A Citation Source Study In Artificial Intelligence

[Technical Report]
HAM-LB-TR-14-01

Robo G. Grey
Ham Laboratories
H.L. Ham Pte Ltd
Hamburg, Hamerica 03977
roboggrey@glys.com

ABSTRACT

In a competitive academic landscape, the perceived influence of their publications has long been a major factor in determining the career success of researchers. Further, not all publication platforms are created equal, with certain venues unavoidably generally acknowledged as more prestiguous than others. With reference to the most recent proceedings of eight recognized conferences in the field of artificial intelligence, we investigate the predictive power of the reference sources of their papers in determining their actual venue, and estimate that papers can be sorted by tier with a confidence level of over 84% from their citations alone.

Categories and Subject Descriptors

H.3.0 [**Information Systems**]: Information Storage and Retrieval—*General*

General Terms

Measurement

Keywords

Citation Analysis, Bibliometrics, Bag-of-words

1. INTRODUCTION

For all their longstanding inadequacies, such as the encouragement of a "publish or perish" culture that incentivizes piecemeal articles and quantity of output [37, 14], citation counts and their derivative statistics remain one of the few widely-used quantitative measures of a researcher's reputation [21]. For example, the h-index [25], a hybrid metric that considers both the productivity and impact of an individual, has become increasingly important in tenure decisions [13, 17], and moreover has been found to correspond fairly well with peer judgment [39]. This is analogous to the *impact factor* of journals and conferences [23, 22], and is a trend that has only become more prevalent with online indexing on comprehensive databases such as *Google Scholar*[32].

As with any well-defined system populated by intelligent autonomous agents, however, such criteria are vulnerable to gaming, ranging from the comparatively innocuous and sometimes-relevant self-citation, to outright blatant (and oft-successful) manipulation [9, 30]. A self-interested author can (perhaps unconsciously) inflate their own h-index significantly, and with little effort [11]. Of course, the most egregious such offences are likely to be discovered eventually; therefore, for this paper, we concentrate on a more subtle question – do publications in selective venues tend to reference similarly-selective venues (and in particular, themselves), as well as the inverse: does the inclusion of high-quality sources correlate with acceptance in such venues? This is doubly important as relevant work is often uncited for various practical reasons [31], and also because of the extremely fine margins involved in achieving acceptance [8].

1.1 Related Work

Various studies based on citation analysis have been used to assess journals [26, 40, 16, 28, 10], despite concerns over unreliable data [6, 18]. Rahm and Thor analyzed citation frequencies for five prominent database venues [35], and concluded that conferences had a substantially higher citation impact than journals in the field (though possibly not for computer science in general [19]). Tsai found that journal rankings can vary greatly depending on the particular metric chosen [38], while Sicilia et. al found that impact factors are however largely consistent across citation databases [36]. Hussain and Swain identified a trend towards collaborative research and multi-authored papers, from the top-ranked computer science journals according to *ScienceDirect* [27]. Freyne et. al confirm the relative status of conferences vis-a-vis journals in computer science by impact factor [20], and also include artificial intelligence venues. However, they do not evaluate the sources of these citations, nor attempt to discover their possible effects.

From our literature review, we believe that our paper is the first to explicitly confront the issue of the *source* of references, and how they pertain to the publication tier of the work they belong to, in citation analysis. We hope that this will motivate discussion as to the strength and significance of the well-known Matthew effect [33] accruing from venue prestige, as well as inspire further investigations utilizing data from other disciplines.

Submitted to Journal of Hamster Information Processing (Feb 2014)

2. DATA AND RESULTS

2.1 Conferences

Judgment on the quality of conferences has always been fraught with sensitivity, and yet it is scarcely deniable that some are viewed as "more prestigious" than others. For understandable reasons, there are few if any authoritative statements on the subject. Thus, to identify a suitable selection of conferences in artificial intelligence and related subfields, we resort to a combination of unofficial lists [2, 3, 4][1], which we consider as peer feedback, and the Field Rating (field-specific impact factor) metric from *Microsoft Academic Search* [5]. We then divided the conferences into three categories (tiers), as follows:

- **Premier** – within the top ten on *Microsoft Academic Search* for its subfield by Field Rating, and in the top-most tier of at least one referenced list
- **Ranked** – within the top 50% on *Microsoft Academic Search* for its subfield by Field Rating, and in the top two tiers of at least one referenced list
- **Regular** – all other conferences on either *Microsoft Academic Search*, or at least one of the referenced lists, for the relevant subfields

This process returned 182 *premier* conference labels, 1319 *ranked* conference labels, and 2111 *regular* conference labels. To cater for common variations, this was expanded to a final collection of 206 *premier*, 1391 *ranked*, and 2162 *regular* labels for matching purposes. As seen from Table 1, there is a broad consensus on the tier of a conference, among the various sources – dissenting opinions are rare. Therefore, we are fairly confident that the tier divisions are reliable.

