Developing a predictive model of editor selectivity in a current awareness service of a large digital library

Thomas Krichel [a], Nisa Bakkalbasi [b,*]

[a] *Palmer School of Library and Information Science, Long Island University, 720 Northern Boulevard, Brookville, NY 11548, USA*
[b] *Purchase College Library, SUNY, 735 Anderson Hill Road, Purchase, NY 10577, USA*

Abstract

We examine editor selectivity in the "NEP: New Economics Papers" current awareness service for the RePEc digital library. The service is powered by volunteer editors who filter new additions to RePEc into subject-specific reports. The intended purpose of this current awareness service is to filter working papers by subject matter without any judgment of its academic quality. We use binary logistic regression analysis to estimate the probability of a paper being included in any of the subject reports as a function of a range of observable values. Our analysis suggests that, contrary to their claims, editors use quality criteria. Such criteria include the series the paper is coming from as well as the reputation of the authors. Our findings suggest that something as simple as a current awareness service can be taken to be the opening step of a peer-review process.

Keywords: Current awareness; SDI (Selective dissemination of information); Digital library; Binary logistic regression analysis

1. Introduction

RePEc (Research Papers in Economics) is a large digital library for economics research. The roots of RePEc go back to 1993, when Thomas Krichel started to collect information about downloadable electronic working papers in economics. Working papers are accounts of recent research results before formal publication. Most economics departments in universities, as well as many other institutions that are involved in economics research—such as central banks and intergovernmental organizations—publish working papers. At the time of writing, over 360 archives based at institutions that issue working papers contribute to RePEc. They provide classic bibliographic data about the papers that they publish, as well as links to the full text. These data are harvested by service providers, who aggregate them to produce services for users who are interested in economics research. The RePEc web site at *http://repec.org* lists the service providers. All RePEc services are available to the public at no charge.

* Corresponding author. Tel: +1-914-251-3415, Fax: +1-914-251-6437.
E-mail addresses: krichel@openlib.org (T. Krichel), nisa.bakkalbasi@purchase.edu (N. Bakkalbasi).

One of the RePEc services is NEP: New Economics Papers. NEP is a human-mediated current awareness service. It was founded by John S. Irons and Thomas Krichel in 1998. It primarily operates through electronic mail. It also has a web homepage at *http://nep.repec.org*. NEP is entirely run by volunteer editors from all corners of the globe. Most of them are PhD students or junior university faculty.

NEP has a simple, two-stage work flow. In the first stage, the general editor collects all new working paper data that have been submitted to RePEc in the previous week. She or he filters out records corresponding to papers that are new to RePEc, but are not new papers. Such records would typically come from new RePEc archives that add a whole back catalog of working papers to RePEc. The remaining records form a NEP report called nep-all. As its name suggests, it contains the new working papers in RePEc from the previous week. Each issue of nep-all is circulated via e-mail to subject editors. This completes the first stage. In the second stage, the subject editors filter every nep-all issue they receive to contain only papers in a certain subject area. When a new subject-specific issue has been created, it is circulated via e-mail to subscribers of the subject report.

Since its inception in 1998, NEP has grown in scale and scope. As RePEc has grown, so has the size of nep-all issues. This is scale growth. On the other hand, over time, more and more subject reports have been created. This is scope growth. At the time of this writing, there are close to 60 distinct subject reports in NEP. Over 11,000 unique e-mail addresses have subscribed to at least one NEP report. Over 30,000 new papers have been announced.

Chu & Krichel (2003) find that NEP is an interesting service model for digital libraries. Still, questions about its sustainability, as a volunteer service, remain. Barrueco Cruz, Krichel, & Trinidad (2003) present a simple empirical assessment of the NEP service. One of the issues they look at is the subject coverage of RePEc as a whole, i.e., if NEP covers all subjects that we find in RePEc. If it does, then we should observe that each working paper in a nep-all issue appears in at least one subject report. Empirically, this conjecture can be examined by looking at the ratio of papers in nep-all that have been announced in at least one NEP subject report. As more and more subject reports have been added, we expect the coverage ratio to improve over time, and in the longer run to reach 100%. Surprisingly, the data reported by Barrueco Cruz, Krichel, & Trinidad (2003) suggest that the coverage ratio has not been improving as more reports have been added, and that certainly remains well below 100%.

