Journal peer review as an information retrieval process

Peer-reviewed author version

Journal peer review as an information retrieval process

L. Bornmann[1]

and

L. Egghe[2,3]

[1.]**Max Planck Society, Hofgartenstr. 8, 80539 Munich, Germany**

[2.] **Universiteit Hasselt, Campus Diepenbeek, Agoralaan, B-3590 Diepenbeek, Belgium**[(*)]

[3.] **Universiteit Antwerpen, Stadscampus, Venusstraat 35, B-2000 Antwerpen, Belgium**

bornmann@gv.mpg.de
leo.egghe@uhasselt.be

ABSTRACT

In editorial peer review systems of journals, one does not always accept the best papers. Due to different human perceptions, the evaluation of papers by peer review (for a journal) can be different from the impact that a paper has after its publication (measured by number of citations received) in this or another journal. This system (and corresponding problems) is similar to the information retrieval process in a documentary system. Also there one retrieves not always the most relevant documents for a certain topic. This is so because the topic is described in the command language of the documentary system and this command does not always completely cover the "real topic" that one wants to describe.

Based on this statement we are applying classical information retrieval evaluation techniques to the evaluation of peer review systems. Basic in such an information retrieval evaluation are the notions of precision and recall and the precision-recall-curve. Such notions are introduced here for the evaluation of peer review systems. The analogues of precision and recall are defined and we construct their curve based on peer review data in the journal *Angewandte Chemie - International Edition* and on citation impact data of accepted papers by this journal or rejected but published elsewhere papers. Conclusions on the value of peer review are drawn.

Introduction

Peer review in journal publishing is the most classical review system and consists of the advice (to the Editor-in-Chief of the journal) of peers on the quality of a submitted paper, prior to publication. Peer review has always been there since there were journals, now more than 300 years ago. Peer review has been well-described in the literature, see e.g. Weller (2001) and Bornmann (2011) and the many references therein. We do not go into the issue of studying different peer review systems since we will not be needed it in this paper. In every peer review system, peer review always leads (possibly after some minor or major revisions) to a final acceptance or rejectance by the Editor-in-Chief of the journal, based on the reviewers' (or peers) advices.

Hence peer review, as any human activity, is depending on human perception and this can lead to different conclusions of acceptance or rejection if different peers are used in the peer review process of a journal - see e.g. Bornmann and Daniel (2009b), Egghe (2010) and Schultz (2010). As a consequence of this, one does not always accept the best papers, as e.g. expressed by citation counts of articles after publication in this or another journal. It is clear that this is (or can be) a serious problem for peer review systems as well as for authors' careers (which depend on correct evaluations of their papers).

To study this phenomenon let us fix any peer-reviewed journal (a large practical example will be given for *Angewandte Chemie - International Edition* in the sequel). In a certain time period, one can determine the complete set of submitted papers in this period. A subset of it consists of the accepted papers in this period. This subset is known by the scientific community since it simply consists of the published papers in this journal in this time period (and acceptance dates are usually added on each paper). However, the set of submitted papers is only known by the services of the Editor-in-Chief (EIC) and must be kept confidential with regard to the authors of rejected papers. Next, of all the accepted papers and of all the rejected papers (but published elsewhere), in order to evaluate the peer review system, we need a way to measure their impact, say by using a normalized (normalized for the different fields) citation count measure (more details will be given in the sequel). Then papers above a certain threshold of this measure can be defined as qualified (Q) (see Bornmann and Daniel (2010a)). It is clear that also the set of qualified papers is unknown by the scientific community and is only known by the services of the EIC on condition that the above mentioned citation scores for accepted and rejected (but published elsewhere) papers are collected.

