Learning from Labeled and Unlabeled Data on a Directed Graph

Dengyong Zhou DENGYONG.ZHOU@TUEBINGEN.MPG.DE
Max Planck Institute for Biological Cybernetics, Spemannstr. 38, 72076 Tübingen, Germany

Jiayuan Huang J9HUANG@CS.UWATERLOO.CA
School of Computer Science, University of Waterloo, Waterloo ON, N2L 3G1, Canada

Bernhard Schölkopf BERNHARD.SCHOELKOPF@TUEBINGEN.MPG.DE
Max Planck Institute for Biological Cybernetics, Spemannstr. 38, 72076 Tübingen, Germany

Abstract

We propose a general framework for learning from labeled and unlabeled data on a directed graph in which the structure of the graph including the directionality of the edges is considered. The time complexity of the algorithm derived from this framework is nearly linear due to recently developed numerical techniques. In the absence of labeled instances, this framework can be utilized as a spectral clustering method for directed graphs, which generalizes the spectral clustering approach for undirected graphs. We have applied our framework to real-world web classification problems and obtained encouraging results.

1. Introduction

Given a directed graph, the vertices in a subset of the graph are labeled. Our problem is to classify the remaining unlabeled vertices. Typical examples of this kind are web page categorization based on hyperlink structure and document classification based on citation graphs (Fig. 1). The main issue to be resolved is to determine how to effectively exploit the structure of directed graphs.

One may assign a label to an unclassified vertex on the basis of the most common label present among the classified neighbors of the vertex. However we want to exploit the structure of the graph globally rather than locally such that the classification or clustering is *consistent* over the whole graph. Such a point of view has been considered previously in the method of Zhou et al. (2005). It is motivated by the framework of *hubs* and *authorities* (Kleinberg, 1999), which separates web pages into two categories and uses the following recursive notion: a hub is a web page with links to many good authorities, while an authority is a web page that receives links from many good hubs. In contrast, the approach that we will present is inspired by the ranking algorithm *PageRank* used by the Google search engine (Page et al., 1998). Different from the the framework of hubs and authorities, PageRank is based on a direct recursion as follows: an authoritative web page is one that receives many links from other authoritative web pages. When the underlying graph is undirected, the approach that we will present reduces to the method of Zhou et al. (2004).

There has been a large amount of activity on how to exploit the link structure of the web for ranking web pages, detecting web communities, finding web pages similar to a given web page or web pages of interest to a given geographical region, and other applications. One may refer to (Henzinger, 2001) for a comprehensive survey. Unlike those work, the present work is on how to classify the unclassified vertices of a directed graph in which some vertices have been classified by globally exploiting the structure of the graph. Classifying a finite set of objects in which some are labeled is called *transductive inference* (Vapnik, 1998). In the absence of labeled instances, our approach reduces to a spectral clustering method for directed graphs, which generalizes the work of Shi and Malik (2000) that may be the most popular spectral clustering scheme for undirected graphs. We would like to mention that understanding how eigenvectors partition a directed graph has been proposed as one of six algorithmic challenges in web search engines by Henzinger (2003). The framework of probabilistic relational models may also be used to

Appearing in *Proceedings of the 22nd International Conference on Machine Learning*, Bonn, Germany, 2005.

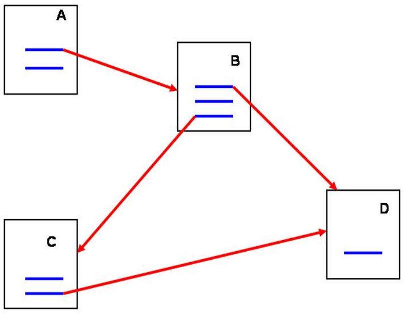


Figure 1. The World Wide Web can be thought of as a directed graph, in which the vertices represent web pages, and the directed edges hyperlinks.

deal with structured data like the web (e.g. Getoor et al. (2002)). In contrast to the spirit of the present work however, it focuses on modeling the probabilistic distribution over the attributes of the related entities in the model.

The structure of this paper is as follows. We first introduce some basic notions from graph theory and Markov chains in Section 2. The framework for learning from directed graphs is presented in Section 3. In the absence of labeled instances, as shown in section 4, this framework can be utilized as a spectral clustering approach for directed graphs. In Section 5, we develop discrete analysis for directed graphs, and characterize this framework in terms of discrete analysis. Experimental results on web classification problems are described in Section 6.

2. Preliminaries

A directed graph $G = (V, E)$ consists of a finite set V, together with a subset $E \subseteq V \times V$. The elements of V are the *vertices* of the graph, and the elements of E are the *edges* of the graph. An edge of a directed graph is an ordered pair $[u, v]$ where u and v are the vertices of the graph. We say that the vertex v is *adjacent from* the vertex u, and the the vertex u is *adjacent to* the vertex v. Moreover, we say that the edge is *incident from* the vertex u and *incident to* the vertex v. When $u = v$ the edge is called a *loop*. A graph is *simple* if it has no loop.

