Innovation or Imitation: The Diffusion of Citations

Chao Min
School of Information Management, Nanjing University, #163 Xianlin Avenue, Nanjing, Jiangsu, China, 210023;
School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN, 47405, USA.
Telephone: +86-15951813844
E-mail: chaomin@iu.edu

Ying Ding
School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN, 47408, USA;
School of Information Management, Wuhan University, Wuhan, Hubei, China, 430072;
University Library, Tongji University, Shanghai, China, 200092.
Telephone: (812) 855-5388
E-mail: dingying@indiana.edu

Jiang Li
Department of Information Resource Management, Zhejiang University, Hangzhou, Zhejiang, China, 310027.
Telephone: +86-18858182670
E-mail: li-jiang@zju.edu.cn

Yi Bu
School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN, 47408, USA;
Telephone: (812) 558-8130
E-mail: buyi@iu.edu

Lei Pei
School of Information Management, Nanjing University, #163 Xianlin Avenue, Nanjing, Jiangsu, China, 210023.
Telephone: +86-13770602872
E-mail: plei@nju.edu.cn

Jianjun Sun
School of Information Management, Nanjing University, #163 Xianlin Avenue, Nanjing, Jiangsu, China, 210023.
Telephone: +86-13905150993
E-mail: sjj@nju.edu.cn

Correspondence concerning this article should be addressed to Dr. Jianjun Sun.

Abstract

Citations in scientific literature are important both for tracking the historical development of scientific ideas and for forecasting research trends. However, the diffusion mechanisms underlying the citation process remain poorly understood, despite the frequent and longstanding use of citation counts for assessment purposes within the scientific community. Here, we extend the study of citation dynamics to a more general diffusion process to understand how citation growth associates with different diffusion patterns. Using a classic diffusion model, we quantify and illustrate specific diffusion mechanisms which have been proven to exert a significant impact on the growth and decay of citation counts. Experiments reveal a positive relation between the "low *p* and low *q*" pattern and high scientific impact. A sharp citation peak produced by rapid change of citation counts, however, has a negative effect on future impact. In addition, we have suggested a simple indicator, *saturation level*, to roughly estimate an individual paper's current stage in the life cycle and its potential to attract future attention. The proposed approach can also be extended to higher levels of aggregation (e.g., individual scientists, journals, institutions), providing further insights into the practice of scientific evaluation.

Keywords: Science of science, diffusion of innovations, diffusion of citations, citation process

Introduction

Like a currency (Yan, Ding, & Cronin et al., 2013) circulating in the scientific community, citations provide a rough measure of academic impact (Moed, 2006; Abramo & D'Angelo, 2016). Citation data, especially the number of citations, have been widely used as a basic metric in many scenarios (Hirsch, 2005; Garfield, 2006; Fersht, 2009) of scientific evaluation. However, the citation process itself (Cronin, 1984), which not only records the trajectories of scientific development (Kuhn, 1962) but also provides implicit clues as to where science will go (Sinatra, Deville & Szell et al., 2015), remains less explored.

Citation is a multifaceted process. For example, Figure 1 shows the citation histories of two imagined papers, both published in the same field and year, with each cited 127 times in the first 15 years post-publication. Intuition tells us that Paper 1 would likely achieve more citations in the future than Paper 2 would, but citation-number-based indicators, such as total citations and average citations per year, would make no distinction between the two papers. This inconsistency is attributable to a trait shared by most currently-used citation metrics, namely, their ignorance of the dynamic nature of citations. A consideration of the problem leads to broader questions: how much do we know about the citation process? Why and how do citations grow and decay? Moreover, how can we safely use citation metrics despite our limited knowledge of this process? Although the citation process may be difficult to predict accurately, we conduct an exploratory study with the help of a quantitative model to broaden our

understanding of the above issues.

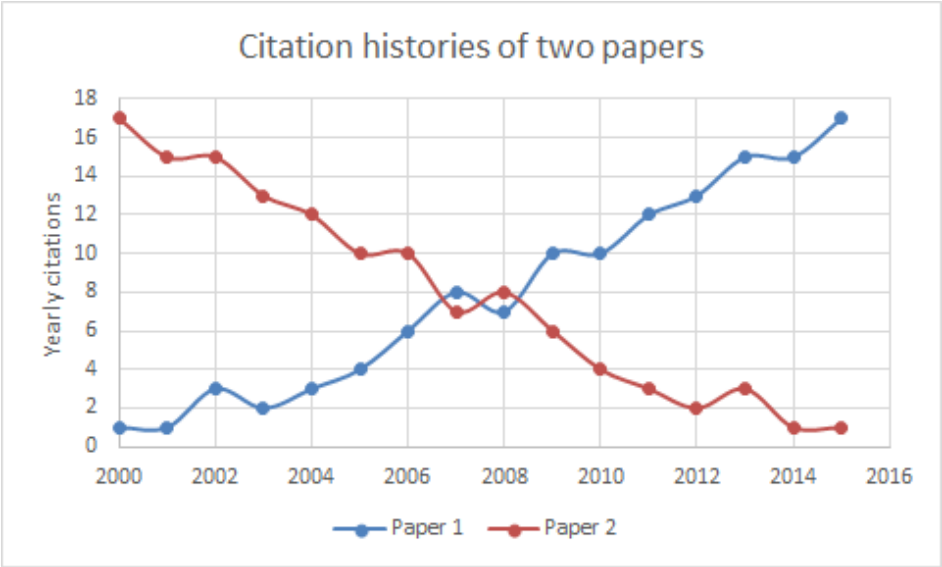

Figure 1: Two imagined papers with identical total citations but distinct citation patterns

In fact, the citation process is closely related to a well-established topic of research: the diffusion of innovations (Rogers, 1995). Consumer adoption of an innovative new product, for example, has long been of both practical and academic interest to scientists. Research assumes that the adoption of a new product is mainly driven by two mechanisms that are related to different buying motives among potential buyers, namely, the *innovation mechanism* and the *imitation mechanism*. The innovation mechanism takes effect when individuals learn of the new product and then decide to buy it irrespective of the influence of others. The imitation mechanism, in contrast, accounts for adoption decisions driven partly by social pressure, which increases with the number of previous adopters. Because of its similarity to the spreading of an epidemic (Bartlett, 1960), the imitation mechanism is often called a contagion effect (Burt, 1987; Young, 2009).

Combining these two effects, Bass (1969) proposed a model that has been used by marketing scientists as a parsimonious way to describe and predict the diffusion of new products. The model involves three important parameters: the innovation coefficient p, the imitation coefficient q, and the market potential m. Historical sales data can be used to estimate these three parameters for a given product, facilitating a rough prediction of future sales.

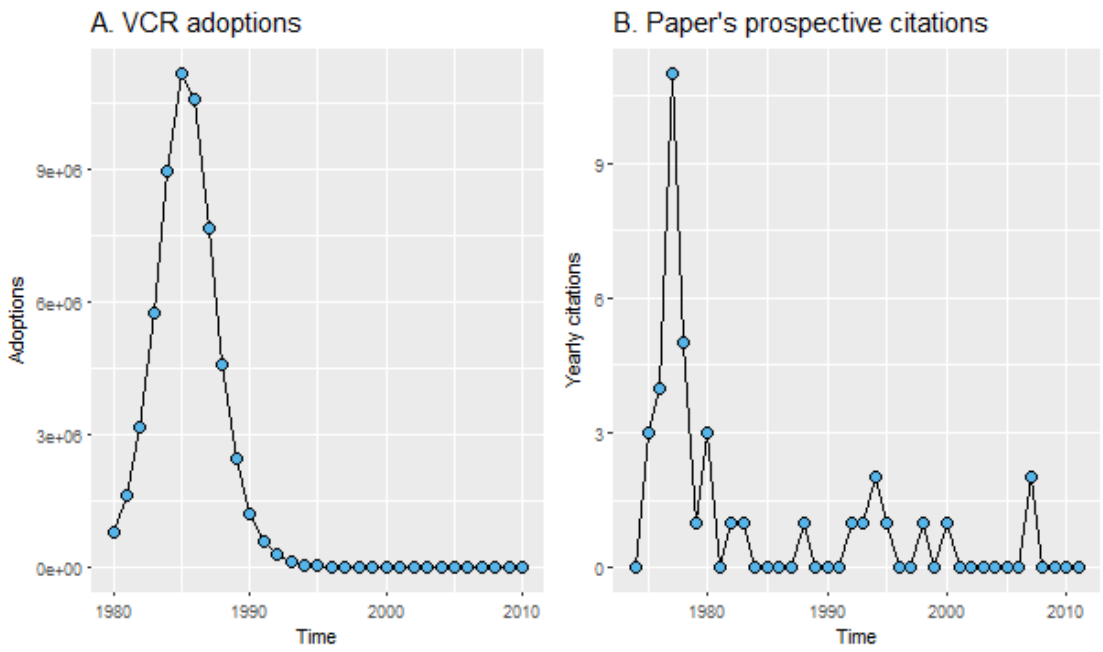

Figure 2: Product sales curve and paper citation curve. (A) Adoptions of VCR in the USA[1]; (B) Yearly citation counts of a scientific article (Flick & Bloch, 1974)