In this paper, we are looking for explanations of this puzzle. We formally investigate subject editors' behavior, specifically to examine what makes a paper "announceable" in a NEP report. The remainder of our paper is organized as follows. In Section 2, we present a conceptual framework. In Section 3, we describe our methodology for developing a predictive model. In Section 4, we discuss our data set. In Section 5, we present our findings. In Section 6, we develop our conclusions and suggest future work.

2. Conceptual Framework

We have two basic theories about editor behavior that aim to explain the static nature of the NEP coverage ratio. We call them the "target theory" and the "quality theory," respectively. Let us examine them in turn.

The target theory starts with the observation that the size of nep-all issues has been highly volatile in the short run, and has been steadily growing in the long run. The theory suggests that, when composing an issue of a subject report, the editors have an implicit issue size in mind. Therefore, if the size of nep-all is large, they will take a narrow interpretation of the subject matter of the report, i.e., they will be choosier as to what papers they include. Thus, the target theory claims that the observed long-run static nature of the coverage ratio comes from the simultaneous effect of scale and scope growth of NEP. Scale growth, all other effects being equal, will reduce the coverage ratio. Scope growth, all other effects being equal, will increase the coverage ratio. The long-run static coverage ratio is the result of both effects canceling each other out.

The quality theory suggests that the subject editors filter for paper quality. There are two types of quality indicators. First, there is the descriptive quality of the record that describes a paper. Some papers are described poorly. They have a meaningless title, and/or no abstract. Second, there is the substantive quality of the paper itself. The paper may be written by authors whom nobody has ever heard of, and/or who are based at institutions with an unenviable research reputation. Whether it is substantive or descriptive, the quality of a paper is likely to be important when it comes to its inclusion in any NEP report.

Surprisingly enough, an e-mail discussion on the private mailing list used by the subject editors has revealed that editors have a uniform view of the quality theory: they reject it. They claim that they perform their work independent of quality considerations. They assert that their only concern is to disseminate new working papers based on the subject matter. Furthermore, they specifically insist that NEP cannot be regarded as a vehicle for a preliminary peer-review.

If the general assessment of the editors is wrong, then NEP may be considered as a first stage in an alternative peer-review system. Such a system may be constructed sitting on top of NEP, as an extended service. To date, RePEc does not engage in peer-review other than vetting the providers of archives. But the idea of quality review through NEP could start to change that, in the long run.

The debate between the two theories also has some short-run stakes for the running of NEP itself. If the target theory is correct, then opening additional specialized report categories should be considered as a way to improve the coverage of NEP. If the quality theory is correct, opening additional report categories will have no effect on coverage. Instead, it may have the adverse effect of making NEP more cumbersome to administer. This question of whether to open more reports or not has been one of the important motivations for the research conducted in this paper.

3. Research Methodology

A simple way to assess the target theory empirically is to see if the coverage ratio declines with the size of a nep-all issue. Barrueco Cruz, Krichel & Trinidad (2003) have a cross-sectional plot of coverage ratio versus size of nep-all. The shape of the plot suggests that this seems to be the case. However, they offer no rigorous statistical test. Even if inferential statistics were used, it would only look at one aspect of the selectivity issue. What we need is an overall model that combines a set of important variables to assess the probability of a working paper being "announced" in any report.

