Scientometrics (2013) 96:467–483
DOI 10.1007/s11192-012-0933-0

Predicting author h-index using characteristics of the co-author network

Christopher McCarty · James W. Jawitz · Allison Hopkins · Alex Goldman

Received: 22 September 2012 / Published online: 28 December 2012
© Akadémiai Kiadó, Budapest, Hungary 2012

Abstract The objective of this work was to test the relationship between characteristics of an author's network of coauthors to identify which enhance the h-index. We randomly selected a sample of 238 authors from the Web of Science, calculated their h-index as well as the h-index of all co-authors from their h-index articles, and calculated an adjacency matrix where the relation between co-authors is the number of articles they published together. Our model was highly predictive of the variability in the h-index ($R^2 = 0.69$). Most of the variance was explained by number of co-authors. Other significant variables were those associated with highly productive co-authors. Contrary to our hypothesis, network structure as measured by components was not predictive. This analysis suggests that the highest h-index will be achieved by working with many co-authors, at least some with high h-indexes themselves. Little improvement in h-index is to be gained by structuring a co-author network to maintain separate research communities.

Keywords Egocentric network · H-index · Co-author network

C. McCarty (✉)
Bureau of Economic and Business Research, University of Florida, 221 Matherly Hall, Gainesville 32611-7145, USA
e-mail: ufchris@ufl.edu

J. W. Jawitz
Soil and Water Science Department, University of Florida, 2169 McCarty Hall, Gainesville, FL 32611, USA
e-mail: jawitz@ufl.edu

A. Hopkins
Department of Family and Community Medicine, University of Arizona, 1450 North Cherry Avenue, Tucson 85724, USA
e-mail: hopkin28@email.arizona.edu

A. Goldman
Department of Sociology, University of Florida, 3219 Turlington Hall, Gainesville 32611-7330, USA
e-mail: alexevasion@gmail.com

Introduction

Over the course of a scientific career there are opportunities for collaboration with other scientists, and wide variability in the extent to which individual scientists choose to collaborate. In this article we explore a range of collaborative behaviors and test the extent to which those behaviors result in a measureable increase in scientific impact. Our objective is to identify those behaviors that will maximize scientific impact.

It is well known that there are cultural differences between disciplines regarding whether scientists collaborate with others (Becher and Trowler 2001). Some disciplines require a team effort in order to conduct even the most basic scientific experiment. Molecular biologists require laboratories with at minimum a senior investigator and lab technicians, and typically employ one or more graduate students and post-doctoral fellows. Scientific efforts such as the Hadron Collider in Switzerland cannot be undertaken without large teams who are often listed among co-authors on published findings. A recent contribution from that group had 2,926 authors (Collaboration et al. 2008). In contrast, there are other disciplines with lesser infrastructural needs, such as theoretical physics, where useful contributions can be made as a sole author.

There are of course other reasons to collaborate that are not driven by infrastructural needs. First and foremost is the synergistic creativity that comes from working with others. Although a physicist such as Albert Einstein published most of his work alone, even he found value in combining his ideas with those of others, particularly in the later part of his career (Einstein and Rosen 1936). There are more practical reasons for collaborating with others. Multiple authors can effectively split the work so more articles can be published. And as will become apparent from our research, co-authors create channels themselves for the dissemination of one's findings.

Although collaboration can enhance scientific output and dissemination, there are reasons not to collaborate. Some disciplines effectively discourage collaboration by assigning more value to sole-authored publications (Becher 2006). There may also be interpersonal and professional issues in working with certain co-authors. Perhaps the division of labor is not equal so that some collaborators become free-riders. Scientists may choose collaborators who ultimately do not agree on the science, which may slow output. In some disciplines scientists may not trust collaborators with their ideas for fear that they will be stolen.

Choosing whether to collaborate, and if so, with how many investigators, is not the only choice. Scientists can choose collaborators based on particular characteristics, such as their own success as scientists or their experience in the field. They can also make choices about who they collaborate with across different publications. By varying the set of collaborators a scientist works with, the structural arrangement of those collaborators around the scientist (the egocentric network) could have positive or negative effects. An examination of the structure of the egocentric co-author network is a particular focus of this research. In this article we will present the results of a model to test the effects of these collaborative behaviors on scientific impact using a random selection of authors from the Web of Science.

Scientific impact—the h-index

There are a variety of ways one could gauge scientific impact. One could think of the lasting impacts that a scientist's contributions have made in the world, such as the

discovery of a polio vaccine by Jonas Salk. Some scientists make lasting contributions without immediate practical application, but lasting theoretical impact. When thinking of the specific behaviors a scientist can engage in, their productivity can include patents or copyrights for discoveries, teaching and mentoring students, presentations to other scientists at conferences or through consulting. But the typical metric of scientific output is some form of analysis of their publication record.

Publication is arguably the major component in the evaluation of an academic scientist. Evaluation for hiring, promotion and grants are heavily weighted by the publication record. Although there is still variability in the way publications are evaluated, quantitative metrics that provide an objective summary of a publication career are increasingly used (Alison et al. 2010).

For this study we will represent scientific impact using the h-index, a measure of the scientific achievement of individual authors first proposed by Jorge Hirsch (2005). Since that time, it has received significant attention from science news editors and researchers working in the area of bibliometrics. Major citation databases, including the Web of Science and Scopus (Bar-Ilan 2008), now include the h-index in their citation reports.

There are many advantages to measuring scientific output using the h-index. It is the first widely used single measure of scientific output that measures both productivity and impact (Hirsch 2005; Roediger 2006). The numbers needed to calculate an author's h-index are easy to acquire and the calculation is easy to perform (Bornmann and Daniel 2007; Hirsch 2005; Roediger 2006). The measure is not inflated by a small number of highly cited papers or a large number of papers with low citation rates (Cronin and Meho 2006; Hirsch 2005), nor does it give extra weight to certain types of published documents (Hirsch 2005). Also, the index does not include arbitrary values which randomly favor or disfavor individuals (Hirsch 2005). Finally, there is low distortion of individual output in co-authored papers compared with the total citation count method, another popular measure of scientific impact (Hirsch 2007).

