Deconstructing the collaborative impact: Article and author characteristics that influence citation count

Lori A. Hurley
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign
501 E. Daniel St.
Champaign, IL 61820-6211

Andrea L. Ogier
Center for Digital Research and Scholarship
University Libraries
Virginia Tech
P.O Box 90001
Blacksburg, VA 24062-90001

Vetle I. Torvik
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign
501 E. Daniel St.
Champaign, IL 61820-6211

ABSTRACT

It is well known that collaborative papers tend to receive more citations than solo-authored papers. Here we try to identify the subtle factors of this collaborative effect by analyzing metadata and citation counts for co-authored papers in the biomedical domain, after accounting for attributes known to be strong predictors of citation count. Article-level metadata were gathered from 98,000 PubMed article records categorized with the term *breast neoplasm*, a topic offering longevity and relevance across biomedical subdisciplines, and yielding a relatively large sample size. Open access citation data was obtained from PubMed Central (PMC). Author-level attributes were encoded from disambiguated author name data in PubMed and appended as article-level attributes of collaborations. A logistic regression model was built to assess the relative weights of these factors as influences on citation counts. As expected, the journal and language of the paper were the strongest predictors. The significance of the number of authors diminished after accounting for other attributes. Some of the more subtle predictors included the group's highest h-index, which was positively correlated, while the diversity of author h-indices, minimum professional age, and author's total unique collaborators were negatively correlated. These observations indicate that smaller collaborations composed of early superstars – young, rapidly successful researchers with relatively high and similar h-indices – may be at least as influential in biomedical research as larger collaborations with different demographics. While minimum h-index was important, the first author's h-index was insignificant, underscoring the importance of the middle authors' publishing history. The gender diversity outcomes suggest that mixed groups may be ideal, and further research in this area is indicated.

Keywords

Bibliometrics, citation analysis, collaboration, impact.

INTRODUCTION

Over the last several decades, there has been a strong shift towards collaborative research in the sciences, much attributed to lowered technological barriers (Abramo, D'Angelo, & Di Costa, 2009). Wuchty, Jones, and Uzzi (2007) indicate that team sizes have nearly doubled, from 1.9 to 3.5 authors per paper and 1.7 to 2.3 inventors per patent. Ioannidis (2008) indicates the average number of co-authors has increased into the range of 7 to 10. Adams, Black, Clemmons, and Stephan (2005) found that team size for a scientific paper increased 50% between 1981 and 1999. The shift from the traditional model of independent authorship is well documented, begging the question of how increased collaboration has affected the quality of scientific research and/or scientific publication. While "collaboration" in practice may signify a breadth and depth of team research beyond joint publication, it is considered to mean co-authorship for the purposes of this study.

Traditionally, scientific impact has been measured through citation indices. There has been much recent controversy about the bibliometric effectiveness of citation counts as an assessment of impact (Hirsch, 2007; Nichols, 2012), with several alternative methods of measurement becoming prevalent, including mean number of citations, total number of citations, and h-index, and even beyond citations to account for things like number of downloads and other alternative metrics (Priem, Taraborelli, Groth & Neylon, 2011). H-index reflects both the researcher's number of publications and the number of citations per publication. While h-indices attempt to correct for the weaknesses of citation indices as a metric, they do not adjust according to collaboration specific factors (Petersen, Riccaboni, Stanley, & Pammolli, 2012). Citation counts continue to serve as a strong compass for tenure and funding decisions (Ioannidis, 2008). Haslam *et al*. (2008) argue, "Citation-based metrics are increasingly used to evaluate researchers, to rank departments and universities, and to assess and advertise the standing of scientific journals." Ioannidis (2008) suggests, "Adoption of metrics that measure and adjust for

co-authorship may offer disincentive against poor authorship practices." The identification of certain collaboration patterns leading to higher citation counts would be considered a significant contribution to bibliometrics and would offer a potential method for normalization of citation numbers in order to arrive at a more accurate tool for impact measurement. Here we leverage an unusually large sample, aiming to identify factors of collaboration that merit further study.

We analyzed several attributes of collaborations as well as particular characteristics of co-authors, seeking patterns that would deepen insights gained from previous studies presenting correlations between collaboration and citation. The objective was to study the relationship between co-authorship and citation numbers, testing previous findings against a larger sample. We hypothesized that both characteristics of collaborations and attributes of co-authors would prove influential in article citation levels.

Literature Review

An examination of the prior research on this topic results in contradictory and incomplete conclusions, likely due to differences in variables across disciplines, in types of collaborations, and in standards used to measure productivity. In this study we addressed some previous limitations by selecting a topic, breast neoplasm, which has a persistent history of relevance across biomedical subdisciplines and offers a relatively large sample size of 98,000 instances of collaboration.

Many studies have focused almost entirely on examining the impact of collaboration size. Several identify a positive correlation between collaboration team size and scientific output (Lee & Bozeman, 2005; Adams *et al*., 2005; Wuchty *et al*., 2007; Fischbach, Putzke, & Schoder, 2011; Gazni & Didegah, 2011; Sooryamoorthy 2009). Sooryamoorthy (2009) in fact states that it is now commonly accepted that co-authorship leads to higher citation rates. Abramo *et al*. (2009) detail the evolution of national policies to incentivize teamwork, based on scientific studies showing increased productivity resulting from collaboration. Lee and Bozeman note that funding agencies "facilitate active research collaboration as part of their funding conditions." However, other authors, such as Bergh, Perry, and Hanke (2006), find that the number of co-authors does not strongly predict citations. Abramo *et al*. (2009) conclude the correlation between the quantity of co-authors and impact to be varied, emerging consistently as strongly positive only in the disciplines of industrial and information engineering (2008). Petersen *et al*. (2011) find a decreasing marginal return as group size increases, which they interpret as indicative of the importance of effective team management. Finally, Haslam and Laham (2009), in their studies of patterns of authorship in social psychology, discover a curvilinear relationship between the proportion of team-authorship and publication impact, indicating that successful scientists should seek a balance of minority roles on team projects and leadership roles on solo works or small collaborations.