We speculate that some observable variables influence the odds that a working paper makes into at least one subject report, i.e., the working paper is "announced". The observable variables we can think of as relevant are:

- the size of the nep-all issue in which the paper appeared
- the length of a title of the paper
- the presence/absence of an abstract to the paper
- the membership of a paper in a series
- the prolificacy of the authors of the paper

Statistically speaking, we conjecture that a percent of the variance in the response variable (i.e., the presence/absence of a paper in any report) can be accounted by predictor variables. If our conjecture turns out to be correct, we can produce a prediction equation that will allow us to predict the probability of a working paper being included in a NEP report. We believe that the most appropriate statistical method for analyzing this relationship is Binary Logistic Regression Analysis (BLRA). There are three reasons for our choice. First, the dependent variable is dichotomous that can suitably be coded with values of 0 and 1. Second, the independent variables are both quantitative and qualitative in nature. Last and most important, BLRA is a flexible technique. BLRA does not require any of the following assumptions commonly made for linear regression analysis to work:

- a linear relationship between the independent variables and the dependent variable
- homoscedasticity of the dependent variable for each level of independent variables
- a normally distributed dependent variable
- normally distributed error terms

According to Hosmer & Lemeshow (2000), BRLA originated in the epidemiological research. It is now heavily used in biomedical research. Its use in information science has not been wide-spread. In fact, having conducted a thorough literature review, we have not found an information science paper that has used this technique. However, in a review of statistical techniques in library and information sciences, Bensman (2001) suggests that, because of the highly skewed probability distributions observed in this discipline, the researcher should look at the biomedical sciences for methodologies used to attack these issues. As the NEP project is breaking new grounds as far as digital library business models are concerned, this paper is breaking new ground in its use of a novel statistical analysis methodology in library and information sciences.

4. Data Set

The data set has been extracted from historic e-mail message archives that contain NEP report issues. In addition, we have used the bibliographic records for the papers referred to in the report issues. The data go back to the inception of NEP in 1998 and contain 32,892 records, with each record corresponding to a paper that was appeared in nep-all. Our response variable (i.e., dependent variable) is called ANNOUNCED. It takes the value 1 if the paper was announced, and 0 if not.

Table 1
The variables identified for exploration

Description	Values	Variable Name
announcement of paper	1 = yes, 0 = no	ANNOUNCED
nep-all size	3 to 803	SIZE
number of characters including space in the title	3 to 1945	TITLE
presence/absence of an abstract	1 = yes, 0 = no	ABSTRACT
average announcement ratio of series	0 to 5.5	SERIES
number of papers the lead author submitted to RePEc archives previously	1 to 284	AUTHOR

Let us now look at the candidate predictor variables (or independent variables) in turn. SIZE corresponds to the size of a nep-all issue and is easily tracked. For TITLE (the length of the title) and ABSTRACT (the presence/absence of an abstract), we refer to the bibliographic record of the working paper. In the very rare cases where the bibliographic information was not available, we dropped the record. For SERIES (the number of different subject-specific reports a series appears in), we use a numeric variable that measures the ratio of the total number of times working papers from a specific series have been announced in subject reports, divided by the total number of working papers from that series that appeared in nep-all. This gives us an overall indication of how well-respected and visible a series is. AUTHOR, the measure of prolificacy of an author, is the most difficult variable to construct and measure. There are at least three problems with constructing such a measure. First, RePEc does not cover the entire economics discipline. Second, it is not easy to know if two similar author names represent the same person. Third, since co-authorship is frequent in economics, one needs to decide how to aggregate the prolificacy of individual authors. To deal with the second problem, RePEc runs an author registration service at *http://authors.repec.org*. This service collects records about the authors and the papers they have written. Authors contact the service to build their own electronic CVs. Such registration is voluntary, of course. We do not have data available for all authors of all papers. Therefore, we need to look for a measure that covers as many papers as possible. After considering several alternatives, we have decided to use a measure we call the "lead author." For each paper in the data set, we take the number of papers in the RePEc database of the registered author with the largest number of papers. Still, due to a high number of unregistered authors, we end up with a significant number of records with missing values. After removing these records, we are left with 10,652 records. Since author registration and appearance of papers in reports are independent events, we conjecture that the removal of a large number of records introduces no sampling bias. To confirm, we analyze the descriptive statistics of both the original data set and the smaller data set. We find that the averages of the other

independent variables stayed approximately the same after the removal of about two thirds of the records. The size of the remaining data is still amply sufficient to conduct our analysis. Table 2 shows a few sample records from the data set before the removal of the records with missing values. In Table 2, HANDLE corresponds to the unique identifier for each record in the data set.

