Productivity in the Internet Mailing Lists: A Bibliometric Analysis

Victor Kuperman
Fourth Dimension Software, Redwood City, CA. E-mail: vkuperman@yahoo.com

The author examines patterns of productivity in the Internet mailing lists, also known as discussion lists or discussion groups. Datasets have been collected from electronic archives of two Internet mailing lists, the LINGUIST and the History of the English Language. Theoretical models widely used in informetric research have been applied to fit the distribution of posted messages over the population of authors. The Generalized Inverse Poisson-Gaussian and Poisson-lognormal distributions show excellent results in both datasets, while Lotka and Yule–Simon distribution demonstrate poor-to-mediocre fits. In the mailing list where moderation and quality control are enforced to a higher degree, i.e., the LINGUIST, Lotka, and Yule–Simon distributions perform better. The findings can be plausibly explained by the lesser applicability of the success-breeds-success model to the information production in the electronic communication media, such as Internet mailing lists, where selectivity of publications is marginal or nonexistent. The hypothesis is preliminary, and needs to be validated against the larger variety of datasets. Characteristics of the quality control, competitiveness, and the reward structure in Internet mailing lists as compared to professional scholarly journals are discussed.

Introduction

Since Lotka (1926), scholarly journals and, to a lesser degree, monographs and patents have constituted the primary field for bibliometric research in scientific productivity. The means of electronic communication (such as Internet mailing lists, newsgroups, forums, and chats) has attracted less attention, despite their increasing popularity in the world of science. Because these channels serve the ever-growing scientific "e-community" as the efficient mechanism of information exchange, they call for examination on the side of informetrics (Bar-Ilan, 1997; Hernandez-Borges, Pareras, & Jimenez, 1997; Hernandez-Borges, Macias, & Torres, 1998; van Raan, 2001; Zelman & Leydesdorff, 2000).

In this article the productivity of science-oriented Internet mailing lists (henceforth, IMLs), also commonly known as discussion groups or discussion lists is studied.[1] Two datasets collected from electronic archives of the LINGUIST (2004a) mailing list and the History of the English Language (HEL-L; 2004a) mailing list demonstrate a highly skewed distribution of postings (i.e., published e-mail messages) over the population of posters (i.e., authors of messages). To test the goodness-of-fit of known informetric distributions and the adequacy of explanatory models, the pools for data collection have been selected to differ in the degree of quality control, with the HEL-L being a less-restrictive mailing list. It is argued that the event of posting a message in an IML is different from conventional scholarly publishing in several important aspects (Bar-Ilan, 1997; Matzat, 1998; van Raan, 2001):

- It has to satisfy a less-rigorous (or nonexistent) publication policy and quality control
- It is only marginally affected by the competition for publishing resources (i.e., space allocation in a journal or book, allowed size and frequency of publications, advertising costs, etc.)
- Its reward structure does not include formal recognition in the community of peers, and thus does not contribute directly to the poster's standing

These factors challenge the notion of "success" as an extraordinary achievement, which is central to a number of explanatory models of intellectual output, most notably, *the cumulative advantage*, or *success-breeds-success*, model associated with the Pareto-like, power-law family of bibliometric regularities (Günther, Levitin, Schapiro, & Wagner, 1996; Simon, 1957).

It is shown that the inverse power laws (i.e., Lotka's, Zipf–Mandelbrot, and Yule–Simon) offer poor-to-mediocre fits to the frequency distribution in datasets. The approximation is better, however, where the publication policy is more rigorous, and thus the event of "success" is more pronounced. At the same time, the Generalized Inverse Gaussian–Poisson

Received June 3, 2004; revised September 20, 2004; accepted November 1, 2004

 ● Published online 24 October 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.20252

[1]For the sake of clarity, no distinction between sciences and humanities, or scientists and scholars, is being made throughout the article.

distribution and Poisson-lognormal distribution, which have been traditionally accounted for by principles other than "success-breeds-success," present excellent fits to the frequency distribution in both data populations. The adequacy of the *cumulative advantage* models of knowledge production is revised in the context of IMLs, and suggestions for the future research are made. Difficulties of the data collection, inherent in any research in "electronic productivity" (Kot, Silverman, & Berg, 2003), restrict this preliminary analysis to only two datasets, therefore the conclusions will require validation against larger and more heterogeneous data pools.

Background and Literature Review

The productivity of authors posting in the Internet mailing lists is an instance of the Information Production Process (Egghe & Rousseau, 1990). Egghe (1990), and Egghe and Rousseau (1990) address the Pareto-like distributions (e.g., laws of Lotka, Zipf–Mandelbrot, Bradford, and Pareto) in terms of the duality formalism of the source-item relationship. Importantly, researchers analyze stability of Lotka's law under a number of factors, which allows for narrowing down the scope of variables that influence the productivity distribution. Thus, Bookstein (1977; 1990a; 1990b) and Egghe and Rousseau (1990) prove that the Lotka-type regularity remains invariant over time. Bookstein (1977, p. 262–263) also argues that Lotka's law, as the inverse power function, is "stable with regard to at least two forms of social change": the ability of society to "modify the caliber" of publishing scientists and its ability "to squeeze out of each individual the maximum amount of research." In particular, Lotka's law is insensitive to the number of active contributors, the rate of their arrival and their departure (Bookstein, 1979).