Prior research provides significant evidence of the existence of regional bias. Rey-Rocha, Martín-Sempere, Martínez-Frías, & López-Vera (2001) and Gazni and Didegah (2011) find that publications with a higher number of foreign collaborators were not as highly cited. Stremersch and Verhoef (2005) conclude that, despite a recent rise in Chinese scholarship and a corresponding decline in U.S. and Canadian work, publications of international scholars are cited less frequently than that of U.S. academics. However, Sooryamoorthy (2009) finds that internationally collaborated work receives more than double the citations as purely domestic collaborations.

Throughout the research, several other variables are examined. Some studies find institutional prestige did not predict impact (Haslam *et al*., 2008). Others, such as the work of Skilton (2009) in the area of natural science, find institutional prestige to have null impact. Haslam *et al*. (2008) determine the following to be strong predictors: author eminence, having a more senior later author, journal prestige, article length, and number and recency of references. In the same study it is concluded that many other variables -- including author gender, nationality, and topic area – do not predict impact. Skilton (2009) contrarily concludes there is good reason to examine demographic variables, such as author age and nationality, more closely.

Haslam *et al*. (2008) acknowledge that their study was limited to one publication year, and suggest that little work has been done to investigate predictors of impact at the article level. Other authors find similar limitations to their sample size and/or coverage, and there was little consistency in methodologies across the research. Abramo *et al*. (2009) conclude there has been no systemic and exhaustive study. A call for additional research echoes across the body of literature on the relationship between collaboration and productivity.

METHODS

Dataset creation

The biomedical domain was chosen due to the availability of disambiguated author name data (Torvik & Smalheiser, 2009; Torvik et al., 2005). For the purposes of this study, 98,082 article metadata records containing *breast neoplasm* in the Medical Subject Headings (MeSH) were extracted from PubMed using the freely available PubMed e-utils tools (http://www.ncbi.nlm.nih.gov). Limiting by topic served to control for topical effects. The topic breast neoplasm was selected for study due to the breadth and longevity of research on breast cancer, therefore insuring availability of a large sample size as well as range of articles across disciplines. The broad reach of this topic is evidenced through the associated MeSH terms, ranging, for

example, from *public health* to *genetic techniques* to *chemistry*. Citation data for the article set was obtained from PubMed Central. PubMed Central (PMC), hosted by the US National Library of Medicine at the National Institute of Health (NIH/NLM), offers an open access electronic archive of full-text biomedical and life sciences literature and contains approximately two million journal articles (http://www.ncbi.nlm.nih.gov/pmc/). Most of these are represented in PubMed as well, an NIH/NLM database of citations and abstracts for more than 22 million biomedical articles (http://www.ncbi.nlm.nih.gov/pubmed).

Python (www.python.org) was used for all data processing. To create the dataset for this project, it was necessary to aggregate metadata from separate files containing author/co-author, article, publication year, gender identification, and inventor identification data. Each instance in the resulting dataset that was used for analysis represents a single collaboration on a particular publication. Each record contains information about the article itself and characteristics of the authors in the collaboration. Article characteristics include publication year and the citation mean of the publication journal (at time of publication), as well as publication type, MeSH headings assigned, and geographical affiliation (typically associated with first author). Author characteristics can be divided into two subcategories: 1) gender / professional age / inventor status, and 2) publishing history. The first subcategory includes data on the gender makeup of the group. In addition to capturing the numbers and percentages of each gender, a gender diversity indicator was encoded and set to a value of (1) if there was a mix of genders in the co-author set, to a value of (-1) if there was high confidence that the author group was composed of a single sex, and to (0) if gender diversity could not be determined. Also in this subcategory are the professional ages of the first and last authors (under the assumption that these are the most distinctive of the group in that the first author leads the research and the last contributes the most experience), and the median, minimum, and maximum professional age of the authors in the collaboration. Attributes in the second subcategory of author attributes, publishing history, include the publishing records of the first and last author: highest number of citations, number of publications, h-index, total number of collaborators during publishing career, and rate of acquiring collaborators. This subcategory also captures the maximum, minimum, and standard of deviation in h-indices of the co-authors. The dataset and dataset documentation are available in the institutional repositories of both Virginia Tech (http://vtechworks.lib.vt.edu/handle/10919/23710) and the University of Illinois (https://ideals.illinois.edu/).

Aggregation of author-level data was based on the Authority dataset (Torvik & Smalheiser, 2009; Torvik et al., 2005), which has author names disambiguated on PubMed articles up to July 2009 with 98% accuracy. More recently, Torvik, Fegley, and Smith (2013) disambiguated across PubMed and the U.S. Patent and Trademark Office (USPTO) database, allowing for the identification of author-inventors. Gender was assigned to authors based on first names (when available) using the model by Smith, Singh, and Torvik (2013). The corresponding tool, called Genni, is available from http://abel.lis.illinois.edu.

Data manipulations

Data manipulations are described below, with histograms of the post-processing data distribution for select attributes displayed in Figure 1.

Gender

For the purposes of gender determination, the names in the author dataset are classified as male, female, neutral, or unknown. Neutral and unknown cases were treated effectively the same; both were classified as unknown. Also, in the case of an author with multiple names of varying gender classes, any gendered result was selected over an unknown or neutral return. If the author name had both male and female components, the majority was chosen, with unknown being assigned in the case of a tie. The attribute containing percentage of authors of unknown genders serves as a confidence measurement.

Topic

Although controlled at the high level of *breast neoplasm*, topical variation was represented in the data through MeSH headings. These were captured at the second level, in order to achieve an optimal amount of variation without creating an untenable quantity of attributes. For example, *Body Regions (A01)* is the level 1 heading, whereas *Abdomen (A01.047)* belongs to level 2. Treating the level 2 MeSH headings as article-level binary attributes ensured that topical bias was not affecting the model.

Instance Filtering

Of 98,070 records, 14,052 instances categorized as reviews were removed in order to capture only collaborations producing original research. 47 instances of articles with authors having a professional age greater than 60, representing an actual age of 80 – 85+ years, were removed under the assumption that scientists typically do not publish before the age of 20 – 25 years. (Professional age was calculated as date of publication for author's first article subtracted from date of publication of the article instance.) These authors, hypothetically publishing well into their eighties, represent either data errors or instances where the author-name disambiguation process was not clear.

