Quantitative Analysis of User-Generated Content on the Web

Xavier Ochoa
Escuela Superior Politécnica del Litoral
Via Perimetral Km. 30.5
Guayaquil, Ecuador
593-4-2269777 (7006)
xavier@cti.espol.edu.ec

Erik Duval
Dept. Computerwetenschappen
Katholieke Universiteit Leuven
Celestijnenlaan 200 A B-3001 Leuven
32-16-327066
Erik.Duval@cs.kuleuven.be

ABSTRACT
User-generated content (UGC) is becoming the most popular and valuable information available on the WWW. However, little serious research has been conducted to measure the properties of its production process. This paper presents an in-depth quantitative analysis of 9 popular websites that are based on different UGC types. The Information Production Process is used as a framework for the analysis. The findings provide for first time strong scientific evidence for previously anecdotic knowledge: UGC production follows "long-tail" distributions and it is marked with a strong "participation inequality". Also, the analysis arrived to unexpected findings: not all the UGC types follow the inverse power-law distribution, and large content collections could be dominated by the presence of ultra-productive users. The analysis results also have implications for the administration of UGC-based websites.

Categories and Subject Descriptors
H.3.5 [**Information Storage and Retrieval**]: Online Information Services – *data sharing, web-based services*

General Terms
Measurement, Economics, Experimentation

Keywords
User-generated content, UGC, IPP, Lotka, Weibull.

1. INTRODUCTION
Technologies that enable internet users, not only to access, but also to contribute content are responsible for a revolution in the way information is created and distributed. Early examples of these technologies are electronic bulletin boards and the Usenet. These extremely popular systems provide users with virtual fora to discuss innumerable topics. No central authority was responsible for the content of each forum. Moderators shaped, but did not create the content. More recently, Web-based repositories invited users to contribute information in order to create a (semi-) public good. For example, the most complete public database of CD album and track information, CDDB [21], was not created by obtaining the information from recording companies. Instead, this database was created combining the submission of individual users that enter the information for their personal cataloguing. Nowadays, a plethora of technologies (wikis, tagging, rating, social networks, etc.), loosely grouped under the name of Web 2.0 [18], enable user participation at a new level. Some Web sites use these technologies to add value to their existing content. For example, user reviews in Amazon (amazon.com) or Digital Photography Review (dpreview.com) while others are entirely made from user contributions, such as user submitted photos on Flickr (flickr.com) or news on Digg (digg.com). User-Generated Content (UGC) is in some fields rendering obsolete (or at least unpopular) traditional professional-generated content. The number of consults to Wikipedia (wikipedia.org) compared with Encyclopedia Britannica (britannica.com) or the number of visits to YouTube (youtube.com) in contrast to BBC Video Service (bbc.co.uk/videonation) are just a few examples of this trend.

UGC, also known as "User-Created Content" or "Consumer Generated Media", is defined in [19] as: "1) Content made publicly available over the Internet, 2) Which reflects a certain amount of creative effort and 3) Which is created outside of professional routines and practices". This definition is not universally accepted (see [13] and [16] for competing definitions). Also, it does not always hold true: some UGC is only available for a closed group or is just repackaging of content without any contribution [3] or is created by professionals as in the case of enterprise sponsored blogs. Nonetheless, this definition reflects the main characteristics shared by very different content types published by internet users.

Contributing UGC is currently a mainstream activity among Web users. According to [9], 35% of the USA Internet users (48 million) have contributed at least once a UGC to the Web. The OECD report on "Participative Web: User-Created Content" [19] also shows similar trends in Europe, Japan and Korea. Despite of its importance for current Web composition, there are open research questions about UGC production not addressed by current literature. What is the average number of contributions by an author? Is the participation distribution the same for different kinds of UGC? Pareto types of rules (inequality) holds true for UGC? What is the impact of production effort in the amount of contribution per user? Another example of the lack of academic research in the field (or maybe a sign of the UGC times) is that most discussions on the topic are conducted through blog postings and comments [12] [23]. To the knowledge of the authors, one of the most rigorous quantitative analysis of the UGC production was conducted for Usenet participation [27] in 1998. More recently, the Participation Inequality rule [22], also known as the 90-9-1 (meaning that 90% of the users do not contribute, 9% contribute a few elements and that 1% contribute a lot) has been used as a rule-of-thumb to measure the UGC production. There has been no study that proves or denies this rule.

Maybe because its economic implications, there has been more research activity on the topic of UGC consumption or popularity.

The assertion that the consume of media had a long-tail [1], inspired the research on the consumption of UGC media. The popularity of different users or content in UGC-based sites, especially YouTube, have been already quantitatively analyzed [3]. While using similar techniques, this paper does not deal with the conmsuption side of UGC, just with its production or publication.

