Graph Ranking and the Cost of Sybil Defense

GWENDOLYN FARACH-COLTON, Rutgers University, USA
MARTÍN FARACH-COLTON*, Rutgers University, USA
LESLIE ANN GOLDBERG, Oxford University, UK
HANNA KOMLÓS*, Rutgers University, USA
JOHN LAPINSKAS, Bristol University, UK
REUT LEVI†, Reichman University, Israel
MOTI MEDINA‡, Bar-Ilan University, Israel
MIGUEL A. MOSTEIRO§, Pace University, USA

Ranking functions such as PageRank assign numeric values (*ranks*) to nodes of graphs, most notably the web graph. Node rankings are an integral part of Internet search algorithms, since they can be used to order the results of queries. However, these ranking functions are famously subject to attacks by spammers, who modify the web graph in order to give their own pages more rank.

We characterize the interplay between rankers and spammers as a game. We define the two critical features of this game, *spam resistance* and *distortion*, based on how spammers spam and how rankers protect against spam. We observe that all the ranking functions that are well-studied in the literature, including the original formulation of PageRank, have poor spam resistance, poor distortion, or both.

Finally, we study Min-PPR, the form of PageRank used at Google itself, but which has received no (theoretical or empirical) treatment in the literature. We prove that Min-PPR has low distortion and high spam resistance. A secondary benefit is that Min-PPR comes with an explicit cost function on nodes that shows how important they are to the spammer; thus a ranker can focus their spam-detection capacity on these vulnerable nodes. Both Min-PPR and its associated cost function are straightforward to compute.

CCS Concepts: • **Theory of computation → Design and analysis of algorithms**.

ACM Reference Format:
Gwendolyn Farach-Colton, Martín Farach-Colton, Leslie Ann Goldberg, Hanna Komlós, John Lapinskas, Reut Levi, Moti Medina, and Miguel A. Mosteiro. 2023. Graph Ranking and the Cost of Sybil Defense. In *Proceedings of the 24th ACM Conference on Economics and Computation (EC '23), July 9–12, 2023, London, United Kingdom.* ACM, New York, NY, USA, 40 pages. https://doi.org/10.1145/3580507.3597782

1 INTRODUCTION

Ranking functions such as PageRank [Brin and Page, 1998] assign numeric values (*ranks*) to nodes of graphs, most notably the web graph. Node rankings are an integral part of Internet search algorithms, since they can be used to order the results of queries. However, these ranking functions are famously

*The author was supported by NSF grants CNS 2118620 and CCF 2106999.
†The author was supported by the Israel Science Foundation under Grant 1867/20.
‡The author was supported by the Israel Science Foundation under Grant 867/19.
§The author was supported by Pace University SRC Award and Kenan Fund.

EC '23, July 9–12, 2023, London, United Kingdom

ACM ISBN 979-8-4007-0104-7/23/07. . . $15.00
https://doi.org/10.1145/3580507.3597782

subject to attacks by spammers, who modify the web graph in order to give their own pages more rank.

In the literature on ranking functions, resistance to spam attacks is treated heuristically [Fogaras et al., 2005, Gyöngyi et al., 2004, Liu et al., 2016, Ng et al., 2001b] or with respect to specific ranking functions versus specific attacks [Cheng and Friedman, 2006, Hopcroft and Sheldon, 2008, Ng et al., 2001a]. Much of the ranking literature has focused on two aspects of fighting spam: link-spam detection [Alvisi et al., 2014, Andersen et al., 2008, Cheng and Friedman, 2006, Gyongyi et al., 2006, Yu, 2011, Yu et al., 2010, 2008]; and heuristically designing ranking functions that are spam resistant [Bhattacharjee and Goel, 2006, Gyöngyi et al., 2004, Hopcroft and Sheldon, 2008, Krishnan and Raj, 2006, Kumar et al., 2006, Liu et al., 2016].

We combine the notion of attack detection and attack resistance into a single two-party game that allows us to reason about ranking functions.

- A spammer can perform a *Sybil attack*[1]:
  - Create as many nodes in the web graph as they like, for free;
  - Change the out-links of all nodes they own, for free;
  - Acquire (by buying, hacking, etc) existing nodes in the web graph, at some cost, except for nodes in a non-empty *trusted set*, which cannot be acquired.
- The ranker can:
  - Modify the ranking function to make it more resistant to spam, subject to quality constraints;
  - Expend effort to detect if nodes have been acquired by spammers, which we model as raising the cost for spammers to acquire those nodes.

In this game, the spammer wants to maximize the sum of rank of its nodes, and naturally the ranker wants to minimize this quantity. The spammer knows the ranking function and the cost function, whereas the ranker does not know what actions the spammer is taking.

Trusted nodes play a critical role. Without trusted nodes, the ranker cannot hope to resist a spam attack, because the spamming model is general enough that the spammer could make a complete (or partial) replica of the web graph, and the ranker would not know which is the correct version. On the other hand, the ranker cannot simply rank trusted nodes and ignore untrusted nodes because: it is prohibitively expensive to certify that all unspammed nodes are trusted; and, restricting rank to only a few nodes would introduce local (ranking-function) *distortions*, formally defined in Section 4.

This rather broad characterization of spam attacks and defense leaves two obvious questions:

- Is there a way to define the spam resistance and distortion of arbitrary ranking functions that is mathematically tractable and matches intuition and practice?
- Do these definitions yield mathematical and practical insight? That is, can we prove that there is a ranking function with both low distortion and high spam resistance? In other words, does the ranker have a good strategy?

Our contribution. In this paper, we answer both questions in the affirmative. In Section 3, we refine the object of the spamming game to be the highest cost-benefit ratio the ranker can force on the spammer, which we call the *spam resistance* of a ranking function. In Section 4, we add a further constraint to the game, which is that the ranker must try to minimize *distortion*. This constraint is meant to avoid trivial ranking functions, such as those that place all rank on a few trusted nodes. These two sections lean heavily on the practical literature on spam resistance in order to come up with a clean and mathematically interesting definition of the spam game that is mathematically tractable.