A *path* in a directed graph is a tuple of vertices $(v_1, v_2, \ldots, v_p)$ with the property that $[v_i, v_{i+1}] \in E$ for $1 \leq i \leq p-1$. We say that a directed graph is *strongly connected* when for every pair of vertices u and v there is a path in which $v_1 = u$ and $v_p = v$. For a strongly connected graph, there is an integer $k \geq 1$ and a unique partition $V = V_0 \cup V_1 \cup \cdots \cup V_{k-1}$ such that for all $0 \leq r \leq k-1$ each edge $[u, v] \in E$ with $u \in V_r$ has $v \in V_{r+1}$, where $V_k = V_0$, and k is maximal, that is, there is no other such partition $V = V_0^{'} \cup \cdots \cup V_{k^{'}-1}^{'}$ with $k^{'} > k$. When $k = 1$, we say that the graph is *aperiodic*; otherwise we say that the graph is *periodic*.

A graph is *weighted* when there is a function $w : E \to \mathbb{R}^+$ which associates a positive value $w([u, v])$ with each edge $[u, v] \in E$. The function w is called a *weight function*. Typically, we can equip a graph with a canonical weight function defined by $w([u, v]) := 1$ at each edge $[u, v] \in E$. Given a weighted directed graph and a vertex v of this graph, the *in-degree function* $d^- : V \to \mathbb{R}^+$ and *out-degree function* $d^+ : V \to \mathbb{R}^+$ are respectively defined by $d^-(v) := \sum_{u \to v} w([u, v])$, and $d^+(v) := \sum_{u \leftarrow v} w([v, u])$, where $u \to v$ denotes the set of vertices adjacent to the vertex v, and $u \leftarrow v$ the set of vertices adjacent from the vertex v.

Let $\mathcal{H}(V)$ denote the space of functions, in which each one $f : V \to \mathbb{R}$ assigns a real value $f(v)$ to each vertex v. A function in $\mathcal{H}(V)$ can be thought of as a column vector in $\mathbb{R}^{|V|}$, where $|V|$ denotes the number of the vertices in V. The function space $\mathcal{H}(V)$ then can be endowed with the standard inner product in $\mathbb{R}^{|V|}$ as $\langle f, g \rangle_{\mathcal{H}(V)} = \sum_{v \in V} f(v)g(v)$ for all $f, g \in \mathcal{H}(V)$. Similarly, define the function space $\mathcal{H}(E)$ consisting of the real-valued functions on edges. When the function space of the inner product is clear in its context, we omit the subscript $\mathcal{H}(V)$ or $\mathcal{H}(E)$.

For a given weighted directed graph, there is a natural random walk on the graph with the transition probability function $p : V \times V \to \mathbb{R}^+$ defined by $p(u, v) = w([u, v])/d^+(u)$ for all $[u, v] \in E$, and 0 otherwise. The random walk on a strongly connected and aperiodic directed graph has a unique *stationary distribution* π, i.e. a unique probability distribution satisfying the *balance equations* $\pi(v) = \sum_{u \to v} \pi(u)p(u, v)$, for all $v \in V$. Moreover, $\pi(v) > 0$ for all $v \in V$.

3. Regularization Framework

Given a directed graph $G = (V, E)$ and a label set $\mathcal{Y} = \{1, -1\}$, the vertices in a subset $S \subset V$ is labeled. The problem is to classify the vertices in the complement of S. The graph G is assumed to be strongly connected and aperiodic. Later we will discuss how to dispose this assumption.

Assume a classification function $f \in \mathcal{H}(V)$, which assigns a label sign $f(v)$ to each vertex $v \in V$. On the one hand, the classification function should be as *smooth* as possible. Specifically, a pair of vertices linked by an edge are likely to have the same label; moreover, vertices lying on a densely linked subgraph are likely to have the same label. Thus we define a functional

$$\Omega(f) := \frac{1}{2} \sum_{[u,v]\in E} \pi(u)p(u,v) \left(\frac{f(u)}{\sqrt{\pi(u)}} - \frac{f(v)}{\sqrt{\pi(v)}} \right)^2, \tag{1}$$

which sums the weighted variation of a function on each edge of the directed graph. On the other hand, the initial label assignment should be changed as little as possible. Let y denote the function in $\mathcal{H}(V)$ defined by $y(v) = 1$ or -1 if vertex v has been labeled as positive or negative respectively, and 0 if it is unlabeled. Then we may consider the optimization problem

$$\underset{f\in\mathcal{H}(V)}{\operatorname{argmin}} \left\{ \Omega(f) + \mu \|f - y\|^2 \right\}, \tag{2}$$

where $\mu > 0$ is the parameter specifying the tradeoff between the two competitive terms.

We will provide the motivations for the functional defined by (1). In the end of this section, this functional will be compared with another choice which may seem more natural. The comparison may make us gain an insight into this functional. In Section 4, it will be shown that this functional may be naturally derived from a combinatorial optimization problem. In Section 5, we will further characterize this functional in terms of discrete analysis on directed graphs.

For an undirected graph $G = (V, E)$, it is well-known that the stationary distribution of the natural random walk has a closed-form expression $\pi(v) = d(v)/\operatorname{vol} V$, where $d(v)$ denotes the degree of the vertex v, and $\operatorname{vol} V = \sum_{u\in V} d(u)$. Substituting the closed-form expression into (1), we have

$$\Omega(f) = \frac{1}{2} \sum_{[u,v]\in E} w([u,v]) \left(\frac{f(u)}{\sqrt{d(u)}} - \frac{f(v)}{\sqrt{d(v)}} \right)^2,$$

which is the regularizer of the transductive inference algorithm of Zhou et al. (2004) operating on undirected graphs.