Examination of various diffusion datasets suggests the cumulative adoption and the period-by-period adoptions are usually portrayed respectively by an S-shaped curve and a Bell-shaped one. This naturally suggests an association between innovation diffusion and the citation process, since temporal citation curves resemble these adoption curves (see, for instance, Figure 2). Since the Bass (1969) model has been widely applied in literature to measure the diffusion of such innovations (Lilien, Rangaswamy & Van den Bulte, 2000) as durables and industrial and agricultural technologies, it seems reasonable to suppose that it might also help us understand the spread of scientific ideas. As the trace of scientific development, citation is traditionally regarded as a proxy for scientific impact, but factors aside from a work's intrinsic value (Petersen, Fortunato & Pan et al., 2014) may also affect citation counts, complicating the task of predicting an individual paper's future citation patterns. Nevertheless, quantitative efforts will surely help us better understand the citation process.

We define *citation diffusion* as the process by which a scientific idea expressed in a publication is communicated through certain channels among the members of a scientific community. This definition provides a wide perspective for us to investigate the citation process, making it easier to understand general citation phenomena, including some seemingly puzzling ones, such as "sleeping beauties" in science (van Raan, 2004; Ke, Ferrara & Radicchi et al., 2015; Li & Shi, 2016; Min, Sun & Pei et al, 2016). In this study, we aim to further our understanding of the citation process by analyzing the citation curves of numerous academic publications with the help of the Bass model, as reinterpreted in the novel context of citation. In doing so, we make the following contributions to the literature:

- Introducing diffusion-of-innovations theory as a tool for analyzing the dynamic process of citation;

- Exploring how diffusion patterns influence scientific impact; and
- Providing a simple method for roughly estimating a paper's present status and potential in garnering scientific impact.

Related work

There is a large amount of previous work on the diffusion of innovations as well as on the study of citation trajectories. Efforts focused on the dynamic process of citations from the diffusion perspective, however, are lacking. In this section, we first give a brief introduction to the diffusion-of-innovations theory (Rogers, 1995), then explicate one of the most successful paradigms to emerge from this theory, namely, the diffusion model posited by Bass (1969). Finally, we review previous efforts in modeling citation histories.

The theory of diffusion of innovations

It is often difficult for an innovation, whether a new idea, technology, product, or service, to gain widespread acceptance, even though the innovation may have obvious advantages. The diffusion-of-innovations theory, proposed by Rogers in the 1960s, seeks to understand the spread of innovations such as new concepts, designs, and products. It has offered guidance in analyzing many practical scenarios, such as the marketing of new products (Mahajan, Muller & Wind, 2000), the spread of social change (Wejnert, 2002), and the penetration of public policies (Simmons & Elkins, 2004).

Diffusion, as defined by Rogers (1995), is "the process by which an innovation is communicated through certain channels over time among the members of a social system." Four main elements (Rogers, 1995) can be derived from this definition: the innovation itself, communication channels, time, and the social system. The innovation can be either tangible (e.g., a new product) or intangible (e.g., a new technology or policy). Whether the innovation is objectively novel is not of great importance. Instead, a person's thoughts about its novelty determines whether it will be adopted or not. The characteristics of the innovation can thus play an essential role in the speed of its adoption by the crowd (Tornatzky & Klein, 1982). Rogers differentiates two channels by which the innovation spreads through the social system: mass media and interpersonal communication. Mass media is an efficient channel that plays a decisive role in the early stage of diffusion, disseminating new knowledge to the majority of potential adopters in a short time. Examples include TV, newspapers, and radio broadcasts. Interpersonal communication is also an important channel in which a person is persuaded by companions to adopt the innovation; this channel is especially important when the crowd shares similar (e.g., socioeconomic or sociocultural) characteristics (Del Vicario, Bessi & Zollo et al., 2016). These two communication channels are also described as two effects which influence the diffusion process in a social system: namely, the *innovation effect* and the *imitation effect*. This conceptualization of diffusion channels has long been widely adopted by those who seek to model the diffusion patterns of various innovations (e.g., Bass, 1969).

The remaining two elements of the diffusion model are time and the social system. Time, as a dimension of innovation-diffusion analysis, allows for a measurement of the

speed by which members of the social system adopt the innovation and reveals differences among those members. The social system itself is the arena in which the diffusion is brought about and reaches the eventual limits of its scope and extent; this system includes such factors as social structure, social norms, and influential individuals. The diffusion-of-innovations theory provides a systematic framework for analyzing and understanding diffusion phenomena in human behavior. As such, it has given rise to an extensive body of research literature.

The Bass diffusion model

The Bass diffusion model (Bass, 1969) is one of the competing paradigms put forth to model the diffusion of new products and technologies. The model's impact is attested by the fact that the original "Bass model" paper has been voted one of the Top 10 Most Influential Papers published in the 50-year history of *Management Science* (Bass, 2004).

According to Rogers' theory (as briefly described above), individuals in a social system can be divided into two groups by the timing of the adoption of an innovation: innovators (defined as the first 2.5 percent of adopters) and imitators (the remaining adopters). Innovators make adoption decisions regardless of other adopters' influence, while imitators are influenced in the timing of adoption by the pressures of the social system. Further formalizing this distinction, the Bass model posits that diffusion patterns can be modeled through two mechanisms: innovation and imitation. It assumes the probability that an individual who has not yet purchased the new product at time *t* will do so in the next small time increment is a linear function of the proportion of the total number of individuals having already purchased:

$\mathrm{h(t)} = p + qF(t)$, where (1)

$h(t)$ represents the *hazard rate of adoption* at time t,

$F(t)$ is a cumulative distribution function of adoptions at time t,

p is the *coefficient of innovation*, corresponding to external influences such as mass media,

q is the *coefficient of imitation*, corresponding to internal influences such as interpersonal communication.

Using the definitions of the density function, $\mathrm{f(t)} = dF(t)/dt$, and of the hazard rate, $\mathrm{h(t)} = f(t)/[1 - F(t)]$, Equation (1) can be reexpressed as:

$$\frac{\mathrm{dF(t)}}{\mathrm{dt}} = [p + qF(t)][1 - F(t)]. \quad (2)$$

Bass then multiplies both sides of Equation (2) by the third parameter m, which is the market potential (final number of adoptions) of the product, and obtains the basic form of the Bass model:

$$n(t) = \frac{dN(t)}{dt} = \left[p + q\frac{N(t)}{m}\right][m - N(t)] = p[m - N(t)] + q\frac{N(t)}{m}[m - N(t)], \quad (3)$$

where $N(t)$ is the number of cumulative adoptions by time t, and $n(t)$ is the number of new adoptions at time t. Solving the differential equation above, two forms of curve equations can be obtained, namely, the S-shaped cumulative curve in Equation (4), and bell-shaped growth curve in Equation (5):

$$N(t) = m[\frac{1-e^{-(p+q)t}}{1+\frac{q}{p}e^{-(p+q)t}}], \quad (4)$$

$$n(t) = m[\frac{p(p+q)^2 e^{-(p+q)t}}{[p+qe^{-(p+q)t}]^2}]. \quad (5)$$

Using the diffusion data of a new product in the initial period, we can estimate the model parameters, and then predict future purchases of the product; for products that lack sales data, the Bass model can still offer a prediction based on the sales histories of similar products. Rogers (1995) provides a summary of attributes influencing an innovation's rate of adoption; these include relative advantage, compatibility, complexity, trialability and observability. In this perspective, diffusion patterns characterized by similar model parameters are also possibly similar in terms of underlying innovation characteristics. In one meta-analysis (Sultan, Farley & Lehmann, 1990), for example, the parameter *p* was found to be generally higher in Europe than in the U.S.; most of the new products were introduced first in the U.S., perhaps making the innovation less risky and leading to faster adoption in Europe. It was also found that industrial/medical innovations have higher q than other innovations, because the adopting units may be under high pressures to adopt quickly.