Several tests have been performed which support the validity of the h-index. Studies from physics, information and consumer sciences, and mathematics report a positive association between h-index and other standard bibliometric measures (Cronin and Meho 2006; Glanzel 2006; Hirsch 2007; Saad 2006). Comparisons of h-index to citation counts and number of papers published in the biological sciences show similar positive results for the validity of h-index (Costas and Bordons 2007; Hirsch 2005; Kelly and Jennions 2006). The h-index compares well to peer-review judgment in chemistry, chemical engineering, and biomedicine (Bornmann and Daniel 2005; Van Raan 2006). In addition, it is resilient to missing and erroneous publication data based on examples from environmental science and management, a Zipf distribution, h-indices of Price medalists, and an analytical model (Rousseau 2007; Vanclay 2007).

However, there are several caveats that are important to consider in the use and interpretation of the h-index. Hirsch (2005) recognized that the h-index, like every other single measure of scientific output, is unable to measure all aspects of scientific impact and should be used in conjunction with other forms of assessment. The h-index is typically considered discipline dependent (Bornmann, Mutz, and Daniel 2008) and is affected by the average number of references in a paper, average number of papers produced, the number of scientists, and the attractiveness of the topic within each field (Bornmann and Daniel 2007). This metric thus tends to favor disciplines in which findings are produced within the context of larger groups and through experimental rather than theoretical research. Equal value is assigned to each author in multiple-author papers, regardless of author sequence or the total number of authors. Books and other alternative forms of publication typically are

not included in h-index calculations. Thus, co-authorship may lead to inflation of h-indices because every author on a paper receives the same amount of credit for it, independent of the actual amount of effort they contributed to producing the work (Roediger 2006). Also, the index is dependent on the number of years a scientist has been publishing scientific papers and may lack sensitivity to performance changes throughout scientists' careers (Rousseau 2008; Sidiropoulos, Katsaros and Manolopoulos 2007). Self-citation can also inflate h-indices (Hirsch 2005; Schreiber 2007; Zhivotovsky and Krutovsky 2008). It must also be noted that lag time between a paper being published and being discovered and cited often varies substantially (Roediger 2006). Efforts to resolve some of these issues have led to the development of several h index variants (Batista et al. 2006; Iglesias and Pecharromán 2007; Imperial and Rodríguez-Navarro 2007; Radicchi, Fortunato and Castellano 2008; Banks 2006; Egghe 2007; Liang 2006; Burrell 2007; Schreiber 2007; Anderson, Hankin and Killworth 2008).

Hirsch (2010) presented a variant called h-bar that penalizes authors who publish with established and productive co-authors. H-bar does not count papers where a co-author has an h-index at or above the number of citations for the paper. This would have the effect of lowering the h-index for those at the beginning of their careers who publish with their advisors or those who continue to publish with established authors. Schubert, Korn and Telcs (2009) developed a network measure called the degree h-index, where the h for a journal is the number of authors (or papers) in the network with a degree of at least h. Schubert (2012) took a different approach by discounting the behavior of publishing repeatedly with the same set of co-authors. His partnership ability index (φ) would reward repeated collaborations with many different coauthors. These variants all incorporate network characteristics of collaborations. As will become clear from our approach, we included network variables that capture these characteristics directly.

Co-authorship and its influence on scientific productivity

One potential benefit of co-authorship is increased scientific productivity, as measured by publication and citation rate (Beaver and Rosen 1979; Melin 2000). Numerous studies have reported a strong positive relationship between co-authorship and scientific productivity (Adams et al. 2005; Börner et al. 2005; Katz and Martin 1997). Studies of Nobel laureates in science (Zuckerman 1967), musicologists (Pao 1982), and scientists from a wide range of disciplines (Persson, Glanzel and Danell 2004) have found a positive association between co-authorship and the number of papers published. Many studies have also found that collaboration between authors from diverse geographical locations (Frenken, Holzl and de Vor 2005; Goldfinch, Dale, and DeRousen 2003; He et al. 2009; Katz and Hicks 1997; Narin et al. 1991; Nemeth and Goncalo 2005), employing institutions (Frenken, Holzl and de Vor 2005; Goldfinch, Dale, and DeRousen 2003; Jones, Wuchty, and Uzzi 2008; Katz and Hicks 1997), and disciplines (Leimu and Koricheva 2005) with multi-disciplinary backgrounds (Skilton 2009) are positively associated with scientific productivity. The benefits of diversity in scientific co-authorship stem from the sharing of different approaches and can also generate an increased readership by tapping into multiple professional social networks (Goldfinch, Dale, and DeRousen 2003; Jones, Wuchty, and Uzzi 2008; Leimu and Koricheva 2005).

Social network studies have furthered our understanding of the relationship between co-authorship and productivity. Studies assessing the relationship between productivity and the position of authors in the co-author network have found that authors who publish with

many different co-authors bridge communication and tend to exhibit higher rates of publication (Börner et al. 2005; Eaton et al. 1999; Kretschmer 2004). Another group of researchers have focused on how productivity is affected by individual and organizational social capital, finding that while individual social capital provides scientists with access to careers in prestigious institutions, establishment in these institutions allows researchers to integrate into existing networks with other top tier institutions, fostering collaborations that lead to greater quantity and quality of publications (Lazega et al. 2008; Lazega et al. 2006; Rodgers and Maranto 1989). In addition, individuals who are higher up the hierarchy in a research institution spend more time defining goals than executing them and have increased access to human and monetary resources (Knorr and Mittermeir 1980). These factors increase researchers' ease in producing publications and help explain the positive association between position and productivity.

Some collaborations do not lead to increased productivity. The timing in a funding cycle (Defazio Lockett, and Wright 2009), and the social distance of the collaboration (Frenken, Holzl and de Vor 2005, Goldfinch, Dale, and DeRousen 2003) can affect whether there is an association between co-authorship and productivity. Several researchers have found that when individual characteristics and work environment variables were controlled for, there was no association between scientific productivity and number of collaborators (Lee and Bozeman 2005; Walters 2006). Glanzel (2002) found that there was an optimal level of co-authors that varied by field and once that level was exceeded, productivity declined. The variation in findings within these different studies shows that the relationship between co-authorship and productivity is still not fully understood (Persson, Glanzel and Danell 2004).

