Jointly published by Akadémiai Kiadó, Budapest
and Kluwer Academic Publishers, Dordrecht

Scientometrics,
Vol. 56, No. 2 (2003) 247–257

Zipf's law and the diversity of biology newsgroups

MARK KOT,[1] EMILY SILVERMAN,[2] CELESTE A. BERG[3]

[1]*Department of Applied Mathematics, University of Washington, Seattle, WA (USA)*
[2]*School of Natural Resources and Environment, University of Michigan, Ann Arbor, MI (USA)*
[3]*Department of Genome Sciences, University of Washington, Seattle,WA (USA)*

Usenet newsgroups provide a popular means of scientific communication. We demonstrate striking order in the diversity of biology newsgroups: Submissions to newsgroups obey a form of Zipf's law, a simple power law for the frequency of posts as a function of the rank, by posting, of contributors. We show that a simple stochastic process, due to *Günther* et al. (1992, 1996), *Levitin* and *Schapiro* (1993), and *Schapiro* (1994), accounts for this pattern and reproduces many of the properties of newsgroups. This model successfully predicts the relative contribution from each poster in terms of the size, the number of posters and total posts, of the newsgroup.

Introduction

The American philologist George Kingsley *Zipf* (1935, 1949) became famous for observing that word frequencies, in a corpus, fall off inversely with word-frequency rank according to a simple power law. Although Zipf's law has lost much of its luster as a "deep law of natural language" (*Miller* et al., 1957; *Li*, 1992), it has reappeared as a frequency–rank relationship in other, more profound, contexts. Examples include page hits observed at web sites (*Huberman* et al., 1998), the sizes of cities (*Zipf*, 1949; *Marsili* and *Zhang*, 1998), and the annual incomes of companies (*Okuyama* et al., 1999).

Zipf's law also has a place in ecology (*Frontier*, 1985) as one of many theoretical species-abundance distributions for communities of plants and animals (*Magurran*, 1988; *Tokeshi*, 1993). Here, Zipf's law is thought to arise from a successional process in which later colonizers are rarer than earlier species (*Frontier*, 1985; *Magurran*, 1988). Unfortunately, this law fails, like many other theoretical distributions, to consistently describe the diversity of real biological communities (e.g., *Wilson* et al., 1996). In contrast, we will show that Zipf's law provides an excellent description of the diversity of many communities of biologists.

Received September 14, 2002.

Address for correspondence:
MARK KOT
Department of Applied Mathematics, Box 352420,
University of Washington, Seattle, WA 98195-2420, USA
E-mail: kot@amath.washington.edu

0138–9130/2003/US $ 20.00

In this paper, we undertake a systematic study of biology newsgroups on the Usenet. The Usenet is a global system of discussion forums or "newsgroups" arranged hierarchically by subject. Individuals post messages or articles to these newsgroups from their computers. Each message contains a header that automatically records information such as the identity of the author and the date and time of posting. Communication is asynchronous (*Osborne*, 1998).

The Usenet has grown from a system of two computers and 15 newsgroups in 1979 (*Hauben* and *Hauben*, 1997) to one that carries gigabytes of messages daily for over 79,000 newsgroups (*Smith*, 1999). Many newsgroups are meant for the use of scientists. For example, the BIOSCI/bionet newsgroups are intended "to promote communication between professionals in the biological sciences" (http: //www.bio.net/docs/biosci.FAQ.html). In spite of the popularity of newsgroups, few papers have examined the role of newsgroups in scientific communication (but see *Bar-Ilan*, 1997).

The paucity of studies of newsgroups may stem from a perception that newsgroups are complicated and that the Usenet is a social institution that "permanently teeters on the brink of chaos" (*Smith*, 1999). In this paper, we suggest that the structure of newsgroups is simpler than it appears. We present data that show that posts to newsgroups obey Zipf's law and that the observed patterns can be explained by a simple stochastic model.

The data

At the start of 2001, we examined 107 BIOSCI/bionet newsgroups located at the newsgroup archive ftp: //ftp.bio.net/pub/BIOSCI/ARCHIVE. We eliminated six of the newsgroups from further consideration for the following reasons: Five of the six newsgroups were not open. The newsgroup bionet.journals.contents distributes tables of contents and is "not for postings by readers." The newsgroup bionet.sci-resources "is used solely to distribute funding agency announcements" and "is not to be used for postings by readers." The three newsgroups bionet.prof-society.afcr, bionet.prof-society.aibs, and bionet.prof-society.biophysics are forums for society announcements with a restricted number of posters. The sixth newsgroup, bionet.molbio.genbank, is an open newsgroup, but read permission for most files was turned off at the time of our study, limiting access to much of the data.

