On the properties of von Neumann kernels for link analysis

Masashi Shimbo · Takahiko Ito · Daichi Mochihashi ·
Yuji Matsumoto

Received: 28 March 2007 / Revised: 1 October 2008 / Accepted: 9 October 2008 / Published online: 12 November 2008
Springer Science+Business Media, LLC 2008

Abstract We study the effectiveness of Kandola et al.'s von Neumann kernels as a link analysis measure. We show that von Neumann kernels subsume Kleinberg's HITS importance at the limit of their parameter range. Because they reduce to co-citation relatedness at the other end of the parameter, von Neumann kernels give us a spectrum of link analysis measures between the two established measures of importance and relatedness. Hence the relative merit of a vertex can be evaluated in terms of varying trade-offs between the global importance and the local relatedness within a single parametric framework. As a generalization of HITS, von Neumann kernels inherit the problem of *topic drift*. When a graph consists of multiple communities each representing a different topic, HITS is known to rank vertices in the most dominant community higher regardless of the query term. This problem persists in von Neumann kernels; when the parameter is biased towards the direction of global importance, they tend to rank vertices in the dominant community uniformly

Editors: Thomas Gärtner and Gemma C. Garriga.

This work was carried out while T. Ito was a Ph.D. student at Nara Institute of Science and Technology.

M. Shimbo (✉) · T. Ito · Y. Matsumoto
Graduate School of Information Science, Nara Institute of Science and Technology, 8916-5 Takayama,
Ikoma, Nara 630-0192, Japan
e-mail: shimbo@is.naist.jp

T. Ito
e-mail: tito@microsoft.com

Y. Matsumoto
e-mail: matsu@is.naist.jp

Present address:

T. Ito
FAST, a Microsoft Subsidiary, Daido-Seimei Kasumigaseki, 1-4-2 Kasumigaseki, Chiyoda-ku, Tokyo
100-0013, Japan

D. Mochihashi
NTT Communication Science Laboratories, Hikaridai 2-4, Keihanna Science City, Kyoto 619-0237,
Japan
e-mail: daichi@cslab.kecl.ntt.co.jp

higher irrespective of the community of the seed vertex relative to which the ranking is computed. To alleviate topic drift, we propose to use of a PLSI-based technique in combination with von Neumann kernels. Experimental results on a citation network of scientific papers demonstrate the characteristics and effectiveness of von Neumann kernels.

Keywords Link analysis · Recommender system · von Neumann kernel · HITS · Topic drift

1 Introduction

PageRank (Brin and Page 1998) and HITS (Hypertext-Induced Topic Search) (Kleinberg 1999) are two of the most popular methods for evaluating the *importance* of web pages. They are used extensively in search engines to compute the ranking of web pages from the network structure induced by hyperlinks.

Even in the early days of bibliometric studies, links (or citations) between documents were already recognized as a major source of information for analyzing the relationship between scientific papers, authors, and journals. A line of research in bibliometrics have dealt with quantifying the *relatedness* of two given documents. Co-citation coupling (Small 1973) is one such classical measure of relatedness still widely used, for example, by the popular scientific literature database CiteSeer (Bollacker et al. 1998) to recommend related papers to the user.

Discussion of these two lines of link analysis measures, importance and relatedness, has remained somewhat independent in the past. The first objective of this paper is to present a unified framework that accounts for both importance and relatedness, and to further define measures intermediate between the two.

Our approach is based on Kandola et al.'s von Neumann kernel (Kandola et al. 2003). We show that this kernel nicely bridges the gap between importance and relatedness. Specifically, it subsumes not only the co-citation and bibliographic coupling (Kessler 1963) relatedness at one end of the parameter range, but also the HITS importance as the other extreme. Between these established relatedness and importance measures lies a spectrum of intermediate link analysis measures, all of which can be obtained by tuning a single parameter of the kernel.

von Neumann kernels thus provide an attractive framework for link analysis, but there are some difficulties in their practical application. In Sect. 4 of this paper, we discuss one of these difficulties, known as *topic drift* (Bharat and Henzinger 1998). This problem is noticeable when a graph consists of multiple communities each addressing a different topic; if von Neumann kernels are applied to such a graph, with the parameter biased towards importance, they assign the highest scores to the vertices in the dominant community irrespective of the seed vertex, the vertex relative to which the ranking is computed.

We propose a method for avoiding topic drift in von Neumann kernels. To this end, we model the generative process of citations borrowing the idea from Cohn and Chang's PHITS (Cohn and Chang 2000), and construct distinct graphs for individual communities. All these 'community graphs' have the same vertex set as the original graph, but their edges are reweighted according to the generative probability of the corresponding citation in the respective community. Applying von Neumann kernels to the community graphs allows us to take communities into consideration even when they are biased towards importance.

In Sect. 5, we discuss some kernels and ranking methods related to von Neumann kernels. We also discuss the connection between the community-based von Neumann kernels we propose and Hofmann's PLSI-based Fisher kernels (Hofmann 2000).

This paper is an integrated and extended version of the work that appeared in (Ito et al. 2005, 2006; Shimbo and Ito 2006), with a focus on von Neumann kernels. The recommendation system simulation in Sect. 6.3 is substantially updated from the preliminary evaluation presented in (Shimbo et al. 2007), with the evaluation criterion modified to be more realistic, and more kernel-based ranking methods are compared.

2 Importance and relatedness measures in link analysis

Most of the existing link analysis measures can be classified into one of the two types:¹ *relatedness* or *importance*.

Relatedness measures quantify the similarity of two vertices in a graph, or the relevance of one vertex to another. Importance, on the other hand, is the measure for ranking a given group of vertices in the order of their significance, impact, or popularity within the group. In this section, we review several link analysis methods of relatedness and importance that are relevant to subsequent discussions. For more comprehensive surveys on link analysis, see (Baldi et al. 2003, Chap. 5) and (Dhyani et al. 2002).

We assume that the vertices of a graph are indexed by natural numbers, and identify a vertex with its index. We also use v_i to denote the i th vertex. Throughout this paper, matrices are denoted by uppercase letters, and vectors are denoted by boldface letters. The symbol T denotes matrix transposition. All matrices are square and all vectors are column vectors unless noted otherwise. For integers i and j , $[A]_{ij}$ represents the (i, j) -element of a matrix A , and $[\mathbf{v}]_i$ represents the i th component of vector $\mathbf{v}$. We denote by $\mathbf{1}$ the vector of all 1's, and by $\mathbf{e}_i$ the unit vector with 1 at the i th component and 0 at the remaining components. I is an identity matrix, and for any matrix A , let $\rho(A)$ be its spectral radius, i.e., the largest modulus of eigenvalues of A .

We consider both non-weighted and weighted graphs, but in all cases, edge weights are positive. For non-weighted graphs, all edges weights are assumed to be 1. The *adjacency matrix* A of a graph is a matrix with elements $[A]_{ij}$ holding the weight of the edge from vertex i to vertex j if such an edge exists, and 0 if not. A graph is undirected if for any pair of vertices (i, j) , $[A]_{ij} = [A]_{ji}$; otherwise it is directed. Thus, the adjacency matrix is symmetric for undirected graphs.

For a path in a graph, its weight is defined as the product of the weights of all edges composing the path, and its length is the (cumulative) number of these edges. The distance between vertices is the minimum length of paths between them.

2.1 Relatedness: co-citation and bibliographic coupling

Link analysis assumes that in a graph modeling a network structure such as bibliographic citations and the web, an edge (e.g., a citation or a hyperlink) between a pair of vertices (papers or web pages) signifies that these vertices are in some sense related. Hence the degree of relatedness can be inferred from the vertex proximity induced by the existence of edges. It is however not assumed that all related vertices are spelled out by those connected by edges, because edges can be omitted for various reasons. The challenge of link analysis is to capture relationship between vertices even if they are not directly connected.

Co-citation (Small 1973) and bibliographic coupling (Kessler 1963) are the standard methods of computing relatedness between documents in a citation network, or, more generally, between vertices in a graph.

¹One notable exception of this classification is White and Smyth's *relative importance* (White and Smyth 2003) we will discuss in Sect. 5.

Co-citation coupling defines relatedness between documents as the number of other documents citing them both. And bibliographic coupling defines relatedness between two documents as the number of common references cited by the two. These measures can be calculated from the adjacency matrix of a citation graph.

Given an adjacency matrix A , the number of co-citations between vertices i and j is given by the (i, j) -element of *co-citation matrix* $A^T A$.

Similarly, *bibliographic coupling matrix* $A A^T$ gives the values of bibliographic coupling. Because these matrices are symmetric, their graph counterparts, the *co-citation graph* and *bibliographic coupling graph*, are undirected, even if the original citation graph is not.

2.2 Importance: Kleinberg's HITS

Because of the difficulty in computing the importance of documents from their contents, citation counts (or, in-degree of vertices) have long been used as the index of document importance. A support for this approach was provided by several researchers; even though citations are made for various reasons, a positive correlation was observed between the number of citations and the significance or impact of the cited work. See the reference (le Pair 1988) for the list of literature on this topic. The recently proposed SALSA algorithm (Lempel and Moran 2001) can be viewed as providing a justification for using the vertex in-degree as an importance index, from a random walk point of view.

Kleinberg's HITS (Hypertext-Induced Topic Search) (Kleinberg 1999), is a more recent and sophisticated method for evaluating document importance. HITS assigns two scores to each document (vertex), called the authority and hub scores. The assumption behind HITS is the presence of *mutual reinforcement* between authorities and hubs. That is, authoritative documents are the ones that are cited by many hub documents, and hub documents are the ones that are citing many authorities. Let A be the adjacency matrix of a graph. The HITS algorithm computes the following recursion over $n = 0, 1, \dots$ starting from $\mathbf{a}_{(0)} = \mathbf{h}_{(0)} = \mathbf{1}$.

$$\mathbf{a}_{(n+1)} = \frac{A^T \mathbf{h}_{(n)}}{|A^T \mathbf{h}_{(n)}|}, \quad \mathbf{h}_{(n+1)} = \frac{A \mathbf{a}_{(n+1)}}{|A \mathbf{a}_{(n+1)}|}. \quad (1)$$

The i th component of the *authority vector* $\lim_{n \rightarrow \infty} \mathbf{a}_{(n)}$ represents the *authority score* of vertex i . Similarly, the *hub vector* $\lim_{n \rightarrow \infty} \mathbf{h}_{(n)}$ gives the *hub scores*. It is well known that the above recursion reduces to the power method for computing dominant eigenvectors. If $A^T A$ and $A A^T$ have a unique dominant eigenvalue, the authority and hub vectors exist and equal to the dominant eigenvectors of $A^T A$ and $A A^T$, respectively.

3 von Neumann kernels as a unified link analysis measure

In this section, we present a formulation of link analysis measures that are an intermediate between importance of vertices and their relatedness.

The concept of an intermediate between importance and relatedness might sound ill-defined, since importance is a measure defined on individual vertices, whereas relatedness is defined between them. However, given an importance score vector $\mathbf{v}$ such as the HITS authority vector, $\mathbf{v} \mathbf{v}^T$ defines a matrix in which every row (and column) i gives a ranking

² $A^T A$ and $A A^T$ have the same set of eigenvalues.

of vertices identical to the one given by $\mathbf{v}$, provided that³ $[\mathbf{v}]_i \neq 0$. Importance can thus be treated as a function over a pair of vertices, or a matrix, as well.

The basis of our formulation is the von Neumann kernel due to Kandola et al. (2003). This kernel is defined over graph vertices, and is symmetric positive semidefinite. In other words, it provides a Gram matrix holding inner products of vertices of a graph.

