The Rich Get Richer: Studying Scholarly Impact in the Emerging Field of Information Visualization

Weimao Ke
School of Information and Library Science
University of North Carolina at Chapel Hill
wke@unc.edu

ABSTRACT

The paper reports on an investigation of the-rich-get-richer effect of scholarly communication in the emerging field of Information Visualization. A dataset containing 31 years' representative publications is used to analyze scholarly impact in terms of citation scores. Rich factors, i.e., variables that carry previous citation scores, are closely examined and their contributions to future citations measured. Based on previous research on citation patterns, a general log-linear regression model is proposed and applied to the prediction of scholarly impact using the rich factors. The analysis supports the "preferential attachment" property, or the-rich-get-richer phenomenon, in citation networks and reveals that the number of citations one has received largely explains the magnitude of future rewards. The implication is that citation-based evaluation of scholarly impact is biased. The large coefficient of determination (R^2) found in the current analysis, to be verified in other domains, is too significant to ignore. This invites thoughts on how Information Science domains like InfoVis can maintain research momentum by rewarding recognized scholars while encouraging new players and novelty.

Categories and Subject Descriptors

H.1.0 [**Information Systems**]: MODELS AND PRINCIPLES—*General*

General Terms

Measurement, Verification

Keywords

scholarly impact, citation analysis, the rich get richer, regression, citation network, preferrential attachment

1. INTRODUCTION AND PURPOSE

Among measures used to analyze various facets of written communication, citations are considered *fair* indicators of use and usefulness of publications [7]. Citation analysis has been widely used to evaluate research productivity and scholarly impact [3].

Previous research on network science [1] and citation analysis [8] proposed that citation networks are small-world networks and subject to "preferential attachment." These networks are scale free – that is, they have a distribution of connectivities that decays with a power law function [1]. In another word, the majority are rarely cited and extremely "poor" whereas highly-cited nodes, i.e., influential publications or scholars, tend to be cited even more – the rich get richer.

The-rich-get-richer effect has various potential implications in the development of a field and requires further empirical investigation. This study, as a priliminary step, aims to verify the the-rich-get-richer effect by directly measuring the "rich" and the "richer" in the emerging field of Information Visualization. It sets out to discover what the "rich" factors are in existing data and whether they help scholarly publications "get richer" over time. A prediction model for scholarly impact will be proposed and applied in the field to evaluate the existence and significance of the effect.

2. DATA COLLECTION

Data come from the IEEE Information Visualization 2004 Contest, now part of the InfoVis Benchmark Repository [6]. The dataset contains major publications and citations within the field over a 31-year period, i.e., from 1974 to 2004, retrieved from the ACM Digital Library. The publication metadata come with title, authors, abstract, keywords, source, references, number of pages, and year of publication. After data cleaning, the dataset contains 613 publications with 1,036 unique authors/scholars and 8,502 references to papers within and without the set.

3. MEASURING IMPACT

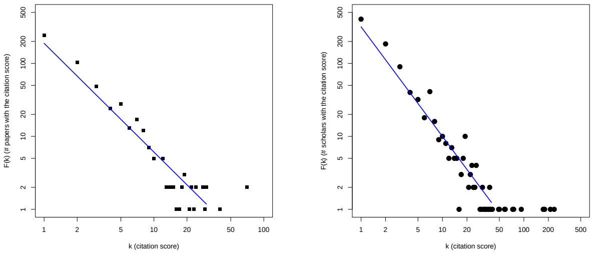
Surely, it is difficult to define and measure impact precisely. For simplicity and data availability, the number of citations one has received is used as an indicator of scholarly impact. In this analysis, a paper's # citations: $C_{paper|i}$ = # papers that cite paper i. This citation score is then divided equally among its authors, i.e., $C_{scholar|ij} = C_{paper|i}/n_i$ for author j of paper i where n_i is the number of authors of the paper.

3.1 Scale-free Citation Network

Research proposed that the-rich-get-richer effect appears in networks that follow a power-law distribution, that is, the probability $P(k)$ that a node in the network connects with k other nodes (e.g., cited by k other papers in a citation network) is roughly proportional to $k^{-\gamma}$ [1]. This type of network is called scale-free because the network structure is independent of its scale of size.

Figure 1 (a) and (b) show distributions of # citations to scholars and publications in the dataset on log-log coordinates, where the X axis denotes citation scores and Y frequencies of the scores. Both figures show a power-law pattern – a roughly linear decreasing trend on log-log coordinates.