The spammer would seem to have too much power in this game, because it has full knowledge and need not concern itself with the quality of the final ranking. Indeed, we find that there is an

[1]The moves by the spammer are quite general and subsume specific attacks treated in the literature [Viswanath et al., 2011]

intuitive tradeoff between spam resistance and distortion in that we observe, rather dispiritingly, that all ranking functions in the literature exhibit either high distortion, low spam resistance, or both.

Interestingly, this leaves open the question of the properties of the type of PageRank that Google itself used. To the best of our knowledge, Google has never used uniform reset PageRank for web-search ranking, in part because of its obvious susceptibility to spammers. Instead, Google has used a variant of PageRank that we call Min-PPR (and which we define in Section 5), because it was intuitively considered to be more spam resistant [Farach-Colton, [n. d.]]. We know of no treatment of Min-PPR in the literature.

In Section 5, we set out our main contribution: we show that this faith in Min-PPR is well placed. We prove that Min-PPR exhibits both low distortion and high spam resistance subject to the mild technical condition of fast mixing. Fast mixing is known to hold in real-world social and semantic networks, and if it does not hold then we prove that no variant of PageRank has low distortion.

No ranking function is spam resistant if the ranker makes no effort to detect spam, so we also show that we can compute a cost function for Min-PPR that establishes the importance of each node in fighting spam. In other words, this cost function shows which nodes the ranker should focus its spam-fighting efforts on.

Proving these results requires a substantial technical contribution: while our definitions of distortion and spam resistance are natural, they are also difficult to analyze. The proof that establishes the distortion of Min-PPR relies on a sophisticated analysis of mixing times, whereas the proof that establishes the spam resistance of Min-PPR relies on an intricate extremal argument. These analyses are left to the Appendices A and B respectively due to space restrictions.

In summary, in this paper we reimagine the arms race between web rankers and web spammers as a two-person game. We define the moves of each party. We show that there is a natural tension between spam resistance and distortion, but we show that the ranker can gain the upper hand; that is, we exhibit a ranking function that has low distortion and high spam resistance.

Roadmap. In Section 2, we introduce preliminary notation. In Section 3, we formally define spam resistance and in Section 4, we formally define distortion. In both sections, we survey existing ranking functions and provide preliminary results. In Section 5, we formally define Min-PPR and give an overview of the main theorems related to its spam resistance and distortion. In Section 6, we explore the algebraic properties of PageRank, which may be of independent interest, and which are used in our main proofs. In Section 7, we discuss related work. In the Appendices A and B, we prove the theorems outlined in Section 5, and in Appendix C, we present experimental results that validate our theoretical analysis.

2 PRELIMINARIES

Let $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ be an n-node graph. Since we will frequently discuss random walks on graphs, we always assume that any node with no outgoing edges to other nodes has a self-loop. For any edge $(u, v) \in \mathbf{E}$, v is an *out-neighbor* of u, and u is an *in-neighbor* of v. For any $v \in \mathbf{V}$, let $N_{\text{in}}(v)$ and $N_{\text{out}}(v)$ denote the set of in-neighbors and out-neighbors of v, respectively. Note that $|N_{\text{out}}(v)| > 0$ for all v, because of the self loops. Let $d_{\text{in}}(v) = |N_{\text{in}}(v)|$, and $d_{\text{out}}(v) = |N_{\text{out}}(v)|$.

We extend the min operator on vectors to component-wise min, so that, for vectors $\mathbf{x_1}, \dots, \mathbf{x_k}$,

$$\min\{\mathbf{x_1}, \dots, \mathbf{x_k}\}[i] = \min\{\mathbf{x_1}[i], \dots, \mathbf{x_k}[i]\}.$$

For any real vector $\mathbf{x}$, let $||\mathbf{x}|| := \sum_i |\mathbf{x}_i|$ be its ℓ_1-norm, and if $\mathbf{x} \neq \mathbf{0}$ then we define $\llbracket \mathbf{x} \rrbracket := \mathbf{x}/||\mathbf{x}||$ to be the standard ℓ_1-normalization. Let the *support* of $\mathbf{x}$ be the set of coordinates i with $\mathbf{x}[i] \neq 0$.

We define a *ranking vector* for a graph $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ to be any function $\hat{\mathbf{x}}\colon \mathbf{V} \to [0, 1]$ such that $\sum_{v \in \mathbf{V}} \hat{\mathbf{x}}[v] = 1$. (We normalise our rankings to facilitate easy comparison.) Note that any ranking vector also defines a probability distribution on $\mathbf{V}$. We also adopt the convention that for all $\mathbf{A} \subseteq \mathbf{V}$, $\hat{\mathbf{x}}[A] = \sum_{v \in \mathbf{A}} \hat{\mathbf{x}}[v]$.

Let $\mathcal{G}$ be the set of all directed graphs on which the uniform random walk is ergodic. For all $\mathbf{G} \in \mathcal{G}$, we write $\mathcal{R}(\mathbf{G})$ for the stationary distribution of the uniform random walk; we call this the *reference rank* of $\mathbf{G}$.

