On Skewed Distributions and Straight Lines: A Case Study on the Wiki Collaboration Network

Osnat Mokryn
School of Computer Science
Tel Aviv Yaffo College
Israel
ossi@mta.ac.il

Alexey Reznik
Microsoft
Redmond, WA
USA
alexrez@microsoft.com

ABSTRACT
In this paper, we present a hypothesis that power laws are found only in datasets sampled from a static data, in which each and every item has gained its maximal importance and is not in the process of changing it during the sampling period. We motivate our hypothesis by examining languages, and word-ranking distribution as it appears in books, and in the Bible. To demonstrate the validity of our hypothesis, we experiment with the Wikipedia edit collaboration network. We find that the dataset fits a skewed distribution. Next, we identify its dynamic part. We then show that when the modified part is removed from the obtained dataset, the remaining static part exhibits a good fit to a power law distribution.

Categories and Subject Descriptors
G.3 [**Probabilities and Statistics**]:

General Terms
Measurement.

Keywords
Skewed distributions; Power law distribution; Trends; Wikipedia; Collaboration networks; Dynamic distributions.

1. INTRODUCTION
Skewed distributions are evident in areas ranging from geophysics to finance and the Internet topology, as well as in the degree distribution of social and real networks. Some measurements of these quantities were often interpreted as power law distributions [1], [2]. However, others have shown that most found distributions are rather skewed [3]–[6]. In a seminal paper, Clauset et al [5] suggested a set of techniques for fitting skewed distributions to a power law, and found that among a variety of datasets they examined, the only good fit for a power law was the words ranking distribution.

Indeed, an exemplary example for a power law distribution is the ranking of words occurrences in books. It was found as early as 1949 by the linguistic Zipf [7]. Interestingly, a later work by Bi et al [3] made a rather unique observation. Unlike many other books, the Bible word distribution deviates from a pure Zipf distribution. One of the main differences of the Bible is that it is a collection of chapters, written over hundreds of years. This is vastly different from most books that are written over a period that usually spans a few years, and seldom reaches a few decades.

Why would the period in which a book was written impact its word ranking distribution? It is the hypothesis of the paper, that pure power law distributions are formed in real life only if they are measured over a static data set or network, rather than over a currently evolving one. Let us look back at our motivating example. Languages are changing very slowly[1]. During the period in which a book is written, words do not change their meaning; do not go out of use, and new words seldom enter the language. Hence, each book can be seen as a sample from a *static* language. The Bible, on the other hand, is a collection of chapters written over hundreds of years. Together, they capture the evolvement of languages, and hence cannot be seen as a sample from a static language. Some words that are never used in one chapter may be often used when a next chapter is written, as new words emerge with time and others become popular. Complementary, words that are used often in one chapter may no longer be of use when a later chapter was written. The Bible, hence, can be seen as a sample from a *dynamic* language. This sample thus capturing the dynamics of the language and hence its word ranking distribution yielding a skewed distribution rather than a straight power law.

We then formalize our hypothesis.

Hypothesis: Power laws are found only in datasets sampled from a static data, in which each and every item has gained its maximal importance and is not in the process of changing it during the sampling period.

To investigate our hypothesis we require a dynamic dataset with a skewed distribution. Our goal is to demonstrate that the removal of its changing or dynamic part indeed yields a power law distribution. The dynamic part of a dataset contains all the items that belong to the dataset but are in the process of changing their importance, or attractiveness. For example, a social network is an ever-changing dataset, in which new relationships are always formed for some or all of the people in the network. It is probably close to impossible to find a period of time in which a sample is possible, yet relations do not change, nor new people join and other leave. The population of cities might have been static over periods in which there were no mobility trends, but vey dynamic during periods of migrations.

