A New Approach for Main Path Analysis: Decay in Knowledge Diffusion

John S. Liu
Graduate Institute of Technology Management, National Taiwan University of Science and Technology, 43, Section 4, Keelung Road, Taipei 10607, Taiwan. E-mail: johnliu@mail.ntust.edu.tw

Chung-Huei Kuan
Graduate Institute of Patent, National Taiwan University of Science and Technology, 43, Section 4, Keelung Road, Taipei 10607, Taiwan. E-mail: maxkuan@mail.ntust.edu.tw

Main path analysis is a powerful tool for extracting the backbones of a directed network and has been applied widely in bibliometric studies. In contrast to the no-decay assumption in the traditional approach, this study proposes a novel technique by assuming that the strength of knowledge decays when knowledge contained in one document is passed on to another document down the citation chain. We propose three decay models, arithmetic decay, geometric decay, and harmonic decay, along with their theoretical properties. In general, results of the proposed decay models depend largely on the local structure of a citation network as opposed to the global structure in the traditional approach. Thus, the significance of citation links and the associated documents that are overemphasized by the global structure in the traditional no-decay approach is treated more properly. For example, the traditional approach commonly assigns high value to documents that heavily reference others, such as review articles. Specifically in the geometric and harmonic decay models, only truly significant review articles will be included in the resulting main paths. We demonstrate this new approach and its properties through the DNA literature citation network.

Introduction

This study proposes a novel approach that modifies the underlying assumption of main path analysis as proposed by Hummon and Doreian (1989) and investigates the properties of the new technique. Main path analysis is a method capable of identifying chains of significant links in an acyclic directed network, thus extracting the bare bones of a large and complicated directed network. The most popular application of this method since its inception has been to "write the history of science," a concept put forward by Garfield, Sher, and Torpie (1964), who suggest that citations among documents can be used in constructing historical maps. Hummon and Doreian (1989) take the concept further with a rigorous methodology and coin the term *main path* for the most significant chain of citations. This method has been used widely in bibliometrics studies, such as mapping technological trajectories (Fontana, Nuvolari, & Verspagen, 2009; Verspagen, 2007), detecting technological changes (Lucio-Arias & Leydesdorff, 2008; Martinelli, 2012), and conducting literature reviews (Bhupatiraju, Nomaler, Triulzi, & Verspagen, 2012; Calero-Medina & Noyons, 2008; Colicchia & Strozzi, 2012; Harris, Beatty, Lecy, Cyr, & Shapiro, 2011; Liu, Lu, Lu, & Lin, 2013).

Main path analysis as proposed by Hummon and Doreian (1989) identifies the most significant citation chain in a citation network in two steps: (a) determine the "traversal count" for each citation link based on its structural position in the citation network and (b) search the significant citation chains according to the value of the traversal count. In determining the traversal count for each citation link, the method assumes that each citation link is an idea or knowledge flow channel and that when these ideas or knowledge flow through the link there is no loss or diminution in their content or strength; that is, the same content of these ideas or the strength of the knowledge will be carried forward forever. In practice, this constant transmission assumption is questionable. Taking the citation of academic articles as an example, when article B cites article A, presumably an idea or knowledge in A flows to B. Later, when another article C cites B, the original idea or knowledge in A will transmit to C, but probably not without any loss in its original content. When yet an even later article, D, cites C, it is reasonable to assume that the content of the original idea or the strength of

Received March 27, 2014; revised July 16, 2014; accepted July 16, 2014

 • Published online 8 January 2015 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/asi.23384

knowledge transmitted is reduced further. The concept that there is decay in strength when knowledge flows through the citation chain is the departure point of this study.

In citation analysis, the notion of strength decay is relevant only when an indirect citation is taken into account. Main path analysis is a well-known method for considering indirect citations. Aside from the main path literature, several previous citation-related studies have considered the effect of indirect citation. Rousseau (1987) discusses a mathematical technique to measure an article's total influence, taking both direct and indirect citations into consideration. Atallah and Rodriguez (2006) propose a cumulative patent citation index that looks at all the direct and indirect citations that have been made to a patent. In an attempt to measure an article's scientific impact, Fragkiadaki, Evangelidis, Samaras, and Dervos (2011) propose an indicator, the *f*-value, that accumulates all direct and indirect citations received by a research article. Hu, Rousseau, and Chen (2011) recommend another total influence indicator for a publication that takes indirect citations into account. Liu, Lu, and Ho (2012) define a scholar's total influence index and mainstream index by looking at indirect citations received by his or her academic work. Several of these works take account of knowledge decay. For example, Atallah and Rodriguez (2006) arithmetically weight each indirect citation according to its distance to the patent it cites in a citation chain, whereas Hu et al. (2011) and Fragkiadaki et al. (2011) assume a multiplier type of weight-reduction scheme.

This study modifies the original no-decay assumption for the main path analysis and introduces three types of decay models, arithmetic, geometric, and harmonic, to enhance the analysis. For each transmission down the citation chain, arithmetic decay incurs a decrease in knowledge strength by a fixed amount, whereas knowledge strength drops by a fixed ratio in geometric decay. The harmonic decay model assumes decay in accordance with harmonic series.

The remaining portion of this article is organized as follows. The next section describes the details of main path analysis and defines "effective" traversal counts, taking decay into consideration, and explores their properties. The section that follows empirically analyzes the characteristics of the new approaches. Discussions and conclusions are provided at the end.

A New Approach to Main Path Analysis

For a given citation network, traditional main path analysis begins by measuring the significance of each citation link in the network. The significance indicator is traversal count, the number of times a link is traversed under a specified situation. It then searches for the main paths by tracking citation links that are relatively significant. When decaying diffusion is taken into consideration, one needs to modify the method of measuring the significance of a citation link. This section presents three new algorithms, search-path arithmetic decay (SPAD), search-path geometric decay (SPGD), and search-path harmonic decay (SPHD), and discusses how these three traversal counts are different from the previous algorithms. It also briefly describes the method to search for main paths.

SPAD, SPGD, and SPHD

Assume that a set of documents, **U**, and the referencing relationships among them, R ($R \subseteq \mathbf{U} \times \mathbf{U}$), constitute a citation network, $\mathbf{N} = (\mathbf{U}, R)$. Thus, (u, v) denotes a citation link in which document v cites document u. One defines *sources* as the documents that are cited but cite no other documents, and *sinks* are documents that cite other documents but are not cited.

Knowledge dissemination in a citation network can be thought of as an imaginary messenger delivering knowledge from one document to the others along citation chains. Imagine a messenger on a mission to send knowledge from a specified origin document to a specified destination document in the citation network. There are many alternative paths between the two documents for the messenger, and the number of these paths depends on the structure of the citation network. If the messenger were to run through all these paths, then the traditional main path method takes the number of times a link is traversed by the messenger as an indication of the link's importance.