For solving the optimization problem (2), we introduce an operator $\Theta : \mathcal{H}(V) \to \mathcal{H}(V)$ defined by

$$\begin{aligned}(\Theta f)(v) \quad = \quad & \frac{1}{2}\Bigg(\sum_{u\to v} \frac{\pi(u)p(u,v)f(u)}{\sqrt{\pi(u)\pi(v)}} \\ & + \sum_{u\leftarrow v} \frac{\pi(v)p(v,u)f(u)}{\sqrt{\pi(v)\pi(u)}} \Bigg). \end{aligned} \tag{3}$$

Let Π denote the diagonal matrix with $\Pi(v,v) = \pi(v)$ for all $v \in V$. Let P denote the transition probability matrix and P^T the transpose of P. Then

$$\Theta = \frac{\Pi^{1/2} P \Pi^{-1/2} + \Pi^{-1/2} P^T \Pi^{1/2}}{2}. \tag{4}$$

Lemma 3.1. *Let $\Delta = I - \Theta$, where I denotes the identity operator. Then $\Omega(f) = \langle f, \Delta f\rangle$.*

Proof. The idea is to use summation by parts, a discrete analogue of the more common integration by parts.

$$\begin{aligned} & \sum_{[u,v]\in E} \pi(u)p(u,v) \left(\frac{f(u)}{\sqrt{\pi(u)}} - \frac{f(v)}{\sqrt{\pi(v)}} \right)^2 \\ = \quad & \frac{1}{2} \sum_{v\in V} \Bigg\{ \sum_{u\to v} \pi(u)p(u,v) \left(\frac{f(u)}{\sqrt{\pi(u)}} - \frac{f(v)}{\sqrt{\pi(v)}} \right)^2 \\ & + \sum_{u\leftarrow v} \pi(v)p(v,u) \left(\frac{f(v)}{\sqrt{\pi(v)}} - \frac{f(u)}{\sqrt{\pi(u)}} \right)^2 \Bigg\} \\ = \quad & \frac{1}{2} \sum_{v\in V} \Bigg\{ \sum_{u\to v} p(u,v) f^2(u) + \sum_{u\to v} \frac{\pi(u)p(u,v)}{\pi(v)} f^2(v) \\ & -2 \sum_{u\to v} \frac{\pi(u)p(u,v)f(u)f(v)}{\sqrt{\pi(u)\pi(v)}} \Bigg\} \\ & + \frac{1}{2} \sum_{v\in V} \Bigg\{ \sum_{u\leftarrow v} p(v,u) f^2(v) + \sum_{u\leftarrow v} \frac{\pi(v)p(v,u)}{\pi(u)} f^2(u) \\ & -2 \sum_{u\leftarrow v} \frac{\pi(v)p(v,u)f(v)f(u)}{\sqrt{\pi(v)\pi(u)}} \Bigg\} \end{aligned}$$

The first term on the right-hand side may be written

$$\begin{aligned} & \sum_{[u,v]\in E} p(u,v) f^2(u) = \sum_{u\in V} \sum_{v\leftarrow u} p(u,v) f^2(u) \\ = \quad & \sum_{u\in V} \left(\sum_{v\leftarrow u} p(u,v) \right) f^2(u) = \sum_{u\in V} f^2(u) = \sum_{v\in V} f^2(v), \end{aligned}$$

and the second term

$$\sum_{v\in V} \left(\sum_{u\to v} \frac{\pi(u)p(u,v)}{\pi(v)} \right) f^2(v) = \sum_{v\in V} f^2(v).$$

Similarly, for the fourth and fifth terms, we can show that

$$\sum_{v\in V} \sum_{u\leftarrow v} p(v,u) f^2(v) = \sum_{v\in V} f^2(v),$$

and

$$\sum_{v\in V} \sum_{u\leftarrow v} \frac{\pi(v)p(v,u)}{\pi(u)} f^2(u) = \sum_{v\in V} f^2(v).$$

respectively. Therefore,

$$\begin{aligned} \Omega(f) \quad = \quad & \sum_{v\in V} \Bigg\{ f^2(v) - \frac{1}{2} \Bigg(\sum_{u\to v} \frac{\pi(u)p(u,v)f(u)f(v)}{\sqrt{\pi(u)\pi(v)}} \\ & + \sum_{u\leftarrow v} \frac{\pi(v)p(v,u)f(v)f(u)}{\sqrt{\pi(v)\pi(u)}} \Bigg) \Bigg\}, \end{aligned}$$

which completes the proof. $\square$

Lemma 3.2. *The eigenvalues of the operator* Θ *are in* $[-1, 1]$, *and the eigenvector with the eigenvalue equal to 1 is* $\sqrt{\pi}$.

Proof. It is easy to see that Θ is *similar* to the operator $\Psi : \mathcal{H}(V) \to \mathcal{H}(V)$ defined by $\Psi = \left(P + \Pi^{-1}P^T\Pi\right)/2$. Hence Θ and Ψ have the same set of eigenvalues. Assume that f is the eigenvector of Ψ with eigenvalue λ. Choose a vertex v such that $|f(v)| = \max_{u\in V} |f(u)|$. Then we can show that $|\lambda| \leq 1$ by

$$\begin{aligned}
|\lambda||f(v)| &= \left|\sum_{u\in V} \Psi(v,u)f(u)\right| \leq \sum_{u\in V} \Psi(v,u)|f(v)| \\
&= \frac{|f(v)|}{2}\left(\sum_{u\leftarrow v} p(v,u) + \sum_{u\rightarrow v} \frac{\pi(u)p(u,v)}{\pi(v)}\right) \\
&= |f(v)|.
\end{aligned}$$

