Learning Random Walks to Rank Nodes in Graphs

Alekh Agarwal ALEKH@CSE.IITB.AC.IN
Soumen Chakrabarti SOUMEN@CSE.IITB.AC.IN
IIT Bombay

Abstract

Ranking nodes in graphs is of much recent interest. Edges, via the graph Laplacian, are used to encourage local smoothness of node scores in SVM-like formulations with generalization guarantees. In contrast, Pagerank variants are based on Markovian random walks. For directed graphs, there is no simple known correspondence between these views of scoring/ranking. Recent scalable algorithms for learning the Pagerank transition probabilities do not have generalization guarantees. In this paper we show some correspondence results between the Laplacian and the Pagerank approaches, and give new generalization guarantees for the latter. We enhance the Pagerank-learning approaches to use an additive margin. We also propose a general framework for rank-sensitive score-learning, and apply it to Laplacian smoothing. Experimental results are promising.

1. Introduction

Learning to rank is of much recent interest. A series of papers (Herbrich et al., 1999; Joachims, 2002; Burges et al., 2005), and even a recent NIPS workshop (Agarwal et al., 2005), are dedicated to ranking instances represented as feature vectors in some feature space. A few variants have been studied: *ordinal regression*, where an instance is assigned a label from an ordered k-level scale; *bipartite ranking*, where instances are relevant or irrelevant, and the job is to rank relevant instances before any irrelevant ones (2-level ordinal regression); and learning from arbitrary *preference pairs* $u \prec v$, meaning that u should be ranked lower than v.

A significant complication is added by the presence of *relationships* (represented by edges in a graph) between the instances (represented by nodes). Labeling of nodes in graphical models is very well-studied (Taskar, 2004; Zhou et al., 2005), but the interest in ranking, motivated partly by Web (Joachims, 2002), XML and database search (Balmin et al., 2004), is relatively nascent. We describe two approaches below. These have known correspondences for undirected graphs but not directed graphs.

Associative networks: The first approach (Agarwal, 2006) adopts the associative Markov network (Taskar, 2004) philosophy: ranks are induced by scores assigned to nodes, and edges hint that the scores must be locally smooth. As in spectral graph partitioning, the smoothness is encouraged via a quadratic penalty term involving the Laplacian of G (Zhou & Schölkopf, 2004). This can be shown (Agarwal, 2006) as equivalent to regularizing the scoring function in an RKHS, which enables drawing on (Agarwal & Niyogi, 2005) to show elegant generalization bounds for bipartite ranking.

Random walks: The second approach involves scoring nodes using the stationary distribution of Markovian random walks (Brin & Page, 1998), and is by far the most popular approach in search applications. In this setting, a directed edge indicates *endorsement*, not necessarily similarity of scores; e.g., thousands of obscure pages link to `www.kernel-machines.org`. Earlier, the transition probabilities used to be tuned by hand; recently, we (Agarwal et al., 2006) proposed methods (henceforth NetRank) to learn the transition parameters from pairwise preferences. Pagerank is modeled as a flow p, and the algorithm minimizes the KL divergence from p to a "reference" flow q, so that p satisfies Markovian balance constraints as well as preference constraints $\prec$. Unfortunately, these methods had no theoretical guarantees of generalization.

Our contributions: Our primary contribution is to consolidate and extend these two views of learning to rank in graphs. We first show two correspondence results in Section 3. Although local smoothness was the guiding concern in using graph Laplacians, they are closely related to Markovian walks for both directed (Chung, 2005) and (trivially) undirected graphs. We show that a Laplacian-based regularizer indeed seeks to preserve Pagerank-based node orders in the absence of preferences. Conversely, if NetRank achieves a low KL divergence, then the Laplacian roughness penalty is low as well.

Using the stability framework of Bousquet and Elis-

Appearing in *Proceedings of the 24th International Conference on Machine Learning*, Corvallis, OR, 2007.

seeff (2002), we next demonstrate, in Section 4, generalization capability of minor variants of NetRank; this gives theoretical justification for their approach. Moreover, we enhance the formulation with an additive margin in Section 5, which improves its accuracy considerably in practice.

In search applications, the perceived quality of ranking algorithms is very sensitive to the top ranks (Matveeva et al., 2006; Rudin, 2006; Burges et al., 2006). Unfortunately, minimizing rank-sensitive loss functions appears much harder than minimizing (a bound on) the number of violated constraints (Joachims, 2002). Our final contribution in Section 6 is a general framework for approaching rank-sensitive losses armed with only pairwise preferences. We demonstrate it with the Laplacian smoothing approach, and generalization bounds from Section 4 carry over.

Benefits: Pagerank variants are widely used for ranking nodes in graphs. Markov balance constraints (Section 2) and linearity have helped harness the massive literature on graph eigensystems to design highly scalable Pagerank algorithms (Jeh & Widom, 2003). But Pagerank transitions are designed by hand, hardly ever learnt from $\prec$. NetRank, as it is, does not give generalization guarantees, and, indeed, does not generalize well in experiments. On the other hand, balance is not ensured by the Laplacian smoothing approach (Agarwal, 2006) which gives formal generalization guarantees. Laplacian smoothing not only involves diagonalizing a large matrix, but it also assigns arbitrary scores to nodes, thus inducing all possible permutations. In contrast, for a given graph, certain node orders may be impossible to achieve via Pagerank. Therefore, the hypothesis space of Pagerank is contained in the hypothesis space of the Laplacian smoothing approach. In preliminary experiments (Section 7), it appears that this increased bias does aid generalization. Given our new evidence of generalization, and our new enhancements to use additive margin and cost/rank-sensitive learning, learning Pagerank flows compares favorably, as a general technique, to Laplacian-based smoothing.

2. Preliminaries and previous work

We set up some notation. The graph is $G = (V, E)$. Instances are nodes, denoted u, v, etc., and also interpreted as matrix/vector indexes in $\{1, \ldots, |V|\}$. A preference pair is written as "$u \prec v$", meaning u is less preferred than v, and should rank lower. For convenience, we use $\prec$ as both a relation and a set. In this paper we do not associate feature vectors with nodes, but the model we use has been extended to incorporate node features such as text (Balmin et al., 2004).