3.1 von Neumann kernels

The *von Neumann kernels* (Kandola et al. 2003) were first proposed as a method for computing document similarity from terms occurring in documents, in a manner akin to the latent semantic analysis (Deerwester et al. 1990). The von Neumann kernels are defined in terms of the term-by-document rectangular matrix X whose (i, j) -element is the (possibly reweighted) frequency of the i th term occurring in document j . We first need the document correlation matrix $K = X^T X$ and term correlation matrix $G = X X^T$. The von Neumann kernel matrices are then defined in terms of K and G as follows.

Definition 1 Let X be a term-by-document matrix, and let $K = X^T X$ and $G = X X^T$. The *von Neumann kernel* matrices with *diffusion factor* $\gamma (\geq 0)$, denoted by $\hat{K}_\gamma$ and $\hat{G}_\gamma$, are defined as the solution to the following system of equations.

$$\hat{K}_\gamma = \gamma X^T \hat{G}_\gamma X + K, \quad \hat{G}_\gamma = \gamma X \hat{K}_\gamma X^T + G. \quad (2)$$

The similarity between documents i and j is given by $[\hat{K}_\gamma]_{ij}$, and the term similarity is given by the elements of $\hat{G}_\gamma$.

When $\gamma < 1/\rho(K)$, the solution exists and is given by

$$\hat{K}_\gamma = K(I - \gamma K)^{-1}, \quad \hat{G}_\gamma = G(I - \gamma G)^{-1}. \quad (3)$$

Shawe-Taylor and Cristianini (2004) presented an equivalent Neumann series representation of these kernels.

$$\hat{K}_\gamma = K \sum_{n=0}^{\infty} \gamma^n K^n, \quad \hat{G}_\gamma = G \sum_{n=0}^{\infty} \gamma^n G^n. \quad (4)$$

3.2 Link analysis with von Neumann kernels

Using the adjacency matrix A of a citation graph in place of the term-by-document matrix X , we obtain the von Neumann kernels based on citation information. We thus have $K = A^T A$ and $G = A A^T$, which coincide with the co-citation and bibliographic coupling matrices, respectively. For convenience, let us introduce the shorthand

$$N_\gamma(B) = B \sum_{n=0}^{\infty} (\gamma B)^n. \quad (5)$$

The von Neumann kernels based on the adjacency matrix A can then be written as

$$\hat{K}_\gamma = N_\gamma(A^T A) = A^T A \sum_{n=0}^{\infty} \gamma^n (A^T A)^n, \quad (6)$$

³The assumption of $[\mathbf{v}]_i \neq 0$ is weaker than it appears. For example, HITS authority vectors satisfy $[\mathbf{v}]_i \neq 0$ for every vertex i whenever the co-citation graph is connected.

$$\hat{G}_\gamma = N_\gamma(AA^T) = AA^T \sum_{n=0}^{\infty} \gamma^n (AA^T)^n. \quad (7)$$

As seen from (6) and (7), $\hat{G}_\gamma$ can be obtained simply by transposing A in (6). We therefore focus on $\hat{K} = N_\gamma(A^T A)$ below.

3.3 Interpretation

Equation (6) shows that the von Neumann kernel matrix $N_\gamma(A^T A)$ is a weighted sum of $(A^T A)^n$ over $n = 1, 2, \dots$. As we can interpret $[A^T A]_{ij}$ as the multiplicity of edges between vertices i and j in the corresponding co-citation graph, $[(A^T A)^n]_{ij}$ represents the number of paths of length n between vertices i and j in the same graph. Thus, we see that each element of the kernel matrix equals the weighted sum of the number of paths between vertices.⁴

Each term $(A^T A)^n$ permits further interpretation as follows. First, it is easy to see that $(A^T A)^n$ with small n captures the degree of proximity between any two vertices. The number $[(A^T A)^n]_{ij}$ of paths between i and j in the co-citation graph is non-zero only if there is a path of length less than or equal to n between them in the co-citation graph.⁵ Since relatedness is a measure based on proximity as we argued in Sect. 2.1, $[(A^T A)^n]_{ij}$ can be interpreted as indicating the relatedness between vertices i and j . In particular, $(A^T A)^1$ is exactly the co-citation matrix.

For larger n values, we have the following theorem.

Theorem 1 *Let $\lambda > 0$ be the dominant eigenvalue of a symmetric positive semidefinite matrix $A^T A$. If λ is a simple eigenvalue (i.e., λ has a multiplicity of one), there exists a unit eigenvector $\mathbf{v}$ corresponding to λ such that $(A^T A/\lambda)^n \rightarrow \mathbf{v}\mathbf{v}^T$ as $n \rightarrow \infty$.*

Proof See Appendix A. □

If $A^T A$ is nonnegative, it has a nonnegative dominant eigenvector. This fact, together with the above theorem, leads to the following corollary.

Corollary 1 *Let the dominant eigenvalue λ of a nonnegative symmetric positive semidefinite matrix $A^T A$ be nonzero and simple, and let $\mathbf{v}$ be a non-negative eigenvector corresponding to λ . For any vertex i with $[\mathbf{v}]_i \neq 0$ and any vertex pair (j, k) such that $[\mathbf{v}]_j > [\mathbf{v}]_k$, there exists an integer m satisfying $[(A^T A)^n]_{ij} > [(A^T A)^n]_{ik}$ for all $n \geq m$.*

Thus, if the co-citation graph (whose adjacency matrix is $A^T A$) is connected, the dominant eigenvector $\mathbf{v}$ of $A^T A$ is equal to the HITS authority vector up to scaling. In this case, Corollary 1 tells us that if we regard a row (or column) vector of $(A^T A)^n$ as a score vector and compute the rankings from the magnitude of its components, the rankings tend to the HITS authority rankings as $n \rightarrow \infty$, irrespective of the chosen row (or column). In other

⁴More generally, if A holds arbitrary non-negative elements, $[A^T A]_{ij}$ represents the sum of the weight of all paths between vertices i and j in the co-citation graph.

⁵For general graphs, $[(A^T A)^n]_{ij}$ is non-zero if and only if a path of length *exactly* n exists between vertices i and j . However, in the co-citation graph induced by $A^T A$, all vertices with a non-zero in-degree in the original citation graph will have a self-loop, and hence there is always a path of length n if there is a path of length less than n between two vertices.

words, the number of paths of length n between i and j is an indicator of the importance of these vertices, provided that n is sufficiently large.

To sum up, we see that summing $(A^T A)^n$ over $n = 1, 2, \dots$ as in (6) can be interpreted as the mixture of relatedness (when n is small) and importance (when n is large). As a special case, the von Neumann kernels subsume co-citation at $\gamma = 0$. Near the opposite extreme of the parameter range, i.e., $\gamma \rightarrow 1/\rho(A^T A)$, the ranking induced by the von Neumann kernels is also identical to the HITS importance, as stated by the following theorem.

Theorem 2 *Let $\lambda > 0$ be the dominant eigenvalue of a symmetric positive semidefinite matrix $A^T A$. If λ is a simple eigenvalue, there exists a unit eigenvector $\mathbf{v}$ corresponding to λ such that*

$$\left(\frac{1}{\lambda} - \gamma\right) N_\gamma(A^T A) \rightarrow \mathbf{v}\mathbf{v}^T \quad \text{as } \gamma \rightarrow \frac{1}{\lambda},$$

where the limit is taken from below.

The proof is again deferred to Appendix A.

A similar argument can be made for the relationship between bibliographic coupling, HITS hub vector (eigenvector of AA^T), and the von Neumann kernel $\hat{G}_\gamma = N_\gamma(AA^T)$.

3.4 A few notes on application

We conclude this section with a few facts that can help applying von Neumann kernels in practice.

Parameter sensitivity analysis The derivative of the von Neumann kernels can be analytically computed at a given point γ by the following simple equation.

$$\frac{\partial N_\gamma(B)}{\partial \gamma} = (N_\gamma(B))^2. \quad (8)$$

This can be exploited for, for instance, parameter tuning and parameter sensitivity analysis.

Reducing memory requirement If we are concerned with the importance of vertices relative to a single vertex i rather than the entire kernel matrix, or if the entire kernel matrix cannot be kept on memory, we can reduce the space requirement by summing $(\gamma A^T A)^n \mathbf{e}_i$ over $n = 1, 2, \dots$ until convergence to obtain the i th column of the von Neumann kernel matrix; recall that $\mathbf{e}_i$ is a unit vector with only 1 at the i th component. This way, an iteration consists only of one matrix-vector multiplication similar to the HITS computation, and one vector summation.

4 Topic drift and von Neumann kernels

In this section, we discuss a major shortcoming of von Neumann kernels known as *topic drift* (Bharat and Henzinger 1998), and propose a solution for this problem.

Topic drift is a phenomenon first observed with HITS. If applied to a graph with multiple communities,⁶ HITS assigns the highest scores to the vertices (documents) in the dominant

⁶To the best of our knowledge, there seems to be no consensus on what constitutes a ‘community’ in the literature. A community in this paper merely means a set of documents in a citation network citing more documents within the community than those outside.

Fig. 1 A citation graph with multiple communities, and the induced co-citation graph

community of the graph. It follows that search engines using HITS as the back-end may output documents unrelated to the user's query, if the query is different from the topic of the dominant community.

Although it was first reported in the context of search engines which assume queries are terms, it is not related to the form of queries; topic drift can be observed even in a pure link analysis setting. In the following example, we assume that the query is posed as a vertex (*seed vertex*) in a given graph, and the rankings of other vertices relative to this seed vertex are to be computed.

Consider the graph of Fig. 1(a) containing two communities. The HITS authority rankings for the vertices of this graph are (from the most important to the least important) $v_1 > v_2 > v_3 > v_4 > v_5 > v_6$. Notice that the ranks of documents in community 2, namely v_4 , v_5 and v_6 , are uniformly lower than those of documents in community 1 (v_1 and v_2).

As von Neumann kernels reduce to HITS when $\gamma \rightarrow 1/\lambda = 1/\rho(A^T A)$, they also inherit the topic drift problem from HITS. For convenience, let us rewrite the formula (5) for von Neumann kernels using a 'normalized' parameter $\beta = \gamma\rho(A^T A)$ which falls into the range $0 \leq \beta < 1$ instead of $0 \leq \gamma < 1/\rho(A^T A)$. Let $B = A^T A$.

$$\overline{N}_\beta(B) = B \sum_{n=0}^{\infty} \left(\frac{\beta}{\rho(B)} \right)^n B^n. \quad (9)$$

Figure 1(b) depicts the co-citation graph induced by the graph of Fig. 1(a). The von Neumann kernel matrix $\overline{N}_{0.99}$ for this graph is shown below.

$$\overline{N}_{0.99} = \begin{pmatrix} 477.37 & 225.98 & 127.64 & 62.70 & 15.33 & 2.90 \\ 225.98 & 108.53 & 59.64 & 29.30 & 7.16 & 1.36 \\ 127.64 & 59.64 & 37.87 & 21.67 & 5.30 & 1.00 \\ 62.70 & 29.30 & 21.67 & 23.74 & 7.34 & 1.39 \\ 15.33 & 7.16 & 5.30 & 7.34 & 5.16 & 2.17 \\ 2.90 & 1.36 & 1.00 & 1.39 & 2.17 & 1.60 \end{pmatrix}. \quad (10)$$

Only the sub-matrix of the kernel matrix for documents v_1 through v_6 is shown. The remaining rows and columns are for isolated vertices in Fig. 1(b), and thus their elements are constantly 0 regardless of β .