The main contribution of this paper is to provide the first in-depth quantitative analysis of different modern UGC production types. This analysis could help answering the research questions mentioned above. First, in Section 2, we frame our research in the context of the Information Production Process (IPP), widely used in Informetrics and Webometrics studies. In Section 3, we explain how we sampled and collected data from 11 different websites that use UGC as an important or exclusive source of content. These data is analyzed in Section 4. First, simple descriptive statistics are obtained. Then, several distributions are fitted to the random variable that represents the number of items contributed per user. Finally, the cumulative distribution of global contribution is analyzed. The implications of these findings are discussed in Section 5. We end this paper with general conclusions.

2. UGC Production as an Information Production Process

Egghe introduced in [5] the concept of Information Production Process (IPP). The objective of the IPP is to establish a quantitative relation between the producers and the information items being produced. An IPP is an informetric system made by triplets of the form (S,I,F): S are the information sources, I are the information items produced and F is production function. The most representative and studied example of an IPP is the paper publication process: the sources are the authors, the items are the papers and the production function is the Lotka distribution [7]. The Lotka distribution, in this case, relates the expected number of authors that have published a given number of papers. Other well-known IPPs [7] are the frequencies of words in a text, number of papers published in a journal, number of inhabitants per city, number of links to a web-site, etc. We will represent the UGC production as an IPP. The contributing users will take the place of the sources; the produced content will be the items; and we will try to establish which function is the most appropriate to establish a relation between the percentage of the sources that publish a given number of items.

The main representation of the IPP is the size-frequency plot [17]. In this plot, the x represents the amount of items and the y represents the probability of a user publishing x items. For empirical data, the y is the relative frequency of users that publish x items. The plot axes could be linear or logarithmic depending on the distribution type. Related plots like rank-frequency or size-frequency with logarithmic binning could also be used to help reducing the noise to graphically fit probable distributions [17]. In this paper we will use analytical methods to distribution fitting and, as a consequence, we will use the size-frequency plot to preserve the original representation of the data.

There are two main benefits that result from this representation: 1) Analysis results of different UGC production types could be easily compared with other source-item relations widely studied in Informetrics (Bibliometrics, Scientometrics and Webometrics) and 2) Previous theoretical knowledge on IPP is directly transferable to UGC production. This last point is especially important given the lack of previous research in this UGC production.

3. DATA COLLECTION AND SAMPLING

As a way to encourage the contribution of content, the majority of websites with a strong UGC component present the contributor username together with the content metadata (title, description, etc.). Moreover, several sites provide user pages where all contributions of a given user are counted and listed. We make use of those features to collect, through web scrapping or API use, the number of items produced by a sample of users for 11 websites that heavily rely on UGC.

Collect the information for all the contributing users is, in most of the cases, not feasible. This is especially true for the method of page scrapping. To still be able to obtain statistically sound conclusions from the collected data, random sampling of at least a 1% of the user base was performed for each site. The 9 websites were chosen based on two criteria: 1) That the site is representative enough of a given UGC type (bookmarks, reviews, metadata, video, news links, etc.) and 2) That the site allows the collection of the requiring sample. For example, while YouTube could be considered the most representative user-generated video site, it has anti-scrapping mechanisms and has a query limit of 500 videos for its API. In this study, we use Revver, another well-known user-generated video site, with a big contributing community, but without the YouTube restrictions.

The following list presents the 11 Websites and the method used to sample the data:

- Furl (furl.com) - Social Bookmarking: Users contribute with links to interesting Websites. The 3500 more recent contributors were taken as the sample.
- Amazon (amazon.com) - Book Reviews: Users contribute with reviews about books sold in the site. The users with more than 10 reviews were taken as the sample.
- LibraryThing (librarything.com) – Book Cataloging: Users contribute with metadata about books. The first 4300 users based on alphabetical ordering were taken as the sample.
- Merlot (merlot.org) – Learning Object Referatory: Users contribute with links and metadata about educational material on the web. All the active users (more than 1 contribution) were taken as the sample.
- Digg (digg.com) – Social News: Users contribute with link to interesting stories. The contributors to the last month stories were taken as the sample. Only stories in that month were considered to count the number of contributions per user.
- SlideShare (slideshare.net) – Presentation Publication: Users contribute with PowerPoint or PDF slide show presentations. The authors of the 5000 most popular presentation were taken as the sample. Only presentations in the 5000 most popular were considered to count the number of contributions per user.
- Scribd (scribd.com) – Document Publication: Users contribute with word-processing documents as Word or PDF. The 15000 most discussed authors were taken as the sample.

- Revver (revver.com) – Video Publication: Users contribute with short videos. The 3255 more recent contributors were taken as the sample.
- Fan Fiction (fanfiction.net) – Literary Publication: Users contribute with stories inspired in existing, professional generated, books, tv series or movies. All members that have published a story related to "Lord of the Rings" were taken as the sample. Only "Lord of the Rings" related stories were considered to count the number of contribution per user.

Data were collected in the week between 5/10/2007 and 10/10/2007. The scrapping tool was implemented by the authors in Java. This application ran on a single thread with time intervals between downloads to avoid being flagged as malware, therefore blocked, by the web application. The final number of users and contributions obtained for each sites is shown in Table 1.