We define PageRank on an arbitrary graph $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ as follows. Let $\varepsilon \in (0,1)$, and let $\hat{\mathbf{r}}$ be a probability distribution on $\mathbf{V}$. We write $\hat{\mathbf{r}}[v]$ for the probability of choosing $v \in \mathbf{V}$ under $\hat{\mathbf{r}}$. The PageRank random walk's state space is $\mathbf{V}$. From any node $v \in \mathbf{V}$, with probability $1-\varepsilon$ it traverses a uniformly random out-edge from v, and with probability ε it moves to a node drawn from $\hat{\mathbf{r}}$; this event is called a *reset*[2]. We denote this random walk model by $(\mathbf{G}, \hat{\mathbf{r}}, \varepsilon)$. The stationary distribution of $(\mathbf{G}, \hat{\mathbf{r}}, \varepsilon)$ is the *PageRank* of $\mathbf{G}$ with *reset probability* ε and *reset vector* $\hat{\mathbf{r}}$. We denote the value of this PageRank at a node $v \in \mathbf{V}$ by $\mathcal{R}(\mathbf{G}, \hat{\mathbf{r}}, \varepsilon)[v]$; note that $\mathcal{R}(\mathbf{G}, \hat{\mathbf{r}}, \varepsilon)$ is a ranking vector. (Since $(\mathbf{G}, \hat{\mathbf{r}}, \varepsilon)$ is ergodic for all choices of parameters, $\mathcal{R}(\mathbf{G}, \hat{\mathbf{r}}, \varepsilon)[v]$ is uniquely defined.)

The original version of PageRank defined by Brin and Page [Page et al., 1999] takes $\hat{\mathbf{r}}$ to be the uniform distribution, so that $\hat{\mathbf{r}}[v] = 1/|\mathbf{V}|$ for all $v \in \mathbf{V}$. We denote this vector by $\hat{\mathbf{u}}$, and call this version of PageRank the *uniform PageRank (UPR)*. *Personalized PageRank (PPR)* instead takes $\hat{\mathbf{r}}$ to be of the form $\hat{\mathbf{r}}[c] = 1$ for some $c \in \mathbf{V}$, which is called the *center node*. Thus each time the walk $(\mathbf{G}, \hat{\mathbf{r}}, \varepsilon)$ resets, it returns deterministically to c. We denote the resulting ranking vector by $\mathcal{R}(\mathbf{G}, c, \varepsilon)$.

We define a *ranking algorithm* to be any algorithm that takes as input an arbitrary graph $\mathbf{H}$ and an arbitrary non-empty *trusted set* $\mathbf{T} \subseteq \mathbf{V_H}$ and outputs a ranking vector. For example, T-PPR$_\varepsilon$ is the version of PPR in which the center vertex is trusted.

We will see in Observation 2 that any ranking algorithm that does not make use of $\mathbf{T}$ has zero spam resistance.

For all graphs $\mathbf{G} \in \mathcal{G}$ and all probability distributions $\hat{\mathbf{p}}$ and $\hat{\mathbf{q}}$ over $\mathbf{V_G}$, we define the *total variation distance* between $\hat{\mathbf{p}}$ and $\hat{\mathbf{q}}$ by

$$d_{\mathrm{TV}}(\hat{\mathbf{p}}, \hat{\mathbf{q}}) := \frac{1}{2} \sum_{v \in \mathbf{V_G}} \left| \hat{\mathbf{p}}[v] - \hat{\mathbf{q}}[v] \right|.$$

In the following definition, let $\hat{\mathbf{p}}_{t,v}$ be the distribution of the uniform random walk on $\mathbf{G}$ from initial state $v \in \mathbf{V_G}$ at time $t \geq 0$. Then for all $\rho > 0$, the (worst-case) *mixing time* of $\mathbf{G}$ to within error ρ is given by

$$\tau_{\mathbf{G}}(\rho) := \min \left\{ t \geq 0 \colon \text{for all } v \in \mathbf{V_G},\ d_{\mathrm{TV}}\big(\hat{\mathbf{p}}_{t,v},\ \mathcal{R}(\mathbf{G})\big) \leq \rho \right\}.$$

Following standard practice, we take ρ to be $1/4$ when we don't specify it.

3 DEFINING SPAM RESISTANCE

We first give some examples of ranking algorithms that show differing levels of spam resistance. We then define spam resistance formally, we note that it matches our intuition on the examples provided, and we prove some preliminary results.

UPR would appear to be trivial to spam. Consider, for example, the effect of making a large set of new nodes with self-loops. The more new nodes the spammer makes (for free), the larger the probability that the PageRank random walk will reset to those nodes, and the more rank the spammer will capture. In fact, this sort of attack is known to be viable in practice [Baeza-Yates et al., 2005, Gyöngyi and Garcia-Molina, 2005].

In 1999, Kleinberg [Kleinberg, 1999] introduced the notion of *hub* and *authority scores* and the HITS algorithm to compute them. Intuitively, a page receives a high authority score if it is pointed

[2] Also known in the literature as *teleportation.*

to by many high-quality hubs. A page receives a high hub score if it points to many high-scoring authorities. Hub scores depend on out-links and are therefore free to spam by creating new nodes and pointing them to nodes with high authority scores. But once a spammer owns many pages with high hub scores, it can create pages with high authority scores, once again for free. Such considerations are far from hypothetical. This vulnerability was already well understood in 2004 [Gyongyi and Garcia-Molina, 2004], and Assano et. al [Asano et al., 2007] report that HITS was unusable by 2007, due to its spammability. Like UPR, we should expect HITS to have no spam resistance, as the spammer can acquire high rank at no cost.

Next, consider *SuperTrust*, which assigns non-zero rank only to trusted nodes, which are known not to belong to spammers. SuperTrust is unspammable, because the spammer receives no rank no matter what they do. (Of course, because in general very few nodes can be fully trusted, in most cases SuperTrust has very high distortion and is not suitable as a ranking algorithm.)

Finally, consider T-PPR$_\varepsilon$, the version of PPR in which the center vertex is trusted and cannot be subverted by a spammer. It is not difficult for the spammer to arrange things so that whenever the PageRank random walk enters a vertex they own, it remains in spammer-owned territory until the next reset; however, the random walk will only enter spammer-owned vertices v at a rate commensurate with v's true PPR. Thus in order to acquire significant rank, the spammer has to invest time and effort into acquiring nodes which already have significant rank, and we should expect T-PPR$_\varepsilon$ to be quite strongly spam resistant.