G has an associated fixed $|V| \times |V|$ transition probability matrix Q, with $Q(u, v) = \Pr(u \to v)$. We assume G has no dead-end nodes. In standard Pagerank $Q(u, v)$ is the reciprocal of the out-degree of u. Each row of Q sums to 1. In each step in the Pagerank random walk model (Brin & Page, 1998), the "random surfer" walks to some neighbor with probability $0 < \alpha < 1$ and teleports with probability $1-\alpha$. A $|V| \times 1$ teleport vector r determines the probability of jumping to a specific u in case of a teleport. Here we assume uniform $r = \vec{1}/|V|$. The steady-state node visit probabilities $\pi \in \mathbb{R}^{|V| \times 1}$ are given by $\pi = \alpha Q^\top \pi + (1-\alpha)\frac{\vec{1}}{|V|}$, which solves to $\pi = (1-\alpha)(\mathbb{I} - \alpha Q^\top)^{-1}\frac{\vec{1}}{|V|}$. For convenience, teleport is often implemented using a dummy node d with (u, d) and (d, u) transitions for each ordinary node u. The resulting graph $\hat{G} = (\hat{V} = V \cup d, \hat{E})$ has transition matrix

$$\hat{Q} = \begin{bmatrix} \alpha Q & (1-\alpha)\vec{1}_{|V| \times 1} \\ \vec{1}/|V| & 0 \end{bmatrix} \in \mathbb{R}^{n \times n} \text{ with } n = |V| + 1$$

and the steady-state $\hat{\pi}$ that satisfies $\hat{\pi} = \hat{Q}^\top \hat{\pi}$ is closely related to π (Langville & Meyer, 2004). While Q and $\hat{Q}$ are very sparse in practice, $(\mathbb{I} - \alpha Q^\top)^{-1}$ is large, dense and never computed explicitly. Henceforth we shall talk about random walks in $\hat{G}$, but continue to use Q and π for notational simplicity. Let $\Pi = \text{diag}(\pi)$.

π and Q induce a *reference circulation* q along each edge of $\hat{G}$, given by $q_{uv} = \pi(u)Q(u, v)$. Note that $\sum_{u,v} q_{uv} = 1$. In NetRank, we seek to optimize a circulation $\{p_{uv}\}$ which stays close to q in terms of KL divergence while trying to satisfy $\prec$:

$$\min_{\substack{\{0 \le p_{uv}\} \\ \{0 \le s_{uv}\}}} \quad \sum_{(u,v) \in \hat{E}} p_{uv} \log \frac{p_{uv}}{q_{uv}} + C \sum_{u \prec v} s_{uv} \qquad \text{(KL)}$$

$$\text{s.t.:} \quad \sum_{(u,v) \in \hat{E}} p_{uv} = 1 \qquad (1)$$

$$\forall v \in \hat{V}: \quad \sum_{(u,v) \in \hat{E}} p_{uv} - \sum_{(v,w) \in \hat{E}} p_{vw} = 0$$

$$\forall v \in V: \quad -\alpha p_{vd} + (1-\alpha) \sum_{(u,v) \in \hat{E}} p_{uv} = 0$$

$$\forall u \prec v: \quad \sum_{(w,u) \in \hat{E}} p_{wu} - \sum_{(w,v) \in \hat{E}} p_{wv} - s_{uv} \le 0 \qquad (2)$$

Note that (2) includes no margin; we will visit this issue in Section 5. The solution p will induce a $n \times 1$ score vector ϕ, where $\phi(v) = \sum_{(u,v) \in \hat{E}} p_{uv}$.

In contrast, from the associative network viewpoint, edge (u, v) connotes similarity, carefully encoded using domain knowledge as a fixed weight $w(u, v)$; in general $w(u, v) \neq w(v, u)$. The total outgoing weight of a node u is $\omega(u) = \sum_{(u,v) \in \hat{E}} w(u, v)$. We assume all $\omega(u) > 0$. W is used to design a *fixed* matrix Q with $Q(u, v) = w(u, v)/\omega(u)$. Q induces $\hat{Q}$, π and Π as before.

Agarwal (2006) assigns to each node u a score $f_u \in \mathbb{R}$, so as to minimize an objective function that balances violations of $\prec$ against the norm of f in a suitable RKHS:

$$\min_{\substack{f:V\to\mathbb{R} \\ s=\{s_{uv}\geq 0: u\prec v\}}} \quad \tfrac{1}{2} f^\top L f + B\sum_{u\prec v} s_{uv} \quad \text{subject to} \qquad \text{(Lap)}$$
$$f_v - f_u \geq 1 - s_{uv} \quad \forall u \prec v$$

L is the directed graph Laplacian fixed by $\hat{Q}$ and Π (which is itself determined by $\hat{Q}$):

$$L = \mathbb{I} - \tfrac{1}{2}(\Pi^{1/2}\hat{Q}\Pi^{-1/2} + \Pi^{-1/2}\hat{Q}^\top\Pi^{1/2}). \quad (3)$$

Optimizing (Lap) requires the computation and storage of the large and dense pseudoinverse L^+. Note that, in general, the solution f need not correspond to any flow, unlike p or q above.

While (Lap) clearly employs L for local smoothness, the involvement of Π in L in (3) hints that there may be connections between solutions of (Lap) and (KL). Unlike in undirected graphs, the connection is not immediate.

3. Laplacian-KL correspondence

Fact 1. *If $\hat{Q}_{(\mathrm{KL})} = \hat{Q}_{(\mathrm{Lap})}$ and $\prec = \varnothing$, the optimal solutions to both* (KL) *and* (Lap) *will order nodes v in decreasing order of $\pi(v)$.*

Proof. If $\prec$ is empty, the objective of (KL) reduces to $\mathrm{KL}(p\|q)$, which is minimized for $p = q$. The regularizer in (Lap), which uses the directed Laplacian of (Chung, 2005), can be rewritten as

$$f^\top L f = \sum_{\{u,v\}\in\hat{E}} \pi(u)\hat{Q}_{uv}\left(\frac{f(u)}{\sqrt{\pi(u)}} - \frac{f(v)}{\sqrt{\pi(v)}}\right)^2$$

It is easy to see that this regularizer is minimized when $f_v \propto \sqrt{\pi(v)}$. □

Therefore, even though the hypothesis space of the associative network view contains that of the random walk view, they coincide in their parsimonious beliefs in the absence of training data.

