

# Strong Localization in Personalized PageRank Vectors

Huda Nassar<sup>1</sup>(✉), Kyle Kloster<sup>2</sup>, and David F. Gleich<sup>1</sup>

<sup>1</sup> Computer Science Department, Purdue University, West Lafayette, USA  
{hnassar,dgleich}@purdue.edu

<sup>2</sup> Mathematics Department, Purdue University, West Lafayette, USA  
kkloste@purdue.edu

**Abstract.** The personalized PageRank diffusion is a fundamental tool in network analysis tasks like community detection and link prediction. It models the spread of a quantity from a set of seed nodes, and it has been observed to stay localized near this seed set. We derive an upper-bound on the number of entries necessary to approximate a personalized PageRank vector in graphs with skewed degree sequences. This bound shows localization under mild assumptions on the maximum and minimum degrees. Experimental results on random graphs with these degree sequences show the bound is loose and support a conjectured bound.

**Keywords:** PageRank · Diffusion · Local algorithms

## 1 Introduction

Personalized PageRank vectors [23] are a ubiquitous tool in data analysis of networks in biology [12, 21] and information-relational domains such as recommender systems and databases [15, 17, 22]. In contrast to the standard PageRank vector, personalized PageRank vectors model a random-walk process on a network that randomly returns to a fixed starting node instead of restarting from a random node in the network as in the traditional PageRank. This process is also called a random-walk with restart.

The stationary distributions of the resulting process are typically called personalized PageRank vectors. We prefer the terms “localized PageRank” or “seeded PageRank” as these choices are not as tied to PageRank’s origins on the web. A seeded PageRank vector depends on three terms: the network modeled as a column-stochastic matrix  $\mathbf{P}$  characterizing the random-walk process, a parameter  $\alpha$  that determines the restart probability  $(1 - \alpha)$ , and a seed node  $s$ . The vector  $\mathbf{e}_s$  is the vector of all zeros with a single 1 in the position corresponding to node  $s$ . The seeded PageRank vector  $\mathbf{x}$  is then the solution of the linear system:

$$(\mathbf{I} - \alpha\mathbf{P})\mathbf{x} = (1 - \alpha)\mathbf{e}_s.$$

---

K. Kloster and D.F. Gleich—Supported by NSF CAREER award CCF-1149756 and DARPA SIMPLEX Code available online <https://github.com/nassarhuda/pprlocal>.

When the network is strongly connected, the solution  $\mathbf{x}$  is non-zero for all nodes. This is because there is a non-zero probability of walking from the seed to any other node in a strongly connected network. Nevertheless, the solution  $\mathbf{x}$  displays a behavior called *localization*. We can attain accurate localized PageRank solutions by truncating small elements of  $\mathbf{x}$  to zero. Put another way, there is a sparse vector  $\mathbf{x}_\varepsilon$  that approximates  $\mathbf{x}$  to an accuracy of  $\varepsilon$ . This behavior is desirable for applications of seeded PageRank because they typically seek to “highlight” a small region related to the seed node  $s$  inside a large graph.

The essential question we study in this paper is: how sparse can we make  $\mathbf{x}_\varepsilon$ ? To be precise, we consider a notion of strong localization,  $\|\mathbf{x}_\varepsilon - \mathbf{x}\|_1 \leq \varepsilon$ , and we focus on the behavior of  $f(\varepsilon) := \min \text{nonzeros}(\mathbf{x}_\varepsilon)$ . Note that  $\mathbf{x}_\varepsilon$  depends on  $\alpha$ , the particular random-walk on the graph  $\mathbf{P}$ , and the seed node  $s$  from which the PageRank diffusion begins. We only consider stochastic matrices  $\mathbf{P}$  that arise from random-walks on strongly-connected graphs. So a more precise statement of our goal is:

$$f_\alpha(\varepsilon) = \max_{\mathbf{P}} \max_s \min_{\mathbf{x}_\varepsilon} \text{nonzeros}(\mathbf{x}_\varepsilon) \text{ where } \|\mathbf{x}_\varepsilon - \mathbf{x}(\alpha, \mathbf{P}, s)\|_1 \leq \varepsilon,$$

and where  $\mathbf{x}(\alpha, \mathbf{P}, s)$  is the seeded PageRank vector  $(1 - \alpha)(\mathbf{I} - \alpha\mathbf{P})^{-1}\mathbf{e}_s$ . The goal is to establish bounds on  $f(\varepsilon)$  that are sublinear in  $n$ , the number of nodes of the graph, because that implies localized solutions to PageRank.

Adversarial localized PageRank constructions exist where the solutions  $\mathbf{x}$  are near the uniform distribution (see Sect. 2). Thus, it is not possible to meaningfully bound  $f(\varepsilon)$  as anything other than  $n$ . It is also known that  $f(\varepsilon)$  is sublinear in  $n$  for graphs that are essentially of bounded maximum degree [6] due to resolvent theory. The case of skewed degrees was open until our result.

We establish an upper-bound on  $f_\alpha(\varepsilon)$  as a function of the rate of decay of the degree sequence,  $1/\varepsilon$ ,  $\alpha$ , the maximum degree  $d$ , and the minimum degree  $\delta$  (Theorem 1). This bound enables us to establish sublinear localization for graphs with growing maximum degrees provided that the other node degrees decay sufficiently rapidly. When we study this bound in random realizations of appropriate networks, it turns out to be loose; hence, we develop a new conjectured bound (Sect. 4).