We now give a definition of spam resistance that coincides with our intuition that HITS and UPR are not at all spam resistant, T-PPR$_\varepsilon$ is quite strongly spam resistant, and SuperTrust has unbounded spam resistance.

A formal definition of spam resistance. In the examples given above, there was no mention of how much it costs to acquire a node. Here, we make explicit the cost model, the changes a spammer can make to a graph, and the cost/benefit ratio a spammer can achieve. It is this cost/benefit ratio that defines spam resistance.

Let $\mathbf{G} = (\mathbf{V_G}, \mathbf{E_G})$, and let $\mathbf{T_G} \subseteq \mathbf{V_G}$ be the set of trusted sites. Then $\mathbf{V_G} \setminus \mathbf{T_G}$ is the set of all sites that might be acquired by the spammer. For all $\mathbf{P} \subseteq \mathbf{V_G} \setminus \mathbf{T_G}$, let $\mathbf{G_P}$ be the set of all graphs obtainable from $\mathbf{G}$ by:

- Adding an arbitrary (possibly empty) set $\mathbf{S}$ of new vertices;
- Changing all the out-edges of vertices in $\mathbf{S} \cup \mathbf{P}$ in an arbitrary fashion.

Thus $\mathbf{G_P}$ is the set of all graphs that the spammer can obtain after acquiring the vertices in $\mathbf{P}$. For any $\mathbf{H} = (\mathbf{V_H}, \mathbf{E_H}) \in \mathbf{G_P}$, let $\mathbf{S_H}$, the set of *spam nodes of* $\mathbf{H}$, be $\mathbf{V_H} \setminus \mathbf{V_G}$.

We are now ready to define the cost and benefit of spamming. Like ranks, we will normalize costs in order to make them comparable.

Let f be a ranking algorithm, so that $f(\mathbf{H}, \mathbf{T})$ returns a rank vector for all graphs $\mathbf{H}$ and all non-empty $\mathbf{T} \subseteq \mathbf{V_H}$. We think of $\mathbf{H}$ as a post-spam graph, belonging to some $\mathbf{G_P}$; thus the ranking algorithm does not know which nodes are owned by the spammer but does know which nodes are trusted. We extend $f(\mathbf{H}, \mathbf{T})$ from vertices to sets in the standard way: For all $\mathbf{X} \subseteq \mathbf{V_H}$, $f(\mathbf{H}, \mathbf{T})[\mathbf{X}] = \sum_{v \in \mathbf{X}} f(\mathbf{H}, \mathbf{T})[v]$.

A *cost function* is a normalized function on $\mathbf{V_G} \setminus \mathbf{T_G}$, i.e. $C(v) \geq 0$ for all $v \in \mathbf{V_G} \setminus \mathbf{T_G}$ and $\sum_{v \in \mathbf{V_G} \setminus \mathbf{T_G}} C(v) = 1$. Let $C_{(\mathbf{G}, \mathbf{T_G})}$ denote the set of all such cost functions.[3] We extend this definition to subsets of nodes as above.

[3]Note that we restrict the spammer to acquiring whole nodes, rather than individual edges. This is because $\mathbf{S}$ could be arbitrarily large, so the set of possible edges between $\mathbf{V_G} \setminus \mathbf{T_G}$ and $\mathbf{S}$ is also arbitrarily large, and under an edge-acquisition model we would be unable to normalize costs. For every ranking algorithm considered in this paper, this technical restriction makes no difference to spammability.

We are ready to define spam resistance, as follows:

DEFINITION 1. *For all classes $\mathcal{A}$ of graphs:*

*(1) For all $\sigma > 0$, a ranking algorithm f is σ-*spam resistant *on $\mathcal{A}$ if, for all $\mathbf{G} = (\mathbf{V_G}, \mathbf{E_G}) \in \mathcal{A}$ and all non-empty $\mathbf{T_G} \subseteq \mathbf{V_G}$, there exists a cost function $C \in C_{(\mathbf{G},\mathbf{T_G})}$ such that, for all $\mathbf{P} \subseteq \mathbf{V_G} \setminus \mathbf{T_G}$ and $\mathbf{H} \in \mathbf{G_P}$ with $f(\mathbf{H}, \mathbf{T_G})[\mathbf{S_H} \cup \mathbf{P}] > 0$,*

$$\frac{C(\mathbf{P})}{f(\mathbf{H}, \mathbf{T_G})[\mathbf{S_H} \cup \mathbf{P}]} \geq \sigma.$$

(2) A ranking algorithm f has unbounded spam resistance *on $\mathcal{A}$ if, for all $\mathbf{G} = (\mathbf{V_G}, \mathbf{E_G}) \in \mathcal{A}$, all non-empty $\mathbf{T_G} \subseteq \mathbf{V_G}$, all $\mathbf{P} \subseteq \mathbf{V_G} \setminus \mathbf{T_G}$, and all $\mathbf{H} \in \mathbf{G_P}$,*

$$f(\mathbf{H}, \mathbf{T_G})[\mathbf{S_H} \cup \mathbf{P}] = 0.$$

(3) A ranking algorithm f has zero spam resistance *on $\mathcal{A}$ if, for all $\mathbf{G} = (\mathbf{V_G}, \mathbf{E_G}) \in \mathcal{A}$, there exists a non-empty $\mathbf{T_G} \subseteq \mathbf{V_G}$ and $\mathbf{H} \in \mathbf{G}_\emptyset$ such that $f(\mathbf{H}, \mathbf{T_G})[\mathbf{S_H}] > 0$.*

Intuition of the cost function. The cost function captures how difficult it is for a spammer to subvert a node in a way that is hidden from the spam-detection efforts of the ranker. Thus, for example, a search engine would be able to substantially increase the cost of a specific node by assigning a human to watch it carefully and to de-index it if they suspected it had been acquired by a spammer, and in fact search engines do manipulate their interaction with spammers by such methods [Sullivan, [n. d.]]. The correct way to view the cost function is as a guarantee that the ranker has a strategy to force a spammer to pay for their rank; using a spam-resistant ranking function by itself is not enough. Thus, when we establish that a ranking function is spam resistant, we must exhibit a cost function that the ranker can efficiently approximate, then use in conjunction with that ranking function.