Next we show a more nontrivial property: any flow p that is "close to" q also induces a smooth scoring function on the nodes of the graph. Following the notation of Smola and Kondor (2003), let $(\lambda_i, \psi_i), i = 1, \ldots, n$ be the spectrum of L with $2 \geq \lambda_1 \geq \lambda_2 \cdots \geq \lambda_n = 0$. Let $r(\lambda)$ be a positive and monotonically increasing function. Define $r(L) = \sum_{i=1}^n r(\lambda_i)\psi_i\psi_i^\top$.

Theorem 2. *Let $\hat{Q}_{(\mathrm{KL})} = \hat{Q}_{(\mathrm{Lap})}$ and p be a valid flow on G, and let q be the reference flow induced by π. Let $f_p(u) = \sqrt{\sum_{\{w:(w,u)\in\hat{E}\}} p_{wu}}$. Then*

$$\mathrm{KL}(p\|q) \leq \epsilon \;\Rightarrow\; f_p^\top r(L) f_p \leq r(2)(1 + (2\epsilon\ln 2)^{1/4})^2.$$

We first quote a useful result from Information Theory (Cover & Thomas, 1991, Lemma 12.6.1).

Lemma 3. *Let p and q be two probability distributions over the same sigma algebra. Then* $\mathrm{KL}(p\|q) \geq \frac{1}{2\ln 2}\|p - q\|_1^2$.

Proof of Theorem 2. First note that $\pi^{1/2}$ is an eigenvector of L with eigenvalue 0, i.e., $\psi_n = \pi^{1/2}$. From Lemma 3, we have that $\mathrm{KL}(p\|q) \leq \epsilon \Rightarrow \|p - q\|_1^2 \leq 2\epsilon\ln 2$. Let $\epsilon_1^2 = 2\epsilon\ln 2$. Then ϵ_1

$$\geq \sum_{(u,v)\in\hat{E}} |p_{uv} - q_{uv}| = \sum_{u\in\hat{V}} \sum_{v:(u,v)\in\hat{E}} |p_{uv} - q_{uv}|$$
$$\geq \sum_{u\in\hat{V}} \left|\textstyle\sum_{v:(u,v)\in\hat{E}} (p_{uv} - q_{uv})\right| = \sum_{u\in\hat{V}} \left|f_p(u)^2 - \pi(u)\right|$$
$$\geq \sum_{u\in\hat{V}} \left|f_p(u) - \sqrt{\pi(u)}\right|^2 = \left(f_p - \sqrt{\pi}\right)^\top \left(f_p - \sqrt{\pi}\right)$$

We now write f_p in terms of the basis spanned by eigenvectors of L, $f_p = \sum_{i=1}^n c_i\psi_i$, and continue:

$$\epsilon_1 \geq \left(\sum_{i=1}^n c_i\psi_i - \psi_n\right)^\top \left(\sum_{i=1}^n c_i\psi_i - \psi_n\right)$$
$$= \sum_{i=1}^{n-1} c_i^2 + (c_n - 1)^2 \quad (4)$$
$$\geq (c_n - 1)^2, \quad \therefore c_n \leq 1 + \sqrt{\epsilon_1} \quad (5)$$

Because all $\lambda_i \leq 2$, we have $f^\top r(L) f =$

$$\sum_{i=1}^n r(\lambda_i)c_i^2 \leq r(2)\left(\sum_{i=1}^{n-1} c_i^2 + (c_n - 1)^2 + 2c_n - 1\right)$$
$$\leq r(2)\left(\epsilon_1 + 2\sqrt{\epsilon_1} + 1\right) = r(2)(1 + \sqrt{\epsilon_1})^2,$$

using (4) and (5). □

Therefore, minimizing KL divergence from p to q amounts to searching for a smooth scoring function. While the optimization proposed in NetRank was intuitive, Theorem 2 gives it theoretical justification.

4. Stability and generalization

It is reassuring that optimization (KL) seeks to also smooth f_p wrt L similar to (Lap), but, ideally, we prefer a direct generalization proof for (KL). In this section we will derive relative generalization bounds using the Algorithmic Stability framework of Bousquet and Elisseeff (2002). (In what follows, $R, R_{\mathrm{emp}}, R_{\mathrm{reg}}$ are true, empirical and regularized risks over preference pairs.) For convenience we modify the regularized objective (KL) superficially to line up with their notation (their N is our KL):

$$R_{\mathrm{reg}}(p) = \underbrace{\tfrac{1}{m}\textstyle\sum_{j=1}^m \ell_{\mathrm{rank}}(p, u_j, v_j)}_{R_{\mathrm{emp}}} + \lambda\,\mathrm{KL}(p\|q)$$

where $m = |\prec|$, the number of preference pairs, and the ranking loss function is

$$\ell_{\text{rank}}(p,u,v) = \begin{cases} \phi(u) - \phi(v), & \text{if } \phi(u) > \phi(v) \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

Note that, in keeping with (2), ℓ_{rank} does not include a margin; we will return to this issue in Section 5. Recall that edge flows $\{p_{uv}\}$ induce node scores ϕ.

Following Bousquet and Elisseeff (2002), we wish to show that an algorithm that finds p to minimize $R_{\text{reg}}(p)$ shows a small gap between empirical risk R_{emp} measured over training $\prec$ and true risk R averaged over random draws of $\prec$, given our choice of ℓ_{rank}. To this end we will show two results.