It is the hope of every EIC that the set of qualified papers is not much different from the set of accepted papers, but that is far from sure. This depends on the quality of the peer review process but, as said, this is a human activity and, hence, is not perfect. So, in general we can say that we have four sets of papers:

1. The set of accepted and qualified papers,
2. The set of accepted and non-qualified papers,

3. The set of rejected and qualified papers and
4. The set of rejected and non-qualified papers

These four sets constitute a partition of the set of submitted papers. The sets 1. and 4. represent correct decisions: the first set consists of the accepted qualified papers (which are well-classified) and the fourth set consists of the rejected non-qualified papers (which are also well-classified). The sets 2. and 3. represent erroneous decisions: the second set consists of accepted non-qualified papers (which are wrongly classified) and the third set consists of rejected qualified papers (which are also wrongly classified).

It is clear that this situation completely resembles the situation of information retrieval (IR) systems, where a researcher is looking for information on a certain topic, by using the command language of a certain documentary system. The topic determines the unknown set of relevant (rel) documents in the system (we assume here that, for every document in the documentary system we can decide in a binary way if the document is relevant (rel) or non-relevant (nrel)). However, in practical IR activities, one cannot check every document in the system: we have to use the command language of the system so that a query can be constructed that replaces the "topic" and with which one can then start the retrieval process. This then leads to a set of retrieved (ret) documents (and hence the complement of this set in the documentary system is the set of the not-retrieved (nret) documents). We will also use the notation rel and nrel for the set of relevant, respectively, not-relevant documents.

Again, it is the hope of any IR searcher that the sets rel and ret are very similar (or equal at its best). But since the used query is a logically formulated version of the real sought topic the sets rel and ret are usually different. We assume that the set ret is completely known since it is the set of retrieved documents. In ordered outputs in decreasing order of expected relevance (such as in search engines) we determine the set ret by cutting off this ranked list after the nth document ($n = 1, 2, 3, ...$). The set rel is unknown unless we check (for research purposes) the entire documentary system. However, for evaluation purposes (see further) one only needs the number of documents in rel and this can be estimated by sampling in the documentary system in order to determine a confidence interval for the fraction of relevant documents in this documentary system.

Let us, briefly, repeat the classical way of evaluation of an IR-system, based on some retrieval results. They are elementary and well-known in IR but we mention them here since we need it further in the evaluation of a peer review system of a journal. When performing a search, as explained above, we have the sets rel, nrel, ret and nret. The Precision (P) of the search is the fraction of retrieved documents that are relevant. The simple formula reads ($|.|$ denotes the number of elements in the set, i.e. the cardinality of the set):

$$P = \frac{|ret \cap rel|}{|ret|} \qquad (1)$$

This indicator is known when performing an IR-search since we know the set of retrieved documents, as assumed. The Recall (R) of the search is the fraction of relevant documents that are retrieved. The simple formula reads:

$$R = \frac{|ret \cap rel|}{|rel|} \qquad (2)$$

As explained above, Recall can be calculated by inspection of the entire documentary system or by performing a sample and deduce by this a confidence interval for the estimation of the fraction of relevant documents for the searched topic. Multiplying this by the database size yields an estimate for the denominator of (2). Other measures (such as Fallout and Miss) can be defined and studied as indicated in Egghe (2007,2008). We do not use them in this paper.

IR evaluation of a documentary system is classically done through the construction of Precision-Recall curves (R-P curves or P-R curves). In Salton and McGill (1987) there is an explicit example (Chapter 5) which can be summarized as follows. Suppose, in respons to a query, the documentary system yields a sequence of documents, in decreasing order of the expected relevance of the documents for the query (orders differ according to the used documentary system and are frequently applied in search engines). For each positive entire number $n = 1, 2, 3, ...$ we can cut the list after the nth document and consider the first n documents as retrieved. So, for each n as above, we can calculate P and R as above and this yields a set of coordinates (R, P) or (P, R) (dependent of what we choose as abscissa and ordinate). This cloud of points is the basis for a generally decreasing Precision-Recall curve (details are not needed for this paper). This Precision-Recall curve informs us what will be the "price" that we pay on R if we want to increase P or on P if we want to increase R.