## 1.1 Related Work on Weak Localization

There is another notion of localization that appears in uses of PageRank for partitioning undirected graphs:

$$\|\mathbf{D}^{-1}(\mathbf{x}_\varepsilon - \mathbf{x})\|_\infty = \max_i |[x_\varepsilon]_i - x_i|/d_i \leq \varepsilon.$$

If this notion is used for a localized Cheeger inequality [1, 10], then we need the additional property that  $0 \leq \mathbf{x}_\varepsilon \leq \mathbf{x}$  element-wise. When restated as a localization result, the famous Andersen-Chung-Lang PageRank partitioning result [1] includes a proof that:

$$\max_{\mathbf{P}} \max_s \min_{\mathbf{x}_\varepsilon} \text{nonzeros}(\mathbf{x}_\varepsilon) \leq \frac{1}{1-\alpha} \frac{1}{\varepsilon}, \text{ where } \|\mathbf{D}^{-1}(\mathbf{x}_\varepsilon - \mathbf{x}(\alpha, \mathbf{P}, s))\|_\infty \leq \varepsilon.$$

This establishes that *any* uniform random walk on a graph satisfies a weak-localization property. The paper also gives a fast algorithm to find these weakly local solutions. More recently, there have appeared a variety of additional weak-localization results on diffusions [13, 19].

## 1.2 Related Work on Functions of Matrices and Diffusions

Localization in diffusions is broadly related to localization in functions of matrices [6]. The results in that literature tend to focus on the case of banded matrices (e.g. [5]), although there are also discussions of more general results in terms of graphs arising from sparse matrices [6]. These same types of decay bounds can apply to a variety of graph diffusion models that involve a stochastic matrix [3, 16], and recent work shows that they may even extend beyond this regime [13]. In the context of the decay of functions of matrices, we advance the literature by proving a localization bound for a particular resolvent function of a matrix that applies to graphs with growing maximum degree.

## 2 A Negative Result for Strong Localization

Here we give an example of a graph that always has a non-local seeded PageRank vector. More concretely, we demonstrate the existence of a personalized PageRank vector that requires  $\Theta(n)$  nonzeros to attain a 1-norm accuracy of  $\varepsilon$ , where  $n$  is the number of nodes in the graph.

The graph is just the undirected star graph on  $n$  nodes. Then the PageRank vector  $\mathbf{x}$  seeded on the center node has value  $1/(1 + \alpha)$  for the center node and  $\alpha/((1 + \alpha)(n - 1))$  for all leaf nodes. Suppose an approximation  $\hat{\mathbf{x}}$  of  $\mathbf{x}$  has  $M$  of these leaf-node entries set to 0. Then the 1-norm error  $\|\mathbf{x} - \hat{\mathbf{x}}\|_1$  would be at least  $M\alpha/((1 + \alpha)(n - 1))$ . Attaining a 1-norm accuracy of  $\varepsilon$  requires  $M\alpha/((1 + \alpha)(n - 1)) < \varepsilon$ , and so the minimum number of entries of the approximate PageRank vector required to be non-zero ( $n - M$ ) is then lower-bounded by  $n(1 - c) + c$ , where  $c = \varepsilon(1 + \alpha)/\alpha$ . Note that this requires  $c \in (0, 1)$ , which holds if  $\varepsilon < \alpha/2$ . Thus, the number of nonzeros required in the approximate PageRank vector must be linear in  $n$ .

## 3 Localization in Personalized PageRank

The example in Sect. 2 demonstrates that there exist seeded PageRank vectors that are *non-local*. Here we show that graphs with a particular type of skewed degree sequence and a growing, but sublinear, maximum degree have seeded PageRank vectors that are always localized, and we give an upper-bound on  $f(\varepsilon)$  for this class of graph. This theorem originates in our recent work on seeded heat kernel vectors [14], and we now employ similar arguments to treat seeded PageRank vectors. Our present analysis yields tighter intermediate inequalities and results in an entirely novel bound for the localization of PageRank.

**Theorem 1.** *Let  $\mathbf{P}$  be a uniform random walk transition matrix of a graph on  $n$  nodes with maximum degree  $d$  and minimum degree  $\delta$ . Additionally, suppose that the  $k$ th largest degree,  $d(k)$ , satisfies  $d(k) \leq \max\{dk^{-p}, \delta\}$ . The Gauss-Southwell coordinate relaxation method applied to the seeded PageRank problem  $(\mathbf{I} - \alpha\mathbf{P})\mathbf{x} = (1 - \alpha)\mathbf{e}_s$  produces an approximation  $\mathbf{x}_\varepsilon$  satisfying  $\|\mathbf{x} - \mathbf{x}_\varepsilon\|_1 < \varepsilon$  having at most  $N$  non-zeros in the solution, where  $N$  satisfies*

$$N = \min \left\{ n, \frac{1}{\delta} C_p \left( \frac{1}{\varepsilon} \right)^{\frac{\delta}{1-\alpha}} \right\}, \quad (1)$$

and where we define  $C_p$  to be

$$\begin{aligned} C_p &:= d(1 + \log d) && \text{if } p = 1 \\ &:= d \left( 1 + \frac{1}{1-p} \left( d^{\frac{1}{p}-1} - 1 \right) \right) && \text{otherwise.} \end{aligned}$$

Note that the upper bound  $N = n$  is trivial as a vector cannot have more non-zeros than entries. Thus,  $d$ ,  $\delta$ ,  $p$ , and  $n$  must satisfy certain conditions to ensure that inequality (1) is not trivial. In particular, for values of  $p < 1$ , it is necessary that  $d = o(n^p)$  for inequality (1) to imply that  $N = o(n)$ . For  $p > 1$ , the bound guarantees sublinear growth of  $N$  as long as  $d = o(n)$ . Additionally, the minimum degree  $\delta$  must be bounded by  $O(\log \log n)$ . Thus we arrive at:

**Corollary 1.** *Let  $G$  be a class of graphs with degree sequences obeying the conditions of Theorem 1 with constant  $\delta$  and  $d = o(n^{\min(p,1)})$ . Then  $f(\varepsilon) = o(n)$ , and seeded PageRank vectors are localized.*

We also note that the theorem implies localized seeded PageRank vectors for any graph with a maximum degree  $d = O(\log \log n)$ .

