Zipf's Law in MOOC Learning Behavior

Chang Men, Xiu Li
Graduate School at Shenzhen
Tsinghua University
Shenzhen, China
{menc15@mails, li.xiu@sz}.tsinghua.edu.cn

Zhihui Du
Department of Computer Science and Technology
Tsinghua University
Beijing, China
duzh@tsinghua.edu.cn

Jason Liu
School of Computing and Information Sciences
Florida International University
Miami, FL, USA

Manli Li, Xiaolei Zhang
Institute of Education
Tsinghua University
Beijing, China

***Abstract*—Learners participating in Massive Open Online Courses (MOOC) have a wide range of backgrounds and motivations. Many MOOC learners sign up the courses to take a brief look; only a few go through the entire content, and even fewer are able to eventually obtain a certificate. We discovered this phenomenon after having examined 76 courses on the xuetangX platform. More specifically, we found that in many courses the learning coverage—one of the metrics used to estimate the learners' active engagement with the online courses—observes a Zipf distribution. We apply the maximum likelihood estimation method to fit the Zipf's law and test our hypothesis using a chi-square test. The result from our study is expected to bring insight to the unique learning behavior on MOOC and thus help improve the effectiveness of MOOC learning platforms and the design of courses.**

Keywords-MOOC; learning coverage; maximum likelihood estimation; Zipf distribution

I. INTRODUCTION

The Massive Open Online Courses, or MOOC, have gained tremendous popularity since 2008 [1]. Besides the three pioneer platforms (Coursera, edX, and Udacity) [2], many other platforms have also been developed around the world, such as Khan Academy in North America, Miriada and Spanish MOOC in Spain, Iversity in German, FutureLearn in England, Open2Study in Australia, Fun in France, Veduca in Brazil, Schoo in Japan, and xuetangX in China [3]. Various universities, including many prestigious ones, nowadays develop and offer MOOC on these platforms. In doing so, MOOC has transformed education beyond the boundary of university campuses.

MOOC has also brought unparalleled opportunities for studying learning behavior. Online learning platforms maintain a rich record on the student population: the demographics, enrollment history, as well as online activities when interacting with the learning platforms. The latter includes browsing behavior, click stream, downloads, video streaming, and so on. Being able to access this data, albeit sanitized and anonymized, provides us the opportunity to analyze learning behavior at an unprecedented scale and detail.

Many researchers have taken interest in studying the learning behavior of MOOC participants. One of the most highlighted issues is how to measure the effectiveness of MOOC in general, given that the student completion rate (the proportion of students obtaining MOOC certificates) is substantially less than traditional online education courses [4]. The release of data points out a very low certification rate with an average less 15%. This problem has generated quite significant research efforts in studying the cause of low certification rate and thereby providing suggestion to improvement strategies (e.g., [5], [6]). MOOC has a large and diverse learner body with different intentions and motivations [7]. Many students engage with the courses and yet choose not to complete the assessments for credits. Consequently, the certification rate cannot be used as a reliable indicator for MOOC [8].

Another highlighted issue in studying the learning behavior is on the difference in the engagement patterns of learners as they interact with the learning platforms. Many researchers use the data collected by the MOOC platforms to define and extract prominent features to describe different learning behaviors and use them to identify different engagement patterns (e.g., [9]–[12]). The focus there is to classify learners into different categories by the engagement patterns and analyze their relationship with performance attributes, student demographics, social activities, and so on.

In this paper, we focus on the distribution of learning coverage. We define learning coverage as the amount of course materials accessed by the students. We found that the statistical distribution of learning coverage has an explicit long-tail feature. In particular, we found that, like many types of natural and man-made events, the learning coverage in MOOC observes the Zipf's law and can thus be approximated with a Zipf distribution.

In our study, we analyzed a dataset provided by xuetangX platform, containing over 40 million entries of event logs. The courses cover a wide range of disciplines,

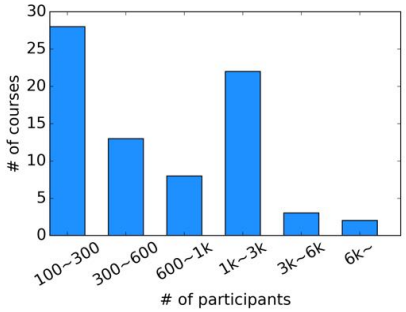

Figure 1. The distribution of course participants.