Indeed, while the definition of spam resistance only requires that some good cost function exist, all our proofs of spam resistance will construct explicit cost functions to direct the ranker's spam-detection effort. Observe that no amount of such effort is sufficient if the ranking function has zero spam resistance and that no such effort is needed if the ranking function has unbounded spam resistance. An examination of SuperTrust and UPR supports this interpretation.

Preliminary results. As a warm-up, to see that trusted sites are necessary, consider any ranking algorithm f whose output does not depend on $\mathbf{T}$. Let $\mathbf{G}$ be an arbitrary graph, and let $\mathbf{H}$ consist of two disjoint copies of $\mathbf{G}$ spanning vertex sets $\mathbf{V}_1$ and $\mathbf{V}_2$. Then given $\mathbf{H}$ as input, without loss of generality f assigns rank at least $1/2$ to $\mathbf{V}_1$. Viewing $\mathbf{H}$ as a spam graph in $\mathbf{H}[\mathbf{V}_2]_\emptyset$, where the base graph $\mathbf{H}[\mathbf{V}_2]$ is the copy of $\mathbf{G}$ spanning $\mathbf{V}_2$, $\mathbf{P} = \emptyset$, and $\mathbf{S_H} = \mathbf{V}_1$, we see that the spammer has attained rank at least $1/2$ without buying any vertices. Thus we have proved the following.

OBSERVATION 2. *Any ranking algorithm that is invariant under membership in $\mathbf{T_G}$ has zero spam resistance on all graph classes.*

Observation 2 implies that UPR and HITS have zero spam resistance. On the positive side, any ranking algorithm that assigns non-zero rank only to vertices in $\mathbf{T_G}$ (that is, any variant of SuperTrust) has unbounded spam resistance; in fact, these are the only ranking algorithms with unbounded spam resistance. Moreover, the trusted variant T-PPR$_\varepsilon$ of PPR also has high spam resistance, as follows.

LEMMA 3. *For all $\varepsilon \in (0, 1)$, T-PPR$_\varepsilon$ is ε-spam resistant on all graph classes. The cost function that establishes this spam resistance is the T-PPR$_\varepsilon$ itself, normalized over untrusted nodes.*

We prove this in Section B, but for now, suppose the spammer acquires a node v with T-PPR$_\varepsilon(v) = x$ and redirects its outward edges into the set $\mathbf{S} \cup \mathbf{P}$ of vertices they own. Then each time the random walk associated with T-PPR$_\varepsilon$ hits v, which happens at rate roughly x, at worst it will stay in $\mathbf{S} \cup \mathbf{P}$

until the walk next resets, which takes $1/\varepsilon$ time on average. So the spammer should acquire rank at most x/ε. As we might expect, Lemma 3 is essentially tight.

OBSERVATION 4. *For all $\varepsilon, \eta \in (0,1)$, T-PPR$_\varepsilon$ is not $\varepsilon \cdot (1+\eta)/(1-\varepsilon)$-spam resistant on the class of all cliques.*

4 DEFINING DISTORTION

In this section, we define the *distortion* of a ranking vector. We first discuss some guiding principles, then give a formal definition, then conclude with a discussion of the distortion of PPR and UPR.

Guiding principles. First, we would like to give an accurate ranking for every site, so our metric should be mostly concerned with the maximum error at any site rather than the total error across all sites. Note in particular that total variation distance from $\mathcal{R}(\mathbf{G})$ does not capture this well — a total variation distance of 0.1 from $\mathcal{R}(\mathbf{G})$ could indicate anything from an additive error of $0.1/|\mathbf{V}_\mathbf{G}|$ at every vertex (a relatively minor error) to an additive error of 0.1 at a single vertex (a very severe error).

Second, we should be concerned with multiplicative error rather than additive error. To see this, suppose $\mathbf{G} = (\mathbf{V}, \mathbf{E}) \in \mathcal{G}$ has n vertices, let $v_1, v_2 \in \mathbf{V}$, and suppose that v_1 has reference rank $1/\sqrt{n}$ and v_2 has reference rank $1/4$. Then intuitively, assigning v_1 a rank of $1/\log n$ is a far more severe mistake than assigning v_2 a rank of $1/4 - 2/\log n$, even though the additive error is smaller. Likewise, assigning v_2 a rank of $1/\sqrt{n}$ would be a far more severe mistake than assigning it a rank of $1/8$, even though the additive errors are comparable.

Third, multiplicative error is only significant when it causes us to make important mistakes in the final site ranking. To illustrate what we mean by this, suppose our ranking function were to assign rank $1/2^{n/2}$ to a node v whose reference rank is $1/2^n$. This would constitute a huge multiplicative error of $2^{n/2}$. However, in practice, $1/2^n$ and $1/2^{n/2}$ are both so small as to be indistinguishable, so this error is unlikely to have much of an impact on query rankings. In general, we can safely ignore multiplicative error on vertices with "insignificant" reference rank as long as we still assign them "insignificant" rank.

