Internet Mathematics Vol. 8, Nos. 1–2: 46–72

Finding and Visualizing Graph Clusters Using PageRank Optimization

Fan Chung and Alexander Tsiatas

Abstract. We give algorithms for finding graph clusters and drawing graphs, highlighting local community structure within the context of a larger network. For a given graph G, we use the personalized PageRank vectors to determine a set of clusters, by optimizing the jumping parameter α subject to several cluster variance measures in order to capture the graph structure according to PageRank. We then give a graph visualization algorithm for the clusters using PageRank-based coordinates. Several drawings of real-world data are given, illustrating the partition and local community structure.

1. Introduction

Finding smaller local communities within a larger graph is a well-studied problem with many applications. For example, advertisers can more effectively serve niche audiences if they can identify their target communities within the larger social web, and viruses in technological or population networks can be effectively quarantined by distributing antidote to local clusters around their origins [Chung et al. 09].

ISSN: 1542-7951/08

There are numerous well-known algorithms for finding clusters within a graph, including k-means [Lloyd 82, MacQueen 67], spectral clustering [Ng et al. 02, Shi and Malik 00], Markov cluster algorithms [Enright et al. 02], and numerous others [Harel and Koren 02, Mancoridis et al. 99, Moody 01, Newman and Girvan 04, Noack 09]. Many of these algorithms require embedding a graph into low-dimensional Euclidean space using pairwise distances, but graph distance-based metrics fail to capture graph structure in real-world networks with small-world phenomena, since all pairs of nodes are connected within short distances. PageRank provides essential structural relationships between nodes and is particularly well suited for clustering analysis. Furthermore, PageRank vectors can be computed more efficiently than performing a dimension reduction for a large graph.

In this paper, we give clustering algorithms, *PageRank-clustering*, that use PageRank vectors to draw attention to local graph structure within a larger network. PageRank was first introduced in [Brin and Page 98] for Web search algorithms. Although the original definition is for the Web graph, PageRank is well defined for any graph. Here, we will use a modified version of PageRank, known as personalized PageRank [Jeh and Widom 03], using a prescribed set of nodes as a seed vector.

PageRank can capture well the quantitative correlations between pairs or subsets of nodes, especially on small-world graphs, where the usual graph distances are all quite small. We use PageRank vectors to define a notion of PageRank distance that provides a natural metric space appropriate for graphs.

A key diffusion parameter in deriving PageRank vectors is the jumping constant α. In our clustering algorithms, we will use α to control the scale of the clustering. In particular, we introduce two variance measures that can be used to automatically find the optimized values for α. We then use PageRank vectors determined by α to guide the selection of a set of centers of mass and use them to find the clusters via PageRank distances. We further apply our clustering algorithm to derive a visualization algorithm that we call *PageRank-display* to effectively display local structure in drawings of large networks.

The paper is organized as follows: The basic definitions for PageRank are given in Section 2. In Section 3, we describe two cluster variance measures using PageRank vectors, and we give clustering algorithms in Section 4, with analysis in Sections 5 and 6. A graph-drawing algorithm is given in the last section and several examples are included.

2. Preliminaries

We consider general undirected graphs $G = (V, E)$ with vertex set V and edge set E. For a vertex v, let d_v denote the *degree* of v, which is the number of *neighbors*

of v. For a set of nodes $T \subseteq V$, the *volume* of T is defined to be $\mathrm{vol}(T) = \sum_{v \in T} d_v$. Let D denote the *diagonal degree matrix* and A the *adjacency matrix* of G, where

$$A_{ij} = \begin{cases} 1 & \text{if } \{v_i, v_j\} \in E, \\ 0 & \text{otherwise.} \end{cases}$$

We consider a typical random walk on G with the *transition probability matrix* defined by $P = D^{-1}A$, and we denote the lazy walk by $W = (I + P)/2$. Let $\pi = \vec{d}/\mathrm{vol}(G)$ denote the stationary distribution of the random walk, if it exists. Personalized PageRank vectors are based on random walks with two governing parameters: a seed vector $\vec{s}$, representing a probability distribution over V, and a jumping constant α, controlling the rate of diffusion. The personalized PageRank vector $\rho(\alpha, \vec{s})$ is defined to be the solution to the following recurrence relation:

$$\rho(\alpha, \vec{s}) = \alpha\vec{s} + (1 - \alpha)\rho(\alpha, \vec{s})W.$$

Here, $\vec{s}$ (and all other vectors) will be treated as row vectors. The original definition of PageRank defined in [Brin and Page 98] is the special case in which the seed vector is the uniform distribution. If $\vec{s}$ is simply the distribution that is 1 for a single node v and 0 elsewhere, we write $\rho(\alpha, v)$.

An alternative expression for the personalized PageRank $\rho(\alpha, \vec{s})$ is a geometric sum of random walks (see [Andersen et al. 06]):

$$\rho(\alpha, \vec{s}) = \alpha \sum_{t=0}^{\infty} (1 - \alpha)^t \vec{s} W^t.$$

In general, it can be computationally expensive to compute PageRank exactly; it requires using the entire graph structure, which can be prohibitive on large networks. Instead, we use an approximate PageRank algorithm as given in [Andersen et al. 06, Chung and Zhao 10]. This approximation algorithm is much more tractable on large networks, because it can be computed using only the local graph structure around the starting seed vector $\vec{s}$. Besides $\vec{s}$ and the jumping constant α, the algorithm requires an approximation parameter ϵ.

For a subset of nodes H in a graph G, the *Cheeger ratio* $h(H)$ is a measure of the cut between H and its complement $\bar{H}$:

$$h(H) = \frac{e(H, \bar{H})}{\min(\mathrm{vol}(H), \mathrm{vol}(\bar{H}))},$$

where $e(H, \bar{H})$ denotes the number of edges $\{u, v\}$ with $u \in H$ and $v \in \bar{H}$.

For a set of points $S = \{s_1, \ldots, s_n\}$ in Euclidean space, the *Voronoi diagram* is a partition of the space into disjoint regions $R_1, \ldots, R_n$ such that each R_i contains s_i and the region of space containing the set of points that are closer to s_i than any other s_j. Voronoi diagrams are well studied in the field of computational

geometry. Here we consider Voronoi diagrams on graphs using PageRank vectors as a notion of closeness.

For two vertices u, v, we define the *PageRank distance* with jumping constant α as

$$\mathrm{dist}_\alpha(u,v) = \left\| \rho(\alpha,u)D^{-1/2} - \rho(\alpha,v)D^{-1/2} \right\|.$$

Throughout this paper, $\|\cdot\|$ denotes the L_2 norm. This choice of the norm allows for differentiation in the optimization process later.

We can further generalize this distance to two probability distributions p and q defined on the vertex set V of G. Namely, the PageRank distance, with jumping constant α, between p and q is defined by

$$\mathrm{dist}_\alpha(p,q) = \sum_{u,v} p(u)q(v)\mathrm{dist}_\alpha(u,v).$$

With this definition, for a subset S of vertices, we can generalize the notion of a center of mass for S to be a probability distribution c. For a given $\epsilon > 0$, we say that c is an ϵ-*center* or *center of mass* for S if

$$\sum_{v\in S} \mathrm{dist}_\alpha(c,v) \le \epsilon.$$

Let C denote a set of k (potential) centers. The goal is for each center c to be a representative center of mass for some cluster of vertices. We let R_c denote the set of all vertices x that are closest to c in terms of PageRank, provided the jumping constant α is given:

$$R_c = \{x \in V : \mathrm{dist}_\alpha(c,x) \le \mathrm{dist}_\alpha(c',x) \text{ for all } c' \in C\}.$$

3. PageRank Variance and Cluster Variance Measures

For a vertex v and a set of centers C, let c_v denote the center that is closest to v, i.e., c_v is the center of mass $c \in C$ such that $v \in R_c$.

We follow the approach using k-means by defining the following evaluative measure for a potential set of k centers C, using PageRank instead of Euclidean distances:

$$\mu(C) = \sum_{v\in V} d_v \|\rho(\alpha,v)D^{-1/2} - \rho(\alpha,c_v)D^{-1/2}\|^2 = \sum_{v\in V} d_v \ \mathrm{dist}_\alpha(v,c_v)^2.$$

Selecting a set of representative centers within a graph is a hard problem, known to be NP-complete [Aloise et al. 09]. There are many approximate and heuristic algorithms used in practice (see [Schaeffer 07]). Here, we will develop

algorithms that use personalized PageRank vectors to select the centers. In the Web graph, links between websites can be interpreted as votes for a website's importance, and PageRank vectors are used to determine which pages are intrinsically more important in the overall graph. Personalized PageRank vectors are local information quantifying the importance of every node to the seed. Thus, the uth component of the personalized PageRank vector $\rho(\alpha, v)$ quantifies how well suited u is to be a representative cluster center for v.