In the next section, these IR evaluation tools will be applied to the evaluation of a peer review system of a journal. IR evaluation methods in the theory of link prediction evaluation (e.g. in social networks) have been applied in Popescul and Ungar (2003), Kashima and Abe (2006) and Guns (2009).

The indicators Recall and Precision will be redefined in a peer review system, introducing the terms "Success Rate" replacing Precision and "Hitting Rate" replacing Recall. We will explain in detail how these indicators are calculated, replacing retrieved documents by "accepted documents" and by replacing relevant documents by "qualified documents". Exact definitions of these notions will be given. We will also describe in full detail how these indicators are calculated in the example of the journal *Angewandte*

Chemie - International Edition. The analogues of the Recall-Precision curves in IR are established on which we base ourselves for evaluative comments on the peer review system of this journal (underlying that the same comments can be given for any peer reviewed journal). It must, however, be stressed that the used data from *Angewandte Chemie - International Edition* (AC-IE, obtained and described in Bornmann and Daniel (2010a)) are very rare (and even unique as far as we are aware) since they are derived from a non-publically available editorial process.

AC-IE is an international journal of the German Chemical Society (Gesellschaft Deutscher Chemiker (GDCh), Frankfurt am Main, Germany) and is published by Wiley-VCH (Weinheim, Germany). It introduced a peer review system in 1982, primarily in conjunction with one of the document types published in the journal, "communications," which are short reports on work in progress or recently concluded experimental or theoretical investigations. AC-IE is one of the prime chemistry journals in the world, with a higher annual Journal Impact Factor (JIF) than the JIFs of comparable journals (at 10.879 in the 2008 Journal Citation Reports (JCR) Science Edition). What the editors of AC-IE look for most of all is the best research in the different areas of chemistry. Submissions that reviewers deem to be of high quality are selected for publication: For most submissions, a manuscript is published only if two external reviewers rate the results of the study reported in the manuscript as important and also recommend publication in the journal (Bornmann and Daniel (2010b)).

Evaluation of a peer review process and example from the journal *Angewandte Chemie - International Edition*

Suppose we have a fixed peer-reviewed journal (any peer-reviewed journal but in our example this will be the AC-IE). Suppose, in a fixed time period, we have a set Ω of submitted papers. In the IR interpretation this is our documentary system. The set of accepted papers (possibly after revisions) is well-known and hence they are the retrieved papers in the IR interpretation. Notationally: $A = ret$. The rejected papers are the complement of the set of accepted papers in the set of submitted papers: $\Omega - A$. Since we are aware (cf. the description in the Introduction) that not all accepted papers are of high citation impact (after publication) which we will call qualified (Q) (exact definitions to follow) we also have the following set: the set of all qualified papers Q (we use the same notation for "qualified papers" and the set of qualified papers). In the IR interpretation this replaces the set of relevant documents: $Q = rel$.

The analogy with IR is clear: the accepted papers are known and are linked with the known set in IR of retrieved papers. The qualified papers are the ones we want and are hence linked with the wanted relevant documents in the documentary system. Finally, the

set of non-qualified papers is denoted by N and it is the complement of the set of qualified papers in the set of submitted papers: $N = \Omega - Q$.

The analogues of recall and precision in IR hence are as follows. Denote by AQ the number of accepted papers that are qualified. This is a short notation (also used in Bornmann and Daniel (2010a)) for $|A \cap Q|$. Denote by AN the number of accepted papers that are non-qualified: $AN = |A \cap N|$. Denote by RQ the number of rejected (but published elsewhere) papers that are qualified: $RQ = |R \cap Q|$. Finally we also have $RN = |R \cap N|$, the number of rejected (but published elsewhere) papers that are non-qualified. The Success rate S was already defined in Bornmann and Daniel (2010a) and is the equivalent of precision P in IR:

$$S = \frac{AQ}{AQ + AN} \tag{3}$$

In this paper we define (new indicator) the Hitting rate H, being the equivalent of R in IR:

$$H = \frac{AQ}{AQ + RQ} \tag{4}$$

The following indicators were also introduced in Bornmann and Daniel (2010a):

$$G = \frac{AQ + RQ}{AQ + RQ + AN + RN} \tag{5}$$

which is the equivalent in IR of "Generality of the topic" (on which one wants documents to retrieve) and

$$G' = \frac{AQ + AN}{AQ + RQ + AN + RN} \tag{6}$$

which is the equivalent in IR of "Generality of the query" (query derived from the topic).

For constructing the S-H curves (being the equivalent of the P-R curves in IR), to be used in the evaluation of the peer review system (as analogue of the evaluation of an IR system) we need that the papers are ordered in decreasing order of the "chance" of acceptance of the papers. This will be described now on the concrete example of the journal AC-IE. We used a dataset described in Bornmann and Daniel (2010a) that consists of 615 papers submitted to AC-IE in the year 2000 of which review scores of 2 reviewers are available. The 615 papers are ranked in decreasing order of these scores (ratings of the reviewers): 12, 11, ... 1, 0 (12 is the best judgment). Within the same rating, we have arranged the papers in two ways: one is the "random" way by means of the manuscript identifier of the paper in the peer review system. Another way is, within the same rating: order the papers in decreasing order of their qualification score. This will be described now. Papers in Ω are the submitted papers to AC-IE. Now we consider only the accepted or rejected (but published elsewhere) papers. One checks the citation score of all these papers (number of citations to these papers) divided by the mean number of citations of all publications in a corresponding subject area (more details in Bornmann and Daniel (2010a)). In this way one can compare the relative citation scores. A score of 1 corresponds to this field average and a score of 1.5 is considered "high impact" and are defined in Bornmann and Daniel (2010a) (and also here): "qualified" (Q) (see here also van Raan, 2004). So for each paper in our list of accepted or rejected (but publishes elsewhere) we know whether it is qualified or not.

As already explained above we have two ordered lists:

- one in decreasing order of reviewers' ratings and, if the ratings are the same, in a random way,

- one in decreasing order of reviewers' ratings and, if the ratings are the same, in decreasing order of their Q-score.