Theorem 4. *For any $m \geq 1$ and any $\delta \in (0,1)$, the following holds with probability at least $1-\delta$ over the random draw of the sample $\prec$ of size m:*

$$R \leq R_{\text{emp}} + \frac{4\ln 2}{\lambda m} + \left(\frac{8\ln 2}{\lambda} + 1\right)\sqrt{\frac{\ln(1/\delta)}{2m}}$$

Theorem 4 does not restrict p in any way except to insist that it is a valid flow. Closer scrutiny in Section 4.2 shows that Theorem 4 can be strengthened by restricting two parameters: the maximum outdegree D in G, and the *eccentricity ratio* in $\hat{G}$:

$$\rho = \max_{u \in \hat{V}} \frac{\max_{v:(u,v)\in\hat{E}} p_{uv}}{\min_{v:(u,v)\in\hat{E}} p_{uv}} \quad \text{(Eccentricity)}$$

Large ρ is bad; for reference flow q, $\rho = 1$.

Theorem 5. *Suppose nodes in G have outdegree at most D, and p is restricted to have eccentricity at most $\rho \geq 1$. Then for any $m \geq 1$ and any $\delta \in (0,1)$, the following bound holds with a probability at least $1-\delta$ over the random draw of the sample $\prec$ of size m:*

$$R \leq R_{\text{emp}} + 2\beta + (4m\beta + 1)\sqrt{\frac{\ln(1/\delta)}{2m}},$$

where β is a function of D, ρ, λ as given by (7).

If preference $u_i \prec v_i$ is dropped from the training set to get $m-1$ preference pairs $\prec^{\setminus i}$, then, instead of p and ϕ, we get $p^{\setminus i}$ and $\phi^{\setminus i}$. Let $\Delta\phi(u) = \phi^{\setminus i}(u) - \phi(u)$.

We can verify that $|\ell_{\text{rank}}(f_1,u,v) - \ell_{\text{rank}}(f_2,u,v)| \leq |f_1(u) - f_2(u)| + |f_1(v) - f_2(v)|$, and from this, we can extend Lemma 20 of Bousquet and Elisseeff (2002) (by setting $t = 1/2$ in their derivation) to the following form suitable for us (we omit the proof):

Lemma 6. *For $i = 1, \ldots, m$,*

$$\begin{aligned} &\text{KL}(p\|q) + \text{KL}(p^{\setminus i}\|q) - 2\,\text{KL}\left(\tfrac{p+p^{\setminus i}}{2} \,\middle\|\, q\right) \\ &\leq \frac{1}{2\lambda m}\left(|\Delta\phi(u_i)| + |\Delta\phi(v_i)|\right) \leq \frac{1}{2\lambda m}\left\|\phi - \phi^{\setminus i}\right\|_1 \end{aligned}$$

Note that the lhs is in terms of p and $p^{\setminus i}$, whereas the rhs is in terms of ϕ and $\phi^{\setminus i}$. To demonstrate generalization, we need to show that $|\ell_{\text{rank}}(p,u,v) - \ell_{\text{rank}}(p^{\setminus i},u,v)|$ is uniformly small (stable).

4.1. No assumption about p

As indicated before, we will first use Lemma 3 and triangle inequality to lower bound the lhs of (6) with some function of ϕ and $\phi^{\setminus i}$, which we will then compare with the rhs $\frac{1}{2\lambda m}\left\|\phi - \phi^{\setminus i}\right\|_1$ so as to derive the following upper bound.

Lemma 7. $\left\|\phi - \phi^{\setminus i}\right\|_1 \leq \frac{2\ln 2}{\lambda m}$.

Proof.

$$\begin{aligned} &\text{KL}(p\|q) + \text{KL}(p^{\setminus i}\|q) - 2\,\text{KL}\left(\tfrac{p+p^{\setminus i}}{2} \,\middle\|\, q\right) \\ &= \text{KL}\left(p \,\|\, \tfrac{p+p^{\setminus i}}{2}\right) + \text{KL}\left(p^{\setminus i} \,\|\, \tfrac{p+p^{\setminus i}}{2}\right) \text{ (Lemma 3)} \\ &\geq \frac{1}{2\ln 2}\left(\left\|p - \frac{p+p^{\setminus i}}{2}\right\|_1^2 + \left\|p^{\setminus i} - \frac{p+p^{\setminus i}}{2}\right\|_1^2\right) \\ &= \frac{1}{4\ln 2}\left\|p - p^{\setminus i}\right\|_1^2 \geq \frac{1}{4\ln 2}\left\|\phi - \phi^{\setminus i}\right\|_1^2. \end{aligned}$$

Combining Lemma 6 with the above, we get

$$\frac{1}{4\ln 2}\left\|\phi - \phi^{\setminus 2}\right\|_1^2 \leq \frac{1}{2\lambda m}\left\|\phi - \phi^{\setminus 2}\right\|_1,$$

from which the desired result follows. □

Proof of Theorem 4. From the definition (6) of ℓ_{rank}, it can be seen that $|\ell_{\text{rank}}(p,u,v) - \ell_{\text{rank}}(p^{\setminus i},u,v)| \leq |\phi(u) - \phi^{\setminus i}(u)| + |\phi(v) - \phi^{\setminus i}(v)| \leq \left\|\phi - \phi^{\setminus i}\right\|_1$, Hence using the above stability of ϕ, we get that ℓ_{rank} is $\frac{2\ln 2}{\lambda m}$-stable, which gives us the result. □

Comparison with bipartite ranking: Agarwal and Niyogi (2005) and Agarwal (2006) assume that positive examples are sampled iid from X_+, negative examples are sampled iid from X_-, and then all (u_-, v_+) pairs are instantiated (but the preferences are not iid over $X \times X$). Their bounds involve $|X_+||X_-|/(|X_+| + |X_-|)$, signifying a loss of effective sample size with class skew. Instead, we assume that m pair-preferences are sampled iid from $X \times X$, so our bounds involve only $m = |\prec|$.

4.2. Crucial parameters constraining p

To go beyond Theorem 4, we need to understand the conditions that affect (KL)'s ability to generalize. $\hat{G}$ exerts very strong influence, and makes the problem much more difficult than ranking feature vectors. In Figure 1(a) (uniform teleport edges hidden and $\{(u,v),(v,u)\}$ shown as $\leftrightarrow$), no training preferences between nodes other than 0 can help generalize to any test nodes disjoint from training nodes. (KL) can fit any consistent $\prec$ and support $(n-1)!$ total orders. In contrast, not all Pagerank rankings are possible in Figure 1(b), e.g. the preference $1 \prec 3 \prec 2 \prec 4$.