Case Definitions

Next, two distinct datasets were created based on different group sizes of co-authors. In both cases, collaborations between fewer than two authors were removed, due to non-existent collaboration dynamics in solo authorship. Also, any collaboration with more than 50 authors was removed, as that was the point of onset for sparse data. This dataset,

titled Case 1, contains over 80,000 instances representing publications produced by 2 – 50 co-authors. Secondly, the lower and upper thresholds for number of authors were reduced, retaining only those instances in the range of three to seven collaborators, based on the hypothesis that much larger groups of authors might be influenced by more varied and complex sociological dynamics. We predicted that narrowing focus onto this set, composed of collaborations larger than pairs but fewer than 8 authors, would reveal different influences than across the larger group. For example, the impact of adding one author to the group was predicted to be much greater for a group of 3 than for a group of 40, where the effect might be negligible. On the other hand, co-author pairs were eliminated because coordination between authors working in a pair seems much less complex than in collaborations of three or more. This second dataset, referred to as Case 2, contains almost 53,000 instances representing publications written by 3 – 7 co-authors.

Finally, instances with publication dates prior to 1987 were deleted, due to the steep increase in instances beginning in that year. The quantity removed was relatively insignificant, ranging from 24 to 45 instances (varying by dataset as those vary in sample size as discussed above.) Removal of instances with publication year prior to 1987 was undertaken after instance removal for the purposes of class balancing as described above. Statistics for quantity impacted might differ if instances had not already been removed for class balancing.

Attribute transformations

Publication year was normalized from 1987 to 2009 to a scale of 1 to 23. Also, for all attributes other than the binary fields (MeSH terms and geographical attributes) and gender percentage fields, the log base 2 was substituted for the original value. (Before taking the log base 2, a Laplace estimator was added, to avoid missing values created by taking the log of any zeroes.) These changes decreased the intercept in our logistic regression models to a range of 2.1 to 4.46, depending on the case being studied.

Attribute reduction

To achieve higher variability in the dataset, binary attributes below or above certain thresholds were removed. These included the following: publication types and MeSH terms with frequency totaling either less than 5% of the total number of instances or greater than 95% of the total instances, and geographical affiliations with frequency totaling either less than 1% of the total number of instances or greater than 99% of the total instances. As a result, publication type attributes were reduced to the following categories: Clinical Trials; English Abstract; Research Support Non-U.S. Gov't; Research Support U.S. Gov't P.H.S; Research Support U.S. Gov't Non-P.H.S.; Research Support, N.I.H., Extramural; Comparative Studies; and Case Reports. MeSH headings attributes were diminished by 96%. Geographical affiliations remaining numbered 19 and 18 for the 2-50 author and 3-7 author datasets respectively, with 15 countries common to both. The following attributes were eliminated because their values were linearly deducible from those of other attributes: number and percentage of unknown gender (equal to the total minus males and females) and first and last authors' years of first publication (deducible from the year of article publication minus professional age).

After these reductions, 93 attributes remained in the 2-50 author dataset, and 92 in the 3-7 case. Attribute evaluators were utilized to attempt to reduce the dimensionality of the attribute space. These tools reduce number of attributes after consideration of the predictive ability of each as well as any correlations between them. Experimentation with different evaluators did not yield any higher performing models using a reduced dataset as input. However, it merits noting the frequently repeated selection of certain attributes by various evaluators/search methods. These include the first and last authors' maximum number of citations as of the time of the article's publication, the citation mean of the journal in that publication year, the presence of an English Abstract (indicating the article in a language other than English), the maximum author h-index and the standard of deviation in author group h-index when the article was published, having a publication type in one of the research categories or being a case report, and certain MeSH types such as *Heterocycle Compounds* and *Genetic Processes*. The attributes chosen through attribute selection classifiers do greatly correlate with the strongest predictors discussed in our results below, thereby reinforcing those outcomes. While it might have been simpler to analyze output with the smaller attribute set, analysis of the larger dataset allowed for the assessment of the more subtle impacts of the author and collaboration characteristics under study.

Impact Class Definitions

Classes were based on the number of citations for the article as of the year 2009. The dataset of articles was divided first into two classes: Class 0 = No Citations (never cited) and Class 1 = 1 or more Citations. These were then balanced through random instance removal executed through a Python script, resulting in 31,011 instances in each class. This provided a method of controlling for prior probabilities of class membership. Class balancing resulted in Zero-R (no rules) classification around 50%, whereas previously the percentage was skewed initially toward the dominant class before classification attempts.

Separate datasets binning citation counts into three classes of citation levels were created as well (through 2009 based on the number of citations an instance received). These were as follows: Class 0 = No Citations, Class 1 = Low (1-4) Citations, and Class 2 = Moderate to High Citations. After balancing the classes, 20,679 instances remained in each category of citation level for the Case 1 dataset. The binning and class balancing procedures were repeated for

the Case 2 dataset (3 to 7 authors). Never-cited instances (Class 0) versus cited instances (Class 1) now contained 21,262 instances each. Binning the same Case 2 dataset into three classes resulted in counts of 12,361 in each class: No Citations (Class 0), Low Citations (Class 1), and Moderate to High Citations (Class 2).

Fitting a model

Models were fitted using the WEKA data mining software (Hall et al., 2009). To establish a baseline, the WEKA ZeroR classifier was utilized, meaning no decision-making rules were applied, effectively mimicking a majority vote. For both cases (regardless of number of authors), ZeroR predicted with 49.99% accuracy for two balanced classes, and with 33.33% accuracy for three balanced classes. Prior to balancing of the classes for Case 1 (2 – 50 authors), ZeroR was classifying with 61.61% correct for the two-class trial, and 38.38% correct for the three-class trial. For the unbalanced datasets in Case 2 (3 - 7 authors), ZeroR correctly classified 59.73% into two classes, and 40.26% into three.

