PageRank as a discrete Green's function

Fan Chung [*]

Abstract

Originally, PageRank was a way to assign quantitative ranking to webpages by Web search engines. In general, PageRank can be viewed as a measure of relative "importance" defined on any given graph. In this paper, we establish the relationship of PageRank with a family of discrete Green's functions. Through this connection, we examine the relationship of the Laplacian, eigenvalues, Cheeger's constants and PageRank. In particular, we investigate the conditions under which the PageRank of an induced subgraph can be used to approximate the PageRank of the whole graph. We also examine the hitting time of a modified random walk and its connection with PageRank via Green's functions.

1 Introduction

PageRank is one of the main methods for determining the ranking of webpages by Web search engines. In 1998, Brin and Page [5] introduced the notion of PageRank as a quantitative ranking of Webpages which entirely relies on the linking structure of the so-called WWW graph, which has all Webpages as the vertices and hyperlinks as the edges. The notion of PageRank is well defined for any given graph and is very useful for examining how similar or far apart the vertices are. We note that examples of graphs derived from real-world networks often exhibit the so-called "small world phenomenon". Namely, any two vertices are joined by a relatively short path. Therefore the usual notion of distance is no longer very useful. Instead, PageRank provides a quantitative measure of relative "importance", for which we will give a detailed definition in the next section. Roughly speaking, PageRank is based on random walks together with a scalar which controls the rate of diffusion. As a result, there are very efficient and robust algorithms for computing and approximating PageRank.

In addition to being a useful tool for Web search, PageRank serves as a valuable method for examining the correlations of pairs and subsets of the vertices in the graph. In [1, 2], PageRank vectors lead to efficient partitioning

[*] University of California, San Diego

algorithms for finding local cuts. Various isoperimetric inequalities for graphs can be derived using PageRank [7]. PageRank is particularly useful since there are very efficient algorithms to compute and to approximate PageRank vectors [1, 4, 9].

In this paper, we wish to examine PageRank from a new perspective as a discrete Green's function. The concept of a Green's function was introduced in a famous essay of George Green in 1728. Since then, Green's function has had profound impact in numerous areas. There are many formulations of Green's function over various topics, ranging from basic functions for solving differential equations with boundary conditions to various types of correlation functions.

In 1999, Yau and the author introduced a discrete Green's function which is defined on graphs. In [8], the Green's function is closely associated with the normalized Laplacian and is useful for solving discrete Laplace equations with boundary condition. Several explicit formulas for Green's functions are given there.

In this paper, we consider a family of Green's functions which correspond to PageRank of a graph. Through this connection, we examine the relationship of subgraphs of a graph with the graph itself. In particular, we focus on the problem of determining the conditions under which the PageRank of an induced subgraph of a graph is a good approximation for that of the whole graph. We will use the Green's functions to derive the required conditions and to answer the above question.

In large information networks which are dynamically changing, we often can only have partial information by observing various subgraphs of the large underlying graph. The above problem can be viewed as an attentive effort for understanding correlations between a graph and its subgraph. Many problems along this line concerning complex networks remain challengingly difficult.

This paper is organized as follows. In Section 2, we define random walks, Laplacian eigenvalues and their implications through Matrix-Tree Theorem. In Section 3, we consider Dirichlet egienvalues, a modified version of Matrix-Tree Theorem and two versions of Cheeger's inequality. In Section 4, we examine the connections between PageRank and Green's function. In Section 5, we discuss the relationship of the Cheeger constant and the PageRank. In Section 6, we investigate the relationship of the PageRank of a graph and the PageRank of its subgraphs. In Section 7, we consider the hitting time in a graph and its relation to the PageRank.

2 Preliminaries

For a graph $G = (V, E)$ with vertex set V and edge set E, a typical random walk on G is defined by the following *transition probability matrix*:

$$P(u, v) \quad = \quad \begin{cases} \frac{1}{d_u} & \text{if } \{u, v\} \in E, \\ 0 & \text{otherwise.} \end{cases}$$

where d_u denotes the degree of vertex u.

For a function $f : V \to \mathbb{R}$, the (discrete) *Laplace* operator Δ is defined by:

$$\Delta f(x) \quad = \quad \frac{1}{d_x} \sum_{y \sim x} (f(x) - f(y)).$$

In other words, we can write $\Delta = I - P$. A symmetrized version of Δ is the Laplacian $\mathcal{L}$ which can be written as:

$$\begin{aligned} \mathcal{L} &= D^{1/2} \Delta D^{-1/2} \\ &= D^{-1/2}(D - A) D^{-1/2} \end{aligned}$$

where D denotes the diagonal matrix with entries $D(x, x) = d_x$ and A denotes the adjacency matrix of G.

Another way to interpret the Laplacian $\mathcal{L}$ is to regard it as a "square":

$$\mathcal{L} = B^* B \tag{1}$$

where B is the matrix whose columns are indexed by the vertices of G, whose rows are indexed by the edges of G and B^* denotes the transpose of B. To define B, we first turn each edge into a directed edge by choosing the direction in any arbitrary way. For a (directed) edge e and a vertex v, we define

$$B(e, v) \quad = \quad \begin{cases} \frac{1}{\sqrt{d_v}} & \text{if } v \text{ is the head of } e, \\ -\frac{1}{\sqrt{d_v}} & \text{if } v \text{ is the tail of } e, \\ 0 & \text{otherwise.} \end{cases}$$

Using the terminology in homology, we can regard B as a *boundary operator* mapping "1-chains" in C_1, defined on edges of a graph, to "0-chains" in C_0, defined on vertices of G. Then B^* is the corresponding *coboundary operator*. This can then be illustrated as follows:

$$\begin{tikzcd} C_1 \arrow[r, shift left, "B"] & C_0 \arrow[l, shift left, "B^*"] \end{tikzcd}$$

Δ and $\mathcal{L}$ have the same eigenvalues, denoted by

$$0 = \lambda_0 \leq \lambda_1 \leq \cdots \leq \lambda_{n-1}.$$

The eigenvalues capture many properties of the graph. For example, the multiplicity of eigenvalue 0 is equal to the number of connected components in G.