	Premier	Ranked	Regular
Premier	182	25	13
Ranked		1294	64
Regular			2034
Total	182	1319	2111

Table 1: Confusion Matrix between Tiers

We also observe that there are many more *ranked* and *regular* conferences than *premier* ones. However, this does not necessarily imply that the distribution of individual papers is also as lopsided, since many *regular* conferences have relatively few publications. Again utilizing data from *Microsoft Academic Search* on our set of conferences, we can obtain a rough estimate of the number of publications represented in each category/tier (as used in Section 2.2).

From the *premier* category, and under the Artificial Intelligence/Machine Learning subfields, we selected the AAAI, ICML, NIPS and UAI conferences to extract references from, partly because their proceedings are freely downloadable from their respective webpages. From the *regular* category, we randomly selected four conferences, which we shall refer to as REG1 to REG4. Their proceedings were obtained from the *IEEE Xplore* and *Springer* databases. We consider only the latest available edition (2013) of all conferences. This gave 918 papers from the *premier* conferences, and 168 papers from the *regular* ones.

2.2 References

The next step was to obtain and process the actual references from the PDF documents. For this, we extracted their plaintext content using Apache PDFBox *ExtractText* [1] from the command line. There has been prior work on reference extraction by Besangi et al. [12], Wellner et al. [41] and Powley et al. [34], among others; we however do not compare our results with theirs, due to our reduced scope in this respect - for example, we do not require in-text citation matching to references.

Conveniently for us, almost all papers began their references section with a "References" header – the few exceptions were manually corrected after detection failure. Since the output from PDFBox retains the formatting of the original document, most references were broken into multiple lines, and therefore had to be reconstructed. Between-reference parsing was straightforward with LaTeX-standard bibliographic styles that utilize square brackets. Otherwise, a heuristic regular-expression based parser (Algorithm 1) was employed to detect the most probable reference line breaks.

Algorithm 1 Algorithm for parsing references

```
buffer[] ← lines(from reference section)
results[] ← $\emptyset$
$N \leftarrow 0$
$L \leftarrow$ average(*lines.length*)
for $i = 0$ to *buffer.length* do
    $C \leftarrow$ *commas in line*
    if $i > 0$ AND $N > 0$ AND *buffer[i].length*$> 0.8L$ AND *buffer[i].isName* AND (*buffer[i-1].length*$< L$ OR $C >=$ 3 OR *buffer[i-1].endswith(.)*) then
        $N \leftarrow 0$
        results[]++
    else
        $N \leftarrow 1$
    end if
    results[] ← *buffer[i]*
end for
return results[]
function *str.isName*
tokens[] ← *str.split(whitespace)*
$M \leftarrow$ min(*tokens.length*,6)
$S \leftarrow 0$
for $i = 0$ to *buffer.length* do
    $U \leftarrow$ *buffer[i-1].lastchar*
    if *tokens[i].length*<= 3 AND (U="." OR U="," OR U=";") then
        $S \leftarrow S + 2$
    else if tokens[i]="and" OR U="." OR U="," OR U=";" OR (i>0 AND *tokens[i].firstchar.isUpperCase*) then
        $S \leftarrow S + 1$
    else if tokens[i].containsDigit then
        $S \leftarrow S - 1$
    end if
end for
$S \leftarrow \frac{S}{M}$
if $S > \frac{1}{3}$ then
    return 1
else
    return 0
end if
end function
```