To evaluate a set of cluster centers in a graph G, we consider two measures that capture the community structure of G with respect to PageRank:

$$\begin{aligned}\Phi(\alpha) &= \sum_{v\in V} d_v \left\|\rho(\alpha, v)D^{-1/2} - \rho(\alpha, \rho(\alpha, v))D^{-1/2}\right\|^2 \\ &= \sum_{v\in V} d_v \ \mathrm{dist}_\alpha(v, \rho(\alpha, v))^2, \\ \Psi(\alpha) &= \sum_{v\in V} d_v \left\|\rho(\alpha, \rho(\alpha, v))D^{-1/2} - \pi D^{-1/2}\right\|^2 \\ &= \sum_{v\in V} d_v \ \mathrm{dist}_\alpha(\rho(\alpha, v), \pi)^2.\end{aligned}$$

The *α-PageRank variance* $\Phi(\alpha)$ measures discrepancies between the personalized PageRank vectors for nodes v and possible centers nearest to v, represented by the probability distribution $\rho(\alpha, v)$. The *α-cluster variance* $\Psi(\alpha)$ measures large discrepancies between personalized PageRank vectors for nodes v and the overall stationary distribution π. If the PageRank variance $\Phi(\alpha)$ is small, then the "guesses" using PageRank vectors for the centers of mass give a good upper bound for the k-means evaluation μ using PageRank distance, indicating the formation of clusters. If the cluster variance $\Psi(\alpha)$ is large, then the centers of mass using the predictions from PageRank vectors are quite far from the stationary distribution, capturing a community structure. Thus, our goal is to find the appropriate α such that $\Phi(\alpha)$ is small but $\Psi(\alpha)$ is large.

For a specific set of centers of mass C, we use the following for an evaluative metric $\Psi_\alpha(C)$, suggesting the structural separation of the communities represented by centers in C:

$$\Psi_\alpha(C) = \sum_{c\in C} \mathrm{vol}(R_c) \left\|\rho(\alpha, c)D^{-1/2} - \pi D^{-1/2}\right\|^2 = \sum_{c\in C} \mathrm{vol}(R_c) \ \mathrm{dist}_\alpha(c, \pi)^2.$$

We remark that this measure is essentially the analogue of k-means in terms of PageRank distance, and it has a similar flavor as a heuristic given in [Dyer and Frieze 85] for the traditional center selection problem. The metrics $\mu(C)$ and

$\Psi_\alpha(C)$ are designed to evaluate a specific set of clusters C, while the measures $\Phi(\alpha)$ and $\Psi(\alpha)$ are well suited to measure a graph's inherent clustered structure.

4. The PageRank-Clustering Algorithms

These evaluative measures give us a way to evaluate a set of community centers, leading to the *PageRank-clustering* algorithms presented here. The problem of finding a set of k centers minimizing $\mu(C)$ is then reduced to the problem of minimizing $\Phi(\alpha)$ while $\Psi(\alpha)$ is large for appropriate α. In particular, for a special class of graphs that consist of k clusters of vertices where each cluster has a bounded Cheeger ratio, the center selection algorithm is guaranteed to be successful with high probability.

A natural question is to find the appropriate α for a given graph, if such α exists and if the graph is clusterable. (The definition for *clusterable* will be given later.) A direct method is to compute the variance metrics for a sample of α and narrow down the range for α using binary search. Here, we give a systematic method for determining the existence of an appropriate α and finding its value by differentiating $\Phi(\alpha)$ and finding roots α satisfying $\Phi'(\alpha) = 0$. It is not too difficult to compute that the derivative of Φ satisfies

$$\Phi'(\alpha) = \frac{1-\alpha}{\alpha^3}\left(\sum_v \left\|g_v(\alpha)D^{-1/2}\right\|^2 - 2\langle g_v(\alpha), \rho(\alpha, g_v(\alpha))D^{-1}\rangle\right), \qquad (4.1)$$

where $g_v(\alpha) = \rho(\alpha, \rho(\alpha, v)(I - W))$. Here, we give two versions of the clustering algorithm. For the sake of clarity, the first PageRank clustering algorithm, Algorithm 1, uses exact PageRank vectors without approximation. The second PageRank clustering algorithm, Algorithm 2, allows for the use of approximate PageRank vectors as well as approximate PageRank variance and cluster variance for faster performance.

We can further reduce the computational complexity by using approximate PageRank vectors in the algorithm *PageRank-clusteringB*.

We remark that using the sharp approximate PageRank algorithm in [Chung and Zhao 10], the error bound δ for PageRank can be set to be quite small, since the time complexity is proportional to $\log(1/\delta)$. If we choose δ to be a negative power of n such as $\delta = \epsilon/n^2$, then approximate PageRank vectors lead to sharp estimates for Φ and Φ' within an error bound of ϵ. Thus for graphs with k clusters, the *PageRank-clusteringB* algorithm will terminate after approximating the roots of Φ', $\mathcal{O}(k \log n)$ approximations of μ and Ψ_α, and $\mathcal{O}(n)$ approximate PageRank computations. With approximation algorithms using sampling, this can be done quite efficiently.

Algorithm 1: PageRank-clusteringA

Input: G, k, ϵ
Output: A set of centers C and partitions S, or nothing
  for all $v \in G$ **do**
    compute $\rho(\alpha, v)$
  end for
  Find the roots of $\Phi'(\alpha)$ (there can be more than one root if G has a layered clustering structure)
  for all roots α **do**
    Compute $\Phi(\alpha)$
    if $\Phi(\alpha) \leq \epsilon$ **then**
      Compute $\Psi(\alpha)$
    else
      Go to the next α
    end if
    if $k < \Psi(\alpha) - 2 - \epsilon$ **then**
      Go to the next α
    else
      Select $c \log n$ sets of k potential centers, randomly chosen according to π
    end if
    for all sets $S = \{v_1, \ldots, v_k\}$ **do**
      Let C be the set of centers of mass where $c_i = \rho(\alpha, v_i)$
      Compute $\mu(C)$ and $\Psi_\alpha(C)$
      if $|\mu(C) - \Phi(\alpha)| \leq \epsilon$ and $|\Psi_\alpha(C) - \Psi(\alpha)| \leq \epsilon$ **then**
        Determine the k Voronoi regions according to the PageRank distances using C and return them
      end if
    end for
  end for

We also note that there might be no clustering output if the conditions set within the algorithms are not satisfied. Indeed, there exist graphs that inherently do not have a k-clustered structure within the error bound that we set for ϵ. Another reason for no output is the probabilistic nature of the above sampling method. We will provide evidence for the correctness of the above algorithm by showing that with high probability, a graph with a k-clustered structure will have outputs that capture its clusters in a feasible manner that we will specify further.

We say a that graph G is (k, h, β, ϵ)-*clusterable* if the vertices of G can be partitioned into k parts such that:

1. Each part S_i has Cheeger ratio at most h.
2. Each S_i has volume at least $\beta \text{vol}(G)/k$ for some constant β.

Algorithm 2: PageRank-clusteringB

Input: G, k, ϵ
Output: A set of centers C and partitions S, or nothing
for all $v \in G$ **do**
compute $\rho(\alpha, v)$
end for
Find the roots of $\Phi'(\alpha)$ within an error bound $\epsilon/2$, by using sampling techniques from [Rudelson and Vershynin 07] involving $\mathcal{O}(\log n)$ nodes, $\log(1/\epsilon)$ values of α and δ-approximate PageRank vectors [Andersen et al. 06, Chung and Zhao 10] where $\delta = \epsilon/n^2$ (there can be more than one root if G has a layered clustering structure)
for all roots α **do**
Approximate $\Phi(\alpha)$
if $\Phi(\alpha) \leq \epsilon$ **then**
Compute $\Psi(\alpha)$
else
Go to the next α
end if
if $k < \Psi(\alpha) - 2 - \epsilon$ **then**
Go to the next α
else
Select $c \log n$ sets of k potential centers, randomly chosen according to π
end if
for all sets $S = \{v_1, \ldots, v_k\}$ **do**
Let C be the set of centers of mass where $c_i = \rho(\alpha, v_i)$
Compute $\mu(C)$ and $\Psi_\alpha(C)$.
if $|\mu(C) - \Phi(\alpha)| \leq \epsilon$ and $|\Psi_\alpha(C) - \Psi(\alpha)| \leq \epsilon$ **then**
Determine the k Voronoi regions according to the PageRank distances using C and return them
end if
end for
end for

3. For each S_i, any subset $S_i' \subset S_i$ with $\mathrm{vol}(S_i') \leq (1-\epsilon)\mathrm{vol}(S_i)$ has its Cheeger ratio at least $c\sqrt{h \log n}$, where $c = 8\sqrt{\beta/k}/\epsilon$.