In addition, we can show that $\Theta\sqrt{\pi} = \sqrt{\pi}$ by

$$\begin{aligned}
&\frac{1}{2}\left(\sum_{u\rightarrow v} \frac{\pi(u)p(u,v)\sqrt{\pi(u)}}{\sqrt{\pi(u)\pi(v)}}\right. \\
&\qquad \left. + \sum_{u\leftarrow v} \frac{\pi(v)p(v,u)\sqrt{\pi(u)}}{\sqrt{\pi(v)\pi(u)}}\right) \\
=\ & \frac{1}{2}\left(\sum_{u\rightarrow v} \frac{\pi(u)p(u,v)}{\sqrt{\pi(v)}} + \sum_{u\leftarrow v} \frac{\pi(v)p(v,u)}{\sqrt{\pi(v)}}\right) \\
=\ & \frac{1}{2}\left(\frac{1}{\sqrt{\pi(v)}}\sum_{u\rightarrow v} \pi(u)p(u,v) + \sqrt{\pi(v)}\sum_{u\leftarrow v} p(v,u)\right) \\
=\ & \sqrt{\pi(v)}.
\end{aligned}$$

□

Theorem 3.3. *The solution of (2) is* $f^* = (1-\alpha)(I - \alpha\Theta)^{-1}y$, *where* $\alpha = 1/(1+\mu)$.

Proof. From Lemma 3.1, differentiating (2) with respect to function f, we have $(I-\Theta)f^* + \mu(f^* - y) = 0$. Define $\alpha = 1/(1+\mu)$. This system may be written $(I - \alpha\Theta)f^* = (1-\alpha)y$. From Lemma 3.2, we easily know that $(I - \alpha\Theta)$ is positive definite and thus invertible. This completes the proof. □

At the beginning of this section, we assume the graph to be strongly connected and aperiodic such that the natural random walk over the graph converges to a unique and positive stationary distribution. Obviously this assumption cannot be guaranteed for a general directed graph. To remedy this problem, we may introduce the so-called *teleporting random walk* (Page et al., 1998) as the replacement of the natural one. Given that we are currently at vertex u with $d^+(u) > 0$, the next step of this random walk proceeds as follows: (1) with probability $1-\eta$ jump to a vertex chosen uniformly at random over the whole vertex set except u; and (2) with probability $\eta w([u,v])/d^+(u)$ jump to a vertex v adjacent from u. If we are at vertex u with $d^+(u) = 0$, just jump to a vertex chosen uniformly at random over the whole vertex set except u.

Algorithm. Given a directed graph $G = (V, E)$ and a label set $\mathcal{Y} = \{1, -1\}$, the vertices in a subset $S \subset V$ are labeled. Then the remaining unlabeled vertices may be classified as follows:

1. Define a random walk over G with a transition probability matrix P such that it has a unique stationary distribution, such as the teleporting random walk.

2. Let Π denote the diagonal matrix with its diagonal elements being the stationary distribution of the random walk. Compute the matrix $\Theta = (\Pi^{1/2}P\Pi^{-1/2} + \Pi^{-1/2}P^T\Pi^{1/2})/2$.

3. Define a function y on V with $y(v) = 1$ or -1 if vertex v is labeled as 1 or -1, and 0 if v is unlabeled. Compute the function $f = (I - \alpha\Theta)^{-1}y$, where α is a parameter in $]0, 1[$, and classify each unlabeled vertex v as $\operatorname{sign} f(v)$.

It is worth mentioning that the approach of Zhou et al. (2005) can also be derived from this algorithmic framework by defining a two-step random walk. Assume a directed graph $G = (V, E)$ with $d^+(v) > 0$ and $d^-(v) > 0$ for all $v \in V$. Given that we are currently at vertex u, the next step of this random walk proceeds as follows: first jump backward to a vertex h adjacent to u with probability $p^-(u,h) = w([h,u])/d^-(u)$; then jump forward to a vertex v adjacent from u with probability $p^+(h,v) = w([h,v])/d^+(h)$. Thus the transition probability from u to v is $p(u,v) = \sum_{h\in V} p^-(u,h)p^+(h,v)$. It is easy to show that the stationary distribution of the random walk is $\pi(v) = d^-(v)/\sum_{u\in V} d^-(u)$ for all $v \in V$. Substituting the quantities of $p(u,v)$ and $\pi(v)$ into (1), we then recover one of the two regularizers proposed by Zhou et al. (2005). The other one can also be recovered simply by reversing this two-step random walk.

Now we discuss implementation issues. The closed form solution shown in Theorem 3.3 involves a matrix inverse. Given an $n \times n$ invertible matrix A, the time required to compute the inverse A^{-1} is generally $O(n^3)$ and the representation of the inverse requires $\Omega(n^2)$ space. Recent progress in numerical analysis (Spielman & Teng, 2003), however, shows that, for an $n \times n$ symmetric positive semi-definite, *diagonally dominant* matrix A with m non-zero entries and a n-vector b, we can obtain a vector $\tilde{x}$ within relative distance ϵ of the solution to $Ax = b$ in time $O(m^{1.31} \log(n\kappa_f(A)/\epsilon)^{O(1)})$, where $\kappa_f(A)$ is the log of the ratio of the largest to smallest non-zero eigenvalue of A. It can be shown that our approach can benefit from this numerical technique. From Theorem 3.3,

$$\left(I - \alpha \frac{\Pi^{1/2} P \Pi^{-1/2} + \Pi^{-1/2} P^T \Pi^{1/2}}{2}\right) f^* = (1-\alpha) y,$$

which may be transformed into

$$\left(\Pi - \alpha \frac{\Pi P + P^T \Pi}{2}\right) (\Pi^{-1/2} f^*) = (1-\alpha) \Pi^{1/2} y.$$

Let $A = \Pi - \alpha \dfrac{\Pi P + P^T \Pi}{2}$. It is easy to verify that A is diagonally dominant.