The Bass model has been successfully applied to estimate the diffusion data of many innovative products, covering a wide range of areas (Lilien, Rangaswamy & Van den Bulte, 2000) such as durables, retail services, industrial technologies, agriculture, education, and pharmaceuticals. However, despite the model's simplicity, some basic assumptions of the Bass model are not consistent with observed reality: in a real market, the market potential is in dynamic variation, marketing strategies exist, new products are upgraded, and different products can be influenced by one another's sales performance. Therefore, some efforts have been made to modify the Bass model (Centrone, Goia & Salinelli, 2007; Norton & Bass, 1987; Bagchi, Kirs & López, 2008; Islam, Fiebig & Meade, 2002; Roberts, Nelson & Morrison, 2005). Nonetheless, research (Chandrasekaran & Tellis, 2007) shows that the Bass model fits actual data almost as well as much more complex models which seek to correct its limitations (Bass, Krishnan, and Jain, 1994). Consistent with the results of decades of subsequent research, the simple Bass model is still preferred to other models in many areas of application.

Modeling citation trajectories

The citation dynamics of a scientific publication are somewhat like the sales data of a new product. By analogy, we can refer to the temporal dynamics of such citations as the *diffusion* of citations. Many efforts (Mingers, 2008; Mingers & Burrell, 2006; Pilkington, 2013; Bouabid, 2011; Nadarajah & Kotz, 2007) have been devoted to fitting various distribution models to these citation data; the models evaluated include exponential, logistic, Gaussian, Gompertz, Weibull, gamma, beta, Pearl logistic, and inverted Gaussian distributions, among others. However, the application of this class of methods has been subject to at least two major limitations. First, the distribution models are usually conducted on citation data in aggregate levels and within a relatively short time span—for example, on the journal level within 10 years (Pilkington, 2013). For citations to individual papers with much longer time periods (e.g., decades), citation patterns would vary so dramatically that it becomes difficult for any single distribution

to fit. Second, the parameters in these mathematical models rarely have corresponding meaning in a citation context; thus, they provide limited help in understanding the citation process. Wang, Song & Barabási (2013) consider individual papers' preferential attachment, aging, and fitness, and derive a mechanistic model for the citation trajectories of those papers, finding that predictable patterns exist in the long-term citation data. One limitation of the model, however, is its inability to account for the citation bump observed in the case of "delayed" papers (Ke, Ferrara & Radicchi et al., 2015). Furthermore, Wang, Mei & Hicks (2014)'s experiment also shows the difficulty in predicting future citations for individual papers.

The Bass model, meanwhile, has been used in several studies of citation dynamics. Franses (2003) was the first to apply the Bass model in citation analysis research, where he fit the model to citations to papers in the journal Econometrica 1987 and found that, on average, the impact of these articles lasted for about 15 years. In a later study (Fok & Franses, 2007) using the diffusion-of-innovations theory, Franses and his colleague tried to explain the process of citation accumulation and the relation between key characteristics of the diffusion process and features of the articles. On the author level, Bjork, Offer & Söderberg (2014) found that the Bass model fits well with the citation trajectories of some Nobel Prize winners in Economics, and that economic knowledge follows the typical innovation cycle. In these studies, the Bass model was viewed as a useful heuristic for understanding the spread of scientific ideas. The nuances of the model's parameters, however, have remained largely unexplored, as have the model's practical implications.

Data and method

Following the practice of previous research (Li et al., 2014; Sun, Min & Li, 2016; Min, Sun & Pei et al., 2016), the empirical analysis in this paper is based on a dataset of essays by 629 Nobel Prize winners in four disciplines: Chemistry, Physics, Physiology or Medicine, and Economic Sciences. Their publications from 1900 to 2000 and citation data until 2011 were collected from the Web of Science database. In total, 58,963 papers and their citation data were obtained with a citation window of at least 11 years.

A scientific idea is a kind of innovation. Thus, if the spread of a fresh scientific idea follows the Bass diffusion paradigm, we should be able to obtain specific values for the parameters p, q, and m when applying the model to publication data. The value of m can be regarded, straightforwardly enough, as the ultimate number of citation counts a paper can get. Interpreting the values of p and q directly as *innovation effect* and *imitation effect*, however, might cause confusion. Therefore, we refrain from imposing an interpretation on p and q here and instead seek the answer from experimental results. Based on the derivation and solution of the Bass equation in the previous section, we applied the model to individual papers to estimate the parameter values for each paper. To guarantee the accuracy and reliability of the empirical analysis, we carefully processed the data in the following manner.

(1) Papers with no more than 19 citations[2] were excluded. This left 28,769 papers in the dataset.

(2) In terms of estimation methods, research has reached a clear consensus that it is non-optimal to use Ordinary Least Squares to estimate the Bass model, but the choice between Non-linear Least (NLS) Squares and Maximum Likelihood Estimation is still not clear (Meade & Islam, 2006)[3]. We opted to use NLS in this study, since the method has gained widespread acceptance in recent works (Van den Bulte & Lilien, 1997).

(3) As initial parameter values were required, we ran the model in loops with different ranges of parameter values until no new results appeared[4]. The results converged after the second loop, but we ran the third and fourth loops and found the same results. A total of 23,399 papers fit the model, among which 22,028 papers have non-negative parameters values and R^2 values. A summary of the 22,028 papers is listed in Table 1.

(4) To get reliable results, we filtered the papers by R^2 values. Empirical analysis in the literature (Bass, 1969) shows that even in instances of low R^2 values (R^2=0.077), the Bass model provides a good description of the general trend of historical data. Therefore, without loss of generality and rigor, we retained papers with $R^2 \geq 0.5$. This condition led to a final dataset of 11,037 papers.

Results

The diffusion of citations

In terms of coefficients *p* and *q*, we find that citation diffusion is roughly similar to the diffusion of new products and technologies. Since 23,399 out of 28,769 papers successfully fit the model, most papers in the dataset follow the Bass diffusion mechanism. Table 1 shows that the median value and the mean value for parameter *p* are 0.032 and 0.035, respectively; corresponding values for parameter *q* are 0.216 and 0.336, respectively. This closely matches the values reported in the marketing diffusion literature, where *p* values are usually between 0.00007 and 0.03 and *q* values tend to lie in the interval from 0.38 to 0.53 (Chandrasekaran & Tellis, 2007). The consistency in model parameters implies that the Bass diffusion mechanism applies to the citation diffusion process for most of the papers, even though it was originally proposed for product diffusion. Meanwhile, it is observed that, generally, the parameter *p* is slightly larger for papers than for products, and the parameter *q* is slightly smaller. However, it remains unclear how *p* and *q* should be interpreted as elements of the citation process.

Table 1 Summary of the estimation results (N=22,028)

	Min	**1st Qu**	**Median**	**Mean**	**3rd Qu**	**Max**
Innovation coefficient *p*	0.000	0.016	0.032	0.035	0.049	0.527
Imitation coefficient *q*	0.000	0.106	0.216	0.336	0.415	4.590
Market potential *m*	5.16	35.08	68.31	225.25	157.21	230283.07
R^2	0.000	0.275	0.501	0.487	0.701	0.991

Interpretation of the parameters *p* and *q*

Table 1 indicates that *q* values are generally much larger than *p* values for papers. The result is not surprising, as this pattern was also observed in many studies on diffusion of innovations (Bass, 1969; Sultan, Farley & Lehmann, 1990; Loh & Venkatraman, 1992; Park, Kim & Lee, 2011). Our concern is how to appropriately interpret these parameter values in this novel context and how to understand their influence on the citation process.

Investigation of the papers' citation curves reveals four major findings:

(1) Small *p* usually indicates a small proportion of citation counts to a paper in the first few years after publication.

(2) Large *p* usually indicates a large proportion of citation counts to a paper in the first few years after publication, often followed by a decreasing trend in citations over time.

(3) A sharp citation peak often appears together with a large *q* value, with yearly citations drastically increasing before the peak and drastically decreasing after the peak.

(4) Papers with simultaneously small *p* and small *q* have many more citations than those with disparate *p* and *q* values. Figure 3 gives a clear illustration of this observation.