If G is connected, then λ_1 is nonzero and is called the spectral gap. In this paper, we assume the graph G is connected. If G is bipartite, then $\lambda_{n-1} = 2$. The reader is referred to [6] for many useful properties related to eigenvalues of the Laplacian.

A classical result in graph theory is the *Matrix-Tree Theorem*, due to Kirchhoff [10], that relates eigenvalues to the number of spanning trees in G. Note that a spanning tree is a subgraph that contains no cycles and the enumeration of spanning trees is one of the main tools in many computational problems.

Here we give a short proof for a modified version of the Matrix-Tree Theorem:

Theorem 1 (Matrix-Tree Theorem) *In a graph $G = (V, E)$, the number of spanning tree $\tau(G)$ satisfies*

$$\prod_{i>0} \lambda_i = \tau(G) \frac{\sum_v d_v}{\prod_v d_v}$$

where λ_i are eigenvalues of the Laplacian of G.

Proof:
For a matrix M and subsets S, T of indices of rows and columns, we denote a submatrix $\mathcal{L}_{S,T}$ of $\mathcal{L}$ by restricting rows and columns of $\mathcal{L}$ to S and T. For a fixed vertex v and the set $S_v = V \setminus \{v\}$, we consider the determinant of $\mathcal{L}_v = \mathcal{L}_{S_v,S_v}$ satisfying the following:

$$\begin{aligned} \det \mathcal{L}_v &= \det(B^*_{E,S_v} B_{E,S_v}) \\ &= \sum_{\substack{T \subseteq E \\ |T|=n-1}} (\det B^*_{T,S_v})(\det B_{T,S_v}) \\ &= \sum_{T \text{ a spanning tree}} \big(\det B_{T,S_v}\big)^2 \end{aligned}$$

where in the last step, we use the fact that $\det B_{T,S_v} = 0$ if T contains a cycle. In the other direction, if T does not contain a cycle, T is a spanning tree and

$$\det B_{T,S_v} = \pm \frac{1}{\sqrt{\prod_{v \in S_v} d_v}}.$$

Thus we have

$$\det \mathcal{L}_v \quad = \quad \frac{\tau(G)}{\prod_{v \in S_v} d_v}$$

Now, we consider the characteristic polynomial $p(x)$ of $\mathcal{L}$

$$p(x) \quad = \quad \det(xI - \mathcal{L}) \; = \; (x - \lambda_0)(x - \lambda_1) \ldots (x - \lambda_{n-1}).$$

The coefficient of the linear term in x is $\prod_{i>0} \lambda_i$, which, on the other hand, is equal to the following, by using a classical result due to Laplace:

$$\begin{aligned} \prod_{i>0} \lambda_i \quad &= \quad \sum_v \det \mathcal{L}_v \\ &= \quad \frac{\sum_v d_v}{\prod_v d_v} \tau(G). \end{aligned}$$

□

Suppose we consider the graph which is a path P_n on n edges. The eigenvalues are $\lambda_k = 1 - \cos \frac{\pi k}{n}$, for $k = 0, \ldots, n$. Since $\tau(P_n) = 1$, the following equality is an immediate consequence of the above theorem.

$$\prod_{k=1}^{n} (1 - \cos \frac{\pi k}{n}) = \frac{n}{2^{n-2}}. \tag{2}$$

3 Dirichlet eigenvalues

For a subset S of vertices in a graph G, there are typically two types of boundary, the vertex boundary and the edge boundary. The *edge boundary* of S, denoted by $\partial(S)$, defined as follows:

$$\partial(S) = \{\{u, v\} \in E(G) \;:\; u \in S \text{ and } v \notin S\}.$$

The *vertex boundary* of S, denoted by $\delta(S)$ is defined by

$$\delta(S) = \{v \in V \;:\; v \notin S \text{ and } v \sim u \in S\}.$$

A function $f : V \to \mathbb{R}$ is said to satisfy the *Dirichlet boundary condition* on S if $f(v) = 0$ for all $v \in \delta(S)$. The operator Δ acting on functions satisfying Dirichlet boundary condition on S can be represented by the submatrix of Δ_S restricted to vertices of S. The *Dirichlet eigenvalues* of S are the eigenvalues of Δ_S or its symmetric version $\mathcal{L}_S$, denoted by $\lambda_{i,S}$ for $i = 1, \ldots, |S|$. In particular, we denote $\lambda_S = \lambda_{1,S}$.

A forest is a subgraph that contains no cycle. For a given subset S with nontrivial vertex boundary, we call a subgraph of S a *rooted forest* F if each connected component of F contains exactly one vertex in $\delta(S)$. We can now give another modified version of Theorem 1, which we call the *Matrix-Forest Theorem.*

Theorem 2 (Matrix-Forest Theorem) *In a graph $G = (V, E)$ with a subset $S \subset V$ with $\delta(S) \neq \emptyset$, the number of rooted spanning tree $\tau_G(S)$ in the induced subgraph of $S \cup \delta(S)$ satisfies*

$$\prod_{i=1}^{|S|} \lambda_{i,S} = \frac{\tau_G(S)}{\prod_v d_v}$$

where $\lambda_{i,S}$ are Dirichlet eigenvalues of S.