We will provide evidence for the correctness of *PageRank-clusteringA* by proving the following theorem:

Theorem 4.1. *Suppose a graph G has a (k, h, β, ϵ)-clustering and $\alpha, \epsilon \in (0, 1)$ satisfy $\epsilon \geq hk/(2\alpha\beta)$. Then with high probability, PageRank-clusteringA returns a set C of k centers with $\Phi(\alpha) \leq \epsilon$, $\Psi(C) > k - 2 - \epsilon$, and the k clusters are near optimal according to the PageRank k-means measure μ with an additive error term ϵ.*

5. Several Facts about PageRank

Before proceeding to show that the PageRank-clustering algorithms are effective for treating clusterable graphs, we will first establish some useful tools for analyzing PageRank vectors. These tools concern the diffusion of PageRank vectors in a subset of nodes with small Cheeger ratio. Before we examine a general mixing inequality involving PageRank vectors, first we consider a diffusion lower bound that is a slightly modified version of the results in [Andersen et al. 06].

Lemma 5.1. [Andersen et al. 06] *For any set S and any constants α, δ in $(0,1]$, there is a subset $S_\alpha \subseteq S$ with volume $\mathrm{vol}(S_\alpha) \geq (1-\delta)\mathrm{vol}(S)$ such that for any vertex $v \in S_\alpha$, the PageRank vector $\rho(\alpha, v)$ satisfies*

$$[\rho(\alpha, v)](S) \geq 1 - \frac{h(S)}{2\alpha\delta}.$$

We use the notation that for a function $f : V \to \mathbb{R}$, we have $f(S) = \sum_{v \in S} f(v)$ for $S \subseteq V$. For a positive real value x, we define

$$f(x) = \max \left\{ \sum_v \frac{\beta_v}{d_v} f(v) : \sum_v \beta_v = x, 0 \leq \beta_v \leq d_v \right\}.$$

This leads to many nice properties of f including, for example, that f is concave and that $f(\mathrm{vol}(S)) \geq f(S)$ (see [Andersen et al. 06, Lovász and Simonovits 93]). We use $[f]$ for clarity when f is a complex vector expression.

Lemma 5.2. *For any set S and any constants α, δ in $(0,1]$, there is a subset $S_\alpha \subseteq S$ with volume $\mathrm{vol}(S_\alpha) \geq (1-\delta)\mathrm{vol}(S)$ such that for any vertex $v \in S_\alpha$, the PageRank vector $\rho(\alpha, \rho(\alpha, v))$ satisfies*

$$[\rho(\alpha, \rho(\alpha, v))](S) \geq 1 - \frac{h(S)}{\alpha\delta}.$$

Proof. The proof is quite similar to that in [Andersen et al. 06]. Let χ_S denote the function of S that assumes the value $\chi_S(x) = d_v/\mathrm{vol}(S)$ if $x \in S$ and 0 otherwise. First we wish to show that

$$[\rho(\alpha, \rho(\alpha, \chi_S))](\bar{S}) \leq h(S)\frac{1-\alpha}{\alpha}.$$

During a single step from $\rho(\alpha, \rho(\alpha, \chi_S))$ to $\rho(\alpha, \rho(\alpha, \chi_S))W$, the amount of probability that moves from S to $\bar{S}$ is bounded from above by

$$[\rho(\alpha, \rho(\alpha, \chi_S))W](\bar{S}) \leq [\rho(\alpha, \rho(\alpha, \chi_S))](\bar{S}) + \frac{1}{2}[\rho(\alpha, \rho(\alpha, \chi_S))](|\delta S|), \qquad (5.1)$$

where $\delta(S)$ denotes the edge boundary of S consisting of edges leaving S. Using the definition of PageRank, we obtain

$$\begin{aligned}[\rho(\alpha, \rho(\alpha, \chi_S))](\bar{S}) &\leq \alpha[\rho(\alpha, \chi_S)](\bar{S}) + (1-\alpha)[\rho(\alpha, \rho(\alpha, \chi_S))W](\bar{S}) \\ &\leq \frac{1-\alpha}{2}h(S) + (1-\alpha)[\rho(\alpha, \rho(\alpha, \chi_S))W](\bar{S})\end{aligned}$$

using [Andersen et al. 06, Theorem 4, inequality (8)]. From (5.1), we have

$$\begin{aligned}[\rho(\alpha, \rho(\alpha, \chi_S))W](\bar{S}) \leq \frac{1-\alpha}{2}h(S) &+ (1-\alpha)[\rho(\alpha, \rho(\alpha, \chi_S))](\bar{S}) \\ &+ \frac{1-\alpha}{2}[\rho(\alpha, \rho(\alpha, \chi_S))](|\delta S|).\end{aligned}$$

This implies

$$[\rho(\alpha, \rho(\alpha, \chi_S))](\bar{S}) = \frac{1-\alpha}{2\alpha}h(S) + \frac{1-\alpha}{2\alpha}[\rho(\alpha, \rho(\alpha, \chi_S))](|\delta S|).$$

Now we use the monotonicity property from [Andersen et al. 06, Lemma 4]; we have

$$[\rho(\alpha, \rho(\alpha, \chi_S))](|\delta S|) \leq [\rho(\alpha, \chi_S)](|\delta(S)|) \leq \chi_S(|\delta(S)|) = \frac{|\delta(S)|}{\mathrm{vol}(S)} = h(S).$$

Thus we have

$$[\rho(\alpha, \rho(\alpha, \chi_S))](\bar{S}) \leq \frac{1-\alpha}{\alpha}h(S).$$

To complete the proof, let S_α denote the set of vertices v in S satisfying

$$[\rho(\alpha, \rho(\alpha, v))](\bar{S}) \leq \frac{h(S)}{\alpha\delta}.$$

Let v be a vertex chosen randomly from the distribution $d_v/\mathrm{vol}(S)$, and define the random variable $X = [\rho(\alpha, \rho(\alpha, v))](\bar{S})$. The linearity property of PageRank vectors implies that

$$\mathrm{E}(X) = [\rho(\alpha, \rho(\alpha, \chi_S))](\bar{S}) \leq \frac{1-\alpha}{\alpha}h(S) \leq \frac{h(S)}{\alpha}.$$

Applying Markov's inequality, we have

$$\Pr[v \notin S_\alpha] \leq \Pr[X \geq \mathrm{E}[X]/\delta] \leq \delta.$$

This completes the proof of Lemma 5.2. □

We will also need the quantitative estimates for PageRank vectors restricted to a subset S of vertices. By considering submatrices W_S restricted to rows and columns associated with vertices in S, we can define the Dirichlet PageRank $\rho_S(\alpha, s)$ for a seed vector defined on S and $0 \leq \alpha < 1$ satisfying

$$\rho_S(\alpha, \vec{s}) = \alpha\vec{s} + (1-\alpha)\rho_S(\alpha, \vec{s})W_S.$$

When α is appropriately chosen, the Dirichlet PageRank is a good estimate of PageRank vectors. Lemma 5 and Theorem 6 in [Chung 10] can be rewritten as follows.

Lemma 5.3. [Chung 10] *Suppose a subset S of vertices has Cheeger ratio $h(S)$ satisfying $\epsilon \geq \frac{(1-\alpha)h(T)}{2\alpha}$, for positive values α, ϵ. Then ρ_S satisfies the following: For any $R \subseteq S$, there is a subset $T \subseteq S$ with $\mathrm{vol}(T) \geq (1-\delta)\mathrm{vol}(S)$ such that for every v in T we have*

$$[\rho(\alpha, v)](R) - [\rho_S(\alpha, v)](R) \leq \sqrt{\frac{\epsilon}{\delta}}.$$

For a probability distribution $f : V \to \mathbb{R}$ and a real value x, we define the Cheeger ratio $h_f(x)$ of f up to x as follows: We order the vertices $v_1, v_2, \ldots,$ from highest to lowest probability per degree, so that $p(v_i)/d(v_i) \geq p(v_{i+1})/d(v_{i+1})$. This produces a collection of sets, called the *segment subsets*, with one set $T_j^f = \{v_1, \ldots, v_j\}$ for each $j \leq n$. For a positive value $x \leq \mathrm{vol}(G)$, we define

$$\begin{aligned} h_f(x) &= \max\{h(T_j^f) : j \text{ satisfies } \mathrm{vol}(T_j^f) \leq x\}, \\ h_f^*(x) &= \max\{h(T_j^f) : j \text{ satisfies } \mathrm{vol}(T_j^f) \leq x(1+h_f(x))\}. \end{aligned} \tag{5.2}$$

Lemma 5.4. [Andersen et al. 06] *For a vertex in G, any constant α in $(0,1]$, and nonnegative integer t, the PageRank vector $\rho(\alpha, v)$ satisfies the following:*

$$[\rho(\alpha, v)](T) - \pi(T) \leq \alpha t + \sqrt{\mathrm{vol}(T)}\left(1 - \frac{\phi^2}{8}\right)^t,$$

where ϕ is the Cheeger ratio $h_f^*(\mathrm{vol}(T))$ with $f = \rho(\alpha, v)$.