Each of these lists are treated in the same way in order to produce the S-H curve (so we obtain two curves): for each $n = 1, 2, 3, ...$, we only consider the first n papers. We consider these n papers as "accepted" (A) and for each of these papers we can determine if they are "qualified" (Q) or not. Hence, for each n we obtain a value for S and H by using formulae (3) and (4). This gives us the S-H curve for the two lists. Their graphs are depicted in Figs. 1 and 2.
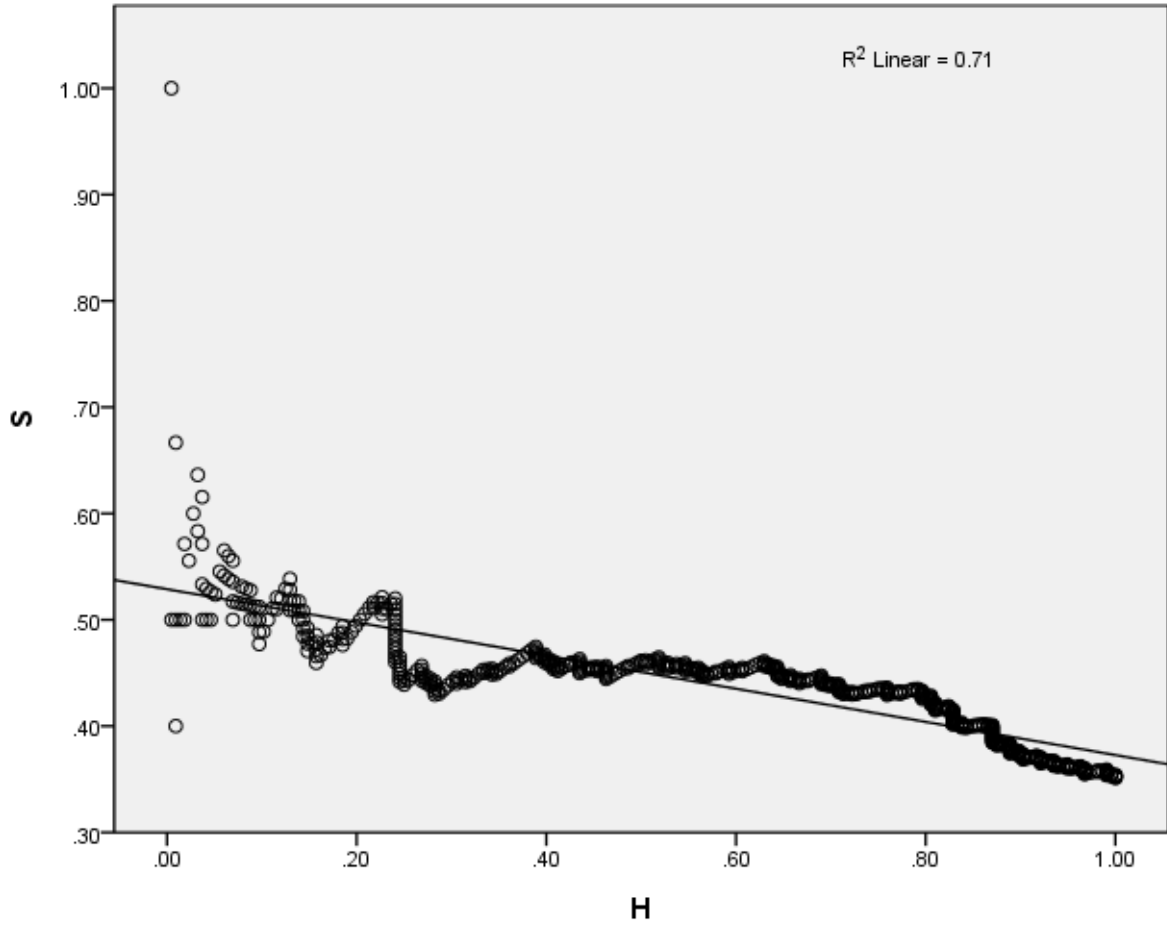
Fig. 1 Scatterplot of S and H ($n = 615$ accepted or rejected, but published elsewhere papers; sorted in decreasing order of reviewers' ratings and sorted randomly – by manuscript identifier – in case of equal reviewers' ratings)
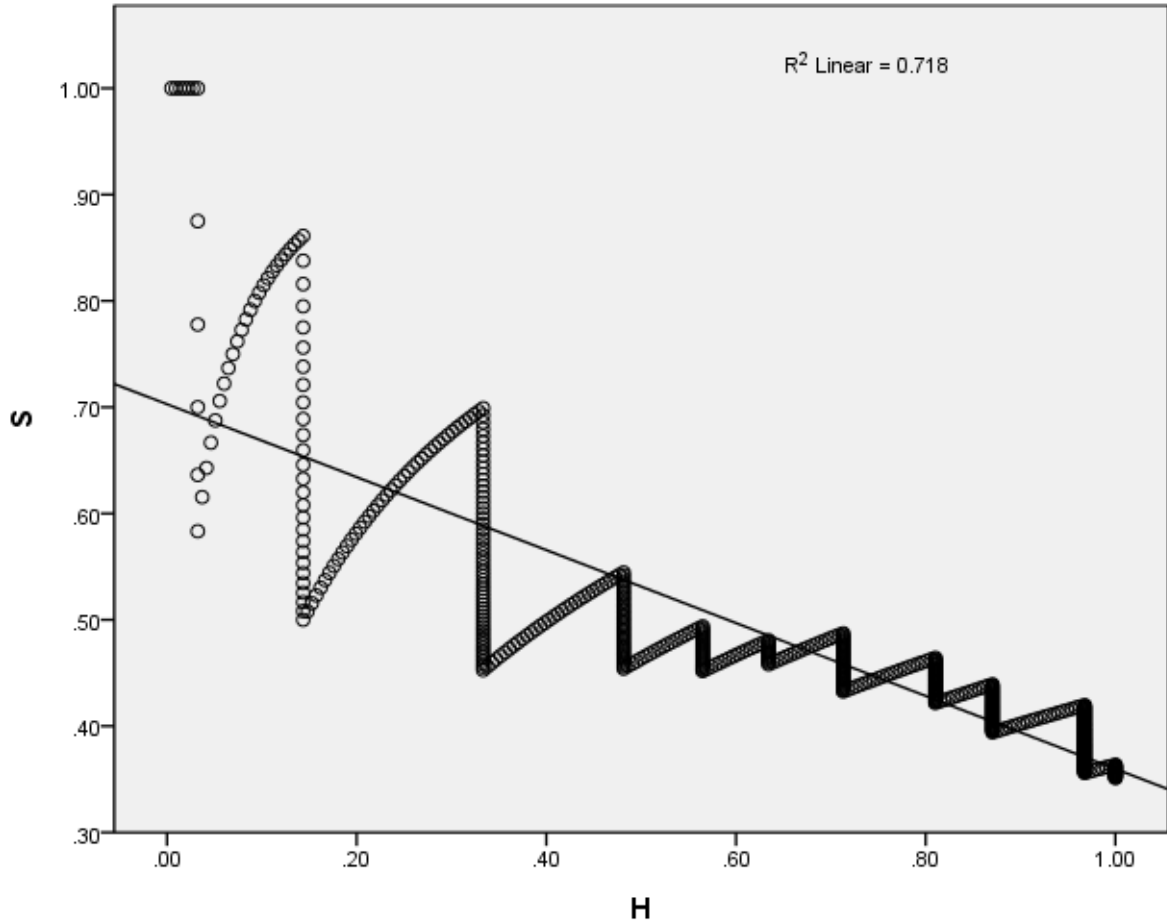
Fig. 2 Scatterplot of S and H with fitted line ($n = 615$ accepted or rejected, but published elsewhere papers; sorted in decreasing order of reviewers' ratings and sorted by normalized citation impact in case of equal reviewers' ratings)