Another example for a dynamic network is the Wikipedia edit collaboration network. Wikipedia is one of the first online projects

WWW'15 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2744716

[1] It is very possible that with the advance of the web, Twitter and blogs, and the global and inter-cultural interactions that arrive

created to facilitate collaborative creation of content. It was launched on January 15, 2001 and rapidly accumulated a large base of contributors that create and update content. Wikipedia can be modeled as a collaboration network of its editors. Each article is a result of collaboration of multiple authors, wherein the number of distinct authors per article has a median of 36.5 and follows a power-law [8].

The dynamics of Wikipedia edit collaborations have been heavily researched [9]–[15]. The vast majority of edits appears in what is termed an "edit-stream", in which the collaborators keep editing the article until a consensus is reached. Edit-streams appear often in bursts, and are referred to as edit-bursts. While edit-bursts do not tend to be long, some may take up to a few weeks [13].

Wikipedia has a very large set of articles. The vast majority of them do not change over large periods of times. Articles may change due to a risen interest of an editor or an exogenous event, such as a news event. Hence, when looking at *editing events,* there are periods during which most articles can be viewed as static, and only a small number of articles are undergoing edit-bursts, and thus form the dynamic part of the dataset for that period. According to our hypothesis, the removal of this dynamic part should result in a good fit of the remainder of the data to a power law distribution.

To demonstrate the validity of our hypothesis, we collected the edit collaboration event-streams of over 300,000 Wikipedia articles. We then grouped them according to their creation date, and created their binned weekly and daily edit event-streams. Indeed, the resulted distributions were skewed. We then identified as the dynamic part articles that were edited within the period preceding the date we collected the data. These articles were candidates to be in an edit burst-event. After the removal of these suspected-dynamic articles, we received a distribution that was a good fit to a power law. This result repeated itself over several groups of pages, and also over the entire obtained dataset, regardless of the articles creation date.

The structure of the paper is as follows. Section 2 surveys shortly the research on power laws and skewed distributions. Section 3 discusses Wikipedia as a collaboration network and its dynamic patterns. It then continues to model the Wikipedia event-streams. Section 4 details the data we obtained, and its characteristics. In Section 5 we present our methodology and experiment results. We discuss the Implications in Section 6.

2. RELATED WORK

Power laws and skewed distributions are a real phenomenon that exists in almost every aspect of society. Zipf [7] introduced in 1949 a word distribution in which the frequency is inversely proportional to the rank of vocabulary words. In the distribution a few of the words occur very often while the rest of the words occur seldom. In a logarithmic-logarithmic scale, the resulted plot of any quantity that follows a Zipf distribution should follow a straight line. However, in reality, many real quantities fail to follow a straight line in the logarithmic-logarithmic scale and adhere to a pure Zipf distribution [3]. For a detailed survey and definitions of the terminology of skewed distributions please refer to the survey presented in [3].

Andriani and McKelvey [16] demonstrated that power laws are an inextricable aspect of how individuals, organizations, economies, and societies work. A recent survey [17] reviews the existence of power laws in real life scenarios, and calls for an agreed upon explanation for their appearance, considering that many of these distributions indeed seem to deviate slightly from a power law. Indeed, pure power laws are seldom found in reality [5]. The skewed distributions found in real data have been heavily researched during the years. The Discrete Gaussian Exponential Distribution was suggested in [3] as a better fit for datasets such as sales data and mobile calls. Mobile calls, however, were further investigated in a subsequent work [6] and found to fit better a different skewed distribution. Another notable example is the Internet infrastructure. In 1999 it was found to show a power law distribution [1]. However, later works showed it follows rather a more skewed distribution [4], [18], [19].

3. A CASE STUDY: WIKIPEDIA

3.1 Modeling Wikipedia as a Collaboration Network

Collaborative (collaboration) network is a network that consists of nodes that represent entities (organizations, people) and edges that represent some sort of collaboration between them. A notable example is the collaborative network between scientists, where an edge represents a paper written together [20]. This mathematical collaboration network is famous for assigning each mathematician an “Erdős number” – the “collaborative distance” between the mathematician and Paul Erdős. Another example of a similar collaborative network is the network of movie actors, which are linked by co-acting in the same movie. Similarly, this network is known for assigning each movie actor a “Bacon number”, a collaborative distance between the actor and the actor Kevin Bacon.