For well understanding this regularization framework, we may compare it with an alternative approach in which the regularizer is defined by

$$\Omega(f) = \sum_{[u,v] \in E} w([u,v]) \left(\frac{f(u)}{\sqrt{d^+(u)}} - \frac{f(v)}{\sqrt{d^-(v)}} \right)^2. \tag{5}$$

A similar closed form solution can be obtained from the corresponding optimization problem. Clearly, for undirected graphs, this functional also reduces to that in (Zhou et al., 2004). At first glance, this functional may look natural, but in the later experiments we will show that the algorithm based on this functional does not work as well as the previous one. This is because the directionality is only slightly taken into account by this functional via the degree normalization such that much valuable information for classification conveyed by the directionality is ignored by the corresponding algorithm. Once we remove the degree normalization from this functional, the resulted functional is totally insensitive to the directionality.

4. Directed Spectral Clustering

In the absence of labeled instances, this framework can be utilized in an unsupervised setting as a spectral clustering method for directed graphs. We first define a combinational partition criterion, which generalizes the normalized cut criterion for undirected graphs (Shi & Malik, 2000). Then relaxing the combinational optimization problem into a real-valued one leads to the functional defined in Section 3.

Given a subset S of the vertices of a directed graph G, define the volume of S by $\operatorname{vol} S := \sum_{v \in S} \pi(v)$. Clearly,

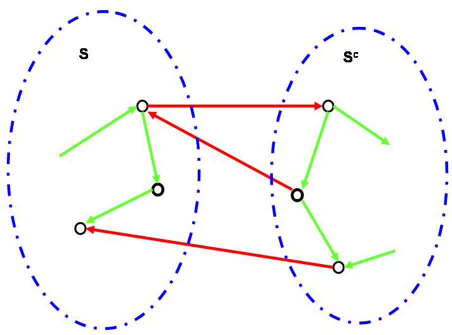


Figure 2. A subset S and its complement S^c. Note that there is only one edge in the out-boundary of S.

$\operatorname{vol} S$ is the probability with which the random walk occupies some vertex in S and consequently $\operatorname{vol} V = 1$. Let S^c denote the complement of S (Fig. 2). The *out-boundary* ∂S of S is defined by $\partial S := \{[u,v] | u \in S, v \in S^c\}$. The value $\operatorname{vol} \partial S := \sum_{[u,v] \in \partial S} \pi(u) p(u,v)$ is called the volume of ∂S. Note that $\operatorname{vol} \partial S$ is the probability with which one sees a jump from S to S^c.

Generalizing the normalized cut criterion for undirected graphs is based on a key observation stated by

Proposition 4.1. $\operatorname{vol} \partial S = \operatorname{vol} \partial S^c$.

Proof. It immediately follows from that the probability with which the random walk leaves a vertex equals the probability with which the random walk arrives at this vertex. Formally, for each vertex v in V, it is easy to see that

$$\sum_{u \to v} \pi(u) p(u,v) - \sum_{u \leftarrow v} \pi(v) p(v,u) = 0.$$

Summing the above equation over the vertices of S (see also Fig. 2), then we have

$$\begin{aligned} &\sum_{v \in S} \left(\sum_{u \to v} \pi(u) p(u,v) - \sum_{u \leftarrow v} \pi(v) p(v,u) \right) \\ = &\sum_{[u,v] \in \partial S^c} \pi(u) p(u,v) - \sum_{[u,v] \in \partial S} \pi(u) p(u,v) = 0, \end{aligned}$$

which completes the proof. □

From Proposition 4.1, we may partition the vertex set of a directed graph into two nonempty parts S and S^c by minimizing

$$\operatorname{Ncut}(S) = \operatorname{vol} \partial S \left(\frac{1}{\operatorname{vol} S} + \frac{1}{\operatorname{vol} S^c} \right), \tag{6}$$

which is a directed generalization of the normalized cut criterion for undirected graphs. Clearly, the ratio of $\operatorname{vol} \partial S$ to $\operatorname{vol} S$ is the probability with which the random walk leaves S in the next step under the condition that it is in fact in S now. Similarly understand the ratio of $\operatorname{vol}\partial S^c$ to $\operatorname{vol} S^c$.