Given the results of ZeroR, it was decided to use the randomly balanced datasets to fit the model, so as to eliminate the possibility of prior probabilities for class artificially influencing results. However, additional experimentation with the unbalanced datasets indicated the following tradeoffs of choosing the balanced datasets. First, classification accuracy is on average about one percentage point lower. Classification accuracy measures how well the model performed during cross-validation within the dataset. For example, the highest classification percentage for the two-class trial was 74.01 for the balanced dataset versus 74.9 for the unbalanced. F-measure is a test for accuracy that takes into account both precision and recall.

Of the models and classifiers applied, the J48 decision tree and logistic regression methods yielded the most accurate classifications, according to F-measure. The J48 decision tree is the WEKA class for generating a pruned or unpruned C4.5 decision tree. (Quinlan, 1993). Decision trees capture highly interactive, non-linear spaces, which is beneficial when attributes have combinatorial effects, but they are not effective for smoothing because they discretize continuous attributes. When a J48 Tree model was used in combination with attribute selection pre-processing, the ROC, F-measure, and classification accuracy were all lower than those achieved by logistic regression by less than 1%, indicating that combinatorial effects are not occurring in this case. In addition, the attribute selection step removed many of the attributes targeted in this study, in favor of the obvious predictors (such as journal citation mean), so J48 was less informative than logistic regression.

	Case 1 2-50 author 2 Class	Case 2 3-7 author 2 Class	Case 1 2-50 author 3 Class	Case 2 3-7 author 3 Class
%Correct	74.01	73.45	58.89	59.40
Precision	.733/.748	.728/.741	.627/.445/.674	.624/.439/.669
Recall	.756/.724	.748/.721	.688/.416/.654	.687/.439/.669
F-measure	.744/.736	.738/.731	.656/.43/.664	.654/.424/.658
ROC	.822/.822	.817/.817	.83/.645/.847	.826/.638/.844

Table 1 WEKA results by class (0/1 and 0/1/2) of logistic regression with ridge parameter of 1.0E-8 coefficients.

The WEKA logistic regression function is a class for building and using a multinomial logistic regression model with a ridge estimator, and is a modified version of the class described by le Cessie and van Houwelingen (1992). In logistic regression, the regression coefficients represent the change in the logit *for* each unit change in the predictor. In an attempt to enhance the performance of the logistic regression model, LogitBoost, a form of additive logistic regression, was attempted using a base classifier of linear regression. While performance did not exceed that of the logistic regression model, the results lend additional credence to our findings from the less complex model.

Logistic regression offers the advantage of output that provides the relative importance of each attribute, in the form of weights and odds ratios. Output from the logistic regression function can be expressed as

$$log\frac{Pr\{Citation\}}{1-Pr\{Citation\}} = w_0 + w_1x_1 +_{\dots} + w_{93}x_{93}$$

where $(x_1, x_2 \dots x_{25}, x_{93})$ are *author attributes*; $x_{26} and\ x_{27}$ are *journal publication year and citation mean*; $(x_{28}, x_{29} \dots x_{35})$ are *publication types*; $(x_{36}, x_{37} \dots x_{73})$ are *MeSH terms*; and $(x_{74}, x_{75} \dots x_{93})$ are *geographical affiliations*. 'w' represents the corresponding weight assigned by the model to each 'x'. Note that the dataset containing 3 to 7 authors has only 92 attributes, due to having a lesser number of geographical affiliations after the filtering process. Attributes having a positive weight in the model are positively correlated with citation.

Fitting a model using logistic regression yielded the best overall accuracy in prediction for the classes in all four datasets (Table 1). Note that the results and discussion that follow are based primarily on the two class (binomial) models built for the Case 1 and Case 2 datasets. These classified instances by no citations versus citations, whereas the three-class model attempted to differentiate low citations from moderate to high citations. Although the trinomial models generally correlated to the binomial models in terms of weight and odds distribution, they classified the instances with much lower accuracy (Table 1). Decreased precision and recall occurred in all three-class models, but was observed to be especially significant in distinguishing the low citation class. Generally, attribute behavior in the three-class trial was consistent with the results of the two-class trial. Therefore, primarily we focus our conclusions and discussion upon the more accurately classified binomial model.