A few types of collaborative networks are defined, characterized by the type of collaboration that forms the links in the network. The first type is a one-time collaboration network, such as co-authoring a scientific paper, or co-acting in a movie. This type of network (graph) is undirected, and can have more than one edge between nodes in case that the entities in question have multiple collaborations (it can also be represented as edges that have weights, corresponding for example to the number of collaborations between the actors). The second type is a long-term collaboration network – in which collaboration spans a long time period or is a common property that the entities share. An example for such a collaborative network is a social network in which the nodes are the people and the edges that connect them represent friendship bonds between them. This type of network creates an undirected (since friendship is usually mutual) graph with only a single edge possible between two nodes. Since the mutual state (friendship) is not limited in time, whenever it ends – the link between the nodes is removed. Another example would be a “club membership” network, which connects people that are members of the same club, even if they are not acquainted.

The third type of collaboration network is a crowd-collaboration network. In this type of network entities (people, organizations) collaborate on a project of some sort, without being directly connected. Examples of such networks vary: software developers collaborating on different open-source projects; various crowd-sourcing projects and wikis, such as Wikipedia, in which editors collaborate on different articles; crowd-funding projects in which users can fund together projects that they like; the aggregated volume of products reviews in a reviews-site, and many more.

The Wikipedia edit collaboration network is modeled as a crowd-collaboration network [10], [21]. The collaborators are modeled as the nodes, and links between them are created if they have collaborated (e.g., edited) the same article. Another approach is to model the network as a two-sided graph, in which there are two types of nodes – the collaborators (editors), and the subjects of their collaboration (articles). In this case, the links will connect the editors with the article they contributed to.

3.2 Identifying Dynamic Patterns in Wikipedia

Wikipedia's success has been attributed right from its beginning to its highly motivated community of maintainers, which drive the project forward [22].

In 2004 Andrew Lih analyzed Wikipedia as a source of journalistic information, and studied the trends in Wikipedia being used as a source in the news media [8] . He analyzed Wikipedia articles in term of total number of edits (which he termed rigor) and number of unique editors (which he termed diversity) in order to gain insights into the dynamics of article popularity. He has established the mean between the articles for both rigor and diversity, and used it as a metric of article's popularity among editors. He had shown that after being cited by a news outlet, article becomes more attractive to editors, and are more likely to measure above the mean on both axes. His findings led him to suggest that a significant amount of activity performed by the Wikipedia contributors is motivated by news and mass media events.

As edits were shown to arrive in bursts, a later study tried indeed to quantify the effect of events on edit bursts in Wikipedia [13]. They have discovered only one pattern that characterizes a stream of edit events (edit event-stream): local clustering, caused by a combination of two factors: intermediate saves performed by the same author and edits that are immediately re-edited by other authors (edit-wars). Other than that, no long-term correlations were discovered in the edit-stream. This may hint that editorial activity is not driven by the shifts in popularity of topics of general interest in society. Although this feels counterintuitive, the authors of [13] claim that the apparent randomness of edit events can be caused by superimposition of few factors: A collaborative network of editors, dedicated to improving the content of Wikipedia, and who coordinate their efforts via discussion pages and to-do lists, control the editing process; external factors (for example a strike of inspiration of one of the editors); continuous process, of few uncoordinated editors that feel responsible for one or more topics, and are constantly updating them.

Taking an article-trajectory approach (as oppose to examining edit collaborations across articles) has shown that news events cause a stream of edits, that is different at first from the common event streams, but converges to that with time [14]. A recent work [15] was able to extract event-related information using a system that its first building block is a burst detection component.

The above leads to the conclusion that identifying articles undergoing edit bursts might enable us to identify the dynamic part of the distribution. For our research we are hence looking for all the edit events of articles created at the same day. To find the dynamic part of the dataset, we then need to identify articles that are currently being edited. In [13] it was found that edit bursts may last as long as a few weeks. This will have an impact on the articles we identify as dynamic, i.e., that are part of an edit-event.