To study diversity within newsgroups, we downloaded and analyzed the contents of the remaining 101 newsgroups. These newsgroups varied from a two-month-old

newsgroup with 5 posts to a 101-month-old newsgroup with 87,636 posts. The distribution of newsgroup sizes was positively skewed, with a median of 1777 posts and an arithmetic mean of 5257.42 posts.

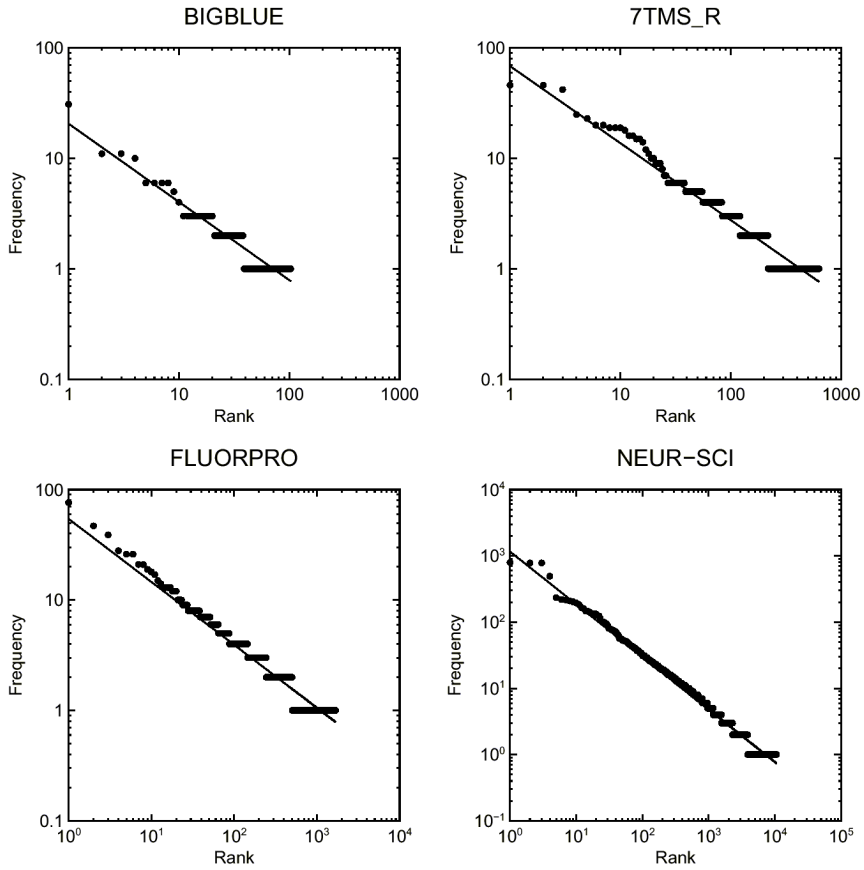

Figure 1. Log-log plots of frequency, as a function of rank, for 4 newsgroups: bigblue/prototype (BIGBLUE), 227 posts from 103 contributors over 51 months; bionet.molbio.proteins.7tms_r (7TMS_R), 1455 posts from 632 contributors over 67 months; bionet.molbio.proteins.fluorescent (FLUORPRO), 3114 posts from 1679 contributors over 69 months; and bionet.neuroscience (NEUR-SCI), 34,517 posts from 10,584 contributors over 101 months. The plots reveal a linear relationship between the logarithms of frequency and rank. We plot a regression line in each case for comparison. These lines have slopes -0.71, -0.70, -0.57, and -0.79

For each newsgroup, we tagged the source of each message with a lower-case version of the user ID or, lacking a clear user ID (1.6% of all cases), with a lower-case version of the entire "From: " line. We then ranked the contributors to each newsgroup by the number of submissions to that newsgroup and examined the frequency of posting by each contributor as a function of his or her rank.

These and subsequent analyses were performed using programs written in Icon (http: //www.cs.arizona.edu/icon), a high-level imperative programming language well-suited for text analysis. The programs are available from the authors on request.