Lemma 5.5. *For subsets S, T of vertices in G with $\mathrm{vol}(S), \mathrm{vol}(T) \leq \mathrm{vol}(G)/2$, any constant α in $(0,1]$, and nonnegative integer t, the Dirichlet PageRank vector $\rho_S(\alpha, v)$ for any vertex v in S satisfies the following:*

(i)

$$[\rho_S(\alpha, v)](T) - [\rho_S(\alpha, \rho_S(\alpha, v))](T) \le \alpha t + \sqrt{\mathrm{vol}(T)} \left(1 - \frac{\phi^2}{8}\right)^t,$$

where ϕ *is the Cheeger ratio* $h_f^*(\mathrm{vol}(T))$ *with* $f = \rho_S(\alpha, v) - \rho_S(\alpha, \rho_S(\alpha, v))$.

(ii)

$$[\rho_S(\alpha, \rho_S(\alpha, v))](T) - [\rho_S(\alpha, v)](T) \le \alpha t + \sqrt{\mathrm{vol}(T)} \left(1 - \frac{\phi'^2}{8}\right)^t,$$

where ϕ' *is the Cheeger ratio* $h_f'^*(\mathrm{vol}(T))$ *with* $f' = \rho_S(\alpha, \rho_S(\alpha, v)) - \rho_S(\alpha, v)$.

(iii) *For two vertices* u *and* v,

$$[\rho_S(\alpha, u)](T) - [\rho_S(\alpha, v)](T) \le \alpha t + \sqrt{\mathrm{vol}(T)} \left(1 - \frac{\phi''^2}{8}\right)^t,$$

where ϕ'' *is the Cheeger ratio* $h_f''^*(\mathrm{vol}(T))$ *with* $f'' = \rho_S(\alpha, u) - \rho_S(\alpha, v)$.

Proof. We prove (i) by induction on t. For $t = 0$, the assertion clearly holds. Suppose the inequality holds for some $t \ge 0$. Let x denote $\mathrm{vol}(T)$. We use [Andersen et al. 06, Lemma 3] and apply the same method using the concavity of f to obtain

$$[\rho_S(\alpha, v)](T) - [\rho_S(\alpha, \rho_S(\alpha, v))](T) = f(T)$$

and

$$\begin{aligned} f(T) &\le \alpha + (1-\alpha)[fW](T) \le \alpha + (1-\alpha)\left(\frac{1}{2}f(x - \phi x) + \frac{1}{2}f(x + \phi x)\right) \\ &\le \alpha + \left(\frac{1}{2}f(x - \phi x) + \frac{1}{2}f(x + \phi x)\right). \end{aligned}$$

Using the induction assumption, we have

$$\begin{aligned} f(T) &\le \alpha(t+1) + \frac{1}{2}\left(\sqrt{x - \phi x} + \sqrt{x + \phi x}\right)\left(1 - \frac{\phi^2}{8}\right)^t \\ &\le \alpha(t+1) + \sqrt{x}\left(1 - \frac{\phi^2}{8}\right)^{t+1}. \end{aligned}$$

This proves (i). We omit the proofs for (ii) and (iii), which can be done in a similar way. The proof of Lemma 5.5 is complete. □

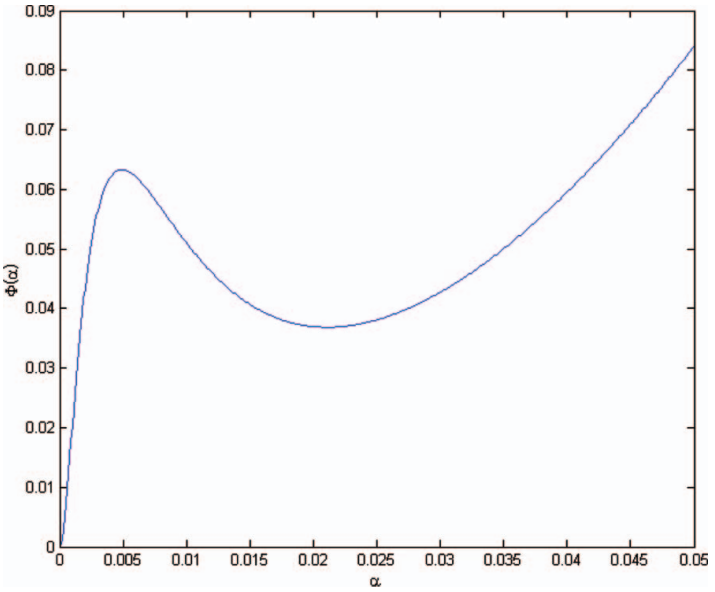

Figure 1. $\Phi(\alpha)$ for the dumbbell graph U (color figure available online).

6. Analyzing PageRank Clustering Algorithms

In this section, we consider an (h, k, β, ϵ)-clusterable graph G, with the following condition:

$$\epsilon \geq \frac{hk}{2\alpha\beta}.$$

Lemma 5.1 implies that in a cluster R of G, most of the vertices u in R have $\rho(\alpha, u)(S) \geq 1 - \epsilon/(2k)$. This fact is essential in the subsequent proof that $\Psi(\alpha) \geq k - 2 - \epsilon$.

We proceed with a series of lemmas that show that if G is (h, k, β, ϵ)-clusterable, then there is an α for which $\Phi(\alpha)$ is small and $\Psi(\alpha)$ is large corresponding to a set of centers chosen from the core of the partitions.

Lemma 6.1. *If a graph G can be partitioned into k clusters having Cheeger ratio at most h and $\epsilon \geq hk/(2\alpha\beta)$, then $\Psi(\alpha) \geq k - 2 - \epsilon$.*

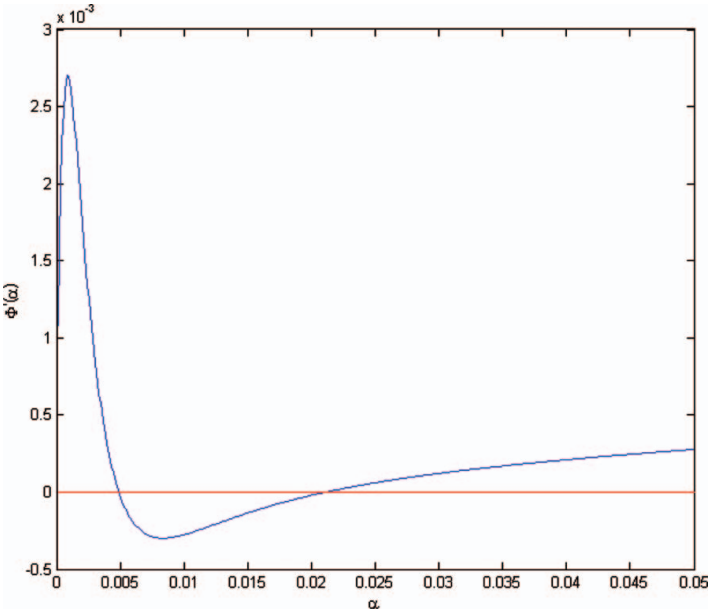

Figure 2. $\Phi'(\alpha)$ for the dumbbell graph U, with the line $y = 0$ for reference (color figure available online).

Proof. Let $S_1, \ldots, S_k$ be a partition of G into k clusters satisfying the conditions of the theorem. Then by definition of Ψ,

$$\begin{aligned}
\Psi(\alpha) &= \sum_{v \in V} d_v \left\| \rho(\alpha, \rho(\alpha, v)) D^{-1/2} - \pi D^{-1/2} \right\|^2 \\
&= \sum_{i=1}^{k} \sum_{v \in S_i} d_v \left\| \rho(\alpha, \rho(\alpha, v)) D^{-1/2} - \pi D^{-1/2} \right\|^2 \\
&= \sum_{i=1}^{k} \sum_{v \in S_i} d_v \sum_{x \in V} \left(\rho(\alpha, \rho(\alpha, v)) D^{-1/2}(x) - \pi D^{-1/2}(x) \right)^2 \\
&\geq \sum_{i=1}^{k} \sum_{v \in S_i} d_v \sum_{x \in S_i} \left(\rho(\alpha, \rho(\alpha, v)) D^{-1/2}(x) - \pi D^{-1/2}(x) \right)^2 \\
&= \sum_{i=1}^{k} \sum_{v \in S_i} d_v \left(\sum_{x \in S_i} \left(\rho(\alpha, \rho(\alpha, v)) D^{-1/2}(x) - \pi D^{-1/2}(x) \right)^2 \sum_{x \in S_i} \frac{d_x}{\mathrm{vol}(S_i)} \right) \\
&= \sum_{i=1}^{k} \sum_{v \in S_i} d_v \left(\sum_{x \in S_i} \frac{1}{d_x} \left(\rho(\alpha, \rho(\alpha, v))(x) - \pi(x) \right)^2 \sum_{x \in S_i} \frac{d_x}{\mathrm{vol}(S_i)} \right).
\end{aligned}$$