It might seem that the excessive overfitting power of Figure 1(a) can be arrested by a degree bound D, but Figure 1(c) shows that approximately $(n/D)!\,D!$ rankings are still possible. To avoid such scenarios, we also need to keep p close to q in another sense: flows going out of a given node should have low eccentricity ρ.

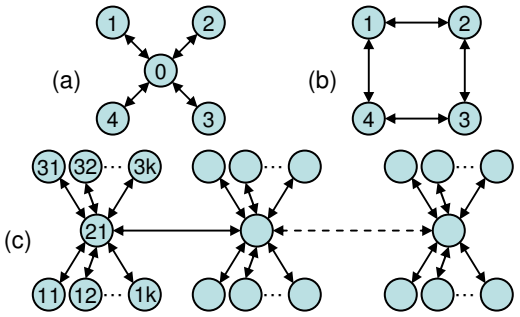

Figure 1. Limits to generalization.

With both D and ρ constrained, we can show that $\|\phi - \pi\|_\infty \leq 2\alpha \min\left\{\frac{(\rho-1)(D-1)}{(D-1)\rho+1-\alpha D}, \frac{(\rho-1)(D-1)}{D-1+\rho-\alpha D\rho}\right\}$ provided $\alpha \max\left\{\frac{(\rho-1)(D-1)}{(D-1)\rho+1-\alpha D}, \frac{(\rho-1)(D-1)}{D-1+\rho-\alpha D\rho}\right\} \leq 1$. This suggests that ρ and D can be effective handles on generalization.

4.3. p constrained by D and ρ

Lemma 8. *Let $p, \tilde{p}$ be valid flows and $\phi, \tilde{\phi}$ be the corresponding node Pageranks. If $p, \tilde{p}$ are constrained by D, ρ, then*

$$\begin{aligned}\mathrm{KL}(p\|\tilde{p}) &\geq \sum_{v\in\hat{V}} \frac{\phi(v)D}{(D-1)\rho+1} \log \frac{\phi(v)(D-1+\rho)}{\tilde{\phi}(v)\rho((D-1)\rho+1)} \\ &\geq \frac{D\|\phi-\tilde{\phi}\|_1^2}{((D-1)\rho+1)2\ln 2} \\ &\quad - \frac{D}{(D-1)\rho+1} \log \frac{\rho((D-1)\rho+1)}{D-1+\rho}\end{aligned}$$

Proof. Consider the flow p_{uv} through edge $(u,v) \in \hat{E}$. Then the flows through all other edges out of u are in the interval $[\frac{p_{uv}}{\rho}, \rho p_{uv}]$ by definition of ρ. Then we can bound the flows as:

$$p_{uv} + (d_u - 1)\rho p_{uv} \geq \phi(u) \Rightarrow p_{uv} \geq \frac{\phi(u)}{(d_u-1)\rho+1}$$

$$\tilde{p}_{uv} + (d_u - 1)\frac{\tilde{p}_{uv}}{\rho} \leq \tilde{\phi}(u) \Rightarrow \tilde{p}_{uv} \leq \frac{\tilde{\phi}(u)\rho}{d_u - 1 + \rho}$$

From which we get $\mathrm{KL}(p\|\tilde{p}) = \sum_{(u,v)\in\hat{E}} p_{uv} \log \frac{p_{uv}}{\tilde{p}_{uv}}$

$$\begin{aligned}&\geq \sum_{(u,v)\in\hat{E}} \frac{\phi(u)}{(d_u-1)\rho+1} \log \frac{\phi(u)}{(d_u-1)\rho+1} \frac{d_u-1+\rho}{\tilde{\phi}(u)\rho} \\ &= \sum_{u\in\hat{V}} \frac{\phi(u)d_u}{(d_u-1)\rho+1} \log \frac{\phi(u)(d_u-1+\rho)}{\tilde{\phi}(u)\rho((d_u-1)\rho+1)} \\ &\geq \sum_{u\in\hat{V}} \frac{\phi(u)D}{(D-1)\rho+1} \log \frac{\phi(u)(D-1+\rho)}{\tilde{\phi}(u)\rho((D-1)\rho+1)},\end{aligned}$$

because $\frac{d_u}{(d_u-1)\rho+1}$ and $\frac{d_u-1+\rho}{(d_u-1)\rho+1}$ are both decreasing function of d_u. Continuing,

$$\begin{aligned}\ldots &= \sum_{u\in\hat{V}} \frac{\phi(u)D}{(D-1)\rho+1} \log \frac{\phi(u)(D-1+\rho)}{\tilde{\phi}(u)\rho((D-1)\rho+1)} \\ &= \frac{D}{(D-1)\rho+1}\mathrm{KL}(\phi\|\tilde{\phi}) + \\ &\quad \frac{D}{((D-1)\rho+1)} \log \frac{D-1+\rho}{\rho((D-1)\rho+1)} \underbrace{\textstyle\sum_{u\in\hat{V}} \phi(u)}_{=1},\end{aligned}$$

which yields the result after another application of Lemma 3 to $\mathrm{KL}(\phi\|\tilde{\phi})$. □

Proof of Theorem 5: Combining Lemma 8 (set $\tilde{p} = p^{\setminus i}$ and $\tilde{\phi} = \phi^{\setminus i}$) with Lemma 6, we get

$$\begin{aligned}\frac{1}{2\lambda m}\left\|\phi - \phi^{\setminus i}\right\|_1 \geq \frac{D\|\phi-\phi^{\setminus i}\|_1^2}{((D-1)\rho+1)2\ln 2} + \\ \frac{D}{(D-1)\rho+1} \log \frac{D-1+\rho}{\rho((D-1)\rho+1)},\end{aligned}$$

a quadratic inequality in $\|\phi - \phi^{\setminus i}\|_1$ that solves to

$$\begin{aligned}&\|\phi-\phi^{\setminus i}\|_1 \leq \frac{c_1}{m} + c_2\sqrt{\frac{1}{4\lambda^2 m^2} + c_3}, \\ &\text{where} \quad c_1 = \tfrac{((D-1)\rho+1)\ln 2}{D\lambda}, \quad c_2 = 2\lambda c_1 \qquad (7)\end{aligned}$$

and $c_3 = \frac{2D^2}{((D-1)\rho+1)^2\ln 2} \log \frac{D-1+\rho}{\rho((D-1)\rho+1)}$. □