Proof: The proof is quite similar to Theorem 1 by considering the determinant of $\mathcal{L}_S$:

$$\begin{aligned} \det \mathcal{L}_S &= \det(B^*_{E,S} B_{E,S}) \\ &= \sum_{\substack{T \subseteq E \\ |T|=n-1}} \det(B_{T,S^*} \det(B_{T,S}) \\ &= \sum_{T \text{ a tree}} \big(\det B_{T,S}\big)^2. \end{aligned}$$

The rest of the proof follows from the fact that $\det B_{T,S_v} = 0$ if T is not a rooted forest and,if T is a rooted forest, we have

$$\det B_{T,S} = \pm \frac{1}{\sqrt{\prod_{v \in S} d_v}}.$$

□

For example, suppose we consider G to be a cycle C_n on n vertices. S consists of all but one fixed vertex in C_n. The Drichelet eigenvalues for S are $1 - \cos \frac{2\pi k}{n}$ for $k = 1, \ldots, n-1$. The above theorem then implies the same equality as in (2).

For a non-emplty subset S of vertices in G, the *volume* of S, denoted by vol (S), is:

$$\text{vol } (S) = \sum_{x \in S} d_x.$$

The volume of G is denoted by vol $(G) = \sum_x d_x$. The *Cheeger ratio* of S, denoted by $h(S)$ is defined by:

$$h(S) = \frac{|\partial(S)|}{\min\{\text{vol } (S), \text{vol } (G) - \text{vol } (S)\}}.$$

The *Cheeger constant* of G, denoted by h_G, is

$$h_G = \min_S h(S).$$

The *Cheeger inequality* is a fundamental isoperimetric inequality that relates the eigenvalue λ_1 of the Laplacian $\mathcal{L}$ of a graph G with the Cheeger constant h_G.

$$2h_G \geq \lambda_1 \geq \frac{h_G^2}{2}.$$

The proof of the above inequality can be found in [6] and, in fact, the arguments in the proof provide the performance bounds for spectral partitioning algorithms that have been widely used.

Here we also define the *Cheeger constant of the subset S* as follows:

$$h^*{}_S = \min_{T \subseteq S} h(T).$$

We called h_S^* the local Cheeger constant of S. Note that the local Cheeger constant of S can be quite *different* from the Cheeger ratio of S.

Here is a variation of the Cheeger inequality associated with a subset S in G, relating the Dirichlet eigenvalue λ_S to the local Cheeger constant λ_S^*.

$$h_G^* \geq \lambda_S \geq \frac{h_S^*}{2}.$$

4 Connections between PageRank and discrete Green's function

The original definition of PageRank is motivated by modeling a typical surfer who goes to a random website with probability α, and goes to a linked site with probability $1-\alpha$ (see [5]). Therefore the PageRank vector r satisfies the following equation:

$$r = \frac{\mathbf{1}}{n} + (1-\alpha) r P.$$

where P is the transition probability matrix of G (as the *WWW*-graph) with n vertices and $\mathbf{1}$ is the all 1's vector. Here both r and $\mathbf{1}$ are represented as row vectors.

We will use the generalized notion of a *personalized PageRank* which has two parameters, a seed vector s, and the *jumping constant* α. Here s can be regarded as an initial probability distribution and α is a positive value that can be used to scale the rate of propagation. The PageRank $\mathrm{pr}(\alpha, s)$ is defined as follows:

$$\mathrm{pr}(\alpha, s) = \alpha s + (1-\alpha)\mathrm{pr}(\alpha, s) W. \tag{3}$$

Here W denotes a lazy walk of G, defined by

$$W = \frac{I+P}{2}.$$

An equivalent definition for $\mathrm{pr}(\alpha, s)$ is the following:

$$\mathrm{pr}(\alpha, s) = \alpha \sum_{k=0}^{\infty} (1-\alpha)^k s W^k. \tag{4}$$

The discrete Green's function is basically the inverse of the Laplacian operating on the space orthogonal to the null space of the Laplacian [8]. Namely,

$$\mathcal{G}\mathcal{L}x = \mathcal{L}\mathcal{G}x = x$$

if x is orthogonal to the eigenfunction ϕ_0 associated with eigenvalue 0 of $\mathcal{L}$.

In [8], a generalized Green's function $\mathcal{G}_\beta$ is introduced with an parameter $\beta \geq 0$ which was useful for the proofs there. $\mathcal{G}_\beta$ is the inverse of $\beta I + \mathcal{L}$. Namely, suppose that ϕ_i, $i = 0, \ldots, n-1$, denote the orthonormal eigenfunctions of $\mathcal{L}$. I.e.,

$$\mathcal{L} = \sum_{i=1}^{n-1} \lambda_i \phi_i^* \phi_i.$$

Then, for $\beta \geq 0$,

$$\mathcal{G}_\beta = \sum_{i=1}^{n-1} \frac{1}{\beta + \lambda_i} \phi_i^* \phi_i.$$

Here we also consider the (unsymmetrized) version $\mathbf{G}_\beta$, for $\beta > 0$, satisfying

$$\mathbf{G}_\beta(\beta I + \Delta) = I \tag{5}$$

Since the PageRank $p = \mathrm{pr}(\alpha, s)$ satisfies

$$p(I - (1-\alpha)W) = \alpha s,$$

we see that $\mathrm{pr}(\alpha, s)$ is the unique solution for x for the following equation:

$$x(\beta I + \Delta) = \beta s,$$

where

$$\beta = \frac{2\alpha}{1-\alpha}. \tag{6}$$

The definition of $\mathbf{G}_\beta$ in (3) and (4) implies the following interpretation for $\mathrm{pr}(\alpha, s)$

$$\mathrm{pr}(\alpha, s) = \beta s \mathbf{G}_\beta \tag{7}$$

where β is related to α as in (6).