Several generative mechanisms have been proposed to account for and model the Lotka-type regularity. The most acknowledged are the success-breeds-success model, and the mathematically equivalent cumulative advantage model (Egghe & Rousseau, 1995; Günther et al., 1996; Koenig & Harrel, 1995; Price, 1963; Price, 1976; Simon, 1957; Tague, 1981). The influential Yule–Simon approach summarizes this model in two stipulations (Simon, 1957):

1. The probability w that the next paper is a paper by a given author who has already published i times is proportional to the number of authors that have contributed exactly i papers to the journal.
2. There is a constant probability, alpha, that the next author is an author who has not previously published in the journal.

From the two assumptions, Simon derives the classical formulation of the power law distribution, which—with modifications—gives rise to Lotka's law, Zipf, Yule, Yule–Simon, Waring, and Zipf–Mandelbrot distributions. Kot et al. (2003) use a "closely related" model developed in (Günther et al., 1996) to explain the conformity of biology newsgroups to Zipf's law. The concept of success has often been associated with such fundamental attributes of the scientific communication, as the quality certification, priority recognition, as well as the ownership reward (e.g. Price, 1963; Stephan, 1996; van Raan, 2001). The insightful review of causal models of knowledge production by Oluic-Vukovic (1997) argues that the success-breeds-success approach, in its generalized form, is capable of generating almost all classical frequency distributions. The notion of publication as success is also utilized in the statistics of exceedances that Huber (1998, 1999, 2002) and Narin (1994) advocate as an alternative to the Yule–Simon model. While using different probabilistic premises than in Simon (1957), Huber's model generates Lotka's law (Braun, Glänzel, & Schubert, 1990; Huber, 1998, 2002; Narin, 1994; Wagner-Döbler, 1995). Below we discuss how the notion of *success* applies to the case of Internet mailing lists with their criteria of acceptance and recognition of achievement. Several alternatives to the success-driven models have been suggested, as well. Stewart (1993, p. 245–246) comes up with the Poisson-lognormal model that relies on the law of proportionate effects where "the underlying propensity to publish is a multiplicative function of many independently distributed factors, such as intelligence, training, motivation, and available resources." Another compound Poisson distribution, the Generalized Inverse Gaussian–Poisson model, is defended by Sichel (1975, 1985, 1990).

Internet mailing lists have only recently become the object of informetric research. Zelman and Leydesdorff (2000) analyze self-organization of e-mail threads to facilitate the knowledge production in IMLs. Several papers (Hernandez-Borges et al., 1997, 1998; Hernandez-Borges et al., 1999) test the quality and reliability of medical IMLs in comparison to relevant professional journals. The science-oriented IMLs are shown to feature qualified authors and are viewed as a valuable complement to other academic sources, such as medical journals and conferences. Matzat (1998) elaborates on the reward structure of publishing in IMLs focusing on the trade-off between one's advantages from such publication and the price paid (i.e., the average time needed for such activity). To our knowledge, productivity patterns have never been studied systematically in the IMLs, yet highly relevant results have been obtained in other electronic communication systems. Bar-Ilan (1997) has first applied bibliometric laws to the USENET newsgroups. Bar-Ilan (1997) reports that "to some extent" Bradford's law holds true in the Usenet newsgroups. Rousseau's study of "sitations" (Rousseau, 1997) proves the applicability of power laws to the distribution of domain names and distribution of links between Web sites. A recent study of biology newsgroups (Kot et al., 2003) finds that submissions obey Zipf's law "as a function of the rank, by posting, of contributors," and offers a stochastic model to predict the diversity patterns in productivity. Current research in electronic communication serves as the methodological basis for the present analysis of knowledge production in Internet mailing lists.

DOI: 10.1002/asi

The Data

This study analyzes the population of electronic messages posted in IMLs. An IML, *Internet Mailing List*, is a list of e-mail addresses identified by a single name and a single e-mail address. When an e-mail message is sent to this generic e-mail address, it is forwarded to all addresses in the list, unless a censoring body restricts its posting. Only those messages posted and archived through IMLs are accounted for in this article. *Author* is here defined as a producer of the message. The mailing list with an overtly stated publication policy is defined as *moderated*. The *publication policy* in an IML is a set of regulations and limitations pertaining to the content and format of acceptable messages. The IMLs lacking such policy are here defined as *nonmoderated*.

Moderation diminishes the notorious Internet noise and enhances the quality level of the mailing list (Hernandez et al., 1997, 1999). Moreover, Hernandez et al. (1997, p. 1) defines the following degrees of control over postings:

> The lowest level of control is restricting postings to members of the list. . .The next level is to subject each posting to a process in which only messages that the manager of the mailing list approves are available by the members of the mailing list. The highest level of control includes both the posting approval and its edition as needed.

One of the research objectives has been to explore publishing patterns in IML populations where the degree of control and the resultant level of the Internet noise differ perceptibly. Within the LINGUIST list, only the *Discussion* category presents threads of scholarly communication in full and is open for the "on-list" original publication and commenting. Thus, the Discussion category has been designated as a source for the data collection. Informally speaking, selecting the Discussion category for data collection is similar to analyzing scholarly papers in journals, while ignoring advertisements, table of contents, calls for papers, job offers, and other materials that may appear in journals, too. It is acknowledged, however, that findings and conclusions of this paper will only relate to the Discussion category, rather than the entire LINGUIST list. The HEL-L applies no mechanism of selecting or categorizing the incoming messages: thus, the data collection is applied to the entire available electronic archive of this IML.