We note that, just like all the bounds derived by Bousquet and Elisseeff (2002), these are relative loss bounds. That is, we bound the probability of the expected loss being very different from the empirical loss for our specific loss function ℓ_{rank} (6). This does not imply a good loss function on some other loss such as the 0-1 ranking loss that counts the number of inversions between two rankings.

However, these bounds do serve to qualitatively justify the minimization of KL divergence from standard Pagerank flow q. This is because the bounds worsen as the eccentricity ρ increases. As Pagerank π is the unique distribution which has the smallest possible $\rho = 1$, it makes sense to minmize distance to obtain better generalization. The bounds also get worse as the largest outdegree D increases, which makes sense in view of Section 4.2.

5. Additive margin in Markov flows

ℓ_{rank} (6) is not an upper-bound on the 0-1 loss, unlike in most max-margin formulations which use a hinge

loss to bound 0-1 loss, which would be

$$\ell_{\text{hinge}}(f,u,v) = \begin{cases} f(u)+1-f(v), & f(u)+1 \geq f(v) \\ 0, & \text{otherwise} \end{cases}$$

for us. This would corresponding to modifying (2) to

$$\forall u \prec v : \underline{1+} \sum_{(w,u)\in\hat{E}} p_{wu} - \sum_{(w,v)\in\hat{E}} p_{wv} - s_{uv} \leq 0 \quad (8)$$

However, with the $\sum p_{uv} = 1$ constraint, this would be awkward on s_{uv}. An arbitrary additive margin, which a SVM can satisfy through adjusting the norm of the score vector, might not be feasible with (KL).

A natural solution would be to relax $\sum p_{uv}$ while keeping $\{p_{uv}\}$ proportional to a valid flow. As long as $\|p\|_1 \geq 1$, this will not upset the (KL) optimization (proof omitted):

Lemma 9. *Let q be a probability distribution and p be an unnormalized distribution, so that $\sum_x p(x) = F$. Then:*

1. $\mathrm{KL}(p\|q) \geq 0$ *if* $F \geq 1$.
2. *For a fixed* $F \geq 1$, $\arg\min_p \mathrm{KL}(p\|q) = F\,q$.

Given this fact, we can change objective (KL) to

$$\min_{\substack{\{p_{uv}\},\{s_{uv}\} \\ F\geq 1}} \sum_{(u,v)\in E'} p_{uv} \log \tfrac{p_{uv}}{q_{uv}} + C \sum_{u\prec v} s_{uv} + C_1 F^2$$

and change (1) to $\displaystyle\sum_{(u,v)\in E'} p_{uv} - F = 0,$

along with using (8) in place of (2). The modified problem can also be solved in the dual using a box-constrained Newton method (e.g., BLMVM). In Section 7 we see that empirically this scheme gives better accuracy than (KL) without margin, and even better than (Lap).

Furthermore, using techniques similar to those in Section 4, we can prove the following:

Theorem 10. *Suppose that $\|p\|_1$ of allowed hypothesis distributions p is at most $\kappa \geq 1$. Then for any $m \geq 1$ and any $\delta \in (0,1)$, the following holds with probability at least $1-\delta$ over the random draw of the sample $\prec$ of size m:*

$$R \leq R_{\text{emp}} + \sqrt{\frac{(\kappa+1)^2 + 12(\kappa+1)m\beta}{2m\delta}}$$

where $\beta \leq \kappa - 1 + r + \sqrt{r(2(\kappa-1)+r)}$, $r = \frac{\kappa \ln 2}{\lambda m}$.

It should be noted here that this a polynomial bound on the error $|R - R_{\text{emp}}|$ unlike the exponential bounds reported in Section 4. A meaningful exponential bound would be possible only if $\beta = o(1/m)$. One should not compare the generalization performances of (KL) with and without margin using these bounds, because they are relative to *different* loss functions ℓ_{rank} and ℓ_{hinge}. ℓ_{rank} makes no connection with 0-1 loss, while ℓ_{hinge} gives an upper bound. Overall, it is reassuring that increasing κ worsens generalization, motivating the $C_1 F^2$ term in the new objective.

6. Cost sensitive ranking framework

Many recent papers (Matveeva et al., 2006; Rudin, 2006; Burges et al., 2006) address ranking applications where high precision at the top ranks is crucial. A common approach is to penalize incorrect training predictions for top-ranked items more severely. Unfortunately, this requires knowledge of absolute ranks. Collecting ordinal targets (for ordinal regression) is reasonable, but collecting *absolute* rank information over large sets of instances is burdensome. *Relative* preference pairs are easier to obtain from click-through and eye-tracking data (Joachims, 2002).

Can we approach cost-sensitive ranking with only pairwise preferences? We build on the following intuition: If our algorithm assigns a top rank to a node, it should have high confidence, and pay more for a mistake. Since high score is a surrogate for top ranks, we can penalize the loss $\ell(f_u, f_v)$ more if f_u and/or f_v is large.

As a general framework, we replace $\ell(f_u, f_v)$ with a function $g(f_u, f_v, \ell(f_u, f_v))$ that satisfies:

- $g(f_u, f_v, \ell(f_u, f_v)) \geq g(f_u, f_v, \ell'(f_u, f_v))$, if $\ell(f_u, f_v) \geq \ell'(f_u, f_v), \ \forall u, v$.
- $g(f_u, f_v, \ell(f_u, f_v)) \geq g(f_u^{'}, f_v^{'}, \ell(f_u^{'}, f_v^{'}))$, if $\ell(f_u, f_v) = \ell(f_u^{'}, f_v^{'})$ and $h(f_u, f_v) \geq h(f_u^{'}, f_v^{'})$ for some function h monotonic in f_u, f_v.