The (i, j) -element of this matrix represents the importance of the j th document relative to the i th document. For example, relative to document v_3 (third row), document v_1 is the most important document because $[\bar{N}_{0.99}]_{3,1} = 127.64$ is the largest in the third row.

Now let us focus on v_6 which belongs to community 2. The ranking for v_6 is given by the sixth row in the matrix, and again, v_1 has the largest value. However, this ranking for v_6 is different from our intuition; because v_1 and v_6 belong to different communities, it would be more natural if the importance score of v_4 , a vertex in the same community as v_6 , were higher than v_1 .

If we increase the diffusion factor (e.g., $\beta = 0.999$), the ranking relative to each document will eventually be identical to that of the HITS authority ranking and determined irrespective of the community where the document belongs.

To prevent the HITS importance rankings from diverging from the topic (or community) the user is interested in, Kleinberg (1999) proposed to first extract documents that include query terms, and apply HITS to the *neighbor graph*, which is a subgraph of the extracted documents and their neighbors.

We may also apply von Neumann kernels to a neighbor graph of documents containing query terms, but this solution requires the query be terms and the document contents be accessible. Depending on the type of data (e.g., citation networks, where paper contents are often not available), it is not always possible to meet this requirement. Moreover, Bharat and Henzinger (1998) pointed out that there can be a discrepancy between queries and the topics of high-ranked documents, even if HITS is applied to a neighbor graph.

4.1 Generative model of citations

In this section, we propose a method for reducing topic drift in von Neumann kernels. Although it uses a technique based on Probabilistic Latent Semantic Indexing (PLSI) (Hofmann 1999, 2001), document contents are not required.

PLSI is a method for modeling the generative process of documents. In PLSI, the probability of word w_j occurring in document d_i is modeled by the following equation.

$$p(d_i, w_j) = \sum_{t=1}^k p(t) p(d_i|t) p(w_j|t), \quad (11)$$

where $t = 1, 2, \dots, k$ represents a latent *topic* of documents. This model of PLSI, also called *aspect model*, assumes that words and documents are conditionally independent given latent topic t . Parameters $p(t)$, $p(d_i|t)$ and $p(w_j|t)$ are estimated from empirical observations with a variant of the Expectation Maximization algorithm.

Cohn and Chang (2000) modeled the generative process of citations analogously to PLSI. In their model, the probability of a citation is defined by using citations in place of words in (11). Thus, the probability of a document i citing another document j is

$$p(d_i, c_j) = \sum_{t=1}^k p(t) p(d_i|t) p(c_j|t), \quad (12)$$

where d_i represents a citation emanating from document i , and c_j represents a citation to document j .

They also proposed to use $p(c_j|t)$, the generative probability of a citation within the community addressing topic t , as the importance of document j in this community. Cohn and Chang tested this method, called Probabilistic HITS, or PHITS for short, on the Cora citation dataset, and demonstrated that it gives intuitive rankings of scientific papers in each inferred community t .

PHITS thus gives the importance rankings in individual communities, but it is not obvious how to compute other quantities provided by von Neumann kernels; i.e., relatedness between two documents, and importance of documents relative to a given document. Below, we exploit the model of (12) in combination with von Neumann kernels to compute these quantities.

4.2 Applying von Neumann kernels to community graphs

To alleviate topic drift in von Neumann kernels, we use the generative model of citations given by (12), and apply von Neumann kernels to the weighted graphs induced by the model. Unlike Kleinberg's neighbor graph method, document contents are not used to derive these community graphs.

The proposed method maintains the property of von Neumann kernels as a mixture of importance and relatedness, and also takes communities into consideration, even when they are biased towards importance. This method takes the number of communities k as input, and consists of the following four steps.

Step 1 Apply PLSI (see (12)) to the citation graph to obtain $p(t|d_i, c_j)$, which is the posterior probability of community t ($t = 1, \dots, k$) given a citation from document i to document j .

With Bayes' rule and PLSI's conditional independence assumption, i.e., $p(d, c|t) = p(d|t)p(c|t)$, $p(t|d_i, c_j)$ can be computed by

$$p(t|d_i, c_j) = \frac{p(d_i, c_j|t)p(t)}{p(d_i, c_j)} = \frac{p(d_i|t)p(c_j|t)p(t)}{\sum_{t=1}^k p(d_i|t)p(c_j|t)p(t)}.$$

Step 2 Create each community graph $\mathcal{G}_t$ for $t = 1, \dots, n$, with the same vertex set as the original citation graph, but assign $[A_t]_{ij} = p(t|d_i, c_j)[A]_{ij}$ as the weight of the edge from vertex i to j . Here, A and A_t are the adjacency matrices of the original citation graph and the t th community graph, respectively.

The intuition here is to distribute the original edge weights to each community graph t in proportion to the probability $p(t|d_i, c_j)$. Notice $\sum_{t=1}^k p(t|d_i, c_j) = 1$.

Step 3 For each t , apply von Neumann kernels to the co-citation matrix $A_t^T A_t$. The resulting von Neumann kernel for community t is

$$\bar{N}_{t,\beta}(A_t^T A_t) = A_t^T A_t \sum_{n=0}^{\infty} \left(\frac{\beta}{\rho(A_t^T A_t)} \right)^n (A_t^T A_t)^n. \quad (13)$$

This equation is identical to (6), except that A_t is used in place of the adjacency matrix A of the original citation graph, and the diffusion parameter γ is rescaled as β .

Step 4 Finally, sum the von Neumann kernels in (13) over all communities t to obtain

$$N_{\beta}^{\text{comm}} = \sum_{t=1}^k \bar{N}_{t,\beta} (A_t^T A_t). \quad (14)$$

This matrix N_{β}^{comm} retains positive semidefiniteness, because the sum of positive semidefinite kernels is still a positive semidefinite kernel (Haussler 1999).

4.3 Illustration

We demonstrate the three steps of our proposed method (Sect. 4.2) with the graph of Fig. 1(a). Let the parameter β for the von Neumann kernel be 0.99, and the hyperparameter (the number of latent communities) k be 2 in PLSI. In Steps 1 and 2, we apply PLSI to Fig. 1(a) and obtain the two community graphs shown in Fig. 2(a) and (b).

In Step 3, von Neumann kernels are applied to each community graph. Notice that vertices v_1 and v_2 , the most important vertices in terms of the HITS authority ranking, are not connected to v_4 , v_5 or v_6 in the two community graphs. Hence the scores for v_1 and v_2 relative to the latter vertices are 0 in the respective kernels.

In Step 4, we compute (14), which is the sum of the von Neumann kernel matrices for communities 1 and 2. This yields the final kernel N_{β}^{comm} .

At $\beta = 0.99$, the resulting kernel matrix is as follows.

$$N_{0.99}^{\text{comm}} = \begin{pmatrix} 489.90 & 235.81 & 91.63 & 0.00 & 0.00 & 0.00 \\ 235.81 & 115.04 & 43.56 & 0.00 & 0.00 & 0.00 \\ 91.63 & 43.56 & 40.82 & 90.96 & 37.59 & 10.18 \\ 0.00 & 0.00 & 90.96 & 375.79 & 157.23 & 42.60 \\ 0.00 & 0.00 & 37.59 & 157.23 & 69.38 & 20.07 \\ 0.00 & 0.00 & 10.18 & 42.60 & 20.07 & 6.70 \end{pmatrix}. \quad (15)$$

The rankings in the 4th to 6th rows are significantly different from those of $\bar{N}_{0.99}$ (see (10)). In the 6th row of $\bar{N}_{0.99}$, the highest score (2.90) was assigned to v_1 , which is a vertex in Community 1, despite v_6 being in Community 2. By contrast, (15) assigns the largest value (42.60) to v_4 , the vertex with the largest number of in-links in Community 2. On the other hand, the ranking relative to vertex v_3 , located between the two communities, is a mixture the importance in two communities by (14).

Fig. 2 Two community graphs induced by the graph of Fig. 1(a). All non-labeled edges have a weight of 1.0

5 Related kernels and link analysis measures

In this section, we discuss some related kernels on graph vertices and link analysis measures.

5.1 Exponential diffusion kernels

Several authors have proposed to use matrix exponentials instead of the Neumann series (Kandola et al. 2003; Kondor and Lafferty 2001; Shawe-Taylor and Cristianini 2004) to define kernels.

Definition 2 Let $\mathcal{G}$ be an undirected graph with positive weights, and B be its adjacency matrix. The *exponential diffusion kernel* matrix E_γ on $\mathcal{G}$, with diffusion factor $\gamma \geq 0$ is given by

$$E_\gamma(B) = \exp(\gamma B) = \sum_{n=0}^{\infty} \frac{(\gamma B)^n}{n!}. \quad (16)$$

Unlike the von Neumann kernels, the series on the right hand side is convergent for all $\gamma \geq 0$, and the exponential diffusion kernels are always positive semidefinite.

Using a technique similar to the one used for the proof of Theorem 2, we can also show that the rankings induced by the limit of the exponential diffusion kernel tends towards the HITS authority ranking as well, if $\mathcal{G}$ is a co-citation graph.

Theorem 3 Let $\mathbf{v}$ be the HITS authority vector of a citation graph A , and λ the dominant eigenvector of $A^T A$. Then the matrix $E_\gamma(A^T A) / \exp(\gamma \lambda)$ converges to $\mathbf{v}\mathbf{v}^T$ as $\gamma \rightarrow \infty$.

Hence, the exponential diffusion kernels are also affected by topic drift, just like the von Neumann kernels. We can consider their community-based variation similar to the community-based von Neumann kernels presented in the previous section, as follows. First let us rescale γ to β , as we did with the community-based von Neumann kernels⁷

$$\bar{E}_\beta(A^T A) = \exp\left(\frac{\beta}{\rho(A^T A)} A^T A\right). \quad (17)$$

And then,

$$E_\beta^{\text{comm}} = \sum_{t=1}^k \bar{E}_\beta(A_t^T A_t). \quad (18)$$

As in Sect. 4.2, A_t is the adjacency matrix of community graph t . In Sect. 6.3, we will test this community-based exponential kernel in a paper recommendation task.

5.2 Laplacian-based kernels

The regularized Laplacian (Chebotarev and Shamis 1997; Smola and Kondor 2003; Zhou and Schölkopf 2004) and the heat diffusion kernel (Chung 1997; Kondor and Lafferty 2001),

⁷Unlike von Neumann kernels, however, rescaled β is not upper-bounded for exponential diffusion kernels. The scaling here is only for the consistency with the community-based von Neumann kernels.

have definitions similar to the von Neumann and exponential diffusion kernels. However, they are defined in terms of the graph Laplacian instead of the raw adjacency matrix. This leads to the property quite different from that of the kernels based on adjacency matrices. For brevity, we only discuss the regularized Laplacian, which like von Neumann kernels is based on the Neumann series.

Definition 3 Let B be the adjacency matrix of an undirected graph $\mathcal{G}$ with positive edge weights. The (combinatorial) *Laplacian* of $\mathcal{G}$ is defined as $L(B) = D(B) - B$, where $D(B)$ is a diagonal matrix with diagonals $[D(B)]_{ii} = \sum_j [B]_{ij}$.

The regularized Laplacian is defined in terms of the graph Laplacian as follows.

Definition 4 Let B be a nonnegative symmetric matrix and let $\gamma \geq 0$. The matrix

$$R_\gamma(B) = (I + \gamma L(B))^{-1} \quad (19)$$

is called the *regularized Laplacian* on the graph induced by B with diffusion factor γ .

The right-hand side of (19) is the closed form of the Neumann series $\sum_n (-\gamma L(B))^n$. However, this series converges only for $\gamma < 1/\rho(L(B))$, whereas $R_\gamma(B) = (I + \gamma L(B))^{-1}$ exists for all $\gamma > 0$ except $\gamma = 1/\rho(L(B))$, and if it exists, it is positive semidefinite.