Ego-centered co-author networks

There are two main approaches to social network analysis, based on either whole or egocentric networks (Wasserman and Faust 1994). Whole network analysis focuses on the interaction between actors within a geographically or socially bounded space. Egocentric network analysis focuses on the social context of a sample of actors and how those characteristics predict something about them. Personal network analysis, a type of ego-centric network, focuses on the social context across social spaces. The majority of co-authorship network studies are whole network studies assessing the relationship between productivity and social position within a discipline or journal (Newman 2004; Moody 2004; Hou, Kretschmer, and Liu 2008). The few ego-centered network studies have mostly focused on citation patterns of specific successful authors (Bar-Ilan 2008; Batagelj and Mrvar 2000; de Castro and Grossman 1999; Moravcsik 1988; Swarna, Kalyane and Kumar 2008; White 2000, 2001), attempting to understand what collaborative behaviors they have engaged in that led to their success.

On occasion, whole fields of study are described using ego-centered bibliometric analyses. Mulchenko et al. (1979) gathered bibliometric data for ten leading chemists and eight renowned physicists. They used this information to characterize the organizational forms of research in the different fields. White and McCain (1998) used author co-citation analysis to describe the domain of information science. Yoshikane and Kageura's (2004) study focused on the development of personal networks in electrical engineering, information processing, polymer science, and biochemistry.

Previous ego-centered bibliometric studies tend to be descriptive with little emphasis on the relationship between co-author and citation patterns on scientific impact. One notable

exception is Moravcsik's (1988) classification of the published works that cited his "citation classic" publication and the impact it had on the citer's research. Börner et al. (2005)(66), suggested that,

> "A closer mathematical and empirical examination of the correlation among the four centrality and impact measures of authors and their relation to prior work in bibliometrics (*e.g. ego-centered bibliometrics*) (Crane 1972; White 2001) is expected to lead to new insights into the co-authorship dynamics."

A recent study by Abbasi et al. (2011) evaluated the network properties of the ego networks of 8,069 authors from 4,837 publications. In addition to network measures such as degree and betweenness centrality they included Burt's structural holes measures of efficiency and constraint. They found a positive association between ego-network structure and both the h-index and the g-index. An important finding was that authors who bridged structural holes by connecting to diverse groups performed better.

Our study was designed to combine bibliometric measures with ego-centered social network measures to assess individual scientific impact as measured by the h-index. An example of this is shown in Fig. 1, which visualizes the network of all co-authors of articles published by the focal author as of 2006 (the year selected as the sample frame for this study). Each node in the graph represents a co-author, with size scaled by the co-author's h-index, and color based on the author's institutional affiliation: red for academic settings and blue for others. This graph shows that co-author (node) 15 is a highly prolific author who also works with many of the other co-authors in this egocentric network.

Our objective with this study was to randomly select a sample of authors from the Web of Science and collect the data to create the egocentric network for each. It is worth noting again that this egocentric network is not bound by disciplines. For example in Fig. 1 the focal author works with co-authors who cross traditional disciplinary boundaries. By understanding how the composition and structure of these egocentric networks do or do not

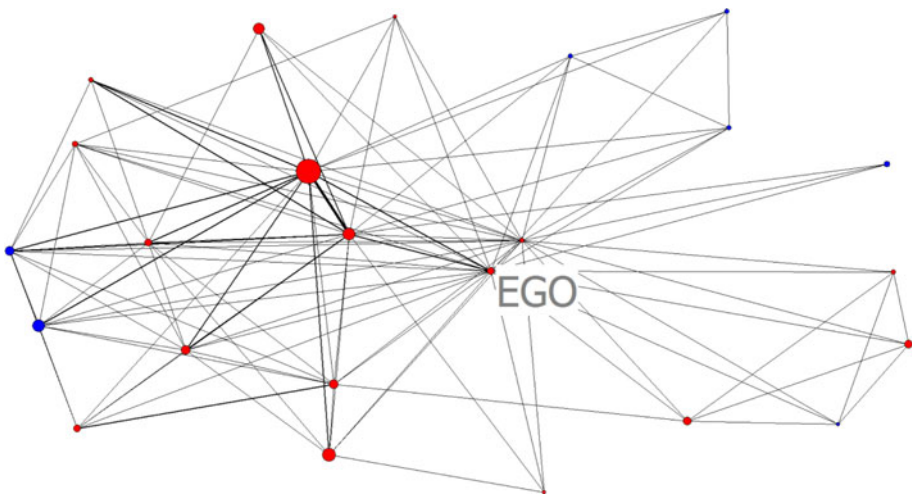

Fig. 1 Example of an author's egocentric network

explain the variability in the h-index of the focal author, we will learn which collaborative behaviors enhance scientific impact.

Methods

We chose 2006 as our sample frame and downloaded citation data from approximately 3.4 million individual publications from the Web of Science into a SAS database. After parsing the author string into individual authors and concatenating the author with their institutional affiliation we de-duplicated the data, leaving approximately 1 million author-affiliation strings. Like all studies of this kind we faced problems of disambiguation. This means that authors often have the same last name and first initial, and therefore appear in the Web of Science as the same person. For some names the effect of this can be severe. Increasingly researchers turn to advanced algorithms to sort this out (Onodera et al. 2011; Tang and Walsh 2010). In our case we devised guidelines based on affiliation, topic and publication record and trained a team of undergraduate and graduate students to disambiguate authors. Since initials, hyphens, and apostrophes all pose problems when attempting to compile a list of all the articles by a unique author indexed in the Web of Science, different versions of their names were used as search criteria. Disciplinary backgrounds of the focal authors and coauthors were also cross-referenced to ensure consistency. Author affiliations were verified by matching email addresses to the institutional affiliation data stored for recent articles or by conducting thorough internet searches. For large h-index values these were double-checked by a second reviewer and discrepancies resolved. Although this was an expensive approach, we believe it is at least as accurate if not more so than the automated solutions available today. H-indexes were then calculated for 594 authors randomly selected from that list.

The distribution of the h-index values is shown in Fig. 2. To our knowledge this is the only existing distribution of h-indexes across the Web of Science. The lowest h-index we recorded was 0 for 28 authors. Although these authors had published, none of their papers had ever been cited and therefore had no scientific impact as measured by the h-index. The highest h-index we recorded was 86.