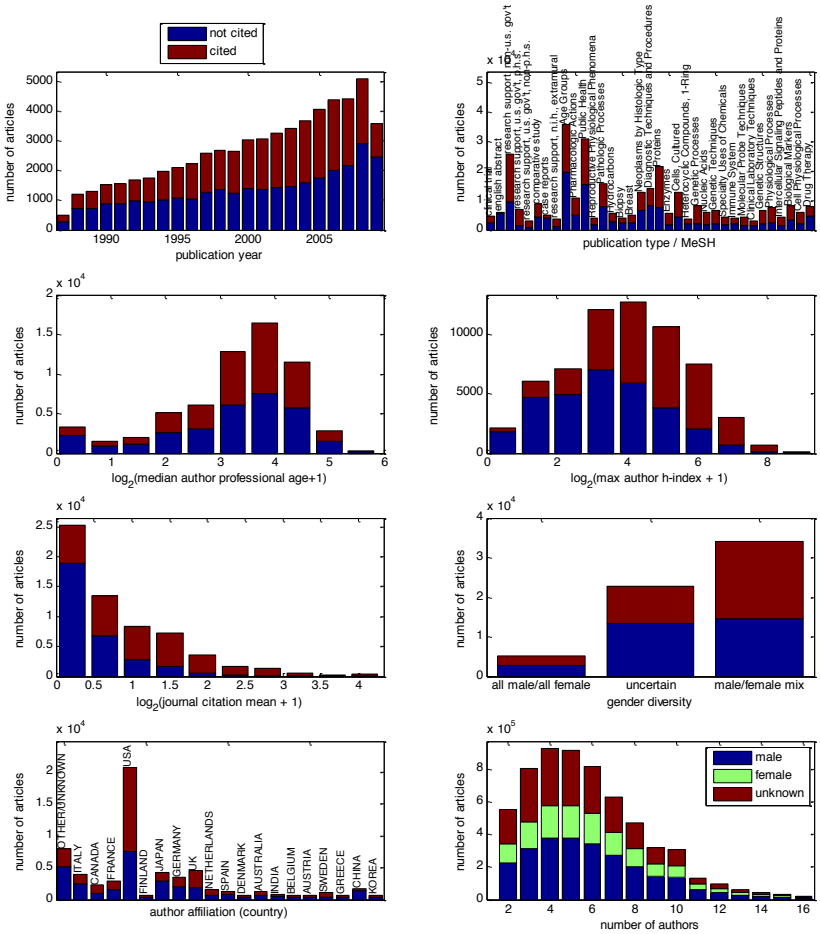

Figure 1 Distributions of select attributes in the 2-50 authors dataset

Because WEKA does not calculate P-values or confidence intervals, Matlab also was used to fit a logistic regression model for the two class cases (citation versus no citation). A P-value represents the probability that a test statistic is significantly different from the null hypothesis; the closer to zero the P-value is, the more likely the data is significant and not different by chance. In order to gain further insight into the significance of individual attributes, Matlab then was used to recursively remove attributes with the largest P-values (most insignificant) until all attributes remaining had P-values of less than 0.05, the threshold for statistical significance. Figures 2-7 reflect the attributes that remain significant after application of the recursive methodology; our conclusions are primarily based upon these, as all attributes with significant P-values have predictive importance.

To ensure that the results were not skewed by edge effects, the model was retested after removing the three most recent years of data. Excluding approximately 13,000 instances with dates spanning 2007 – 2009 did not result in any significant changes to the outcome (data not shown).

RESULTS AND DISCUSSION

A full table of logistic regression parameter estimates for the full model and after attribute reduction is available in the institutional repositories of Virginia Tech (http://vtechworks.lib.vt.edu/handle/10919/23710) and the University of Illinois (https://ideals.illinois.edu/). Table 2 displays the full model and post-reduction results for author attributes. Only attributes having P-values less than 0.05 are displayed in Figures 2 - 7, as these are interpreted to be statistically significant. An attribute with positive weight has a positive effect on citation, while an attribute with negative weight correlates to a negative effect on citation.

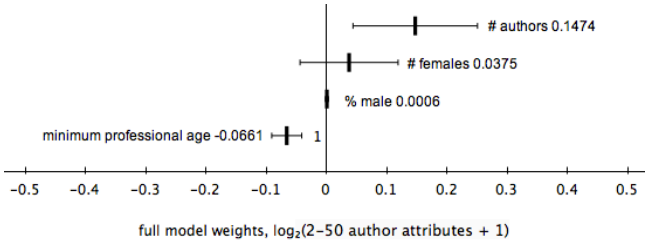

Figure 2 Author attributes for 2-50 author set

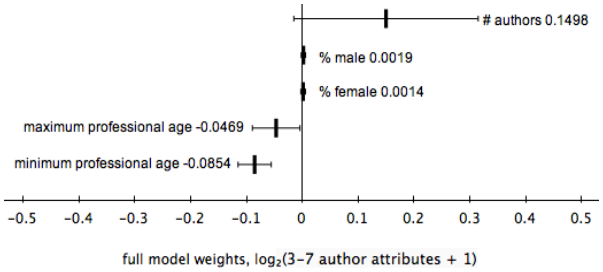

Figure 3 Author attributes for 3-7 author set

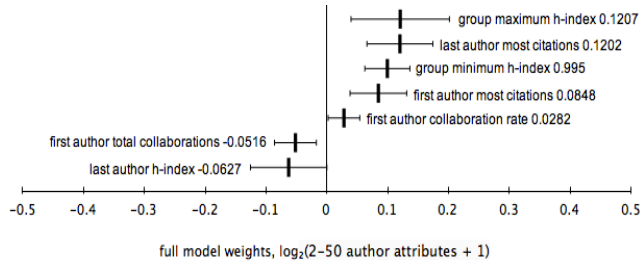

Figure 4 Publication history for 2-5 author set

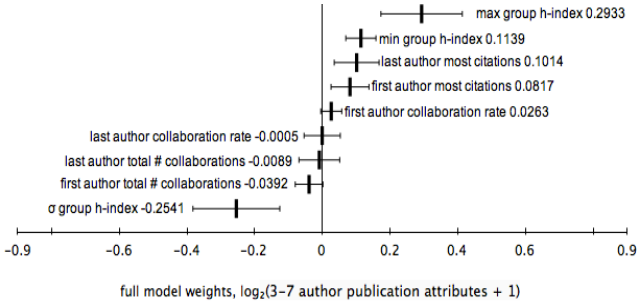

Figure 5 Publication history for 3-7 author set

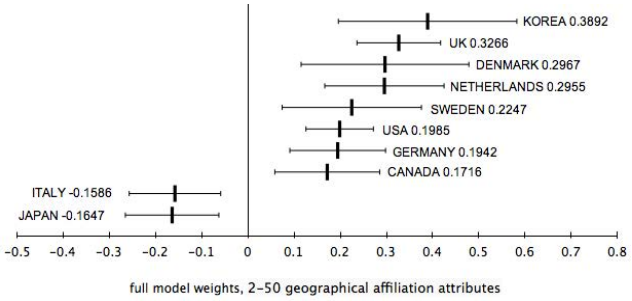

Figure 6 Geographical attr. for 2-50 author set

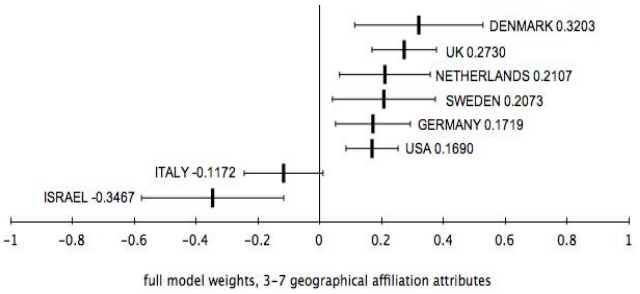

Figure 7 Geographical attr. for 3-7 author set

Figures 2 – 7 Logistic regression parameter estimates and 95% confidence intervals for full model Attributes shown are those remaining as significant after attribute reduction