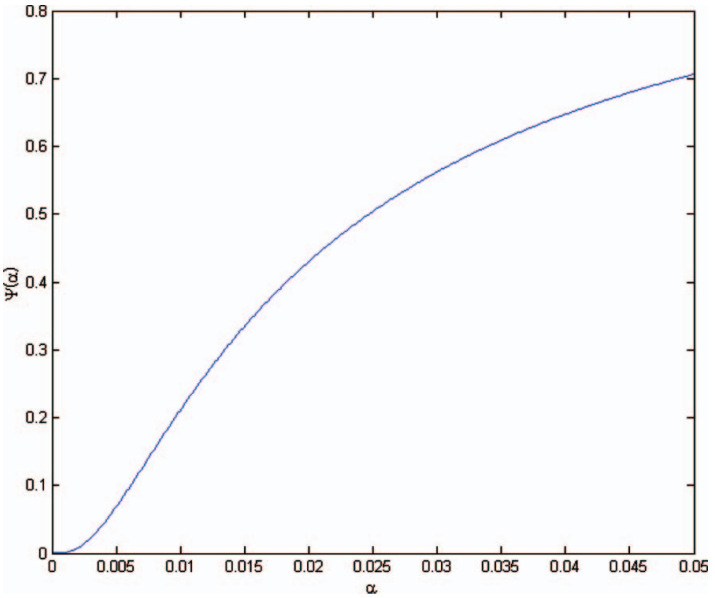

Figure 3. $\Psi(\alpha)$ for the dumbbell graph U (color figure available online).

Using the Cauchy–Schwarz inequality and then Lemma 5.2, we have

$$\begin{aligned}\Psi(\alpha) &\geq \sum_{i=1}^{k} \sum_{v \in S_i} \frac{d_v}{\mathrm{vol}(S_i)} \left(\sum_{x \in S_i} (\rho(\alpha, \rho(\alpha, v))(x) - \pi(x)) \right)^2 \\ &\geq \sum_{i=1}^{k} \sum_{v \in S_i} \frac{d_v}{\mathrm{vol}(S_i)} \left(1 - \frac{\epsilon}{2} - \frac{\mathrm{vol}(S_i)}{\mathrm{vol}(G)} \right)^2 = \sum_{i=1}^{k} \left(1 - \frac{\epsilon}{2} - \frac{\mathrm{vol}(S_i)}{\mathrm{vol}(G)} \right)^2 \\ &\geq \frac{1}{k} \left(\sum_{i=1}^{k} \left(1 - \frac{\epsilon}{2} - \frac{\mathrm{vol}(S_i)}{\mathrm{vol}(G)} \right) \right)^2 = \frac{1}{k} \left(k - 1 - \frac{\epsilon}{2} \right)^2 \geq k - 2 - \epsilon.\end{aligned}$$

□

We have shown that if G has a clustered structure, then there is an α for which $\Psi(\alpha)$ is large. We will also show that our algorithm will also yield $\Phi(\alpha) \leq \epsilon$.

Lemma 6.2. *If G is (k, h, β, ϵ)-clusterable, then we have $\Phi(\alpha) \leq \epsilon$.*

Proof. The proof follows from preceding lemmas. Within each cluster S of G, we first use Lemma 5.2, which implies there is a subset S' of S such that

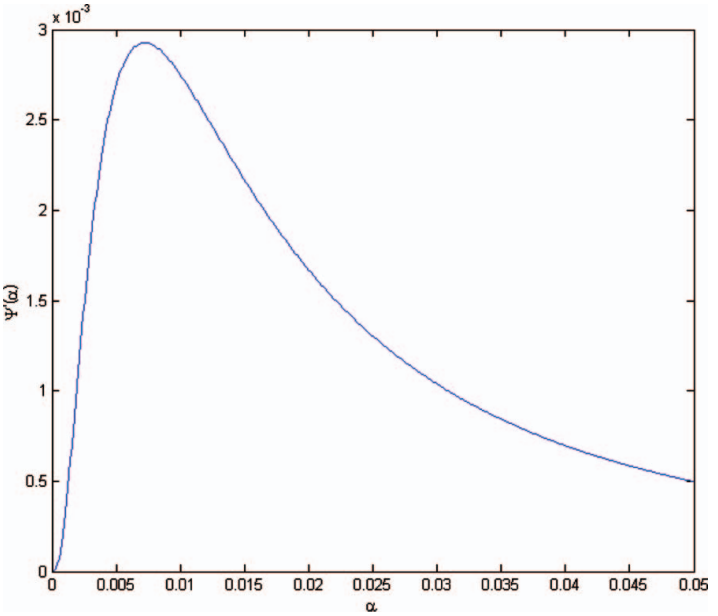

Figure 4. $\Psi'(\alpha)$ for the dumbbell graph U (color figure available online).

$[\rho(\alpha, v)](S) \geq 1 - \epsilon/k$ and $\mathrm{vol}(S') \geq (1 - \delta)\mathrm{vol}(S)$, since S has Cheeger ratio at most h.

We can apply Lemma 5.3 to approximate PageRank vectors $\rho(\alpha, v)$ by the Dirichlet PageRank vectors $\rho_S(\alpha, v)$.

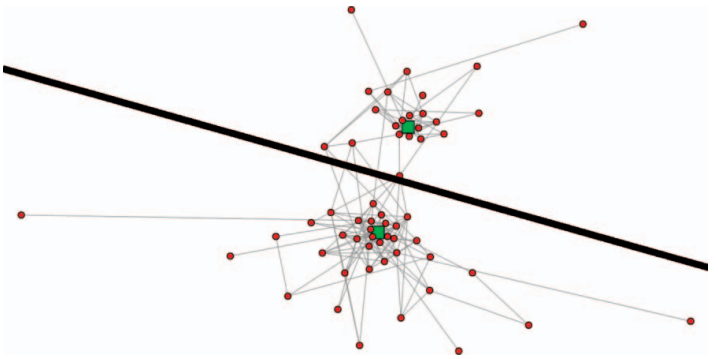

Figure 5. Results of *PageRank-Display* ($\alpha = 0.03$) on the dolphin social network [Lusseau et al. 03], separating the dolphins into two communities (color figure available online).

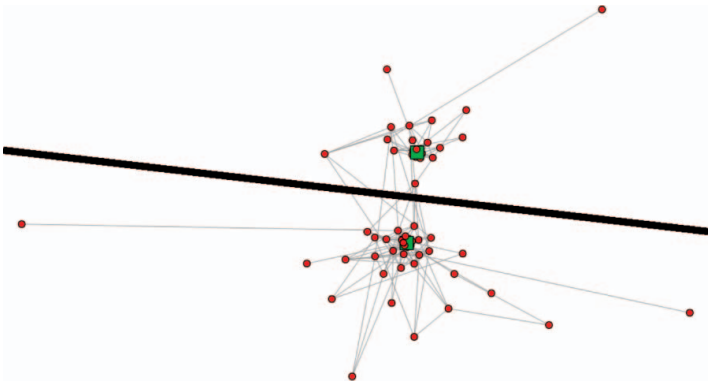

Figure 6. Results of *PageRank-Display* ($\alpha = 0.3$) on the dolphin social network [Lusseau et al. 03], separating the dolphins into two communities (color figure available online).

From the definition of a (k, h, β, ϵ)-clusterable graph, each subset T of S has Cheeger ratio at least $c\sqrt{h \log n}$. This allows us to use Lemma 5.5 for any segment subset T_j^f, as defined in (5.2), with volume at most $(1 - \epsilon/2)\mathrm{vol}(S)$ defined by the function f as in Lemma 5.5. Altogether, we have that for any subset $R \subset S$ with $\mathrm{vol}(R) \leq (1 - \epsilon/2)\mathrm{vol}(S)$,

$$|[\rho(\alpha, v)](R) - [\rho(\alpha, \rho(\alpha, v)](R)| \leq \alpha t + \sqrt{n} e^{-(c^2 th \log n)/8} \leq \frac{\epsilon}{4}$$

by the assumption that $c = 8\sqrt{\beta/k}/\epsilon$, and choosing $t = 1/(hc^2)$. This implies that for any subset $R \subset S$ and any vertex v, we have

$$|[\rho(\alpha, v)](R) - [\rho(\alpha, \rho(\alpha, v)](R)| \leq \frac{\epsilon}{2}.$$

Thus the total variation distance between the two PageRank vectors is

$$\Delta_{TV}(\alpha) = \max_v \max_{R \subseteq S} [\rho(\alpha, v) - \rho(\alpha, \rho(\alpha, v))](R) \leq \frac{\epsilon}{2}.$$