Based on such an analogy, we formulate the search path link count (SPLC) algorithm originally defined by Hummon and Doreian (1989) as follows.[1] For a simple messenger mission that delivers knowledge from a specified origin document, o, to a specified destination document, d, without decay, the traversal count $w_{o \to d}(u,v)$ for the link (u, v) is

$$w_{o \to d}(u, v) = \sum_{\text{traversal } i}^{\text{all traversals}} S_i(u, v), \tag{1}$$

where $s_i(u,v) = 1$ is a constant. Furthermore, assume that the messenger's mission is to deliver knowledge through all possible paths from all the ancestors of u (documents leading to u, including itself) to all the sinks. Hence, the significance of a link (u, v) under the SPLC algorithm is

$$w_{SPLC}(u, v) = \sum\nolimits_{\text{all combinations of ancestors and sinks}} w_{o \to d}(u, v). \tag{2.1}$$

We note that $w_{SPLC}(u,v)$ can also be expressed as multiples of the number of inflow paths to and the number of outflow paths from the link (u, v).

$$w_{SPLC}(u, v) = N^-_{\text{ancestors as origins}}(u) \cdot N^+_{\text{sinks as destinations}}(v), \tag{2.2}$$

[1]Other than SPLC, similar traversal count algorithms discussed in the main path literature include search path node pair (SPNP), node pair projection count (NPPC), and search path count (SPC). We discuss only SPLC, because among these algorithms it is conceptually closest to the new algorithms proposed in this study.

DOI: 10.1002/asi

where

$$N^{-}_{\text{ancestors as origins}}(u)=\begin{cases}1,\text{ if } u \text{ is an ancestor}\\ \sum_{\text{all ancestors}} \text{number of paths leading}\\ \text{from an ancestor to } u\text{, otherwise.}\end{cases} \quad (2.3)$$

$$N^{+}_{\text{sinks as destinations}}(v)=\begin{cases}1,\text{ if } v \text{ is a sink}\\ \sum_{\text{all sinks}} \text{number of paths leaving}\\ \text{from } v \text{ to a sink, otherwise.}\end{cases} \quad (2.4)$$

There is an important assumption behind such an approach: The strength of knowledge stays the same even though it has been passed down the citation chain for several generations. It is reasonable to believe that the strength of the particular piece of knowledge that the messenger carries decays by a certain amount each time the knowledge is passed on to a document further down the citation chain. We therefore propose to replace the traditional traversal count with the "effective traversal count," defining it as the sum of the effective knowledge strength on all traverses. Furthermore, for enhanced flexibility, we propose that knowledge can decay either arithmetically, geometrically, or harmonically. For each passing on, the strength of knowledge decreases by an "arithmetic decay factor," f, in arithmetic decay and diminishes by a "geometric decay factor," r, in geometric decay, whereas in harmonic decay the knowledge strength of each generation follows a simple harmonic sequence in which no decay parameter is needed. Both f and r are numbers between 0 and 1. Effective knowledge strength will be subtracted down to 0 in arithmetic decay, but will only be reduced to a small number in geometric and harmonic decays.

For a simple messenger mission that delivers knowledge from a specified origin document, o, to a specified destination document, d, at arithmetic decay factor f or geometric decay factor r, paralleling Equation 1, the effective traversal count for the link (u, v) is

$$w_{o\to d,a}(u,v,f)=\sum_{\text{traversal } i}^{\text{all traversals}} S_{i,a}(u,v,f),\text{ for arithmetic decay,} \quad (3.1)$$

$$w_{o\to d,g}(u,v,r)=\sum_{\text{traversal } i}^{\text{all traversals}} S_{i,g}(u,v,r),\ \text{for geometric decay, and} \quad (3.2)$$

$$w_{o\to d,h}(u,v,)=\sum_{\text{traversal } i}^{\text{all traversals}} S_{i,h}(u,v),\text{ for harmonic decay,} \quad (3.3)$$

where

$$S_{i,a}(u,v,f)=\begin{cases}1,\text{ if } u \text{ is the origin}\\ 1-n_i f,\text{ if } 1-n_i f>0,\ n_i \text{ is } u'\text{s network}\\ \text{distance to the origin for traversal } i,\\ 0,\text{ otherwise,}\end{cases} \quad (4.1)$$

$$S_{i,g}(u,v,r)=\begin{cases}1,\text{ if } u \text{ is the origin}\\ r^{n_i},\text{ otherwise, } n_i \text{ is } u'\text{s network}\\ \text{distance to the origin for traversal } i,\end{cases} \quad (4.2)$$

and

$$S_{i,h}(u,v)=\begin{cases}1,\text{ if } u \text{ is the origin}\\ 1/n_i,\text{ otherwise, } n_i \text{ is } u'\text{s network}\\ \text{distance to the origin for traversal } i,\end{cases} \quad (4.3)$$

In Equations 3.1, 3.2, and 3.3, $s_{i,a}(u,v,f)$, $s_{i,g}(u,v,r)$, and $s_{i,h}(u,v)$, respectively, are summed over all traversals by the messenger when he runs through all possible paths from a specified origin to a specified destination. It should be noted that the messenger will stop delivering when knowledge strength drops to 0. For the SPAD algorithm, this dramatically reduces the number of possible paths under the situation of a large decay because in general the number of possible paths generally grows exponentially with the number of generations for which the knowledge is able to diffuse. For the SPGD and SPHD algorithms, the number of possible paths is not affected, because knowledge strength does not fall exactly to 0.

As in the SPLC algorithm, one assumes that the messenger's mission is to deliver knowledge through all possible paths from all the ancestors of u (documents leading to u, including itself) to all the sinks. In such a case, the effective traversal count is, respectively, the sum of $w_{o\to d,a}(u,v,f)$, $w_{o\to d,g}(u,v,r)$, or $w_{o\to d,h}(u,v)$, where the origins are all the ancestors of u and the destinations are all the sinks. The significance of a link (u, v) under the SPAD, SPGD, and SPHD algorithms is, respectively,

$$w_{SPAD}(u,v,f)=\sum\nolimits_{\text{all combinations of ancestors and sinks}} w_{o\to d,a}(u,v,f), \quad (5.1)$$

$$w_{SPGD}(u,v,r)=\sum\nolimits_{\text{all combinations of ancestors and sinks}} w_{o\to d,g}(u,v,r),\text{ and} \quad (5.2)$$

$$w_{SPHD}(u,v)=\sum\nolimits_{\text{all combinations of ancestors and sinks}} w_{o\to d,h}(u,v). \quad (5.3)$$

In parallel with the expression of Equation 2.2, these traversal counts except for $w_{SPAD}(u,v,f)$ can also be expressed in the following forms.

$$w_{SPGD}(u,v,r)\,E^{-}_{\text{ancestors as origins}}(u)\cdot N^{+}_{\text{sinks as destinations}}(v),\text{ and} \quad (5.4)$$

$$w_{SPHD}(u,v)\,E^{-}_{h,\text{ ancestors as origins}}(u)\cdot N^{+}_{\text{sinks as destinations}}(v), \quad (5.5)$$

where $E^{-}_{g,\text{ ancestors as origins}}(u)$ and $E^{+}_{h,\text{ ancestors as origins}}(u)$ are the effective number of paths leading from all ancestors to u. Here, $E^{+}_{g,\text{ ancestors as origins}}(u)$ and $E^{+}_{h,\text{ ancestors as origins}}(u)$ are conceptual numbers that can be calculated indirectly after traversal counts are obtained from Equations 5.2 and 5.3, respectively.