As in IR both clouds of points have a decreasing tendency. The first graph is smoother than the second one since for the first graph we use random order in case of equal reviewer's ratings while in the second one each decrease occurs when, within an equal reviewers' rating, we go from qualified to non-qualified papers. As in IR, these decreasing curves indicate the "price" we have to pay in Hitting rate when we want the Success rate to be higher and vice-versa. This is rather "bad news" for any peer review system since, if we want to have "sufficient" qualified papers (this are the papers we are seeking) we will "automatically" have to publish non-qualified papers as well due to the peer review system of the peer reviewed journal! With this IR-like approach this is shown very clearly.

Similar results to the AC-IE peer review system could be found by the calculation of the extent of type I and type II errors of the editors' selection decisions. Bornmann and

Daniel (2009a) show that the decisions regarding 15% of the papers demonstrate a type I error (accepted papers that did not perform as well as or worse than the average rejected, but published elsewhere paper). Moreover, the decisions regarding 15% of the papers concerned a type II error (rejected papers that performed equal to or above the average accepted paper).

Conclusions and suggestions for further research

In this paper we noticed that the classical evaluation techniques in IR (using Precision and Recall and their P-R curves) can be used for the evaluation of a peer review system of a journal, provided we have data on reviewers' scores and on later citation impact of papers that are accepted for publication in a journal or rejected for this journal but published elsewhere.