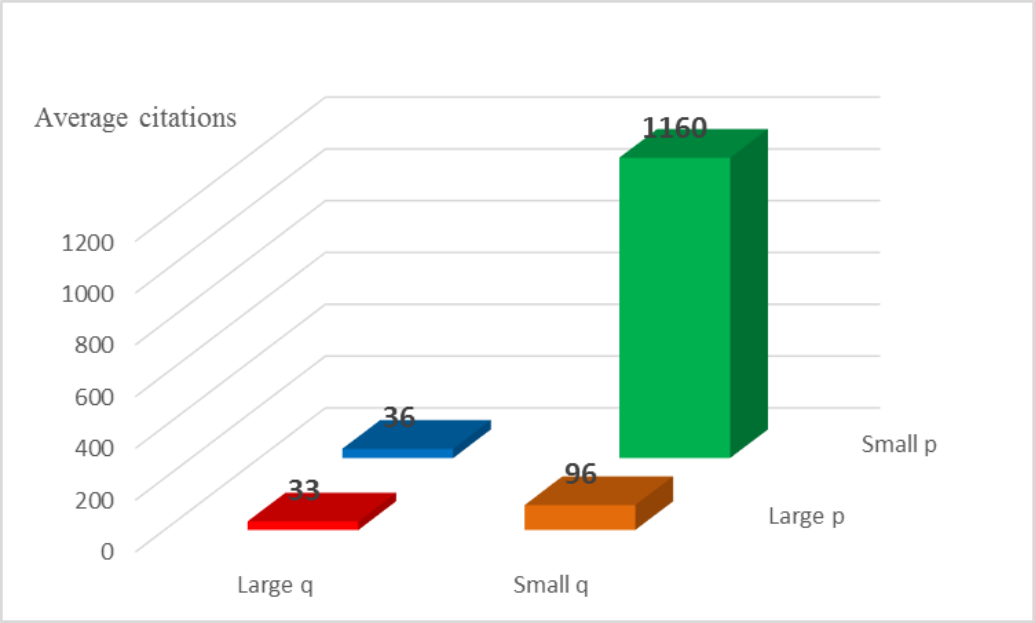

Figure 3: Papers with simultaneously small *p* and small *q* have many more citations

To see the characteristics of the parameters *p* and *q*, we select from the dataset representative papers with very small and very large *p*, *q* values[5], and graph the citation curves of these papers in Figure 4.

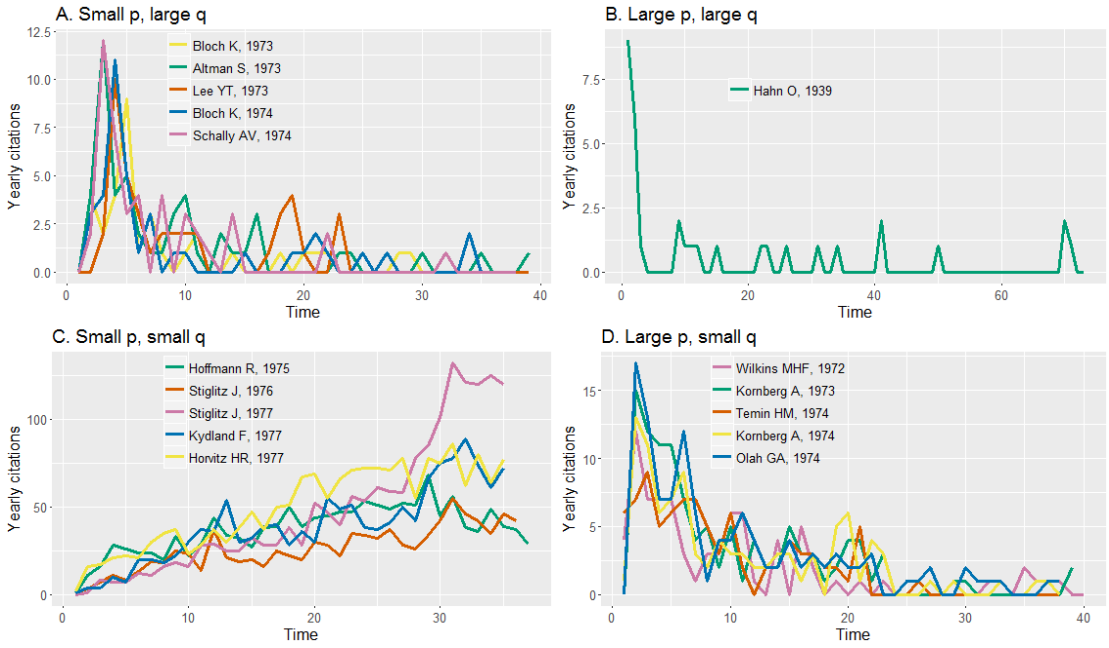

Figure 4: Temporal citation curves of papers with very small and very large *p*, *q* values
Papers in Panel 4C have both small *p* value and small *q* value; those in 4D have large *p* value but small *q* value; in 4A, small *p* value but large *q* value, and in 4B, large *p* value and large *q* value. In other words, across the four panels of Figure 4, *p* increases along the horizontal axis and *q* increases along the vertical axis. In all but Panel 4B, five instances that were published in approximately the same year in the 1970s are selected for illustration.

150 papers have *p* value and *q* value both smaller than the corresponding 10th percentile values. Interestingly, the citation curve of this kind of paper generally follows a prominent trend: overall, yearly citations maintain a persistent increase after publication. Figure 4C shows five typical citation curves of these papers. What's more, these papers generally have a very large number of total citations compared to other papers in the dataset. For these papers, the median of total citations is 647 and the mean is 1,160, while the 90th percentile of total citations for all papers in the dataset is only 429.

Papers with large *p* value but small *q* value (Figure 4D) exhibit almost the opposite trend of papers in Figure 4C. These papers receive the highest yearly citations in the early stages but display a sustained downward trend in subsequent years. A total of 260 papers fall into this category; their impact is mediocre, with an average total citations of 96.

Papers with small *p* value but large *q* value seem to have a citation pattern closer to that of a typical paper, whose citations begin to decrease several years after publication (Moed, Burger, & Frankfort et al., 1985; Amin & Mabe, 2003). Figure 4A depicts five instances of 334 papers in this category. Their citations grow from a relatively low level to the highest level and then go through a decline similar to that found in in Figure 4D. However, the two categories of papers are different in two ways. First, there exists an obvious rise in yearly citations prior to the citation peak for papers

with small *p* but large *q* (Figure 4A). Second, this type of paper often exhibits an unusually steep citation peak, much higher than other fluctuations in the same citation curve. With an average of 36 total citations, however, these papers have even less of an impact than those with large *p* and small *q*.

Papers falling into the last category (Figure 4B), with both large *p* and large *q* values, are relatively infrequent. There is only one instance in the dataset whose *p* value is 0.119171 and *q* value is 2.054208. The paper got some attention shortly after publication and reached citation peak in the year of publication after which its yearly citations rapidly declined and remained at a low level. Unlike other papers that lose citations quickly and then stay uncited, this paper still maintained a small amount of citations even decades after publication.

Given the observations above, we interpret the parameters *p* and *q* in the citation context as follows.

Parameter *p* reflects the ***proportion*** of citations a paper receives in the ***early stage*** after its publication, relative to its entire lifecycle. A large *p* value indicates a paper gets a large proportion of its citations shortly after publication, which also means it has limited potential to obtain more citations in the remaining portion of the lifecycle (Figure 4B, D). A small *p* value signifies a small proportion of early citations compared with the eventual total; on the other hand, it indicates the paper has a high potential to obtain more citations as time goes on (Figure 4A, C). Therefore, parameter *p* to some extent represents a paper's potential to achieve future citations: the smaller a paper's parameter *p*, the more potential it has to achieve future citations; the larger the *p* value, the more citations the paper has exhausted in the early stage.

In addition to potential, however, a paper also needs persistence to achieve great impact. (The examples in Fig 4A show the fate of papers which have the former trait, but not the latter.) The quality of persistence is, to an extent, captured by parameter *q*. In Eq. 1, *q* is the slope of the math formula of the probability that a researcher who has not yet added a paper in her reference list will do so in the next small time increment. A large *q* value can certainly increase the probability that a paper will receive a new citation, but this also speeds up the paper's obsolescence and death, since the probability can't exceed 1. As more and more researchers (F(t) in Eq. 1) cite the paper, the probability that a new researcher will do so (h(t) in Eq. 1) increases, but this condition ends when h(t) reaches 1 and the ultimate citation potential *m* (the third parameter in the Bass model) is achieved. A large *q* value thus indicates quick death, or poor persistence (Figure 4A, B). A small *q* value, however, provides more time for a paper to persistently accumulate citations (Figure 4C, D).

Papers in Figure 4C illustrate how the Bass model parameters influence the citation dynamics. They also provide a clue to the underlying mechanism: high-impact papers often have both good potential and good persistence.