Table 1. Information about sampled UGC-based sites. Presents the mnemonic code for each site, the type of content published and the total number of users and contributions considered in the study

Code	Website	UGC Type	User (Sources)	Contrib. (Items)
FR	Furl	Bookmarks	3,500	808,520
AM	Amazon	Reviews	82,365	3'100,671
LT	LibraryThing	Book Metadata	4,300	355,630
MR	Merlot	LO Metadata	2,675	17,379
DG	Digg	News	55,388	196,896
SS	SlideShare	Presentations	2,383	5,000
SC	Scribd	Documents	15,000	175,850
RV	Revver	Videos	3,255	69,519
FF	Fan Fiction	Stories	7,451	17,624

The data was not filtered in any way. All the sampled users were taken into account. Outliers were preserved because the data intrinsic distribution was not known. The data were converted to comma separated text files and they are available for download[1] for the interested reader.

4. QUANTITATIVE ANALYSIS

4.1 Simple Descriptive Statistics

The simplest analysis that can be conducted with the data is to obtain common descriptive statistics. We analyze X, the random variable that represents the number of items produced by each user. We will use this analysis to analytically test if the distribution of X is symmetric around a central value (for example normally distributed) or if it presents a left or right tail. The values necessary for this analysis are the quartiles, mean and skewness of X. Those values are presented in Table 2 for each of the 9 empirical data sets.

The values of the quartiles already present a very asymmetric distribution for all the data sets. In the case of Digg, at least 50% of the users have only produced 1 news item each, while users in the forth quartile have contributed from 3 to 451 news. The mean value also provide evidence for the inequality. For all the 9 data sets, more than 75% of contributors had produced less than the mean number of items. The final piece of evidence is the skewness value. High and positive skewness means that the distribution has a right long tail. All the empirical data sets representing the UGC production of the 9 Websites are not symmetric.

[1] http://www.cti.espol.edu.ec/Learnometrics/files/datawww.zip

Due to its asymmetric nature, traditional values of mean and standard deviation cannot be used to describe the data. For example, Scribd users, in average, have contributed 2.01 presentations. However, more than 75% of them have contributed 2 or less. A user even has contributed 44,060 presentations, more than 100 standard deviations from the mean. This dispersion would be nearly impossible in a normal distribution.

Table 2. Descriptive statistics of the sampled Websites.

C.	Min	Q1	Q2	Q3	Max	Mean	Sk.
FR	1	3	18	91	24,920	231.7	11.10
AM	10	12	18	31	14,950	37.7	43.19
LT	1	3	16	70	4,503	82.72	8.36
MR	1	1	1	3	2,094	6.50	31.39
DG	1	1	1	3	451	3.55	16.57
SS	1	1	1	2	55	2.01	6.94
SC	1	1	1	2	44,060	11.76	87.05
RV	2	4	9	24	2,823	29.52	16.54
FF	1	1	1	2	112	2.37	8.96

To visualize the analytical findings, the data was plotted as size-frequency. First, lineal axes were used (see Fig. 1). The shape for the 9 data sets was the characteristic L of long tail distributions. To gain a better insight of the type of distribution, the data was again plotted as size-frequency but with logarithmic axes (see Fig. 2). On log-log axes, the plot showed an almost-straight-line for all the data sets. The log-log plot also confirms the assumption made by IPP that the production function is strictly decreasing. At the end, the tail becomes wider and noisier. This effect is due to the discrete nature of the data (an item can only be produced by one source) and also to the fact that at high production numbers, it is difficult to find two users with the same amount of published items.

The linear or almost linear shape of the log-log size-frequency plot suggests that the data follow an inverse-power law or similar distribution. In the following subsection several of the most common distributions for Informetric distributions will be fitted against the empirical data.

4.2 Distribution of Items per Source

The first candidate distribution when the data have an inverse linear shape in the log-log plot is the inverse power-law [17] (also known as Lotka distribution). However, other distributions, like LogNormal, Weibull or Exponential could also present a similar behavior over some decades of the log-log plot [4]. In order to analytically determine the most adequate distribution for studied data sets, we will fit and compare several distributions using the method proposed by Clauset et al. in [4] to fit long-tailed distributions. For readability reasons we will briefly describe this method. 1) the parameters of each distribution are estimated from the empirical data through Maximum Likelihood Estimation (MLE) [8]. 2) the likelihood ratio test [26] is used to establish which of the competing distributions provide a better fit to the data. 3) the Kolmogorov-Smirnov (K-S) test [15] is applied to the best fitted distribution to establish if it is a good approximation to the data.

We will fit 7 statistical distributions to the 9 empirical data sets: Lotka, Lotka with exponential cut-off, Yule, LogNormal, Weibull, Exponential and Poisson. These distributions were chosen because they have been used to explain IPPs before or because they have an almost-linear behavior for some decades of the log-log plot. The probability mass function of these distributions is presented in Table 3. The fitting was performed using the R

statistical software and a custom adaptation of the code provided in [4]. A copy of the R procedures is available [2] for the repeatability of the analysis. The results of the distribution fitting can be seen in Table 4. Each data set is listed with its most probable distribution, the fitted parameters and the D value of the K-S test. The D value is used to determine if the selected distribution is a good fit for the data.

