical society*, 551-572.

Bornmann, L. (2014a). Validity of altmetrics data for measuring societal impact: A study using data from Altmetric and F1000Prime. Journal of informetrics, 8(4), 935-950. https://doi.org/10.1016/j.joi.2014.09.007

Bornmann, L. (2014b). Do altmetrics point to the broader impact of research? An overview of benefits and disadvantages of altmetrics. Journal of informetrics, 8(4), 895-903.

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

Campanario, J. M., & Coslado, M. A. (2011). Benford's law and citations, articles and impact factors of scientific journals. *Scientometrics*, *88*(2), 421-432.

Cerioli, A., Barabesi, L., Cerasa, A., Menegatti, M., & Perrotta, D. (2019). Newcomb–Benford law and the detection of frauds in international trade. Proceedings of the National Academy of Sciences, 116(1), 106-115.

Cerqueti, R., Maggi, M., & Riccioni, J. (2022). Statistical methods for decision support systems in finance: how Benford's law predicts financial risk. Annals of Operations Research, 1-25.
Cheung, M. K. (2013). Altmetrics: Too soon for use in assessment. Nature, 494(7436), 176-176.
Costas, R., Zahedi, Z., & Wouters, P. (2015). 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.
Crocetti, E., & Randi, G. (2016). Using the Benford's law as a first step to assess the quality of the cancer registry data. Frontiers in public health, 4, 225.
Durtschi, C., Hillison, W., & Pacini, C. (2004). The effective use of Benford's law to assist in detecting fraud in accounting data. Journal of forensic accounting, 5(1), 17-34.
Finch, T., O'Hanlon, N., & Dudley, S. P. (2017). Tweeting birds: Online mentions predict future citations in ornithology. Royal Society Open Science, 4(11). https://doi.org/10.1098/rsos.171371
Garcovich, D., & Adobes Martin, M. (2020). Measuring the social impact of research in Paediatric Dentistry: An Altmetric study. International Journal of Paediatric Dentistry, 30(1), 66-74.
Golbeck, J. (2015). Benford's law applies to online social networks. PloS one, 10(8), e0135169.
Gonzalez-Garcia, M. J., & Pastor, M. G. C. (2009). Benford's law and macroeconomic data quality. International Monetary Fund.
Haustein, S. (2014). 17 Readership Metrics. Beyond bibliometrics: Harnessing multidimensional indicators of scholarly impact, 327.
Haustein, S. (2016). Grand challenges in altmetrics: heterogeneity, data quality and dependencies. Scientometrics, 108(1), 413–423. https://doi.org/10.1007/s11192-016-1910-9
Haustein, S., Bowman, T. D., Holmberg, K., Tsou, A., & Sugimoto, C. R. (2016). Tweets as Impact Indicators: Examining the Implications of Automated "bot" Accounts on Twitter. Journal of the Association for Information Science and Technology, 67(1), 232–238. https://doi.org/10.1002/asi
Haustein, S., Bowman, T. D., Holmberg, K., Tsou, A., Sugimoto, C. R., & Larivière, V. (2016). Tweets as impact indicators: Examining the implications of automated "bot" accounts on T witter. Journal of the Association for Information Science and Technology, 67(1), 232-238.
Haustein, S., Peters, I., Bar-ilan, J., Priem, J., Shema, H., Jens, T., & Terliesner, J. (2014a). Coverage and adoption of altmetrics sources in the bibliometric community. Scientometrics, 101(2), 1145–1163. https://doi.org/10.1007/s11192-013-1221-3
Haustein, S., Peters, I., Sugimoto, C. R., Thelwall, M., & Larivière, V. (2014b). 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.
Herrmannova, D., Stahl, C. G., & Patton, R. M. (2018). Do citations and readership identify seminal Publications? Scientometrics, 115(1), 239–262. https://doi.org/10.1007/s11192-018-2669-y
Hill, T. P. (1995a). Base-invariance implies Benford's law. Proceedings of the American Mathematical Society, 123(3), 887-895.
Hill, T. P. (1995b). The significant-digit phenomenon. The American Mathematical Monthly, 102(4), 322-327.
Horton, J., Kumar, D. K., & Wood, A. (2020). Detecting academic fraud using Benford law: The case of Professor James Hunton. Research Policy, 49(8), 104084.
Huang, Y., Niu, Z., & Yang, C. (2020). Testing firm-level data quality in China against Benford's Law. Economics Letters, 192, 109182.
Idrovo, A. J., & Manrique-Hernández, E. F. (2020). <? covid19?> Data Quality of Chinese Surveillance of COVID-19: Objective Analysis Based on WHO's Situation Reports. Asia Pacific Journal of Public Health, 32(4), 165-167.
Iorliam, A., Ho, A. T., Poh, N., & Shi, Y. Q. (2014, March). Do biometric images follow Benford's law?. In 2nd International Workshop on Biometrics and Forensics (pp. 1-6). IEEE.
Kaiser, M. (2019). Benford' Law As An Indicator Of Survey Reliability—Can We Trust Our Data?. Journal of Economic Surveys, 33(5), 1602-1618.
Karmakar, M., Banshal, S. K., & Singh, V. K. (2020). Does presence of social media plugins in a journal website results in higher social media attention of its research publications? Scientometrics, Unpublished draft. https://doi.org/https://doi.org/10.1007/s11192-020-03574-7