Virtually all newsgroups showed a clear, linear relationship between the logarithm of frequency and the logarithm of rank. Figure 1 shows log-log plots of frequency as a function of rank for 4 representative newsgroups of differing sizes. Surprisingly, these newsgroups satisfy a form of Zipf's law,

$$f_R = \frac{\beta}{R^\gamma}, \tag{1}$$

for the frequency f_R of contributions as a function of the rank R of the poster. The exponent γ is the negative of the slope of the log-log plot of rank and frequency; β is a newsgroup-specific normalization constant. Coefficients of determination for the 101 newsgroups ranged from a minimum of 0.548 to a maximum of 0.96, with an arithmetic mean of 0.876.

As with lexical data (*Mandelbrot*, 1953, 1961; *Baayen*, 2001), better fits are possible (especially at low ranks) using generalizations of Zipf's law that contain more parameters.

A simple stochastic model

Günther et al. (1992, 1996), *Levitin* and *Schapiro* (1993), and *Schapiro* (1994) proposed a simple, nonstationary, branching process that gives rise to Zipf's law. We found this process to be an excellent model for the growth and evolution of newsgroups.

In the present context, the model assumes that a newsgroup contains $N_k(N)$ messages from contributor k at the time of the Nth posting to the newsgroup. Also, at this time, $A(N)$ individuals have contributed to the newsgroup, so that $k = 1, 2, \ldots, A(N)$. N_k and A are integer-valued random variables. N, the number of posts, indexes time because the process is updated with each new contribution.

Let a new message arrive. The process assigns this new message, post $N+1$, to a new participant, contributor $A(N)+1$, with probability $c(N)$. The probability that the new

message is due to *old* contributor k is, in turn, proportional to [1-$c(N)$] and to the relative frequency of k's submissions to the newsgroup just before the new message,

$$\Pr\{N_k(N+1) = n_k + 1 \,|\, N_k(N) = n_k\} = \frac{[1-c(N)]\, n_k}{N}. \qquad (2)$$

Here, n_k is a particular value of the random variable N_k. The process starts with the second message.

Parameter estimation

For most newsgroups, the odds that a post is due to a new contributor remain remarkably constant over time (data not shown). We therefore took $c(N)$ to be a newsgroup-specific constant, c. Since the process starts with the second message, we subtracted one from the number of posters, one from the number of posts, and estimated c by taking the ratio of these two differences. For the 101 newsgroups, c ranged from a minimum of 0.2 to a maximum of 0.75 with an arithmetic mean of 0.469 (see Figure 2). We applied the rapid test of *David* et al. (1954) and failed to reject the null hypothesis that the values of c are normally distributed.

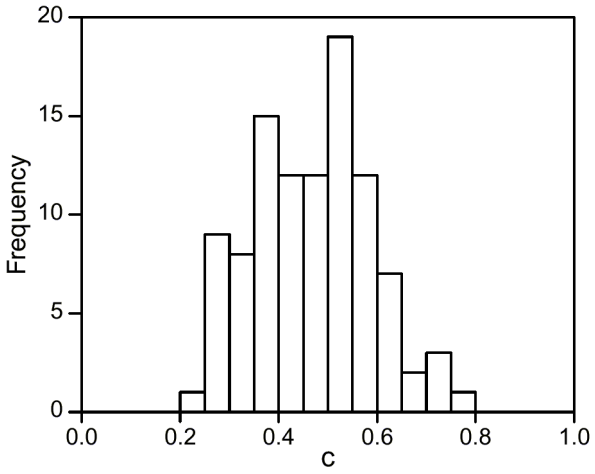

Figure 2. A histogram for the frequency of c, the probability that a post is due to a new contributor, for 101 newsgroups. For each newsgroup, we subtracted one from the number of posters and one from the number of posts. We then estimated c by taking the ratio of these two differences

Miscounts of news items and contributors can distort the estimation of c. To determine the extent of this potential bias, we hand-tagged bionet.drosophila, a typical newsgroup. This tedious and time-consuming procedure involved ascertaining the identity of each poster, in some cases by emailing contributors. Due to internal documents that were mistaken for separate posts, our computer programs counted 8 too many posts, for an error rate of 0.0014. More importantly, many contributors regularly employed more than one user ID or changed their login names at the time of a move. Spammers routinely changed their identities. Consequently, our programs counted 301 too many contributors, for an error rate of 0.12. Our estimate of c for the machine-tagged data (0.47) differed from our estimate for the hand-tagged data (0.42) with a relative error (0.12) that closely paralleled our error in counting the number of contributors. In spite of these differences, the frequency–rank data for the machine-tagged and hand-tagged newsgroups each obeyed Zipf's law. Subsequent analyses (not shown) revealed that the stochastic process fits both hand-tagged and machine-tagged data sets equally well.