In the following, we show that the functional (1) can be recovered from (6). Define an indicator function $h \in \mathcal{H}(V)$ by $h(v) = 1$ if $v \in S$, and -1 if $v \in S^c$. Denote by ν the volume of S. Clearly, we have $0 < \nu < 1$ due to $S \subset G$. Then (6) may be written

$$\mathrm{Ncut}(S) = \frac{\sum\limits_{[u,v]\in E} \pi(u)p(u,v)(h(u)-h(v))^2}{8\nu(1-\nu)}.$$

Define another function $g \in \mathcal{H}(V)$ by $g(v) = 2(1-\nu)$ if $v \in S$, and -2ν if $v \in S^c$. We easily know that $\operatorname{sign} g(v) = \operatorname{sign} h(v)$ for all $v \in V$ and $h(u) - h(v) = g(u) - g(v)$ for all $u, v \in V$. Moreover, it is not hard to see that $\sum_{v\in V} \pi(v)g(v) = 0$, and $\sum_{v\in V} \pi(v)g^2(v) = 4\nu(1-\nu)$. Therefore

$$\mathrm{Ncut}(S) = \frac{\sum\limits_{[u,v]\in E} \pi(u)p(u,v)(g(u)-g(v))^2}{2\sum\limits_{v\in V} \pi(v)g^2(v)}.$$

Define another function $f = \sqrt{\pi}g$. Then the above equation may be further transformed into

$$\mathrm{Ncut}(S) = \frac{\sum\limits_{[u,v]\in E} \pi(u)p(u,v)\left(\frac{f(u)}{\sqrt{\pi(u)}} - \frac{f(v)}{\sqrt{\pi(v)}}\right)^2}{2\langle f, f\rangle}.$$

If we allow the function f to take arbitrary real values, then the graph partition problem (6) becomes

$$\underset{f\in\mathcal{H}(V)}{\operatorname{argmin}}\, \Omega(f) \qquad (7)$$
$$\text{subject to} \quad \|f\| = 1,\ \langle f, \sqrt{\pi}\rangle = 0.$$

From Lemma 3.2, it is easy to see that the solution of (7) is an eigenvector of the operator Θ corresponding the second largest eigenvalue.

Algorithm. Given a directed graph $G = (V, E)$, it may be partitioned into two parts as follows:

1. Define a random walk over G with a transition probability matrix P such that it has a unique stationary distribution.

2. Let Π denote the diagonal matrix with its diagonal elements being the stationary distribution of the random walk. Form the matrix $\Theta = (\Pi^{1/2}P\Pi^{-1/2} + \Pi^{-1/2}P^T\Pi^{1/2})/2$.

3. Compute an eigenvector Φ of Θ corresponding to the second largest eigenvalue, and then partition the vertex set V of G into the two parts $S = \{v \in V|\Phi(v) \geq 0\}$ and $S^c = \{v \in V|\Phi(v) < 0\}$.

It is easy to extend this approach to k-partition. Assume a k-partition to be $V = V_1 \cup V_2 \cup \cdots \cup V_k$, where $V_i \cap V_j = \emptyset$ for all $1 \leq i, j \leq k$. Let P_k denote a k-partition. Then we may obtain a k-partition by minimizing

$$\mathrm{Ncut}(P_k) = \sum_{1\leq i\leq k} \frac{\operatorname{vol}\partial V_i}{\operatorname{vol} V_i}. \qquad (8)$$

It is not hard to show that the solution of the corresponding relaxed optimization problem of (8) can be any orthonormal basis for the linear space spanned by the eigenvectors of Θ pertaining to the k largest eigenvalues.

5. Discrete Analysis

We develop discrete analysis on directed graphs. The regularization framework in Section 3 is then reconstructed and generalized using discrete analysis. This work is the discrete analogue of classic regularization theory (Tikhonov & Arsenin, 1977; Wahba, 1990).

We define the *graph gradient* to be an operator $\nabla : \mathcal{H}(V) \to \mathcal{H}(E)$ which satisfies

$$(\nabla f)([u,v]) := \sqrt{\pi(u)p(u,v)}\left(\frac{f(v)}{\sqrt{\pi(v)}} - \frac{f(u)}{\sqrt{\pi(u)}}\right). \qquad (9)$$

For an undirected graph, equation (9) reduces to

$$(\nabla f)([u,v]) = \sqrt{\frac{w([u,v])}{d(v)}}f(v) - \sqrt{\frac{w([u,v])}{d(u)}}f(u).$$

We define the graph gradient of function f at each vertex v as $\nabla f(v) := \{(\nabla f)([v,u])|[v,u] \in E\}$, which is also denoted by $\nabla_v f$. Then the norm of the graph gradient ∇f at v is defined by

$$\|\nabla_v f\| := \left(\sum_{u\leftarrow v} (\nabla f)^2([v,u])\right)^{\frac{1}{2}}, \qquad (10)$$

and the *p-Dirichlet form* of function f is defined by

$$\Omega_p(f) := \frac{1}{2}\sum_{v\in V} \|\nabla_v f\|^p,\ p \in [1, \infty[. \qquad (11)$$

Note that $\Omega_2(f) = \Omega(f)$. Intuitively, the norm of the graph gradient measures the smoothness of a function around a vertex, and the p-Dirichlet form the smoothness of a function over the whole graph. In addition, we define $\|\nabla f([v,u])\| := \|\nabla_v f\|$. Note that $\|\nabla f\|$ is defined in the space $\mathcal{H}(E)$ as $\|\nabla f\| = \langle \nabla f, \nabla f\rangle^{1/2}_{\mathcal{H}(E)}$.