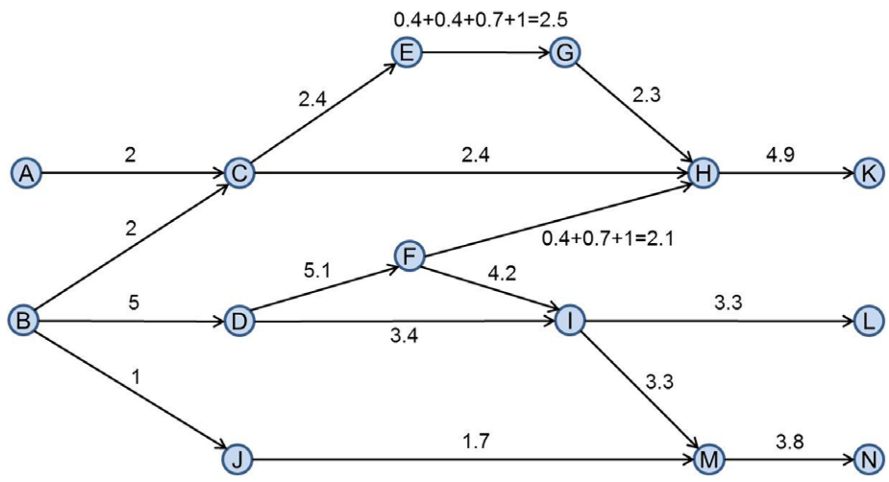

FIG. 1. A simple citation network. The number attached to each link is the w_{SPAD} value at an arithmetic decay factor of 0.3. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

We use the simple network in Figure 1 to demonstrate how the SPAD algorithm calculates $w_{SPAD}(u,v,f)$. The network consists of 14 nodes in which A and B are the sources, and K, L, and N are the sinks. Assume the arithmetic decay factor f is 0.3; that is, knowledge strength decreases by a fixed amount, 0.3, each time after crossing over a node. For link (F,H), node F has three ancestors (including itself), B, D, and F. Taking B as the origin, $s_{B\to K,a}(F,H,0.3)$ is equal to 0.4. Taking D and F as the origin, $s_{D\to K,a}(F,H,0.3)$ and $s_{F\to K,a}(F,H,0.3)$ are equal to 0.7 and 1, respectively. Therefore, $w_{SPAD}(F,H,0.3)$ is $0.4+0.7+1=2.1$. For link (E,G), the four ancestors of node E (including itself) are A, B, C, and E. Both $s_{A\to K,a}(E,G,0.3)$ and $s_{B\to K,a}(E,G,0.3)$ contribute 0.4, whereas $s_{C\to K,a}(E,G,0.3)$ and $s_{E\to K,a}(E,G,0.3)$, respectively, contribute 0.7 and 1. Consequently, $w_{SPAD}(E,G,0.3)$ is $0.4+0.4+0.7+1=2.5$. The calculation of $w_{SPAD}(u,v,f)$ for the rest of the links can be straightforwardly obtained by referencing Table 1, which enumerates all the paths originating from all the nodes except for the sinks and presents for each path the corresponding $s_{i,a}(u,v,f)$ at $f=0.3$ for all (u,v). The bottom row shows the Equation 5.1 results for each link, which are also indicated in Figure 1.

Figure 2 demonstrates the calculation of $w_{SPGD}(u,v,r)$. We set the geometric decay factor to 1/2. To calculate the effective traversal count of the link (F,H), taking B, D, and F as the origin, $s_{B\to K,g}(F,H,1/2)$, $s_{D\to K,g}(F,H,1/2)$, and $s_{F\to K,g}(F,H,1/2)$ are, respectively, equal to 1/4, 1/2, and 1. Therefore, $w_{SPGD}(F,H,1/2)$ is $1/4+1/2+1=1.75$. For link (E,G), the four ancestors of node E contribute $s_{A\to K,g}(E,G,1/2)=1/4$ $s_{B\to K,g}(E,G,1/2)=1/4$ $s_{C\to K,g}(E,G,1/2)=1/2$, and $s_{E\to K,g}(E,G,1/2)=1$, respectively. Thus, $w_{SPGD}(E,G,1/2)$ equals $1/4+1/4+1/2+1=2$. Figure 2 highlights w_{SPGD} of the rest of the network links.

The calculation for $w_{SPGD}(u,v)$ is quite similar to that for $w_{SPGD}(u,v,r)$. It is not difficult to see that $w_{SPHD}(F,H,1/2)=1/3+1/2+1=1.83$ and $w_{SPHD}(F,H,1/2)=1/3+1/3+1/2+1=2.17$. Figure 3 shows the SPHD values of the example network.

Properties of SPAD, SPGD, and SPHD

This section discusses the properties of SPAD, SPGD, and SPHD, both structurally independent and structurally dependent. Structurally independent properties refer to those that always hold true no matter how the network nodes are connected. Structurally dependent properties are those that prevail only for certain network connections.

Structurally independent properties. It is straightforward to derive from equation sets 3, 4, and 5 that

$$w_{SPAD}(u,v,f_1) > w_{SPAD}(u,v,f_2),\ \text{if } f_1 < f_2, \text{and} \tag{6.1}$$

$$w_{SPGD}(u,v,r_1) < w_{SPGD}(u,v,r_2),\ \text{if } r_1 < r_2. \tag{6.2}$$

Thus, for each citation link, w_{SPAD} decreases when the arithmetic decay factor increases, whereas w_{SPGD} decreases with the geometric decay factor. Explicitly, both the SPAD and the SPGD values for a citation link become smaller when there is a greater decay.

Comparing the SPAD, SPGD, and SPHD values against the value of SPLC, one obtains two additional properties. First, because of decay, $s_{i,a}(u,v,f) =< s_i(u,v)$, $s_{i,g}(u,v,r) =< s_i(u,v)$, and $s_{i,h}(u,v,r) =< s_i(u,v)$. Therefore, w_{SPAD}, w_{SPGD}, and w_{SPHD} for each link are always smaller than or equal to the traversal counts obtained by the SPLC algorithm; that is,

$$w_{SPAD}(u,v,f) =< w_{SPLC}(u,v), \tag{7.1}$$

$$w_{SPGD}(u,v,r) =< w_{SPLC}(u,v), \text{and} \tag{7.2}$$

DOI: 10.1002/asi

TABLE 1. Effective traversal for each link of the citation network in Figure 1 at an arithmetic decay factor of 0.3.

Links	(A,C)	(B,C)	(B,D)	(B,J)	(C,E)	(C,H)	(D,F)	(D,I)	(E,G)	(F,H)	(F,I)	(G,H)	(H,K)	(I,L)	(I,M)	(J,M)	(M,N)
A-C-E-G-H-K	1				0.7				0.4			0.1	0				
A-C-H-K	1					0.7							0.4				
B-C-E-G-H-K		1			0.7				0.4			0.1	0				
B-C-H-K		1				0.7							0.4				
B-D-F-H-K			1				0.7			0.4			0.1				
B-D-F-I-L			1				0.7				0.4			0.1			
B-D-F-I-M-N			1				0.7				0.4				0.1		0
B-D-I-L			1					0.7						0.4			
B-D-I-M-N			1					0.7							0.4		0.1
B-J-M-N				1												0.7	0.4
C-E-G-H-K					1				0.7			0.4	0.1				
C-H-K						1							0.7				
D-F-H-K							1			0.7			0.4				
D-F-I-L							1				0.7			0.4			
D-I-L								1						0.7			
D-F-I-M-N							1				0.7				0.4		0.1
D-I-M-N								1							0.7		0.4
E-G-H-K									1			0.7	0.4				
F-H-K										1			0.7				
F-I-L											1			0.7			
F-I-M-N											1				0.7		0.4
G-H-K												1	0.7				
H-K													1				
I-L														1			
I-M-N															1		0.7
J-M-N																1	0.7
M-N																	1
w_{SPAD}	2	2	5	1	2.4	2.4	5.1	3.4	2.5	2.1	4.2	2.3	4.9	3.3	3.3	1.7	3.8