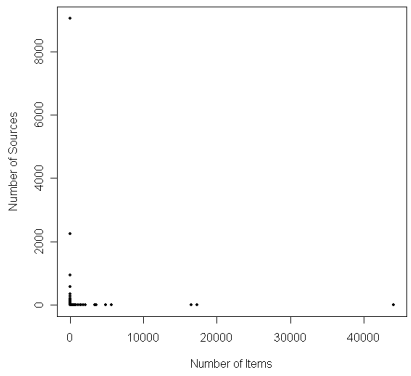

Figure 1. Linear Size-Frequency Plot for the Scribd data set. The L shape suggests a distribution with long tail

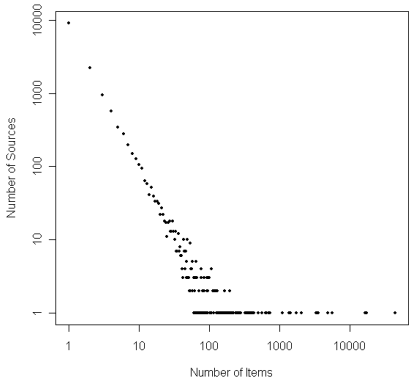

Figure 2. Log-Log Size-Frequency Plot for the Scribd data set. The plot presents an inverse linear relation between the number of items published and the number of sources with that yield.

According to the best-fitting distribution, the IPPs could be classified in two groups: 1) IPPs where the production function is Lotka or Lotka with exponential cut-off. Amazon Reviews, Merlot, Digg, SlideShare and Scribd belong to this group. 2) IPPs where Weibull is the production function. Furl, LibratyThing and Revver are part of this group.

If the empirical and theoretical distributions are plotted as log-log frequency-size, the difference between the two groups is clear (Fig. 3 and 4). The first group, explained by the Lokta distribution, presents a narrow linear behaviour during the upper and middle part and finishes with a "fat-tail". IPPs explained by Weibull differ from the linear behavior and present a "fat-belly" that the Lotka distribution, even with exponential cut-off, could not explain.

Table 3. Formulas for the candidate distribution tested against the empirical data. The constant values are omitted for clarity.

Distribution	Probability Mass Function (without constants)
Lotka	$x^{-\alpha}$
Lotka with exponential cut-off	$x^{-\alpha}e^{-\lambda x}$
LogNormal	$\frac{1}{x}\exp\left(-\frac{(\ln(x-\mu))^2}{2\sigma^2}\right)$
Weibull	$x^{\beta-1}e^{-\lambda x^{\beta}}$
Exponential	$e^{-\lambda x}$
Yule	$\frac{\Gamma(x)}{\Gamma(x-\alpha)}$
Poisson	$\mu / x!$

The different distributions could be attribute to different ways in which the contributor base grow, as well as how the rate of production of each contributor increase over time. Egghe in [7] proved that the Lotka distribution is the result of an exponential increase in the number of sources and the exponential increase in the rate of production of each source. It can also be proved that the Weibull distribution is the result of an exponential increase in the number of sources together with a polynomial (x^n) increase in the rate of production of each source. The proof of this assertion is similar to the one produced by Egghe, but is not included in this paper due to space constrains. Further analysis is needed in order to corroborate if the mathematical explanations agree with the empirical growth in the contributor base and the rate of production of each contributor.

Table 4. The best fitted distribution for each empirical data set, the fitted parameters for the distribution, the Kolmogorov-Smirnoff D value and the final conclusion of the goodness of fit.

C.	Best fitted Distribution	Parameters	K-S D	Good Fit
FR	Weibull	$\lambda = 0.24$ $\beta = 4.87$	2.5×10^{-2}	Yes
AM	Lotka with cut-off	$\alpha = 2.11$ $\lambda = 6.8 \times 10^{-4}$	1.3×10^{-2}	Yes
LT	Weibull	$\lambda = 0.35$ $\beta = 12.29$	2.6×10^{-2}	Yes
MR	Lotka with cut-off	$\alpha = 1.87$ $\lambda = 6 \times 10^{-4}$	2.2×10^{-2}	Yes
DG	Lotka with cut-off	$\alpha = 1.91$ $\lambda = 7 \times 10^{-3}$	2.8×10^{-3}	Yes
SS	Lotka with cut-off	$\alpha = 2.06$ $\lambda = 3 \times 10^{-2}$	4.7×10^{-3}	Yes
SC	Lotka	$\alpha = 1.97$	6.2×10^{-3}	Yes
RV	Weibull	$\lambda = 0.33$ $\beta = 1.69$	1.2×10^{-2}	Yes
FF	Lotka with cut-off	$\alpha = 1.86$ $\lambda = 4 \times 10^{-2}$	8.1×10^{-3}	Yes

The next step in the analysis is to determine if the distribution of the number of items contributed by each source affect the overall distribution of the contribution.