The electronic archives of two IMLs in the field of language studies (the LINGUIST and the History of the English Language List, HEL-L) have been accessed at http://www.linguistlist.org/issues/master.html and http://listserv.linguistlist.org/archives/hel-l.html. The electronic archives of the two IMLs are powered by the ListServe software and are searchable with the ListServe search engine. For both IMLs, messages are stored in weekly, monthly, or yearly subarchives. The two datasets represent complete electronic archives that were available at the time of data collection. It is unknown how many messages were rejected or modified by the moderators' committee of the LINGUIST list, or by the owner of the HEL-L during the observation period. Dierick (1992) claims that only the *random* sampling by *source* (that is, by author) gives reliable results, and the sample size should account for at least 10% of the complete dataset. The datasets in question satisfy Dierick's requirement, because no selection either by source, or by item, has been made in any of the two document populations.

Much as described in Kot et al. (2003), the data collection procedure involved ascertaining the identity of posters and hand tagging in both document populations. About 0.5% of posts in each list were excluded from the count; they represented internal documents published by mistake, rather than valid items. In less than 10% of messages where authorship was uncertain, identity marks have been looked for in the message body, e-mail address or signature. E-mailing contributors was sometimes necessary. This way of identifying authors heavily relies on a toilsome ad hoc process, which is geared up by the nonuniform presentation format of the messages in electronic archives (Kot et al., 2003). Zelman and Leydesdorff (2000) lament the lack of the "standardized form of archiving mailing list output" that makes research of IMLs particularly labor-intensive. The percentage of messages with uncertain author identity (8% in the LINGUIST list, 7% in the HEL-L) is comparable with the 5% of items "unknown authorship" in Pao's 1979 study and 12% in Kot et al.'s (2003). It is possible that even after the identification procedure some messages end up attributed to wrong authors. The distortion, if any, is likely to increase the total number of authors and reduce the number of high scorers (Huber, 1999).

The data analysis follows the validation procedure developed in Pao (1985, 1986) and Nicholls (1986, 1989). The method of the authors' count was normal. Values of high scorers were not excluded from calculations. A number of theoretical models widely employed in bibliometrics were tested for goodness-of-fit against the resulting frequency distributions: Lotka, Yule, Yule–Simon, Mandelbrot–Zipf, generalized Waring, negative binomial, lognormal, logarithmic series, generalized Poisson, Generalized Inverse Gaussian–Poisson (GIGP), Poisson-lognormal (PLN) and geometric. As suggested in Ajiferuke, Wolfram, and Xie, (2004), with the models that have a natural origin of zero (e.g., negative binomial) the zero-truncated version was tested, where each outcome of $f(x)$ for x greater than zero is divided by $1 - f(0)$. The theoretical functions that provide best fits for the empirical data are Lotka, Yule–Simon, PLN, and GIGP. Parameters of Lotka's law were estimated by the maximum likelihood method (Rousseau & Rousseau, 2000); same method was applied for parameter estimation in PLN with the help of software kindly provided by Professor John A. Stewart (cf. Stewart, 1993). Parameters of Yule–Simon and GIGP models were estimated by the least-squares method (Baayen, 2001). The chi-square goodness-of-fit test has been chosen; it is more reliable for discrete data distributions than the Kolmogorov– Smirnov test (Sichel, 1990). Data bins were collapsed to maintain at least the count of four in each bin.

DOI: 10.1002/asi

The functional forms of distributions reported in this article are as follows:

1. Lotka's law

$$p(x) = \alpha x^{-\beta},$$

2. Generalized Inverse Gaussian–Poisson distribution (GIGP), adopted in this form from (Baayen, 2001, p. 89):

$$p(x) = \frac{(2/\alpha\beta)^{\gamma}}{2K_{\gamma}(\beta)} x^{\gamma-1} \quad \exp\left\{-\frac{x}{\alpha} - \frac{\beta^2\alpha}{4x}\right\},$$

where $-\infty < \gamma < \infty$, $\beta \geq 0$, $\alpha \geq 0$ for $x = 1, 2, \ldots$ and $K_a\{z\}$ is the modified Bessel function of the second kind of order a and argument z.

3. Poisson-lognormal distribution:

$$p(x) = \frac{1}{\alpha\sqrt{2\pi}} \frac{1}{x!} \int_0^{\infty} e^{-\delta}\, \delta^{x-1} \quad \exp\left\{-\frac{(\ln\delta - \beta)^2}{2\alpha^2}\right\} d\delta,$$

where β is the mean, and α is the standard deviation of the normally distribution of logged δ.

4. Yule-Simon distribution:

$$p(x) = \frac{(\alpha + 1)\Gamma(x)\Gamma(\alpha + 1)}{\Gamma(\alpha + x + 1)},$$

where $\Gamma(x)$ denotes the Gamma function, and $\alpha > 0$ for $x = 1, 2, 3 \ldots$

Bibliometric indicators of the list functioning have been assessed, e.g., the date of inception, the duration, as well as the number of authors and subscribers. To enable the discussion of productivity patterns in IMLs, publication policies, and authors' instructions have been obtained from the owners of the LINGUIST list and the HEL-L.