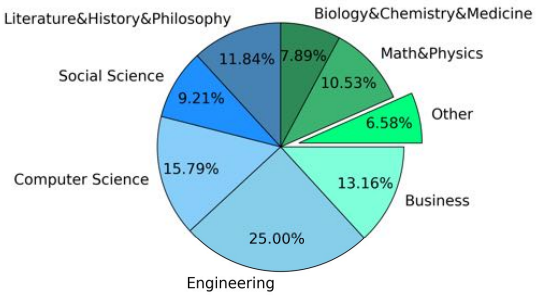

Figure 2. Courses across disciplines.

including mathematics, computer science, engineering, physics, chemistry, philosophy, history, business and so on. The results show that in over half of the courses the students' learning coverage on MOOC follows a Zipf distribution with only slight differences between the courses (in the exponent parameter), which we believe can be attributed to the inherent features of specific courses, such as their level of difficulty and popularity.

Our study is a first of its kind in that we explore and derive the statistical distribution of students' learning behavior by analyzing large datasets from MOOC with detailed event logs capturing the users' interaction with the online learning platforms. Our study is the first to show the existence of a Zipf distribution in the student engagement patterns. Our study can bring further insight of the unique learning behaviors of MOOC and thus can help both MOOC developers and course providers improve the effectiveness of the learning platforms and the design of the courses.

II. Related Work

A. MOOC Learning Behavior

Multidimensional data composed of user profiles and learning activities has been made available for researchers in education and data science. There have been studies attempting to establish relationships between students' background, motivation, and performance (e.g., [13], [14]). Many researchers classify students and activities according to the level of engagement with the online courses. For example, Perna et al. [15] define "starters" as those who register for a course no later than one week after its start date. Ho et al. [16] divide students into three types: "registrant" as any registered user in a course, "participant" as a registrant who has accessed the content of a course, and "explorer" as a participant who has accessed more than half of a course's content. Anderson et al. [11] classify users into five categories based on their accomplishment in the assignments: "viewers", "all-rounders", "solvers", "collectors" and "bystanders". Here, the collectors refer to those who primarily download lectures, while the bystanders refer to those with very low level of activities. Similarly, Kizilcec, Piech, and Schneider [9] define four types of learning patterns: "on track", "auditing", "behind", and "out". Evan et al. [12] define three types of activities: "engagement" refers to any activity such as downloading materials or watching lecture; "persistence" refers to engagement for a prolonged duration; and "completion" refers to persistence to the end of the course. Our study is also focused on student engagement. Our definition of learning coverage is a quantitative measure of student engagement in a particular course. We discuss learning coverage in detail in section IV-A.

B. Zipf's Law

Zipf's law builds on a fundamental premise that the occurrences of many types of natural and man-made events can be approximated with a Zipf distribution. Initially, Zipf's law was applied in the context of language studies. It states that given some corpus of natural language utterances, the frequency of any word is inversely proportional to its rank in the frequency table [17]. Thereafter, Zipf's law has been proven applicable to similar phenomena in various areas, such as population in cities, visits to websites, company size, science article citations, as well as natural and physical phenomena [18].

A Zipf distribution can be defined as $f(\mathrm{r}) = \mathrm{Cr}^{-\alpha}$, where $C = \left(\sum_{r=1}^{k} r^{-\alpha}\right)^{-1}$, and α is the exponent parameter with a positive value. In the classic version of a Zipf distribution, the exponent α is 1. If we plot the Zipf distribution, frequency versus rank, in log-log scale, the result is a line and the slope of the line is $-\alpha$. Because of this, most of the authors that claim the Zipf's law patterns (e.g., [18]–[23]) use linear regression to examine the linearity between log(frequency) and log(rank). The better the linearity is, the closer it is to a Zipf distribution.

This procedure, however, considers the intercept as a nuisance parameter, omitting the fact that it is related to α. More precisely, the intercept should be equal to $\log(C)$. Moreover, linear regression through ordinary least squares is inefficient in this case, given that r is an integer [24]. A better method to fit the Zipf's law for empirical data is to use the maximum likelihood estimation (MLE), which has been proven effective in practice for similar distributions, such as Zipf-Mandelbrot law [25] and power law [26]. In our study, we also use MLE to estimate the exponent parameter in the Zipf distribution and check the goodness of fit by performing

III. Dataset

In this study, we use a dataset provided by xuetangX (http: //www.xuetangx.com/) which contains data of 76 courses held by Tsinghua University in year 2014 and 2015.