At the limit $\gamma \rightarrow \infty$, the regularized Laplacian exhibits a property strikingly different from that of the von Neumann kernels: if the graph represented by adjacency matrix B is connected, $R_\gamma(B)$ tends to a uniform matrix as $\gamma \rightarrow \infty$; see Theorem 6 in Appendix A. In other words, in the limit $\gamma \rightarrow \infty$, the regularized Laplacian regards all vertices equally important, or equally related to each other. This is contrastive to von Neumann kernels, as they give rankings identical to HITS near the upper bound of the admissible parameter range.

In (Ito et al. 2005), we have compared the properties of the regularized Laplacian and von Neumann kernels, and argued that the regularized Laplacian remains a relatedness measure throughout their parameter range, unlike von Neumann kernels.

Another interesting kernel based on the Laplacian of graphs is Saerens et al.'s commute time kernels (Saerens et al. 2004; Fouss et al. 2007) defined as the pseudoinverse of the Laplacian matrix. They showed that the distance measure provided by this kernel has a natural interpretation as the *commute time*, defined as the expected number of steps required for a random walker starting from a given vertex to reach another given vertex for the first time, and then to go back to the starting vertex.

Nadler et al. (2006) recently proposed a measure called diffusion map, which is also related to the graph Laplacian.

5.3 Smola and Kondor's kernel-based regularization framework

Smola and Kondor (2003) pointed out that all the Laplacian-based kernels discussed above can be viewed as a form of regularization of Laplacian eigenvalues. For example, the regularized Laplacian maps each Laplacian eigenvalue λ to $1/(1 + \gamma\lambda)$, whereas the commute-time kernel maps λ to $1/\lambda$ for $\lambda \neq 0$ and $\lambda = 0$ to 0.

They also mention the connection between Laplacian kernels and some importance computation methods including HITS. In this respect, Smola and Kondor's formulation appears quite different from the one presented in this paper. They state that for a given vertex in a

regular graph, its HITS score is given by the length of its corresponding vector in the feature space induced by kernels based on the (normalized) Laplacian.

In the formulation presented in this paper, by contrast, the HITS importance scores are obtained as an extremum of von Neumann kernels defined in terms of the adjacency matrix instead of the Laplacian. In addition, our formulation takes the components of the kernel matrices directly as the score indicating relatedness or importance, instead of the distance of vertices in the kernel-induced feature space.

5.4 Relative importance

Relative importance is a new class of link analysis measure recently proposed by White and Smyth (2003). This measure is defined as the ‘importance of vertices in a graph relative to one or more root vertices.’ The notion of relative importance contrasts with the ‘global’ importance measures such as HITS and PageRank.

White and Smyth made a convincing argument that simply applying global importance algorithms to a subgraph surrounding the root vertices does not yield a precise estimate of relative importance, because the root vertices are not given any special preference during importance computation. This can be thought of as one reason *topic drift* persists in HITS even if Kleinberg’s neighbor graph approach is taken. We have discussed topic drift in HITS and its implication on von Neumann kernels in Sect. 4.

von Neumann kernels fit naturally as relative importance, and as a bonus clarify the relationship between relative importance and relatedness (namely, the co-citation and bibliographic coupling relatedness), an issue not addressed in White and Smyth’s work. With von Neumann kernels, relative importance is formulated explicitly as an intermediate between relatedness and global importance.

5.5 Hofmann’s PLSI-based Fisher kernels

Hofmann (2000) used his PLSI model to define a kernel, but in a quite different way from how we used it in Sect. 4. We briefly discuss Hofmann’s work and the connection between his method and ours.

The Fisher kernel (Jaakkola and Haussler 1999) is a general technique to obtain a kernel from generative models. To derive a Fisher kernel from a generative model, we need to compute the Fisher score $u(d, \theta)$, the gradient of the log-likelihood function for data d with respect to parameters θ . Given the Fisher score $u(d, \theta)$, the Fisher kernel is given by following equation.

$$F(d_i, d_j) = u(d_i; \theta)^T I(\theta)^{-1} u(d_j; \theta).$$

Here, $I(\theta)$ is the Fisher information matrix, typically approximated by an identity matrix.

Hofmann proposed a Fisher kernel based on PLSI for computing document similarity. The expectation of the log-likelihood in PLSI is given as follows.

$$\log p(d_i) = \sum_j \hat{p}(w_j | d_i) \sum_{t=1}^k \log p(w_j | t) p(t | d_i)$$

where $p(w_j | t)$ and $\hat{p}(w_j | d_i)$ respectively represent the probability that word w_j is generated from topic t , and the empirical probability of word w_j observed in document d_i . Hofmann

computed the derivatives of the log-likelihood function of PLSI with respect to parameters $\rho_{jt} = 2\sqrt{p(w_j|t)}$ and $\rho_t = 2\sqrt{p(t)}$ to yield two types of Fisher kernels. For the first parameter ρ_{jt} , the derivative is given by

$$\frac{\partial \log p(d_i)}{\partial \rho_{jt}} = \frac{\hat{p}(w_j|d_i)p(t|d_i, w_j)}{\sqrt{p(w_j|t)}},$$

so the resulting Fisher kernel is

$$\bar{F}(d, d') = \sum_{j=1} \hat{p}(w_j|d) \hat{p}(w_j|d') \sum_{t=1}^k \frac{p(t|d, w_j)p(t|d', w_j)}{p(w_j|t)}.$$

Hofmann argued that this is an inner product of empirical distribution (the $\hat{p}$ factors) but common words contribute only if they occur in the same topical context; the latter statement follows from the rightmost summation of the product of posterior probabilities. In parallel to PHITS, we can define the Fisher kernels for a citation graph by replacing word w_j with c_j , a citation to document j . The Fisher kernel $\bar{F}$ based on ρ_{jt} in this case can be written as a matrix

$$\bar{F} = \sum_{t=1}^k \bar{A}_t^T \bar{A}_t, \quad (20)$$

where the (i, j) -element of $\bar{A}_t$ is $\hat{p}(c_j|d_i)p(t|d_i, c_j)/\sqrt{p(c_j|t)}$ if there is a citation from vertex i to j , and 0 otherwise.

As seen from the term $\bar{A}_t^T \bar{A}_t$ in (20), this Fisher kernel essentially computes the sum of (reweighted) co-citation counts for communities $t = 1, \dots, k$; recall that $p(t|d_i, c_j)$ is the weight assigned to edge (i, j) in the t th community graph (see Step 2 of Sect. 4.2). A subtle difference between the graphs induced by $\bar{A}_t$ and A_t is that the edge weights in the former are reweighted by $\hat{p}(c_j|d_i)/\sqrt{p(c_j|t)}$.

Because computation in (20) for each community t is identical to co-citation coupling (up to reweighting), the score is zero for any pair of vertices not directly co-cited by other vertices. In contrast, the community-based von Neumann kernel N_β^{comm} (see (14)) proposed in the previous section computes the weighted sum of all *paths* between documents, so they assign non-zero weights to any document pair as long as they are connected in a community graph; see the discussion on von Neumann kernels in Sect. 3.3.

Hofmann also proposed another Fisher kernel, this time based on the derivative with respect to parameter $2\sqrt{\rho_t}$. This Fisher kernel is defined as

$$\tilde{F}(d, d') = \sum_{t=1}^k p(t|d)p(t|d')/p(t),$$

and is the (weighted) inner product in the low dimensional factor (topic) representation. This kernel is apparently intended to address the above limitation of kernel $\bar{F}$, but it lacks a clear interpretation as a link analysis measure, as it does not involve citation c .

6 Evaluation with bibliographic citation data

We carried out an empirical evaluation of von Neumann kernels and other graph-based ranking methods, using a citation network of papers on natural language processing (NLP).

This citation network was built from scanned bibliographic references in the papers from major NLP journals and conference proceedings. Matching each scanned items against an existing database of NLP literature and filtering out the unmatched items resulted in a citation graph with 2867 vertices (papers) and 6939 edges (citations).⁸

After the co-citation graph was built from the citation graph, von Neumann and other kernel matrices were computed with MATLAB and GNU Octave from the closed form representation (e.g., $A^T A (I - \gamma A^T A)^{-1}$ for von Neumann kernels) using the built-in matrix inversion and eigenvalue functions.

Each kernel matrix was treated as a ranking method by taking the i th row vector of the matrix as the score vector for the i th vertex (paper). Given the i th score vector, or the rankings induced thereof, we call the i th vertex the *seed vertex* (or *seed paper*) of these rankings.

6.1 Plain von Neumann kernels

In Table 1, we present the ranking lists produced by the von Neumann kernels under various diffusion parameters, for a specific seed paper ‘Empirical studies in discourse’ by M.A. Walker and J.D. Moore, *Computational Linguistics* 23(1):1–12, 1997. These ranking lists are meant to illustrate the characteristics of the von Neumann kernels under different parameter values.

This specific seed paper was chosen for illustration, for the following reason. It is a paper on discourse processing, and presumably, the papers most related to it must be those on discourse processing as well. Discourse processing is not the dominant sub-field of NLP, and as a result, papers on discourse are not ranked among the top 10 HITS authority list (see column H in Table 1). Hence we can expect the set of globally authoritative papers to be quite distinct from the set of papers most related to this specific paper on discourse. Consequently, given a ranking list computed for this paper, we can observe whether it is inclined towards (global) importance or (local) relatedness, by comparing the ranking with those of HITS (global importance) and co-citation (local relatedness). Another reason for choosing a discourse paper is that it is relatively easy to tell discourse papers from the rest, because terms unique to discourse are often included in the paper title. These terms are ‘discourse,’ ‘dialogue,’ ‘(discourse) structure analysis,’ and ‘(discourse) segmentation,’ for example.

Table 1 lists all the 23 papers that are ranked among the top 10 in at least one of the ranking lists shown in the left-hand side of the table. Bold figures indicate the top-10 ranks, with 1 representing the highest rank.

The seed paper is boldfaced, and the papers in the same field as the seed, namely discourse processing, are underlined. From the underlined titles and column CC (co-citation coupling), we see that except for one paper, the papers with nonzero co-citation scores indeed deal with discourse processing. The only exception is the most authoritative paper, the one on a standard NLP dataset (“Penn Treebank”). A ‘–’ in column CC indicates that the paper was not co-cited with the seed paper.

Instead of the diffusion factor parameter γ introduced in (2)–(7), we present its normalized version β defined by $\beta = \gamma\lambda$, where λ is the spectral radius of the co-citation matrix. Since the admissible range of γ is $0 \leq \gamma < 1/\lambda$, we have $0 \leq \beta < 1$. This normalized parameter β was also used in Sect. 4 to define the community-based von Neumann kernels.

⁸The data can be downloaded from <http://cl.naist.jp/~shimbo/citationdata.html>.