[2] http://www.cti.espol.edu.ec/Learnometrics/files/codewww.zip

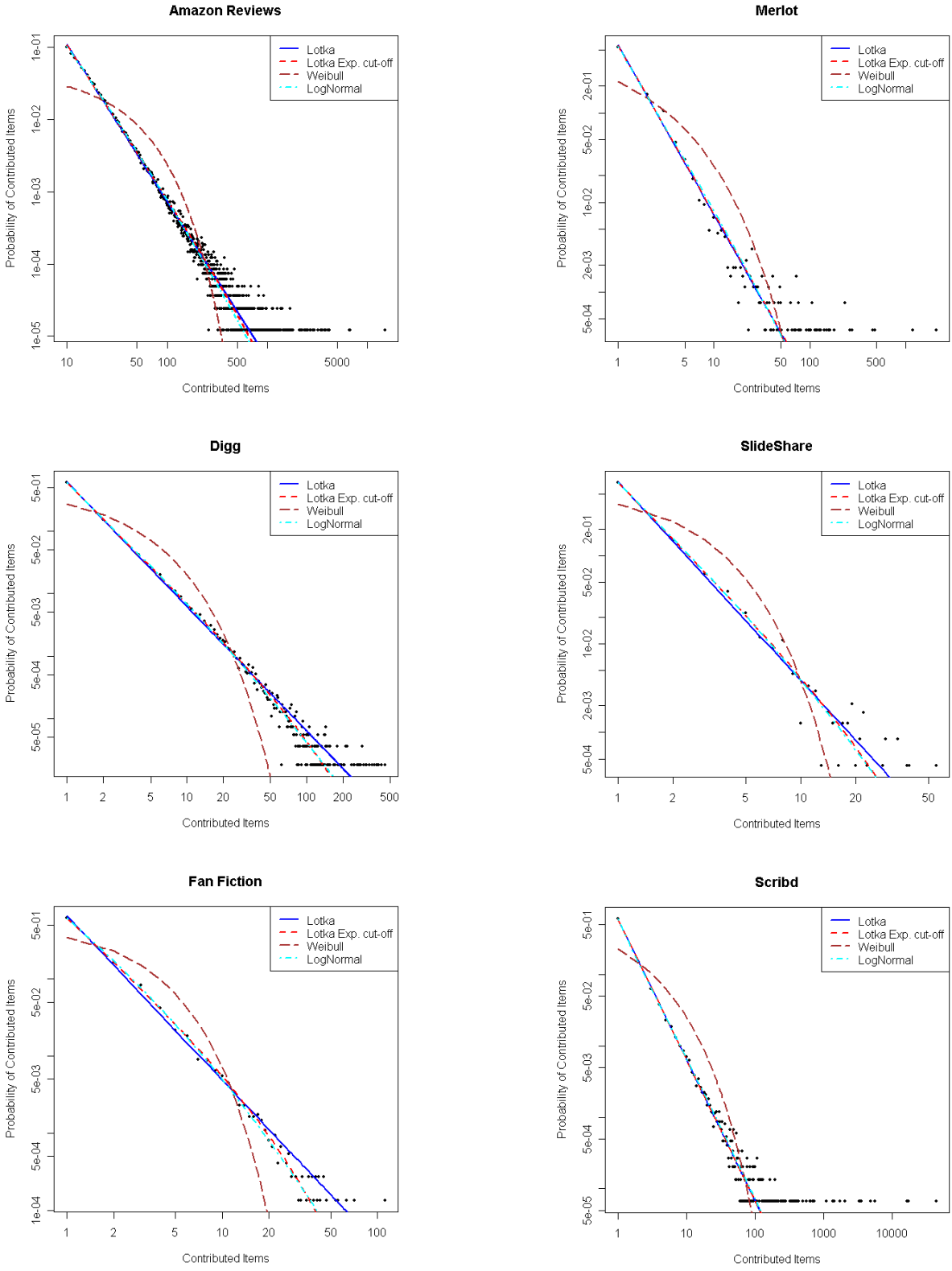

Figure 3. Log-Log Size-Frequency Plot for the "fat-tail" UGC production processes. The points represent the empirical data, while the lines represent the best fitting of the different distributions. Lotka and Lotka with Exponential cut-off are the best fitting distributions for this group.