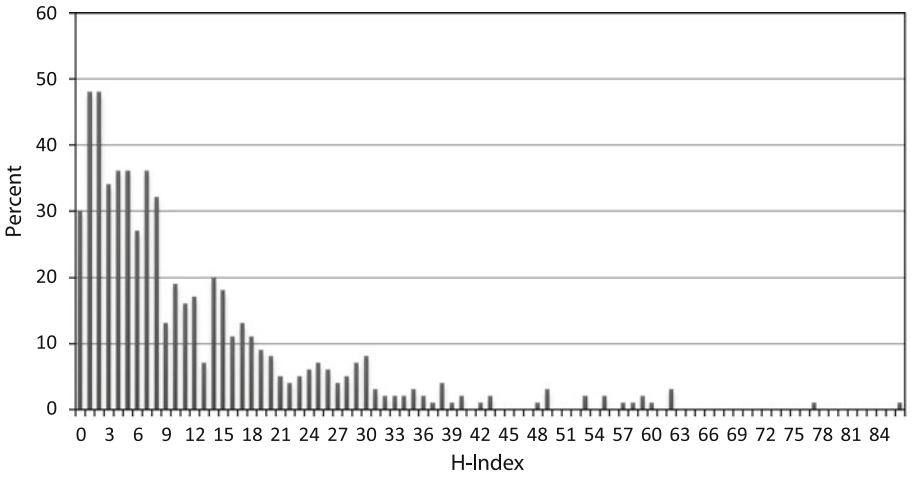

Fig. 2 Distribution of h-index in a random sample of 594 authors from the Web of Science

The h-index distribution in Fig. 2 is skewed to the right with a mean of 11.7 (sd 12.5). As one would expect, higher h-indexes become increasingly rare, although this distribution is somewhat uneven. This unevenness may reflect differences between disciplines where there are varying standards in team size. Another unexpected result is that the mode is not 0. Although we expected the typical case to be that authors publish but are not cited, the modal value was tied between 1 and 2. This may reflect the tendency for authors to cite their own previous work in at least one subsequent publication.

Next we turned to the task of creating the egocentric networks for the authors. Our objective was to map the co-authored publications between every pair of co-authors. One of the network variables that we hypothesized might affect a focal author's h-index was the scientific stature of the co-authors, as measured by the h-index. Thus it was necessary to determine the h-index not only for the 594 focal authors (network egos), but also for all of their co-authors (network alters). A pilot with a few of the authors made it clear that it would be cost-prohibitive to do this with all 594 egos. We decided to sample 250 egos from the list of 594 for the remaining analyses.

An important simplification we implemented was that for all focal authors we only considered the co-authors in articles that contributed to their h-index. Eleven authors had an h-Index of 0 and were dropped from the remainder of the analysis. Four authors were found to have a co-author network of greater than 340 (one had almost 4,000). For those four we selected a random sample of 50 of their co-authors to be used in our analysis. Further Web of Science searches were then used to determine how many times each focal author's coauthors had published with each other in order to build a relational matrix for each author. The mean number of alters was 38.8. In each matrix the cell intersecting two co-authors was populated with the number of articles they had published together. Each matrix was symmetric (undirected). We also created an attribute matrix for each co-author consisting of the variables described in the next section. These data were individually loaded into Borgatti, Everett and Freeman (2002) for network analysis.

Variables

With the goal of explaining the h-index as the dependent variable, we selected a set of independent variables that represented types of collaborative behavior we wanted to test (Table 1). Network size (Netsize), reflects the total number of authors published with, while average authors per article (AvgAuthors) reflects the size of the teams the author typically worked with. We hypothesize that the h-index will increase as network size and average authors per paper increase.

We used several variables in an attempt to capture the structural properties of the co-author networks. Figure 3 shows four examples of structural variability as it relates to the h-index. Both low-h examples show cohesiveness compared to the high-h examples which show sub-grouping. One way of measuring sub-groupings is with components, which are sets of nodes that are tied, either directly or indirectly. If two groups of co-authors are in separate components it means they never publish together and are only linked through the focal author. This would not likely arise from work conducted in one lab as there would tend to be overlap between authors moving through the lab. It would instead reflect a conscious effort on the part of the focal author to work with separate research communities. We hypothesized that the h-index will increase with the number of components of size three or more; that is ego maintains separate communities of co-authors. Components of size one are called isolates. We analyzed this as a different variable because the behavior

Table 1 Description of explanatory variables reflecting an author's collaborative behavior

Variable name	Description	Behavior—Publish with…	Transformation
Number of co-authors			
Netsize	Number of authors across all h-index articles	Many different authors	Logarithmic
AvgAuthors	Average authors per article	Large team	Logarithmic (Var+1)
Structure of collaborations			
Components	Number of components with ego removed	Disconnected groups	None
Isolates	Number of isolates with ego removed	Disconnected co-authors	Removed outlier with 37 isolates
Betweenness	Normalized mean betweenness	Different connected groups	Logarithmic
Hierarchy	Extent to which co-authors are brokered by single co-author	A highly brokering co-author	None
MeanTie	Average number of articles published between co-authors	Co-authors who are prolific	Logarithmic
Characteristics of co-authors			
Academic	Proportion co-authors in academic setting	Academics	None
MeanAlterH	Average h-index of co-authors	High h-index co-authors	None
MaxAlterH	Maximum h-index among co-authors	One high h-index co-author	Logarithmic
HofMostEVC	H-index of most eigenvector central co-author	One high h-index co-author who is highly connected	Logarithmic

of working with multiple single disconnected authors is in many ways more flexible than working with groups, but may be more resource intensive. By working with many different isolated co-authors the focal author may maximize their exposure to different research areas. We hypothesize that as isolates increase the h-index will increase.

Node betweenness is a widely used network metric that measures the extent to which the network exhibits brokering. The betweenness centrality of a given node is the number of shortest paths it lies on after calculating the shortest path between every pair of nodes. Nodes high in betweenness are in the position to broker information and ideas. A network with high mean normalized betweenness would have different groups (like components) without the condition that they are completely disconnected. In other words, a focal author might choose to collaborate with a set of authors within a research community who are loosely connected. We hypothesize that as betweenness increases the h-index will increase.