Note that $\sqrt{\Phi(\alpha)}$ is just the so-called χ-square distance Δ_χ. Using the same technique as in [Aldous and Fill 12], we have

$$\Delta_{TV} \leq \Delta_\chi \leq \sqrt{1 - (1 - 2\Delta_{TV})^2}.$$

Thus, we conclude that $\Phi(\alpha) \leq \epsilon$, as desired. □

We will also show that the sampling methods in *PageRank-clusteringA* will ensure that with high probability, the cluster centers $\{c_1, \ldots, c_k\}$ will include one from the core of each of k partitions in a clusterable graph:

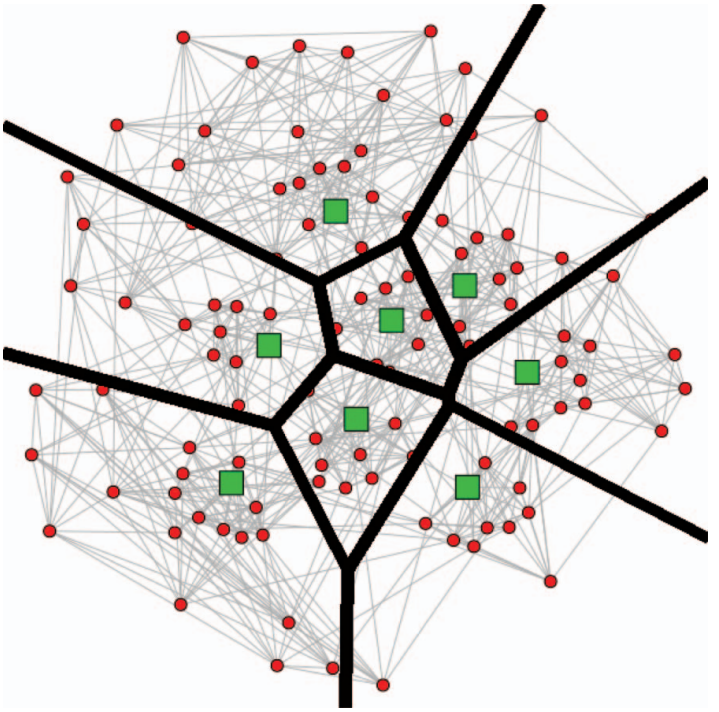

Figure 7. Results of *PageRank-Display* ($\alpha = 0.1$) on the football game network [Girvan and Newman 02], highlighting eight of the major collegiate conferences (color figure available online).

Lemma 6.3. *Suppose G is (h, k, β, ϵ)-clusterable, and $c \log n$ sets of k potential centers are chosen from G according to the stationary distribution π, where c is some absolute constant. With probability $1 - o(1)$, at least one set will contain one vertex from the core of each of the k clusters.*

Proof. Let $S_1, \ldots, S_k$ be a partition of (h, k, β, ϵ)-clusterable G, and let S_i' be the core of S_i. Suppose vertices $C = \{c_1, \ldots, c_k\}$ are chosen randomly according to π, and let $E(C)$ be the event that each $c_i \in S_i'$. Then we have

$$\Pr[\mathrm{E}(C)] \geq \prod_{i=1}^{k} \Pr[c_i \in S_i'] = \prod_{i=1}^{k} \frac{\mathrm{vol}(S_i')}{\mathrm{vol}(G)} \geq \prod_{i=1}^{k} \frac{(1-\epsilon)\mathrm{vol}(S_i')}{\mathrm{vol}(G)}$$
$$\geq \prod_{i=1}^{k} \frac{(1-\epsilon)\beta \mathrm{vol}(G)}{k\mathrm{vol}(G)} = \left(\frac{\beta(1-\epsilon)}{k}\right)^k.$$

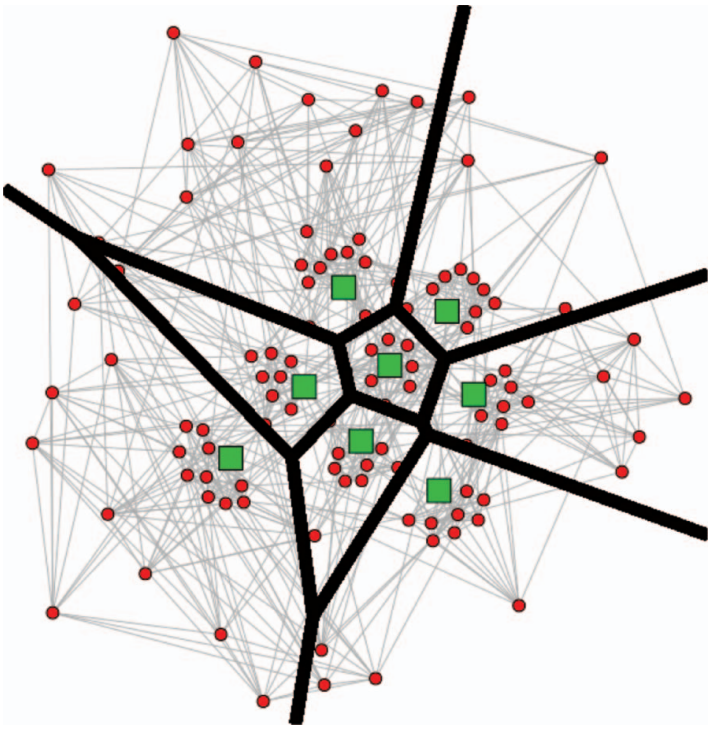

Figure 8. Results of *PageRank-Display* ($\alpha = 0.3$) on the football game network [Girvan and Newman 02], highlighting eight of the major collegiate conferences (color figure available online).

If $c \log n$ sets $C_1, \ldots, C_{c \log n}$ of k centers are sampled independently, the probability that at least one contains each $c_i \in S_i'$ is

$$\Pr[E(C_1) \vee \cdots \vee E(C_{c \log n})] \geq 1 - \prod_{i=1}^{c \log n} \Pr[\neg E(C_i)]$$

$$= 1 - \prod_{i=1}^{c \log n} (1 - \Pr[E(C_i)]) \geq 1 - \left(1 - \left(\frac{\beta(1-\epsilon)}{k}\right)^k\right)^{c \log n} = 1 - o(1).$$

□

This series of lemmas then leads to the proof of Theorem 4.1, showing the correctness of *PageRank-clusteringA*.

Proof of Theorem 4.1. We note that $\rho(0, s) = \pi$ and $\rho(1, s) = s$ for any distribution s. This implies that $\Phi(0) = \Phi(1) = \Psi(0) = 0$ and $\Psi(1) = n - 1$. It is not hard to check that Ψ is an increasing function, since $\Psi'(\alpha) > 0$ for $\alpha \in (0, 1]$. The function

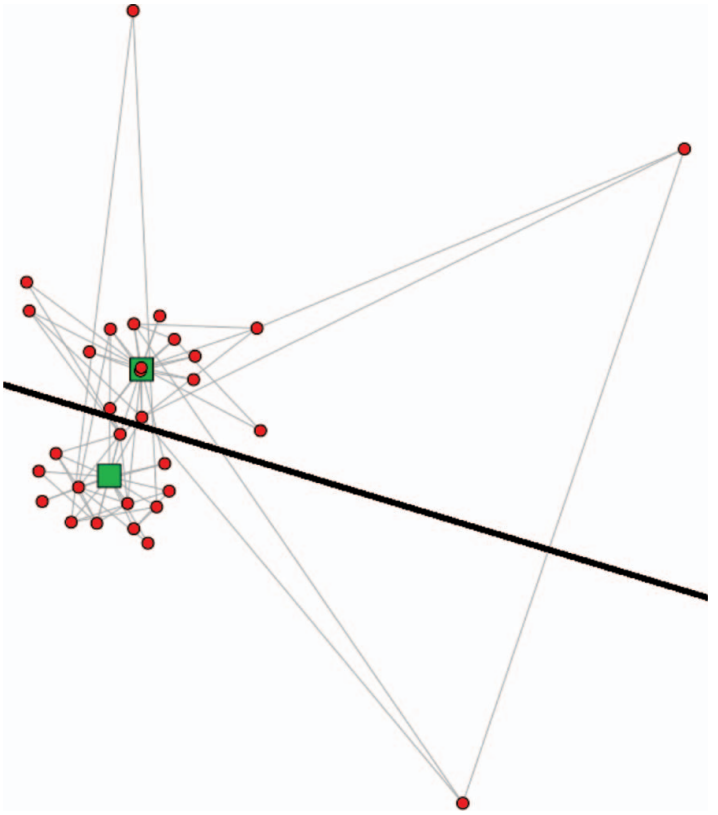

Figure 9. Results of *PageRank-Display* on Zachary's karate network [Zachary 77] (color figure available online).

of particular interest is Φ. Since we wish to find α such that Φ is small, it suffices to check the roots of Φ' for an α such that $\Phi(\alpha) < \epsilon$, which our algorithm does. Such an α exists due to Lemma 6.2.