Here we state a useful recurrence for $\mathbf{G}_\beta$ which is an alternate way to express (5).

$$(1+\beta)\mathbf{G}_\beta = I + \mathbf{G}_\beta P. \tag{8}$$

For a subset T of vertices in G, suppose the induced subgraph on T, denoted by G', is connected. We can extend the definitions and define the PageRank and Green's function for G'. Namely, the personalized PageRank restricted to T with a seed vector s, and the jumping constant α, is defined as follows:

$$\mathrm{pr}'(\alpha, s) = \alpha s + (1-\alpha)\mathrm{pr}'(\alpha, s) W_T.$$

Or, equivalently,

$$\mathrm{pr}'(\alpha, s) = \alpha \sum_{i=0}^{\infty} (1-\alpha)^k s W_T^k. \tag{9}$$

The Green's function $\mathbf{G}'_\beta$ restricted to T is defined as follows:

$$\mathbf{G}'_\beta(\beta I_T + \Delta_T) = I \tag{10}$$

and $\mathrm{pr}'(\alpha, s)$ is the unique solution for x for the following equation:

$$x(\beta I_T + \Delta_T) = \beta s,$$

where β is as defined in (6). Therefore for $\mathrm{pr}'(\alpha, s)$ for the induced subgraph on T, we have

$$\mathrm{pr}'(\alpha, s) = \beta s G'_\beta.$$

From the definition, for any two vertices u and v in T, we have

$$\mathbf{G}'_\beta(u, v) \le \mathbf{G}(u, v)$$

since $W_T^k \le W^k$ for any $k \ge 0$.

We remark that for the PageRank and Green's function which are operating on functions defined on T., we extend them to functions defined on the vertex set of G with Dirichlet boundary condition for T.

From (4), (9) and the definitions, we have the following fact which will be useful later.

Lemma 1 *For vertex v and $s : T \to \mathbb{R}$, we have*
(1)

$$\mathrm{pr}'(\alpha, s)(v) \le \mathrm{pr}(\alpha, s)(v).$$

(2) The transpose P^* of P satisfies $P^* = DPD^{-1}$. Therefore we have $W^* = DWD^{-1}$ and for any positive integer k, we have $(W^*)^k = DW^kD^{-1}$.

$$\mathbf{G}_\beta^* = D\mathbf{G}_\beta D^{-1}.$$

(3) Let D_T be the restriction of D to rows and columns indexed by vertices of T. We have

$$\mathbf{G}'^*_\beta = D_T\mathbf{G}'_\beta D_T^{-1}.$$

5 Relating the Cheeger constant to the PageRank

For a subset S, we consider the probability distribution which is

$$f_S(v) \quad = \quad \begin{cases} \frac{d_v}{\mathrm{vol}\ (S)} & \text{if } \ v \in S, \\ 0 & \text{otherwise.} \end{cases}$$

Note that we can write $f_S = \chi_S D/\mathrm{vol}\ (S)$ where χ_S is the characteristic function of S. Also, for a function f, we write $f(S) = \sum_{v\in S} f(v)$.

Here we state several facts that relate the Cheeger constant to PageRank.

Lemma 2 *For a subset S, the probability function f_S satisfies*

$$\mathrm{pr}(\alpha, f_S)(S) \geq 1 - \frac{1-\alpha}{2\alpha}h(S).$$

Proof: We use (7), (6) and (8), we have

$$\begin{aligned}
\mathrm{pr}(\alpha, f_S)(S) &= \beta f_S \mathbf{G}_\beta \chi_S^* \\
&= f_S\big(I - \mathbf{G}_\beta(I-P)\big)\chi_S^* \\
&= 1 - \frac{1}{\beta\mathrm{vol}\ (S)}\chi_S(\beta\mathbf{G}_\beta^*)(D-A)\chi_S^* \\
&\geq 1 - \frac{1}{\beta\mathrm{vol}\ (S)}\chi_S(D-A)\chi_S^* \\
&= 1 - \frac{|\partial(S)|}{\beta\mathrm{vol}\ (S)} \\
&= 1 - \frac{1}{\beta}h(S)
\end{aligned}$$

□

Lemma 3 *For a subset S of vertices in G with* vol $(S) \le$ vol $(G)/2$*, there is a subset $T \subset S$ with* vol $(T) \ge$ vol $(S)/2$ *such that for any $u \in T$, the personalized pagerank* $\mathrm{pr}(\alpha, u)$ *satisfies that*

$$\mathrm{pr}(\alpha, u)(S) \ge 1 - \frac{(1-\alpha)h_S}{\alpha}.$$

Proof: We consider a subset T' of S defined by

$$T' = \{v \in S \ : \ \mathrm{pr}(\alpha, v)(\bar{S}) \ge \frac{1-\alpha}{\alpha} h_S\}.$$

and we have

$$\begin{aligned} \mathrm{pr}(\alpha, f_S)(\bar{S}) &= \sum_{v \in S} \frac{d_v}{\mathrm{vol}\ (S)} \mathrm{pr}(\alpha, v)(\bar{S}) \\ &\ge \sum_{v \in T'} \frac{d_v}{\mathrm{vol}\ (S)} \mathrm{pr}(\alpha, v)(\bar{S}) \\ &\ge \frac{(1-\alpha)\mathrm{vol}\ (T')}{\alpha \mathrm{vol}\ (S)} h_S. \end{aligned}$$

On the other hand,

$$\begin{aligned} \mathrm{pr}(\alpha, f_S)(\bar{S}) &= 1 - \mathrm{pr}(\alpha, f_S)(S) \\ &\le \frac{1-\alpha}{2\alpha} h_S. \end{aligned}$$