3.3 Wikipedia as a Case Study

Our hypothesis assumes that power laws arise from static datasets, i.e., datasets in which each item has gained a maximal popularity or importance, and remains at this level. Items do not lose importance, nor do they gain importance. In the motivating word distribution example, we referred to the dynamics of languages. Our claim was that as languages change very slowly, over hundreds of years, each book was a sample of a static language. We refer to a language as static when all of the words in the language gain a certain popularity or importance, and keep the same level of importance during the period in which the book is written. New words do not appear, and existing words do not lose or gain importance, nor disappear.

The Wikipedia edit collaboration network is not a static network. There are always some articles that are undergoing edits, and changes. However, we are not interested in the entire network. For our purposes, only the edit history of each article is important. Hence, we are interested in the edits event-stream of each article, from its creation day. Then, we need to define a period of time, for which we would like to sample a static dataset. I.e., in our case, a dataset in which every item (article) has gained its full importance. In the case of edits, this would be the maximal number of edits up until this period. Dynamic items are then articles that are being edited during current period, and hence their importance (i.e., number of edits) is currently changing.

4. DATA AND MODELING

The MediaWiki project makes all its data and history records available in the form of database-dumps: a set of records in the form of bz2[2] compressed xml files[3]. We used the data-set from 02/07/2012. Wikipedia dumps consist of a few file-sets. For the needs of this work we have used the file-sets containing all the versions of every article in Wikipedia accompanied with the editors metadata. The first 306,740 in the Wikipedia dump, which is sorted by page id, were selected. Wikipedia contains quite a lot of articles that were created automatically by scripts from external data sources. Such a script has created most (or even all) of the 8752 articles at Feb 25th 2002, and at Oct.18th 2002.

For each article, then, we obtained all of its edit events: we first obtain for each article its title and id, and then the series of edit events, including: the time of the edit and the username of the editor (or the IP address, if the user wasn't logged in). For each article we then create its event-stream. The edit events per article were then aggregated daily, or weekly.

Figure 1 shows the weekly binned edit distribution of all collected articles. For each article, all the edits done during each week were aggregated and binned. The distribution is a skewed distribution that deviates from a pure power law. For example, of all the 306,740 articles we've collected from Wikipedia, (y=100) of them had any edits during their (x=100) week of existence. This means that from all articles that were 500 weeks old, only 10 articles are edited each week.

To avoid time-dependent fluctuations, we further divided the data to groups of articles created at the same date, and examined their edit-stream from their creation date forward. Figure 2 shows the

[2] bz2 implements the Burrows–Wheeler algorithm.

[3]These files can be freely downloaded from http:/dumps.MediaWiki.org/enwiki/

weekly-binned edit distribution of the 8752 articles created at Feb.25th 2002. The distribution is skewed.

Over all we collected and generated the edit event stream information for 306740 articles. For each articles the edit information spanned the period from its creation and right up to the time the data was collected.