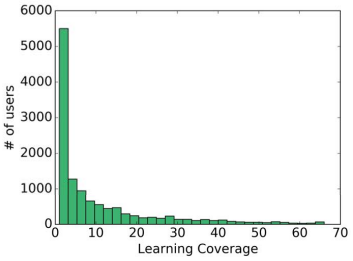

Figure 3. Histogram of learning coverage of a course.

The dataset contains information on individual users and courses, as well as the event logs of all users' online activities. There are more than 40 million event log entries in the dataset.

Fig. 1 shows the distribution of the number of participants for the courses. Here, participants refers users who have enrolled in and have accessed course content [16]. The minimum number of participants for a course is 101, and the maximum number of participants is 9,668. There are 27 courses with more than 1,000 participants.

Fig. 2 shows the distribution of courses in various disciplines. Courses in the dataset are already labeled with their subject areas as part of the course's summary information. As we can see, engineering, which is a mix of many subjects including electronic engineering, mechanical engineering, environmental engineering, and so on, has the largest number of courses (25%).

IV. METHOD

A. Learning Coverage

We define learning coverage as the amount of content a learner has accessed. On MOOC, the content is usually organized as a multi-level tree: each course contains several chapters, a chapter contains several sections, and a section contains various materials, including texts, videos, assignments and quizzes. Conceptually, one can calculate the learning coverage at different granularities (chapter, section, or specific content within a section). The xuetangX dataset contains event logs that record users' online activities with the learning platform. Using the event logs we can locate the specific section for each event. This enables us to count how many sections a learner has accessed.

As a typical example, Fig. 3 shows the histogram of learning coverage for the course, Financial Analysis and Decision Making, on xuetangX. Other courses would give a similar distribution. The number of students generally decreases as the learning coverage increases, but not monotonically. Using method presented in [26], we test the learning coverage for power law. However, the null hypothesis that the learning coverage fits power law is rejected in all 76 courses. We believe that the absence of monotonicity is a major cause for the rejection.

A long-tail feature of the frequency-rank distribution can be speculated. Consequently, we test the learning coverage for the Zipf's law, which describes the relationship between frequency and rank. We sort the frequency of each learning coverage in descending order, and then conduct linear regression to the frequency versus the rank in log-log scale as a pre-experiment. The results show that the learning coverage fits well with a Zipf distribution consistently. For all 76 courses, the estimated α ranges from 1.0018 to 2.2503. And for all but 3 courses, the R-squared value is larger than 90%, indicating a high goodness of fit. The result is encouraging. We decide to use the maximum likelihood method for a more effective and accurate estimation of the Zipf's law.

TABLE I. STATISTICS OF α AND p-value

Statistic	Mean	Min.	1Q	Median	3Q	Max.
α	1.3068	0.8915	1.2107	1.2998	1.3709	1.9752
p-value	0.3707	0.0000	0.0000	0.1863	0.8420	1.0000

B. Fitting Zipf's Law

Formally, a random variable X is Zipf distributed with parameter α $\left(X \sim Zipf_{\alpha}\right)$, if for a given $\alpha \in \mathbb{R}$,

$$p_{\alpha,r} = P\left(X = x_r\right) = \frac{C}{r^{\alpha}}, r \in \{1,2,...,k\}, \quad (1)$$

where x_r is the r^{th} frequent element, and C is the normalization factor: $C = \left(\sum_{r-1}^{k} r^{-\alpha}\right)^{-1}$.

Consider the observed sample $x = \left(n_1, n_2, ..., n_k\right)$ from a course, with n_i being the frequency of the i^{th} learning coverage, $n_1 \geq n_2 \geq ... \geq n_k$. Let $n = \sum_{i=1}^{k} n_i$. We obtain the likelihood function for sample x as follows:

$$l_{\alpha}\left(x\right) = \frac{n!}{n_1!n_2!...n_k!}\prod_{i=1}^{k}\left(p_{\alpha,i}\right)^{n_i}, \quad (2)$$

which gives the probability of the observed sample supposedly from a Zipf distribution with parameter α.