Therefore, we have vol $(T') \le$ vol $(S)/2$. We define $T = S \setminus T'$. Therefore for u in T, we have

$$\mathrm{pr}(\alpha, u)(S) = 1 - \chi_u(\beta \mathbf{G}_\beta)(\bar{S}) \ge 1 - \frac{1-\alpha}{\alpha} h_S$$

as claimed. □

The following inequality is proved in [7] which improves slightly an earlier result in [1]:

Lemma 4 *For any positive integer k, a jumping constant $\alpha, 0 \le \alpha \le 1$, a subset S of vertices, and a vertex $u \in S$, the personalized pagerank* $\mathrm{pr}(\alpha, u)$ *satisfies*

$$\mathrm{pr}(\alpha, u)(S) - \pi(S) \le \Big(1 - (1-\alpha)^k + \sqrt{\frac{\mathrm{vol}\ (S)}{d_u}} \big(1 - \frac{\gamma_\alpha^2}{8}\big)^k (1-\alpha)^k\Big)(1 - \pi(S)),$$

where γ_α is the minimum Cheeger ratio determined by a sweep of $\mathrm{pr}(\alpha, u)$.

We are now ready to relate to Cheeger constant to PageRank.

Theorem 3 *For a subset S in G with* vol $(S) \leq$ vol $(G)/2$*, there is a subset $T \subset S$ with* vol $(T) \geq$ vol $(S)/2$ *such that for any $u \in T$, we have*

$$h_S \geq \frac{\alpha}{2} \geq \frac{\gamma_u^2}{16 \log(\text{vol }(G))}.$$

Proof: Let S denote the subset achieving the Cheeger constant h_G. By combining Lemma 3 and Lemma 4, there is a T with vol $(T) \geq$ vol $(S)/2$ such that for $u \in T$, we have

$$1 - \frac{(1-\alpha)h_S}{\alpha} - \pi(S) \leq \left(1 - (1-\alpha)^k + \frac{\text{vol }(S)}{d_u}(1 - \frac{\gamma_u^2}{8})^k (1-\alpha)^k\right)(1 - \pi(S)).$$

This implies

$$\frac{h_S(1-\alpha)}{\alpha(1-\pi(S))} \geq (1-\alpha)^k \big(1 - \frac{\text{vol }(S)}{d_u}(1 - \frac{\gamma_\alpha^2}{8})^k\big) \qquad (11)$$

We then choose k and α as follows:

$$k = \lceil \frac{16 \log(\text{vol }(G))}{\gamma_\alpha^2} \rceil, \quad \alpha = \frac{1}{k}.$$

Then (11) implies

$$h_S \geq \frac{\alpha}{2} \geq \frac{\gamma_u^2}{32 \log(\text{vol }(G))},$$

as claimed. □

6 Relating the PageRank of a graph to that of its subgraphs

Let T denote a subset of vertices in G and we consider the induced subgraph on T, denoted by G'. Suppose G' is connected. We wish to show that the PageRank of G' and the PageRank of the whole graph G are close to each other if the jumping constant is within a certain range.

One underlying motivation is the case that T is a community within a graph G. Although it is not easy to define "community", for our purpose here a community can be regarded as an induced subgraph on T so that the Cheeger constant $h(T)$ is relatively small (in comparison with some given quantity or the Cheeger constants of its own subsets). We here will only consider the mathematical analysis of the PageRank of an induced subgraph without further going into various aspects or applications of identifying communities.

Let $\mathbf{G}_\beta$ and $\mathbf{G}'_\beta$ denote the Green's functions of G and G', respectively. Note that the PageRanks of G and G' are closely related to their Green's functions as in equations (7) and (11). Will it be possible to bound the difference of $\mathbf{G}_\beta - \mathbf{G}'_\beta$? If we do the usual worst-case analysis, the answer is negative except for some trivial upper bounds. However, the answer can depend on many factors, such as the Cheeger constant of T, the choice of β, as well as the probabilistic nature of the problem. For example, we would like to show that for a fixed subset R of T, a random vertex v has the difference of its Green's functions evaluated at R, e.g.,

$$(\mathbf{G}_\beta - \mathbf{G}'_\beta)(v, R) = \sum_{u \in R} \mathbf{G}_\beta(v, u) - \sum_{u \in R} \mathbf{G}'_\beta(v, u),$$

being very small (as a function of β and the Cheeger constant of T).

To do so, we will first define a probabilistic space on T. In a series of lemmas, we derive bounds for expected values and variances of some formulations using $\mathbf{G}_\beta - \mathbf{G}'_\beta$. The inequalities from the lemmas then lead to the main theorems later on.

We choose each vertex v in T with probability $d_v/\text{vol }(T)$ where d_v is the degree of v in G. We define f_T as follows:

$$f_T(u) \quad = \quad \begin{cases} \frac{d_u}{\text{vol }(T)} & \text{if } u \in T, \\ 0 & \text{otherwise.} \end{cases}$$

We will first show that the expected value of $(\mathbf{G}_\beta - \mathbf{G}'_\beta)(v, T)$ is small:

Lemma 5 *For a subset T of vertices in G and $\beta > 0$, the Green's functions $\mathbf{G}_\beta$ for G and $\mathbf{G}'_\beta$ for the induced subgraph G'on T satisfy*

$$\mathbf{E}_T\big((\mathbf{G}_\beta - \mathbf{G}'_\beta)(v, T)\big) \quad \leq \quad \frac{h(T)}{\beta^2}. \tag{12}$$