General Observations

For all cases, the journal citation mean (at the time of the article's publication) consistently had the strongest impact on whether an article received citations. Of the author attributes (Figures 2 and 3), the number of authors was the most predictive; however, its influence was greatly reduced when accounting for all the other variables. Weight was 0.43 for the 2-50 dataset when looking at number of authors alone, but dropped to 0.17 in the recursive full model (and decreased from 0.38 to 0.10 in the 3-7 dataset). In the 2-50 author dataset, the weight of the number of authors attribute was only slightly greater than that of the maximum h-index, while in the 3-7 dataset the maximum h-index was actually twice as important as the number of authors. The group minimum h-index was also predictive of citation. The attribute that was most closely correlated with lack of citations was the publication type English Abstract, with a weight of -1.49 for the 2-50 dataset and -1.5 for the 3-7 case, indicating an article was written in a language other than English. However, several non-English speaking countries had a positive correlation with citation in PubMed; this is a linguistic rather than cultural feature. Certain MeSH topics, such as those associated with the field of genetics, appeared highly influential before accounting for the other attributes, after which point they carried only average weight amongst the significant MeSH terms.

Direction of influence changed when comparing full models to single-attribute models. Often referred to as Simpson's Paradox, this illustrates the danger of not including a large range of attributes, especially when

	2-50 AUTHORS				3-7 AUTHORS			
	WEIGHTS (P-VALUES)			ODDS RATIOS	WEIGHTS (P-VALUES)			ODDS RATIOS
ATTRIBUTE	Single Attribute Models	Full Model	After AttrIbute Reduction	Full Model	Single Attribute Models	Full Model	After Attribute Reduction	Full Model
AUTHOR AGE/GENDER/INVENTOR STATUS								
Diversity Ind	0.3875 (0.00)	-0.0068 (0.77)		0.9932	0.2792 (0.00)	0.0169 (0.62)		1.0171
Number of authors	0.4334 (0.00)	0.1474 (0.00)	0.1767 (0.00)	1.1588	0.3890 (0.00)	0.1498 (0.07)	0.1075 (0.01)	1.1616
Number of males	0.2873 (0.00)	0.0360 (0.46)		1.0367	0.2276 (0.00)	-0.0395 (0.60)		0.9613
Number of females	0.3823 (0.00)	0.0375 (0.36)	0.0543 (0.00)	1.0382	0.3245 (0.00)	-0.0064 (0.93)		0.9936
Percentage male	0.0042 (0.00)	0.0006 (0.64)	0.0014 (0.00)	1.0006	0.0043 (0.00)	0.0019 (0.33)	0.0011 (0.02)	1.0019
Percentage female	0.0093 (0.00)	0.0006 (0.62)		1.0006	0.0090 (0.00)	0.0014 (0.52)	0.0017 (0.00)	1.0014
Inventor(s) present	0.6189 (0.00)	0.0538 (0.12)		1.0553	0.5434 (0.00)	0.0133 (0.76)		1.0134
Percentage inventors	0.0144 (0.00)	-0.0011 (-0.25)		0.9989	0.0146 (0.00)	0.0003 (0.80)		1.0003
Median professional age	0.1181 (0.00)	0.0123 (0.29)		1.0124	0.1321 (0.00)	0.0140 (0.29)		1.0141
Minimum professional age	-0.0051 (0.44)	-0.0661 (0.00)	-0.072 (0.00)	0.9360	0.0477 (0.00)	-0.0854 (0.00)	-0.0935 (0.00)	0.9181
Maximum professional age	0.2618 (0.00)	-0.0174 (0.34)		0.9827	0.2174 (0.00)	-0.0469 (0.03)	-0.0539 (0.00)	0.9542
First author's professional age	0.0074 (0.13)	-0.0258 (0.06)		0.9746	0.0049 (0.40)	-0.0225 (0.17)		0.9778
Last author's professional age	0.1487 (0.00)	-0.0167 (0.40)	-0.0336 (0.00)	0.9834	0.1465 (0.00)	-0.0182 (0.43)		0.9820
AUTHOR PUBLICATION HISTORY								
Group minimum h-index	0.3771 (0.00)	0.0995 (0.00)	0.1170 (0.00)	1.1047	0.4766 (0.00)	0.1139 (0.00)	0.1240 (0.00)	1.1206
Group maximum h-index	0.4749 (0.00)	0.1207 (0.00)	0.0701 (0.00)	1.1283	0.4704 (0.00)	0.2933 (0.00)	0.3236 (0.00)	1.3408
Std of deviation in h-index	0.5347 (0.00)	-0.0626 (0.15)		0.9393	0.5215 (0.00)	-0.2541 (0.00)	-0.274 (0.00)	0.7757
First author h-index	0.2966 (0.00)	0.0327 (0.26)		1.0333	0.2942 (0.00)	0.0326 (0.35)		1.0331
Last author h-index	0.3523 (0.00)	-0.0627 (0.05)	-0.0774 (0.00)	0.9392	0.3448 (0.00)	-0.0457 (0.24)		0.9553
First author total publications	0.0440 (0.00)	0.0030 (0.86)		1.0030	0.0387 (0.00)	-0.0105 (0.62)		0.9896
Last author total publications	0.1445 (0.00)	-0.0047 (0.78)		0.9953	0.1397 (0.00)	-0.0074 (0.72)		0.9926
First author most citations	0.3129 (0.00)	0.0848 (0.00)	0.1104 (0.00)	1.0885	0.3108 (0.00)	0.0817 (0.00)	0.1044 (0.00)	1.0852
Last author most citations	0.3497 (0.00)	0.1202 (0.00)	0.1251 (0.00)	1.1277	0.3414 (0.00)	0.1014 (0.00)	0.0635 (0.00)	1.1067
First author total collaborators	0.0558 (0.00)	-0.0516 (0.00)	-0.0652 (0.00)	0.9497	0.0427 (0.00)	-0.0392 (0.06)	-0.0601 (0.00)	0.9616
First author collaboration rate	0.1276 (0.00)	0.0282 (0.03)	0.0376 (0.00)	1.0286	0.0975 (0.00)	0.0263 (0.09)	0.0346 (0.00)	1.0267
Last author total collaborators	0.1445 (0.00)	-0.0183 (0.48)		0.9819	0.1340 (0.00)	-0.0089 (0.77)	-0.028 (0.00)	0.9911
Last author collaboration rate	0.2639 (0.00)	0.0054 (0.81)		1.0054	0.2347 (0.00)	-0.0005 (0.98)		0.9995