The method of MLE estimates α by finding a value of α that maximizes $l_{\alpha}\left(x\right)$. For ease of calculation, we maximize log-likelihood function, which is

$$\ln\left(l_{\alpha}\left(x\right)\right) = -n\ln\left(\sum_{i=1}^{k} i^{-\alpha}\right) - \sum_{i=1}^{k}\alpha n_i \ln\left(i\right) + \sum_{i=1}^{n}\ln\left(i\right) - \sum_{i=1}^{k}\sum_{j=1}^{n_i}\ln\left(j\right). \quad (3)$$

Therefore, the gradient with respect to α of log-likelihood function is:

$$\frac{\partial \ln\left(l_{\alpha}\left(x\right)\right)}{\partial \alpha} = -\sum_{i=1}^{k} n_i \ln\left(i\right) + n\left(\sum_{i=1}^{k} i^{-\alpha}\right)^{-1}\sum_{i=1}^{k} i^{-\alpha}\ln\left(i\right). \quad (4)$$

We can use gradient descent to obtain the optimal parameter, $\hat{\alpha}$, which maximizes the log-likelihood function.

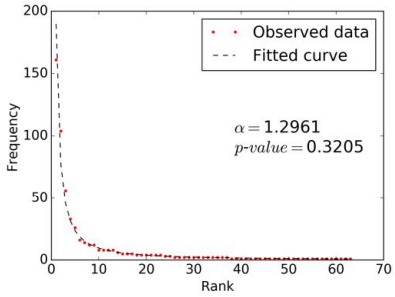

Figure 4. Fitting result of learning coverage for the course *Medical Parasitology*.

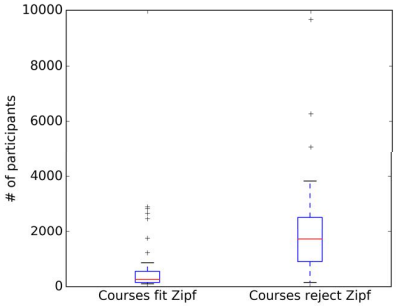

Figure 5. The number of participants for courses fitting and rejecting Zipf distriution.

C. *Goodness-of-fit Test*

The method described above allows us to fit a Zipf distribution to a given dataset and provide an estimate—$\hat{\alpha}$. Now we need to determine whether the sample data are consistent with the hypothesized distribution, in this case, a Zipf distribution with the parameter $\hat{\alpha}$. For this, we use the chi-square goodness-of-fit test [27]. The null hypothesis for the test is as follows:

H_0 : The data of learning coverage is consistent with a Zipf distribution with parameter $\hat{\alpha}$.

We can calculate chi-square statistic as:

$$\chi^2 = \sum_{i=1}^{k} \frac{\left(n_i - np_{\hat{\alpha},i}\right)^2}{np_{\hat{\alpha},i}}. \quad (5)$$

We can get a p-value with $k-1$ degree of freedom. The p-value is the probability that a chi-square statistic is more extreme than the calculated value from (5). We set the significance level to be 0.01. That is, if p-value<0.01, we reject the null hypothesis that the data of learning coverage is Zipf distributed with the parameter $\hat{\alpha}$; otherwise, the null hypothesis is not rejected.

V. RESULTS

We calculate the maximum likelihood estimation of α based on the log-likelihood function (3) and its gradient (4). For finding the optimal value of α that maximizes log-likelihood function, we use the MATLAB built-in function, called *fminunc*, as the optimization solver with the initial α set to be 1.5. After obtaining $\hat{\alpha}$, we conduct a chi-square test to determine whether the observed data fit with the Zipf's law. The procedure above is conducted for all 76 courses. (The code is available upon request.) The results of $\hat{\alpha}$ and p-value for the 76 courses are summarized in Table I. $\hat{\alpha}$ ranges from 0.8915 to 1.9752, which is largely consistent with our previous study using the linear regression method. Fig. 4 shows the results of parameter estimation together with the observed data in a course whose $\hat{\alpha}$ is close to the median. The results show that the learning coverage of 47 courses is likely to fit with the Zipf's law, which accounts for 61.84%.

The results also show that over 25% of the courses have a p-value approximate to zero, leading to a definite rejection of the null hypothesis. Fig. 5 compares the difference of the number of participants between the courses fitting a Zipf distribution and those not. We observe that the courses with more than 3,000 participants all reject the Zipf's law. On the other hand, most courses with less than 1,000 participants are likely to fit with the Zipf's law. We have tried to focus on other differences (e.g., disciplines, semesters) between the courses that report different conclusions, but so far no meaningful fundamental difference has been found. A previous study that uses a similar method for testing the Zipf-Mandelbrot model [25] claimed that the size of dataset should be no larger than 3,000 for the maximum likelihood method. The rejection of the null hypothesis may be caused by the limitation of the method.