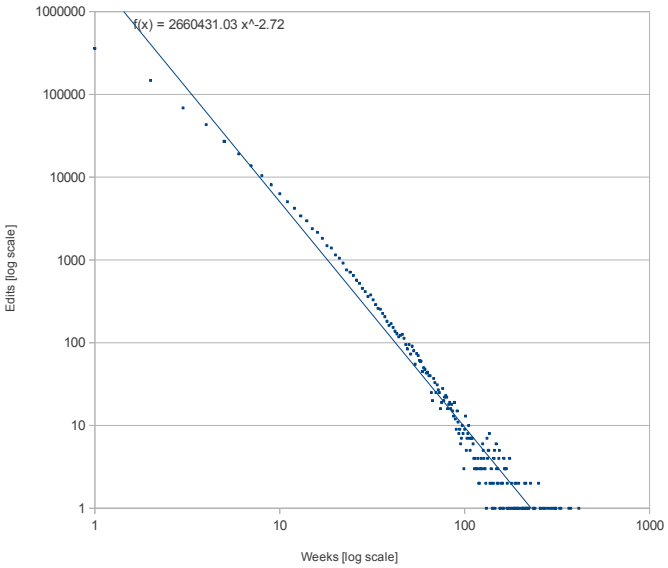

Figure 1 Weekly binned edits distribution of all 306,740 collected articles

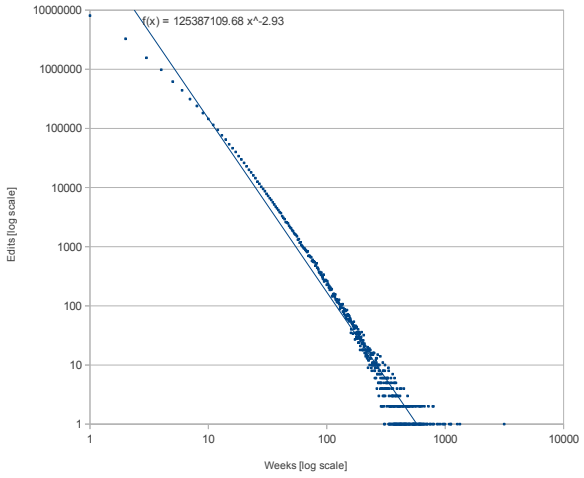

Figure 2 Weekly binned edits distribution of all 8752 articles created at Feb. 25th 2002

5. METHODOLOGY AND EXPERIMENT RESULTS

We present here our findings over the Wikipedia collaboration edits event-stream datasets.

First, we identify the dynamic part of the dataset. Recall that our definition of a static dataset requires that during a period of time, the dataset is static and each item in the dataset has obtained its maximal importance, which remains unchanged for the duration of the sample. In our case this requires a period of time, in which no article in the dataset undergoes any edits (changes). This requires also that articles are not undergoing any edit-burst. However, an article that is undergoing an edit-burst may seem unchanged during a short period of time, as sometimes there may be a periods of hours or days in which it is not changed during the burst [13]. The article has not gained its maximal number of edits, as there are edits that are part of the current edit-burst, which have not occurred yet. Reasons may vary, from an editor being delayed to an offline discussion on the needed edits between collaborating editors.

We then look to identify a period of time that is longer than the longest edit-burst.

Setting the day in which the Wikipedia data was collected as the latest possible updating date, we look to define the longest period of time prior to it in which an edit-burst could occur. In other words, we are looking for the latest date, in which an edit-burst could have started, for any article.

Based on the findings presented in [13], we set this period of time to four weeks.

Thus, we identified as the dynamic part of the datasets articles that were edited (even once) during the last four weeks before obtaining the data. In this manner, we define the sampling period to four weeks, and the static part of the dataset to all the articles that have not been edited at all during that period of time, and hence have maintained their maximal importance, i.e., number of edits, static during this period.

Figure 3 shows the edits distributions of the static part of the dataset of the articles created at feb.25th 2002. Out of the original 8752 gathered articles, 2,728 were found as dynamic during the sampling period, e.g., they had at least one edit in the last four weeks of the period. After their removal the remaining 6024 indeed exhibit a good fit to a power law.

Figures 4 and 5 show the edits distributions of the dataset of pages created at Oct.18th 2002, and only its static part, corresponding. It is interesting to note that only a small fraction of the pages were identified as dynamic during the period examined out of the dataset (16%). Indeed, the whole dataset does not deviate much from a power law to start with, and fits well to a power law after the removal of the dynamic part.

The above may lead to the theory, that the skewer the distribution the bigger is the dynamic part in it.

We then examine the entire dataset of articles, regardless of the date an article was created. Figure 1 depicts the overall distribution.