Hierarchy is an egocentric network measure developed by Burt (1992). Unlike components, isolates and betweenness where the focal author (ego) is removed from the graph, with hierarchy they are included. High levels of hierarchy indicate that the network tends to be dominated structurally by one or a few nodes who are not the focal author. This would occur if, for example, an author worked with a prolific co-author who tended to work with all of the focal author's co-authors. We consider hierarchy a measure of what we call the Godfather Effect; that is working with a highly productive and cited co-author. We hypothesize that as hierarchy increases, the h-index will increase as the focal author is associated with highly cited articles.

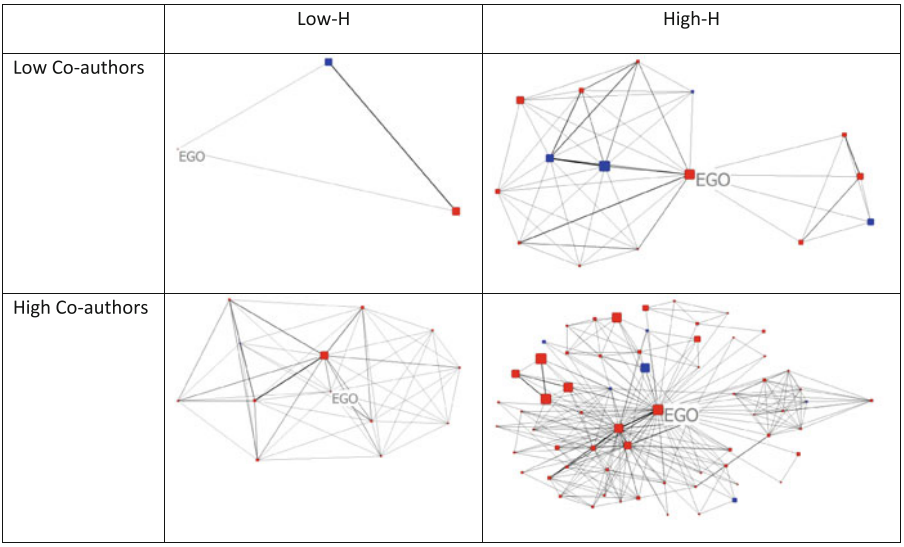

Fig. 3 Examples of structural configurations across low and high H and co-authors

Mean tie strength is a measure of interaction between co-authors. We hypothesize that as mean tie strength increases the h-index will increase, reflecting more publication.

One of our research objectives was to test the extent to which publishing with non-academics would affect the transmission of knowledge and scientific impact. We hypothesize that extending collaboration to the non-academic sector could lead to more impact in a variety of ways. Non-academics often have access to practical data and circumstances to test findings that academics cannot replicate in a lab or through grant funding. Non-academics also may synthesize information that circulates in the academic community with applied experiences to create new types of knowledge. Those in the academic sector may provide new funding sources and potentially lead to new avenues for data dissemination. Therefore we hypothesized that higher proportions of non-academics in the co-author network would lead to a higher h-index. Non-academics are operationalized as those with a non-academic affiliation; that is not at a university or college.

The mean alter h-index is an overall measure of the productivity and impact of the focal author's publication context. The maximum h-index of the focal author's co-authors is an alternative measure of the Godfather Effect, and perhaps a more direct measure. We hypothesized that both would be positively related to the h-index of the focal author.

The last variable in Table 1 is the h-index of the most eigenvector central co-author (HofMostEVC). Eigenvector centrality is a measure of the social position of a node that uses the valued data; that is the number of articles two authors share, rather than the presence or absence of a relationship. Eigenvector centrality measures how connected a node is to other nodes that are connected. It is focused more on position within the entire network structure rather than the local network structure. The h-index of the most eigenvector central co-author is thus a third measure of the Godfather Effect. It is the only variable that combines the structural property of a co-author with an attribute. We hypothesize that larger values of this variable will be associated with higher h-indexes for the focal author.

Although the network data were symmetric (the tie between authors is simply the number of articles on which they collaborated), they were not dichotomous (the ties between authors were valued data, not binary). Many of the network measures we used (betweenness, components, isolates and hierarchy) are graph-based and require transformation to binary data before calculation (Wasserman and Faust 1994). This raises questions as to the appropriate level to define whether a tie exists. We experimented with different strategies, but ultimately decided to use a cutoff of two articles, as the mean number of articles co-authors had published together was 2.18. Glanzel (2012) suggests an alternative approach to defining the cut-off for an edge based on the h-index of the entire graph. This approach is particularly useful for identifying core nodes in a whole network, such as a set of authors or papers within a discipline. In this study we were comparing egocentric networks, which would have resulted in different h-indexes for each graph, and thus a different cutoff for each graph. We decided to use the same cutoff to improve comparability.

Table 1 also shows the data transformations that were made to each variable to correct for non-normal distributions. Most required either no transformation or a log transformation. Average authors per article had a log transformation of the variable plus 1 as we could not calculate the log of 0. The variable isolates presented a very special case. The variable was normally distributed with the exception of one very large outlier. This represented a transplant specialist who had published with many single authors. This is a common circumstance with bio-statisticians who provide assistance with analysis as a co-author and are therefore on many unrelated articles. In this case we decided to remove the isolate.

Table 2 shows the correlation matrix between the eleven independent variables. There are three correlations that appear particularly high, introducing the possibility of multicollinearity. These are network size with the maximum alter h-index, maximum alter h-index with mean alter h-index and mean alter h-index with the h-index of the most eigenvector central alter. In the following models we decided to remove the maximum alter h-index and the h-index of the most eigenvector central alters, leaving nine variables in the model.

Results

The results of our regression models are presented in Table 3. The first column shows bivariate models for each of the nine variables with the focal author's h-index. The important role network size will play in our model is very clear. The second column shows a multivariate model using only the variables that were significant at $p \leq 0.05$ in the bivariate models. Of these five, only four were significant when included in the model together. Components, one of our key structural variables, was not significant in the multivariate model.

The final model is depicted in the third column of Table 3. The overall R^2 for the model was 0.69. Fifty-nine percent of this variance is explained by network size alone. Of the remaining three variables only hierarchy reflects a structural property of the network. Recall that this is the extent to which the author publishes with a co-author who tends to publish with the other authors. We think of this variable as representing the Godfather Effect, rather than a structural collaborative behavior such as components or betweenness that would reflect collaborations across groups.

