BENFORD'S LAW AND ITS APPLICATION TO MODERN INFORMATION SECURITY

Mahyar A. Amouzegar, Department of Economics and Finance, University of New Orleans, New Orleans, LA 70148, 504-280-6595, mahyar@uno.edu

Khosrow Moshirvaziri, Department of Information Systems, California State University, Long Beach, Long Beach, CA 90840, 562-985-7965, moshir@csulb.edu

Don Snyder, The RAND Corporation, Santa Monica, CA 904, 310-451-6913, snyder@rand.org

ABSTRACT

Modern society's dependence on accuracy of government supported, publically available or proprietary corporate databases is increasing rapidly. This has become particularly true since the advent of computer technology and popularity of the Internet to support social and economic related activities. Naturally, there has been a parallel effort by both state actors and criminal elements (benign or otherwise) to exploit public and private networks for financial, criminal or political gain. The common defense is through protection of layers of cyber network and other means to create resilient and reliable operations. However, no network is ever fully safe from intrusion and concerns for data integrity has become a paramount issue, in particular in financial sector. This paper aims at exposing and extending upon the so-called Benford's law to provide a new look at how we can recognize intentional data corruption.

INTRODUCTION

Falsified numbers in tax returns, invoice payment records, expense account claims, and many other settings often display patterns that are not present in legitimate records. In fact, there is a certain pattern in the way a large group (list) of numbers behave that may be somewhat counter intuitive. One would expect that the ten digits occur with equal frequency. In fact, why would one digit be favored over another? Yet, it has been shown in many situations (both naturally- occurring or human-generated) the first digits of numbers in a dataset (e.g., legitimate records) often follow a distribution similar to the table below.

Table 1: Distribution of the occurrence of the first digit

First digit	1	2	3	4	5	6	7	8	9
Probability	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046

That is, the number 1 appears as the leading significant digit[1] about 30% of the time, while 9 appears as the most significant digit less than 5% of the time. This distribution is known as Benford's law, also called the "First Digit Law." It is interesting to note that this result applies to a surprisingly large number of data sets, including street addresses, certain stock market data, Internal Revenue Service files, electricity bills, death rates, and lengths of rivers [2].

[1] The *first significant digit* in a number is the first nonzero digit when reading from left to right. For example, the first significant digit of 218.81 is 2 and that of 0.0375 is 3. The first significant digit is always nonzero, but second and higher significant digits can be 0. The second significant digit of 0.102 is 0.

Mathematically, the probability of an occurrence of the particular leading significant digit is

$$P(D = d) = log_{10}\left(1 + \frac{1}{d}\right) \quad (1)$$

Where, D denotes the first significant digit and d is any integer in the set $\{1, 2, 3, \dots, 9\}$.

Benford's law can be generalized for the k[th] significant digit. In other words, a random continuous variable *X*, with realization *x*, is said to follow Benford's law, if the application of the first-k-significant-digits function, $D_k(x)$, yields a Benford-distributed discreet variable with density function given by (2) below, where $\mathbb{Z}_+$ denotes the set of positive real integers.

$$P(d_k) = P(D_k(X) = d_k) = \log_{10}(1 + d_k^{-1}), \quad \forall k \in \mathbb{Z}_+ \quad (2)$$

With $$D_k(x) = \left\lfloor |x| \cdot 10^{(-1 \cdot \lfloor \log_{10} \cdot |x| \rfloor + k - 1)} \right\rfloor, \quad (3)$$

Where $$d_k \in \{10^{k-1}, 10^{k-1} + 1, \dots, 10^k - 1\}$$

Datasets following this distribution are said to be *Benford*. If a dataset is Benford, then by equation (2) there is approximately a 30% chance that the first significant digit of any datum in that dataset is 1, about an 18% chance that the first significant digit is 2, and so on. When altered fraudulently, Benford datasets depart from this pattern, a fact that is used in fraud detection [3]. These predicted probabilities, especially for the first significant digit, have been shown to hold for both theoretical distributions and some real data sets [5].