Suppose α is a root of Φ'. To find k clusters, we can further restrict ourselves to the case of $\Psi(\alpha) \geq k - 2 - \epsilon$ by Lemma 6.1.

We note that by sampling $c \log n$ sets of k vertices from π, for sufficiently large c, the values $\mu(C)$ and $\Psi(C)$ for one such random set of k centers are close to $\Phi(\alpha)$ and $\Psi(\alpha)$, respectively, with high probability (exponentially decreasing depending on c and β) by probabilistic concentration arguments. In this context, the upper bound ϵ for $\mu(C)$ implies that the set consisting of distributions $\rho(\alpha, c)$ for $c \in C$ serves well as the set of centers of mass. Thus, the resulting Voronoi regions using C give the desired clusters. This proves the correctness of our clustering algorithm with high probability for (k, h, β, ϵ)-clusterable graphs. □

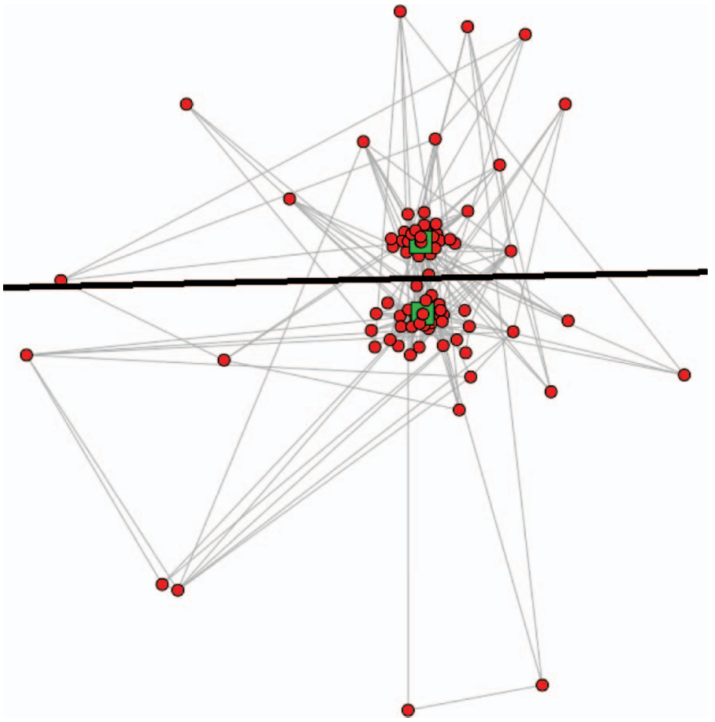

Figure 10. Results of *PageRank-Display* on a network of political books about the 2004 US presidential election [Krebs 11]. Edges are present between two books if they were frequently purchased together (color figure available online).

To illustrate *PageRank-clusteringB*, we consider a dumbbell graph U as an example. This graph U has two complete graphs K_{20} connected by a single edge, yielding a Cheeger ratio of $h \approx 0.0026$. Plotting $\Phi(\alpha)$ (Figure 1) and its derivative (Figure 2) shows that there is a local minimum near $\alpha \approx 0.018$. When Ψ is large, many individual nodes have personalized PageRank vectors that differ greatly from the overall distribution. This indicates that there are many nodes that are more representative of a small cluster than the entire graph. By plotting $\Psi(\alpha)$ (Figure 3) and its derivative (Figure 4), we can see that there is a distinct inflection point in the plot of Ψ for the dumbbell graph U as well.

7. A Graph-Drawing Algorithm Using PageRank

The visualization of complex graphs provides many computational challenges. Graphs such as the World Wide Web and social networks are known to exhibit

ubiquitous structure, including power-law distributions, small-world phenomena, and a community structure [Albert et al. 99, Broder et al. 00, Faloutsos et al. 99]. With large graphs, it is easy for such intricate structures to be lost in the sheer quantity of nodes and edges, which can result in drawings that reflect a network's size but not necessarily its structure.

Given a set of nodes S, we can extract communities around each node and determine the layout of the graph using personalized PageRank. The arrangement can be done using a force-based graph layout algorithm such as the Kamada–Kawai algorithm [Kamada and Kawai 89]. The goal is to capture local communities; we can do this by assigning edges $\{s, v\}$ for each $s \in S$ and $v \in V \setminus S$ with weight inversely proportional to the personalized PageRank. In this way, unrelated nodes with low PageRank will be forced to be distant, and close communities will remain close together. We also add edges $\{s, s'\}$ for $s, s' \in S$ with large weight to encourage separation of the individual communities. We use an implementation from Graphviz [Gansner and North 00].

We note that because force-based algorithms are simulations, they do not guarantee the exact cluster structure, but we will illustrate that it works well in practice. Additionally, there are algorithms specifically designed for clustered graph visualization [Eades and Feng 96, Parker et al. 98] and highlighting high-ranking nodes [Brandes and Cornelsen 01], but they impose a considerable amount of artificial hierarchical structure on the drawing and often require precomputing the clusters. Once we have a layout for all the nodes in the graph, we can partition them using a Voronoi diagram. We compute the Voronoi diagram efficiently using Fortune's algorithm [Fortune 86].

We tie together personalized PageRank and Voronoi diagrams in the algorithm *PageRank-display*, Algorithm 3.

The jumping constant α is associated with the scale of the clustering. We can determine α either by trial and error or by optimizing Φ and Ψ as in Section 4. As long as G is connected, the PageRank vector will be nonzero on every vertex. Using the algorithms from [Andersen et al. 06, Chung and Zhao 10], the approximation factor ϵ acts as a cutoff, and any node v with PageRank less than ϵd_v will be assigned zero. This is advantageous because the support of the approximate PageRank vector will be limited to the local community containing its seed. In *PageRank-display*, we give weights to the edges equal to $1/p_s(v)$, but this is problematic if $p_s(v) = 0$. In that case, we omit the edge from G' entirely.

We remark that the selection of ϵ will influence the size of the local communities: the subset of nodes with nonzero approximate PageRank has volume at most $\frac{2}{(1-\alpha)\epsilon}$ (see [Andersen et al. 06]). This implies that a good selection of ϵ is $O(\frac{|S|}{(1-\alpha)\mathrm{vol}(G)})$.

Algorithm 3: PageRank-display

Input: $G = (V, E)$, S, α, ϵ
Output: A graph drawing
for all $s \in S$ **do**
Compute an approximate PageRank vector $p_s = \rho(\alpha, s)$
end for
Let G' be a graph with vertex set V
for all $s \in S$ and $v \in V \setminus S$ **do**
Add an edge $\{s, v\}$ to G' with weight $1/p_s(v)$, as long as $p_s(v) > 0$
end for
for all $s, s' \in S$ **do**
Add an edge $\{s, s'\}$ to G' with weight $10 \times \max_{s,v} 1/p_s(v)$
end for
Use a force-based display algorithm on G' to determine coordinates c_v for each $v \in V$
Compute the Voronoi diagram on S
Draw G using the coordinates c_v, highlighting S with a different color, and overlaying the Voronoi diagram

We also remark that the selection of S is important. If S contains vertices that are not part of communities or two nodes in the same community, then there will be no structure to display. In general, the selection of S is similar to the geometric problem of finding a set of points with minimum covering radius, which can be intractable (see [Guruswami et al. 05]). There are several algorithms that can automatically choose S, including *PageRank-clustering* as presented here.

We used our algorithm to demonstrate and highlight the existence of local structure in two real-world data sets. The first data set is a social network among 62 dolphins [Lusseau et al. 03]. While the graph exhibits traditional network structure such as small-world phenomena, one can see in Figures 5 and 6 that the dolphins can be divided into two communities, with just a few connected to both sides. Note that with larger α, the far-flung nodes become more isolated, making the communities appear denser.

A more interesting example is shown in Figures 7 and 8. The vertices represent 114 NCAA Division I American collegiate football teams, with edges connecting two teams if they played against each other during the year 2000 football season. The league is divided into many smaller conferences of up to 12 teams; for each team, about half of its games are played against conference opponents, and the rest are played against nonconference teams. An appropriate selection of the eight highlighted teams in Figures 7 and 8 reveal a partition that separates their eight respective conferences, and teams from the remaining conferences are placed on the periphery of the drawing. Here, the larger α is more effective, since

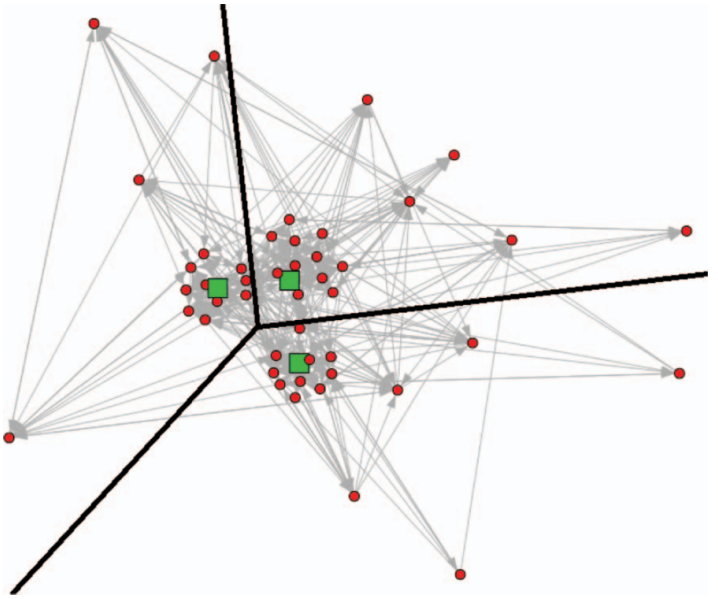

Figure 11. Results of *PageRank-Display* on a network of US Air Force flying teams [de Nooy et al. 04, Chapter 4] (color figure available online).

the PageRank is more concentrated near the community centers. Several more graph-drawing examples are shown in Figures 9 through 11.