### 3.1 Our Class of Skewed Degree Sequences

We wish to make a few remarks about the class of skewed degree sequences where our results apply. Perhaps the most well-known is the power-law degree distribution where the probability that a node has degree  $k$  is proportional to  $k^{-\gamma}$ . These power-laws can be related to our skewed sequences with  $p = 1/(\gamma - 1)$  and  $d = O(n^p)$  [2]. This setting renders our bound trivial with  $n$  nonzeros. Nevertheless, there is evidence that some real-world networks exhibit our type of skewed degrees [11] where the bound is asymptotically non-trivial.

### 3.2 Deriving the Bound

Getting back to the proof, our goal is an  $\varepsilon$ -approximation,  $\mathbf{x}_\varepsilon$ , to the equation  $(\mathbf{I} - \alpha\mathbf{P})\mathbf{x} = (1 - \alpha)\mathbf{e}_s$  for a seed  $s$ . Given an approximation,  $\hat{\mathbf{x}}$ , we can express the error in terms of the residual vector  $\mathbf{r} = (1 - \alpha)\mathbf{e}_s - (\mathbf{I} - \alpha\mathbf{P})\hat{\mathbf{x}}$  as follows:

$$\mathbf{x} - \hat{\mathbf{x}} = (\mathbf{I} - \alpha\mathbf{P})^{-1} \mathbf{r}. \quad (2)$$

Using this relationship, we can bound our approximation’s 1-norm accuracy,  $\|\mathbf{x} - \hat{\mathbf{x}}\|_1$ , with the quantity  $\frac{1}{1-\alpha} \|\mathbf{r}\|_1$ . This is because the column-stochasticity of  $\mathbf{P}$  implies that  $\|(\mathbf{I} - \alpha\mathbf{P})^{-1}\|_1 = \frac{1}{1-\alpha}$ . Guaranteeing a 1-norm error  $\|\mathbf{x} - \hat{\mathbf{x}}\|_1 < \varepsilon$  is then a matter of ensuring that  $\|\mathbf{r}\|_1 < (1-\alpha)\varepsilon$  holds. To bound the residual norm, we look more closely at a particular method for producing the approximation.

*The Gauss-Southwell iteration.* The Gauss-Southwell algorithm is a coordinate relaxation method for solving a linear system akin to the Gauss-Seidel linear solver. When solving a linear system, the Gauss-Southwell method proceeds by updating the entry of the approximate solution that corresponds to the largest magnitude entry of the residual,  $\mathbf{r}$ . We describe the Gauss-Southwell update as it is used to solve the seeded PageRank linear system.

The algorithm begins by setting the initial solution  $\mathbf{x}^{(0)} = 0$  and  $\mathbf{r}^{(0)} = (1-\alpha)\mathbf{e}_s$ . In step  $k$ , let  $j = j(k)$  be the entry of  $\mathbf{r}^{(k)}$  with the largest magnitude, and let  $m = |\mathbf{r}_j^{(k)}|$ . We update the solution  $\mathbf{x}^{(k)}$  and residual as follows:

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + m\mathbf{e}_j \tag{3}$$

$$\mathbf{r}^{(k+1)} = \mathbf{e}_s - (\mathbf{I} - \alpha\mathbf{P})\mathbf{x}^{(k+1)}, \tag{4}$$

and the residual update can be expanded to  $\mathbf{r}^{(k+1)} = \mathbf{r}^{(k)} - m\mathbf{e}_j + m\alpha\mathbf{P}\mathbf{e}_j$ . Since each update to the solution  $\mathbf{x}^{(k)}$  alters exactly one entry of the vector, the index  $k$  is an upper-bound on the number of non-zeros in the solution.

This application of Gauss-Southwell to seeded PageRank-style problems has appeared numerous times in recent literature [7, 8, 18, 20]. In at least one instance ([8], Sect. 5.2) the authors showed that the residual and solution vector stay nonnegative throughout this process, assuming the seed vector is nonnegative (which, in our context, it is). So the 1-norm of the residual can be expressed as  $\|\mathbf{r}^{(k+1)}\|_1 = \mathbf{e}^T \mathbf{r}^{(k+1)}$ , where  $\mathbf{e}$  is the vector of all ones. Expanding the residual in terms of the iterative update presented above, we can write the residual norm as  $\mathbf{e}^T (\mathbf{r}^{(k)} - m\mathbf{e}_j + m\alpha\mathbf{P}\mathbf{e}_j)$ . Then, denoting  $\|\mathbf{r}^{(k)}\|_1$  by  $r_k$ , yields the recurrence  $r_{k+1} = r_k - m(1-\alpha)$ .

Next observe that since  $m$  is the largest magnitude entry in  $\mathbf{r}$ , it is larger than the average value of  $\mathbf{r}$ . Let  $Z(k)$  denote the number of nonzero entries in  $\mathbf{r}^{(k)}$ ; then the average value can be expressed as  $r_k/Z(k)$ . Hence, we have  $m \geq r_k/Z(k)$ , and so we can bound  $r_k - m(1-\alpha)$  above by  $r_k - r_k(1-\alpha)/Z(k)$ . Thus,  $r_{k+1} \leq r_k (1 - (1-\alpha)/Z(k))$ , and we can recur to find:

$$r_{k+1} \leq r_0 \prod_{t=0}^k \left(1 - \frac{1-\alpha}{Z(t)}\right), \tag{5}$$

where  $r_0 = (1-\alpha)$  because  $\mathbf{r}_0 = (1-\alpha)\mathbf{e}_s$ . Then, using the fact that  $\log(1-x) \leq -x$  for  $x < 1$ , we note:

$$r_{k+1} \leq (1-\alpha) \prod_{t=0}^k \left(1 - \frac{1-\alpha}{Z(t)}\right) \leq (1-\alpha) \exp\left(- (1-\alpha) \sum_{t=0}^k \frac{1}{Z(t)}\right). \tag{6}$$

To progress from here we need some control over the quantity  $Z(t)$  and this is where our skewed degree sequence enters the proof.