Assessing the model

If the model of *Günther* et al. (1992) accurately describes the growth and evolution of newsgroups, the slopes of the log-log plots of frequency versus rank for the proposed stochastic process should agree with the corresponding slopes for real newsgroups. *Günther* et al. (1992, 1996), *Levitin* and *Schapiro* (1993), and *Schapiro* (1994) previously showed that for $c(N) = c$, a constant, and for $c \ll 1$, the expected value of the slope in the Zipf's plot of the stochastic process is asymptotically ($N \gg 1$) equal to $-1+c$. Since many of our newsgroups do not satisfy $c \ll 1$ and $N \gg 1$, we computed the expected slopes directly by simulating the model for each newsgroup 1000 times using the estimated c and the total number of postings for that newsgroup. For each simulation, we determined the slope using least squares; we averaged the resulting slopes and compared them to the slopes for the real newsgroups (Figure 3). The expected and observed slopes are in close agreement ($r = 0.96$). The stochastic process accurately predicts the slopes of the log-log plots of rank and frequency for real newsgroups.

Our simulations also revealed significant deviations from the asymptotic slope $-1+c$ for small and medium-sized newsgroups. Figure 4 shows expected slopes for newsgroups with 2000 postings for different values of c. This curve is nonlinear for both low and high c. For low c, too few news items are present to allow convergence to the asymptotic slope. (Increasing N removes this nonlinearity.)

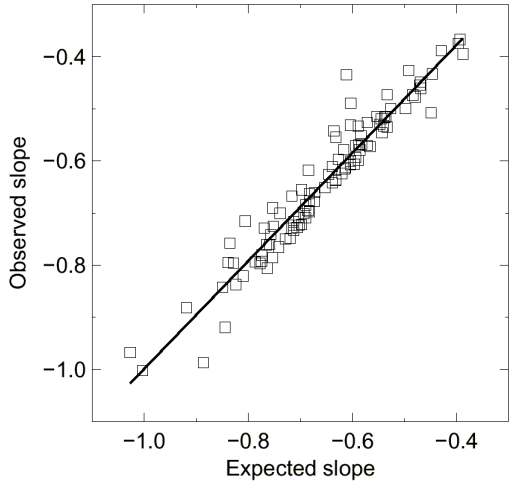

Figure 3. A scatter diagram of expected and observed slopes for 101 newsgroups. For each newsgroup, the slope of the log-log plot of frequency and rank was compared to an expected slope. The latter was obtained by simulating a simple stochastic process 1000 times using parameters for the newsgroup. The observed and the expected slopes are quite close. The best-fit line through the data has a slope of 1.03±0.05, an intercept of 0.04±0.03 and a correlation coefficient of 0.96

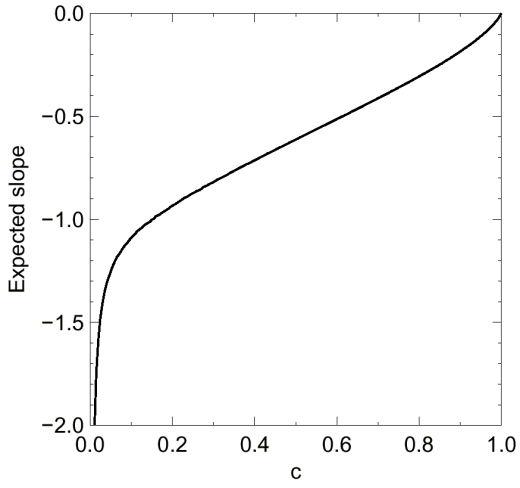

Figure 4. Expected slope as a function of c for simulated newsgroups of 2000 posts. The curve was obtained by simulating a simple stochastic process 1000 times for each of 200 c values on the interval (0, 1]. Note that the relationship is linear between $c = 0.2$ and $c = 0.75$