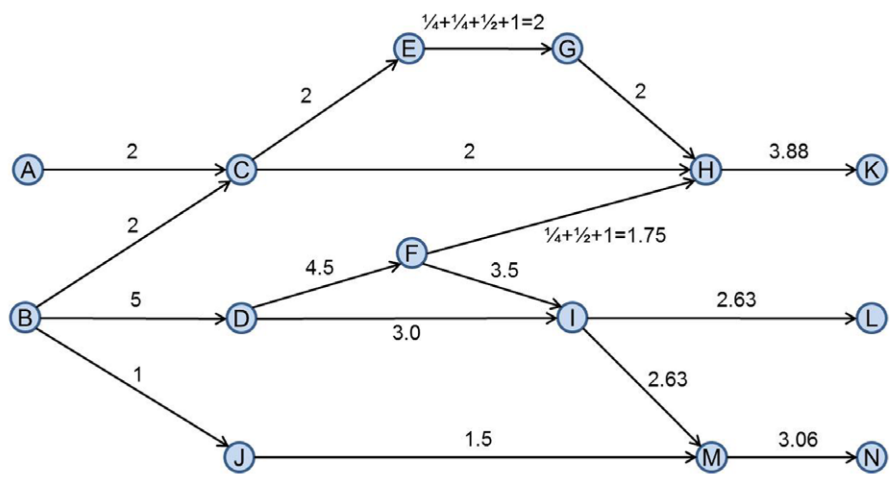

FIG. 2. A simple citation network. The number attached to each link is the w_{SPGD} value at a geometric decay factor of 1/2. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

$$w_{SPHD}(u,v) =< w_{SPLC}(u,v). \tag{7.3}$$

Furthermore, when there is no decay, at an arithmetic decay factor of 0 for SPAD or a geometric decay factor of 1 for SPGD, the traversal counts of both algorithms become the same as that of SPLC; that is,

$$w_{SPAD}(u,v,0) = w_{SPLC}(u,v), \text{ and} \tag{8.1}$$

$$w_{SPGD}(u,v,1) = w_{SPLC}(u,v). \tag{8.2}$$

Another extreme case is that of complete decay. For SPAD, when the arithmetic decay factor is 1, the SPAD

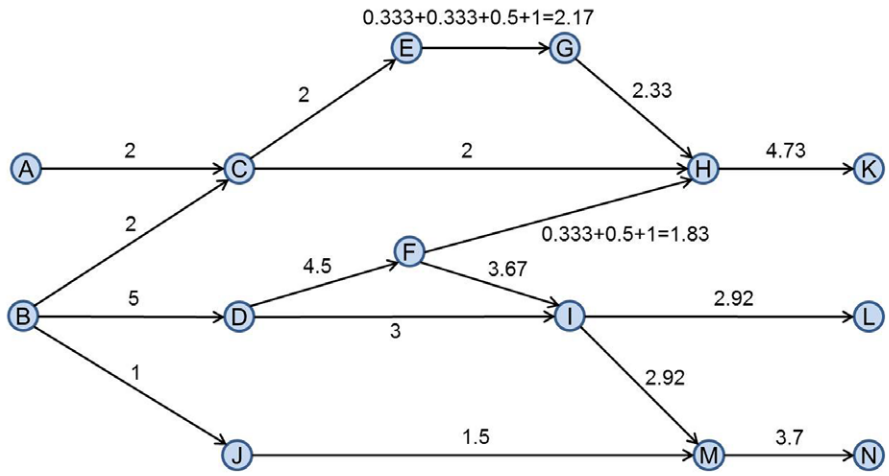

FIG. 3. A simple citation network. The number attached to each link is the w_{SPHD} value. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

value of a link is 1 for all links. This is because no knowledge can be diffused over the destination node such that each link is traversed only once.

We summarize the first four structurally independent properties for SPAD, SPGD, and SPHD below.

> Property 1: SPAD and SPGD values for each link decrease with increasing decay strength, if they do not not stay the same.

> Property 2: SPAD, SPGD, and SPHD values for each link are no greater than the value of SPLC.

> Property 3: When there is no decay, both SPAD and SPGD values become the same as that of SPLC for each link.

> Property 4: When there is complete decay, SPAD value is equal to 1 for all links.

Property 4 describes how, when there is complete decay, the SPAD value equals 1 for all links. However, when the decay is nearly complete, that is, the arithmetic decay factor is only marginally less than 1, the SPAD value for a link becomes virtually the same as the degree of the destination node. We thus encounter the fifth property.

> Property 5: Under the situation of nearly complete decay (arithmetic decay factor approaching 1), SPAD values for a link are virtually the same as the outdegree of the destination node except when the node is a sink.

In the language of bibliometrics, the SPAD value for a citation link is roughly the same as the citation count of the citing document under the situation when decay is slightly incomplete. In other words, when there is nearly complete arithmetic decay, the effective traversal counts of the links pointing toward a document are roughly equal to the number of times the document is cited. It is as if one takes only direct citations into account and uses citation count directly as the effective traversal count. Therefore, if we want to examine the main paths that consider only the effect of direct citations, then we set the arithmetic decay factor to a number very close to 1 and run the SPAD algorithm before searching the main paths.

Equation 5.2 predicts the situation for SPGD when decay is nearly complete, with a very small geometric decay factor. Because the contributions from all the ancestors except for u itself are nearly 0, this is effectively equivalent to the situation when there is only one ancestor, (u). Thus, SPGD values are virtually the same as the number of outflow paths of the destination nodes.

> Property 6: Under the situation of nearly complete decay (geometric decay factor approaching 0), SPGD values for a link diminish to the number of outflow paths of the destination node.

Properties 5 and 6 underscore a fundamental difference between the arithmetic decay and the geometric decay models. Both models reduce the dependence of effective traversal counts on the network structure, that is, the connectivity of all of node *u's* ancestors and all of node *v's* descendants, but the way in which they depend on the network structure is very different. For arithmetic decay, the knowledge content eventually becomes nil and stops disseminating, whereas knowledge continues its diffusion endlessly under geometric decay. This is the reason why, under the situation of nearly complete decay, arithmetic decay decreases the SPLC value of a link to the outdegree of the destination node, whereas a very strong geometric decay reduces that to the number of outflow paths.

Structurally dependent properties. The network structure as seen from a node illustrates the number of the node's direct connections and the paths flowing into and out of it. The properties discussed above are independent of the structure of a citation network and hold true for all types of network structures. It is more complicated to examine structurally dependent characteristics, but, for networks with specific characteristics, one property can be expected.

DOI: 10.1002/asi

Properties 2 and 3 indicate that SPGD and SPHD values for each citation link (u, v) are equal to or less than the value of SPLC, but how does the level of difference vary with the structural characteristic of each link? From Equations 2.2 and 5.4, we have

$$\frac{w_{SPGD}(u,v,r)}{w_{SPLC}(u,v)} = \frac{E^{-}_{g,\text{ ancestors as origins}}(u)}{N^{-}_{\text{ancestors as origins}}(u)}. \qquad (9)$$

The ratio between the values of SPGD and SPLC for any link (u, v) is sensitive to and decreases with the increasing number of inflow paths, $N^{-}(u)$. Thus, relatively speaking, the decay has a greater effect on those links with a greater number of inflow paths. The same reasoning applies for the ratio between SPHD and SPLC. We thus obtain a structurally dependent property as follows.