Kössler, W., Lenz, H. J., & Wang, X. D. (2019, August). Is the Benford Law Useful for Data Quality Assessment? In *International Workshop on Intelligent Statistical Quality Control* (pp. 391-406). Springer, Cham.
Lee, K. B., Han, S., & Jeong, Y. (2020). COVID-19, flattening the curve, and Benford's law. *Physica A: Statistical Mechanics and its Applications*, *559*, 125090.
Lin, J. (2012, June). A case study in anti-gaming mechanisms for altmetrics: PLOS ALMs and DataTrust. In paper, altmetrics12 ACM Web Science Conference, Evanston, IL.
Marcus, A., & Oransky, I. (2011). The paper is not sacred. Nature, 480(7378), 449-450.
Mebane, W. R. (2011). Comment on "Benford's Law and the detection of election fraud". Political Analysis, 19(3), 269-272.
Mir, T. A. (2016). Citations to articles citing Benford's law: A Benford analysis. *arXiv preprint arXiv:1602.01205*.
Mohammadi, E., & Thelwall, M. (2014). M endeley readership altmetrics for the social sciences and humanities: Research evaluation and knowledge flows. Journal of the Association for Information Science and Technology, 65(8), 1627-1638.
Mohammadi, E., & Thelwall, M. (2019). Readership Data and Research Impact. In Springer Handbook of Science and Technology Indicators (pp. 761–779). Springer.
Morales, H. R., Porporato, M., & Epelbaum, N. (2022). Benford's law for integrity tests of high-volume databases: a case study of internal audit in a state-owned enterprise. Journal of Economics, Finance and Administrative Science, 27(53), 154-174.
Natashekara, K. (2022). COVID-19 cases in India and Kerala: a Benford's law analysis. Journal of Public Health, 44(2), e287-e288.
Newcomb, S. (1881). Note on the frequency of use of the different digits in natural numbers. *American Journal of mathematics*, *4*(1), 39-40.
Nigrini, M. J. (2012). Benford's Law: Applications for forensic accounting, auditing, and fraud detection (Vol. 586). John Wiley & Sons.
Ortega, J. L. (2016). To be or not to be on Twitter, and its relationship with the tweeting and citation of research papers. Scientometrics, 109(2), 1353–1364. https://doi.org/10.1007/s11192-016-2113-0
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
Pericchi, L., & Torres, D. (2011). Quick anomaly detection by the Newcomb—Benford Law, with applications to electoral processes data from the USA, Puerto Rico and Venezuela. Statistical science, 502-516.
Peters, I., Kraker, P., Lex, E., Gumpenberger, C., & Gorraiz, J. (2016). Research data explored: an extended analysis of citations. Scientometrics, 107(2), 723–744. https://doi.org/10.1007/s11192-016-1887-4
Pietronero, L., Tosatti, E., Tosatti, V., & Vespignani, A. (2001). Explaining the uneven distribution of numbers in nature: the laws of Benford and Zipf. *Physica A: Statistical Mechanics and its Applications*, *293*(1-2), 297-304.
Priem, J., Piwowar, H. A., & Hemminger, B. M. (2012). Altmetrics in the wild: Using social media to explore scholarly impact. arXiv preprint arXiv:1203.4745.
Priem, J., Taraborelli, D., Groth, P., & Neylon, C. (2011). Altmetrics: A manifesto. http://altmetrics.org/manifesto/
Riccioni, J., & Cerqueti, R. (2018). Regular paths in financial markets: Investigating the Benford's law. Chaos, Solitons & Fractals, 107, 186-194.
Roemer, R. C., & Borchardt, R. (2015). Issues, controversies, and opportunities for altmetrics. Library Technology Reports, 51(5), 20-30.
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
Shema, H., Bar-Ilan, J., & Thelwall, M. (2012). Research blogs and the discussion of scholarly information. PloS one, 7(5), e35869.

Silva, L., & Figueiredo Filho, D. (2021). Using Benford's law to assess the quality of COVID-19 register data in Brazil. *Journal of public health*, *43*(1), 107-110.
Snijder, R. (2016). Revisiting an open access monograph experiment: measuring citations and tweets 5 years later. Scientometrics, 109(3), 1855–1875. https://doi.org/10.1007/s11192-016-2160-6
Sotudeh, H., Mazarei, Z., & Mirzabeigi, M. (2015). CiteULike bookmarks are correlated to citations at journal and author levels in library and information science. Scientometrics, 105(3), 2237–2248. https://doi.org/10.1007/s11192-015-1745-9
Strielkowski, W., & Chigisheva, O. (2018). Research functionality and academic publishing: Gaming with altmetrics in the digital age. Economics & Sociology, 11(4), 306.
Tahamtan, I., & Bornmann, L. (2020). Altmetrics and societal impact measurements: Match or mismatch? A literature review. El profesional de la información (EPI), 29(1).
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. (2021). Measuring societal impacts of research with altmetrics? Common problems and mistakes. Journal of Economic Surveys, 35(5), 1302-1314.
Thelwall, M., & Nevill, T. (2018). Could scientists use Altmetric. com scores to predict longer term citation counts? Journal of Informetrics, 12(1), 237–248.
Thelwall, M., Haustein, S., Larivière, V., & Sugimoto, C. R. (2013). Do altmetrics work? Twitter and ten other social web services. PloS One, 8(5), e64841.
Thelwall, M., Kousha, K., Dinsmore, A., & Dolby, K. (2015). Alternative metric indicators for funding scheme evaluations. Aslib Journal of Information Management, 68(1), 2–18.
Tošić, A., & Vičič, J. (2021). Use of Benford's law on academic publishing networks. *Journal of Informetrics*, *15*(3), 101163.
Zahedi, Z., Costas, R., & Wouters, P. (2014). How well developed are altmetrics? A cross-disciplinary analysis of the presence of 'alternative metrics' in scientific publications. Scientometrics, 101(2), 1491–1513. https://doi.org/10.1007/s11192-014-1264-0