When should we expect a dataset to be Benford? There is no clear answer. However, it can be proven mathematically that taking any distribution and repeatedly multiplying or dividing by random numbers or raising it to a random integral power numerous times converges to a Benford distribution. In addition, a distribution that is Benford remains Benford under multiplication, division, and the raising to a power (scale and base invariance) [2, 5]. In fact, a distribution that is scale invariant is always Benford. This invariance means that if a distribution is Benford when expressed in one set of units, it is Benford when expressed in any units.

Researchers have laid out some interesting properties that tend to lead to a dataset being Benford, though satisfying these conditions will not guarantee it:

- Numbers coming from mathematical combinations of other numbers;
- Transaction-level data (as opposed to aggregated data);
- Large datasets that span multiple orders of magnitude in values;
- Data for which the mean is greater than the median (hence, positively skewed); and
- Scale invariance.

Datasets that are less likely to be Benford are those composed of assigned or sequential numbers (e.g., telephone numbers), data that are influenced by psychological factors (e.g., prices set at $29.99), data with a large number of firm-specific numbers (accounts set up to record refunds of a fixed price), or data with a built-in minimum or maximum. Data that are presented as percentages rather than raw values are also less likely to be Benford. Data that have a fixed number of digits for each entry are often not Benford.

Although the mathematical proof is beyond the need of this article, intuitively the law is not difficult to understand. Consider a stock portfolio with a current market value of $1,000.000. For the first significant digit to turn from "1" to "2", it will have to double in size. That is, the portfolio value needs to grow 100 percent. Now, for the first digit to become "3," then the portfolio only needs to grow by 50 percent. And of course, for the first digit to become "4", the portfolio needs to only grow by 33%. Therefore, it is clear to see that in many distributions of financial data, which measure the size of anything from a purchase order to stock market returns, the first digit one is much further from two than eight is from nine. Thus, the observed finding is that for these distributions, smaller values of the first significant digits are much more likely than larger values [3].

Table 2 presents the first, second, third and fourth digit portions of Benford's law, generated using the formula above (note how in higher orders, the frequency of digits converge to a more uniform distribution).

Table 2: First, Second, Third, and Fourth Digit Proportions of Benford's Law

	Position Frequency			
Digit	**1st digit**	**2nd digit**	**3rd digit**	**4th digit**
0		0.119679	0.101784	0.100180
1	0.301030	0.113890	0.101376	0.100140
2	0.176091	0.108821	0.100972	0.100100
3	0.124939	0.104330	0.100573	0.100006
4	0.096910	0.100308	0.100178	0.100200
5	0.079181	0.096677	0.099788	0.099980
6	0.066947	0.093375	0.099401	0.099940
7	0.057992	0.090352	0.099019	0.099900
8	0.051153	0.087570	0.098641	0.099860
9	0.045757	0.084997	0.098267	0.099820

ORIGIN OF BENFORD'S LAW

In 1881, Simon Newcomb, an astronomer and mathematician, discovered the statistical principle that has become known as Benford's law. He observed that the earlier pages of logarithm books, used at that time to carry out logarithmic calculations, were considerably more worn in the beginning pages which dealt with low digits and progressively less worn on the pages dealing with higher digits. This led him to formulate the principle that, in any list of numbers taken from an arbitrary set of data, more numbers will tend to begin with "1" than with any other digit. The obvious conclusion was that more numbers exist which begin with the numeral one than with larger numbers.

Newcomb calculated that the probability that a number has any particular non-zero first digit d is: $\text{Probability}(D = d) = \log_{10}\left(1 + \frac{1}{d}\right)$. Newcomb provided no theoretical explanation for the phenomena he described and his article went virtually unnoticed. Then, almost 50 years later, Frank Benford, a physicist, also noticed that the first few pages of his logarithm books were more worn than the last few. He came to the same conclusion Newcomb had arrived at years prior; that people more often looked up numbers that began with low digits rather than high ones. He also

posited that there were more numbers that began with the lower digits. Benford collected more than 20,000 observations from such diverse data sets as areas of rivers, atomic weights of elements, and numbers appearing in Reader's Digest articles [1]. Benford found that numbers consistently fell into a pattern with low digits occurring more frequently in the first position than larger digits. The mathematical tenet defining the frequency of digits became known as Benford's law.