Diffusion patterns and scientific impact

In the Bass model, *m* stands for the ultimate market potential a new product can achieve. By analogy, in the context of citation diffusion, *m* can be regarded as the total number of citations a paper can obtain through its lifetime. The strong linear trend in

Figure 5A suggests a significant positive relation (correlation coefficient = 0.350, $p < 0.001$) between total citations and *m*. Although the total citations of most papers are located close to *m*, some are below *m*, which suggests that they still have great potential for achieving future impact. It may be that those papers have not yet reached the peak of their citation curves; some papers published late need more time to reach their peaks. To keep our results grounded in the papers' observed impact, we use total citations instead of *m* to analyze the relation between scientific impact and diffusion patterns.

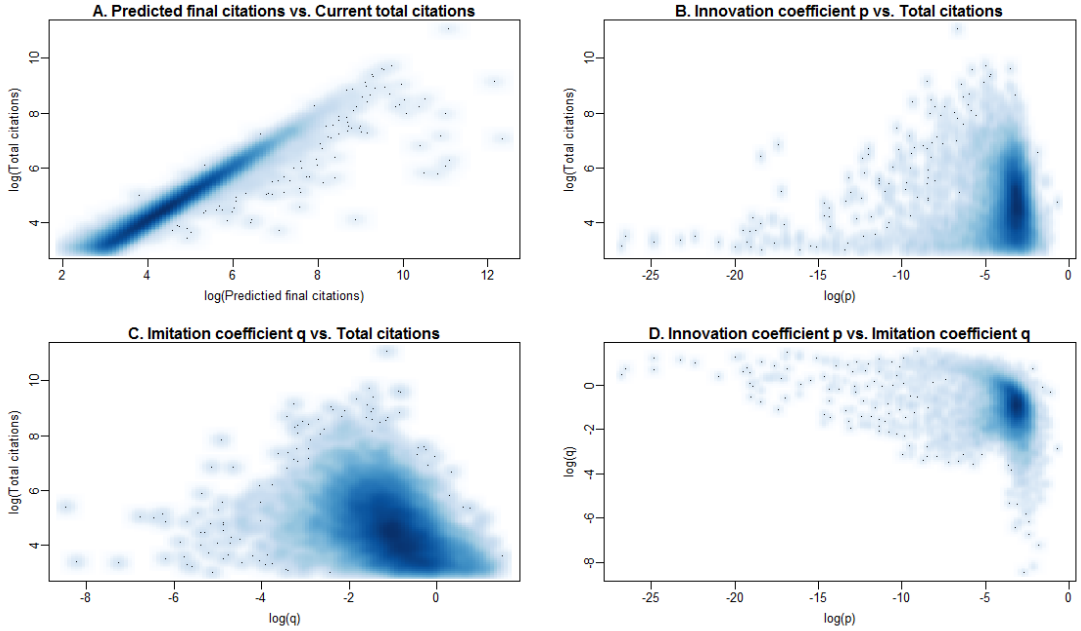

Figure 5: Scatter plots for predicted final citations (*m*), current citations, parameter *p*, and parameter *q*

Figure 5 also shows the relations among total citation count, parameter *p*, and parameter *q*. Parameter *p* has no significant relationship to total citation count (Figure 5B): papers with a given *p* value may have various total citation counts, and large citation counts can appear with a wide range of *p* values. The Pearson correlation test validates this observation with a nearly negligible correlation coefficient (0.025, $p<0.01$). This indicates that early citation proportion (*p*) is, by itself, a poor predictor of a paper's long-term scientific impact.

Parameter *q* has a somewhat clearer relationship to citation count than does *p*. Figure 5C indicates that total citation count generally decreases as the value of *q* increases. Again, this observation is borne out by the Pearson correlation test, which yields a correlation coefficient of -0.395 ($p<0.001$). This lends weight to our observation that rapid citation accrual may hasten the death of a paper, thus leading to lower total citation counts. Furthermore, Figures 5B & C indicate that *q* has a more direct and significant effect than *p* when their joint influence on total citations is considered.

Figure 5D shows that the majority of articles have high *p* and high *q*. Connecting these parameters with total citations (see Figure 6), it is clear that papers with high total citations (light-colored areas) are mainly distributed where both *p* and *q* are relatively

small. To further test this pattern, we select two groups of papers from the dataset: a high-impact group of papers with no fewer than 664 total citations each, and a mediocre-impact group of papers with no more than 24 total citations[6]. Figure 7 shows a significant difference between the two groups: high-impact papers tend to have both smaller *p* and smaller *q* than mediocre-impact papers. The gap in *q* values is even larger than the gap in *p* values.

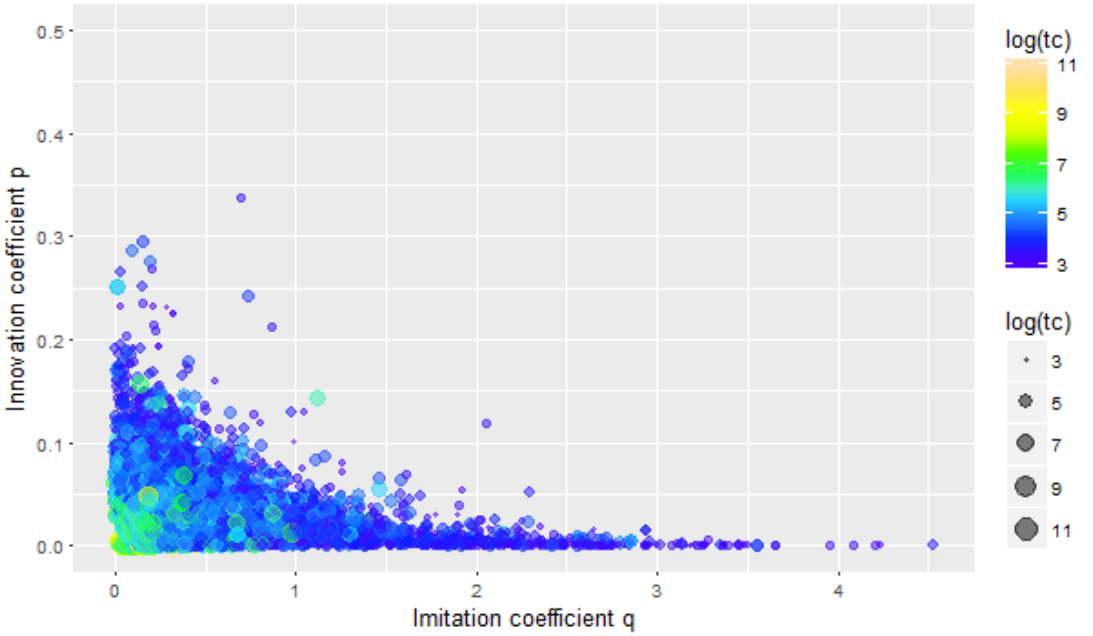

Figure 6: Innovation coefficient *p*, imitation coefficient *q*, and total citations

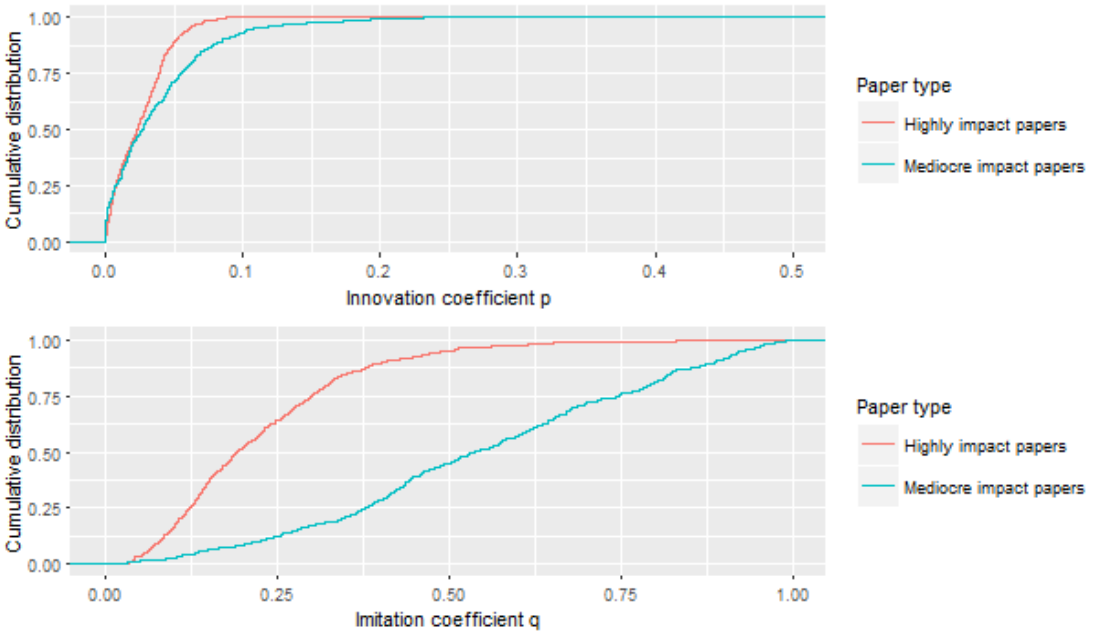

Figure 7: Cumulative distributions of parameters *p* and *q* for two groups of papers