Formal definition. Let $\delta > 0$. For all n-vertex graphs $\mathbf{G} = (\mathbf{V}, \mathbf{E}) \in \mathcal{G}$, all ranking vectors $\hat{\mathbf{x}}\colon \mathbf{V} \to [0,1]$, and all vertices $v \in \mathbf{V}$, define the *stretch* and *contraction of* $\hat{\mathbf{x}}$ *on* v by

$$\mathrm{Stretch}_\delta(\hat{\mathbf{x}}, \mathbf{G}, v) = \frac{\max\{\hat{\mathbf{x}}[v], 1/n^\delta\}}{\max\{\mathcal{R}(\mathbf{G})[v],\ 1/n^\delta\}} \qquad \mathrm{Cont}_\delta(\hat{\mathbf{x}}, \mathbf{G}, v) = \frac{\max\{\mathcal{R}(\mathbf{G})[v], 1/n^\delta\}}{\max\{\hat{\mathbf{x}}[v],\ 1/n^\delta\}}.$$

We then define[4] the *distortion of* $\hat{\mathbf{x}}$ *on* v by

$$\mathbf{D}_\delta(\hat{\mathbf{x}}, \mathbf{G}, v) = \max\{\mathrm{Stretch}_\delta(\hat{\mathbf{x}}, \mathbf{G}, v),\ \mathrm{Cont}_\delta(\hat{\mathbf{x}}, \mathbf{G}, v)\},$$

and the *distortion of* $\hat{\mathbf{x}}$ *on* $\mathbf{G}$ by

$$\mathbf{D}_\delta(\hat{\mathbf{x}}, \mathbf{G}) = \max\{\mathbf{D}_\delta(\hat{\mathbf{x}}, \mathbf{G}, v) \colon v \in \mathbf{V}\}.$$

We pause for a moment to map these definitions back onto our intuition. We take our reference rank threshold for a vertex to be "significant" to be $1/n^\delta$. Observe that if a vertex v is significant, and $\hat{\mathbf{x}}$ ranks it as significant, then the distortion of $\hat{\mathbf{x}}$ on v is simply the approximation ratio between its $\hat{\mathbf{x}}$-rank and its reference rank. The maxima in the definitions of Stretch and Cont capture the idea that we should disregard any multiplicative error below the significance threshold. For example, if a

[4]The names of our accuracy metrics are taken from the theory of metric space embedding. In this theory, distortion is stretch times contraction, but since we consider normalized ranking functions, we instead take the distortion to be the maximum of the stretch and contraction.

vertex is insignificant, and $\hat{\mathbf{x}}$ ranks it as insignificant, then the distortion of $\hat{\mathbf{x}}$ on v is 1, i.e. as low as possible.

It is natural to ask which vertices should we consider insignificant. That is, how should we choose δ? It will turn out that for our purposes, it doesn't actually matter — our main distortion bounds hold for any choice of δ, so we leave this question to future work. Note, however, that since the total reference rank is 1, any choice of $\delta < 1$ will leave at most an $n^{\delta-1} = o(1)$ proportion of nodes above the significance threshold; for this reason we shall take $\delta \geq 1$.

Restricting to fast-mixing graphs. Before we discuss specific examples, we note the following intuitive requirement for the output of any PageRank to have low distortion on its input graph $\mathbf{G} \in \mathcal{G}$: After a reset, on average, the distribution of its random walk should have time to converge to $\mathcal{R}(\mathbf{G})$ before the next reset. Thus for a reset vector $\hat{\mathbf{r}}$ and a reset probability ε to yield an effective PageRank on $\mathbf{G}$, the mixing time of the uniform random walk on $\mathbf{G}$ whose initial state is drawn from $\hat{\mathbf{r}}$ should be less than $1/\varepsilon$. As an example of what might otherwise go wrong, consider the case where $\mathbf{G}$ is an n-vertex directed cycle — where unless $\hat{\mathbf{r}}$ is close to uniform, or ε is very small, the resulting PageRank will be biased away from segments with low mass in $\hat{\mathbf{r}}$.

Fortunately, real networks are very often fast mixing (at least on the giant component). Experimental studies [Albert et al., 1999, Broder et al., 2000, Faloutsos et al., 1999] have demonstrated that the degrees of the web graph are power-law distributed; Gkantsidis, Mihail and Saberi [Gkantsidis et al., 2003] prove that n-vertex random power-law graphs have $O(\log n)$ mixing time [Gkantsidis et al., 2003] with high probability. While power-law random graphs are of foundational importance, there are many other models for random "web-like" graphs, including graphs such as the Facebook graph, which may not admit a power-law degree distribution [Gjoka et al., 2010, Ugander et al., 2011]. Among the most well-known such models are preferential attachment [Barabási and Albert, 1999], which exhibits $O(\log n)$ mixing time with high probability [Cooper and Frieze, 2007, Mihail et al., 2006], and the Newman–Watts small world model [Newman and Watts, 1999], which exhibits $O(\log^2 n)$ mixing time with high probability [Addario-Berry and Lei, 2012]. Fast mixing is also a common assumption in the literature on defenses against Sybil attacks [Mohaisen et al., 2010, Viswanath et al., 2011]. Some important models, such as random hyperbolic graphs [Krioukov et al., 2010], do not exhibit worst-case fast mixing [Kiwi and Mitsche, 2018], though it is not known if they have fast average-case mixing time.

Mohaisen, Yun and Kim [Mohaisen et al., 2010] offer an explanation for why some models exhibit fast worst-case mixing and others do not, by characterizing two kinds of social networks: those in which nodes are linked based on real acquaintance, such as DBLP, and those that do not have this requirement, such as Facebook. They argue based on experimental evidence that DBLP-like networks are slowly mixing compared to the Facebook-like networks in the worst case, but are nevertheless still fast-mixing in the average case.