Table 2
Data set sample

HANDLE	ANNOUNCED	SIZE	TITLE	ABSTRACT	SERIES	AUTHOR
RePEc:jku:econwp:2001_05	0	230	88	1	1.083	NA
RePEc:nbr:nberwo:9361	1	175	42	1	1.619	NA
RePEc:fip:fedfap:2002-02	1	433	54	1	1.519	95
RePEc:wop:wobaiy:2957	0	803	66	0	1.917	NA
RePEc:cbr:cbrwps:wp207	1	433	74	1	1.405	NA

All our calculations use the R[1] language and environment. The computer code, as well as our data set, is available on request.

5. Findings

5.1. Exploratory Data Analysis

We begin our study by examining the data set in order to describe the main characteristics of each variable. We start with a frequency count of the response variable ANNOUNCED, reported in Table 3. It shows that nearly 78% of the 10,652 working papers are included in at least one subject-specific report.

Table 3
Frequencies of responses

ANNOUNCED	
0 = no	1 = yes
2373	8279

Table 4 displays the descriptive statistics for the quantitative predictor variables. It does not include the qualitative variable ABSTRACT. Initial intuition suggests constructing ABSTRACT as the number of characters in each abstract. However, proceeding in that way, we encounter a wide range of values [0, 11295], with the value 0 occurring very frequently. Conventional measures of central tendency and variance are meaningless in this context. Therefore we make ABSTRACT a categorical variable and encode it as 0 (no abstract) and 1 (has abstract) within each record. This is a common procedure used in statistical analysis to remove the large variation in a predictor variable while maintaining the functional relationship between the response variable and the predictor variable.

[1] R is a language and environment for statistical computing and graphics. For more information visit http://www.r-project.org/.

Table 4
Descriptive statistics for quantitative predictor variables

	SIZE	TITLE	SERIES	AUTHOR
Minimum	3.0	3.0	0.000	1.0
1st quartile	125.0	48.0	1.245	11.0
Median	202.0	63.0	1.480	26.0
Mean	240.4	66.2	1.450	40.5
Standard deviation	160.5	32.0	0.421	42.3
Coefficient of variation [a]	0.668	0.483	0.290	1.044
3rd quartile	306.0	82.0	1.619	55.0
Maximum	803.0	1945.0	5.500	284.0

[a] Coefficient of Variation = Standard Deviation / Mean

For each quantitative predictor variable, we look at different measures of central tendency (i.e., mean and median) and dispersion (i.e., standard deviation, coefficient of variation, range). Two quantitative predictor variables, TITLE and SERIES, have their mean and median close to each other and their variation is small, implying a symmetrical distribution. For these two variables, the mean is an appropriate measure to determine their typical values for observation. Therefore, a typical working paper title contains 66 characters and the average announcement ratio for a series is 1.45.

For the variables AUTHOR and SIZE, there are differences between the mean and the median. The dispersion measures for those two variables indicate a high variation, due to the frequency of extreme values. To illustrate, there are many authors who have written 3 or 4 papers. There is only one author (Nobel laureate Joseph E. Stiglitz) with 284 papers. But by the very fact that he appears as a prolific author on so many papers, he introduces an upward distortion for the average number of papers that an author has written. The same scenario is valid for SIZE. The mean is highly influenced by extreme values. Therefore, for these two variables, the median qualifies as a more appropriate measure of central tendency than the mean. Thus, we conclude that an average lead author has written 26 papers, and that the typical nep-all issue contains 202 working papers.