Table 2 Correlation matrix between the eleven independent variables

	AvgAuthors	Components	Isolates	Betweenness	Hierarchy	MeanTie	Academic	MeanAlterh	MaxAlterh	HofMostEVC
Netsize	0.3581	0.2128	0.0695	−0.1108	0.3151	−0.3539	−0.0885	0.5258	0.7595	0.593
AvgAuthors		−0.1195	−0.0607	−0.226	−0.1241	0.3116	−0.0674	0.3074	0.2715	0.288
Components			0.161	0.2629	0.1462	−0.3478	−0.0059	0.1269	0.2363	0.194
Isolates				−0.0188	0.1172	−0.2392	−0.0707	0.0181	0.0508	0.0814
Betweenness					0.1066	−0.1779	0.042	−0.0572	0.0004	−0.0629
Hierarchy						−0.2079	0.0791	0.1332	0.2504	0.1783
MeanTie							0.0289	0.0112	−0.2465	−0.1693
Academic								0.0568	−0.0008	0.0226
Meanalterh									0.774	0.701

Table 3 Results of regression models

	Bivariate models			Multivariate model		Final model		
	Coefficient	Prob > |t|	R-square	Coefficient	Prob > |t|	Coefficient	Prob > |t|	Partial R-square
Netsize	0.73	0.001	0.60	0.51	0.0001	0.51	0.0001	0.59
AvgAuthors	0.03	0.7526	0	–	–			
Components	0.32	0.0001	0.09	0.07	0.0784	–	–	
Isolates	0.08	0.2992	0	–	–			
Betweenness	0.15	0.1360	0	–	–			
Hierarchy	2.93	0.0001	0.24	1.53	0.0001	1.55	0.0001	0.07
Meantie	−0.54	0.0001	0.19	−0.21	0.0001	−0.23	0.0001	0.02
Academic	0.32	0.2991	0	–	–			
Meanalterh	0.08	0.0001	0.21	0.024	0.0019	0.02	0.001	0.01

Discussion

Our objective with this article was to understand which collaborative behaviors contributed most to the scientific impact of a focal author across all disciplines represented in the Web of Science. We were particularly interested in the egocentric network properties of co-author networks, specifically whether actively seeking groups of loosely connected (betweenness) or disconnected co-authors (components and isolates) improved impact. While we hypothesized that maintaining separate communities would enhance h-index, that was not borne out by the data. We expected the total number of co-authors (network size) to be a contributing variable, and it was. Yet we were surprised by the magnitude of this effect. Simply put network size matters. We believe this finding has far-reaching implications.

Many disciplines, particularly in the social sciences, value sole authorship. Indeed, some academic departments expect one or more sole-authored publications for consideration for tenure and promotion. This is not the case in other disciplines. Despite the high value placed on sole-authorship as a measure of scientific independence, this will more often than not result in lower scientific impact, at least as measured by the h-index. As stated earlier in this article, the tendency for those who publish to cite their own work will favor the impact of someone who publishes with many people. While one can debate whether wide citations are a measure of scientific impact, it is hard to argue that widely cited work does not have a higher probability of future citations. Sole-authored publications must rely more on other channels to spread impact, such as notoriety through word-of-mouth, conference presentations, teaching and consulting or other applied efforts.

In addition to network size, hierarchy and mean alter h-index were also significant. These are variables representing the publication patterns of the co-authors. The positive correlation with hierarchy suggests there are citation rewards for working with a Godfather or Godmother who publishes with many of the other co-authors. This may represent the circumstance where a successful principal investigator has a large and well-funded laboratory working on a program of research where faculty, postdocs, and graduate students tend to cite each other. The positive coefficient with mean alter h-index suggests that focal authors benefit from publishing with co-authors who have high impact.

The negative but significant association between the mean number of articles co-authors share is very likely related to network size. The larger the network, the more likely it is that ties between any pair of co-authors would be null, thus bringing down the average. It is worth noting that mean number of articles and network size were negatively correlated in Table 2, but not to such an extent that it warranted exclusion from the model.

Conclusion

Based on these results the best advice for increasing scientific impact is to publish with as many co-authors as possible with a preference towards co-authors who are already highly cited. Given the relationship with hierarchy, those in fields such as biochemistry with large labs will be rewarded by publishing with their principal investigator. It is difficult to imagine how that behavior can be systematically replicated in some social science disciplines where authors do not tend to work in teams.

Limitations

This study has several limitations. As has already been mentioned, the h-index does not consider some forms of publishing that may lead to impact. The h-index is typically calculated from articles indexed in a database such as the Web of Science. Like many such databases the Web of Science does not index all journals, particularly those that do not rely on peer-review. The Web of Science also excludes books and book chapters which are a key channel of output for some disciplines. The h-index may therefore vary based on the source for the citation data (Bar-Ilan 2008).

Our study compared h-indexes across disciplines. Hirsch pointed out that comparison of h-indexes across disciplines may not be valid as they differ in terms of the types of publications they value and expectations for number and length of articles. In this study we were looking for regularities across disciplines. We believe that the overwhelming effect of network size would exist in any discipline, regardless of other cultural factors. The sample size for this study was relatively small for the network variables. Perhaps advances in creating algorithms to disambiguate will be able to calculate these network variables with little expense and more accurately. For this study the size was constrained by the cost of collecting the network data. Despite the small sample size we found some clear effects. These effects were found using variables that were highly skewed and required transformation.

References

Abbasi, A., Chung, K., & Hossain, L. (2011). Egocentric analysis of co-authorship network structure, position and performance. *Information Processing and Management, 48*, 671–679.

Abbott, A., Cyranoski, D., Jones, N., Maher, B., Schiermeier, Q., & Van Noorden, R. (2010). Metrics: do metrics matter? *Nature, 465*, 860–862.

Adams, J. D., Black, G. C., Clemmons, J. R., & Stephan, P. E. (2005). Scientific teams and institutional collaborations: evidence from US universities, 1981–1999. *Research Policy, 34*, 259–285.

Anderson, T. R., Hankin, R. K. S., & Killworth, P. D. (2008). Beyond the Durfee square: enhancing the h-index to score total publication output. *Scientometrics, 76*, 577–588.