[1] A* and A considered as a single (top) tier for CORE

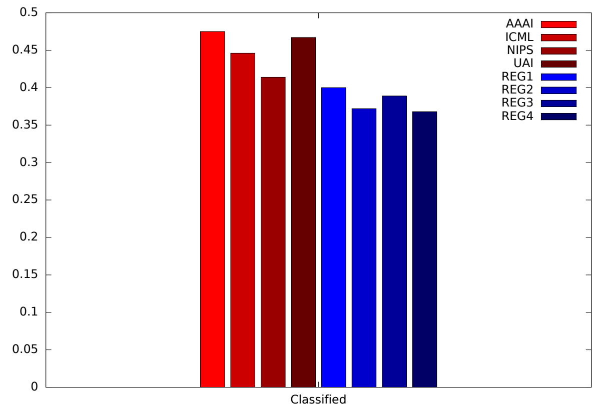


Figure 1: Classifiable References

	Reference Source							
	AAAI	ICML	NIPS	UAI	REG1	REG2	REG3	REG4
AAAI	**235**	124	145	71	0	0	0	0
ICML	34	563	**733**	95	0	0	0	0
NIPS	59	493	**777**	117	0	0	0	0
UAI	32	94	**167**	130	0	0	0	0
REG1	9	**12**	7	3	0	0	0	0
REG2	**2**	0	0	0	0	1	0	1
REG3	6	4	2	1	0	0	**8**	0
REG4	**33**	2	8	4	0	0	0	0

Table 2: Reference Matrix (Absolute Counts)

This parsing was not perfect due to the existence of inherently ambiguous tokens such as editorships and publisher names, and sometimes inconsistent styling even within a single references section. Remaining mistakes were manually corrected as far as was possible. This produced 21368 (avg. 23.3/paper) references from the *premier* category conferences, and 2936 references (avg 17.5/paper) from the *regular* category conferences.

For each reference, we attempt to automatically match it with the category that its source belongs to. We achieve this by expression-matching substrings of each reference against the conference labels, from the longest (most specific) to shortest label. Of course, this is not possible when the source is not from one of the conferences in our labelled set, for instance when they are from a conference in a non-Computer Science area, a journal, or some other modality; for such references, we refer to them as "unclassified". About 40% of the references from each conference can be matched with a registered conference label (i.e. "classified") this way (see Figure 1). From this point on, we consider only "classified" references in our analysis.

As illustrated in Figure 2, the reference sources of *premier* and *regular* conference papers are extremely well-separated – both overwhelmingly tended to cite papers from a similar conference tier to themselves. A more detailed breakdown among the eight conferences that were specifically studied is displayed in Table 2, which shows the raw total number of references that each conference had from each other conference. A normalised version, where the reference count from each conference is scaled by the historic number of publications for that conference (as taken from *Microsoft Academic Search*) is given in Table 3. From this, it can be seen that the citation of past papers from the same conference is a common, and possibly expected, practice, at least among *premier* conferences.

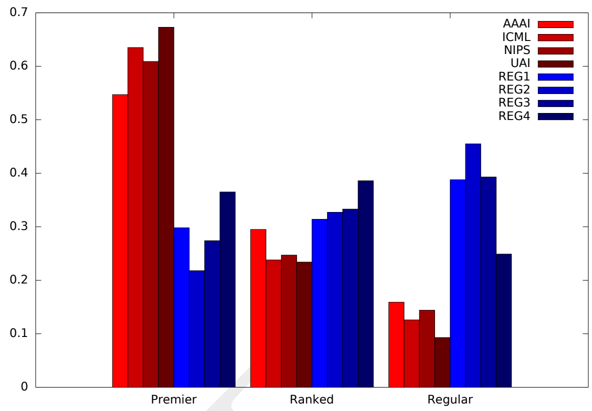


Figure 2: Category Distribution of Reference Sources

	Reference Source							
	AAAI	ICML	NIPS	UAI	REG1	REG2	REG3	REG4
AAAI	4.35	**5.03**	3.38	4.52	0.00	0.00	0.00	0.00
ICML	0.63	**22.84**	17.10	6.04	0.00	0.00	0.00	0.00
NIPS	1.09	**20.00**	18.13	7.44	0.00	0.00	0.00	0.00
UAI	0.59	3.81	3.90	**8.27**	0.00	0.00	0.00	0.00
REG1	0.17	**0.49**	0.16	0.19	0.00	0.00	0.00	0.00
REG2	0.04	0.00	0.00	0.00	0.00	**0.98**	0.00	0.46
REG3	0.11	0.16	0.05	0.06	0.00	0.00	**1.47**	0.00
REG4	**0.61**	0.08	0.19	0.25	0.00	0.00	0.00	0.00

Table 3: Reference Matrix (Relative Counts)

2.3 Predictive Power

From the data presented, it appears possible that meaningful correlations can be found between the characteristics of a paper's collection of references, and its publication venue. We formulate two questions, in order of difficulty: Firstly, how well can we determine the probable *tier* of a paper from its references alone? Secondly, how well can we determine the probable *conference* it belongs to, again from its references alone? To answer this, we examine the results that can be achieved from various combinations of representations of the reference sets with classifiers.