Potential of scientific impact

If *m*, the third parameter in the Bass model, is used to estimate the market potential of a new product, can it also estimate future total citations? We choose papers with the largest *m* values and plot their actual citation curves (Figure 8), which exhibit diverse citation trends. Yearly citations in Figures 8B and 8C show continuously increasing trends in spite of some slight perturbations. The paper in Figure 8F goes through a citation surge and then keeps increasing. The paper in Figure 8A, however, goes downhill after climbing to a local peak, but then comes back up for a second surge. In Figures 8D and 8E, however, the papers seem to have completed the citation cycle. In general, we observe a huge difference between *m* and current total citations for papers with increasing citation trends (Figures 8A, B, C & F); for papers with a roughly complete citation cycles (Figures 8D & E), *m* is much closer to the current number of total citations. To differentiate papers with great future potential from those with great current impact, we propose the *saturation level*:

$$\text{saturation level} = \text{current total citations} / \text{m} \tag{7}$$

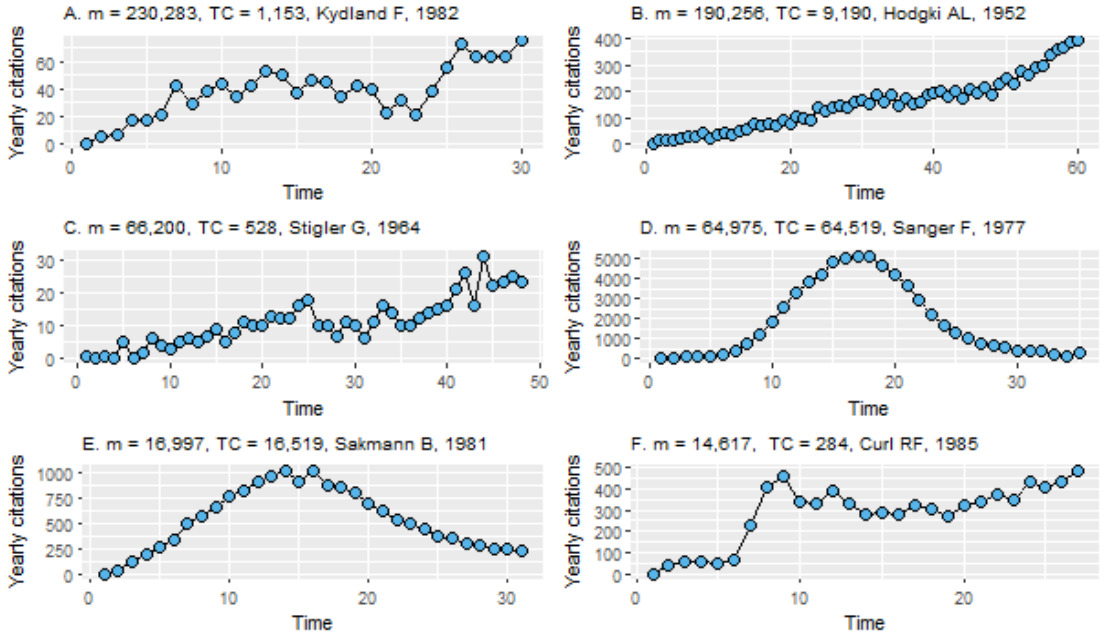

Figure 8: Actual citation curves for papers with top *m* values. (*m* stands for predicted final citations; TC stands for current total citations)

With the saturation level, we can roughly estimate the potential scientific impact for a given article. Papers with low saturation levels have great potential to achieve more citations in the future, e.g. papers in Figures 9A and 9B, with saturation levels of 0.50% and 0.65%, respectively, exhibit a high capacity for future citations. Papers with saturation level approaching 100%, in contrast, have almost exhausted their potential, as is shown in Figures 9C and 9D. Interestingly, papers can reach saturation levels even higher than 100%. Two instances (327.42% and 205.57%) are shown in Figures 9E and 9F respectively. Both papers undergo a sharp citation peak after which they attain relatively few new citations. Their citation curves will likely converge to 0 soon, but in the meantime, the papers continue to receive citations at a low rate. These papers show

potential exceeding that predicted by the Bass model.

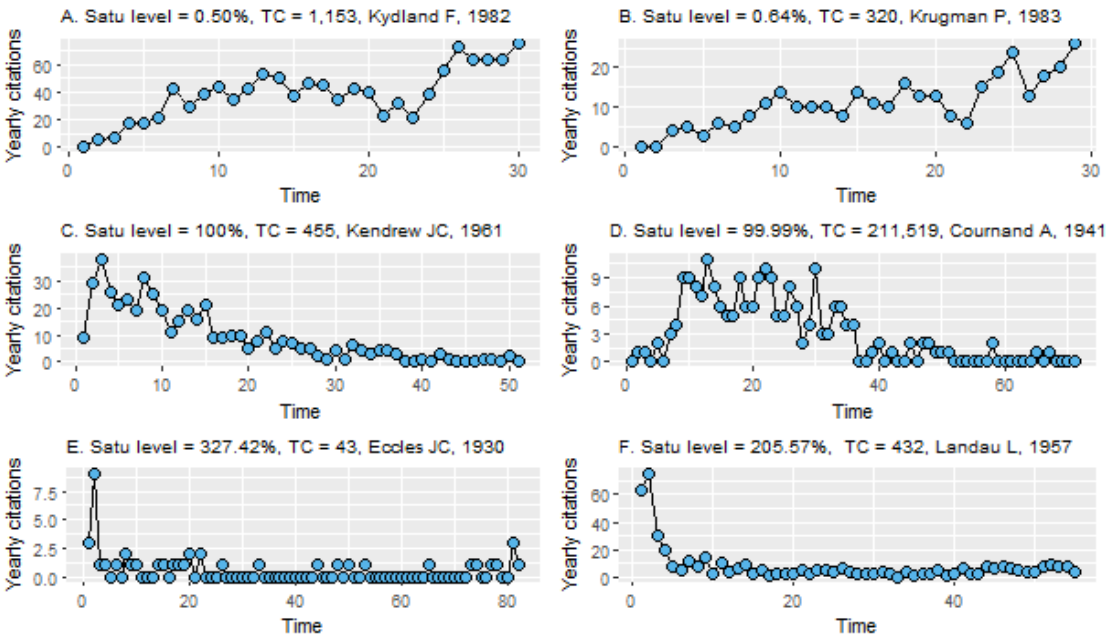

Figure 9: Papers with different saturation levels. (Satu level stands for saturation level; TC stands for current total citations)

Model reevaluation

We tested the Bass model on another dataset that was extracted from the American Physical Society journal series (hereafter APS dataset). We randomly selected 50,000 APS papers and then applied Bass model on this dataset. Citation patterns similar to those in Figures 3 & 4 were found, suggesting the main findings of this study can be validated by different datasets.

In addition, further investigation of the cumulative distribution confirms that the three parameters are field-specific. In terms of the parameter *p*, papers in the natural sciences (e.g., Medicine, Physics, and Chemistry) are similar, while Economic Sciences papers have even smaller values. The parameter *q*, in contrast, separates the four disciplines clearly: Economic Sciences < Chemistry < Medicine < Physics. In terms of the parameter *m*, Chemistry < Physics < Medicine < Economic Sciences. Since the APS papers from Physics provide similar experimental observations, and the field-related differences don't hamper our main conclusions, we have not repeated the experiments separately for every discipline. Moreover, the age of the papers shows limited impact on the three parameters, since further experiments show that the values of *p*, *q*, and *m* are almost randomly distributed throughout different ages.