Property 7: The deviation (in terms of ratio) of SPGD and SPHD values from that of SPLC value is greater for those links with a higher number of inflow paths.

Property 7 is rather interesting because it implies that the significance of links emanating from a document that references a large number of other documents will be weakened if the SPLC algorithm is replaced by the SPGD or SPHD algorithm. Review articles typically reference many more documents than general types of articles. The links extending from them tend to have a high traversal count by the SPLC algorithm, which makes review articles more likely to appear on the main path even though they may not be highly cited. Property 7 suggests that SPGD and SPHD algorithms weight the significance of a review article more properly than the SPLC algorithm. By increasing the decay strength (providing a smaller geometric decay factor), one reduces the effective traversal count of the links extending out from a review article.

Searching for Main Paths

This study does not suggest modifying the second step of traditional main path analysis, which searches the main paths based on traversal counts. For the new approach, the search procedure is the same except that the search is now based on the effective traversal counts rather than the traditional traversal counts. The literature proposes several search algorithms (Batagelj, 2003; Hummon & Doreian, 1989; Liu & Lu, 2012). The local algorithm highlights significance at a particular point in time and tracks the most significant citation link at every possible splitting point. The global algorithm emphasizes the overall importance of citation chain and suggests the path with the largest overall traversal count.

In this study we apply the key-route algorithm (Liu & Lu, 2012) to explore the effects of SPAD, SPGD, and SPHD. The key-route algorithm guarantees the inclusion of a given key-route. A key-route is a link in the citation network with high (effective) traversal counts. The key-route algorithm begins by searching forward from the end node of a selected key-route until a sink is hit and searching backward from the start node of the same key-route until a source is encountered, and then it pieces together the given key-route and the two paths extending from the key-route. The search forward and backward can be based on either the local or global algorithm. Upon providing multiple key-routes and executing the procedure multiple times, one obtains multiple key-route main paths. One particular benefit of the key-route algorithm is that it controls the level of detail in the main paths by providing a different number of key-routes.

For the citation network depicted in Figure 1, by selecting the top link, (D,F), as the key-route, we obtain the key-route main path B-D-F-I-M-N if the global search is adopted or key-route main paths B-D-F-I-M-N and B-D-F-D-L if the local search is assumed.

Empirical Verification and Results

This section empirically examines the properties and the effects of decay diffusion. We take the DNA literature citation network shown by Hummon and Doreian (1989) as our empirical verification target. This citation network was originally presented by Garfield et al. (1964) and subsequently investigated by Hummon and Doreian (1989). It is constructed of 40 DNA milestone papers published in the period from 1820 to 1962. Among the 40 papers, one "isolate" and two "islands" (each containing two papers) are disconnected from all the other papers. Here, we analyze only the largest connected component that consists of the remaining 35 papers. Figure 4 presents this largest connected component, which contains a total of 59 citation links as well as eight source nodes (3, 6, 10, 11, 14, 16, 23, 26) and six sink nodes (4, 28, 37, 38, 39, 40). Each node in the figure is labeled beginning with an identification number followed by the author name(s) and the publishing year of the document. Table 2 presents basic network measures, including the number of inflow/outflow paths and traversal counts under different algorithms for the top 20 citation links having the largest SPLC values. The numbers of inflow and outflow paths are, respectively, $N^{-}(u)$ and $N^{+}(v)$ defined in Equations 2.3 and 2.4.

Empirical Results on the Structurally Independent Properties

The structurally independent properties described previously are obvious from observing traversal counts in Table 2. For example, the SPAD values for the citation link (*27Watson_Crick1953, 32Ochoa1955*) diminish gradually from 362.08 to 5.03 when the arithmetic decay factor increases from 0.01 to 0.99, exactly as predicted by property 1. The fact that all of the SPAD, SPGD, and SPHD values for each link are equal to or less than the SPLC values conforms to property 2. Property 3 is obviously compliant, because the SPAD value with a 0 arithmetic decay factor and the SPGD value under a geometric decay factor of 1 are exactly the same as the SPLC values for each citation link. An extreme

DOI: 10.1002/asi

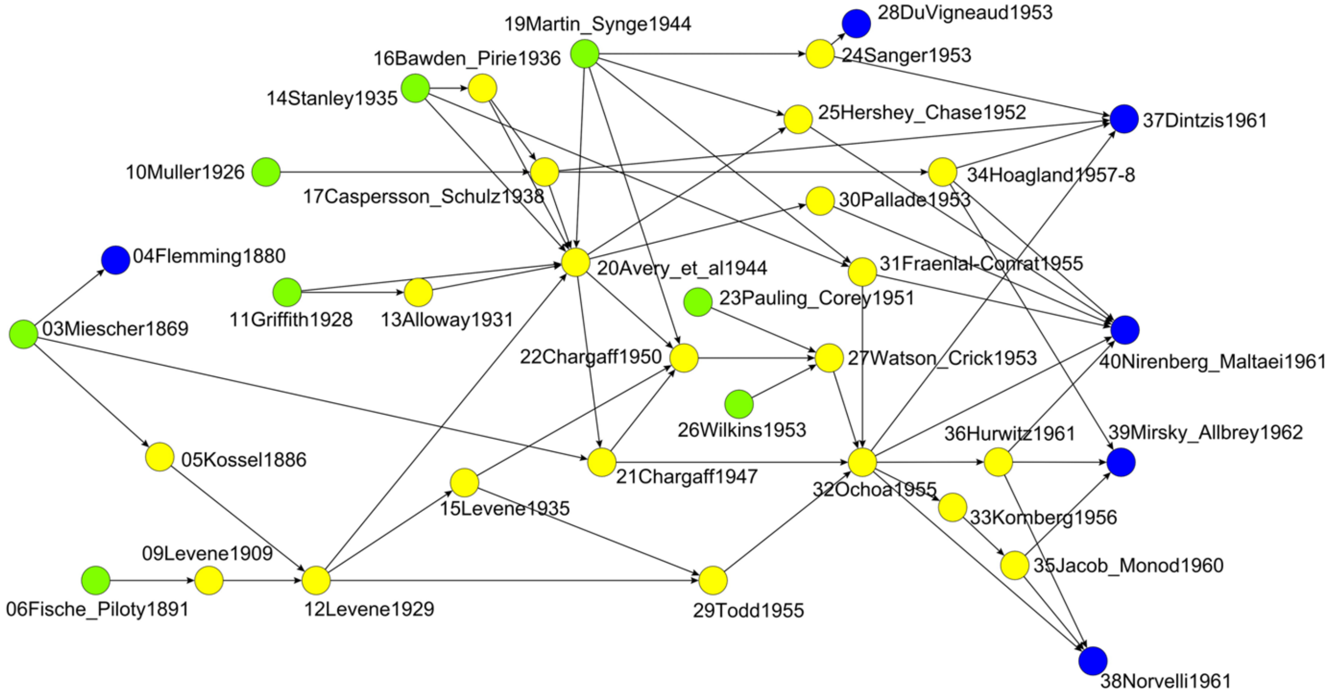

FIG. 4. DNA theory citation network. Redrawn Figure 1 from Hummon and Doreian (1989). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

case for SPAD meets the prediction of property 4 where SPAD values are all 1 at complete decay ($f = 1$).