Table 1 Paper rankings relative to the paper ‘Empirical studies in discourse.’ HITS (H), co-citation (CC), and von Neumann kernels ($\beta = 0.01, 0.95, 0.9999$)

0.01	0.95	0.9999	H	CC	Title
2	1	1	1	2	Building a large annotated corpus of English: the Penn Treebank
12	3	2	2	–	A stochastic parts program and noun phrase parser for unrestricted text
8	2	3	3	–	Statistical decision-tree models for parsing
9	4	4	4	–	A new statistical parser based on bigram lexical dependencies
62	10	5	5	–	Unsupervised word sense disambiguation rivaling supervised methods
365	9	6	6	–	Word-sense disambiguation using statistical models of Roget’s categories trained
374	15	9	7	–	The mathematics of statistical machine translation
11	7	7	8	–	Three generative, lexicalized models for statistical parsing
21	11	8	9	–	Transformation-based error-driven learning and natural language processing
63	14	10	10	–	Integrating multiple knowledge sources to disambiguate word sense
3	5	25	50	2	<u>Attention, intentions, and the structure of discourse</u>
15	47	63	59	–	<u>Multi-paragraph segmentation of expository text</u>
4	8	41	76	2	<u>Assessing agreement on classification tasks: the kappa statistic</u>
19	59	77	77	–	<u>Lexical cohesion computed by thesaural relations as an indicator of the structure of text</u>
10	60	90	96	–	<u>Centering: a framework for modeling the local coherence of discourse</u>
14	92	114	115	–	<u>Combining multiple knowledge sources for discourse segmentation</u>
32	120	162	162	–	<u>Text segmentation based on similarity between words</u>
15	126	182	198	–	<u>A prosodic analysis of discourse segments in direction-giving monologue</u>
5	173	94	201	2	<u>The reliability of a dialogue structure coding scheme</u>
6	23	156	668	2	<u>Message Understanding Conference tests of discourse processing</u>
62	472	691	778	–	<u>Discourse segmentation by human and automated means</u>
1	6	95	896	1	<u>Empirical studies in discourse</u>
7	267	205	1304	2	<u>Effects of variable initiative . . . in human-computer spoken natural language dialogue</u>

Looking at the columns for the von Neumann kernels in Table 1, we see that at $\beta = 0.01$, the top-ranked paper is the seed paper itself, followed by six papers that are co-cited with the seed. The high correlation with co-citation counts suggests that at this small value of β , the kernel provides a relatedness measure. However, we already see a symptom of topic drift in the rankings assigned to papers not co-cited with the seed paper; even though there are many discourse papers in the graph, those ranked 8th and 9th are the globally authoritative

Table 2 K-min distance between the von Neumann kernels and HITS, averaged over the 2280 seed vertices in the largest connected components

β	0.1	0.5	0.9	0.99	0.999	0.9999	0.99999
Average K-min	87.3	86.3	72.0	26.4	5.5	1.1	0.0

papers with the HITS rankings of 3rd and 4th, respectively (‘Statistical decision-tree models for parsing’ and ‘A new statistical parser based on bigram lexical dependencies’). These are papers on parsing, which is the most popular sub-field of NLP.

As β is increased, the rankings are more inclined towards that of HITS. At $\beta = 0.9999$, the two top-10 lists are mostly identical, and accordingly, the seed paper drops its ranking to 95th.

The trend observed above is consistent with the role of β we argued in Sect. 3.3; as β tends to 1, the measure induced by the von Neumann kernel deviates away from the co-citation relatedness and approaches the HITS importance.

We can show that this trend is not specific to this seed paper. Table 2 lists the average minimizing Kendall (K-min) distance (Fagin et al. 2003) between the top-10 paper lists produced by HITS and the von Neumann kernel for all 2280 seed papers in the largest connected component of the co-citation graph.

Our use of the K-min distance as a measure of correlation between rankings follows White and Smyth (2003). K-min is a modification to Kendall’s Tau to be applicable to partial rankings, such as top- n ranking lists. As such, K-min is a practical measure more suitable for evaluating recommender systems than Kendall’s Tau, in that the highest ranked items are considered to be more relevant than those in lower ranked items. A small K-min distance means the two ranking lists are similar; it is equal to 0 if all top-10 items are identical, and takes the maximum value of 100 if there are no common items in the top-10 lists.

In Table 2, the average is taken only over 2280 vertices in the largest connected components of the co-citation graph. These are the vertices to which HITS assigns nonzero scores, and the scores for those in all other components are uniformly zero. Theorem 2 guarantees the ranking produced by von Neumann kernels for these vertices (with nonzero HITS scores) to approach that of HITS. Thus, by taking average only over these vertices, it is expected that the average K-min distance should vanish as $\beta \rightarrow 1$. As seen from Table 2, this is indeed the case, and the rankings induced by the von Neumann kernels are biased towards the HITS rankings as β is increased. It is worth noting that the K-min distance makes a drastic change towards 0 near the upper bound of β .

6.2 Community-based von Neumann kernels

We now focus on the community-based von Neumann kernels, the von Neumann kernels combined with PLSI-based community decomposition proposed in Sect. 4. In the experiments throughout Sect. 6.2, we set the hyperparameter k (the number of communities) of PLSI to 5.

6.2.1 Comparison of the rankings relative to “Empirical studies in discourse”

Table 3 lists the ranking list of the community-based kernels relative to the same seed paper (‘Empirical studies in discourse’) used in Sect. 6.1, to contrast the community-based von Neumann kernels (see (14)) with the plain von Neumann kernels. The table lists the same papers listed in Table 1, the ranking lists produced by the plain von Neumann kernels.

At $\beta = 0.01$, most of top-ranked papers are discourse papers as indicated by the underlined paper titles.

Even at $\beta = 0.95$, the top-ranked papers still retain the same topic as the seed paper, which is contrastive to the plain von Neumann kernel at $\beta = 0.95$ (see Table 1). The only non-discourse paper in the top-10 list is the most authoritative paper in the entire citation graph (according to HITS), ‘Building a large annotated corpus of English: the Penn Treebank.’ This is not a discourse paper, but its inclusion may still be understandable since it is also co-cited with the seed paper, as seen from column CC.

The trend of discourse paper being ranked higher than the rest continues to the rankings at $\beta = 0.9999$. Here, the top-10 list is occupied by discourse papers (although 9th and 10th papers are not shown in the table to make the comparison with Table 1 easier). In this list, even the Penn Treebank paper is moved out of the top-10.

The two globally important papers on parsing discussed in Sect. 6.1, namely, ‘Statistical decision-tree models for parsing’ and ‘A new statistical parser based on bigram lexical dependencies,’ never enter the top-10 rankings for the community-based von Neumann kernels.

6.2.2 Correlation with HITS

Table 4 shows the K-min distance (averaged over all seed vertices) between the top-10 lists induced by our community-based von Neumann kernels and the global HITS authority rankings. It also lists the average K-min distance between the ranking list of the kernels computed for individual seed vertex, and the HITS authority rankings computed for the *principal* community graph of the same seed vertex. Here, the principal community t of a seed vertex is determined by $\arg \max_{t=1}^5 p(t|d)$, i.e., the community t in which $p(t|d)$ is the highest for the vertex d . The probability $p(t|d)$ is computed with PLSI which we also used in Step 1 of the kernel computation in Sect. 4.2.

The results show that as β is increased, the rankings induced by the community-based von Neumann kernels approach the HITS authority rankings in the principal community of the seed vertices, even though this is not guaranteed; there may be vertices which belongs to multiple communities, and for these vertices, the rankings may be the mixture of multiple HITS rankings in these communities. On the other hand, the distance between the community-based von Neumann kernel and plain HITS does not decrease with increased β . This reflects the fact that the rankings for the seed vertices whose principal community is different from the globally most authoritative vertex do not necessarily converge to the global HITS authority rankings.

6.3 Paper recommendation system simulation

Our next experiment is a simulation of a paper recommendation system. The experiment is designed with the following scenario in mind: A student, who is interested in a research field he has no knowledge of, attempts to carry out a survey of the field. His only clue to the field is a few papers he happens to own, so he inputs these papers as the ‘seeds’ to the paper recommendation system, expecting to obtain a prioritized list of relevant papers in the field.

To simulate the above scenario, we design our experiment on the basis of several published survey papers and the references cited within them. Each of these survey papers is assumed to be citing many of the most relevant papers of the surveyed field. Hence if we select a few of these cited papers as the seed papers that our fictitious student has input to a recommendation system, the performance of the systems (which, in our case, is the ranking methods such as the von Neumann kernels) can be evaluated in terms of how many of the remaining papers cited by the survey paper are high-ranked by the respective systems.

Table 3 The paper rankings generated by the community-based von Neumann kernels for ‘Empirical studies in discourse.’ HITS (H), co-citation (CC), and community-based von Neumann kernels with $\beta = 0.01, 0.95, 0.9999$

0.01	0.95	0.9999	H	CC	Title
6	6	371	1	2	Building a large annotated corpus of English: the Penn Treebank
362	154	98	2	–	A stochastic parts program and noun phrase parser for unrestricted text
359	102	548	3	–	Statistical decision-tree models for parsing
349	103	441	4	–	A new statistical parser based on bigram lexical dependencies
419	108	43	5	–	Unsupervised word sense disambiguation rivaling supervised methods
346	104	40	6	–	Word-sense disambiguation using statistical models of Roget’s categories trained
99	18	223	7	–	The mathematics of statistical machine translation
367	153	589	8	–	Three generative, lexicalized models for statistical parsing
237	307	209	9	–	Transformation-based error-driven learning and natural language processing
429	123	587	10	–	Integrating multiple knowledge sources to disambiguate word sense
5	4	1	50	2	<u>Attention, intentions, and the structure of discourse</u>
84	8	2	59	–	<u>Multi-paragraph segmentation of expository text</u>
3	3	6	76	2	<u>Assessing agreement on classification tasks: the kappa statistic</u>
82	9	3	77	–	<u>Lexical cohesion computed by thesaural relations as an indicator of the structure of text</u>
8	7	20	96	–	<u>Centering: a framework for modeling the local coherence of discourse</u>
81	10	4	115	–	<u>Combining multiple knowledge sources for discourse segmentation</u>
106	13	8	162	–	<u>Text segmentation based on similarity between words</u>
409	11	5	198	–	<u>A prosodic analysis of discourse segments in direction-giving monologue</u>
2	2	242	201	2	<u>The reliability of a dialogue structure coding scheme</u>
7	15	518	668	2	<u>Message Understanding Conference tests of discourse processing</u>
471	51	7	778	–	<u>Discourse segmentation by human and automated means</u>
1	1	224	896	1	<u>Empirical studies in discourse</u>
4	5	136	1309	2	<u>Effects of variable initiative . . . in human-computer spoken natural language dialogue</u>

6.3.1 Method and evaluation metric

Again, we used the same citation graph consisting of 2867 vertices used earlier in Sect. 6.

Aside from the citation graph, we collected twelve survey papers, ten from the chapters of the book “*Handbook of Natural Language Processing*” (Dale et al. 2000), one tutorial

Table 4 Averages of K-min distances between the community-based von Neumann kernels, HITS, and HITS applied to the principal community graph of each seed vertex

β	0.01	0.1	0.5	0.9	0.95	0.99	0.999	0.9999	0.99999
HITS (entire graph)	93.7	92.9	92.0	85.4	82.4	78.2	78.2	75.7	75.1
HITS (principal community of seeds)	85.1	83.7	82.2	68.3	60.3	42.2	22.6	13.6	8.4

paper on robust parsing presented at the *COLING* conference, and one introductory paper on text summarization for the special issue of the *Computational Linguistics* journal. The surveyed fields of these papers range over various subtopics of NLP, from part-of-speech tagging, parsing, and word-sense disambiguation to text summarization, machine translation and discourse processing. None of these survey papers appear in the citation graph, and this poses a challenging task.

We repeated the following procedure for each of the 12 survey papers and each ranking method. Let $\mathcal{V}$ be the set of all vertices (papers) in the citation graph. Suppose that we are given the i th survey paper whose citation set is denoted by $\mathcal{C}_i (\subset \mathcal{V})$.