Results

Comparative characteristics of the functioning and usage of the HEL-L and Discussion category of LINGUIST are presented in Table 1.[2] All numbers in Table 1 and in the further discussion are valid as of October 10, 2001.

The activity and participation rates in the Discussion category of the LINGUIST list reveal that this information resource is more popular and utilized than the HEL-L. Besides having a longer history of activity, the Discussion category features a larger average monthly volume of messages than the HEL-L (46.4 vs. 36.4). While the number of subscribers (11477 in Discussion vs. 253 in HEL-L) reflects the relative exposure of the list to the respective scientific community, it is the number of contributing posters that affects the study of productivity more directly. The number of messages per author is higher in the HEL-L than in the Discussion (6.03 vs. 3.62), yet the Discussion category leans on a much wider base of authors than the HEL-L does (1385 vs. 501). The number of subscribers who posted in HEL-L (501) is twice as large as the number of current subscribers (253); regrettably, no information is available on the total of the HEL-L subscribers throughout its lifetime.

TABLE 1. Descriptions of document populations.

	Discussion section of LINGUIST	HEL-L
Date of inception	Dec. 1990	Jan. 1994
Duration of covered period	108 months	83 months
Total number of emails	5016	3023
Number of subscribers	11477	253
Number of authors	1385	501
Average messages per author	3.62	6.03
Average messages per month	46.4	36.4
Degree of quality control	Highest	Lowest

Productivity distribution, i.e., distribution of postings over the population of posters, has a highly skewed shape in both populations of messages. The observed data from the Discussion category of the LINGUIST list, as well as theoretical distributions, are reported in Table 2.

Figure 1 plots the empirical data and theoretical models on the log-linear scale.

The empirical dataset from the HEL-L (February 1994–January 2001) and data from theoretical approximations are presented in Table 3.

Figure 2 plots the observed distribution and theoretical models on the log-linear scale.

To submit a message for publication in either mailing list, the poster needs a subscription to the list. Whether the subscriber's message is accepted for posting, is further governed by the list's publication policy. Excerpted below are the publication guidelines provided by the owners of the LINGUIST list in "General Policies" (LINGUIST, 2004b):

> All postings are subject to editor approval. Nothing is mailed to the subscriber list—or appears on the Website—until an editor has reviewed and approved the submission . . . We occasionally do light editing of messages to bring them into conformity with our policies. In general, we post only messages with substantial linguistic content or with content which will be of wide interest within the discipline . . . In general, we do not post commercial advertisements.

Additional regulations for the Discussion topic (LINGUIST, 2004c) state "readers may submit a question, hypothesis, or issue for debate and discussion by fellow linguists . . . Discussions may be ended at the discretion of the moderators; a "last call" notice is always given." Another unmentioned restraint enforced in the Discussion category is the originality of the suggested discussion subject. The issue raised by a subscriber is reviewed by moderators, and may

[2]Data in Table 1 relate to the entire LINGUIST list and not only to the Discussion category in question. No information is available as to the number of subscribers that read or write to the Discussion category proper. In addition to the numbers in this column, the LINGUIST counts another 77 "concealed" subscribers, the HEL-L—another 18 "digested" subscribers.

DOI: 10.1002/asi

TABLE 2. Distribution of messages in Discussion category of the LINGUIST list.

Number of messages	Observed frequency	Expected Lotka	Expected GIGP	Expected PLN	Expected Yule–Simon
1	753	801.9150	753.8112	749.83	748.459
2	243	211.9097	240.8193	236.88	200.1775
3	104	97.2872	110.1208	112.78	98.5535
4	56	55.9981	62.716	65.38	60.2424
5	49	36.4843	40.5984	42.46	41.1952
6	29	25.7086	28.4908	29.68	30.1908
7	22	19.1223	21.1246	21.85	23.1965
8	13	14.7977	16.2983	16.71	18.446
9	13	11.8027	12.9585	13.16	15.0577
10	10	9.6411	10.5482	10.62	12.5482
11	5	8.0289	8.75	8.73	10.6332
12	12	6.7936	7.3719	7.29	9.1359
13	5	5.8258	6.2918	6.17	7.9412
14	5	5.0532	5.4293	5.28	6.9716
15	5	4.4262	4.7295	4.57	6.173
16	5	3.9104	4.1538	3.99	5.507
17	6	3.4807	3.6744	3.51	4.9453
18–19	9	5.9303	6.1996	5.87	8.5231
20–24	11	10.7326	10.9494	10.21	15.7516
25–29	9	7.2149	7.0658	6.5	10.8363
30–49	12	14.6721	12.9728	11.94	22.6601
50–130	9	14.1975	8.2981	8.84	22.5486
Sample Mean	3.6217				
Sample *SD*	7.9548				
α	—	0.579	0.0145	2.3431	0.5496
β	—	1.920	0.0732	−3.1074	—
γ	—	—	−0.7613	—	—
χ^2	—	25.3877	11.9588	13.7174	34.7325
dF	—	20	20	20	20
$P(\chi^2)$	—	0.1870	0.9175	0.8445	0.0216

be withdrawn at their discretion if deemed repetitive.[3] There is no data available as of how many messages are refused publication in the LINGUIST, or are revised.