Banks, M. G. (2006). An extension of the Hirsch index: indexing scientific topics and compounds. *Scientometrics, 69*, 161–168.
Bar-Ilan, J. (2008). Which h-index?—A comparison of WoS, Scopus and Google Scholar. *Scientometrics, 74*, 257–271.
Batagelj, V., & Mrvar, A. (2000). Some analyses of Erdos collaboration graph. *Social Networks, 22*, 173–186.
Batista, P. D., Campiteli, M. G., Kinouchi, O., & Martinez, A. S. (2006). Is it possible to compare researchers with different scientific interests? *Scientometrics, 68*, 179–189.
Beaver, D. D., & Rosen, R. (1979). Studies in Scientific Collaboration 2. Scientific co-authorship, research productivity and visibility in the French scientific elite, 1799–1830. *Scientometrics*, 1:133–149.
Becher, T. (2006). Disciplinary discourse. *Studies in Higher Education, 12*(3), 261–274.
Becher, T., & Trowler, P. (2001). *Academic tribes and territories: intellectual enquiry and the culture of disciplines*. Buckingham: SRHE and Open University Press.
Borgatti, S. P., Everett, M. G., & Freeman, L. C. (2002). *Ucinet for windows: software for social network analysis*. Harvard: Analytic Technologies.
Börner, K., Dall'Asta, L., Ke, W., & Vespignani, A. (2005). Studying the emerging global brain: analyzing and visualizing the impact of co-authorship teams. *Complexity, Special issue on Understanding Complex Systems, 10*, 57–67.
Bornmann, L., & Daniel, H. D. (2005). Does the h-index for ranking of scientists really work? *Scientometrics, 65*, 391–392.
Bornmann, L., & Daniel, H. D. (2007). What do we know about the h index? *Journal of the American Society for Information Science and Technology, 58*, 1381–1385.
Bornmann, L., Mutz, R., & Daniel, H. D. (2008). 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*, 830–837.
Burrell, Q. L. (2007). Hirsch index or Hirsch rate? Some thoughts arising from Liang's data. *Scientometrics, 73*, 19–28.
Burt, R. S. (1992). *Structural Holes: the social structure of competition*. Cambridge: Harvard University Press.
Collaboration, The Atlas, Aad, G., Abat, E., Abdallah, J., Abdelalim, A. A., Abdesselam, A., Abdinov, O., Abi, B. A. et al. (2008). "The ATLAS experiment at the CERN large hadron collider". Journal of Instrumentation 3 (08).
Costas, R., & Bordons, M. (2007). The h-index: advantages, limitations and its relation with other bibliometric indicators at the micro level. *Journal of Informetrics, 1*, 193–203.
Crane, D. (1972). *Invisble colleges: diffusion of knowledge in scientific communities*. Chicago: University of Chicago Press.
Cronin, B., & Meho, L. (2006). Using the h-index to rank influential information scientists. *Journal of the American Society for Information Science and Technology, 57*, 1275–1278.
de Castro, R., & Grossman, J. W. (1999). Famous trails to Paul Erdos. *Mathematical Intelligencer, 21*, 51–63.
Defazio, D., Lockett, A., & Wright, M. (2009). Funding incentives, collaborative dynamics and scientific productivity: evidence from the EU framework program. *Research Policy, 38*, 293–305.
Eaton, J. P., Ward, J. C., Kumar, A., & Reingen, P. H. (1999). Structural analysis of co-author relationships and author productivity in selected outlets for consumer behavior research. *Journal of Consumer Psychology, 8*, 39–59.
Egghe, L. (2007). Dynamic h-index: the Hirsch index in function of time. *Journal of the American Society for Information Science and Technology, 58*, 452–454.
Einstein, A., & Rosen, N. (1936). Two-body problem in general relativity theory. *Physical Review, 49*(5), 0404–0405.
Frenken, K., Holzl, W., & de Vor, F. (2005). The citation impact of research collaborations: the case of European biotechnology and applied microbiology (1988–2002). *Journal of Engineering and Technology Management, 22*, 9–30.
Glanzel, W. (2002). Coauthorship patterns and trends in the sciences (1980–1998): a bibliometric study with implications for database indexing and search strategies. *Library Trends, 50*, 461–473.
Glanzel, W. (2006). On the h-index—A mathematical approach to a new measure of publication activity and citation impact. *Scientometrics, 67*, 315–321.
Glanzel, W. (2012). The role of core documents in bibliometric network analysis and their relation with h-type indices. *Scientometrics, 93*, 113–123.
Goldfinch, S., Dale, T., & DeRousen, K. (2003). Science from the periphery: collaboration, networks and 'periphery effects' in the citation of New Zealand Crown Research Institute articles, 1995–2000. *Scientometrics, 57*, 321–337.

He, Z. L., Geng, X. S., & Campbell-Hunt, C. (2009). Research collaboration and research output: a longitudinal study of 65 biomedical scientists in a New Zealand university. *Research Policy, 38*, 306–317.

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

Hirsch, J. E. (2007). Does the h index have predictive power? *Proceedings of the National Academy of Sciences of the United States of America, 104*, 19193–19198.

Hirsch, J. E. (2010). An index to quantify an individual's scientific research output that take into account the effect of multiple coauthorship. *Scientometrics, 85*, 741–754.

Hou, H., Kretschmer, H., & Liu, Z. (2008). The structure of scientific collaboration networks in Scientometrics. *Scientometrics, 75*, 189–202.

Iglesias, J. E., & Pecharromán, C. (2007). Scaling the h-index for different scientific ISI fields. *Scientometrics, 73*, 303–320.

Imperial, J., & Rodríguez-Navarro, A. (2007). Usefulness of Hirsch's h-index to evaluate scientific research in Spain. *Scientometrics, 71*, 271–282.

Jones, B. F., Wuchty, S., & Uzzi, B. (2008). Multi-university research teams: shifting impact, geography, and stratification in science. *Science, 322*, 1259–1262.

Katz, J. S., & Hicks, D. (1997). How much is a collaboration worth? A calibrated bibliometric model. *Scientometrics, 40*, 541–554.

Katz, J. S., & Martin, B. R. (1997). What is research collaboration? *Research Policy, 26*, 1–18.

Kelly, C. D., & Jennions, M. D. (2006). The h index and career assessment by numbers. *Trends in Ecology & Evolution, 21*, 167–170.

Knorr, K. D., & Mittermeir, R. (1980). Publication productivity and professional position—cross-national evidence on the role of organizations. *Scientometrics, 2*, 95–120.