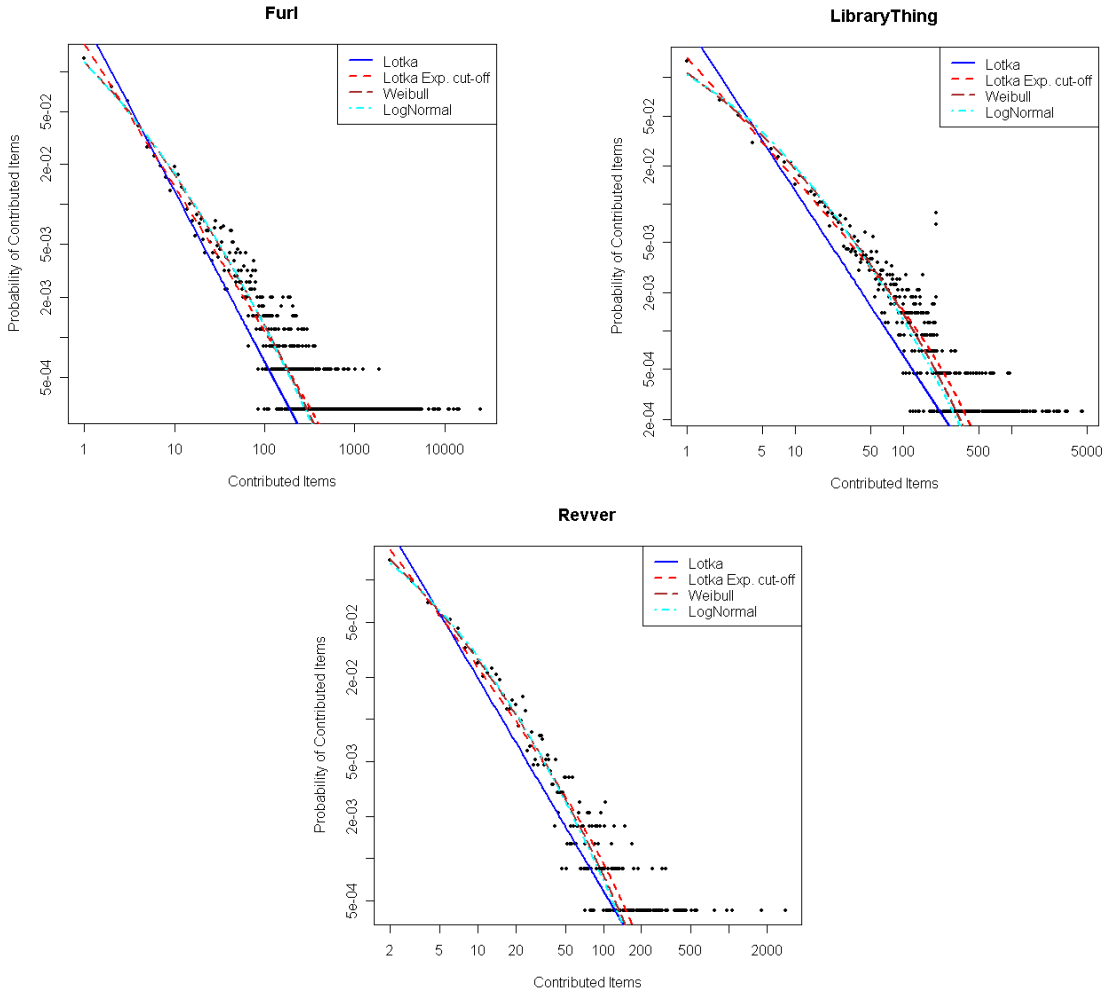

Figure 4. Log-Log Size-Frequency Plot for the "fat-belly" UGC production processes. The points represent the empirical data, while the lines represent the best fitting of the different distributions. Weibull is the best fitting distribution.

4.3 Distribution of the Contribution

The empirical data could be represented through its cumulative mass function (CDF). The function shape can be used to estimate the differences in contribution from different segments of users.

The CDF value is taken at fixed proportions of the number of sources. These values represent which proportion of items has been contributed by the corresponding proportion of sources. Table 5 shows the CDF taken for the most prolific source, at 1%, 10%, 20%, 40%, 60% and 80% of the amount of sources. The IPPs are grouped according to their production function.

Three groups can be inferred from the data. 1) Amazon Reviews, Digg, FanFiction and SlideShare seem to have a similar distribution of contribution. The 10% of the users contribute from 40% to 60% of the content. 2) a group is integrated by the "fat-belly" IPPs: Furl, LibraryThing and Revver. In those cases, the 10% of the users contribute between 60% and 80% of the content. 3) Scribd and Merlot form the third group. In this group, the most prolific sources seem to have a big impact in the overall number of items contributed (25% and 12% respectively for the most prolific contributor). For these IPPs, the 1% of the sources generate from 40% to 70% of the content.

Table 5. The cumulative contribution of different user segments

C.	First	1%	10%	20%	40%	60%	80%
"Fat-tail" IPPs – Lotka							
AM	0.5%	20%	50%	63%	78%	87%	94%
DG	0.2%	23%	58%	70%	83%	89%	94%
FF	0.6%	13%	43%	57%	75%	83%	92%
MR	12%	44%	75%	82%	90%	94%	97%
SC	25%	74%	87%	91%	95%	97%	98%
SS	1.1%	12%	41%	56%	71%	81%	90%
"Fat-belly" IPPs – Weibull							
FR	3.0%	32%	82%	91%	98%	99%	100%
LT	1.2%	23%	64%	81%	94%	98%	100%
RV	4.1%	23%	61%	76%	89%	95%	98%

This results confirm the rule-of-thumb rule known as "Participation Inequality" that suggest that 90% of the content is generated by 10% of the contributors.