At slightly incomplete decay, ($f = 0.99$), SPAD values are marginally greater than the outdegree (number of citations) of the destination node as predicted by property 5. For example, node 32Ochoa1955's outdegree is 5, and the SPAD values at an arithmetic decay factor of 0.99 of all the links that have 32Ochoa1955 as their destination node are either 5.02 or 5.03. One can therefore set the arithmetic decay factor to 0.99 to find the main paths that consider only direct citations.

For SPGD, a slightly incomplete decay, ($r = 0.01$), gives values marginally equal to the number of outflow paths of the destination node. From Table 2, we can see that the number of outflow paths for both 32Ochoa1955 and 27Watson_crick1953 is eight, and the links that have them as the destination nodes all have SPGD values marginally greater than eight for a geometric decay factor of 0.01, meeting the prediction of property 6.

Empirical Results on the Structurally Dependent Properties

Property 7 presents how the deviation (in terms of ratio) of SPGD and SPHD values from the value of SPLC is greater for links with a larger number of inflow paths. This property can be observed by comparing the SPLC and SPGD values of the links (*20Avery_et_al1944, 22Chargaff1950*) and (*12Levene1929, 20Avery_et_al1944*). The number of the inflow paths of the former (17) is much greater than that of the latter (five). The two links have roughly the same SPLC values, 136 and 130, respectively, yet the SPGD value (at a geometric decay factor of 1/4) of the link (*20Avery-_et_al1944*, *22Chargaff1950*) is smaller than that of the link (*12Levene1929*, *20Avery_et_al1944*) at 25.38 and 42.25, respectively. The SPHD value of the link (*20Avery-_et_al1944*, *22Chargaff1950*) is also smaller than that of the link (*12Levene1929*, *20Avery_et_al1944*) at 58.00 and 69.33, respectively. A direct result of this property is that the traversal count ranking of these two links is reversed after introducing decay through the SPGD or SPHD algorithm.

Empirical Results on the Main Paths

To compare the effects of different algorithms and associated decay parameters with the end results, we use the key-route search approach to find the main paths for the cases SPLC and SPAD at $f = 0.25$, SPAD at $f = 0.99$, SPGD at $r = 0.5$, SPGD at $r = 0.01$, and SPHD. The path search extends out from the top three key-routes with the global method. Figure 5 presents the resulting main paths for the DNA literature development for these six cases. In the figure the numbers associated with links are their effective traversal counts.

The main paths obtained from the SPLC algorithm exhibit a convergent–divergent form. Two paths beginning from 03Miescher1869 and 06Fische_Piloty1891 quickly merge to 12Levene1929. A dominant path follows, which is a chain consisting of the top three key-routes, (*27Watson_Crick1953*, *32Ochoa1955*), (*22Chargaff1950*, *27Watson_Crick1953*), and (*20Avery_et_al1944*, *21Chargaff1947*). At the end, the path diverges from

DOI: 10.1002/asi

TABLE 2. Basic values for the top 20 citation links with the largest SPLC values.

	From → to	Inflow paths	Outflow paths	No. of citations	SPLC	SPAD $f = 0$	SPAD $f = 0.01$	SPAD $f = 1/4$	SPAD $f = 0.99$	SPAD $f = 1$	SPGD $r = 1$	SPGD $r = 1/2$	SPGD $r = 1/4$	SPGD $r = 0.01$	SPHD
1	27watson_crick1953 → 32ochoa1955	47	8	5	376	376	362.08	36.25	5.03	1.00	376	48.63	17.81	8.24	92.76
2	22chargaff1950 → 27watson_crick1953	44	8	1	352	352	341.76	30.75	1.04	1.00	352	65.25	23.24	8.33	102.27
3	20avery_et_al1944 → 21chargaff1947	17	16	2	272	272	267.52	46.25	2.07	1.00	272	102.00	50.75	17.13	116.00
4	32ochoa1955 → 36hurwitz1961	82	3	3	246	246	236.07	29.00	3.04	1.00	246	27.21	8.65	3.12	56.34
5	33kornberg1956 → 35jacob_monod1960	83	2	2	166	166	157.74	9.75	2.01	1.00	166	11.07	3.44	2.02	31.56
6	32ochoa1955 → 33kornberg1956	82	2	1	164	164	157.38	16.00	1.04	1.00	164	18.14	5.76	2.08	37.56
7	21chargaff1947 → 22chargaff1950	19	8	1	152	152	148.32	11.50	1.02	1.00	152	37.50	16.34	8.17	51.47
8	21chargaff1947 → 32ochoa1955	19	8	5	152	152	148.32	37.50	5.02	1.00	152	37.50	16.34	8.17	51.47
9	20avery_et_al1944 → 22chargaff1950	17	8	1	136	136	133.76	14.00	1.07	1.00	136	51.00	25.38	8.56	58.00
10	12levene1929 → 20avery_et_al1944	5	26	4	130	130	128.44	20.50	4.02	1.00	130	65.00	42.25	26.53	69.33
11	17caspersson_schulz1938 → 20avery_et_al1944	4	26	4	104	104	102.96	18.50	4.02	1.00	104	58.50	40.63	26.52	60.67
12	5kossel1886 → 12levene1929	2	50	3	100	100	99.5	17.25	3.01	1.00	100	75.00	62.50	50.50	75.00
13	9levene1909 → 12levene1929	2	50	3	100	100	99.5	17.25	3.01	1.00	100	75.00	62.50	50.50	75.00
14	29todd1955 → 32ochoa1955	12	8	5	96	96	93.76	27.00	5.02	1.00	96	27.00	14.06	8.16	35.20
15	35jacob_monod1960 → 38norvelli1961	84	1	1	84	84	79.04	3.25	1.01	1.00	84	3.77	1.43	1.01	13.85
16	35jacob_monod1960 → 39mirsky_allbrey1962	84	1	1	84	84	79.04	3.25	1.01	1.00	84	3.77	1.43	1.01	13.85
17	36hurwitz1961 → 38norvelli1961	83	1	0	83	83	78.87	6.00	1.01	1.00	83	5.54	1.72	1.01	15.78
18	36hurwitz1961 → 39mirsky_allbrey1962	83	1	0	83	83	78.87	6.00	1.01	1.00	83	5.54	1.72	1.01	15.78
19	36hurwitz1961 → 40nirenberg_maltaei1961	83	1	0	83	83	78.87	6.00	1.01	1.00	83	5.54	1.72	1.01	15.78
20	32ochoa1955 → 37dintzis1961	82	1	0	82	82	78.69	12.00	1.04	1.00	82	9.07	2.88	1.04	18.78