The HEL-L has only an embrionic publication policy (History of the English Language, 2004b). As stated in a personal letter from the HEL-L administrator (April 14, 2001), "One key difference between the two lists [the LINGUIST and the HEL-L] is that HEL-L is not moderated . . . Messages aren't vetted prior to distribution."

In other words, any message sent to the HEL-L by the list subscriber will automatically be posted and archived.

Discussion

Datasets examined demonstrate a highly skewed frequency distribution characteristic of the information production processes. Generalized Inverse Gaussian–Poisson and Poisson-lognormal models demonstrate excellent fits in both datasets (92% and 84% in LINGUIST vs. 98% and 96% in HEL-L, respectively). The Pareto-like models, i.e., Lotka and Yule–Simon distributions, generally fare less well, and provide better results for the data from LINGUIST. Thus, Lotka's law is an adequate approximation (19%) to the

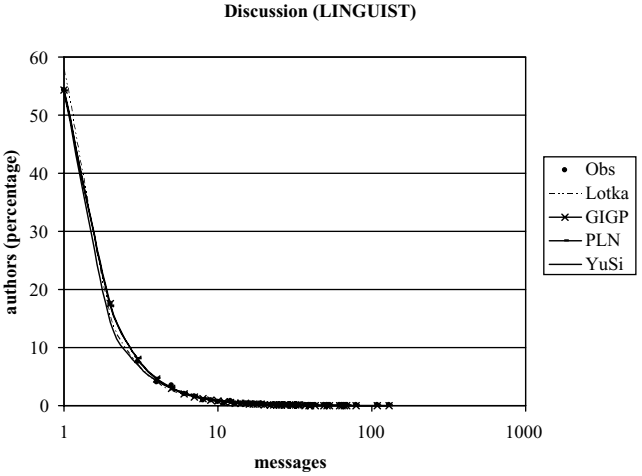

FIG. 1. Productivity distribution in Discussion category of the LINGUIST list.

[3] Consider the excerpt of the response from the LINGUIST moderator to our query (September 3, 2001):

Dear Victor Kuperman,

Thank you for your message. I am reluctant to post it, however, because we just had a very lively discussion on this topic in May. It begins in Linguist Issue 12.1462 and continues for quite a few issues after that . . . If this turns out to be less than satisfactory, please let me know . . .

LINGUIST managing editor

DOI: 10.1002/asi

TABLE 3. Distribution of messages in the HEL-L.

Number of messages	Observed frequency	Expected Lotka	Expected GIGP	Expected PLN	Expected Yule–Simon
1	192	243.8868	192.0618	195.98	192.6266
2	96	75.0650	95.9357	89.45	72.0869
3	55	37.6775	52.6664	50.63	40.774
4	30	23.1040	32.378	32.45	27.1288
5	17	15.8103	21.7602	22.51	19.7062
6	17	11.5966	15.6084	16.5	15.1303
7	13	8.9232	11.745	12.58	12.0709
8	12	7.1111	9.1632	9.89	9.9055
9	10	5.8207	7.3524	7.97	8.3063
10	5	4.8662	6.0327	6.54	7.0858
11–12	10	7.7076	9.3163	10.08	11.4942
13–19	16	15.8929	18.0694	19.2	24.7504
20–29	10	10.9542	11.2325	11.32	17.7812
30–49	11	9.9067	8.819	8.22	16.3015
50–171	7	13.1829	7.9387	6.59	20.5787
Sample Mean	6.0339				
Sample *SD*	15.7642				
α	—	0.4868	0.0519	1.9143	0.7937
β	—	1.7000	0.1087	−0.8396	—
γ	—	—	−0.8069	—	—
χ^2	—	41.5215	4.6588	5.4840	32.6553
d*F*	—	13	13	13	13
P(χ^2)	—	0.00008	0.9820	0.9629	0.0019

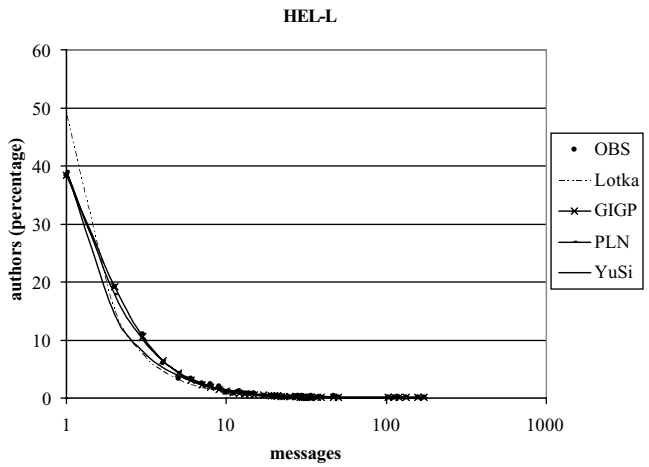

FIG. 2. Productivity distribution in the HEL-L.

LINGUIST data, yet it fails to reasonably fit the HEL-L data. The Yule–Simon model offers poor fits of 2% versus 0.2% in the LINGUIST and HEL-L datasets, respectively.