5.2. Outlier analysis

Visual examination of the descriptive statistics shows some questionable observations. For example, there appears a paper title with as few as 3 characters and another with as many as 1945 characters. One nep-all issue containing 803 working papers immediately raises eyebrows. To detect striking deviations as potential outliers, we carefully examine lists of data points with values greater than three standard deviations from the mean. According to the empirical rule, the interval $(\bar{y} \pm 3s)$, where $\bar{y}$ is the mean of a variable and s is its standard deviation, contains virtually all the observations if the shape of the distribution is nearly bell-shaped. Although empirical rule furnishes us with a practical way of obtaining potential outliers, it does not appear to work well for variables that are not bell-shaped.

In general, dealing with outliers is difficult and a matter of judgment. We find that in some cases the outliers are extraordinary occurrences. In other cases, they are likely to be errors in the data. Whenever we encounter a suspicious observation, we attempt to correct it by looking at the original record. For example most of the titles with less than

10 characters turn out to be acronyms for instructional data sets. These should not have appeared in nep-all issues at all. We drop the erroneous occurrences that cannot be corrected and keep what we assess to be extraordinary occurrences. At the end of this process, we identify 86 records as outliers. This allows us to use 10,566 records for building the binary logistics model.

5.3. Inferential Data Analysis

We use the "Design Library of Modeling Functions" in R to build the regression model. Prior to building the model, we perform a test to see whether the predictor variables are correlated to each other. Based on a Pearson correlation test among the 5 predictor variables, we conclude that there is no significant pair-wise correlation among any of the pairs of the predictor variables. Lack of correlation among the predictor variables increases our confidence that there is a higher likelihood for each predictor variable to contribute to the final prediction equation independently.

5.3.1 Fitting and testing the model

The results of the logistic regression analysis are shown in Table 5.

Table 5
Estimated coefficients for multiple logistic regression model

	Coefficient	S.E.	Wald	P
Intercept	–1.1202	0.1268499	–8.83	0.0000
SIZE	–0.0008	0.0001454	–5.61	0.0000
TITLE	0.0038	0.0009651	3.89	0.0000
ABSTRACT	0.3067	0.0634233	4.84	0.0001
SERIES	1.4434	0.0696371	20.73	0.0000
AUTHOR	0.0025	0.0006381	3.87	0.0001
	Model χ_2	d.f.	p-value	
	690.94	5	0	

First, we need to assess how the overall model works. To that end, we perform the likelihood ratio test for the overall significance of the five coefficients for the independent variables. That is

$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$

H_A : At least one coefficient is not equal to 0

where β_i is the coefficient for each predictor variable, respectively.

The likelihood ratio test statistic takes the value $G_M = 690.94$, where G_M is referred to as the Model χ^2. This value leads us to reject the null hypothesis at virtually any significance level. We conclude that at least one of the five coefficients is different from zero and that, together, SIZE, TITLE, ABSTRACT, SERIES, and AUTHOR, are significant predictors of ANNOUNCED. Next, we use the Wald statistics for the individual coefficients to test the significance of the variables in the model.

$H_0 : \beta_i = 0$
$H_A : \beta i \neq 0$

The Wald test statistics W are the ratio of each coefficient to its standard error.

$$W_i = \frac{\hat{\beta}_i}{S\hat{E}(\hat{\beta}_i)}, \qquad \text{where } \beta_i \text{ is the coefficient for each predictor variable}$$

Based on the evidence contained in the data, at a significance level of $\alpha = 0.05$, we reject the null hypothesis for each of the five coefficients and conclude that each of the predictor variables is significant. Table 5 contains the details.