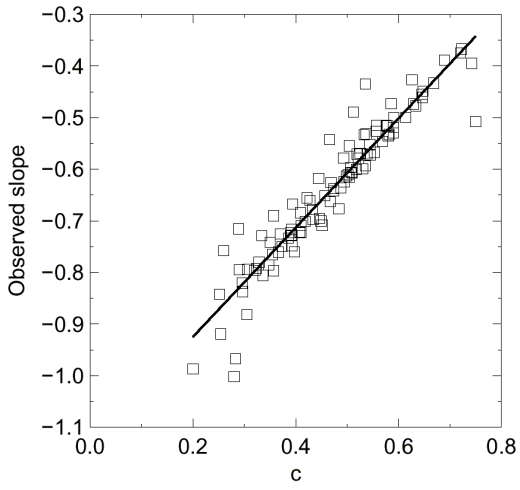

Figure 5. A scatter diagram of slope and c for 101 newsgroups. For each newsgroup, the slope of the log-log plot of frequency and rank was compared to the estimated value of c for that newsgroup. The plot reveals a linear relationship between the two variates over most of the range

Even so, for intermediate c values the slope is a linear function of c. The slopes for the real newsgroups also show a linear dependence on c (Figure 5). Linear regression of the data in Figure 5 suggests that the BIOSCI/bionet newsgroups follow the simple power law

$$f_R = \beta R^{-1.14+1.06c} . \tag{3}$$

In principle, one can predict how often a ranked individual will post to a newsgroup knowing only the number of contributors and the total contributions to the newsgroup!

In spite of this overall regularity, we could see differences between real newsgroups and simulations. To determine the extent of these differences, we chose a newsgroup and counted the number of individuals that posted once, twice, and so on. We computed this same "frequency spectrum" (*Baayen*, 2001) for a simulation of the newsgroup. We then performed a Kolmogorov–Smirnov test (*Kanji*, 1999) comparing the two spectra. We performed this analysis for each newsgroup and repeated the process 100 times. The real and simulated spectra differed significantly at the 0.05 level in 37% of the 101×100 simulations. For 27 of the 101 newsgroups, real and simulated spectra never differed significantly; for 17 newsgroups, these spectra always differed.

Conclusion and discussion

Bionet newsgroups exhibit striking statistical regularities in their growth and evolution. These patterns occur despite large amounts of internet noise. For the hand-tagged newsgroup bionet.drosophila, this noise included spam (5.6% of all posts), replies to religious or sacrilegious cross-postings (2.7% of all posts), and scores of high school students asking basic questions about fruitfly biology (8–9% of all posts).

The observed regularities are readily explained by a simple stochastic model that assumes a constant influx of new contributors and that grants most older participants a greater probability of posting news. This model excels at predicting the slopes of the log-log plots of the frequency of posting as a function of the rank of the poster for real newsgroups. It predicts frequency spectra less accurately.

Although the assumptions of the *Günther* et al. (1992) model are intuitive, it is still quite surprising that this stochastic model accurately reproduces the patterns of diversity for newsgroups. Similar attempts to develop models that predict the relative abundance of species in ecology have, by and large, failed (*Hubbell*, 2001).

We also considered the possibility that the observed statistical patterns might be explained by the oft-cited model of *Simon* (1955). The models of *Simon* (1955) and of *Günther* et al. (1992) are closely related; the former is a more general, but asymptotic, version of the latter. *Simon* derives a *Yule* (1924) distribution. This Yule distribution predicts frequency spectra less accurately than our simulations. (Using a Kolmogorov–Smirnov goodness-of-fit test, the spectra of 50% of the newsgroups differed significantly from *Simon*'s Yule distribution at the 0.05 level.) This disparity suggests that some newsgroups have not been in existence long enough to reach their stationary distributions. With either model, a critical role is played by c (*Simon*'s α), the probability that a new message comes from a new poster. Arguably, most of the differences in the diversity of newsgroups are due to differences in this single parameter.

The parameter c is almost identical to the poster-to-post ratio. *Smith* (1999) suggested that the poster-to-post ratio is a "good rough measure" of the quality of interaction in a newsgroup: A poster-to-post ratio close to zero indicates few active participants; a poster-to-post ratio close to one indicates no dialogue. Our own research suggests that the poster-to-post ratio (more properly c) is more than just a good rough measure of interaction. It is a useful summary statistic that allows one to predict the relative contribution of posters to a newsgroup. It is thus a natural starting point for comparative studies that examine the success, failure, and stability of newsgroups, the diffusion and crossposting of information, and social and scientific networking.

The regular growth and predictable diversity observed in bionet newsgroups also appear in other processes. An analysis of our own email suggests that Zipf's law governs email archives as well as newsgroups.

*

We thank the scores of contributors to bionet.drosophila who responded so quickly to our queries about their login names.