1. Generate all possible combination of a given size m ($m = 1, 2$, or 3) from $\mathcal{C}_i$, which is the papers cited by the given survey paper. Let the generated combinations be $\text{Comb}_m(\mathcal{C}_i)$. These will be used as the seed papers input to the system.
2. For each combination $S \in \text{Comb}_m(\mathcal{C}_i)$, input S to the ranking method. Hence S is the set of seed papers. Obtain the rankings of all remaining papers $\mathcal{V} - S$ in the citation graph.
3. At this point, $\bar{S} = \mathcal{C}_i - S$ is the set of papers that are referenced by the i th survey paper but not used as the seeds (i.e., not input to the ranking method).
A good recommendation system (ranking method) is supposed to rank papers in $\bar{S}$ higher than the rest, because these papers are assumed to have high relevance to the surveyed field (because they are cited by the survey paper).
4. Calculate the number of papers in $\bar{S} = \mathcal{C}_i - S$ that are included in the top- n rankings, for various n .

The evaluation metric is the micro-average recall score, where average is taken over all possible combination of seed papers for each of the 12 survey papers. For a ranking method R , let $\mathcal{R}_n(S)$ represent the set of papers in the top- n list output by the system $\mathcal{R}$ upon input seeds S . The score for this method $\mathcal{R}$, under a fixed seed set size m and restricted to top n list (after taking micro average) is thus

$$\text{recall}_m(\mathcal{R}, n) = \frac{\sum_{i=1}^{12} \sum_{S \in \text{Comb}_m(\mathcal{C}_i)} |\mathcal{R}_n(S) \cap (\mathcal{C}_i - S)|}{\sum_{i=1}^{12} |\text{Comb}_m(\mathcal{C}_i)| \cdot |\mathcal{C}_i - S|}. \quad (21)$$

Except co-citation coupling which may output less than n items if there are insufficient number of co-citations, all tested methods output equal number (n) of items, and the precision score is proportional to the recall score.

Note that the preliminary experimental results reported in (Shimbo et al. 2007) counted seed papers as well as non-seed papers as the correct output, if seed papers are output in a top n list. But since seed paper are those input to the system by the user, recommending these back to the user does not actually make sense. Hence, the evaluation in this paper is modified so that the top n papers are to be selected only from non-seed papers.

6.3.2 Compared algorithms

We computed various kernel matrices for the citation graph, and two baseline methods (HITS and co-citation rankings) for comparison. For kernel-based methods, if there is only one seed paper i , the rankings are computed from the i th row vector of the kernel matrix, in the same manner as Sects. 6.1 and 6.2. For the case of two or three seed papers, the rankings are based on the sum of the row vectors in the kernel matrix corresponding to the seeds.

The compared algorithms are summarized below, together with their abbreviations.

- HITS: HITS authoritative rankings based on the dominant eigenvector of $A^T A$. HITS is presented as the baseline of comparison. Notice that HITS, which does not have the concept of seed vertices, gives a single global ranking list irrespective of the seeds.
- CC: Rankings based on the number of co-citations with the seed paper(s). This is another baseline method.
- vNK/ k : the community-based von Neumann kernel N_β^{comm} with k communities introduced in Sect. 4.2. vNK/1 is Kandola et al.’s original von Neumann kernel. We tried the number of communities $k = 1, 5, 10, 15, 20$, and 30. These are regarded as different methods, to see the influence of the number of communities. The diffusion factor was chosen from $\beta = 0.001, 0.01, 0.05, 0.1, 0.3, 0.5, 0.7, 0.9, 0.95, 0.99$, and 0.999, so that the system achieved the best performance in terms of the average recall score over all combinations of seed size m and the size of the ranking list n reported in Table 5.
- EDK/ k : the exponential diffusion kernels $\bar{E}_\beta$ and their community-based variation E_β^{comm} introduced in Sect. 5.1; EDK/1 is the vanilla exponential diffusion kernels given by (17). For number of communities $k \geq 2$, the kernels are given by (18). We tried the following values of β : 0.001, 0.01, 0.05, 0.1, 0.3, 0.5, 0.7, 0.9, 0.95, 0.99, 0.999, 1.1, 2, 3, 4, 5, 10, 100, and 1000. The best β was selected in the same manner as vNK/ k , and the number of communities k tried were $k = 1, 5, 10, 15, 20, 30$.
- RL: the regularized Laplacian of (19). Similar to the above kernels, the diffusion factor was rescaled by $\beta = \gamma\lambda$, where λ is the spectral radius of the Laplacian of the co-citation graph. We tested the same values of β we used for the exponential kernels EDK/ k .
- MFA: Matrix-Forest-based Algorithm (Chebotarev and Shamis 1997) is defined as $(I + L)^{-1}$, and hence it is in fact a special case of the regularized Laplacian with $\gamma = 1$ in (19). The spectral radius of the Laplacian of the co-citation graph is 996.92, and thus MFA in this graph is equivalent to the regularized Laplacian with $\beta = 996.92$.
- CT: Commute-time kernel (Saerens et al. 2004), or the pseudoinverse of the Laplacian matrix. CT and MFA are known to perform well in a collaborative filtering task (Fouss et al. 2006), which is quite attractive since they have no parameters to tune.

6.3.3 Results

Table 5 summarizes the micro average recall given by (21), for various top- n ranking lists with $n = 10, 20, 30, 40$, and 50.

For each kernel method with diffusion parameter β , only the best results (in terms of macro-average recall of all the reported settings of m and n in the table) over tested parameter β are shown.⁹ The parenthesized figures right of the method names are the chosen value of β . See Sect. 6.3.2 for the abbreviation of the methods.

The following can be observed from the table.

⁹In practice, parameters must be chosen via cross-validation, but since this is extremely time-consuming in our setting, the results reported here are for the best parameter values.

Table 5 Average recall scores for various top- n lists and seed size m . Bold figures indicate the best scores in each (n, m) combination

			n				
			10	20	30	40	50
1 seed ($m = 1$)	Baselines	HITS	5.3	8.8	12.7	15.4	19.5
		CC	18.1	27.7	32.9	36.8	40.2
	von Neumann	vNK/1 (0.01)	20.8	31.6	38.3	43.9	49.2
		vNK/5 (0.01)	18.1	30.0	37.2	42.7	47.4
		vNK/10 (0.7)	19.4	31.6	38.7	44.5	49.5
		vNK/15 (0.9)	21.0	34.0	41.6	48.3	54.8
		vNK/20 (0.7)	20.8	32.5	38.0	42.8	47.7
		vNK/30 (0.7)	19.3	29.9	37.1	42.0	47.6
	Exponential	Exp/1 (0.05)	20.8	31.6	38.3	43.9	49.2
		Exp/5 (0.95)	17.5	28.3	35.5	41.3	46.3
		Exp/10 (3.0)	18.9	30.7	38.5	44.3	49.3
		Exp/15 (4.0)	20.2	32.6	40.1	47.2	53.4
		Exp/20 (3.0)	19.5	30.2	37.5	42.5	47.1
		Exp/30 (5.0)	18.0	27.2	34.2	39.6	44.8
	Laplacian	RL (2.0)	16.1	27.5	35.1	41.3	48.1
		MFA	4.7	12.5	20.5	28.2	34.3
		CT	4.6	12.1	19.2	26.6	31.9
2 seeds ($m = 2$)	Baselines	HITS	5.0	9.8	14.2	17.0	21.5
		CC	23.8	35.9	42.8	48.5	53.1
	von Neumann	vNK/1 (0.01)	24.6	36.9	44.6	51.7	58.0
		vNK/5 (0.01)	23.3	35.8	44.7	51.2	56.3
		vNK/10 (0.7)	23.1	36.2	43.6	52.1	59.2
		vNK/15 (0.9)	24.0	38.3	46.6	57.1	65.9
		vNK/20 (0.7)	25.4	38.3	46.0	52.2	58.0
		vNK/30 (0.7)	23.2	36.6	45.3	52.4	58.2
	Exponential	Exp/1 (0.05)	24.6	36.9	44.6	51.7	58.0
		Exp/5 (0.95)	21.0	32.7	40.3	47.3	53.1
		Exp/10 (3.0)	22.8	35.6	43.1	51.6	58.4
		Exp/15 (4.0)	23.3	36.5	44.1	53.8	61.1
		Exp/20 (3.0)	22.8	35.5	43.4	50.1	55.9
		Exp/30 (5.0)	21.3	33.0	40.4	46.5	51.5
	Laplacian	RL (2.0)	23.3	36.3	44.8	51.9	57.9
		MFA	6.1	15.5	24.5	32.7	40.4
		CT	5.6	13.8	22.3	30.6	37.7
3 seeds ($m = 3$)	Baselines	HITS	4.8	10.8	15.9	18.9	23.7
		CC	26.9	39.9	48.1	54.9	60.4
	von Neumann	vNK/1 (0.01)	26.7	39.8	48.1	56.3	63.6
		vNK/5 (0.01)	26.2	39.4	48.4	56.0	62.1
		vNK/10 (0.7)	25.3	38.5	46.2	56.9	65.7
		vNK/15 (0.9)	25.7	40.0	49.0	62.3	71.9

Table 5 (Continued)

		<i>n</i>				
		10	20	30	40	50
Exponential	vNK/20 (0.7)	28.0	41.3	49.7	57.0	63.7
	vNK/30 (0.7)	25.2	39.5	49.8	59.2	65.3
	Exp/1 (0.05)	26.7	39.8	48.1	56.3	63.6
	Exp/5 (0.95)	23.8	35.9	43.3	50.9	57.4
	Exp/10 (3.0)	25.5	38.5	45.1	54.6	62.1
	Exp/15 (4.0)	25.3	38.5	46.2	56.3	64.3
	Exp/20 (3.0)	25.1	38.7	46.7	54.1	60.4
	Exp/30 (5.0)	23.8	36.5	43.9	50.0	55.1
Laplacian	RL (2.0)	27.2	40.7	50.0	57.5	63.8
	MFA	6.8	16.2	25.9	34.6	42.5
	CT	6.1	14.1	23.3	31.6	39.2

- When no community decomposition was performed, i.e., $k = 1$, kernels based on the adjacency matrix (vNK/1 and EDK/1) performed better than the regularized Laplacian (RL) when $m = 1$, but they are about even when $m = 2$ and RL was better when $m = 3$. In fact, with the seed size $m = 3$, even the co-citation ranking (CC) performed pretty well.
- The best recall scores were obtained by the community-based von Neumann kernel vNK/ k except one situation with $m = 3$ and $n = 30$ in which the regularized Laplacian (RL) was the best. In most cases, the optimal number of communities for vNK/ k was $k = 15$.
- The exponential diffusion kernels EDK/ k showed performance close to that of the von Neumann kernels vNK/ k . Their performance was equal when no community decomposition was performed (i.e., EDK/1 and vNK/1). For exponential diffusion kernels, however, community decomposition was not as effective as with the von Neumann kernels.
- Even co-citation performed moderately well, especially when $m = 3$.
- Non-parametric Laplacian-based kernels, namely, the commute-time kernels (CT) and MFA, did not perform well in this experiment.

It seems that the poor performance of MFA in this experiment can be attributed to the large parameter value $\beta = 996.92$ when it is viewed as an instance of the regularized Laplacian. We observed that the commute-time kernel also output rankings similar to those by MFA.

On the other hand, these kernels, MFA and the commute-time kernel, are known to perform quite well in many collaborative filtering tasks (Fouss et al. 2006). What characterizes this discrepancy, as well as which of the adjacency-matrix-based kernels or the Laplacian-based kernels should be chosen for a given task, is an interesting future research question.

One conceivable factor is the presence of bias towards importance rankings in adjacency-matrix based kernels. This paper recommendation task requires ranking methods which choose important papers (in each individual research field), so this may be the reason adjacency-matrix based kernels worked better than the Laplacian-based kernels in this task. Indeed, Fouss et al. (2007) reported negative correlation between the commute-time kernels and the best-seller (or, importance) rankings in their experiments.