These findings can be plausibly explained by the unsuitability of the cumulative advantage model (argued to be the causal process for Pareto-like distributions) to the conditions of information production in IMLs. The key notion of this model is the *success*, or the event of surpassing the quality threshold and producing an extraordinary achievement, such as a scientific paper, monograph, or patent (cf. Huber, 2002; Koenig & Harrel, 1995; Price, 1963, 1976). Furthermore, under the Yule-Simon approach (Simon, 1957) accumulation of previous achievements makes this event proportionally more probable. The explanation of Price's urn simulation (Koenig & Harell, 1995, p. 387) emphasizes this point:

> When you draw, if the failure ball is drawn (F), you are out of the game, but if you draw a success ball (S) you add another paper to your credit, you get to continue in the game, and you put an additional success ball (S) into the urn for future drawings; so that odds get better with each succeeding round you play.

This assumption is particularly vulnerable in IMLs because more often than not, a message is automatically published in the mailing list as long as its contributor is a subscriber. Thus, the probability of posting a new message in such an IML may be equal for novices and veteran list participants, which is the case in the HEL-L. The probability of one's publication is claimed in a Yule–Simon model to depend on one's publication history. However, the reliance on the author's identity, reputation, or history of prior publications, is impeded by the uncertainty of authorship. Internet mailing lists do not obligate the poster to publicize his or her identity. The instability of e-identity reduces one's ability to enjoy an increasing scientific reputation in a mailing list, and does not allow readers and moderators to acknowledge one's reputation as an accrual of achievements. Likewise, the priority recognition, as well as the ownership reward cannot be procured unambiguously (van Raan, 2001). Ascertaining the

DOI: 10.1002/asi

authorship during the data collection reveals how many authors may remain unrevealed, unless a special effort is made to identify them (12% in the data sample of Kot et al., 2003). Generative mechanisms of the cumulative advantage type are likely to fail in a publishing environment that ignores the history of achievements when permitting a new achievement. It is recognized, however, that a more extensive quantitative examination of IMLs bibliometric behavior in various datasets is required to further substantiate this preliminary hypothesis.

To reiterate, the concept of *success* presupposes an existence of selectivity (most frequently, the quality control) and competitiveness in the publishing medium. As mentioned in Bar-Ilan's study of newsgroups (1997, p. 46), selectivity and competitiveness constitute the two basic dissimilarities between the electronic media communication, such as newsgroups or IMLs, and the scientific literature. While there is a strict mechanism of refereeing in journals, patent offices, or publishing houses, IMLs are rarely moderated and often publish any posting submitted by subscribers. Also, the competition for publishing in either the LINGUIST or the HEL-L is virtually nullified due to the unlimited frequency or number of simultaneous publications, practically unlimited space allocated to individual publications, marginality of advertising and publishing costs, as well as the availability of the storage space and archiving facilities. In other words, the specific character of the IML functioning implies that (a) models less sensitive to the probability of acceptance of individual publications should generate better results in IMLs, than the success-breeds-success model, and (b) the success-breeds-success model should have a more adequate goodness-of-fit in IMLs where a higher quality threshold or competitiveness is observed.

Statement (a) is supported by the fact that compound Poisson distributions, such as GIGP and PLN, demonstrate much better results than Lotka and Yule–Simon distributions in both analyzed datasets. Neither the Poisson-lognormal model, nor the Generalized Inverse Gaussian–Poisson model takes into account the probability of acceptance of an individual contribution. The common assumption for the two models is that the latent propensities for scientists to publish Rt papers in time t follow the Poisson distribution (Allison, 1980). The latent propensity indicates the potential of a scientist to create a publishable informational item, which has not yet come to realization: such as, the zero productivity of a fresh PhD holder. The models differ in hypothesizing how the individual rates of publishing R are distributed: following the lognormal distribution (Stewart, 1993) or the Generalized Inverse Gaussian law (Sichel, 1985). Explaining the possible causal mechanism, the PLN refers to a multifarious variety of factors (intelligence, training, motivation and available resources) interacting in a way that "a weakness in any one factor reduces the effects of all the other factors" (Stewart, 1993, p. 246). The model that generates GIGP is less articulated, yet it clearly refutes the notion of success, since the zero score of publications is considered along with other observable scores. Thus, their results being superior to those shown by Pareto-like models is an expected finding in the datasets taken from IMLs.

Statement (b) is also asserted by the empirical data of this study. The Discussion category of LINGUIST and the HEL-L both fail to live up to requirements imposed on scientific contributions by editorial boards, publishing houses, or patent agencies. Still, it is obvious from the publication policy of the lists that a contribution to the moderated Discussion category needs to meet several restrictive conditions, which are not enforced in the nonmoderated HEL-L. To apply the classification from (Hernandez et al., 1997), the LINGUIST list makes a claim of the highest level of control over postings where the editorial approval and occasional editing are mandatory, while the HEL-L maintains the lowest level of control only requiring an author to subscribe. Thus, the LINGUIST list comes closer to the quality standards and refereeing policies accepted in professional literature. This higher threshold of quality required from a publishable posting, may account for better fits of Lotka's law and the Yule–Simon model.