Finally, we attempt to obtain the best fitting model with the least number of parameters. To that end, we run different models using different subsets of the predictor variables. After a thorough comparison of various models, we conclude that we cannot exclude any of the variables from the final model. Let x be a vector representing values of the predictor variables. Then, the final logistic regression model, which gives us the estimated logistic probability, can be expressed as

$$(1) \qquad P(\text{ANNOUNCED} = 1 \mid x) = \frac{e^{\hat{g}(x)}}{1 + e^{\hat{g}(x)}}$$

where the estimated logit is given by the following expression:

$$(2) \qquad \hat{g}(x) = -1.1202 - 0.0008*\text{SIZE} + 0.0038*\text{TITLE} + 0.3067*\text{ABSTRACT} + 1.4434*\text{SERIES} + 0.0025*\text{AUTHOR}$$

5.3.2 Interpreting the fitted logistic regression model

In this section, we discuss the interpretation of the estimated coefficients in the model. Equation (1) gives us a probability for the event occurring given all the values of the predictors. Equation (2) looks like a linear regression model as it is commonly understood. In such a linear regression equation, the coefficients are interpreted as the rate of change in the dependent variable associated with one-unit change in the respective independent variables. In the logistic regression model however, the slope coefficient represents the change in the logit corresponding to a change of one unit in the independent variable. Therefore the relationship between the independent and the dependent variable is less intuitive. One commonly used measure of association is called the odds ratio, commonly abbreviated as OR. Roughly speaking, OR is a measure of the degree of association between each predictor variable and outcome. It is obtained by transforming the estimated coefficients. There are different ways of expressing odds ratio depending on the different types of predictor variables in the model, i.e., dichotomous, polychotomous, or continuous.

Table 6 contains the estimated OR values for our predictor variables. Interpretation of our categorical variable ABSTRACT is pretty straightforward. The OR for

the ABSTRACT coefficient is 1.36 with a 95% confidence interval of [1.20, 1.54]. This suggests that in the presence of an abstract, a working paper is 1.36 times more likely to be included in at least one subject category than in the absence of an abstract. In order to draw a meaningful interpretation for our continuous variables, we need to create intervals, which will allow us to observe an impact of "c" units of change in the independent variable as opposed to one-unit of change, which does not offer any practical inference for our study. Therefore, we create three intervals using the quartiles as cut off points before obtaining the odds ratios. The three intervals are the following:

1. 1st quartile
2. 2nd and 3rd quartiles combined
3. 4th quartile

Table 6
Estimated odds ratios and 95% confidence intervals for predictor variables

	Interval	Difference (c)	Odds Ratio (OR)	95% CI over OR
SIZE	3 – 125	122	0.91	[0.87, 0.94]
	125 – 306	181	0.86	[0.82, 0.91]
	306 – 803	497	0.67	[0.58, 0.77]
TITLE	4 – 48	44	1.18	[1.09, 1.28]
	48 – 82	34	1.14	[1.07, 1.21]
	82 – 218	136	1.67	[1.29, 2.15]
ABSTRACT	N/A	N/A	1.36	[1.20, 1.54]
SERIES	0 – 1.25	1.25	6.08	[5.12, 7.21]
	1.25 – 1.62	0.39	1.70	[1.62, 1.79]
	1.62 – 5.5	3.88	270.91	[159.51, 460.12]
AUTHOR	1 – 11	10	1.02	[1.01, 1.04]
	11 – 54	43	1.70	[1.62, 1.79]
	54 – 284	230	1.76	[1.32, 2.35]

Let us review each continuous predictor variable keeping the difference or "c" units of change within each interval in mind. The estimated OR for SIZE suggests that, in the first interval, where an increase of 122 papers occurs, the odds of a paper being announced in at least one subject category increases 0.91 times. In other words, an estimated OR of approximately 1 indicates that a working paper with a 122 increase in SIZE is equally likely to be announced or not to be announced. As it is shown in Table 6, the odds ratio reduces as the difference in SIZE increases for the next two intervals. We notice a significant drop in the odds ratio for the third interval where the increase in SIZE is 497, indicating that working papers from large nep-all issues are less likely to be included in subject-specific reports. When we examine the relationship between the two columns, "c" and OR for the TITLE variable we notice the odds for being "announced" increase as the title gets longer. More specifically, an increase of 136 characters in the title increases the odds of a working paper being included in a subject-specific report 1.67 times. Similarly, the OR estimates for SERIES and AUTHOR suggest that, as the respective "c" units of change increases, there are significant corresponding increases in the likelihood of a working paper being included in at least one subject-specific report.