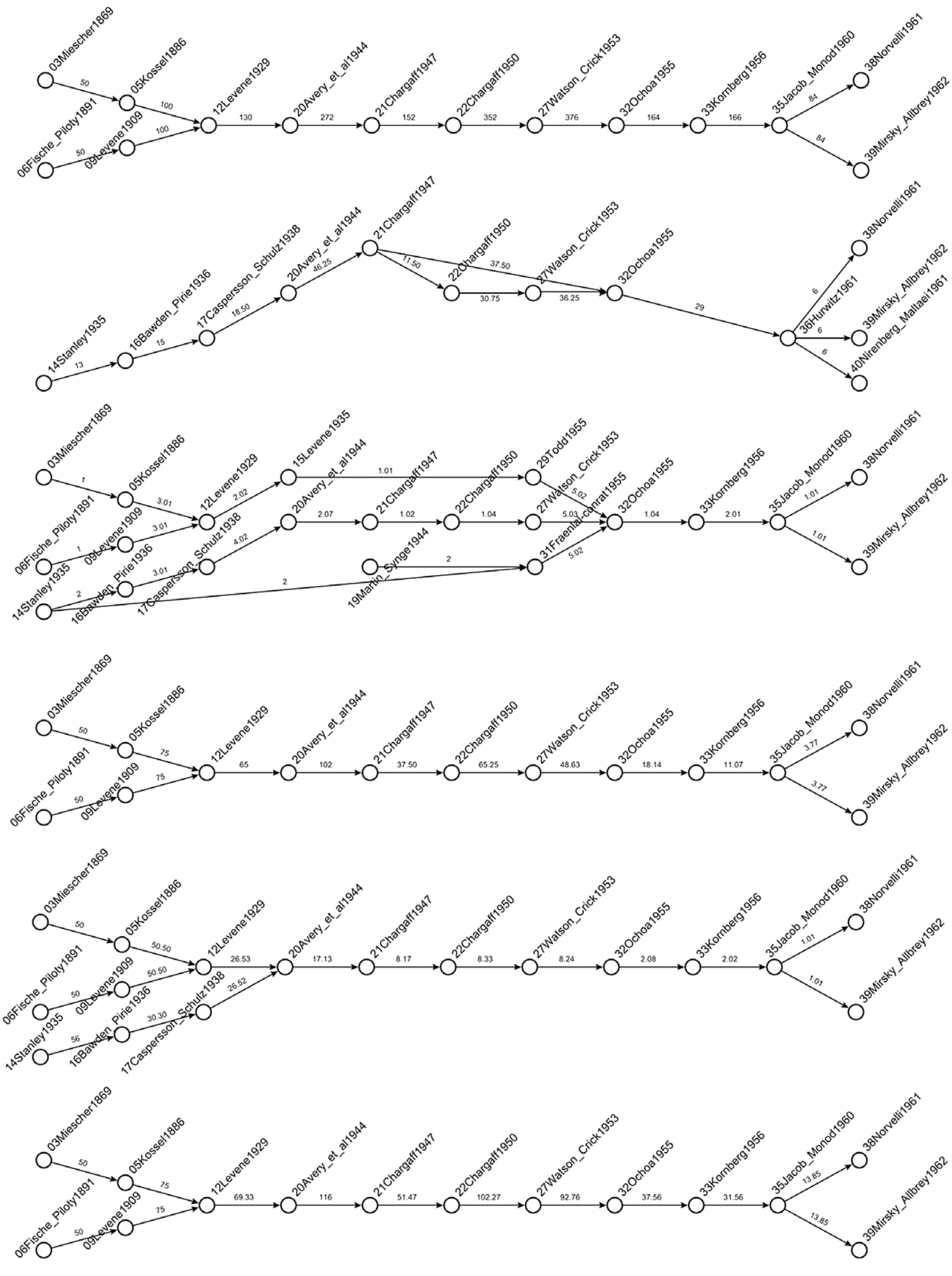

FIG. 5. Main paths of the DNA literature citation network for various traversal count algorithms.

DOI: 10.1002/asi

35Jacob_Monod1960 to 38Norvelli1961 and 39Mirsky _Allbrey1962. We note that there are small variations between these SPLC main paths and the one highlighted in Figure 4 of Hummon and Doreian (1989), but the backbone of these paths (12Levene1929 → 20Avery_et_al1944 → 21Chargaff1947 → 22Chargaff1950 → 27Watson_Crick 1953 → 32Ochoa1955) is the same.[2]

In all six cases, the chain 20Avery_et_al1944 → 21Chargaff1947 → 22Chargaff1950 → 27Watson_Crick1953 → 32Ochoa1955 appears repeatedly. In addition, the paths of SPGD at $r = 0.5$ and SPHD are exactly the same as the path of the SPLC algorithm, although the effective traversal counts for each citation link are very different. In the following paragraphs, we discuss the paths for the cases of SPAD at $f = 0.25$, SPAD at $f = 0.99$, and SPGD at $r = 0.01$, which exhibit paths different from the path of the SPLC algorithm.

SPAD at $f = 0.25$ is a reasonable setting, which assumes that knowledge strength reduces to 0 after passing through four generations. In this case, node 14Stanley1935 rather than 03Miesher1869 or 06Fische_Piloty1891 leads the main path. From the structure of the DNA citation network, this is quite reasonable because the traversal count of the link (*14Stanley1935, 16Bawden_Pirie1936*) is greater than that of the links (*3Miescher1869*, *5Kossel1886*) and (*6Fische_piloty1981*, *9Levene1909*). As a matter of fact, this is true in all six cases.

SPAD at $f = 0.99$ is an extreme case that essentially directly takes the citation counts as the traversal counts without considering the effect of indirect citations. The main paths in this case include many more paths than those of all the other cases, making it difficult to highlight dominant development paths. For a large-scale citation network, this approach will provide very complex main paths, thus defeating the whole purpose of main path analysis.

SPGD at $r = 0.01$ is another extreme case with a very strong geometric decay. It deviates slightly from the main paths generated by SPLC, mainly in the early stage of development. In general, the main paths generated by the SPGD algorithm at various geometric decay factors do not deviate much from those generated by the SPLC algorithm.

All the main paths discussed here contain more papers than the no-decay case presented in Figure 4 of Hummon and Doreian (1989). This is in contrast to the expectation that the new decay model would highlight fewer papers. Three reasons account for such results. First, Figure 4 in Hummon and Doreian (1989) presents only the path beginning from node 3 and ignores all other possibile paths. A complete analysis would also highlight the sequence 06Fische_Piloty1891 → 09Levene1909 → 12Levene1929 because it has as much importance as 03Miescher1869 → 05Kossel1886 → 12Levene1929 based on SPLC values. Second, we believe that there is a lower bound on the number of papers on the main path. The time element frequently comes into play and prevents further reduction on the number of papers on the main path. For example, there is a small possibility that two crucial papers farther apart in time will be linked directly on the main path, because they are very likely to be linked indirectly through several intermediate papers that may not be the crucial paper. Main path analysis usually preserves the connectivity role of the intermediate papers. One therefore may not be able to reduce further the number of papers on the main path in such a situation, which seems to be the case for the DNA literature. The sequence of the core literature (20Avery_et_al1944 → 21Chargaff1947 → 22Chargaff1950 → 27Watson_Crick 1953 → 32Ochoa1955) is not likely to be compressed further. The third reason why the main paths in Figure 5 do not highlight fewer papers is that they are the key-route main paths (Liu & Lu, 2012). Key-route main paths by definition present more papers than the traditional "priority first search" algorithm. We show key-route main paths in order to examine more details of the effects of decay.

[2]The main reason for the difference between the SPLC main paths in Figure 5 of this study and the one highlighted in Figure 4 of Hummon and Doreian (1989) is that this study adopts the key-route search algorithm, whereas, in Hummon and Doreian's (1989, p. 61) work, "priority first search was used to trace the main path from node 3." In other words, besides applying the priority first search algorithm, Hummon and Doreian (1989) trace only the path from node 3. The reason that we choose to use the key-route search algorithm is that it provides more details of the citation network, so readers can gain more insight from the traversal count values among different decay models.