2.3.1 *Representations*

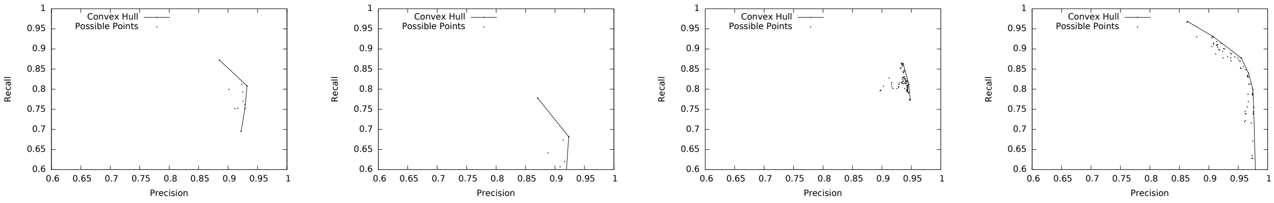
- **R1** – four features, corresponding to the percentage of "classified" references, and the percentage of *premier*, *ranked* and *regular* sources among the "classified" references

- **R2** – 3667 features, corresponding to the four features from **R1**, and a (large, sparse) vector where each known conference label is hashed to a unique vector index. This is in effect a "bag of words" model, which has seen prominent usage in document analysis [29] and computer vision. For citation analysis in particular, references have been studied for plagarism detection [24] and content clustering using their context [7], but not specifically for venue prediction

2.3.2 *Classifiers*

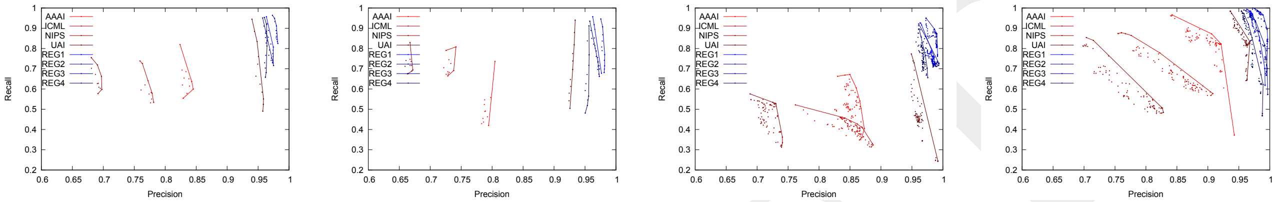
- **kNN** – the k-nearest neighbour classifier, with Euclidean distance metric

- **SVM** – the radial basis function support vector machine classifier. We use the LIBSVM[15] implementation, which includes support for sparse vectors



(a) kNN-R1 (b) kNN-R2 (c) SVM-R1 (d) SVM-R2

Figure 3: Tier prediction results, precision-recall curves



(a) kNN-R1 (b) kNN-R2 (c) SVM-R1 (d) SVM-R2

Figure 4: Conference prediction results, precision-recall curves

2.3.3 *Tier Prediction*

We perform five-fold cross-validation on the available reference data as follows: firstly, for each conference, we randomly divide its papers into five folds, to ensure that the set of papers from which the references used in training are obtained from are always independent of those used for evaluation. The feature vectors of references from *premier* conferences are labelled as “1”, and those from *regular* conferences as “0”. For each fold, its training data consists of all references from the other four folds. Since there will be many more references from the *premier* class in the training set, the references from the *regular* class will be oversampled such that both classes are equally represented.

Since the performance of both classifiers is dependant on parameters (k for **kNN**, and $\{c, g\}$ for **SVM**), we exhaustively search the respective parameter spaces
$k = \{1, 3, \ldots, 21\}$ and $c = \{2^{-5}, \ldots, 2^{15}\}, g = \{2^{-15}, \ldots, 2^{3}\}$, and plot the achievable convex hulls obtained (Figure 3). It is observed that SVM with “bag of words” data (SVM-R2) produces the best results. F-measure is maximized with SVM-R2 (Figure 3(d)) at a level of 0.917; at this point, sensitivity for *premier* papers is 0.921 (839/918), while sensitivity for *regular* papers is 0.571 (96/168). Since maximizing f-measure is biased towards identifying *premier* papers due to their larger number, we also consider the Balanced Correct-classification Rate (BCR), which is defined as the product of the sensitivities of each class. BCR is maximized at a level of 0.710; at this point, sensitivity for *premier* papers is 0.840 (771/918), while sensitivity for *regular* papers is 0.845 (142/168).