In summary, we presented statistical evidence that the predictive model based on the five predictor variables described in this section has potential to offer practical value in assessing the likelihood of a working paper submitted to NEP being included in at least one subject-specific report.

6. Conclusions and future work

In this paper, we have set up and successfully tested a statistical model of editor selectivity in a current awareness service. The most important conclusion of our work is that the quality theory about editor behavior cannot be dismissed. That is, despite their assertion to the contrary, the subject editors appear to take into account the reputation of the series and the prolificacy of authors. Although we can only measure the authors' prolificacy in a very rough way, there still is statistical evidence of an editor bias. As more authors register with RePEc, we will get better data to assess our model again.

An important result form our work will be to examine the list of true-negative observations: those records that, according to our model, were likely to be included in a NEP report, but did not make it into any. This analysis, with the help of the subject editors, will allow us to identify possible gaps in the subject coverage of NEP.

We also conclude that the BLRA technique is suitable for building a quantitative model of editor selectivity. With the help of another set of prediction equations, we intend to examine each new nep-all issue automatically and develop a forecast for each subject editor regarding the inclusion of a given working paper in a specific subject area. This will make it easier for the editors to scrutinize the new working papers. The ultimate aim would be a recursive system where each new forecast is based on the evidence of the previous editorial judgment. Such a system will undoubtedly make the work of the subject editors easier, and keep NEP on a path of sustainability.

References

Barrueco Cruz, J. M., Krichel, T., & Trinidad, J. C. (2003). *Organizing current awareness in a large digital library*. Presented at the 2003 Conference on Users in Electronic Information Environments in Espoo, Finland, September 8-9, 2003, http://openlib.org/home/ krichel/papers/espoo.pdf.

Bensman, S. J. (2000). Probability distributions in Library and Information Science: A historical and practitioner viewpoint. *Journal of the American Society for Information Science and Technology,* 51(9), 816-833.

Bookstein, A. (2001). Implications of ambiguity for scientometric measurement. *Journal of the American Society for Information Science and Technology,* 52(1): 74-79.

Chu, H. & Krichel T. (2003). NEP: Current awareness service of the RePEc Digital Library. *D-Lib Magazine, 9*(12). http://www.dlib.org/ dlib/december03/chu/12chu.html

Hosmer, A. W. & Lemeshow S. (2000). *Applied logistic regression*. New York, USA: John Wiley & Sons

Maindonald, J. & Braun, J. (2003). *Data analysis and graphics using R – an example-based approach.* Cambridge, UK: Cambridge University Press.

Acknowledgements

We are grateful to Dr. Stephen J. Bensman, Louisiana State University, and Dr. Sune Karlsson, Stockholm School of Economics, for helpful comments on an earlier version of this paper.

Thomas Krichel is an Assistant Professor in Palmer School of Library and Information at Long Island University. He studied Economics and Social Sciences at the universities of Toulouse, Paris, Exeter and Leicester. Between February 1993 and April 2001 he lectured in the Department of Economics at the University of Surrey. In 1993 he founded NetEc, a consortium of internet projects for academic economists. In 1997, he founded the RePEc data set to document Economics. His main area of work is the

development open digital libraries for scholarly communication. His homepage is http://openlib.org/home/krichel.

Nisa Bakkalbasi is a Science Librarian in Purchase College Library at State University of New York. Her responsibilities include information literacy instruction, collection development, reference, and faculty liaison to the Natural Sciences Department. She holds an M.L.I.S. from Long Island University and an M.S. in Applied Statistics from the University of Alabama. Prior to joining Purchase College, Nisa worked for many years as an independent consultant providing consulting services in model building and data analysis.