The first property enforces that g will be well-behaved *wrt* the loss function. The second introduces cost-sensitivity. The exact way in which differential penalty is incurred depends on the functions h and g. In addition, we usually like g to be *convex* in f_u and f_v so that we can perform efficient optimization.

If we further assume that there is a constant γ such that $g(f_u, f_v, \ell(f_u, f_v)) \geq \gamma h(f_u, f_v)\ell(f_u, f_v)$, then we can easily prove that:

$$\begin{aligned} &\Pr(g(f_u, f_v, \ell(f_u, f_v)) \geq \epsilon) \leq \delta(\epsilon) \Rightarrow \\ &\Pr(\ell(f_u, f_v) \geq \epsilon \wedge h(f_u, f_v) \geq \theta) \leq \delta(\epsilon\gamma\theta) \end{aligned} \quad (9)$$

As the function δ would be decreasing in ϵ for any reasonable g, the above result shows that the probability of the loss being large on a pair with a high cost is small, and thus an algorithm minimizing g would be cost-sensitive.

One embodiment of the above general principle would be $g_{\max}(f_u, f_v, \ell(f_u, f_v)) = (\max(f_u, f_v) + \ell(f_u, f_v))^2$, which satisfies both properties, with $h(f_u, f_v) = \max(f_u, f_v)$, which demonstrates, via (9), that $g_{\max}$ is cost-sensitive with $\gamma = 2$. To implement it, we in-

troduce variable t_{uv}, assert inequalities

$$\forall u \prec v : t_{uv} \geq \sum_{(w,u)\in\hat{E}} p_{wu}; \quad t_{uv} \geq \sum_{(w,v)\in\hat{E}} p_{wv},$$

and replace $C\sum_{u\prec v} s_{uv}$ by $C\sum_{u\prec v}(t_{uv}+s_{uv})^2$. The function is clearly convex in all its arguments.

The generalization bounds for the loss function ℓ in the stability setup carry over to the function g with the σ-admissibility of ℓ replaced by that of g. Consider the function $g_{\max}$, for example. For functions f, f' taking values in an interval $[a,b]$, loss function in (6), $g_{\max}$ satisfies $|(\max(f_u,f_v)+s_{uv})^2-(\max(f'_u,f'_v)+s_{uv})^2| \leq \underbrace{\max\{4(2b-a),2(3b-a)\}}_{=\sigma}(|f_u-f'_u|+|f_v-f'_v|)$. This gives us $\|g-g^{\setminus i}\|_\infty \leq 2\sigma\beta$, where β is a bound on $\|f-f^{\setminus i}\|_\infty$. The latter can be bounded by plugging the σ-admissibility of g in the techniques of Section 4.

Despite the desirable properties of $g_{\max}$, the dual problem contains many equality constraints in this setup. Taking a cue from augmented Lagrangian appraoches, we add a quadratic penalty barrier for every constraint to the objective function. A stiff penalty ensures that a minimizer of the augmented objective is also approximately feasible *wrt* the dual constraints.

7. Experiments

Synthetic graphs: We used RMAT (Chakrabarti et al., 2004) to generate graphs with 1000–4000 nodes and 4000–16000 edges resembling real social networks.

Real graphs: We also performed experiments on biological cellular networks used by Jeong et al. (2000). The dataset consists of small directed graphs that resemble social networks in degree distribution.

Preferences: Our goal was to see how well the learning algorithms can identify a hidden favored "personalized" community (Jeh & Widom, 2003) from $\prec$. To this end, we first computed reference scores π. Then we chose random hidden seed nodes and routed a large (0.1–0.8) teleport into the seed. This gave us the hidden ground truth ϕ^* for all nodes, from which we sampled pairs to prepare $\prec$ for training and testing. To remove transitivity artifacts, we ensured these are node-disjoint.

Evaluation: The algorithm must estimate ϕ and p to induce a ranking. Accuracy is measured in terms of the fraction of test pairs whose rankings violates the hidden ϕ^*. Results are averaged over six randomly-chosen hidden communities, and cross-validated for C and/or C_1.

Preference satisfaction results: Figure 2 shows that (KL) with additive margin compares favorably with (Lap) in terms of accuracy. (KL) without additive margin performs poorly.

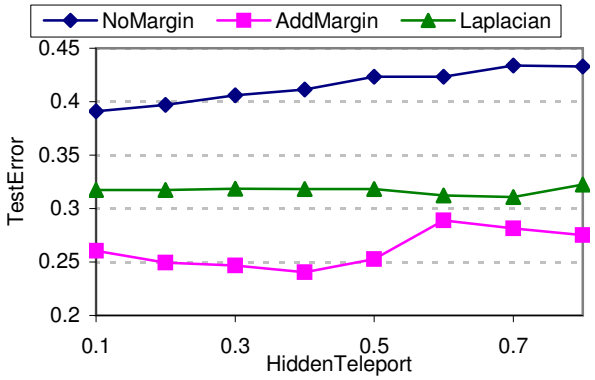

Figure 2. Comparison over synthetic graphs of the accuracy of Laplacian- and KL-based approaches without and with additive margin. The x-axis is the fraction of teleport diverted into the hidden favored community.

Our numbers for (Lap) cannot be compared with those reported by Agarwal (2006), for two reasons. First, Agarwal (2006) designed edge weights $w(u,v)$ by hand, with consideration to domain knowledge and node features, to *fix* the Laplacian; in (KL), flows p_{uv} are estimated as part of learning from $\prec$. Second, Agarwal (2006) optimized for ordinal regression (3- or 5-partite ranking) tasks, while we generate arbitrary preference pairs.

(KL) with additive margin works well also on cellular networks, as shown in Figure 3. The training and test preferences were generated synthetically as described above.