### 3.3 Using the degree sequence

We show that for a graph with this kind of skewed degree sequence, the number of entries in the residual obeys:

$$Z(t) \leq C_p + \delta t, \tag{7}$$

where the term  $C_p$  is defined in the statement of Theorem 1. A similar analysis was presented in [14], but the current presentation improves the bound on  $C_p$ . This bound is proved below, but first we use this bound on  $Z(t)$  to control the bound on  $r_k$ . Lemma 5.6 from [14] implies that

$$\sum_{t=0}^k \frac{1}{Z(t)} \geq \frac{1}{\delta} \log \left( \frac{(\delta(k+1) + C_p)}{C_p} \right)$$

and so, plugging into (6), we can bound

$$r_{k+1} \leq (1 - \alpha) \exp \left( -\frac{(1-\alpha)}{\delta} \log \left( \frac{(\delta(k+1) + C_p)}{C_p} \right) \right),$$

which simplifies to  $r_{k+1} \leq (1 - \alpha) \left( \frac{(\delta(k+1) + C_p)}{C_p} \right)^{-(\alpha-1)/\delta}$ . Finally, to guarantee  $r_k < \varepsilon(1 - \alpha)$ , it suffices to choose  $k$  so that  $\left( \frac{(\delta k + C_p)}{C_p} \right)^{(\alpha-1)/\delta} \leq \varepsilon$ . This holds if and only if  $(\delta k + C_p) \geq C_p (1/\varepsilon)^{\delta/(\alpha-1)}$  holds, which is guaranteed by  $k \geq \frac{1}{\delta} C_p (1/\varepsilon)^{\delta/(1-\alpha)}$ . Thus,  $k = \frac{1}{\delta} C_p (1/\varepsilon)^{\delta/(1-\alpha)}$  steps will produce an  $\varepsilon$ -approximation. Each step introduces at most one non-zero, which implies that if  $k < n$ , then there is an approximation  $\mathbf{x}_\varepsilon$  with  $N = k < n$  non-zeros. If  $k \geq n$ , then this analysis produces the trivial bound  $N = n$ .

*Proving the degree sequence bound.* Here we prove the inequality in (7) used in the proof above. For additional details, see the proof of Lemma 5.5 in [14], which is similar but results in a slightly worse bound. First, observe that the number of nonzeros in the residual after  $t$  steps is bounded above by the sum of the largest  $t$  degrees,  $Z(t) \leq \sum_{k=1}^t d(k)$ . When we substitute the decay bound  $d(k) \leq dk^{-p}$  into this expression,  $d(k)$  is only a positive integer when  $k \leq (d/\delta)^{1/p}$ . Hence, we split the summation  $Z(t) \leq \sum_{k=1}^t d(k)$  into two pieces,

$$Z(t) \leq \sum_{k=1}^t d(k) \leq \left( \sum_{k=1}^{\lfloor (d/\delta)^{1/p} \rfloor} dk^{-p} \right) + \sum_{k=\lfloor (d/\delta)^{1/p} \rfloor + 1}^t \delta.$$

We want to prove that this implies  $Z(t) \leq C_p + \delta t$ . The second summand is always less than  $\delta t$ . The first summand can be bounded above by  $d \left( 1 + \int_1^{(d/\delta)^{1/p}} x^{-p} dx \right)$  using a right-hand integral rule. This integral is straightforward to bound above with the quantity  $C_p$  defined in Theorem 1. This completes the proof.

## 4 Experiments

We present experimental results on the localization of seeded PageRank vectors on random graphs that have our skewed degree sequence and compare the actual sparsity with the predictions of our theoretical bound. This involves generating random graphs with the given skewed degree sequence (Sect. 4.1) and then comparing the experimental localization with our theoretical bound (Sect. 4.3). The bound is not particularly accurate, and so we conjecture a new bound that better predicts the behavior witnessed (Sect. 4.4).

### 4.1 Generating the Graphs

For experimental comparison, we wanted a test suite of graphs with varying but specific sizes and degree sequences. To produce these graphs, we use the Bayati-Kim-Saberi procedure [4] for generating undirected graphs with a prescribed degree sequences. The degree sequences used follow our description in Theorem 1 precisely. We choose the maximum degree  $d$  to be  $n^{1/3}$  or  $n^{1/2}$  and the minimum degree to be  $\delta = 2$ . We use several values for the decay exponent  $p$ , stated below.

After generating the degree sequence, we use the Erdős-Gallai conditions and the Havel-Hakimi algorithm to check if it is graphical. If the previously generated sequence fails, we perturb the sequence slightly and recheck the conditions. It often fails because the sequence has an odd sum; to resolve this state it suffices to increase the degree of one of the nodes with minimum degree by 1. Lastly, we verify that the graph contains a large connected component. We proceed once a graph has been generated that meets the above conditions and has a largest connected component that includes at least  $n(1 - 10^{-2})$  nodes.