Finally, Table 6 summarizes the results of the Wilcoxon signed rank test (Siegel and Castellan 1988) (of macro recall scores over 12 survey papers), with the null hypothesis H_0 :

Table 6 p values of the Wilcoxon sign ranks test for recall scores over 12 survey papers for vNK/15 (single tail; in favor of vNK/15). The results are for $m = 1$ (one seed). Only values $p < 0.05$ are shown

Top n	HITS	CC	vNK	Exp	Exp/15	RL	MFA	CT
10	0.00342	–	–	–	–	0.00684	0.00049	0.00049
20	0.00049	–	–	–	–	0.03857	0.00122	0.00073
30	0.00098	0.02100	–	–	–	–	0.00049	0.00049
40	0.00098	0.01611	–	–	–	–	0.00122	0.00098
50	0.00244	0.02686	–	–	–	–	0.00488	0.00244

there is no performance difference between vNK/15 (0.9) (i.e., community-based von Neumann kernel with $\beta = 0.9$) and the ranking methods listed above the table; alternate hypothesis H_1 : vNK/15 outperforms the methods above the table. Only values with significance level $p < 0.05$ are shown. The table is for the seed set size $m = 1$ only, since for $m = 2$ and $m = 3$, seeds are overlapping. For more results on the significance between all the methods involved not just vNK/15, see Appendix B.

The results of Table 6 show that vNK/15 significantly outperformed HITS, MFA, and CT on all tested n , CC (co-citation) on all but $n = 10$, and RL (regularized Laplacian) on $n = 10$ and 20.

The test did not show the statistical significance of vNK/15 (community-based von Neumann kernels) over vNK/1 (plain von Neumann kernels), even though the obtained micro-average recalls were consistently higher for vNK/15 when the seed size $m = 1$. We also conducted a paired sign test, by interpreting each of the results for all possible seed-answer pair (S, c) as a trial, where S is a singleton set of seed (since the test is again for seed size $m = 1$), and $c \in \mathcal{C} - S$ is one of the non-seed papers cited in the survey paper. Thus we counted a + if vNK/15 correctly included c in its top- n list but vNK/1 did not, and a – if vNK/1 listed c in its top- n list but not vNK/15. The sign test showed the results are statistically significant in favor of vNK/15 for all tested ns at $p < 0.01$ level.

7 Conclusions

In this paper, we have analyzed the characteristics of the von Neumann kernels.

The major contributions of this paper are as follows. (i) We have clarified the relationship between von Neumann kernels and HITS. (ii) We have pointed out that the von Neumann kernels are prone to topic drift, and the community structure of the graph must be taken into account prior to the application of the kernels. To this end, we have proposed to build a ‘community graph’ corresponding to each community in the graph, using a PLSI-based graph decomposition. This decomposition borrows idea from Cohn and Chang’s PHITS method (Cohn and Chang 2000), and does not require the contents of documents be available. The resulting community graphs can be effectively combined with von Neumann kernels, to obtain various ranking methods respecting the community structure underlying the graph. (iii) We have presented experimental results demonstrating the effectiveness of von Neumann kernels in a simulated recommendation task.

Although most of the previous work on kernel-based link analysis use Laplacian-based kernels (Fouss et al. 2007; Saerens et al. 2004; Smola and Kondor 2003; Zhou and Schölkopf 2004), the unified framework of importance and relatedness provided by von Neumann kernels is also attractive for many link analysis applications.

In particular, we believe that von Neumann kernels have a great potential for recommendation systems. Consider the paper recommendation task we simulated in Sect. 6.3. In this scenario, the degree of the user's acquaintance with the research field affects how much the system should bias its decision towards authoritative papers in the field. A researcher who is an expert in the field would not be happy if the system recommended the most authoritative paper in the field, since he is likely to have already known the paper. Thus, in this case, it would be reasonable to bias the ranking method towards relatedness measures. On the other hand, recommending the most well-known papers should be acceptable if the user is not familiar with the field. Such adjustment can be easily done with the single parameter of von Neumann kernels.

In future work, we will apply the community-based von Neumann kernels to other data sets. A more challenging problem is the automatic determination of the number of communities in community-based von Neumann and exponential kernels. A simple approach is to use cross-validation. In the clustering literature, eigengaps of the Laplacian matrix have been used to deduce appropriate number of clusters (Zelnik-Manor and Perona 2005). We plan to check if similar approach is effective for PLSI-based graph decomposition we used for the community-based von Neumann kernels.

Automatic tuning of a suitable diffusion factor is another issue in practical application of parametric kernel-based ranking methods. In this respect, Acharyya and Ghosh (2003) pursued an automatic parameter tuning method in their parametric vertex importance measure similar to the von Neumann kernels. Their optimality criterion was the stability of the parameters, but since it is only one of many criteria for optimality, we believe there is much room for further research in this area. Cristianini et al.'s work on kernel alignment (Cristianini et al. 2002; Kandola et al. 2002) is a more goal-directed approach that optimizes the parameter toward an objective matrix. It is applicable to von Neumann kernels in the presence of such a matrix and looks promising.

Acknowledgements We are grateful to Eric Nichols for proofreading, and to Taku Kudo for discussion in the early stage of this work. We would also like to thank three anonymous reviewers for detailed and helpful comments that greatly improved the paper.

Appendix A: Proofs

This appendix provides the proofs of the theorems that appeared in Sect. 3.3. We also present a theorem needed for the discussion of Sect. 5.2.

Theorem 4 (Theorem 1) *Let $\lambda > 0$ be the dominant eigenvalue of a symmetric positive semidefinite matrix $A^T A$. If λ is a simple eigenvalue (i.e., λ has a multiplicity of one), there exists a unit eigenvector $\mathbf{v}$ corresponding to λ such that $(A^T A / \lambda)^n \rightarrow \mathbf{v}\mathbf{v}^T$ as $n \rightarrow \infty$.*

Proof Let $A^T A = \sum_{i=1}^m \lambda_i \mathbf{v}_i \mathbf{v}_i^T$ be a spectral projection of $A^T A$, with $\lambda = \lambda_1 > 0$ and $\mathbf{v} = \mathbf{v}_1$ being the corresponding eigenvector. By raising both sides to the power of n , we have for each $n = 1, 2, \dots$,

$$(A^T A)^n = \sum_{i=1}^m \lambda_i^n \mathbf{v}_i \mathbf{v}_i^T,$$

since $\mathbf{v}_i^T \mathbf{v}_i = 1$ for every i , and $\mathbf{v}_i^T \mathbf{v}_j = 0$ for every $i \neq j$. Dividing both sides of the equation by λ_1^n yields

$$\left(\frac{A^T A}{\lambda_1}\right)^n = \mathbf{v}_1 \mathbf{v}_1^T + \sum_{i=2}^m \left(\frac{\lambda_i}{\lambda_1}\right)^n \mathbf{v}_i \mathbf{v}_i^T.$$

Because $\lambda = \lambda_1 > \lambda_i$ for every $i = 2, \dots, m$, the second term on the right-hand side vanishes as $n \rightarrow \infty$. $\square$

Theorem 5 (Theorem 2) *Let $\lambda > 0$ be the dominant eigenvalue of a symmetric positive semidefinite matrix $A^T A$. If λ is a simple eigenvalue, there exists a unit eigenvector $\mathbf{v}$ corresponding to λ such that*

$$\left(\frac{1}{\lambda} - \gamma\right) N_\gamma(A^T A) \rightarrow \mathbf{v} \mathbf{v}^T \quad \text{as } \gamma \rightarrow \frac{1}{\lambda},$$

where the limit is taken from below.

Proof Let $A^T A = \sum_{i=1}^m \lambda_i \mathbf{v}_i \mathbf{v}_i^T$ be a spectral projection of $A^T A$, with the dominant eigenpair $\lambda = \lambda_1$ and $\mathbf{v} = \mathbf{v}_1$. Suppose $\gamma < 1/\lambda_1$. By the infinite series representation of the von Neumann kernels (see (6)), we have

$$N_\gamma(A^T A) = \sum_{i=1}^m \left(\sum_{n=1}^{\infty} \gamma^{n-1} \lambda_i^n \right) \mathbf{v}_i \mathbf{v}_i^T = \sum_{i=1}^m \frac{\lambda_i}{1 - \gamma \lambda_i} \mathbf{v}_i \mathbf{v}_i^T.$$

Multiplying both sides by $(1/\lambda_1 - \gamma)$ and rearranging terms yields

$$\left(\frac{1}{\lambda_1} - \gamma\right) N_\gamma(A^T A) = \mathbf{v}_1 \mathbf{v}_1^T + \sum_{i=2}^m \frac{\lambda_i (1 - \gamma \lambda_1)}{\lambda_1 (1 - \gamma \lambda_i)} \mathbf{v}_i \mathbf{v}_i^T,$$

and the second term on the right-hand side vanishes as $\gamma \rightarrow (1/\lambda_1)$ from below. $\square$

The following theorem is for the discussion of Sect. 5.2.

Theorem 6 *Consider an undirected connected graph with positive edge weights, and let B be its adjacency matrix. Then, the regularized Laplacian $R_\gamma(B)$ given by (19) tends to a uniform matrix as $\gamma \rightarrow \infty$.*

Proof Let m be the number of vertices in the graph, and let $\{(\lambda_i, \mathbf{v}_i)\}_{i=1}^m$ be the eigenpairs of $L(B)$. As we have assumed the graph represented by B is connected, first note the well-known fact that the Laplacian matrix has an eigenpair $(\lambda_1, \mathbf{v}_1) = (0, \mathbf{1}/\sqrt{m})$, and all other eigenvalues λ_i ($i \neq 1$) are nonzero. Further let $L(B) = P \Lambda P^{-1}$ be the spectral decomposition of $L(B)$, with $P = [\mathbf{v}_1, \dots, \mathbf{v}_m]$ an orthogonal matrix with each $\mathbf{v}_i$ representing a unit eigenvector corresponding to λ_i , and Λ a diagonal matrix with eigenvalues λ_i as the diagonal element $[\Lambda]_{i,i}$. It follows that $I + \gamma L(B) = P I P^{-1} + \gamma P \Lambda P^{-1} = P (I + \gamma \Lambda) P^{-1}$ and hence

$$R_\gamma(B) = (I + \gamma L(B))^{-1} = P (I + \gamma \Lambda)^{-1} P^{-1}.$$

Here, notice that $(I + \gamma \Lambda)^{-1}$ is a diagonal matrix with diagonal elements $[(I + \gamma \Lambda)^{-1}]_{i,i} = 1/(1 + \gamma \lambda_i)$. In particular, since $\lambda_1 = 0$, we have $[(I + \gamma \Lambda)^{-1}]_{1,1} = 1$, a constant. Taking the limit $\gamma \rightarrow \infty$ yields

$$\lim_{\gamma \rightarrow \infty} [(I + \gamma \Lambda)^{-1}]_{i,i} = \lim_{\gamma \rightarrow \infty} \frac{1}{1 + \gamma \lambda_i} = \begin{cases} 1 & \text{if } i = 1, \\ 0 & \text{otherwise.} \end{cases}$$

Plugging this back to $R_\gamma(B) = P(I + \gamma \Lambda)^{-1}P^{-1}$ and recalling that $P = [\mathbf{v}_1, \dots, \mathbf{v}_m]$, we have $\lim_{\gamma \rightarrow \infty} R_\gamma(B) = \mathbf{v}_1 \mathbf{v}_1^T = \mathbf{1}\mathbf{1}^T/m$. $\square$

Appendix B: Full results of Wilcoxon sign ranks test

Table 7 shows the full results of the Wilcoxon sign rank test discussed in Sect. 6.3.3.