Oluic-Vukovic (1997, p. 839–840) states that "there is no unambiguous solution to the important problem of determining the underlying probability mechanism producing the observed regularities, since several theoretical hypotheses . . . could be advanced, each leading under fairly general conditions, to the equivalent description." Further research in causal mechanisms of output in mailing lists should continue testing the correlation between the selectivity of an IML and the goodness-of-fit of informetric distributions. The method of quantifying the degree of quality control in a given mailing list needs elaboration. The ratio of rejected contributions to the number of published postings, or the ratio of "noise" messages (e.g. spam, commercial advertisement, subscribe/unsubscribe requests, or holiday greetings) to the total of messages may prove useful to this end.

Conclusion

This, to our knowledge, is the first approach to theoretical models and casual mechanisms of productivity in Internet mailing lists. A number of bibliometric distributions have been applied to two samples from IMLs (the Discussion category of the LINGUIST list and the HEL-L). The lists have been deliberately chosen to represent different levels of quality control, with the HEL-L being less restrictive. Generalized Inverse Gaussian–Poisson and Poisson-lognormal were found to provide excellent fits to both datasets, while Lotka's law and Yule–Simon distribution showed poor-to-mediocre goodness-of-fit. These results confirmed our tentative hypothesis that with the barrier of competitive acceptance lowered or nonexistent in IMLs, the event of publication ceases to be a distinctive achievement such as producing a scholarly paper, or a patented invention. Therefore, distributions traditionally explained by the cumulative advantage principle or the success-breeds-success model (i.e., Lotka and Yule–Simon) produce overall much worse results than

DOI: 10.1002/asi

distributions generated by other models. At the same time, Lotka's and Yule–Simon's laws perform demonstrably better in the sample where selectivity is maintained in a stricter way. Internet mailing lists differ drastically from non-electronic means of scholarly publishing in terms of competitiveness, selectivity, and the reward structure. Comparative study of research variables, such as the quality control, length, and intensity of functioning, and rate of participation, needs to be carried out to reach definitive conclusions regarding the bibliometric patterns of intellectual output in IMLs.

Acknowledgments

This paper is based on the thesis submitted for the partial fulfillment of the MLS degree requirements at the Graduate School of Library, Archive and Information Studies at the Hebrew University of Jerusalem, Israel. The author wishes to thank his supervisors, Professor Peritz and Dr. Bar-Ilan, for valuable guidance and help. Thanks are due to two anonymous reviewers of the *JASIST* for insightful comments and suggestions, Professor Ajiferuke of the University of Western Ontario, Canada, for consultations, and Professor Stewart of the Hartford University, USA, for the provision of the statistical software.

References

Ajiferuke, I., Wolfram, D., & Xie, H. (2004, June). Brief communication: Modelling website visitation and resource usage characteristics by IP address data. Paper presented at the Proceedings of the Canadian Association for Information Science 2004, Annual Conference, Winnipeg, Canada. Retrieved August 30, 2004, from http://www.cais-csi.ca/proceedings/2004/ajiferuke_2004.pdf

Allison, P.D. (1980). Inequality and scientific productivity. Social Studies of Science, 10, 163–179.

Baayen, R.H. (2001). Word frequency distributions. Dordrecht: Kluwer Academic Publishers.

Bar-Ilan, J. (1997). The 'mad cow disease', USENET newsgroups and bibliometric laws. Scientometrics, 39(1), 29–55.

Bookstein, A. (1977). Patterns of scientific productivity and social change: A discussion of Lotka's law and bibliometric symmetry. Journal of American Society for Information Science, 28(4), 206–210.

Bookstein, A. (1979). Explanations of the bibliometric laws. Collection Management, 3(2), 151–162.

Bookstein, A. (1990a). Informetric distributions, part I: Unified overview. Journal of the American Society for Information Science, 41, 368–375.

Bookstein, A. (1990b). Informetric distributions, part II: Resilience to ambiguity. Journal of the American Society for Information Science, 41, 376–386.

Braun, T., Glänzel, W., & Schubert, A. (1990). Publication productivity: From frequency distributions to scientometric indicators. Journal of Information Science, 16, 1–8.

Dierick, J.C.J. (1992). Determining the Lotka parameters by sampling. Scientometrics, 25(1), 115–148.

Egghe, L. (1990). The duality of informetric systems with applications to empirical laws. Journal of Information Science, 16, 17–27.

Egghe, L., & Rousseau, R. (1990). Introduction to informetrics. Amsterdam: Elsevier.

Egghe, L., & Rousseau, R. (1995). Generalized success-breeds-success principle leading to time-dependent informetric distributions. Journal of American Society for Information Science, 46, 426–445.

Günther, R., Levitin, L., Schapiro, B., & Wagner, P. (1996). Zipf's law and the effect of ranking on probability distributions. International Journal of Theoretical Physics, 35, 395–417.

History of the English Language (HEL-L). (2004a). Archives of the history of English language list. Retrieved May 19, 2004, from http://listserv.linguistlist.org/archives/hel-l.html

History of the English Language (HEL-L). (2004b). About HEL-L. Retrieved May 19, 2004, from http://wiz.cath.vt.edu/mailman/listinfo/hel-l

Hernandez-Borges, A.A., Pareras, L.G., & Jimenez, A. (1997). Comparative analysis of pediatric mailing lists on the internet. Pediatrics, 100(2), e8.