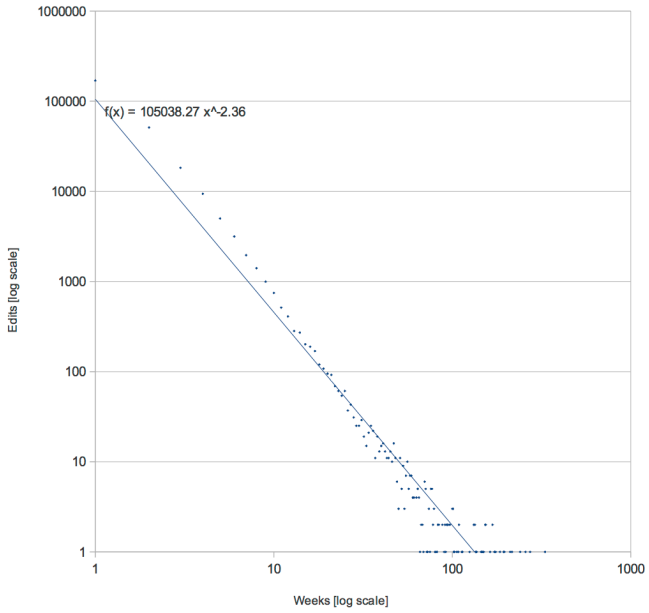

Figure 3 Weekly binned edits distribution of 6024 static articles out of original 8752 created at Feb. 25th 2002.

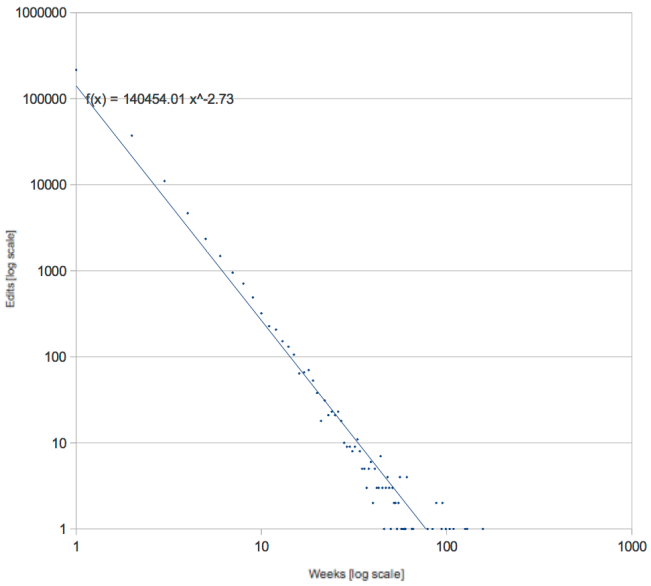

Figure 5 Weekly binned edits distributions of 7349 static articles out of original 8752 created at Oct. 18th 2002

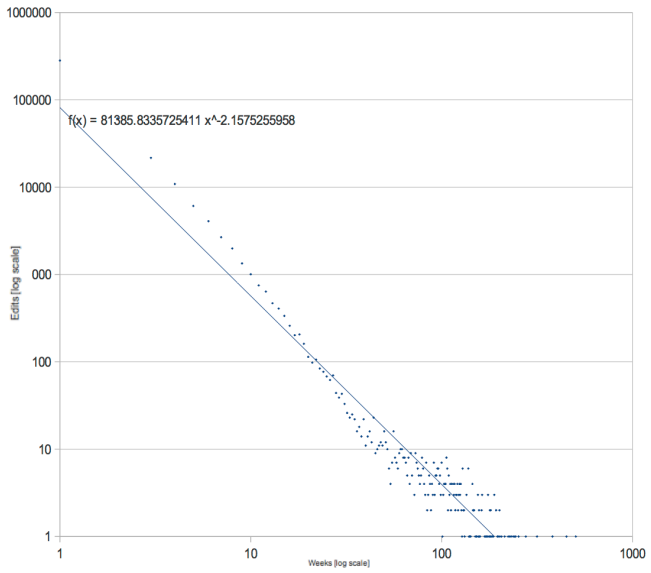

Figure 4 Weekly binned edits distributions of 8662 pages created at Oct. 18th 2002