As we now show, PPR and UPR do not in general output low-distortion ranking vectors even on very fast-mixing graphs, but a mixing time a little lower than $1/\varepsilon$ will suffice for Min-PPR (see Section 5). For any function $T\colon \mathbb{N} \to [0, \infty)$, let

$$\mathcal{G}_T := \big\{\mathbf{G} \in \mathcal{G}\colon \tau_{\mathbf{G}}(1/4) \leq T(|\mathbf{V}_{\mathbf{G}}|)\big\}.$$

PPR usually has high distortion. Consider PPR with center $c \in \mathbf{V}$ and reset probability ε. Since the PPR random walk resets to c with probability ε at each step, we have $\mathcal{R}(\mathbf{G}, c, \varepsilon)[c] \geq \varepsilon$, so

$$\mathrm{D}_\delta(\mathcal{R}(\mathbf{G}, c, \varepsilon), \mathbf{G}) \geq \mathrm{Stretch}_\delta(\mathcal{R}(\mathbf{G}, c, \varepsilon), \mathbf{G}, c) \geq \varepsilon / \max\{\mathcal{R}(\mathbf{G})[c], 1/|\mathbf{V}_{\mathbf{G}}|^\delta\}.$$

Thus PPR has high distortion unless either ε is unrealistically small or c happens to be a vertex with extremely high reference rank. Indeed, since the total reference rank of $\mathbf{G}$ is 1, we obtain the following.

OBSERVATION 5. *Let $\delta \geq 1$, $\varepsilon \in (0,1)$, and $t \leq \varepsilon n^{\delta}$. Then for any n-vertex graph $\mathbf{G} \in \mathcal{G}$, there are at most t/ε vertices $c \in \mathbf{V}_{\mathbf{G}}$ such that $\mathbf{D}_{\delta}(\mathcal{R}(\mathbf{G}, c, \varepsilon), \mathbf{G}) \leq t$.*

Moreover, a clique is a simple example of an n-vertex graph on which the output of PPR has distortion at least εn for any choice of center. (Note that all cliques are contained in $\mathcal{G}_1$.) While there do exist specific graphs and center choices for which the output of PPR has low distortion, we have no reason to believe that these specific inputs are relevant for real-world use.

UPR can have high distortion. For any $\mathbf{G} \in \mathcal{G}$, $v \in \mathbf{V}$, and $(u, v) \in \mathbf{E}$, replacing (u, v) with a suitably large collection of internally vertex-disjoint two-edge paths from u to v inflates $\mathcal{R}(\mathbf{G}, \hat{\mathbf{u}}, \varepsilon)[v]$ to a near-arbitrary extent while leaving $\mathcal{R}(\mathbf{G})[v]$ almost unchanged. This construction does not significantly affect the mixing time of $\mathbf{G}$, so we conclude the following.

OBSERVATION 6. *Let $\delta \geq 1$ and $\varepsilon \in (0,1)$. Then there exist infinitely many graphs $\mathbf{G} = (\mathbf{V}, \mathbf{E}) \in \mathcal{G}_4$ with $\mathbf{D}_{\delta}(\mathcal{R}(\mathbf{G}, \hat{\mathbf{u}}, \varepsilon), \mathbf{G}) \geq \frac{1}{2}\varepsilon(1-\varepsilon)|\mathbf{V}_{\mathbf{G}}|^{\delta}$.*

5 MIN-PPR HAS LOW DISTORTION AND HIGH SPAM RESISTANCE

In this section, we introduce Min-PPR, which is the idea that ranks should be determined by running PPR multiple times, with different trusted centers. Recall that this is the actual version of PageRank used in Google. The rank of a site w should be the minimum of the ranks that result (appropriately normalised). We will show how to turn the Min-PPR idea into a family of ranking algorithms, T-Min-PPR$_{k,\varepsilon}$, as Min-PPR on k trusted centers with reset probability ε, though we will need some constraints on how the centers are selected. We then outline the theorems needed to establish that it has low distortion and high spam resistance. These theorems are proven in Appendices A and B.

Specifically, we will show that, given a small enough ε and a large enough k, T-Min-PPR$_{k,\varepsilon}$ is $(\varepsilon/3k)$-spam resistant. It turns out (see the discussion following the statement of Theorem 11) that T-PPR$_{\varepsilon}$ is ε-spam resistant but not (2ε)-spam resistant. Thus, we see that T-Min-PPR$_{k,\varepsilon}$ resists spam almost as well as T-PPR$_{\varepsilon}$. Moreover, we prove that there are many possible choices of $\mathbf{T}_{\mathbf{G}}$ such that T-Min-PPR$_{k,\varepsilon}(\mathbf{G}, \mathbf{T}_{\mathbf{G}})$ has $1 + o(1)$ distortion, so that T-Min-PPR$_{k,\varepsilon}$ is accurate on the pre-spam graph. Finally, we will show that T-Min-PPR$_{k,\varepsilon}$ is a PageRank, by the closure properties of PageRank that we establish below. Hence, it fits neatly into the existing methods and heuristics of the field.

Defining T-Min-PPR. In order to compute $\mathcal{R}_{\min}(\mathbf{G}, \mathbf{K}, \varepsilon)$, we choose an arbitrary subset $\mathbf{K}$ of $\mathbf{T}$ of size $\min\{k, |\mathbf{T}|\}$ which does not depend on $\mathbf{G}$ and output

$$\mathcal{R}_{\min}(\mathbf{G}, \mathbf{K}, \varepsilon)[w] = \llbracket \min\{\mathcal{R}(\mathbf{G}, c, \varepsilon)[w] : c \in \mathbf{K}\} \rrbracket.$$

There is a technical difficulty here: if $\mathbf{G}$ has been disconnected by the spammer and the centers $\mathbf{K}$ are badly chosen, it might be that for all vertices w we have $\min\{\mathcal{R}(\mathbf{G}, c, \varepsilon)[w] : c \in \mathbf{K}\} = 0$, in which case the normalisation above is invalid and $\mathcal{R}_{\min}(\mathbf{G}, \mathbf{K}, \varepsilon)$ does not output a ranking vector. To deal with this, we introduce the following technical restriction.