Hernandez-Borges, A.A., Macias, P., & Torres, A. (1998). Are medical mailing lists reliable sources of professional advice? Medical Informatics and Internet in Medicine, 23(3), 231–236.

Hernandez-Borges, A.A., Macias-Cervi, P., Gaspar-Guardado, M.A., Torres-Alvarez de Arcaya, M.L., Ruiz-Rabaza, A., & Ormazabal-Ramos, C. (1999). Assessing the relative quality of anesthesiology and critical care medicine internet mailing lists. Anesthesiology and Analgesia, 89(2), 520–525.

Huber, J.C. (1998). The underlying process generating Lotka's law and the statistics of exceedances. Information Processing & Management, 34(4), 471–487.

Huber, J.C. (1999). Cumulative advantage and success-breeds-success: The value of time pattern analysis. Journal of the American Society for Information Science, 49(5), 471–476.

Huber, J.C. (2002). A new model that generates Lotka's law. Journal of American Society for Information Science, 53(3), 209–219.

Koenig, M., & Harrell, T. (1995). Brief communication: Lotka's law, Price's urn, and electronic publishing. Journal of American Society for Information Science, 46(5), 386–388.

Kot, M., Silverman, E., & Berg, C.A. (2003). Zipf's law and the diversity of biology newsgroups. Scientometrics, 56(2), 247–257.

LINGUIST. (2004a). Archives of the LINGUIST list. Retrieved May 19, 2004, from http://www.linguistlist.org/issues/master.html

LINGUIST. (2004b). General policies of the LINGUIST list. Retrieved May 19, 2004, from http://linguistlist.org/LL/posting-help2.html#General

LINGUIST. (2004c). Discussion topic of the LINGUIST list. Retrieved May 19, 2004, from http://linguistlist.org/LL/posting-help2.html#Disc

Lotka, A.J. (1926). The frequency distribution of scientific productivity. Journal of the Washington Academy of Sciences, 16, 317–323.

Matzat, U. (1998, March). Informal academic communication and the use of internet discussion groups by scientists. Paper presented at IRISS '98, Bristol, England. Retrieved May 19, 2004, from http://www.sosig.ac.uk/iriss/abstracts/iriss19.htm

Narin, F. (1994). Patent bibliometrics. Scientometrics, 30(1), 147–155.

Nicholls, P.T. (1986). Empirical validation of Lotka's law. Information Processing & Management, 22(5), 417–419.

Nicholls, P.T. (1989). Bibliometric modelling processes and the empirical validity of Lotka's law. Journal of American Society for Information Science, 40(6), 379–385.

Oluic-Vukovic, V. (1997). Bradford's distribution: From the classical bibliometric 'law' to the more general stochastic models. Journal of American Society for Information Science, 48(9), 833–842.

Pao, M.L. (1979). Bibliometrics and computational musicology. Collection Management, 3(1), 97–109.

Pao, M.L. (1985). Lotka's law: A testing procedure. Information Processing & Management, 21(4), 305–320.

Pao, M.L. (1986). An empirical examination of Lotka's law. Journal of American Society for Information Science, 37(1), 26–33.

Price, D.D.S. (1963). Little science, big science. New York: Columbia University Press.

Price, D.D.S. (1976). A general theory of bibliometric and other cumulative advantage processes. Journal of American Society for Information Science, 27(5), 292–306.

Rousseau, R. (1997). Sitations: An exploratory study. Cybermetrics, 1(1). Retrieved May 19, 2004, from http://www.cindoc.csic.es/cybermetrics/articles/v1i1p1.html

DOI: 10.1002/asi

Rousseau, B., & Rousseau, R. (2000), LOTKA: A program to fit a power law distribution to observed frequency data. Cybermetrics, 4(1). Retrieved May 19, 2004, from http://www.cindoc.csic.es/cybermetrics/articles/v4i1p4.html

Sichel, H.S. (1975). On a distribution law for word frequencies. Journal of the American Statistical Association, 70, 542–547.

Sichel, H.S. (1985). A bibliometric distribution which really works. Journal of American Society for Information Science, 36(5), 314–321.

Sichel, H.S. (1990). Anatomy of the generalized inverse Poisson–Gaussian distribution with special applications to bibliometric studies. Information Processing & Management, 28(1), 5–12.

Simon, H.A. (1957). Models of man, social and rational. New York: Wiley.

Stephan, P.E. (1996). Economics of science. Journal of Economic Literature, 34(3), 1199–1235.

Stewart, J.A. (1993). The Poisson-lognormal model for bibliometric/scientometric distributions. Information Processing & Management, 30(2), 239–251.

Tague, J. (1981). Success-breeds-success phenomenon and bibliometric processes. Journal of American Society for Information Science, 32, 280–286.

van Raan, A.F.J. (2001). Bibliometrics and internet: Some observations and expectations. Scientometrics, 50(1), 59–63.

Wagner-Döbler, R. (1995). Where has the cumulative advantage gone? Some observations about the frequency distribution of scientific productivity, of duration of scientific participation, and of speed of publication. Scientometrics, 32, 123–132.

Zelman, A., & Leydesdorff, L. (2000). Threaded e-mail messages in self-organization and science & technology studies oriented mailing lists. Scientometrics, 48(3), 361–380.

DOI: 10.1002/asi