Table 7 Full results of the Wilcoxon sign ranks test for recall scores on top- n lists (p values; single tail; in favor of the method on the left over the one at the top). Only values $p < 0.05$ are shown

(a) Top-10 lists

	HITS	CC	vNK/1	vNK/15	EDK/1	EDK/15	RL	MFA	CT
HITS	–	–	–	–	–	–	–	–	–
CC	0.00610	–	–	–	–	–	0.00049	0.00024	0.00049
vNK/1	0.00122	0.00928	–	–	–	–	0.00122	0.00024	0.00024
vNK/15	0.00342	–	–	–	–	–	0.00684	0.00049	0.00049
EDK/1	0.00122	0.00928	–	–	–	–	0.00122	0.00024	0.00024
EDK/15	0.00342	–	–	–	–	–	0.00977	0.00049	0.00049
RL	–	–	–	–	–	–	–	0.00049	0.00049
MFA	–	–	–	–	–	–	–	–	–
CT	–	–	–	–	–	–	–	–	–

(b) Top-20 lists

	HITS	CC	vNK/1	vNK/15	EDK/1	EDK/15	RL	MFA	CT
HITS	–	–	–	–	–	–	–	–	–
CC	0.00049	–	–	–	–	–	–	0.00049	0.00049
vNK/1	0.00049	0.00342	–	–	–	–	0.00293	0.00049	0.00049
vNK/15	0.00049	–	–	–	–	–	0.03857	0.00122	0.00073
EDK/1	0.00049	0.00342	–	–	–	–	0.00293	0.00049	0.00049
EDK/15	0.00098	–	–	–	–	–	0.03223	0.00073	0.00073
RL	0.00073	–	–	–	–	–	–	0.00049	0.00049
MFA	–	–	–	–	–	–	–	–	–
CT	–	–	–	–	–	–	–	–	–

Table 7 (Continued)

(c) Top-30 lists

	HITS	CC	vNK/1	vNK/15	EDK/1	EDK/15	RL	MFA	CT
HITS	–	–	–	–	–	–	–	–	–
CC	0.00073	–	–	–	–	–	–	0.00684	0.00488
vNK/1	0.00073	0.00098	–	–	–	–	0.02734	0.00098	0.00098
vNK/15	0.00098	0.02100	–	–	–	–	–	0.00049	0.00049
EDK/1	0.00073	0.00098	–	–	–	–	0.02441	0.00098	0.00098
EDK/15	0.00122	0.01050	–	–	–	–	–	0.00049	0.00049
RL	0.00073	–	–	–	–	–	–	0.00049	0.00049
MFA	–	–	–	–	–	–	–	–	0.01953
CT	–	–	–	–	–	–	–	–	–

(d) Top-40 lists

	HITS	CC	vNK/1	vNK/15	EDK/1	EDK/15	RL	MFA	CT
HITS	–	–	–	–	–	–	–	–	–
CC	0.00122	–	–	–	–	–	–	0.03857	0.02124
vNK/1	0.00049	0.00049	–	–	–	–	0.03857	0.00464	0.00464
vNK/15	0.00098	0.01611	–	–	–	–	–	0.00122	0.00098
EDK/1	0.00049	0.00049	–	–	–	–	0.03857	0.00464	0.00464
EDK/15	0.00073	0.00488	–	–	–	–	–	0.00073	0.00049
RL	0.00049	0.00928	–	–	–	–	–	0.00244	0.00244
MFA	–	–	–	–	–	–	–	–	0.00586
CT	–	–	–	–	–	–	–	–	–

(e) Top-50 lists

	HITS	CC	vNK/1	vNK/15	EDK/1	EDK/15	RL	MFA	CT
HITS	–	–	–	–	–	–	–	–	–
CC	0.00171	–	–	–	–	–	–	–	0.03857
vNK/1	0.00024	0.00049	–	–	–	–	–	0.00684	0.00244
vNK/15	0.00244	0.02686	–	–	–	–	–	0.00488	0.00244
EDK/1	0.00024	0.00024	–	–	–	–	–	0.00684	0.00244
EDK/15	0.00122	0.00293	–	–	–	–	–	0.00244	0.00171
RL	0.00049	0.00024	–	–	–	–	–	0.00098	0.00098
MFA	–	–	–	–	–	–	–	–	0.00195
CT	–	–	–	–	–	–	–	–	–

References

- Acharyya, S., & Ghosh, J. (2003). A maximum entropy framework for higher order link analysis on discrete graphs. In *Workshop on link analysis for detecting complex behavior (LinkKDD 2003)*.
- Baldi, P., Frascioni, P., & Smyth, P. (2003). *Modeling the Internet and the web: probabilistic methods and algorithms*. New York: Wiley.
- Bharat, K., & Henzinger, M. R. (1998). Improved algorithms for topic distillation in a hyperlinked environment. In *Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval*, Melbourne, Australia.

- Bollacker, K. D., Lawrence, S., & Giles, C. L. (1998). CiteSeer: an autonomous web agent for automatic retrieval and identification of interesting publications. In *Proceedings of the 2nd international ACM conference on autonomous agents* (pp. 116–123).
- Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual (web) search engine. *Computer Network and ISDN Systems*, 30(1–7), 107–117.
- Chebotarev, P. Y., & Shamis, E. V. (1997). The matrix-foreset theorem and measuring relations in small social groups. *Automation and Remote Control*, 58(9), 1505–1514.
- Chung, F. R. K. (1997). *Spectral graph theory*. Providence: American Mathematical Society.
- Cohn, D., & Chang, H. (2000). Learning to probabilistically identify authoritative documents. In *Proceedings of the 17th international conference on machine learning (ICML'00)* (pp. 167–174), Stanford, CA, USA.
- Cristianini, N., Kandola, J., & Elissee, A. (2002). On kernel-target alignment. In *Advances in neural information processing systems* (Vol. 14, pp. 367–373). Cambridge: MIT Press.
- Dale, R., Moisl, H., & Somers, H. (Eds.). (2000). *Handbook of natural language processing*. New York: Marcel Dekker.
- Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., & Harshman, R. A. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6), 391–407.
- Dhyani, D., Ng, W. K., & Bhowmick, S. S. (2002). A survey of web metrics. *ACM Computing Surveys*, 34(4), 469–503.
- Fagin, R., Kumar, R., & Sivakumar, D. (2003). Comparing top k lists. *SIAM Journal on Discrete Mathematics*, 17(1), 134–160.
- Fouss, F., Pirotte, A., Renders, J. M., & Saerens, M. (2007). Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 19(3), 355–369.
- Fouss, F., Yen, L., Dupont, P., & Saerens, M. (2006). An experimental investigation of graph kernels on a collaborative recommendation task. In *Proceedings of the 2006 IEEE international conference on data mining (ICDM'06)* (pp. 863–868), Hong Kong.
- Haussler, D. (1999). *Convolution kernels on discrete structures* (Technical Report UCSC-CRL-99-10). University of California at Santa Cruz.
- Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd ACM SIGIR conference on research and development in information retrieval* (pp. 50–57), Berkeley, CA, USA.
- Hofmann, T. (2000). Learning the similarity of documents: An information-geometric approach to document retrieval and categorization. In *Advances in neural information processing systems* (Vol. 12, pp. 914–920). Cambridge: MIT Press.
- Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1/2), 177–196.
- Ito, T., Shimbo, M., Kudo, T., & Matsumoto, Y. (2005). Application of kernels to link analysis. In *Proceedings of the 11th ACM SIGKDD international conference on knowledge discovery and data mining (KDD'05)* (pp. 586–592), Chicago, IL, USA.
- Ito, T., Shimbo, M., Mochihashi, D., & Matsumoto, Y. (2006). Exploring communities with kernel-based citation analysis. In *Proceedings of the 10th European conference on principles and practice of knowledge discovery in databases (PKDD'06)* (pp. 235–246). Berlin: Springer.
- Jaakkola, T., & Haussler, D. (1999). Exploiting generative models in discriminative classifiers. In *Advances in neural information processing systems* (Vol. 11, pp. 487–493). Cambridge: MIT Press.
- Kandola, J., Shawe-Taylor, J., & Cristianini, N. (2002). *Optimizing kernel alignment over combinations of kernels* (Technical Report NC-TR-2002-121). NeuroCOLT.
- Kandola, J., Shawe-Taylor, J., & Cristianini, N. (2003). Learning semantic similarity. In *Advances in neural information processing systems* (Vol. 15, pp. 673–680). Cambridge: MIT Press.
- Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation*, 14(1), 10–25.
- Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46, 604–632.
- Kondor, R., & Lafferty, J. (2001). Diffusion kernels on graphs and other discrete input spaces. In *Proceedings of the 18th international conference on machine learning (ICML'01)* (pp. 21–24), Williamstown, MA, USA.
- le Pair, C. (1988). The citation gap of applicable science. In A. F. J. van Raan (Ed.), *Handbook of quantitative studies of science and technology* (pp. 537–553). Amsterdam: Elsevier.
- Lempel, R., & Moran, S. (2001). The stochastic approach for link-structure analysis. *ACM Transactions on Information Systems*, 19(2), 131–160.
- Nadler, B., Lafon, S., Coifman, R., & Kevrekidis, I. (2006). Diffusion maps, spectral clustering and eigenfunctions of Fokker-Planck operators. In *Advances in neural information processing systems* (Vol. 18, pp. 955–962). Cambridge: MIT Press.

- Saerens, M., Fouss, F., Yen, L., & Dupont, P. (2004). The principal component analysis of a graph, and its relationship to spectral clustering. In *Lecture notes in artificial intelligence: Vol. 3202. Proceedings of 15th European conference on machine learning (ECML'04)* (pp. 371–383), Pisa, 2004. Berlin: Springer.
- Shawe-Taylor, J., & Cristianini, N. (2004). *Kernel methods for pattern analysis*. Cambridge: Cambridge University Press.
- Shimbo, M., & Ito, T. (2006). Kernels as link analysis measures. In D. J. Cook & L. B. Holder (Eds.), *Mining graph data* (Chap. 12). New York: Wiley.
- Shimbo, M., Ito, T., & Matsumoto, Y. (2007). Evaluation of kernel-based link analysis measures on research paper recommendation. In *Proceedings of the ACM/IEEE joint conference on digital libraries (JCDL 2007)* (pp. 354–355), Vancouver, British Columbia, Canada, 2007.
- Siegel, S., & Castellan, J. N. (1988). *Nonparametric statistics for the behavioral sciences* (2nd ed.). Boston: McGraw Hill.
- Small, H. (1973). Co-citation in the scientific literature: a new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 24, 265–269.
- Smola, A. J., & Kondor, R. (2003). Kernels and regularization of graphs. In *Lecture notes in artificial intelligence: Vol. 2777. Proceedings of the 16th annual conference on learning theory and 7th kernel workshop (COLT/Kernel'03)* (pp. 144–158). Berlin: Springer.
- White, S., & Smyth, P. (2003). Algorithms for estimating relative importance in networks. In *Proceedings of the 9th ACM SIGKDD international conference on knowledge discovery and data mining (KDD'03)* (pp. 266–275), Washington, DC, USA.
- Zelnik-Manor, L., & Perona, P. (2005). Self-tuning spectral clustering. In *Advances in neural information processing systems* (Vol. 17). Cambridge: MIT Press.
- Zhou, D., & Schölkopf, B. (2004). A regularization framework for learning from graph data. In *Proceedings of the workshop on statistical relational learning*.