Discussion

Unlike the SPLC and SPHD algorithms, which require no arbitrary parameters, the calculation of $w_{SPAD}(u,v,f)$ requires an arithmetic decay factor, f, and $w_{SPGD}(u,v,r)$ involves a geometric decay factor, r. Both are arbitrary numbers, and it is up to analysts to provide the one decay factor that they think will fit their context. In their study of indirect patent citations, Atallah and Rodriguez (2006) suggest an arithmetic decay scheme and set the arithmetic decay factor for a specific citation chain to be $1/D$, where D is the length of that citation chain at the time of investigation. The setting reduces the knowledge strength to zero at the end of any citation chain. The arithmetic decay factor for such a scheme is dynamic in time because the length citation chain increases when there is a new reference to the article at the end. This decay scheme also involves more mathematics, because one has to assess the arithmetic decay factor for each citation chain. The SPAD algorithm as proposed here assumes that the arithmetic decay factor is a constant across all citation chains and for all times. The difficulty of the scheme that this study proposes, nevertheless, lies in how to determine a proper arithmetic decay factor. Possibilities include adopting the reciprocal of the longest length among all citation chains (network diameter) or taking the average length of all citation chains as the arithmetic decay factor.

For geometric decay, whereas one heuristic approach assumes that knowledge decays to half of its original strength for each generation and sets r to 1/2, Fragkiadaki et al. (2011) choose to use the value 1/2.2. This value was computed empirically from data in the CiteSeer database,

DOI: 10.1002/asi

which covers papers mostly in the area of computer and information science. They found that “for each 1-gen citation an article receives from within our database, it also receives 2.22 2-gen citations” (Fragkiadaki et al. 2011, p. 678). The rationale is not fully explained, but it seems to assume that when an article is cited multiple times knowledge strength split equally to the citing article, which is certainly an issue requiring further discussion. Moreover, the empirically determined value should be valid only within the context that generates it and should not be treated as a universal constant. All in all, the best way to determine the arithmetic decay factor and geometric decay factor remains an interesting topic for future research.

When considering indirect citations, the number of citation generations to consider is also an open issue. Rousseau (1987, p. 227) suggests heuristically that “two to four will be reasonable, especially considering the large matrices one has to handle.” Hu et al. (2011) and Fragkiadaki et al. (2011), in their empirical demonstrations, trace down to three generations. The definitions for SPAD, SPGD, and SPHD algorithms in the current study, nevertheless, do not include this arbitrary parameter and trace down infinitely until the knowledge strength diminishes to 0. The advantage is that it takes all indirect citations into consideration no matter how insignificant they are. The downside is that it uses a lot of computing power when the citation network is large. In practice, the knowledge strength in the SPGD and SPHD algorithms never drops to 0, so all indirect citations are taken into account. On the other hand, the SPAD algorithm has knowledge strength diminishing to 0 quickly for a medium level of an arithmetic decay factor and as a result considers only a few citation generations.

Conclusions

We have introduced here the concept of knowledge decay in main path analysis. The concept modifies the traditional assumption that knowledge emanates in citation networks without losing any of its strength or content. We have also proposed three types of decay models, arithmetic decay, geometric decay, and harmonic decay, as well as the formulae that define the associated effective traversal counts. The properties of these proposed algorithms are presented and fully discussed. Finally, the effect of knowledge decay on main path analysis is explored using the DNA literature citation network. The empirical results provide practical references for researchers who would like to consider knowledge decay when applying main path analysis.

The contribution of this study lies in two areas. First, it suggests that one should reconsider the previous assumption that knowledge emanates without any decay in main path analysis and provides alternatives to that end. The new decay approach is closer to real-world situations and therefore finds more reasonable main paths. Second, this study provides a methodological formulation and highlights the properties of the new approach, thus providing a sound basis for further development of main path analysis.

Acknowledgments

We thank two anonymous reviewers for their constructive comments, which have greatly improved the accuracy and readability of this article. This work is partially supported by Taiwan’s Ministry of Science and Technology grants NSC 102-2410-H-011-008-MY3.

References

Atallah, G., & Rodriguez, G. (2006). Indirect patent citations. Scientometrics, 67(3), 437–465.

Batagelj, V. (2003). Efficient algorithms for citation network analysis. Preprint series. Ljubljana: University of Ljubljana, Institute of Mathematics, Physics and Mechanics Department of Theoretical Computer Science.

Bhupatiraju, S., Nomaler, O., Triulzi, G., & Verspagen, B. (2012). Knowledge flows—Analyzing the core literature of innovation, entrepreneurship and science and technology studies. Research Policy, 41(7), 1205–1218.

Calero-Medina, C., & Noyons, E.C.M. (2008). Combining mapping and citation network analysis for a better understanding of the scientific development: The case of the absorptive capacity field. Journal of Informetrics, 2(4), 272–279.

Colicchia, C., & Strozzi, F. (2012). Supply chain risk management: A new methodology for a systematic literature review. Supply Chain Management: An International Journal, 17(4), 403–418.

Fontana, R., Nuvolari, A., & Verspagen, B. (2009). Mapping technological trajectories as patent citation networks. An application to data communication standards. Economics of Innovation and New Technology, 18(4), 311–336.

Fragkiadaki, E., Evangelidis, G., Samaras, N., & Dervos, D.A. (2011). *f*-Value: Measuring an article’s scientific impact. Scientometrics, 86(3), 671–686.

Garfield, E., Sher, I.H., & Torpie, R.J. (1964). The use of citation data in writing the history of science. Philadelphia: Institute for Scientific Information.

Harris, J.K., Beatty, K.E., Lecy, J.D., Cyr, J.M., & Shapiro, R.M. (2011). Mapping the multidisciplinary field of public health services and systems research. American Journal of Preventive Medicine, 41(1), 105–111.

Hu, X.J., Rousseau, R., & Chen, J. (2011). On the definition of forward and backward citation generations. Journal of Informetrics, 5(1), 27–36.

Hummon, N.P., & Doreian, P. (1989). Connectivity in a citation network: The development of DNA theory. Social Networks, 11(1), 39–63.

Liu, J.S., & Lu, L.Y.Y. (2012). An integrated approach for main path analysis: Development of the Hirsch index as an example. Journal of the American Society for Information Science and Technology, 63(3), 528–542.

Liu, J.S., Lu, L.Y.Y., & Ho, M.H.C. (2012). Total influence and mainstream measures for scientific researchers. Journal of Informetrics, 6(4), 496–504.

Liu, J.S., Lu, L.Y.Y., Lu, W.M., & Lin, B.J.Y. (2013). Data envelopment analysis 1978–2010: A citation-based literature survey. OMEGA: The International Journal of Management Science, 41(1), 3–15.

Lucio-Arias, D., & Leydesdorff, L. (2008). Main-path analysis and path-dependent transitions in HistCite (TM)-based historiograms. Journal of the American Society for Information Science and Technology, 59(12), 1948–1962.

Martinelli, A. (2012). An emerging paradigm or just another trajectory? Understanding the nature of technological changes using engineering heuristics in the telecommunications switching industry. Research Policy, 41(2), 414–429.

Rousseau, R. (1987). The Gozinto theorem: Using citations to determine influences on a scientific publication. Scientometrics, 11(3–4), 217–229.

Verspagen, B. (2007). Mapping technological trajectories as patent citation networks: A study on the history of fuel cell research. Advances in Complex Systems, 10(1), 93–115.