We define the *graph divergence* to be an operator $\operatorname{div} : \mathcal{H}(E) \to \mathcal{H}(V)$ which satisfies

$$\langle \nabla f, g\rangle_{\mathcal{H}(E)} = \langle f, -\operatorname{div} g\rangle_{\mathcal{H}(V)} \qquad (12)$$

for any two functions f and g in $\mathcal{H}(E)$. Equation (12) is the discrete analogue of the Stokes' theorem [1]. It is not hard to show that

$$(\operatorname{div} g)(v) = \frac{1}{\sqrt{\pi(v)}}\left(\sum_{u \leftarrow v} \sqrt{\pi(v)p(v,u)}\,g([v,u]) - \sum_{u \rightarrow v} \sqrt{\pi(u)p(u,v)}\,g([u,v])\right). \quad (13)$$

Intuitively, we may think of the graph divergence $(\operatorname{div} g)(v)$ as the net outflow of the function g at the vertex v. For a function $c : E \rightarrow \mathbb{R}$ defined by $c([u,v]) = \sqrt{\pi(u)p(u,v)}$, it follows from equation (13) that $(\operatorname{div} c)(v) = 0$ at any vertex v in V.

We define the *graph Laplacian* to be an operator $\Delta : \mathcal{H}(V) \rightarrow \mathcal{H}(V)$ which satisfies [2]

$$\Delta f := -\frac{1}{2}\operatorname{div}(\nabla f). \quad (14)$$

Clearly, Δ is linear and self-adjoint. In addition, we have $\Omega_2(f) = \langle f, \Delta f\rangle$, which implies that Δ is positive semi-definite. From (9) and (13), we have

$$(\Delta f)(v) = f(v) - \frac{1}{2}\left(\sum_{u \rightarrow v} \frac{\pi(u)p(u,v)f(u)}{\sqrt{\pi(u)\pi(v)}} + \sum_{u \leftarrow v} \frac{\pi(v)p(v,u)f(u)}{\sqrt{\pi(v)\pi(u)}}\right). \quad (15)$$

In matrix notation, Δ can be written as

$$\Delta = I - \frac{\Pi^{1/2}P\Pi^{-1/2} + \Pi^{-1/2}P^T\Pi^{1/2}}{2}, \quad (16)$$

which is just the Laplacian matrix for directed graphs proposed by Chung (2005). For an undirected graph, equation (16) clearly reduces to the Laplacian matrix for undirected graphs (Chung, 1997).

We define the *graph p-Laplacian* to be an operator $\Delta_p : \mathcal{H}(V) \rightarrow \mathcal{H}(V)$ which satisfies

$$\Delta_p f := -\frac{1}{2}\operatorname{div}(\|\nabla f\|^{p-2}\nabla f). \quad (17)$$

Clearly, $\Delta_2 = \Delta$, and $\Delta_p (p \neq 2)$ is nonlinear. In particular, the case $p = 1$ is the discrete analogue of the mean curvature operator. Now let us consider the general optimization problem

$$\operatorname*{argmin}_{f \in \mathcal{H}(V)} \left\{\Omega_p(f) + \mu\|f - y\|^2\right\}. \quad (18)$$

[1]Given a compact Riemannian manifold (M, g) with a function $f \in C^\infty(M)$ and a vector field $X \in \mathcal{X}(M)$, it follows from the Stokes' theorem that $\int_M \langle \nabla f, X\rangle = -\int_M (\operatorname{div} X) f$.

[2]The Laplace-Beltrami operator $\Delta : C^\infty(M) \rightarrow C^\infty(M)$ is defined by $\Delta f = -\operatorname{div}(\nabla f)$. The additional factor 1/2 in (14) is due to edges being oriented.

Let f^* denote the solution of (18). It is not hard to show that

$$p\Delta_p f^* + 2\mu(f^* - y) = 0. \quad (19)$$

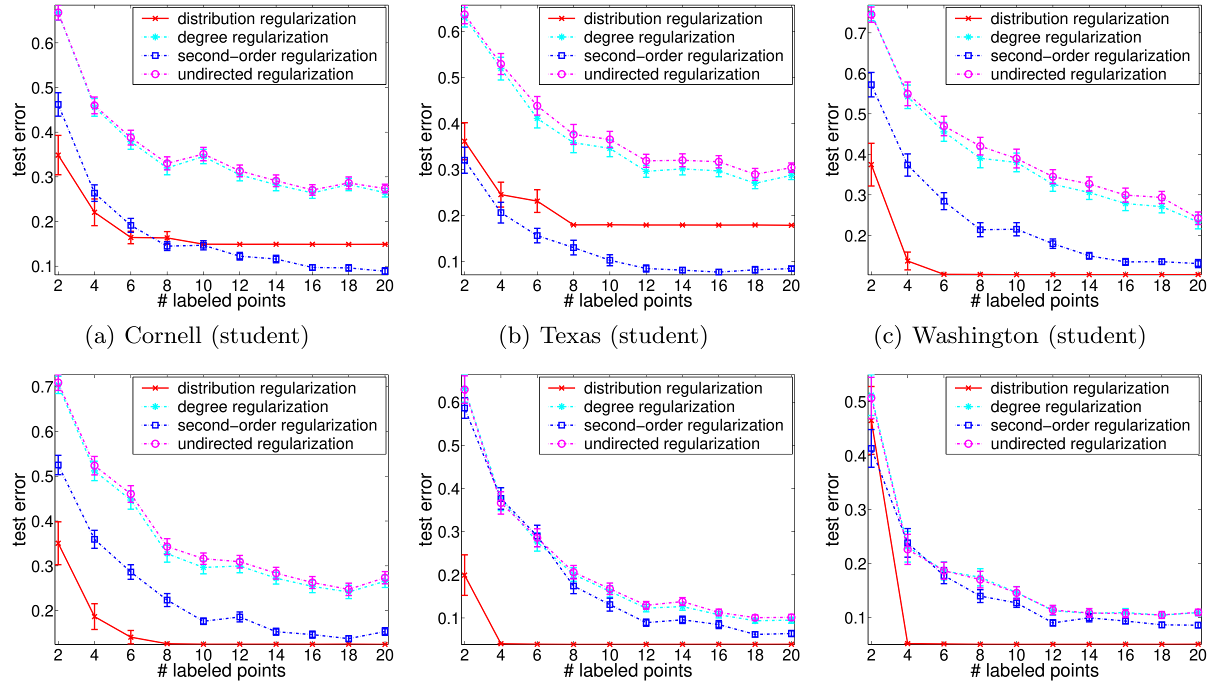
When $p = 2$, as we have shown before, the above equation leads to the closed-form solution in Theorem 3.3. In other cases, we are not aware of any closed-form solution.