The Bass model, we find, is a better descriptive model than it is a predictive model. Therefore, we should lower the expectations of the model's long-term predictability. Just as Bass (1969) stated in the original paper, the model is intended for new product growth and thus should be regarded as a short-term model. Wang et al. (2013) have already shown that their model offers better performance than the Bass model in terms of long-term prediction. We also generated a prediction of citations 20 years later based

on a 10-year training period. Figure 10 shows results similar to those of Wang et al. (2013), suggesting that the Bass model tends to underestimate long-term citations. Therefore, we suggest that the Bass model should be used to describe certain characteristics of citation process if so desired, instead of making accurate predictions.

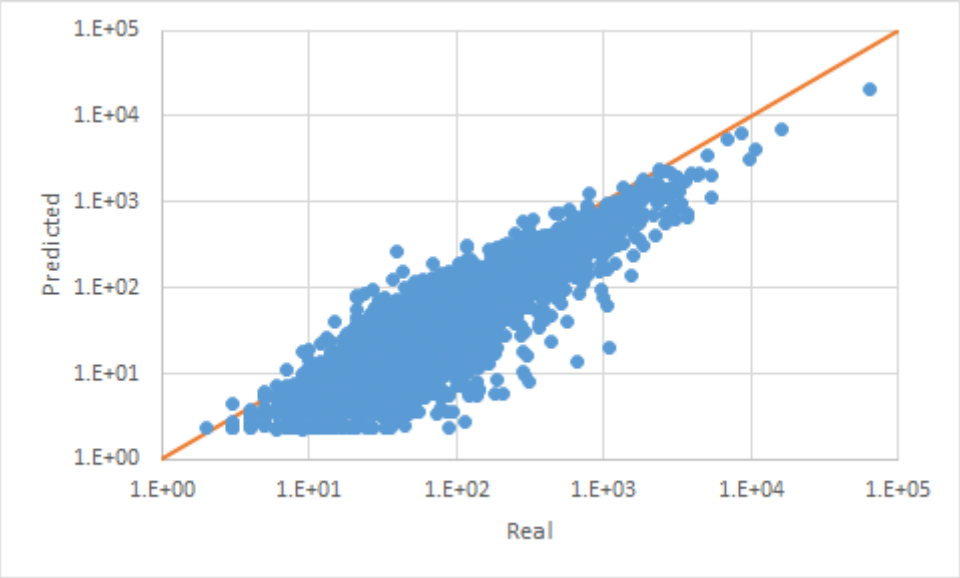

Figure 10: Predicting citations 20 years after publication based on a 10-year training period

Discussion and Conclusion

This paper studies the temporal diffusion of citations to scientific papers in the dynamic process of scientific communication. Scientific communication is complicated, multifaceted, and driven by patterns of discovery and adoption—traits which suggest a comparison to the already well-studied evolution and diffusion processes of innovative ideas, technologies, and products (Rogers, 1995). Although diffusion research in marketing concerns how to speed up the rate of adoptions and increase the number of final sales, we aim to find diffusion mechanisms of scientific ideas and new ways to evaluate scientific outputs. The theory of diffusion of innovations is applied to the dynamic process of citations to better understand the mechanism of scientific communication as well as the course of scientific advancement. The classic Bass model (Bass, 1969) is used to estimate the citation diffusion data of tens of thousands of papers by Nobel Prize laureates. Notably, this study quantifies and presents two mechanisms in the diffusion process of scientific ideas, corresponding to the *innovation effect* and *imitation effect* in marketing diffusion literature. By investigating two diffusion parameters p and q, and the diffusion mechanism of different types of papers, we show that in the citation realm p and q respectively associate, to some extent, with a paper's potential and persistence in achieving future citations. For the diffusion of breakthrough innovations, both early take-off and late growth are slower but more persistent than the diffusion of mediocre innovations. Moreover, our proposed metric of *saturation level* shows a good ability to distinguish papers with great impact potential from those which have almost exhausted their scientific impact.

This study introduces diffusion-of-innovations theory to the analysis of citation

dynamics. Although there has been a vast range of diffusion literature in such scientific fields as marketing, public policy, and sociology, the diffusion of scientific citations remains relatively less explored (Zhai, Ding & Wang, forthcoming).

To conclude, we discuss the dynamic process of citation from a diffusion perspective. Beyond citation count, which measures the overall impact of scientific outcome, we can gain more insights from the citation diffusion process and understand the nuances of scientific impact. The Bass model, which has achieved great success in marketing research, is applied to citation dynamics in this study. Results show that low citation counts of a paper usually associates with large early citation proportion (high value of p) or quick citation obsolescence (high value of q), which might indicate compatibility with traditional knowledge and thus a lack of novelty. In contrast, low values of p and q indicate high impact and more citations. The third parameter m, interpreted as the final number of citations a paper will achieve, is reasonable, since a paper will eventually die (or infinitely approach death) as time goes on. Quantitative measures, such as saturation level, could be considered to roughly estimate the potential impact of scientific output. Although the Bass model is not as good as WSB (Wang, Song & Barabási, 2013) model at predicting long-term impact (actually it was originally proposed as a short-term model for new products), what we emphasize here is its descriptive power and simplicity: it elegantly integrates the potential and persistency of a paper in terms of attracting citations into a single formula and successfully differentiates papers with distinct citation patterns (e.g., Figure 1 & Figure 4).

There are important limitations to this study. First, although we draw an analogy between citations and innovations, the two are different in certain essential respects, such the characteristics of their respective communication channels, suppliers, and adopter populations. These differences would be of great interest in future research. The second limitation is a manifestation of the first one: despite the parsimony of the Bass model, not all papers in the dataset follow the Bass diffusion mechanism. We ran the model in various ranges of parameter spaces, but 18.67% of the papers tested still didn't fit the model. The reason why the Bass model failed to fit these papers requires further investigation. A third concern is the limited long-term predictive power of the Bass model. Therefore, we suggest the model be used as a descriptive one, that is, one that quantifies and describes certain characteristics of scientific outputs. In future studies, we would also intend to inspect the citation process using other tools of diffusion research, such as adopting behavior and diffusion networks.

Acknowledgement

This research was supported by the National Natural Science Foundation of China (NSFC No 71273125), the Major Bidding Program of National Social Science Foundation of China (No 16ZDA224) and the China Scholarship Council. The authors are grateful to Alessandro Flammini, Filippo Radicchi, and Santo Fortunato, Xiaoran Yan, and Yong-Yeol Ahn for helpful discussions, and Tianjiu Yin for help process the data. The authors would like to thank the anonymous reviewers for their careful reading of the manuscript and their many insightful comments and suggestions, which have greatly improved the quality of the paper.

References

Abramo, G., & D'Angelo, C. A. (2016). A farewell to the MNCS and like size-independent indicators. *Journal of Informetrics, 10*(2), 646-651.

Amin, M., & Mabe, M. (2000). Impact factors: Use and abuse. *Perspectives in Publishing, 1*(1), 1-6.

Bartlett, M. S. (1960). *Stochastic population models in ecology and epidemiology*. London: Methuen.

Bass, F. M. (1969). A new product growth for model consumer durables. *Management Science, 15*(5), 215-227.

Bass, F. M. (2004). Comments on "A new product growth for model consumer durables the Bass model". *Management Science, 50*(12), 1833-1840.

Bass, F. M., Krishnan, T. V., & Jain, D. C. (1994). Why the Bass model fits without decision variables. *Marketing Science, 13*(3), 203-223.

Bjork, S., Offer, A., & Söderberg, G. (2014). Time series citation data: The Nobel Prize in economics. *Scientometrics, 98*(1), 185-196.

Bouabid, H. (2011). Revisiting citation aging: A model for citation distribution and life-cycle prediction. *Scientometrics, 88*(1), 199-211.

Burt, R. S. (1987). Social contagion and innovation: Cohesion versus structural equivalence. *American Journal of Sociology, 92*(6), 1287-1335.

Centrone, F., Goia, A., & Salinelli, E. (2007). Demographic processes in a model of innovation diffusion with dynamic market. *Technological Forecasting and Social Change, 74*(3), 247-266.

Chandrasekaran, D., & Tellis, G. J. (2007). A critical review of marketing research on diffusion of new products. *Review of Marketing Research, 3*, 39-80.

Cronin, B. (1984). *The citation process: The role and significance of citations in scientific communication*. London: Taylor Graham.

Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., & Quattrociocchi, W. (2016). The spreading of misinformation online. *Proceedings of the National Academy of Sciences, 113*(3), 554-559.