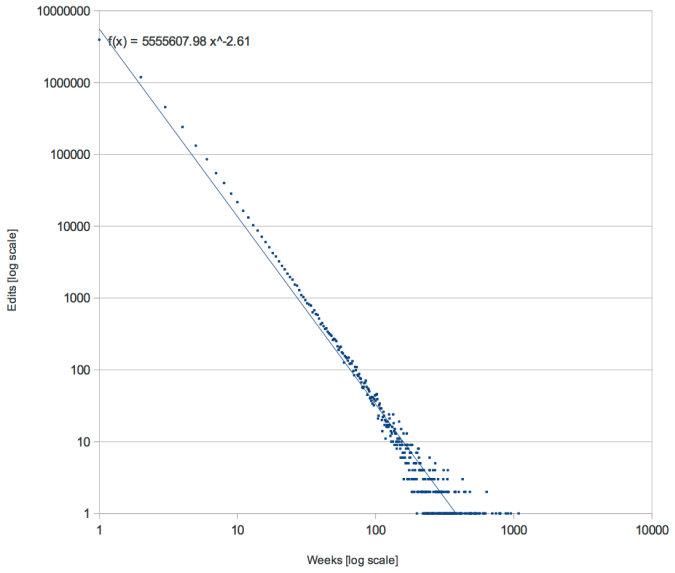

Figure 6 Weekly edit distribution of the entire dataset without the articles that were edit in the last four weeks. The static part contains 175,708 articles

Figure 6 depicts the binned edits weekly distribution of all pages not undergoing any edit during the last four weeks, hence static. Indeed, we see a good fit to a power law. It is significantly better than the original one of the whole dataset, depicted in Figure 1. The dynamic part in this case accounts for almost 43% of the data. Again, we see that the bigger the dynamic part of the data, the more skewed is the distribution.

6. CONCLUSIONS AND DISCUSSION

We have demonstrated that when sampling a non-changing (static) dataset, in which all items have acquired a maximal unchanged importance for duration of the sampling period, the resulted distribution gives a good fit to a power law. We further showed that when the data contains items that are changing their importance (in our case, the number of edits) during the sampling period, the resulted distribution is skewed. We further

demonstrated using the Wikipedia edits collaboration datasets that the bigger the dynamic part of the dataset, the more skewed is the distribution compared to a power law on the logarithmic scale.

An interesting question then arises. Let us take for example the mobile call dataset [6]. The dataset contains all calls made between all the people involved. Shouldn't it be a power law, then, if it contains complete information? We would argue that the answer is no. While the dataset contains all phone calls made between the people, the real network we are measuring and considering in this case is the result of the relations between the people. As the relations evolve during the sampling period (the period in which the data was collected), some relations have reached their peak, while others are transitioning. Some relations transition towards a stronger bond that might entail a closer relationships and a higher rates of calls, while other relations weaken, and are in the process of having a lower number of calls over time. We claim that the dataset is not complete, as it does not capture the changing relationships dynamics, that affect the rate by which people communicate.

7. ACKNOWLEDGMENTS

The authors would like to thank Jean Bolot, Yuval Shavitt, and Christiophe Diot, for interesting discussions and their important remarks on this work.

8. REFERENCES

[1] M. Faloutsos, P. Faloutsos, and C. Faloutsos, “On power-law relationships of the internet topology,” in *ACM SIGCOMM Computer Communication Review*, 1999, vol. 29, no. 4, pp. 251–262.

[2] L. A. Adamic and B. A. Huberman, “Power-law distribution of the world wide web,” *Science (80-.).*, vol. 287, no. 5461, p. 2115, 2000.

[3] Z. Bi, C. Faloutsos, and F. Korn, “The DGX distribution for mining massive, skewed data,” in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, 2001, pp. 17–26.

[4] P. Mahadevan, D. Krioukov, M. Fomenkov, X. Dimitropoulos, A. Vahdat, and others, “The Internet AS-level topology: three data sources and one definitive metric,” *ACM SIGCOMM Comput. Commun. Rev.*, vol. 36, no. 1, pp. 17–26, 2006.