Proof: We note that

$$\mathbf{E}_T\big((\mathbf{G}_\beta - \mathbf{G}'_\beta)(T)\big) \quad = \quad f_T(\mathbf{G}_\beta - \mathbf{G}'_\beta)\mathbf{1}_T^*$$

where $\mathbf{1}$ is the all 1's function (as a row vector) and f_T is as defined before. From (5) and (10), we have

$$\begin{aligned} f_T \mathbf{G}_\beta(\beta I + \Delta) \quad &= \quad f_T \\ f_T \mathbf{G}'_\beta(\beta I_T + \Delta_T) \quad &= \quad f_T. \end{aligned}$$

By taking the difference of the above two equations, we have

$$\big((f_T\mathbf{G}_\beta)_T - f_T\mathbf{G}'_\beta\big)(\beta I_T + \Delta_T) - (f_T\mathbf{G}_\beta)_{\bar{T}} P_{\bar{T},T} = 0$$

where $\bar{T}$ is the complement of T, g_T is the restriction of a function g to T and $P_{\bar{T},T}$ is the restriction of P to rows restricted to $\bar{T}$ and columns to T.

Therefore we have

$$\big((f_T\mathbf{G}_\beta)_T - f_T\mathbf{G}'_\beta\big)(\beta I_T + \Delta_T) \quad = \quad (f_T\mathbf{G}_\beta)_{\bar{T}} P_{\bar{T},T}.$$

By multiplying both sides of the above equation by $\mathbf{G}'_\beta$, we have

$$(f_T\mathbf{G}_\beta)_T - f_T\mathbf{G}'_\beta \quad = \quad (f_T\mathbf{G}_\beta)_{\bar{T}} P_{\bar{T},T}\mathbf{G}'_\beta$$

which then leads to

$$\big((f_T\mathbf{G}_\beta)_T - f_T\mathbf{G}'_\beta\big)\mathbf{1}^* \quad = \quad (f_T\mathbf{G}_\beta)_{\bar{T}} P_{\bar{T},T}\mathbf{G}'_\beta\mathbf{1}^*.$$

We note that it follows from the definition of P that

$$\mathbf{1}_T D P(v) \leq d_v.$$

This then implied that for every integer k,

$$\mathbf{1}_T D W^k(v) \leq d_v.$$

By using (4) and (7), we have

$$\begin{aligned} \beta\mathbf{1}_T D\mathbf{G}_\beta(v) &\leq \frac{\beta}{1+\beta}\sum_{k\geq 0}\mathbf{1}_T D\Big(\frac{W}{1+\beta}\Big)^k \\ &\leq d_v. \end{aligned}$$

Furthermore, we have

$$\beta\mathbf{1}_T D\mathbf{G}_\beta W(v) \quad \leq \quad d_v$$

By using Lemma 1, we have

$$\begin{aligned} \beta\mathbf{G}'_\beta\mathbf{1}_T^*(v) &= \beta(\mathbf{1}_T D_T\mathbf{G}'_\beta D_T^{-1})^*(v) \\ &\leq \mathbf{1}_T \leq \mathbf{1}_T(v), \end{aligned}$$

and

$$\beta(f_T\mathbf{G}_\beta)_{\bar{T}} \quad \leq \quad \frac{1}{\text{vol }(T)}(\mathbf{1}D)_{\bar{T}}$$

since the corresponding inequalities are componentwise.

Therefore, we have

$$\begin{aligned} \big((f_T\mathbf{G}_\beta)_T - f_T\mathbf{G}'_\beta\big)\mathbf{1}^* &= (f_T\mathbf{G}_\beta)_{\bar{T}} P_{\bar{T},T}\mathbf{G}'_\beta\mathbf{1}^* \\ &\leq \frac{1}{\beta\text{vol }(T)}(\mathbf{1}D)_{\bar{T}} P_{\bar{T},T}\mathbf{G}'_\beta\mathbf{1}^* \\ &\leq \frac{1}{\beta^2\text{vol }(T)}(\mathbf{1}D)_{\bar{T}} P_{\bar{T},T}\mathbf{1}^* \\ &= \frac{e(T,\bar{T})}{\beta^2\text{vol }(T)} \\ &= \frac{h(T)}{\beta^2}. \end{aligned}$$

This proves Lemma 12. □

Immediately we have the following:

Theorem 4 *For positive values α, ϵ satisfying*

$$\epsilon \geq \frac{(1-\alpha)h(T)}{2\alpha},$$

the expected value of the difference of personalized PageRank in the induced subgraph on T and personalized PageRank in G, with a seed at a random vertex v, evaluated at R, satisfies

$$\mathrm{E}[|\ \mathrm{pr}(v,\alpha)(R) - \mathrm{pr}'(v,\alpha)(R)\ |] \leq \epsilon,$$

for any subset R of T.

Proof: Since $\mathrm{pr}(v,\alpha)(R) \geq \mathrm{pr}'(v,\alpha)(R)$, it is enough to deal with $\mathrm{pr}(v,\alpha)(R) - \mathrm{pr}'(v,\alpha)(R)$. We see that for $\beta = 2\alpha/(1-\alpha)$, we have

$$\begin{aligned} \mathrm{E}[\mathrm{pr}(v,\alpha)(R) - \mathrm{pr}'(v,\alpha)(R)] &= \sum_v \frac{d_v}{\mathrm{vol}\ (T)}\big(\mathrm{pr}(v,\alpha)(R) - \mathrm{pr}'(v,\alpha)(R)\big) \\ &= \beta f_T(\mathbf{G}_\beta - \mathbf{G}'_\beta)\mathbf{1}^*_R. \end{aligned}$$