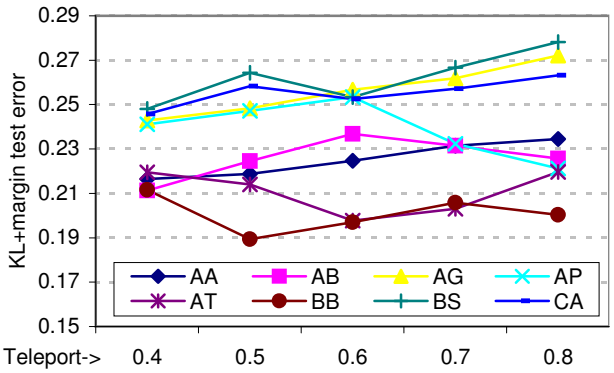

Figure 3. Accuracy of (KL) with additive margin for various cellular graphs.

Cost-sensitive ranking results: In the cost-sensitive setting, only pair preferences are presented to the algorithm, even though the trainer knows the ranks induced by ϕ^*. The algorithm estimates ϕ and uses it for ranking. Let k be a rank cutoff, T_k^* the true top-k nodes using ϕ^*, and T_k those reported using ϕ. The "precision at k" is defined as $|T_k^u \cap \hat{T}_k^u|/k$.

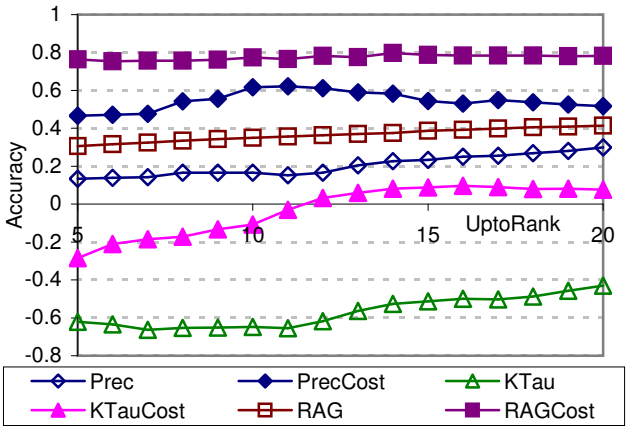

Figure 4. Effect of cost-sensitive optimization on three top-k accuracy measures. The x-axis is the rank k up to which accuracy is measured.

The "relative average goodness" or "RAG at k" is $\left(\sum_{v\in T_k}\phi^*(v)\right)\ /\ \left(\sum_{v\in T_k^*}\phi^*(v)\right)$. We also measure Kendall's τ between (ordered) T_k and T_k^*. Figure 4 shows the results—cost-sensitive (Lap) is consistently better than baseline (Lap) wrt all three criteria.

8. Conclusion

We analyzed and enhanced algorithms to learn Pagerank-style random walks for ranking nodes in graphs, and drew correspondences with a recent algorithm that uses graph Laplacians. The latter does not ensure Markov balance, essential for many Pagerank optimizations, and involves the diagonalization of a large matrix. Furthermore, generalization power of the latter approach is expressed in terms of the somewhat inscrutable quantity $\max_u L^+_{uu}$, whereas, for Pagerank, generalization can be expressed wrt intuitive parameters D and ρ. Pagerank learning uses a hypothesis space that is a strict subset of the Laplacian smoothing approach. This increased bias seems to help in practice for the kind of ranking tasks we consider.

References

Agarwal, A., Chakrabarti, S., & Aggarwal, S. (2006). Learning to rank networked entities. *SIGKDD Conference* (pp. 14–23).

Agarwal, S. (2006). Ranking on graph data. *ICML* (pp. 25–32).

Agarwal, S., Cortes, C., & Herbrich, R. (Eds.). (2005). *Learning to rank*, NIPS Workshop.

Agarwal, S., & Niyogi, P. (2005). Stability and generalization of bipartite ranking algorithms. *COLT* (pp. 32–47). Bertinoro.

Balmin, A., Hristidis, V., & Papakonstantinou, Y. (2004). Authority-based keyword queries in databases using ObjectRank. *VLDB*. Toronto.

Bousquet, O., & Elisseeff, A. (2002). Stability and generalization. *Journal of Machine Learning Research*, *2*, 499–526.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *WWW Conference*.

Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., & Hullender, G. (2005). Learning to rank using gradient descent. *ICML*.

Burges, C. J. C., Ragno, R., & Le, Q. V. (2006). Learning to rank with nonsmooth cost functions. *NIPS*.

Chakrabarti, D., Zhan, Y., & Faloutsos, C. (2004). R-MAT: A recursive model for graph mining. *ICDM*.

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

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. John Wiley and Sons, Inc.

Herbrich, R., Graepel, T., & Obermayer, K. (1999). Support vector learning for ordinal regression. *International Conference on Artificial Neural Networks* (pp. 97–102).

Jeh, G., & Widom, J. (2003). Scaling personalized web search. *WWW Conference* (pp. 271–279).

Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., & Barabási, A.-L. (2000). The large-scale organization of metabolic networks. *Nature*, *407*, 651–654.

Joachims, T. (2002). Optimizing search engines using clickthrough data. *SIGKDD Conference*. ACM.

Langville, A. N., & Meyer, C. D. (2004). Deeper inside PageRank. *Internet Mathematics*, *1*, 335–380.

Matveeva, I., Burges, C., Burkard, T., Laucius, A., & Wong, L. (2006). High accuracy retrieval with multiple nested ranker. *SIGIR Conference* (pp. 437–444). Seattle, Washington, USA.

Rudin, C. (2006). Ranking with a p-norm push. *COLT* (pp. 589–604).

Smola, A., & Kondor, R. (2003). Kernels and regularization on graphs. *COLT* (pp. 144–158).

Taskar, B. (2004). *Learning structured prediction models: A large margin approach*. Doctoral dissertation, Stanford University.

Zhou, D., Huang, J., & Schölkopf, B. (2005). Learning from labeled and unlabeled data on a directed graph. *ICML* (pp. 1041–1048).

Zhou, D., & Schölkopf, B. (2004). A regularization framework for learning from graph data. *ICML Workshop on Statistical Relational Learning*.