6. Experiments

We address the web categorization task on the WebKB dataset (see http://www-2.cs.cmu.edu/~webkb/). We only consider a subset containing the pages from the four universities Cornell, Texas, Washington and Wisconsin, from which we remove the isolated pages, i.e., the pages which have no incoming and outgoing links, resulting in 858, 825, 1195 and 1238 pages respectively, for a total of 4116. These pages have been manually classified into the following seven categories: *student*, *faculty*, *staff*, *department*, *course*, *project* and *other*. We may assign a weight to each hyperlink according to the textual content or the anchor text. However, here we are only interested in how much we can obtain from link structure only and hence adopt the canonical weight function.

We compare the approach in Section 3 with its counterpart based on equation (5). Moreover, we also compare both methods with the schemes in (Zhou et al., 2005; Zhou et al., 2004). For the last approach, we transform a directed graph into an undirected one by defining a symmetric weight function as $w([u,v]) = 1$ if $[u,v]$ or $[v,u]$ in E. To distinguish among these approaches, we refer to them as *distribution regularization*, *degree regularization*, *second-order regularization* and *undirected regularization* respectively. As we have shown, both the distribution and degree regularization approaches are generalizations of the undirected regularization method.

The investigated task is to discriminate the student pages in a university from the non-student pages in the same university. We further remove the isolated pages in each university. Other categories including faculty and course are considered as well. For all approaches, the regularization parameter is set to $\alpha = 0.1$ as in (Zhou et al., 2005). In the distribution regularization approach, we adopt the teleporting random walk with a small jumping probability $\eta = 0.01$ for obtaining a unique and positive stationary distribution. The testing errors are averaged over 50 trials. In each trial, it is randomly decided which of the training points get labeled. If there is no labeled point existed for some class, we sample again. The experimental re-



(a) Cornell (student) (b) Texas (student) (c) Washington (student)

(d) Wisconsin (student) (e) Cornell (faculty) (f) Cornell (course)

Figure 3. Classification on the WebKB dataset. Fig. (a)-(d) depict the test errors of the regularization approaches on the classification problem of student vs. non-student in each university. Fig. (e)-(f) illustrate the test errors of these methods on the classification problems of faculty vs. non-faculty and course vs. non-course in Cornell University.

sults are shown in Fig. 3. The distribution regularization approach shows significantly improved results in comparison to the degree regularization method. Furthermore, the distribution regularization approach is comparable with the second-order regularization one. In contrast, the degree regularization approach shows similar performance to the undirected regularization one. Therefore we can conclude that the degree regularization approach almost does not take the directionality into account.

References

Chung, F. (1997). *Spectral graph theory*. No. 92 in CBMS Regional Conference Series in Mathematics. American Mathematical Society, Providence, RI.

Chung, F. (2005). Laplacian and the Cheeger inequality for directed graphs. *Annals of Combinatorics*, *9*, 1–19.

Getoor, L., Friedman, N., Koller, D., & Taskar, B. (2002). Learning probabilistic models of link structure. *Journal of Machine Learning Research*, *3*, 679–708.

Henzinger, M. (2001). Hyperlink analysis for the web. *IEEE Internet Computing*, *5*, 45–50.

Henzinger, M. (2003). Algorithmic challenges in web search engines. *Internet Mathematics*, *1*, 115–123.

Kleinberg, J. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, *46*, 604–632.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1998). *The PageRank citation ranking: bring order to the web* (Technical Report). Stanford University.

Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*, 888–905.

Spielman, D., & Teng, S. (2003). Solving sparse, symmetric, diagonally-dominant linear systems in time $o(m^{1.31})$. *Proc. 44th Annual IEEE Symposium on Foundations of Computer Science*.

Tikhonov, A., & Arsenin, V. (1977). *Solutions of ill-posed problems*. W. H. Winston, Washington, DC.

Vapnik, V. (1998). *Statistical learning theory*. Wiley, NY.

Wahba, G. (1990). *Spline models for observational data*. No. 59 in CBMS-NSF Regional Conference Series in Applied Mathematics. SIAM, Philadelphia.

Zhou, D., Bousquet, O., Lal, T., Weston, J., & Schölkopf, B. (2004). Learning with local and global consistency. *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA.

Zhou, D., Schölkopf, B., & Hofmann, T. (2005). Semi-supervised learning on directed graphs. *Advances in Neural Information Processing Systems 17*. MIT Press, Cambridge, MA.