Since for any two vertices u and v,

$$\mathbf{G}_\beta(u,v) \geq \mathbf{G}'_\beta(u,v),$$

we have

$$\begin{aligned} \mathrm{E}[\mathrm{pr}(v,\alpha)(R) - \mathrm{pr}'(v,\alpha)(R)] &= \beta f_T(\mathbf{G}_\beta - \mathbf{G}'_\beta)\mathbf{1}^*_R \\ &\leq \beta f_T(\mathbf{G}_\beta - \mathbf{G}'_\beta)\mathbf{1}^*_T \\ &\leq \frac{h(T)}{\beta} \\ &= \frac{2\alpha h(S)}{\alpha} \\ &\leq \epsilon \end{aligned}$$

by Lemma 12. □

Next, we wish to say the difference of PageRanks in T and G are close not only on average but also in the following stronger sense:

Theorem 5 *Suppose that positive values α, ϵ satisfy*

$$\epsilon \geq \frac{(1-\alpha)h(T)}{2\alpha}.$$

Then we have

$$\sum_{v\in T} \frac{d_v}{\text{vol }(T)} \big(\text{pr}(v,\alpha)(T) - \text{pr}'(v,\alpha)(T)\big)^2 \le \frac{\epsilon}{2}. \tag{13}$$

Proof: We can write

$$\begin{aligned} \text{pr}(v,\alpha)(T) - \text{pr}'(v,\alpha)(T) &= \beta\chi_v(\mathbf{G}_\beta - \mathbf{G}'_\beta)\mathbf{1}_T \\ &= \beta\mathbf{1}_T D(\mathbf{G}_\beta - \mathbf{G}'_\beta)D^{-1}(v) \end{aligned}$$

We consider

$$\begin{aligned} & \sum_{v\in T} \frac{d_v}{\text{vol }(T)} \big(\text{pr}(v,\alpha)(T) - \text{pr}'(v,\alpha)(T)\big)^2 \\ = \; & \sum_{v\in T} \frac{\beta^2 d_v}{\text{vol }(T)} \big(\mathbf{1}_T D(\mathbf{G}_\beta - \mathbf{G}'_\beta)D^{-1}(v)\big)^2 \\ = \; & \beta^2(f_T\mathbf{G}_\beta - f_T\mathbf{G}'_\beta)D^{-1}(\mathbf{G}_\beta - \mathbf{G}'_\beta)^* D\mathbf{1}_T^* \\ \le \; & \beta^2(f_T\mathbf{G}_\beta - f_T\mathbf{G}'_\beta)D^{-1}\mathbf{G}_\beta^* D\mathbf{1}_T^* \\ = \; & \beta^2(f_T\mathbf{G}_\beta - f_T\mathbf{G}'_\beta)\mathbf{G}_\beta(T) \\ = \; & \beta(f_T\mathbf{G}_\beta - f_T\mathbf{G}'_\beta)(T) \\ \le \; & \frac{\epsilon}{2} \end{aligned}$$

by Lemma 12. □

Theorem 6 *Suppose positive values α, ϵ satisfy*

$$\epsilon \ge \frac{(1-\alpha)h(T)}{2\alpha}.$$

For any subset R in T, there is a subset T' of T with $\text{vol }(T') \ge \text{vol }(T)/2$ so that for every v in T' we have

$$\text{pr}(v,\alpha)(R) - \text{pr}'(v,\alpha)(R) \le \sqrt{\epsilon}.$$

Proof:

Let T'' denote the subset consisting of all v in T satisfying

$$\text{pr}(v,\alpha)(R) - \text{pr}'(v,\alpha)(R) \ge \sqrt{\epsilon}.$$

From (13), we have

$$\begin{aligned} \sum_{v\in T''} \frac{d_v}{\text{vol }(T)} \big(\text{pr}(v,\alpha)(R) - \text{pr}'(v,\alpha)(R)\big)^2 &\le \sum_{v\in T} \frac{d_v}{\text{vol }(T)} \big(\text{pr}(v,\alpha)(T) - \text{pr}'(v,\alpha)(T)\big)^2 \\ &\le \frac{\epsilon}{2}. \end{aligned}$$

On the other hand, we have

$$\begin{aligned}\sum_{v\in T''}\frac{d_v}{\text{vol }(T)}\big(\text{pr}(v,\alpha)(R)-\text{pr}'(v,\alpha)(R)\big)^2 &\geq \sum_{v\in T''}\frac{d_v}{\text{vol }(T)}\epsilon\\ &\geq \frac{\text{vol }(T'')}{\text{vol }(T)}\epsilon.\end{aligned}$$

Together, we obtain

$$\text{vol }(T'')\leq\frac{\text{vol }(T)}{2}.$$

By taking $T'=T\setminus T''$, the statement of Theorem 5 holds, subject to proving inequality (13). This completes the proof of Theorem 5. □

We note that Theorem 6 has immediate algorithmic implications. It asserts that if α is in the specified range, it is enough to use the Green's function of the induced subgraph T as an estimate for the Green's function for the whole graph for a large subset of T.

7 The PageRank and the hitting time

In this section, we will consider the connections of PageRank and the hitting time via the Green's function. For two vertices u and v, the *hitting time* for a random walk starting at u is the number of steps when v is first reached (see [12]). There has been previous research relating hitting time to the discrete Green's function (e.g.,[8]), as well as the examining various probabilistic aspects of PageRank [3, 11]. Some modified versions of PageRank were proposed, such as using the frequency of visits by random walks [3] or using some variations of hitting time [11]. Here we will give a brief description of a parameterized family of hitting time. The connection with PageRank then follows.