### 4.2 Measuring the Non-zeros

Given a graph, we first use the power method to compute a PageRank vector, seeded on the node with largest degree, to high-accuracy (1-norm error bounded by  $10^{-12}$ ). This requires  $\lfloor (\log(\varepsilon/2))/(\log(\alpha)) \rfloor$  iterations based on the geometric convergence rate of  $\alpha$ . We then study vectors  $\mathbf{x}_\varepsilon$  satisfying  $\|\mathbf{x}_\varepsilon - \mathbf{x}\|_1 \leq \varepsilon$ , for accuracies  $\varepsilon = \{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$ . To count the number of nonzeros in a vector  $\mathbf{x}_\varepsilon$  for a particular accuracy  $\varepsilon$ , we first recall:

$$f_\alpha(\varepsilon) = \max_{\mathbf{P}} \max_s \min_{\mathbf{x}_\varepsilon} \text{nonzeros}(\mathbf{x}_\varepsilon) \text{ where } \|\mathbf{x}_\varepsilon - \mathbf{x}(\alpha, \mathbf{P}, s)\|_1 \leq \varepsilon.$$

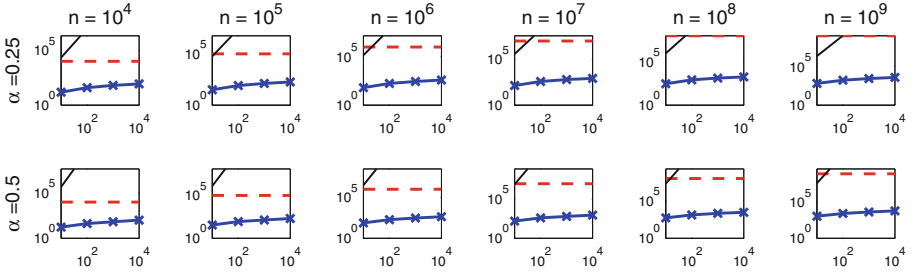
Thus, we need to compute  $\mathbf{x}_\varepsilon$  in a way that includes as many zeros as possible, subject to the constraint that the 1-norm of the difference between  $\mathbf{x}_\varepsilon$  and  $\mathbf{x}$  stays bounded by  $\varepsilon$ . The idea is to generate  $\mathbf{x}_\varepsilon$  by computing  $\mathbf{x}$  and deleting its smallest entries. The following steps illustrate our process to accomplish this:

- Compute the PageRank vector  $\mathbf{x}$  with accuracy  $10^{-12}$  via the power method.
- Sort  $\mathbf{x}$  in ascending order.
- Determine the largest index  $j$  so that  $(\sum_{k=1}^j \mathbf{x}_k) \leq \varepsilon$ .
- Truncate these  $j$  entries to 0. Then  $\mathbf{x}_\varepsilon$  contains  $n - j$  nonzeros.

### 4.3 Testing the Theoretical Bound

To test the effectiveness of our theoretical bound in Theorem 1 we generate graphs with decay exponent  $p = 0.95$ , with different sizes  $n = \{10^4, \dots, 10^9\}$ , and with maximum degree  $d = n^{1/3}$  and minimum degree  $\delta = 2$ . Then we solve the seeded PageRank system, seeded on the node of maximum degree, with  $\alpha = \{0.25, 0.5\}$ .

For a more compact notation, let  $\text{nonzeros}(\mathbf{x}_\varepsilon) = \text{nnz}(\mathbf{x}_\varepsilon)$  be the empirical number of nonzeros produced by our experiment. Figure 1 shows how  $\text{nnz}(\mathbf{x}_\varepsilon)$  varies with  $n$  and  $\alpha$ . Our theory gives a non-trivial bound once  $n > 10^6$  for  $\alpha = 0.25$  and  $n > 10^8$  for  $\alpha = 0.5$ . These values of  $\alpha$  (or near relatives) have been used in the literature [9, 24]. That said, the theoretical bound stays far from the plot of the sparsity of the  $\varepsilon$ -approximate diffusion. Since the theoretical bound behaves poorly even on the extreme points of the parameter settings, we wished for a tighter, empirical bound.



**Fig. 1.** A log-log plot of the quantity  $\text{nnz}(\mathbf{x}_\varepsilon)$  versus  $1/\varepsilon$  obtained for different experiments for  $\alpha = \{0.25, 0.5\}$ . We fix  $p = 0.95$  for all plots, and run experiments on graphs of sizes  $\{10^4, 10^5, 10^6, 10^7, 10^8, 10^9\}$ . We choose  $d = n^{1/3}$  and  $\delta = 2$ . The red dashed line represents a vector with all non-zeros present. The solid black line shows the bound predicted by Theorem 1. The blue curve shows the actual number of non-zeros found (Color figure online).

### 4.4 Empirical Non-zero Analysis

In this section, we develop a new bound that better predicts the scaling behavior of the number of nonzeros in  $\mathbf{x}_\varepsilon$  as other parameters vary. We do this by studying the relationships among  $\text{nnz}(\mathbf{x}_\varepsilon)$ ,  $\varepsilon$ , and  $p$  in a parametric study. Our goal is to find a function  $g$  where

$$\frac{\text{nnz}(\mathbf{x}_\varepsilon)}{d \log(d)} \text{ scales like } g(\alpha, \varepsilon, p). \quad (8)$$

(The choice of  $d \log(d)$  was inspired by our theoretical bound.) We first fix  $n = 10^6$ ,  $d = n^{1/2}$ , and  $p = 0.95$  and generate a graph as mentioned in Sect. 4.1. We then solve the PageRank problem and find the number of nonzeros for different  $\varepsilon$  values as mentioned in Sect. 4.2. We use  $\alpha = \{0.25, 0.3, 0.5, 0.65, 0.85\}$  and count the number of nonzeros in the diffusion vector based on four accuracy settings,



$\varepsilon = \{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$ . We then generate a log-log plot of  $\frac{\text{nnz}(\mathbf{x}_\varepsilon)}{d \log(d)}$  versus  $1/\varepsilon$  for the different values of  $\alpha$ . The outcome is illustrated in Fig. 2 (left).