Table 2 logistic regression parameter estimates for full model, single attribute model, and after attribute reduction

interpreting the intrinsic effects of weak predictors without first factoring out strong predictors in the model. In this case, attributes that switched direction of importance from positive correlation with citation to negative in the full model included the last author's professional age and h-index, the first author's total number of collaborators, and several MeSH terms. There were also some MeSH terms that appeared negatively correlated when evaluated individually, but became positive predictors of citation in the full model. Likewise, Germany and Korea were negative before accounting for the full range of attributes but are positive after. Evaluating only geographical affiliation, out of context of the other variables, produced distorted results.

Author Attributes

For both the 2-50 and the 3-7 datasets, the highest number of citations received (by both the first and last authors) was predictive of citation (Figures 4 and 5). Doubling the number of authors resulted in a 16% and 13% increase in probability of the article being cited (for the 2-50 and 3-7 groups respectively). Interestingly, for the 2-50 author group, an increase in the minimum professional age of the group, or in the professional age of the last author, resulted in a greater likelihood of the article never being cited. The case might be made that younger researchers are pursuing more innovative investigations. Note, however, that the median and maximum professional ages were insignificant, as was that of the first author, indicating perhaps that the role of the middle authors may be undervalued. Furthermore, an increase in the last author's h-index was linked to lack of citation, while the group minimum and maximum h-indices predicted citation (Figure 4). The first author's h-index was insignificant. Assuming the last author's h-index is often the group maximum, the contradictory direction of influence of these two attributes shifts greater importance to the minimum h-index. Doubling the minimum h-index results in 10% greater probability of citation. Since the first author's h-index was insignificant, we can conclude that a collaborating group is as strong as its weakest link; in other words, the previous publishing successes of all co-authors are important to achieving citation.

These same author attributes behaved somewhat differently for the 3-7 author collaborations (Figure 5). In that case, the group maximum h-index was the strongest predictor of citation, almost three times stronger than any other author attribute, with odds of citation increasing by 34% when doubling the attribute. Also, the minimum h-index emerged as an important factor, increasing probability of being cited by 12%. However, an increase in the standard of deviation of h-index was highly negatively correlated (citation 22% less likely after doubling this attribute); also negatively weighted were the minimum and maximum professional ages. Combining these observations about age and h-index suggests that smaller collaborations composed of early superstars – young, rapidly successful researchers with relatively high and similar h-indices – might prove to be highly productive in biomedical research.

The total number of unique collaborators for the first and last author was negatively correlated with citation in the 3-7 author case, and negative for first author but insignificant

for the last author in the 2-50 dataset. The collaboration rate, defined as the number of unique collaborators divided by professional age, was positively weighted for first author but insignificant for last author in both datasets. These results may further reinforce the importance of the participation of younger researchers. Accruing a high number of collaborators often correlates with career length, and this is especially true for authors in sub disciplines tending toward smaller collaborations (which would explain the insignificance of the attribute in the 2-50 author dataset). However, gaining a high total of collaborators over a short period of time results in a high collaboration rate, and is perhaps indicative of early productivity.

In considering the gender diversity of the author group, results differed across the two datasets (Figures 2 and 3). While the number of males was insignificant for both cases, the number of females was influential in the 2-50 case. Interestingly, the percentage of males was also predictive of citation within that dataset. This suggests that a large, male-dominated collaboration would increase the probability of citation by adding a small number of female researchers to the group. However, for the 3-7 dataset, both the number of females and the number of males were insignificant, while higher percentages of each predicted citation. The paradoxical behavior of these attributes suggests that a group of mixed gender may be preferable, and indicate that gender diversity in collaboration may be an interesting area for further research.

Geographical Affiliations

Eight of the eleven countries that emerged as statistically significant were positively correlated with citation (Figures 6 and 7). This is unsurprising, as the list is composed of relatively large, first-world countries in which research is emphasized and more commonly funded. Of these, Korea had the strongest positive weight, explainable under the assumption that Korea is not as common in the datasets, therefore increasing the likelihood of ending up at an extreme. The same reasoning of small sample size might explain Israel's highly negative correlation in the 3-7 author set (while appearing insignificant in the 2-50). The U.S., in contrast, with the largest number of articles in the datasets, and thereby representing the greatest range of citations levels, had a lesser weight than almost all the other countries due to an averaging effect. The only country of affiliation that was consistently negatively weighted was Italy. Japan was negatively correlated for the 2-50 dataset and insignificant for the 3-7 author group.

CONCLUSION

As suggested by previous research, adding an author to an article positively impacts the number of citations the article receives; however, the influence of this attribute is greatly lessened after accounting for the other author and article attributes. While the number of collaborating co-authors is a factor in citation counts, other highly influential factors were found. The maximum h-index of the author group emerged as strongly predictive of citation, with a weight approaching that of the number of authors in the 2-50 author dataset, and far surpassing it in the 3-7 case. When combined with the finding that a smaller standard of deviation in h-index correlates with citations in the 3-7 group, we conclude that smaller superstar groups may be likely to achieve comparable citations levels with large, highly networked collaborations. In addition, there is much evidence to support the importance of the participation of young researchers. Negative correlations emerged between citation and both professional age and total number of collaborators, while collaboration rate, associated with a shorter career, predicted citation. Minimum h-index was important but the first author's h-index was insignificant, suggesting that greater importance should be attributed to the middle authors' publishing history. The gender diversity outcomes suggest that mixed groups may be ideal, and further research in this area is indicated.