For instance, on Figure 1 (a) the distribution of paper citation scores, when X (citation score) is very small (e.g., 1), Y (frequency) is extremely large. That means, the vast majority of the papers have very small citation scores. On the other hand, when X (citation score) is extremely large, Y (frequency) is very small, meaning that only a tiny portion of the papers have obtained large citation scores. In another word, very few are extremely rich and the majority are extremely poor.



(a) paper citation

(b) scholar citation

Figure 1: Power-law distributions of citation scores (straight lines a reference to the eye)

3.2 Log Transformation of Citation Scores

The linear distributions on log-log coordinates suggest a log transformation of citation scores. Preliminary analysis using the Box-cox test [2] also supports the usefulness of a log transformation. As we will see in Section 5, most of the predictor variables will involve previous # citations in various ways. Therefore, log-transformation of x (independent) variables also appears reasonable. All this leads to the use of a log-linear regression model in Section 4.

4. MODEL

The objective of the study is to build a model for scholarly impact and to use regression analysis to study the-rich-get-richer effect. As suggested in the discussion above, log transformation of dependent and independent variables is desirable for this model, which is presented below.

Let y_i be the number of citations of an observation (paper) i. Suppose we have independent variables $x_{ij} \in [x_{i1}, x_{i2}, .., x_{ip}]$ where p is the number of independent variables. The proposed model is:

$$y_i = (\beta_0^* * \prod_{j=1}^{p} x_{ij}^{\beta_j}) * \epsilon^* \tag{1}$$

where ϵ^* is the stochastic disturbance or error term. Applying log function to the both sides of the equation:

$$\begin{aligned} log(y_i) &= log((\beta_0^* * \prod_{j=1}^{p} x_{ij}^{\beta_j}) * \epsilon^*) & (2) \\ &= log(\beta_0^*) + \sum_{j=1}^{p}(\beta_j * log(x_{ij})) + log(\epsilon^*) & (3) \end{aligned}$$

Let $log(\beta_o^*)$ be β_0 and $log(\epsilon^*)$ be ϵ, the model becomes:

$$log(y_i) = \beta_0 + \sum_{j=1}^{p}(\beta_j * log(x_{ij})) + \epsilon \tag{4}$$

This is a linear regression model for the log-transformed variables. Using generalized linear regression models [5] will allow estimation of the β coefficients, i.e., elasticity of each independent variable's contribution to citations.

5. "RICH" FACTORS

A set of candidate rich factors will be discussed below and used for the model. By "rich," the study refers to variables that carried a certain amount of previous citations associated with the publication being analyzed.

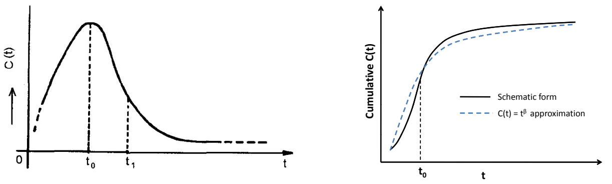
5.1 Rich Factor Selection

In citation networks, the-rich-get-richer proposal implies that a highly cited paper will get more citations in the future. Scholars who have been well recognized will attract particular attention and get cited even more. Similarly, high impact journals will continue to dominate. In addition, scholarly references often indicate knowledge flows from the cited publications to the current work. Publications that attach to the rich ones by referring to them might potentially become richer.

All this has suggested as predictor variables the previous impact factors (the richness) of authors, publication venue, and referred works of a paper (see Section 5.3 for the variables). The model also has to include a time factor, i.e., *age* of a publication, because obviously citations accumulate over time.

5.2 Aging Effect

Research has shown that a citation decay curve over time consists of two parts, as shown in Figure 2 (a): 1) an increase of citations during first couple of years, followed by 2) gradual decline of citations when the paper gets older [4]. Its cumulative form is shown in Figure 2 (b). For model simplicity and consistency, this study uses the functional form of τ^{β}, where τ is age and β an exponent to be estimated, to model accumulative citations. Although this is not the perfect way to model the aging effect, it does capture the decaying pattern of citations over time (see Figure 2 (b)). This also fits in well with the entire model, in which the exponent β can be estimated using log-linear regression.