For a graph G, we define a matrix Q with columns and rows indexed by vertices of G as follows:

$$Q(u,v)\quad=\quad\begin{cases}\text{E(the hitting time from } u \text{ to } v) & \text{if } u\neq v,\\ 0 & \text{if } u=v.\end{cases}$$

Let $R(u)$ denote the returning time at u which is the number of steps in a random walk starting at u and first returning to u. It is not difficult to show that $E(R(u))=\text{vol }(G)/d_u$, the inverse of the stationary distribution. For a random walk with transition probability matrix P, we have the following recurrence relation, for $u\neq v$:

$$Q(u,v)\quad=\quad\sum_{\substack{w\\ u\sim w}}\frac{1}{d_u}Q(w,v)+1.$$

Hence, we have

$$Q = PQ + J - R$$

where J is the all 1's matrix and R is a diagonal matrix with diagonal entries $E(R(v))$.

Now we consider an α-jumping random walk for some α with $0 < \alpha < 1$. This random walk moves from a vertex v, with probability α, to a random vertex and with probability $1 - \alpha$ follows the lazy random walk W. Namely, theq transition probability matrix W_α of the α-jumping random walk satisfies

$$W_\alpha(x, y) = (1 - \alpha)W(x, y) + \alpha\pi(y).$$

where $\pi(y) = d_y/\text{vol}\,(G)$ and a random vertex u is chosen with probability $\pi(u)$. Let Q_α denote the hitting time matrix corresponding to the α-jumping random walk W_α. Then we have

$$Q_\alpha = (1 - \alpha)(WQ_\alpha - R) + (J - I) + \alpha Q' \tag{14}$$

where for $x \neq y$, we have

$$Q'(x, y) = \sum_u \pi(u)Q_\alpha(u, y) = (\pi Q)(y)$$

and $Q'(x, x) = 0$.

Note that, from (14) we have

$$\begin{aligned} \alpha\pi Q' &= \pi Q_\alpha - (1 - \alpha)\pi(WQ_\alpha - R) - \pi(J - I) \\ &= \pi Q_\alpha - (1 - \alpha)\pi Q_\alpha + (1 - \alpha)\mathbf{1} - \mathbf{1} + \pi \\ &= \alpha\pi Q_\alpha - \alpha\mathbf{1} + \pi. \end{aligned}$$

On the other hand,

$$\pi Q' = \pi Q_\alpha - \pi Q_\alpha \Pi$$

where Π is the diagonal matrix with entries $\Pi(x, x) = \pi(x)$. This implies

$$(\pi Q_\alpha)(x) = \pi(x)^{-1} - \frac{1}{\alpha}.$$

Thus,

$$Q' = \mathbf{1}^*\pi^{-1} - \frac{1}{\alpha}J - R + \frac{1}{\alpha}I.$$

We can rewrite (14) as follows:

$$Q_\alpha = (1 - \alpha)WQ_\alpha + \alpha\mathbf{1}^*\pi^{-1} - \alpha R.$$

From (4) and (7), we have

$$Q_\alpha = \beta\mathbf{G}_\beta\big(\mathbf{1}^*\pi^{-1} - R\big).$$

Therefore we have proved the following :

Theorem 7 *The hitting time Q_α for the α-jumping random walk satisfies*

$$Q_\alpha = \beta \mathbf{G}_\beta \big(\mathbf{1}^* \pi^{-1} - R \big) \tag{15}$$

where $\beta = 2\alpha/(1-\alpha)$, R is the diagonal matrix with $R(v,v) = \pi^{-1}(v)$ and $\mathbf{1}^*\pi^{-1}$ is the matrix product of the column vector $\mathbf{1}^*$ and the row vector π^{-1}.

We note that the equation in (15) immediately relates the hitting time with PageRank as in (7). In [11], Hopcroft and Sheldon proposed to use a variation of hitting time as a 'manipulation-resistant' alternative for PageRank. Here we can make a similar argument for Q_α. Indeed, it would be interesting to see how the ranking determined by the hitting time Q_α compared with that by PageRank.

References

[1] R. Andersen, F. Chung and K. Lang, Local graph partitioning using PageRank vectors, *Proceedings of the 47th Annual IEEE Symposium on Founation of Computer Science (FOCS'2006)*, 475–486.

[2] R. Andersen, F. Chung and K. Lang, Detecting sharp drops in PageRank and a simplified local partitioning algorithm, *Theory and Applications of Models of Computation, Proceedings of TAMC 2007*, 1–12.

[3] Y. Bao, G. Feng, T.-Y. Liu, Z.-M. Ma, and Y. Wang, Ranking websites: a probabilistic view, *Internet Math.*, **3**, (2007).

[4] P. Berkhin, Bookmark-coloring approach to personalized pagerank computing, *Internet Math.*, **3**, (2006), 41–62.

[5] S. Brin and L. Page, The anatomy of a large-scale hypertextual Web search engine, *Computer Networks and ISDN Systems*, **30 (1-7)**, (1998), 107–117.

[6] F. Chung, *Spectral Graph Theory*, AMS Publications, 1997.

[7] F. Chung, Four proofs of the Cheeger inequality and graph partition algorithms, *Proceedings of ICCM*, 2007.

[8] F. Chung and S.-T. Yau, Coverings, heat kernels and spanning trees, *Electronic Journal of Combinatorics* **6** (1999), #R12.

[9] G. Jeh and J. Widom, Scaling personalized web search, *Proceedings of the 12th World Wide Web Conference WWW*, (2003), 271–279.

[10] F. Kirchhoff, Über die Auflösung der Gleichungen, auf welche man bei der Untersuchung der linearen Verteilung galvanischer Ströme geführt wird, *Ann. Phys. chem.* **72** (1847), 497–508

[11] J. Hopcroft and D. Sheldon, Manipulation-resistant reputations using hitting time, WAW 2007, *LNCS 4863*, 68–81.

[12] L. Lovász, Random walks on Graphs, *Combinatorics, Paul Erdős is Eighty*, Vol. 2, Bolyai Society Mathematical Studies, 2, Keszthely (Hungary), 1993, 1–46.