The learning coverage is intrinsically related to how students perceive and carry on with the courses. In particular, the exponent parameter α can be regarded as an indicator of the student retention in the course. Higher α means that more students are either dropping out from the course early or simply taking in less content overall. That is, students are less engaged with an online course with a higher α than that of a lower α. It would be interesting to correlate this α parameter with course's content quality and alternative teaching methods (for example, more forum activities) to evaluate potential improvements in student retention. We leave the investigation for future work.

VI. CONCLUSION

To measure learner's engagement in MOOC, we introduce a new metric, called learning coverage, to estimate the amount of course content accessed by the learners. It is a measure of how far a learner has advanced into the course. By analyzing the dataset provided by the MOOC platform, specifically xuetangX, we calculate the learning coverage for 76 courses of various disciplines. We discover that the learning coverage distribution observes a clear long-tail feature. To confirm the observation, we apply the MLE method to fit the Zipf's law and conduct a chi-square goodness-of-fit test. We found that the learning coverage is likely to fit with a Zipf distribution in about 62% of the courses. The exponent parameter for the Zipf distribution can be used as an inherent feature of the course, representing in some degree the student retention in the course, and therefore a reflection of the course's difficulty and popularity.

Our study can be improved in several ways. First, the research context for this study considered learners on one Chinese platform as an under-researched and under-represented group due to data restrictions. The language-limited population of the study represents a limitation. Further work is needed to examine whether the results observed here can be generalized to learners on other MOOC platforms and learning contexts. Besides this, our method can be improved by using better methods to test the goodness-of-fit. The chi-square test depends on an adequate sample size for the approximations to be valid.

This study is our first attempt to analyze massive activity records data from the MOOC platforms. More in-depth studies on discovering knowledge behind the data are warranted. Not only will they provide us with methods for evaluating the effectiveness of learning on MOOC, but also they will provide the educators and MOOC providers with the basis for further improving the learning platforms and teaching methods. In future work, we would like to investigate other methods that can capture the learning behavior of the students more accurately, and therefore more accurately represent the learning behavior of the students.

ACKNOWLEDGMENT

This research is supported in part by National Natural Science Foundation of China (Nos. 61440057, 61272087, 61363019, 61073008 and 71171121), MOE research center for online education foundation (No. 2016ZD302), the Sci-Tech Interdisciplinary Innovation and Cooperation Team Program of the Chinese Academy of Sciences, the Specialized Research Fund for State Key Laboratories, and the National Key Research and Development Program of China (No. 2016YFB1000602), National "863" High Technology Research & Development Program of China (863 Project No. 2012AA09A408), and Shenzhen Science and Technology Project (Project No. JCYJ 20151117173236192 and CXZZ 20140902110505864).

REFERENCES

[1] T. R. Liyanagunawardena, A. A. Adams, and S. A. Williams, "MOOCs: A systematic study of the published literature 2008-2012," The International Review of Research in Open and Distributed Learning, vol. 14, no. 3, pp. 202–227, 2013.

[2] D. T. Seaton, Y. Bergner, I. Chuang, P. Mitros, and D. E. Pritchard, "Who does what in a massive open online course?" Communications of the ACM, vol. 57, no. 4, pp. 58–65, 2014.

[3] J. Qiu, J. Tang, T. X. Liu, J. Gong, C. Zhang, Q. Zhang, and Y. Xue, "Modeling and predicting learning behavior in MOOCs," in Proceedings of the 9th ACM International Conference on Web Search and Data Mining, 2016, pp. 93–102.

[4] R. Saadatdoost, A. T. H. Sim, H. Jafarkarimi, and J. MeiHee, "Exploring MOOC from education and information systems perspectives: a short literature review," Educational Review, vol. 67, no. 4, pp. 505–518,2015.

[5] H. Khalil and M. Ebner, "MOOCs completion rates and possible methods to improve retention – a literature review," in World Conference on Educational Multimedia, Hypermedia and Telecommunications, no. 1, 2014, pp. 1305–1313.

[6] M. Gaebel, MOOCs: Massive open online courses. EUA, 2014.

[7] L. Breslow, D. E. Pritchard, J. DeBoer, G. S. Stump, A. D. Ho, and D. T. Seaton, "Studying learning in the worldwide classroom: Research into edx's first MOOC," Research & Practice in Assessment, vol. 8, 2013.