(a) schematic form (b) cumulative approximation

Figure 2: Citation aging & approximation

5.3 Variables

Previous discussions suggested the following variables with hypotheses stated in Section 6. Dependent variable: $logCite$ - Citation score of a publication as of year 2004. Independent variables: 1) $logAuthCite$ - Previous[1] citations to authors, 2) $logRefCite$ - Previous[1] citations to referred works, 3) $logVnCite$ - Average previous[1] citations to publication venue, and 4) $logAge$ - Age of publication as of year 2004. Note that it was in 2004 the data were collected and all variables will be log transformed.

6. HYPOTHESES

H1: A paper that refers to more famous/highly-cited works potentially attracts more citations.

H2: The more previous citations to a paper's authors, the more may the paper's future citations be.

H3: The more prestigious the venue in which a paper is published, the more citations will the paper gain in the future.

H4: Number of citations of a paper accumulates over time.

7. RESULTS

Regressing $logCite$ against $logAuthCite$, $logRefCite$, $logVnCite$, and $logAge$ produces the results in Table 1, which shows all significant coefficients and supports the four hypotheses. The more previous citations a paper's authors have received, the more citations the paper will gain in the future. A paper that refers to more highly-cited works potentially attracts more citations. The established impact of a publication venue also has a positive effect on future citation scores.

Although both $AuthCite$ and $RefCite$ have positive coefficients, $AuthCite$ seems to contribute more toward citation scores (exponent 0.325 vs. 0.0395). For example, when the previous citation score to authors increases 100% (doubled), while holding the other variables constant, the expected citation score of a paper will increase $2^{0.325} - 1 = 0.2527$, or about 25%. Referring to good works does help the current work get more citations, but to a limited extent. People seem to pay more attention to works written by prestigious scholars than works that refer to influential others.

Table 1: Scholarly Impact: Parameter Estimates

Variable	DF	β Est	StdErr	t	$Pr > \|t\|$	
Intercept	1	-0.335	0.050	-6.7	$4.3E^{-11}$	$***$
logAuthCite	1	0.326	0.020	16.3	$< 2E^{-16}$	$***$
logVnCite	1	0.081	0.037	2.2	0.026	$*$
logRefCite	1	0.040	0.014	2.9	0.004	$**$
logAge	1	0.573	0.048	11.9	$< 2E^{-16}$	$***$

Signif. codes: 0 '$***$' 0.001 '$**$' 0.01 '$*$' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.238 on 608 degrees of freedom
Multiple R-Squared: 0.4726, Adjusted R-squared: 0.4691
F-statistic: 136.2 on 4 and 608 DF, p-value: $< 2.2E^{-16}$

8. CONCLUSION

This study builds a regression model for scholarly impact and uses a dataset containing 31 years of publications to test the-rich-get-richer effect. Results support the claim that success tends to breed success in scholarly communication and that the small number of rich factors (predictor variables) explained a large portion of variance in citation scores. The implication is that citation-based evaluation of scholarly impact is biased. The large coefficient of determination (R^2) found in the current analysis, to be verified in other domains, is too significant to ignore. This invites thoughts on how a domains like InfoVis can maintain research momentum by rewarding recognized scholars while encouraging new players and novelty.

Acknowledgments

I appreciate helpful discussions with Thomas Carsey, Frank Acito, Paul Solomon, John Paolillo and Javed Mostafa, and valuable comments from Cassidy R. Sugimoto, Jeff Harden, Ellen Gutman, Xi Niu and anonymous reviewers.

9. REFERENCES

[1] R. Albert and A.-L. Barabasi. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47–97, 2002.

[2] G. E. P. Box and D. R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society*, 26(2):211–252, 1964.

[3] B. Cronin and K. Overfelt. Citation-based auditing of academic performance. *Journal of the American Society for Information Science*, 45(2):61–72, 1994.

[4] B. M. Gupta. Analysis of distribution of the age of citations in theoretical population genetics. *Scientometrics*, 40(1):139–162, September 1997.

[5] M. H. Kutner, C. J. Nachtsheim, and J. Neter. *Applied Linear Regression Models.* McGraw-Hill/Irwin, revised edition, 2004.

[6] C. Plaisant, J. Fekete, and G. Grinstein. Promoting insight based evaluation of visualizations: From contest to benchmark repository. Technical report, HCIL, University of Maryland, 2004.

[7] A. Pritchard. Statistical bibliography or bibliometrics? *Journal of Documentation*, 25:348–349, 1969.

[8] S. Redner. How popular is your paper? an empirical study of the citation distribution. *The European Physical Journal B - Condensed Matter and Complex Systems*, 4(2):131–134, July 1998.

[1] *Previous* refers to the time before a paper was published.