2.3.4 *Conference Prediction*

We adopt the same methodology for conference prediction as for tier prediction. For each conference, for each representation-classifier combination, we train a model to predict whether a paper belongs to that particular conference, or not (i.e. belongs to one of the seven other conferences). This is much more challenging than tier prediction, since each conference has to be distinguished from many other conferences covering the same general subject area, three of them moreover of the same tier (Figure 4). Despite this, better-than-chance prediction is still possible, as detailed in Table 4 (SVM-R2; *Pos* represents the sensitivity for the conference, and *Neg* for the other seven conferences):

	Conference Targeted							
	AAAI	ICML	NIPS	UAI	REG1	REG2	REG3	REG4
BCR	0.57	0.48	0.41	0.51	0.60	0.78	0.69	0.54
Pos	0.69	0.79	0.71	0.62	0.85	0.90	0.82	0.80
Neg	0.82	0.62	0.58	0.83	0.70	0.88	0.84	0.67

Table 4: Conference prediction sensitivities (max BCR)

3. CONCLUSIONS AND FUTURE WORK

Our findings suggest that even *completely disregarding* a paper’s content, we can gain a pretty good idea of the quality of conference that it was (will be?) presented at, from its references alone. While this phenomena is probably not causative – a poor paper that takes pains to cite only from top conferences[2] obviously remains unlikely to be accepted at those same conferences as-is – its existence is still worth explaining. Specialization may play a part, though all conferences studied overlapped heavily in scope. Also, if the widespread belief that papers are submitted to progressively less-selective conferences until accepted is true, then there is little prior justification for why the characteristics of their reference sources would differ so significantly.

This is however but an exploratory work, and there remains much space for examination; does this effect hold with more conferences, with conferences from other fields, and with journals? Would considering the tier of referenced journals too, improve predictive power further? How strongly might the *choice* of referenced sources affect acceptance?

4. ACKNOWLEDGMENTS

This work was supported in part by a kind human, who shall remain unnamed for political reasons.

[2] as this very paper does

5. REFERENCES

[1] Apache pdfbox. `http://pdfbox.apache.org/`.

[2] Computer science conference rankings. `http://webdocs.cs.ualberta.ca/~zaiane/htmldocs/ConfRanking.html`.

[3] Computer science department conference rankings. `http://www.comp.nus.edu.sg/~wang06/SoC%20Conference%20Ranking.htm`.

[4] Core computer science conference rank. `http://lipn.univ-paris13.fr/~bennani/CSRank.html`.

[5] Microsoft academic search. `http://academic.research.microsoft.com/`.

[6] D. Adam. Citation analysis: The counting house. *Nature*, 415(6873):726–729, 2002.

[7] B. Aljaber, N. Stokes, J. Bailey, and J. Pei. Document clustering of scientific texts using citation contexts. *Information Retrieval*, 13(2):101–131, 2010.

[8] T. Anderson. Conference reviewing considered harmful. *ACM SIGOPS Operating Systems Review*, 43(2):108–116, 2009.

[9] D. N. Arnold and K. K. Fowler. Nefarious numbers. *Notices of the AMS*, 58(3):434–437, 2011.

[10] S. J. Barnes. Assessing the value of is journals. *Communications of the ACM*, 48(1):110–112, 2005.

[11] C. Bartneck and S. Kokkelmans. Detecting h-index manipulation through self-citation analysis. *Scientometrics*, 87(1):85–98, 2011.

[12] D. Besagni, A. Belaïd, and N. Benet. A segmentation method for bibliographic references by contextual tagging of fields. In *Document Analysis and Recognition, 2003. Proceedings. Seventh International Conference on*, pages 384–388. IEEE, 2003.

[13] L. Bornmann, R. Mutz, and H.-D. Daniel. Are there better indices for evaluation purposes than the h index? a comparison of nine different variants of the h index using data from biomedicine. *Journal of the American Society for Information Science and Technology*, 59(5):830–837, 2008.