However, it wasn't until 1995 that T. P. Hill, provided a proof for Benford's law as well as demonstrating how it applied to stock market data, census statistics, and certain accounting data. He noted that Benford's distribution, like the normal distribution, is an empirically observable phenomenon. Hill's proof relies on the fact that the numbers in sets that conform to the Benford distribution are second generation distributions, that is, combinations of other distributions [5].

APPLICATION OF BENFORD'S LAW TO DETECTING ANOMALIES

Public and private sectors are highly dependent on information systems to carry out their missions and business functions. Moreover, these dependencies have made business and government supply and information process highly vulnerable to cyber-attacks. The problem is not just denial of service, or malicious firmware, which are of course of concern, but also subtle corruption of data that might impact the operation of these enterprises. The sheer size and complexity of many of the supply chain systems, for example, place demands on knowledge of the identity of parts, stock levels, part locations—and many other data—that exceed human capacity. In these systems, the absence of reliable data, might force many key functions to halt or at minimum decrease trust in the accuracy of information.

Data Error and impact - Of course, data errors in any operations are inevitable. Errors occur routinely from everyday mistakes. For the most part, these day-to-day errors do not have significant negative operational impacts as most systems and there are processes that have evolved to handle them. Furthermore, the randomness of routine errors makes it unlikely that any one error will cascade into a major operational problem. However, significant impacts are possible, as experience has shown. A skilled, determined, and knowledgeable adversary could potentially wreak far more damage by deliberately corrupting data that are unlikely to be detected as anomalous, yet targeting the attack to have a significant negative impact on operations. An adversary (whether internal or external or domestic or foreign) might choose this kind of targeted attack by corruption, over data destruction or denial of access to data in order to maintain a longer foothold in the systems or to mask attribution. Of course, regardless of whether data is corrupted by attack or random error, operational system should be sufficiently resilient and robust to data corruption to continue providing adequate support.

Election Case - An interesting use of Benford's law was after the controversial 2009 Iran's election, when Mahmoud Ahmadinejad, running against three challengers, won with an overwhelming majority, despite the pre-election expectations. The Ministry of the Interior (MOI) published a table of the numbers of votes received by each candidate for the 366 voting areas. The MOI's data vary from about 10^4 to 10^6, which suggested the possibility of the dataset being Benford. Boudewijn Roukema [6] used the available data and Benford's law to show, amongst other issues, one of the losing candidates had a significant excess of vote counts starting with the digit 7. He concluded, "[the] most consistent way to explain these results would appear to be the hypothesis of artificial interference in the official results."

Benford's law doesn't apply to every dataset but if one falls within the requirements to be Benford then it provides a great vehicle to detect deviation to what the data should be.

Pricing dataset - Secondary Item Requirements System (SIRS), and Central Secondary Item Stratifications System (CSIS), which together are known as the Requirements Management System (RMS) store data on the Requirement Data Bank (RDB) for the Air Force's maintenance and repairs. In this instance, the term "requirements" refers to the necessary supplies to complete the maintenance and repair of the weapons systems. SIRS computes spare parts requirements on an aggregate basis. Certain dataset in this system have shown interesting resemblance to an ideal Benford distribution. Numerous goodness-of-fit and other statistical tests confirm that the data is Benford and hence easily detectable if there any malicious data manipulation.

Social Networks dataset - Data collected from various social networks showed another interesting application of Benford's law [4]. Data was collected from the following social networks with number of users:

Pinterest = 40 Million users	Twitter = 78,000 users	Facebook = 18,000 users
LiveJournal = 45,000 users	Google Plus = 20,000 users	

Where numbers of friends or followers of each user had been counted and then determined how the first digits were distributed. Every dataset, expect for one (Pinterest) followed Benford's law. From the discussion above, the fact that most of these datasets are Benford shouldn't be a surprise since the dataset was generated naturally or organically grown. It became clear that Pinterest users are required to follow five or more "interests" as a part of their registration process; this creates at least five initial followers for each user affecting the entire distribution of first significant digits. It was further considered network created by user's friends, which is known as Egocentric Networks. The correlation between a user's egocentric network and Benford's law was measured and the result was that for majority of people, this correlation was greater than 0.9, which means that they conformed to Benford's law. In case of Twitter, only 170 users out of 21,000 had a correlation lower than 0.5. Further investigation showed some of the accounts were spam and most of them were a part of a Russian bots' network who behaved in similar way. The purpose of these accounts was not clear but they were certainly suspicious.