[5] A. Clauset, C. R. Shalizi, and M. E. J. Newman, “Power-law distributions in empirical data,” *SIAM Rev.*, vol. 51, no. 4, pp. 661–703, 2009.

[6] M. Seshadri, S. Machiraju, A. Sridharan, J. Bolot, C. Faloutsos, and J. Leskove, “Mobile call graphs: beyond power-law and lognormal distributions,” in *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2008, pp. 596–604.

[7] G. K. Zipf, “Human behavior and the principle of least effort.,” *Addison Wesley, Cambridge, Massachusetts*, 1949.

[8] A. Lih, “Wikipedia as participatory journalism: Reliable sources? metrics for evaluating collaborative media as a news resource,” *Nature*, 2004.

[9] A. Capocci, V. D. P. Servedio, F. Colaiori, L. S. Buriol, D. Donato, S. Leonardi, and G. Caldarelli, “Preferential attachment in the growth of social networks: The internet encyclopedia Wikipedia,” *Phys. Rev. E*, vol. 74, no. 3, p. 36116, Sep. 2006.

[10] U. Brandes, P. Kenis, J. Lerner, and D. van Raaij, “Network analysis of collaboration structure in Wikipedia,” in *Proceedings of the 18th international conference on World wide web*, 2009, pp. 731–740.

[11] G. Wu, M. Harrigan, and P. Cunningham, “Characterizing wikipedia pages using edit network motif profiles,” in *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, 2011, pp. 45–52.

[12] J. Liu and S. Ram, “Who does what: Collaboration patterns in the wikipedia and their impact on article quality,” *ACM Trans. Manag. Inf. Syst.*, vol. 2, no. 2, p. 11, 2011.

[13] M. Kämpf, S. Tismer, J. W. Kantelhardt, and L. Muchnik, “Fluctuations in Wikipedia access-rate and edit-event data,” *Phys. A Stat. Mech. its Appl.*, 2012.

[14] B. Keegan, D. Gergle, and N. Contractor, “Staying in the loop: structure and dynamics of Wikipedia’s breaking news collaborations,” in *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*, 2012, p. 1.

[15] M. Georgescu, N. Kanhabua, D. Krause, W. Nejdl, and S. Siersdorfer, “Extracting event-related information from article updates in wikipedia,” in *Advances in Information Retrieval*, Springer, 2013, pp. 254–266.

[16] P. Andriani and B. McKelvey, “Perspective-From Gaussian to Paretian Thinking: Causes and Implications of Power Laws in Organizations,” *Organ. Sci.*, vol. 20, no. 6, pp. 1053–1071, 2009.

[17] C. Pinto, A. Mendes Lopes, and J. A. Machado, “A review of power laws in real life phenomena,” *Commun. Nonlinear Sci. Numer. Simul.*, vol. 17, no. 9, pp. 3558–3578, 2012.

[18] Q. Chen, H. Chang, R. Govindan, S. Jamin, S. Shenker, and W. Willinger, “The Origin of Power-Laws in Internet Topologies Revisited,” in *{IEEE} Infocom 2002*, 2002.

[19] S. Carmi, S. Havlin, S. Kirkpatrick, Y. Shavitt, and E. Shir, “A model of internet topology using k-shell decomposition,” *Proc. Natl. Acad. Sci.*, vol. 104, no. 27, pp. 11150–11154, 2007.

[20] M. E. J. Newman, “The structure of scientific collaboration networks,” *Proc. Natl. Acad. Sci.*, vol. 98, no. 2, pp. 404–409, 2001.

[21] R. M. Kimmons, “Understanding collaboration in Wikipedia,” *First Monday*, vol. 16, no. 12, 2011.

[22] A. Ciffolilli, “Phantom authority, self-selective recruitment and retention of members in virtual communities: The case of Wikipedia,” *first monday*, vol. 8, no. 12, 2003.