[14] F. Casati, F. Giunchiglia, and M. Marchese. Publish and perish: why the current publication and review model is killing research and wasting your money. *Ubiquity*, 2007(January):3, 2007.

[15] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[16] C. H. Cheng, A. Kumar, J. G. Motwani, A. Reisman, and M. S. Madan. A citation analysis of the technology innovation management journals. *Engineering Management, IEEE Transactions on*, 46(1):4–13, 1999.

[17] R. Ciriminna and M. Pagliaro. On the use of the h-index in evaluating chemical research. *Chemistry Central Journal*, 7(1):132, 2013.

[18] M. E. Falagas, E. I. Pitsouni, G. A. Malietzis, and G. Pappas. Comparison of pubmed, scopus, web of science, and google scholar: strengths and weaknesses. *The FASEB Journal*, 22(2):338–342, 2008.

[19] M. Franceschet. The role of conference publications in cs. *Communications of the ACM*, 53(12):129–132, 2010.

[20] J. Freyne, L. Coyle, B. Smyth, and P. Cunningham. Relative status of journal and conference publications in computer science. *Communications of the ACM*, 53(11):124–132, 2010.

[21] E. Garfield. Is citation analysis a legitimate evaluation tool? *Scientometrics*, 1(4):359–375, 1979.

[22] E. Garfield. The history and meaning of the journal impact factor. *Jama*, 295(1):90–93, 2006.

[23] E. Garfield et al. Citation analysis as a tool in journal evaluation. American Association for the Advancement of Science, 1972.

[24] B. Gipp and J. Beel. Citation based plagiarism detection: a new approach to identify plagiarized work language independently. In *Proceedings of the 21st ACM conference on Hypertext and hypermedia*, pages 273–274. ACM, 2010.

[25] J. E. Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences of the United States of America*, 102(46):16569–16572, 2005.

[26] C. W. Holsapple, L. E. Johnson, H. Manakyan, and J. Tanner. A citation analysis of business computing research journals. *Information & Management*, 25(5):231–244, 1993.

[27] A. Hussain and D.-K. Swain. A citation analysis of top research papers of computer science. *International Research: Journal of Library and Information Science*, 1(2), 2011.

[28] P. Katerattanakul, B. Han, and S. Hong. Objective quality ranking of computing journals. *Communications of the ACM*, 46(10):111–114, 2003.

[29] R. Kosala and H. Blockeel. Web mining research: A survey. *ACM Sigkdd Explorations Newsletter*, 2(1):1–15, 2000.

[30] E. D. Lopez-Cozar, N. Robinson-Garcia, and D. Torres-Salinas. Manipulating google scholar citations and google scholar metrics: Simple, easy and tempting. *arXiv preprint arXiv:1212.0638*, 2012.

[31] M. H. MacRoberts and B. R. MacRoberts. Problems of citation analysis: A study of uncited and seldom-cited influences. *Journal of the American Society for Information Science and Technology*, 61(1):1–12, 2010.

[32] L. I. Meho. The rise and rise of citation analysis. *arXiv preprint physics/0701012*, 2006.

[33] R. K. Merton. The matthew effect in science. *Science*, 159(3810):56–63, 1968.

[34] B. Powley and R. Dale. Evidence-based information extraction for high accuracy citation and author name identification. In *Large Scale Semantic Access to Content (Text, Image, Video, and Sound)*, pages 618–632, 2007.

[35] E. Rahm and A. Thor. Citation analysis of database publications. *ACM Sigmod Record*, 34(4):48–53, 2005.

[36] M.-A. Sicilia, S. Sánchez-Alonso, and E. García-Barriocanal. Comparing impact factors from two different citation databases: the case of computer science. *Journal of Informetrics*, 5(4):698–704, 2011.

[37] P. Smith. *Killing the spirit: Higher education in America.* ERIC, 1990.

[38] C.-F. Tsai. Citation impact analysis of top ranked computer science journals and their rankings. *Journal of Informetrics*, 8(2):318–328, 2014.

[39] A. F. Van Raan. Comparison of the hirsch-index with standard bibliometric indicators and with peer judgment for 147 chemistry research groups. *scientometrics*, 67(3):491–502, 2006.

[40] R. J. Vokurka. The relative importance of journals used in operations management research a citation analysis. *Journal of Operations Management*, 14(4):345–355, 1996.

[41] B. Wellner, A. McCallum, F. Peng, and M. Hay. An integrated, conditional model of information extraction and coreference with application to citation matching. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 593–601, 2004.