Kretschmer, H. (2004). Author productivity and geodesic distance in bibliographic co-authorship networks, and visibility on the Web. *Scientometrics, 60*, 409–420.

Lazega, E., Jourda, M.-T., Mounier, L., & Stofer, R. (2008). Catching up with big fish in the big pond? Multi-level network analysis through linked design. *Social Networks, 30*, 159–176.

Lazega, E., Mounier, L., Jourda, M.-T., & Stofer, R. (2006). Organizational vs. personal social capital in scientists' performance: a multi-level network study of elite French cancer researchers (1996–1998). *Scientometrics, 67*, 27–44.

Lee, S., & Bozeman, B. (2005). The impact of research collaboration on scientific productivity. *Social Studies of Science, 35*, 673–702.

Leimu, R., & Koricheva, J. (2005). Does scientific collaboration increase the impact of ecological articles? *BioScience, 55*, 438–443.

Liang, L. M. (2006). h-index sequence and h-index matrix: constructions and applications. *Scientometrics, 69*, 153–159.

Melin, G. (2000). Pragmatism and self-organization—research collaboration on the individual level. *Research Policy, 29*, 31–40.

Moody, J. (2004). The structure of a social science collaboration network: disciplinary cohesion from 1963 to 1999. *American Sociological Review, 69*, 213–238.

Moravcsik, M. J. (1988). Citation context classification of a citation classic concerning citation context classification. *Social Studies of Science, 18*, 515–521.

Mulchenko, Z. M., Granovsky, Y. V., & Strakhov, A. B. (1979). Scientometrical characteristics on information activities of leading scientists. *Scientometrics, 1*, 307–325.

Narin, F., Stevens, K., & Whitlow, E. S. (1991). Scientific co-operation in Europe and the citation of multinationally authored papers. *Scientometrics, 21*, 313–323.

Nemeth, C. J., & Goncalo, J. A. (2005). Creative collaborations from afar: the benefits of independent authors. *Creativity Research Journal, 17*, 1–8.

Newman, M. E. J. (2004). Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences of the United States of America, 101*, 5200–5205.

Onodera, N., Iwasawa, M., Midorikawa, N., Yoshikane, F., Amano, K., Ootani, Y., et al. (2011). *Journal of the American Society for Information Science and Technology, 62*(4), 677–690.

Pao, M. L. (1982). Collaboration in computational musicology. *Journal of the American Society for Information Science, 33*, 38–43.

Persson, O., Glanzel, W., & Danell, R. (2004). Inflationary bibliometric values: the role of scientific collaboration and the need for relative indicators in evaluative studies. *Scientometrics, 60*, 421–432.

Radicchi, F., Fortunato, S., & Castellano, C. (2008). Universality of citation distributions: toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences of the United States of America, 105*, 17268–17272.

Rodgers, R. C., & Maranto, C. L. (1989). Causal-models of publishing productivity in psychology. *Journal of Applied Psychology, 74*, 636–649.
Roediger, H. L. III. (2006). The h index in science: a new measure of scholarly contribution. *The Academic Observer* 19.
Rousseau, R. (2007). The influence of missing publications on the Hirsch index. *Journal of Informetrics, 1*, 2–7.
Rousseau, R. (2008). Reflections on recent developments of the h-index and h-type indices. *COLLNET Journal of Scientometrics and Information Management, 2*, 1–8.
Saad, G. (2006). Exploring the h-index at the author and journal levels using bibliometric data of productive consumer scholars and business-related journals respectively. *Scientometrics, 69*, 117–120.
Schreiber, M. 2007. Self-citation corrections for the Hirsch index. *Epl* 78.
Schubert, A. (2012). A Hirsch-type index of co-author partnership ability. *Scientometrics, 91*, 303–308.
Schubert, A., Korn, A., & Telcs, A. (2009). Hirsch-type indices for characterizing networks. *Scientometrics, 78*, 375–382.
Sidiropoulos, A., Katsaros, D., & Manolopoulos, Y. (2007). Generalized Hirsch h-index for disclosing latent facts in citation networks. *Scientometrics, 72*, 253–280.
Skilton, P. F. (2009). Does the human capital of teams of natural science authors predict citation frequency? *Scientometrics, 78*, 525–542.
Swarna, T., Kalyane, V. L., & Kumar, V. (2008). Homi Jehangir Bhabha: his collaborators, citation identity, and his citation image makers. *Malaysian Journal of Library & Information Science, 13*, 49–67.
Tang, L., & Walsh, J. P. (2010). Bibliometric fingerprints: name disambiguation based on approximate structure equivalence of cognitive map. *Scientometrics, 84*(3), 763–784.
Van Raan, A. F. J. (2006). Comparison of the Hirsch-index with standard bibliometric indicators and with peer judgment for 147 chemistry research groups. *Scientometrics, 67*, 491–502.
Vanclay, J. K. (2007). On the robustness of the h-index. *Journal of the American Society for Information Science and Technology, 58*, 1547–1550.
Walters, G. D. (2006). Predicting subsequent citations to articles published in twelve crime-psychology journals: author impact versus journal impact. *Scientometrics, 69*, 499–510.
Wasserman, S., & Faust, K. (1994). *Social network analysis: methods and applications. Structural analysis in the social sciences*. New York: Cambridge University Press.
White, H. D. (2000). "Toward ego-centered citation analysis", In: B. Cronin & H. B. Atkins (Eds.), *The web of knowledge: a Festschrift in honor of Eugene Garfield*. pp. 475–496. Medford, NJ: Information Today.
White, H. D. (2001). Author-centered bibliometrics through CAMEOs: characterizations automatically made and edited online. *Scientometrics, 50*, 607–637.
White, H. D., & McCain, K. W. (1998). Visualizing a discipline: an author co-citation analysis of information science, 1972–1995. *Journal of the American Society for Information Science, 49*, 327–355.
Yoshikane, F., & Kageura, K. (2004). Comparative analysis of coauthorship networks of different domains: the growth and change of networks. *Scientometrics, 60*, 433–444.
Zhivotovsky, L. A., & Krutovsky, K. V. (2008). Self-citation can inflate h-index. *Scientometrics, 77*, 373–375.
Zuckerman, H. (1967). Nobel laureates in science—patterns of productivity, collaboration, and authorship. *American Sociological Review, 32*, 391–403.