Limitations and Value

The purpose of this study is to serve as a starting point for exploring the effect of group author characteristics on citation counts within PubMed Central; while PubMed and PubMed Central represent a valuable disciplinary and scholarly ecosystem for citation-based bibliometric study, further research is required to test whether our results hold across a global scale. The value of this type of exploration of author/article attribute space lies in not just confirming or challenging of previously held notions about collaboration patterns, but the large scale, both in attributes and observations, permits identification of more subtle and unexpectedly influential attributes.

Further Research

We suggest that additional research into the impact of gender diversity on productivity might be warranted. For example, our categorization of collaboration as same-sex or mixed-gender does not account for sociological patterns such as "queen bee" or "alpha male", which may occur when there are same-sex members within a mixed-gender group. For further discussion, see Raghubir and Valenzuela (2010) or Parker (2010). Future research might examine collaboration impacts through a comparison of solo publications to collaborative works completed by the same author. Such a strategy would leverage the natural controls in place when focusing upon a single author operating in multiple modes. Finally, it would be interesting to slice the dataset temporally, under the hypothesis that spikes in collaboration might correspond to surges in technological and cultural advances.

Acknowledgements

Funding for this study was provided in part by NSF Science of Science and Innovation Policy (SciSIP) program award 0965341. The authors wish to thank Dr. Godmar Back for technical advice and assistance.

REFERENCES

Abramo, G., D'Angelo, C.A., & Di Costa, F. (2009). Research Collaboration and Productivity: Is There Correlation? *Higher Education 57*(2), 155-171.

Adams, J. D., Black, G.C., Clemmons, J.R., & Stephan, P.E. (2005). Scientific Teams and Institutional Collaborations: Evidence from U.S. Universities, 1981–1999. *Research Policy 34*(3), 259-85.

Bergh, D.D., Perry, J., & Hanke, R. (2006). Some Predictors of SMJ Article Impact. *Strategic Management Journal, 27*(1), 81-100.

Fischbach, K., Putzke, J., & Schoder, D. (2011) Co-authorship Networks in Electronic Markets Research. *Electron Markets 21*, 19-40.

Gazni, A., & Didegah, F. (2011). Investigating Different Types of Research Collaboration and Citation Impact: A Case Study of Harvard University's Publications. *Scientometrics 87*(2), 251-265.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explorations 11*(1).

Haslam, N., Ban, L., Kaufmann, L., Loughnan, S., Peters, K., Whelan, J., & Wilson, S. (2008). What Makes an Article Influential? Predicting Impact in Social and Personality Psychology. *Scientometrics 76*(1), 169-85.

Haslam, N., & Simon Laham, S. (2009). Early-Career Scientific Achievement and Patterns of Authorship: The Mixed Blessings of Publication Leadership and Collaboration. *Research Evaluation 18*(5), 405-10.

Hirsch, J.E. (2007) Does the h Index Have Predictive Power? *PNAS* 104(49), 19193-19198.

Ioannidis, J.P.A. (2008). Measuring Co-Authorship and Networking-Adjusted Scientific Impact. *PLoS ONE 3*(7), e2778.

Le Cessie, S., & Van Houwelingen, J.C. (1992). Ridge Estimators in Logistic Regression. *Applied Statistics*.41(1), 191-201.

Nichols, D. (2012, May 4). Metrics for Openness. CIRSS Seminar Series, University of Illinois at Urbana-Champaign.

Quinlan, R. (1993). *C4.5: Programs for Machine Learning*.

Parker, J. (2010). An Empirical Examination of the Roles of Ability and Gender in Collaborative Homework Assignments. *The Journal of Economic Education 41*(1), 15–30.

Petersen, A. M., Riccaboni, M., Stanley, H. E., & Pammolli, F. (2012). Persistence and uncertainty in the academic career. *Proceedings of the National Academy of Sciences*, *109*(14), 5213-5218.

Priem, J., Taraborelli, D., Groth, P., & Neylon, C. (2011) altmetrics: a manifesto. Retrieved from altmetrics.org/manifesto

Raghubir, P., & Valenzuela, A. (2010). Male Female Dynamics in Groups: A Field Study of The Weakest Link**.** *Small Group Research* 2010 *41*: 41

Rey-Rocha, J., & Martín-Sempere, M.J., Martínez-Frías, J., & López-Vera, F. (2001). Some Misuses of Journal Impact Factor in Research Evaluation. *Cortex 37*(4), 595-597.

Skilton, P.F. (2009). Does the Human Capital of Teams of Natural Science Authors Predict Citation Frequency? *Scientometrics 78*(3), 525-542.

Smith B.N., Singh M., Torvik V.I. (2013). A search engine approach to estimating temporal changes in gender orientation of first names. Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries. JCDL '13, July 22-26, Indianapolis, IN, USA. 199-208.

Sooho, L., & Bozeman, B. (2005). The Impact of Research Collaboration on Scientific Productivity. *Social Studies of Science 35.5,* 673-702.

Sooryamoorthy, R. (2009). Do Types of Collaboration Change Citation? Collaboration and Citation Patterns of South African Science Publications. *Scientometrics 81.1*, 177-93.

Stremersch, S., & Verhoef, P.C. (2005). Globalization of Authorship in the Marketing Discipline: Does It Help or Hinder the Field? *Marketing Science 24*(4), 585-594.

Torvik V.I., Fegley B.D., Smith B.N. (2013). Identifying biomedical author-inventors by probabilistic disambiguation and linking names across PubMed and USPTO. Working paper Graduate School of Library and Information Science.

Torvik, V.I., & Smalheiser, N.R. (2009). Author Name Disambiguation in MEDLINE. *ACM Transactions on Knowledge Discovery from Data 3*(3), 11.

Torvik V.I., Weeber M., Swanson D.R., Smalheiser N.R. (2005). A probabilistic similarity metric for Medline records: A model for author name disambiguation. *Journal of the American Society of Information Science and Technology* 56(2), 140–158.

Wuchty, S., Jones, B.F., & Uzzi, B. (2007). The Increasing Dominance of Teams in Production of Knowledge. *Science 316*, 1036.