Practical data on 615 submitted papers to AC-IE yield reviewers' ratings and qualified scores based on later relative citation impact. As in IR processes, we conclude that, due to the imperfect peer review process (based on human evaluation), if we want to publish a high amount of qualified papers (the ones we seek), one will also accept several non-qualified papers as well. The effect is clearly illustrated via the decreasing S-H curves (Success rate - Hitting rate curves) being the equivalent of P-R curves (Precision - Recall curves) in IR.

It is clear that such data are hard to get since they belong to the confidential area of an Editor-in-Chief (EIC) regarding the list of submitted papers and the reviewers' ratings. In addition to this one need to "follow" these submitted papers (accepted or rejected (but published elsewhere)) to establish their citation impact on which we then can base ourselves to determine the qualifiedness or un-qualifiedness of these papers. We hope that in the future similar studies of evaluation of peer review processes will be possible. This can only be realized in close collaboration with EICs of peer reviewed journals who work with a rating system for reviewers and on the condition that the concrete details about the submitted papers are kept confidential within such a research.

References

L. Bornmann (2011). Scientific peer review. Annual Review of Information Science and Technology, 45, 199-245.

L. Bornmann and H.-D. Daniel (2009a). Extent of type I and type II errors in editorial decisions: a case study on *Angewandte Chemie - International Edition*. *Journal of Informetrics, 3*, 348-352.

L. Bornmann and H.-D. Daniel (2009b). The luck of the referee draw: the effect of exchanging reviews. Learned Publishing 22(2), 117-125.

L. Bornmann and H.-D. Daniel (2010a). The manuscript reviewing process - empirical research on review requests, review sequences and decision rules in peer review. Library & Information Science Research, 32(1), 5-12.

L. Bornmann and H.-D. Daniel (2010b). The usefulness of peer review for selecting manuscripts for publication: A utility analysis taking as an example a high-impact journal. PLoS ONE, 5(4), e11344.

L. Egghe (2007). Existence theorem of the quadruple (P,R,F,M): Precision, Recall, Fallout and Miss. Information Processing and Management 43(1), 265-272.

L. Egghe (2008). The measures precision, recall, fallout and miss in function of the number of retrieved documents and their mutual interrelations. Information Processing and Management 44(2), 856-876.

L. Egghe (2010). Study of some Editor-in-Chief decision schemes. Annals of Library and Information Studies 57(3), 184-195.

R. Guns (2009). Generalizing link prediction: collaboration at the University of Antwerp as a case study. In: A. Grove (ed.), ASIST 2009: Proceedings of the 72nd ASIS&T Annual Meeting, Vancouver, B.C., Canada, November 6-11, 2009. ASIS&T: Silver Spring, Md., USA.

H. Kashima and N. Abe (2006). A parameterized probabilistic model of network evolution for supervised link prediction. In: Proc. of the 2006 IEEE International Conference on Data Mining (ICDM 2006), 340-349.

A. Popescul and L.H. Ungar (2003). Structural logistic regression for link analysis. In S. Džeroski et al (eds.). Proc. of the 2nd Workshop on Multi-Relational Data Mining (MRDM-2003), 92-106.

A. F. J. van Raan (2004). Measuring science. Capita selecta of current main issues. In H. F. Moed, W. Glänzel & U. Schmoch (Eds.), Handbook of quantitative science and technology research. The use of publication and patent statistics in studies of S&T systems (pp. 19-50). Dordrecht, The Netherlands: Kluwer Academic Publishers.

G. Salton and M.J. McGill (1987). Introduction to modern Information Retrieval, McGraw-Hill, Auckland, New Zealand.

D.M. Schultz (2010). Are three heads better than two? How the number of reviewers and editor behavior affect the rejection rate. Scientometrics 84(2), 277-292.

A.C. Weller (2001). Editorial Peer Review. Its Strenghts and Weaknesses. ASIST Monograph Series, Information Today, Inc., Medford, New Jersey, USA.