Academic project - We examined close to 3,141 set of data extracted from dozens of projects. Participants were required to generate a dataset using a certain algorithm with a unique "seed". The algorithm was designed to create a Benford dataset if everyone followed the instructions verbatim and did not use shortcuts in order to save time. The shortcuts would mimic subtle manipulation of data. Figure 1 illustrates the results for the first digit test as compared to the expected Benford's results.

Of course, there are several tools to test for the goodness-of-fitness since the idea of using Benford's law is to quickly and easily identify data that are not Benford when they are supposed to be. For our dataset, we used the Chi-square Test, which indicated the dataset was not conforming to Benford. However, another test, the so called the Mean Absolute Deviation (MAD) test, indicated marginally actable to acceptable conformity. And finally, *Z*-test told us the deltas were too wide for conformity. Untimely, the point is not whether a dataset is fully conforming to

Benford's law but rather if it gives us pause to accept the data as is and look for possible anomalies. We expected a close correlation and the goodness of fit tests have created doubts. In fact, it was easy to identify the data that was manipulated because of shortcuts or other means. In total about 5% of the data was "corrupted".

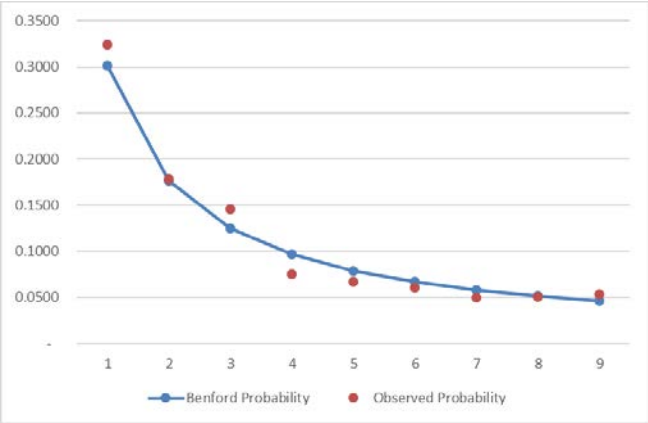

Figure 1: First Digit Test

CONCLUDING REMARKS

In recent years, we have witnessed a dramatic rise in hackers' activities and destructive attacks, from in and out of the country, on various social, financial enterprises, political, and government networks. It is quite difficult to detect fraudulent and suspicious activities as hackers equip themselves with sophisticated tools in carrying out their malicious attacks. The framework presented by implementation of Benford's law has proven to have important implications for social network forensics. Models built upon the framework introduced herein has been very promising as implemented on Twitter's and Facebook networks [4]. This makes Benford's law to be one of the available effective tools in the war against fraud and suspicious activity on social networks. The use of the technique and applications of Benford's law to social media is a new tool for analyzing user behavior, understanding when and why natural deviations may occur, and ultimately detecting when anomalies occur.

REFERENCES

[1] Benford, Frank, The law of anomalous numbers. Proceedings of the American Philosophical Society, Vol. 78, No. 4, 1938.
[2] Berger, Arno, and Theodore P. Hill, *An Introduction to Benford's Law*, Princeton, New Jersey: Princeton University Press, 2015.
[3] Durtschi, Cindy, et al., "The Effective Use of Benford's Law to Assist in Detecting Fraud in Accounting Data," Journal *of Forensic Accounting*, Vol. 5, 2004.
[4] Golbeck, Jennifer, Social and Information Networks (cs.SI); Physics and Society, DOI: 10.1371/ journal.pone.0135169
[5] Hill, Theodore, P., "A Statistical Derivation of the Significant-Digit Law," *Statistical Science*, Vol. 10, No. 4, 1995.
[6] Roukema, Boudewijn, A First-Digit Anomaly in the 2009 Iranian Presidential Election, *Journal of Applied Statistics*, Vol. 41(1), 2014.