From Fig. 2, we can see that as  $\alpha$  increases, the values  $\frac{\text{nnz}(\mathbf{x}_\varepsilon)}{d \log(d)}$  also increase, interestingly, nearly as a linear shift. Since we have seen that  $\frac{\text{nnz}(\mathbf{x}_\varepsilon)}{d \log(d)}$  seems to vary inversely with  $(1 - \alpha)$ , we specialize the form of  $g$  as  $g(\alpha, \varepsilon, p) = \frac{c_1}{(1-\alpha)} \cdot g_2(\varepsilon, p)$ .

We similarly derive a relation between  $\frac{\text{nnz}(\mathbf{x}_\varepsilon)}{d \log(d)}$  and  $p$ . Here, we fix  $n = 10^6$  and  $d = n^{1/2}$  then generate graphs with different decay exponents  $p$ , namely:  $p = \{0.5, 0.75, 0.95\}$ . We report the results in Fig. 2 (right). We can see that as the value of  $p$  increases, the curves  $\frac{\text{nnz}(\mathbf{x}_\varepsilon)}{d \log(d)}$  appear to grow much more slowly. Furthermore, the difference between the curves becomes exponential as  $1/\varepsilon$  increases. This leads us to think of the relation between  $p$  and  $\frac{\text{nnz}(\mathbf{x}_\varepsilon)}{d \log(d)}$  as an exponential function in terms of  $1/\varepsilon$ . Also, since  $p$  and  $\frac{\text{nnz}(\mathbf{x}_\varepsilon)}{d \log(d)}$  are inversely related, we consider  $1/p$  rather than  $p$ . Therefore, we arrive at a relationship of the form:

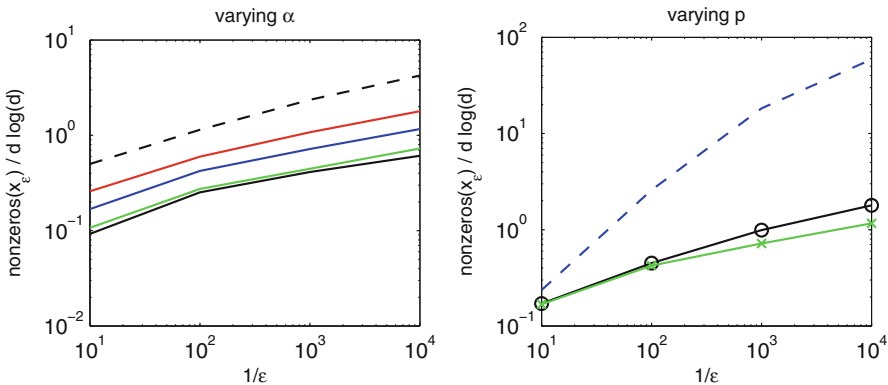
$$g(\alpha, \varepsilon, p) = \frac{c_1}{(1-\alpha)} \left(\frac{1}{\varepsilon}\right)^{c_2/p^{c_3}}$$

for some constants  $c_1, c_2, c_3$ . After experimenting with the above bound, we found that the best results were achieved at  $c_1 = 0.2, c_2 = 0.25, c_3 = 2$ .

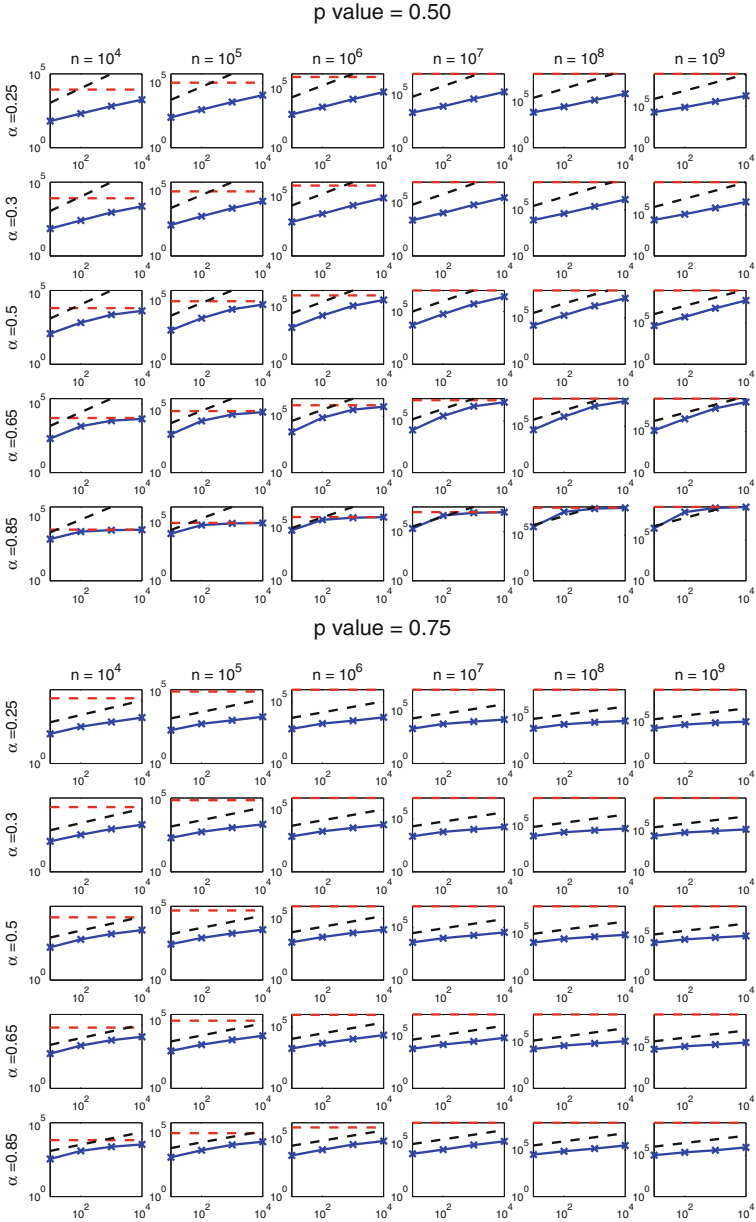
### 4.5 Results

The experimental scaling bound derived in Sect. 4.4 is now:

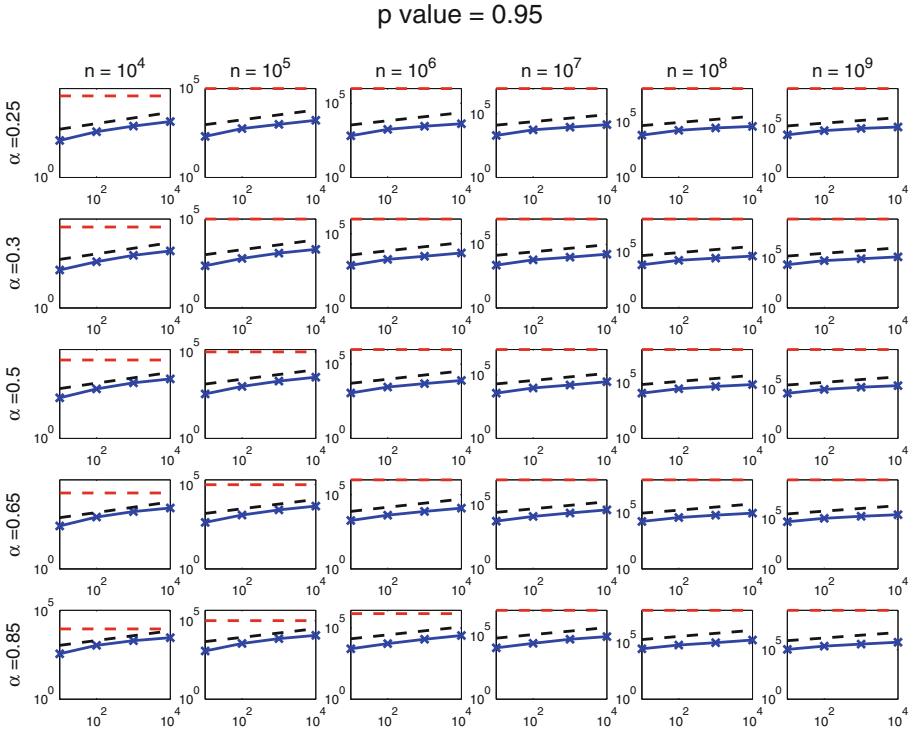
$$\text{nnz}(\mathbf{x}_\varepsilon) \leq d \log(d) \frac{0.2}{(1-\alpha)} \left(\frac{1}{\varepsilon}\right)^{1/(2p)^2}. \tag{9}$$



**Fig. 2.** Log-log plots of  $\frac{\text{nnz}(\mathbf{x}_\varepsilon)}{d \log(d)}$  versus  $1/\varepsilon$  obtained on graphs of size  $n = 10^6$  with  $d = n^{1/2}$  as  $\alpha$  and  $p$  vary. At left,  $p$  is fixed to  $p = 0.95$  and the black, green, blue, red, and dashed black curves represent  $\frac{\text{nnz}(\mathbf{x}_\varepsilon)}{d \log(d)}$  for  $\alpha = \{0.25, 0.3, 0.5, 0.65, 0.85\}$  respectively. At right,  $\alpha$  is fixed to  $\alpha = 0.5$  and the dashed blue, black, and green curves represent  $\frac{\text{nnz}(\mathbf{x}_\varepsilon)}{d \log(d)}$  for  $p = \{0.5, 0.75, 0.95\}$ , respectively (Color figure online).



**Fig. 3.** Each sub-plot has x-axis representing  $1/\epsilon$  and y-axis representing the number of non-zeros present in a diffusion vector of 1-norm accuracy  $\epsilon$ . The red dashed line represents a vector with all non-zeros present. The black dashed line shows our predicted bound (9). The blue curve shows the actual number of non-zeros found. As graphs get bigger (i.e. the fourth to sixth columns) the theoretical bound (black line) almost exactly predicts the locality of the  $\epsilon$ -approximate diffusion (Color figure online).



**Fig. 4.** Each sub-plot has x-axis representing  $1/\varepsilon$  and y-axis representing the number of non-zeros present in a diffusion vector of 1-norm accuracy  $\varepsilon$ . The red dashed line represents a vector with all non-zeros present. The black dashed line shows our predicted bound (9). The blue curve shows the actual number of non-zeros found. As graphs get bigger (i.e. the fourth to sixth columns) the theoretical bound (black line) almost exactly predicts the locality of the  $\varepsilon$ -approximate diffusion (Color figure online).

In what follows, we demonstrate the effectiveness of this bound in describing the localization of seeded PageRank vectors computed with different values of  $\alpha$ , on graphs with skewed degree sequences with varying decay exponents.

For each set of parameters (graph size  $n$ ,  $d = n^{1/2}$ , decay exponent  $p$ , and PageRank constant  $\alpha$ ), the plots in Figs. 3 and 4 display the number of nonzeros needed to approximate a PageRank vector with 1-norm accuracy  $\varepsilon$  as a function of  $1/\varepsilon$ . The blue curve represents the actual number of nonzeros required in the  $\varepsilon$ -approximation. Each plot also has a black dashed line showing the prediction by our conjectured bound (9). We note that our conjectured bound appears to properly bound the empirical scaling in all of the plots well; although, it fails to provide a true bound for some.