[8] A. D. Ho, J. Reich, S. O. Nesterko, D. T. Seaton, T. Mullaney, J. Waldo, and I. Chuang, "Harvardx and MITx: The first year of open online courses, fall 2012-summer 2013," Ho, AD, Reich, J., Nesterko, S., Seaton, DT, Mullaney, T., Waldo, J., & Chuang, I.(2014). HarvardX and MITx: The first year of open online courses (HarvardX and MITx Working Paper No. 1), 2014.

[9] R. F. Kizilcec, C. Piech, and E. Schneider, "Deconstructing disengagement: analyzing learner subpopulations in massive open online courses," in Proceedings of the 3rd International Conference on Learning Analytics and Knowledge, 2013, pp. 170–179.

[10] T. Phan, S. G. McNeil, and B. R. Robin, "Students' patterns of engagement and course performance in a Massive Open Online Course," Computers & Education, vol. 95, pp. 36–44, 2016.

[11] A. Anderson, D. Huttenlocher, J. Kleinberg, and J. Leskovec, "Engaging with massive online courses," in Proceedings of the 23rd international conference on World wide web, 2014, pp. 687–698.

[12] B. J. Evans, R. B. Baker, and T. S. Dee, "Persistence patterns in Massive Open Online Courses (MOOCs)," The Journal of Higher Education, vol. 87, no. 2, pp. 206–242, 2016.

[13] M. Liu, J. Kang, and E. Mckelroy, "Examining learners' perspective of taking a MOOC: reasons, excitement, and perception of usefulness," Educational Media International, vol. 52, no. 2, pp. 129–146, 2015.

[14] N. Hood, A. Littlejohn, and C. Milligan, "Context counts: How learners' contexts influence learning in a MOOC," Computers & Education, vol. 91, pp. 83–91, 2015.

[15] L. W. Perna, A. Ruby, R. F. Boruch, N. Wang, J. Scull, S. Ahmad, and C. Evans, "Moving through MOOCs: understanding the progression of users in Massive Open Online Courses," Educational Researcher, vol. 43, no. 9, pp. 421–423, 2014.

[16] A. D. Ho, I. Chuang, J. Reich, C. A. Coleman, J. Whitehill, C. G. Northcutt, J. J. Williams, J. D. Hansen, G. Lopez, and R. Petersen, "Harvardx and MITx: Two years of open online courses fall 2012summer 2014," Available at SSRN 2586847, 2015.

[17] Wikipedia, "Zipf's law," https://en.wikipedia.org/wiki/Zipf%27s law.

[18] W. Li, "Zipfs law everywhere," Glottometrics, vol. 5, pp. 14–21, 2002.

[19] X. Gabaix, "Zipf's law for cities: an explanation," Quarterly journal of Economics, pp. 739–767, 1999.

[20] Y. Fujiwara, "Zipf law in firms bankruptcy," Physica A: Statistical Mechanics and its Applications, vol. 337, no. 1, pp. 219–230, 2004.

[21] L. A. Adamic and B. A. Huberman, "Zipfs law and the internet," Glottometrics, vol. 3, no. 1, pp. 143–150, 2002.

[22] K. Okuyama, M. Takayasu, and H. Takayasu, "Zipf's law in income distribution of companies," Physica A: Statistical Mechanics and its Applications, vol. 269, no. 1, pp. 125–131, 1999.

[23] T. Yamakami, "A Zipf-like distribution of popularity and hits in the mobile web pages with short life time," in 2006 Seventh International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT'06), 2006, pp. 240–243.

[24] C. M. Urz´ua, "Testing for zipfs law: A common pitfall," Economics Letters, vol. 112, no. 3, pp. 254–255, 2011.

[25] F. Izs´ak, "Maximum likelihood estimation for constrained parameters of multinomial distributionsapplication to Zipf–Mandelbrot models," Computational statistics & data analysis, vol. 51, no. 3, pp. 1575–1583, 2006.

[26] A. Clauset, C. R. Shalizi, and M. E. Newman, "Power-law distributions in empirical data," SIAM review, vol. 51, no. 4, pp. 661–703, 2009.

[27] H. Chernoff, E. Lehmann et al., "The use of maximum likelihood estimates in $\chi 2$ tests for goodness of fit," The Annals of Mathematical Statistics, vol. 25, no. 3, pp. 579–586, 1954