References

BAAYEN, R. H. (2001), *Word Frequency Distributions.* Kluwer Academic Publishers, Dordrecht, Netherlands.

BAR-ILAN, J. (1997), The "mad cow" disease, Usenet newsgroups and bibliometric laws. *Scientometrics*, 39 : 29–55.

DAVID, H. A., HARTLEY, H. O., PEARSON, E. S. (1954), The distribution of the ratio, in a single normal sample, of range to standard deviation. *Biometrika*, 41 : 482–493.

FRONTIER, S. (1985), Diversity and structure in aquatic ecosystems. *Oceanography and Marine Biology: An Annual Review*, 23 : 253–312.

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.

GÜNTHER, R., SCHAPIRO, B., WAGNER, P. (1992), Physical complexity and Zipf's law. *International Journal of Theoretical Physics*, 31 : 525–543.

HAUBEN, M., HAUBEN, R. (1997), *Netizens: On the History and Impact of Usenet and the Internet.* IEEE Computer Society Press, Los Alamitos, California, USA.

HUBBELL, S. P. (2001), *The Unified Neutral Theory of Biodiversity and Biogeography.* Princeton University Press, Princeton, New Jersey, USA.

HUBERMAN, B. A., PIROLLI, P. L. T., PITKOW, J. E., LUKOSE, R. M. (1998), Strong regularities in World Wide Web surfing. *Science*, 280 : 95–97.

KANJI, G. K. (1999), *100 Statistical Tests.* Sage Publications, London, UK.

LEVITIN, L. B., SCHAPIRO, B. (1993), Zipf's law and information complexity in an evolutionary system. *Proceedings IEEE International Symposium on Information Theory*, 76.

LI, W. (1992), Random texts exhibit Zipf's-law-like word frequency distribution. *IEEE Transactions on Information Theory*, 38 : 1842–1845.

MANDELBROT, B. (1953), An information theory of the statistical structure of language. In: W. E. JACKSON (Ed.), *Communication Theory*, Academic Press, New York, New York, USA, pp. 486–502.

MANDELBROT, B. (1961), On the theory of word frequencies and on related Markovian models of discourse. In: R. JAKOBSON (Ed.), *Structure of Language and its Mathematical Aspects*, American Mathematical Society, Providence, Rhode Island, USA, pp. 190–219.

MAGURRAN, A. E. (1988), *Ecological Diversity and Its Measurement.* Princeton University Press, Princeton, New Jersey, USA.

MARSILI, M., ZHANG, Y.-C. (1998), Interacting individuals leading to Zipf's law. *Physical Review Letters*, 80 : 2741–2744.

MILLER, G. A., NEWMAN, E. B., FRIEDMAN, E. A. (1957), Some effects of intermittent silence. *American Journal of Psychology*, 70 : 311–313.

OKUYAMA, K., TAKAYASU, M., TAKAYASU, H. (1999), Zipf's law in income distribution of companies. *Physica A*, 269 : 125–131.

OSBORNE, L. N. (1998), Topic development in USENET newsgroups. *Journal of the American Society for Information Science*, 49 :1010–1016.

SCHAPIRO, B. (1994), An approach to the physics of complexity. *Chaos, Solitons and Fractals*, 4 : 115–123.

SIMON, H. A. (1955), On a class of skew distribution functions. *Biometrika*, 42 : 425–440.

SMITH, M. A. (1999), Invisible crowds in cyberspace: mapping the social structure of the Usenet. In: M. A. SMITH, P. KOLLOCK (Eds), *Communities in Cyberspace*, Routledge, London, UK, pp. 195–219.

TOKESHI, M. (1993), Species abundance patterns and community structure. *Advances in Ecological Research*, 24 : 111–186.

WILSON, J. B., WELLS, T. C. E., TRUEMAN, I. C., JONES, G., ATKINSON, M. D., CRAWLEY, M. J., DODD, M. E., SILVERTOWN, J. (1996), Are there assembly rules for plant species abundance? An investigation in relation to soil resources and successional trends. *Journal of Ecology*, 84 : 527–538.

YULE, G. U. (1924), A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F. R. S. *Philosophical Transactions B*, 213 : 21.

ZIPF, G. K. (1935), *The Psycho-Biology of Language.* Houghton Mifflin, Boston, Massachusetts, USA.

ZIPF, G. K. (1949), *Human Behavior and the Principle of Least Effort.* Addison-Wesley Publishing Company, Cambridge, Massachusetts, USA.