References

[Albert et al. 99] R. Albert, A.-L. Barabási, and H. Jeong. "Diameter of the World Wide Web." *Nature* 401 (1999), 130–131.

[Aldous and Fill 12] D. Aldous and J. Fill. "Reversible Markov Chains and Random Graphs on Graphs." In preparation, 2012.

[Aloise et al. 09] D. Aloise, A. Deshpande, P. Hansen, and P. Popat. "NP-Hardness of Euclidean Sum-of-Squares Clustering." *Machine Learning* 75:2 (2009), 245–248.

[Andersen and Chung 07] R. Andersen and F. Chung. "Detecting Sharp Drops in PageRank and a Simplified Local Partitioning Algorithm." In *Proceedings of the 4th International Conference Theory and Applications of Models of Computation*, pp. 1–12, 2007.

[Andersen et al. 06] R. Andersen, F. Chung, and K. Lang. "Local Graph Partitioning Using PageRank Vectors." In *Proceedings of the 47th Annual IEEE Symposium on Foundation of Computer Science* (FOCS 2006), pp. 475–486, 2006.

[Brandes and Cornelsen 01] U. Brandes and S. Cornelsen. "Visual Ranking of Link Structures." In *Proceedings of the 7th International Workshop on Algorithms and Data Structures* (2001), pp. 222–233, 2001.

[Brin and Page 98] S. Brin and L. Page. "The Anatomy of a Large-Scale Hypertextual Web Search Engine." *Computer Networks and ISDN Systems* 30 (1998), 107–117.

[Broder et al. 00] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. "Graph Structure in the Web." *Computer Networks* 33 (2000), 1–6.

[Chung 10] F. Chung. "PageRank as a Discrete Green's Function, Geometry and Analysis, I." *ALM* 17 (2010), 285–302.

[Chung and Zhao 10] F. Chung and W. Zhao. "A Sharp PageRank Algorithm with Applications to Edge Ranking and Graph Sparsification." In *Proceedings of WAW 2010*, pp. 2–14, 2010.

[Chung et al. 09] F. Chung, P. Horn, and A. Tsiatas. "Distributing Antidote Using PageRank Vectors." *Internet Mathematics* 6:2 (2009), 237–254.

[de Nooy et al. 04] W. de Nooy, A. Mrvar, and V. Batagelj. *Exploratory Social Network Analysis with Pajek.* Cambridge: Cambridge University Press, 2004.

[Dyer and Frieze 85] M. E. Dyer and A.M. Frieze. "A Simple Heuristic for the p-Centre Problem." *Operations Research Letters* 3:6 (1985), 285–288.

[Eades and Feng 96] P. Eades and Q. Feng. "Multilevel Visualization of Clustered Graphs." In *Proceedings of the International Symposium on Graph Drawing*, pp. 101–112, 1996.

[Enright et al. 02] A. J. Enright, S. Van Dongen, and C. A. Ouzounis. "An Efficient Algorithm for Large-Scale Detection of Protein Families." *Nucleic Acids Research* 30:7 (2002), 1575–1584.

[Faloutsos et al. 99] M. Faloutsos, P. Faloutsos, and C. Faloutsos. "On Power-Law Relationships of the Internet Topology." In *Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication* (SIGCOMM 1999), pp. 251–262, 1999.

[Fortune 86] S. Fortune. "A Sweepline Algorithm for Voronoi Diagrams." In *Proceedings of the Second Annual Symposium on Computational Geometry*, pp. 313–322, 1986.

[Gansner and North 00] E. Gansner and C. North. "An Open Graph Visualization System and Its Applications to Software Engineering." *Software—Practice and Experience* 30:11 (2000), 1203–1233.

[Girvan and Newman 02] M. Girvan and M. E. J. Newman. "Community Structure in Social and Biological Networks." *Proceedings of the National Academy of Sciences* 99:12 (2002), 7821–7826.

[Guruswami et al. 05] V. Guruswami, D. Micciancio, and O. Regev. "The Complexity of the Covering Radius Problem on Lattices and Codes." *Computational Complexity* 14:2 (2005), 90–120.

[Harel and Koren 02] D. Harel and Y. Koren. "Graph Drawing by High-Dimensional Embedding." In *Proceedings of the 10th International Symposium on Graph Drawing* pp. 207–219, 2002.

[Jeh and Widom 03] G. Jeh and J. Widom. "Scaling Personalized Web Search." In *Proceedings of the 12th International Conference on World Wide Web*, pp. 271–279, 2003.

[Kamada and Kawai 89] T. Kamada and S. Kawai. "An Algorithm for Drawing General Undirected Graphs." *Information Processing Letters* 31:1 (1989), 7–15.

[Krebs 11] V. Krebs. "Social Network Analysis Software & Services for Organizations, Communities, and Their Consultants." Available online (http://www.orgnet.com/ and http://www-personal.umich.edu/~mejn/netdata/), 2011.

[Lloyd 82] S. Lloyd. "Least Square Quantization in PCM." *IEEE Transactions on Information Theory* 28:2 (1982), 129–137.

[Lovász and Simonovits 93] L. Lovász and M. Simonovits. "Random Walks in a Convex Body and an Improved Volume Algorithm." *Random Structures and Algorithms* 4 (1993), 359–412.

[Lusseau et al. 03] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson. "The Bottlenose Dolphin Community of Doubtful Sound Features a Large Proportion of Long-Lasting Associations." *Behavioral Ecology and Sociobiology* 54:4 (2003), 396–405.

[MacQueen 67] J. MacQueen. "Some Methods for Classification and Analysis of Multivariate Observations." In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297, 1967.

[Mancoridis et al. 99] S. Mancoridis, B. S. Mitchell, Y. Chen, and E. R. Gansner. "Bunch: A Clustering Tool for the Recovery and Maintenance of Software System Structures." In *Proceedings of the IEEE International Conference on Software Maintenance*, pp. 50–59, 1999.

[Moody 01] J. Moody. "Peer Influence Groups: Identifying Dense Clusters in Large Networks." *Social Networks* 23:4 (2001), 261–283.

[Newman and Girvan 04] M. E. J. Newman and M. Girvan. "Finding and Evaluating Community Structure in Networks." *Physical Review E* 69 (2004), 026113.

[Ng et al. 02] A. Ng, M. Jordan, and Y. Weiss. "On Spectral Clustering: Analysis and an Algorithm." *Advances in Neural Information Processing Systems* 14:2 (2002), 849–856.

[Noack 09] A. Noack. "Modularity Clustering Is Force-Directed Layout." *Physical Review E* 79 (2009), 026102.

[Parker et al. 98] G. Parker, G. Franck, and C. Ware. "Visualization of Large Nested Graphs in 3D: Navigation and Interaction." *Journal of Visual Languages and Computing* 9:3 (1998), 299–317.

[Rudelson and Vershynin 07] M. Rudelson and R. Vershynin. "Sampling from Large Matrices: An Approach through Geometric Functional Analysis." *Journal of the ACM* 54:4 (2007), Article 21.

[Schaeffer 07] S. E. Schaeffer. "Graph Clustering." *Computer Science Review* 1:1 (2007), 27–64.

[Shi and Malik 00] J. Shi and J. Malik. "Normalized Cuts and Image Segmentation." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22:8 (2000), 888–905.

[Zachary 77] W. W. Zachary. "An Information Flow Model for Conflict and Fission in Small Groups." *Journal of Anthropological Research* 33 (1977), 452–473.

Fan Chung, Department of Computer Science and Engineering, University of California, San Diego, 9500 Gilman Drive, Mail Code 0404, La Jolla, CA 92093-0404, USA (fan@cs.ucsd.edu)

Alexander Tsiatas, Department of Computer Science and Engineering, University of California, San Diego, 9500 Gilman Drive, Mail Code 0404, La Jolla, CA 92093-0404, USA (atsiatas@cs.ucsd.edu)

Received April 1, 2011; accepted June 9, 2011.