5. IMPLICATION OF FINDINGS

The knowledge extracted from the previous analysis should have implications in the way that UGC production is understood and managed. Following is a list of inferences that can be drawn:

There is no such thing as an average user. UGC production is not a normal distributed process. From the contributing users, the majority contribute few items, whereas few contribute a lot. As has been proved in [7], the mean, when the α parameter of Lotka is less than 2 (most of our cases), is a meaningless measurement. A system that consists of logarithmic levels, similar to the one used to classify economic strata [20], should be a better way to describe the user base.

The production of different UGC types is similar, but not the same. In the 9 sampled Websites we found two clearly different distributions. Seven data sets can be classified as "fat-tail" IPPs where the distribution follows very closely the Lotka distribution or straight line in the log-log plot. Three data sets had a "fat-belly", a pronounced curvature in the middle of their range. These sets were best fitted by a not-straight distribution as Weibull. Mathematical analysis suggests that this different groups are created by difference in the change of the rate of contribution.

Pareto also applies to UGC production. The 80/20 Pareto rule (more like 76/20 in our average case) is a good rule-of-thumb to establish the distribution of the contribution of UGC. Nevertheless, as any rule-of-thumb, it has to be used as a guide, but not to replace a real measurement. The actual distribution of individual UGC data is very sensible to extreme contributors as in the case of Scribd and Merlot. Those extreme contributors cannot be considered outliers, because they are natural occurrences due to the power-law distribution [25].

"Fat-tail" UGC production is similar to professional production. The range of the alpha values found for "fat-tail" UGC production goes from 1.86 to 2.11. This is consistent with the findings of previous studies for professional/academic books [11] and scientific papers [4] production. Also, according to the original work of Lotka [14] on papers published in Journals, alpha is approximately 2.

The contribution effort has no effect in the distribution of the contribution. The contribution of Digg News, Amazon Book Reviews and Fan Fiction Stories follows a similar distribution (Lotka with exponential cut-off with $\alpha \approx 2$) even if the effort required to contribute a link to a news item is much lower than to create a multi-page literary story. The distribution of "fat-belly" IPPs is also insensible to the effort (For example Furl bookmarks against Revver videos).

The amount of published items has no effect in the distribution of the contribution. The distribution type of UGC production is not dependent on the collection original size. Merlot contributions, with almost 18.000 learning objects in total, follow a similar distribution than Amazon book review contributions, having more than 30'000.000 book reviews in total. This property is called "self-similarity" and it is a characteristic of power-law distributions [6]. Unfortunately, the Weibull-based distributions have a similar number of items and no conclusions could be inferred from our analysis.

If you have a "fat-belly", take care of your star users. If we retain just the 10% of the most productive users in a normal "fat-belly" scenario we retain more than the 60% of the material. On the other hand, in normal "Fat-tail" scenarios, 10% of the most productive users could only represent the 40%. Maybe this is one of the reasons why Netscape Propeller (netscape.com) is not as successful as Digg (a "fat-tail" IPP) even if it paid the 50 most prolific users from Digg [2] to publish in Propeller. According to our measurement, those 50 users (0.1%) only contribute a 6% of the content.

Informetrics can help us to understand UGC production. The shape and distribution of most of the sampled UGC production processes has already been found in other Information Production Processes. Due to their strictly decreasing production function, the lotkaian informetrics (one of the most developed branches of Informetrics [24]) can be used to study the characteristics and properties of UGC production.

UGC production can help us to understand Informetrics. Weibull is not a traditional informetric distribution for production process. The finding of some examples where this distribution fits the data is interesting for Informetric research. Exploring the richness of variety of UGC could provide relevant lessons to understand other IPPs.

6. CONCLUSIONS

This quantitative analysis about user-generated content (UGC) production for 9 different types of communities confirms, for first time, what has been anecdotical knowledge: That amateur users contribute online material in a similar manner (same type of size-frequency distributions) than traditional authors contribute more established media forms as scientific papers or books, even if the entry barriers and publishing channels are completely different. Few exceptional users produce 2 to 3 orders of magnitude more items than the majority of the user base. This result provides empirical test to the second part of the "Participation Inequallity" rule which establishes that 90% of the contributing users produce few items, while 10% produce most of the content. The analysis was also able to capture the fact that single users were responsible for more than the 10% of the content, producing sometimes 4 or 5 orders of magnitude more items than the majority of the users. This effect, almost impossible to see in Gaussian (Normal) distributions, is not uncommon under power-law conditions and has received the popular name of black-swan [25].