Fersht, A. (2009). The most influential journals: Impact Factor and Eigenfactor. *Proceedings of the National Academy of Sciences, 106*(17), 6883–6884.

Flick, P. K., & Bloch, K. (1974). In vitro alterations of the product distribution of the fatty acid synthetase from Mycobacterium phlei. *Journal of Biological Chemistry, 249*(4), 1031-1036.

Fok, D., & Franses, P. H. (2007). Modeling the diffusion of scientific publications. *Journal of Econometrics, 139*(2), 376-390.

Franses, P. H. (2003). The diffusion of scientific publications: The case of Econometrica, 1987. *Scientometrics, 56*(1), 29-42.

Garfield, E. (2006). The history and meaning of the journal impact factor. *JAMA, 295*(1), 90-93.

Golosovsky, M., & Solomon, S. (2017). Growing complex network of citations of scientific papers: Modeling and measurements. *Physical Review E, 95*(1), 012324.

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

Islam, T., Fiebig, D. G., & Meade, N. (2002). Modelling multinational telecommunications demand with limited data. *International Journal of Forecasting*, *18*(4), 605-624.

Ke, Q., Ferrara, E., Radicchi, F., & Flammini, A. (2015). Defining and identifying sleeping beauties in science. *Proceedings of the National Academy of Sciences*, *112*(24), 7426-7431.

Kuhn, T. (1962). *The structure of scientific revolutions*. Chicago, IL: University of Chicago Press.

Lekvall, P., & Wahlbin, C. (1973). A study of some assumptions underlying innovation diffusion functions. *The Swedish Journal of Economics*, 362-377.

Li, J., & Shi, D. (2016). Sleeping beauties in genius work: When were they awakened? *Journal of the Association for Information Science and Technology*, *67*(2), 432-440.

Li, J., Shi, D., Zhao, S. X., & Fred, Y. Y. (2014). A study of the "heartbeat spectra" for "sleeping beauties." *Journal of Informetrics*, *8*(3), 493-502.

Lilien, G. L., Rangaswamy, A., & Van den Bulte, C. (2000). Diffusion models: managerial applications and software. In V. Mahajan, E. Muller & Y. Wind (Eds.), *New product diffusion models (pp. 295-311)*. New York: Kluwer Academic.

Loh, L., & Venkatraman, N. (1992). Diffusion of information technology outsourcing: Influence sources and the Kodak effect. *Information Systems Research*, *3*(4), 334-358.

Mahajan, V., Muller, E., & Wind, Y. (Eds.). (2000). *New-product diffusion models* (Vol. 11). New York: Springer Science & Business Media.

Massiani, J., & Gohs, A. (2015). The choice of Bass model coefficients to forecast diffusion for innovative products: An empirical investigation for new automotive technologies. *Research in Transportation Economics*, *50*, 17-28.

Meade, N., & Islam, T. (2006). Modelling and forecasting the diffusion of innovation–A 25-year review. *International Journal of Forecasting*, *22*(3), 519-545.

Min, C., Sun, J., Pei, L., & Ding, Y. (2016). Measuring delayed recognition for papers: Uneven weighted summation and total citations. *Journal of Informetrics*, *10*(4), 1153-1165.

Mingers, J. (2008). Exploring the dynamics of journal citations: modelling with S-curves. *Journal of the Operational Research Society*, *59*(8), 1013-1025.

Mingers, J., & Burrell, Q. L. (2006). Modeling citation behavior in management science journals. *Information Processing & Management*, *42*(6), 1451-1464.

Moed, H. F., Burger, W. J. M., Frankfort, J. G., & van Raan, A. F. (1985). The use of bibliometric data for the measurement of university research performance. *Research Policy*, *14*(3), 131-149.

Moed, H. F. (2006). Citation analysis in research evaluation (Vol. 9). New York: Springer Science & Business Media.

Nadarajah, S., & Kotz, S. (2007). Models for citation behavior. *Scientometrics*, *72*(2), 291-305.

Norton, J. A., & Bass, F. M. (1987). A diffusion theory model of adoption and substitution for successive generations of high-technology products. *Management Science*, *33*(9), 1069-1086.

Park, S. Y., Kim, J. W., & Lee, D. H. (2011). Development of a market penetration forecasting model for Hydrogen Fuel Cell Vehicles considering infrastructure and cost reduction effects. *Energy Policy*, *39*(6), 3307-3315.

Petersen, A. M., Fortunato, S., Pan, R. K., Kaski, K., Penner, O., Rungi, A., & Pammolli, F. (2014). Reputation and impact in academic careers. *Proceedings of the National Academy of Sciences*, *111*(43), 15316-15321.

Pilkington, A. (2013). Modeling citation diffusion: Innovation management literature. *International Journal of Innovation and Technology Management*, *10*(01), 1350004.

Roberts, J. H., Nelson, C. J., & Morrison, P. D. (2005). A prelaunch diffusion model for evaluating market defense strategies. *Marketing Science*, *24*(1), 150-164.

Rogers, E. M. (1995). *Diffusion of innovations*. New York: Simon and Schuster.

Simmons, B. A., & Elkins, Z. (2004). The globalization of liberalization: Policy diffusion in the international political economy. *American Political Science Review*, *98*(1), 171-189.

Sinatra, R., Deville, P., Szell, M., Wang, D., & Barabási, A. L. (2015). A century of physics. *Nature Physics*, *11*(10), 791-796.

Sultan, F., Farley, J. U., & Lehmann, D. R. (1990). A meta-analysis of applications of diffusion models. *Journal of Marketing Research*, *27*(1), 70-77.

Sun, J., Min, C., & Li, J. (2016). A vector for measuring obsolescence of scientific articles. *Scientometrics*, *107*(2), 745-757.

Tornatzky, L. G., & Klein, K. J. (1982). Innovation characteristics and innovation adoption-implementation: A meta-analysis of findings. *IEEE Transactions on Engineering Management*, *EM-29*(1), 28-45.

Van den Bulte, C. (2002). Want to know how diffusion speed varies across countries and products? Try using a Bass model. *PDMA Visions*, *26*(4), 12-15.

Van den Bulte, C., & Lilien, G. L. (1997). Bias and systematic change in the parameter estimates of macro-level diffusion models. *Marketing Science*, *16*(4), 338-353.

van Raan, A. F. (2004). Sleeping beauties in science. *Scientometrics*, *59*(3), 467-472.

Wang, D., Song, C., & Barabási, A. L. (2013). Quantifying long-term scientific impact. *Science*, *342*(6154), 127-132.

Wang, J., Mei, Y., & Hicks, D. (2014). Comment on "Quantifying long-term scientific impact". Science, 345(6193), 149-149.

Wejnert, B. (2002). Integrating models of diffusion of innovations: A conceptual framework. *Annual Review of Sociology*, *28*(1), 297-326.

Yan, E., Ding, Y., Cronin, B., & Leydesdorff, L. (2013). A bird's-eye view of scientific trading: Dependency relations among fields of science. *Journal of Informetrics*, *7*(2), 249-264.

Young, H. P. (2009). Innovation diffusion in heterogeneous populations: Contagion, social influence, and social learning. *American Economic Review*, *99*(5), 1899-1924.

Zhai, Y., Ding, Y., Wang, F. (in press). Measuring the diffusion of an innovation: A citation analysis. *Journal of the Association for Information Science and Technology*.

---

[1] source: Bass's Basement Research Institute, http://bassbasement.org/BassModel/Default.aspx

[2] 19 is the median value for papers in Physics; median values for Chemistry, Physiology or Medicine, and Economic Sciences are 21, 40, and 20, respectively.

[3] Please refer to Meade & Islam (2006) for more details about estimation methods.

[4] In the first loop, p = [0.0004, 0.004], increment = 0.0004, q = [0.3, 1], increment = 0.1; in the second loop, p = [0.00009, 0.0009], increment = 0.00009, q = [0.1, 0.3], increment = 0.02; in the third loop, p = [0.01, 0.06], increment = 0.005, q = [0.3, 0.5], increment = 0.02; in the fourth loop, p = [0.01, 0.06], increment = 0.005, q = [0.7, 1.2], increment = 0.05.

[5] 10^{th} and 90^{th} percentile values are used to represent "very small" and "very large" values. In the dataset, the 10^{th} and 90^{th} percentile values of innovation coefficient p are 0.007734 and 0. 073187 respectively, and the 10^{th} and 90^{th} percentile values of imitation coefficient q are 0. 118974 and 0.929482 respectively.

[6] 664 is the 95^{th} percentile value, and 24 is the 5^{th} percentile value, of total citations for all papers.