## 5 Discussion

We have shown that seeded PageRank vectors, though not localized on all graphs, must behave locally on graphs with degree sequences that decay sufficiently

rapidly. Our experiments show our theoretical bound to be terribly loose. In some sense this is to be expected as our algorithmic analysis is worst case. However, it isn't clear that any real-world graphs realize these worst-case scenarios. We thus plan to continue our study of simple graph models to identify empirical and theoretical localization bounds based on the parameters of the models. This will include a theoretical justification or revisit of the empirically derived bound. It will also include new studies of Chung-Lu graphs as well as the Havel-Hakimi construction itself. Finally, we also plan to explore the impact of local clustering. Our conjecture is that this should exert a powerful localization effect beyond that due to the degree sequence.

One open question sparked by our work regards the relationship between localized solutions and constant or shrinking average distance in graphs. It is well known that social networks appear to have shrinking or constant effective diameters. Existing results in the theory of localization of functions of matrices imply that a precise *bound* on diameter would force delocalization as the graph grows. Although the localization theory says nothing about average distance or small effective diameters, it hints that the solutions would delocalize. However, solutions often localize nicely in real-world networks, and we wish to understand the origins of the empirical localization behavior more fully. Another open question regards whether localization is possible on graphs with a power-law degree distribution. Our current analysis is insufficient for this case.

## References

1. Andersen, R., Chung, F., Lang, K.: Local graph partitioning using PageRank vectors. In: FOCS 2006 (2006)
2. Avrachenkov, K., Litvak, N., Sokol, M., Towsley, D.: Quick detection of nodes with large degrees. In: Bonato, A., Janssen, J. (eds.) WAW 2012. LNCS, vol. 7323, pp. 54–65. Springer, Heidelberg (2012)
3. Baeza-Yates, R., Boldi, P., Castillo, C.: Generalizing PageRank: damping functions for link-based ranking algorithms. In: SIGIR 2006, pp. 308–315 (2006)
4. Bayati, M., Kim, J., Saberi, A.: A sequential algorithm for generating random graphs. *Algorithmica* **58**(4), 860–910 (2010)
5. Benzi, M., Razouk, N.: Decay bounds and  $O(n)$  algorithms for approximating functions of sparse matrices. *ETNA* **28**, 16–39 (2007)
6. Benzi, M., Boito, P., Razouk, N.: Decay properties of spectral projectors with applications to electronic structure. *SIAM Rev.* **55**(1), 3–64 (2013)
7. Berkhin, P.: Bookmark-coloring algorithm for personalized PageRank computing. *Internet Math.* **3**(1), 41–62 (2007)
8. Bonchi, F., Esfandiari, P., Gleich, D.F., Greif, C., Lakshmanan, L.V.: Fast matrix computations for pairwise and columnwise commute times and Katz scores. *Internet Math.* **8**(1–2), 73–112 (2012)
9. Chen, P., Xie, H., Maslov, S., Redner, S.: Finding scientific gems with Google pagerank algorithm. *J. Informetrics* **1**(1), 8–15 (2007)
10. Chung, F.: The heat kernel as the PageRank of a graph. *Proc. Natl. Acad. Sci.* **104**(50), 19735–19740 (2007)
11. Faloutsos, M., Faloutsos, P., Faloutsos, C.: On power-law relationships of the internet topology. In: ACM SIGCOMM Computer Communication Review (1999)

12. Freschi, V.: Protein function prediction from interaction networks using a random walk ranking algorithm. In: BIBE, pp. 42–48 (2007)
13. Ghosh, R., Teng, S.-H., Lerman, K., Yan, X.: The interplay between dynamics and networks: centrality, communities, and cheeger inequality, pp. 1406–1415 (2014)
14. Gleich, D.F., Kloster, K.: Sublinear column-wise actions of the matrix exponential on social networks. *Internet Math.* **11**(4–5), 352–384 (2015)
15. Gori, M., Pucci, A.: ItemRank: a random-walk based scoring algorithm for recommender engines. In: IJCAI, pp. 2766–2771 (2007)
16. Huberman, B.A., Pirolli, P.L.T., Pitkow, J.E., Lukose, R.M.: Strong regularities in World Wide Web surfing. *Science* **280**(5360), 95–97 (1998)
17. Jain, A., Pantel, P.: Factrank: random walks on a web of facts. In: COLING, pp. 501–509 (2010)
18. Jeh, G., Widom, J.: Scaling personalized web search. In: WWW, pp. 271–279 (2003)
19. Kloster, K., Gleich, D.F.: Heat kernel based community detection. In: KDD, pp. 1386–1395 (2014)
20. McSherry, F.: A uniform approach to accelerated PageRank computation. In: WWW, pp. 575–582 (2005)
21. Morrison, J.L., Breitling, R., Higham, D.J., Gilbert, D.R.: Generank: using search engine technology for the analysis of microarray experiments. *BMC Bioinformatics* **6**(1), 233 (2005)
22. Nie, Z., Zhang, Y., Wen, J.R., Ma, W.Y.: Object-level ranking: bringing order to web objects. In: WWW, pp. 567–574 (2005)
23. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank citation ranking: bringing order to the web. Technical Report 1999–66, Stanford University (1999)
24. Winter, C., Kristiansen, G., Kersting, S., Roy, J., Aust, D., Knsel, T., Rmmele, P., Jahnke, B., Hentrich, V., Rckert, F., Niedergethmann, M., Weichert, W., Bahra, M., Schlitt, H.J., Settmacher, U., Friess, H., Bchler, M., Saeger, H.D., Schroeder, M., Pilarsky, C., Grtzmann, R.: Google goes cancer: improving outcome prediction for cancer patients by network-based ranking of marker genes. *PLoS Comput. Biol.* **8**(5), e1002511 (2012)