A closer look to UGC production distributions also provides evidence of differences with established Informetric distributions. The "fat-belly" UGC production process cannot be explained purely by the "success-breed-success" or the "preferential attachment" mechanism used to justify Lotka distributions. While this paper suggest a difference in the change of the rate of production as the cause of this characteristic, further theoretical work in the lines of [10] and [6] should be conducted in order to find mathematical explanations and empirical corroborations to the fitting of the Weibull distribution for a Information Production Process. The study of more examples of different types of UGC production, together with their measurment in different communities of practice should provide valuable empirical information to extend the understanding about Information Production Processes.

Finally, the calculations made in the analysis phase were performed in the R statistical package. The same calculations can be used by the owners or administrators of UGC-based sites in order to obtain deep knowledge of how their users contribute and

what could be the best strategy to increase the production of the site while making a rationale use of the resources.

7. REFERENCES

[1] Anderson, C. The long tail. Hyperion. (2006)

[2] Calacanis, J. Kevin Rose cracks. http://www.calacanis.com/2006/07/25/kevin-rose-cracks-or-how-to-know-when-youve-won-the-debate/. Retrived October 2, 2007

[3] Cha, M., Kwak, H., Rodriguez, P., Ahn, Y. Y., & Moon, S. I Tube, You Tube, Everybody Tubes: Analyzing the World's Largest User Generated Content Video System. Proc. ACM Internet Measurement Conference (IMC) (San Diego, CA, October 2007)

[4] Clauset, A., Shalizi, C. R., & Newman, M. E. J.. Power-law distributions in empirical data. Reviews of Modern Physics (2007)

[5] Egghe, L.. The duality of informetric systems with applications to the empirical laws. *Journal of Information Science*, *16*(1), (1990), 17

[6] Egghe, L. The power of power laws and an interpretation of Lotkaian informetric systems as self-similar fractals. *Journal of the American Society for Information Science and Technology*, *56*(7), (2005), 669-675.

[7] Egghe, L. L. Power Laws in the Information Production Process: Lotkaian informetrics. (Oxford, UK. 2005), Elsevier

[8] Goldstein, M. L., Morris, S. A., & Yen, G. G. Problems with fitting to the power-law distribution. The European Physical Journal B-Condensed Matter, 41(2), (2004), 255-258.

[9] Horrigan, J. B. Home Broadband Adoption 2006. *Pew Internet & American Life Project* (Washington, DC. 2006)

[10] Huber, J. C. A new model that generates Lotka's law. *Journal of the American Society for Information Science and Technology*, *53*(3), (2002), 209-219.

[11] Huber, J. C., & Wagner-Döbler, R. Scientific production: A statistical analysis of authors in mathematical logic. *Scientometrics*, *50*(2), (2001), 323-337

[12] Karp, S. The User-Generated Content Myth - Publishing 2.0. http://publishing2.com/2007/10/26/the-user-generated-content-myth/. Retrieved October 2, 2007

[13] Koskinen, I. User-generated content in mobile multimedia: empirical evidence from user studies. Proceedings of International Multimedia and Expo, (2003)

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

[15] Massey Jr, F. J. The Kolmogorov-Smirnov Test for Goodness of Fit. *Journal of the American Statistical Association*, *46*(253), (1951), 68-78.

[16] Miller, P. Web 2.0: Building the New Library. *Ariadne*, *45 (*(2005).

[17] Newman, M. E. J. Power laws, Pareto distributions and Zipf's law. Contemporary Physics, 46(5), (2005)

[18] O'Reilly, T. What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. O'Reilly Media, Inc. (2005) http://www. oreillynet. com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20. html

[19] OECD. Participative Web and User-Created Content: Web 2.0, Wikis and Social Networking Edition complete. *OCDE Information Sciences and Technologies* (October 2007). http://www.oecd.org/document/40/0,3343,en_2649_34223_39428648_1_1_1_1,00.html

[20] Olson Jr, M. The Principle of" Fiscal Equivalence": The Division of Responsibilities among Different Levels of Government. The American Economic Review, 59(2), (1969), 479-487.

[21] Pachet, F. (2005). Knowledge Management and Musical Metadata. *Encyclopedia of Knowledge Management*. Idea Group. (2005)

[22] Nielsen, J. Participation Inequality: Lurkers vs. Contributors in Internet Communities. http://www.useit.com/alertbox/participation_inequality.html. Retrieved on October 2, 2007.

[23] Powazek, D. Death to User-Generated Content. http://www.powazek.com/2006/04/000576.html Retrieved on October 2, 2007.

[24] Tague-Sutcliffe, J. An Introduction to Informetrics. *Information Processing and Management*, *28*(1), (1992), 1-4.

[25] Taleb, N. The Black Swan: The Impact of the Highly Improbable. (2007). Random House.

[26] Vuong, Q. H. Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. Econometrica, 57(2), (1989), 307-333.

[27] Whittaker, S., Terveen, L., Hill, W., & Cherny, L. The dynamics of mass interaction. *Proceedings of the 1998 ACM conference on Computer supported cooperative work*, (1998), 257-264.