DEFINITION 7. *Let $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ be a (directed) graph and let $\mathbf{K} \subseteq \mathbf{V}$. We say that $\mathbf{K}$ is* coherent *if it is non-empty and, for some $w \in \mathbf{V}$, there is a path in $\mathbf{G}$ from each vertex in $\mathbf{K}$ to w.*

If $\mathbf{G} \in \mathcal{G}$ then every non-empty subset of $\mathbf{V}$ is coherent, but to turn Definition 7 into a family of ranking algorithms we must consider what happens if the spammer breaks this coherence. For each positive integer k and each real number $\varepsilon \in (0,1)$, we now define the ranking algorithm T-Min-PPR$_{k,\varepsilon}$ as follows. (As in T-PPR$_{\varepsilon}$, the "T" stands for "trusted".)

Let $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ be a graph and let $\mathbf{T} \subseteq \mathbf{V}$ be a non-empty set of trusted nodes. The algorithm chooses an arbitrary subset $\mathbf{K}$ of $\mathbf{T}$ of size $\min\{k, |\mathbf{T}|\}$ which does not depend on $\mathbf{G}$. It then chooses an arbitrary maximum-size set $\mathbf{K}' \subseteq \mathbf{K}$ which is coherent in $\mathbf{G}$, and outputs $\mathcal{R}_{\min}(\mathbf{G}, \mathbf{K}', \varepsilon)$. (Note that some choice of $\mathbf{K}'$ must exist, since any 1-vertex set is coherent, and that we have $\mathbf{K}' = \mathbf{K}$ unless the spammer has disrupted the coherence of $\mathbf{K}$.)

Perhaps surprisingly, the output of T-Min-PPR$_{k,\varepsilon}$ is not just a normalized minimum of PageRanks, but a PageRank in itself.

LEMMA 8. *Let $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ be a graph, let $\varepsilon \in (0, 1)$, and let $\mathbf{K} \subseteq \mathbf{V}$ be coherent. Then there exists a reset vector $\hat{\mathbf{r}}$ such that $\mathcal{R}_{\min}(\mathbf{G}, \mathbf{K}, \varepsilon) = \mathcal{R}(\mathbf{G}, \hat{\mathbf{r}}, \varepsilon)$.*

In fact, we prove a stronger result as Theorem 18: the class of PageRanks is closed under normalized component-wise min whenever the component-wise min is not identically zero.

Min-PPR has low distortion. Our first result says that on suitably fast-mixing graphs, *any* PageRank has low contraction.

LEMMA 9. *Let $\delta > 0$, let $\varepsilon \in (0, 1)$, and let $T(n)$ be any function such that, for all n,*

$$0 \leq T(n) \leq 1/(2\varepsilon(3 + \delta \log_2 n)).$$

Then for all n-vertex graphs $\mathbf{G} \in \mathcal{G}_T$, all reset vectors $\hat{\mathbf{r}}$ on $\mathbf{G}$, and all $y \in \mathbf{V}_{\mathbf{G}}$, we have $\mathrm{Cont}_\delta(\mathcal{R}(\mathbf{G}, \hat{\mathbf{r}}, \varepsilon), \mathbf{G}, y) \leq 1 + 2\varepsilon T(n)(3 + \delta \log_2 n)$.

Lemma 9 confirms our intuition about the behavior of PPR: On fast-mixing graphs, its inaccuracy is solely the result of a large spike of bias around its trusted center, which Min-PPR can correct for. We will use this result in our analysis of Min-PPR. As a corollary (see Theorem 26), we see that any PageRank has total variation distance at most $\varepsilon T(n)(3 + \log_2 n)$ from $\mathcal{R}(\mathbf{G})$.

We cannot hope for $\mathcal{R}_{\min}(\mathbf{G}, \mathbf{K}, \varepsilon)$ to have low distortion for an arbitrary (coherent) choice of $\mathbf{K}$, since, if the vertices of $\mathbf{K}$ are clustered together, then their distortion spikes may overlap and cause Min-PPR to suffer the same distortion as PPR. But it is nevertheless true that good choices of $\mathbf{K}$ are very common and easy to find.

THEOREM 10 (**MAIN RESULT**). *Let $\delta \geq 1$. Let $\varepsilon \in (0, 1)$ and let $T(n) \leq 1/(210\varepsilon\delta \log_2 n)$. Let $\mathbf{G} \in \mathcal{G}$ be an n-vertex graph with $n \geq 3$, and suppose that the worst-case mixing time of G is at most $T(n)$. Let $k \geq 1$, let $\hat{\mathbf{r}}$ be an arbitrary reset vector, let $X_1, \dots, X_k$ be drawn independently from $\mathbf{V}_{\mathbf{G}}$ with probabilities given by $\mathcal{R}(\mathbf{G}, \hat{\mathbf{r}}, \varepsilon)$, and let $\mathbf{K} = \{X_1, \dots, X_k\}$. Then with probability at least $1 - 4^{-k}n$, the distortion of* T-Min-PPR$_{k,\varepsilon}(\mathbf{G}, \mathbf{K})$ *satisfies* $\mathbf{D}_\delta(\mathcal{R}_{\min}(\mathbf{G}, \mathbf{K}, \varepsilon), \mathbf{G}) \leq 1 + 210\varepsilon\delta T(n) \log_2 n$.

Thus according to $\mathcal{R}(\mathbf{G})$, UPR, or any other PageRank, $\Theta(\log n)$-sized sets of centers giving rise to accurate Min-PPR's are very common. (The reason we can afford to be so flexible in our choice of distribution is that all PageRanks are close in total variation distance, as stated above.) Note that the dependence of Theorem 10 on our significance parameter δ is very mild.

Recall that Theorem 10 requires fast worst-case mixing, i.e. that the uniform random walk mixes quickly from every vertex in $\mathbf{G}$. While this is a common assumption, as discussed in Section 4, some web-like graphs may exhibit only fast average-case mixing. For this reason, as Theorem 31, we prove a version of Theorem 10 which requires only fast average-case mixing from our chosen centers. We prove the result by altering T-Min-PPR$_{k,\varepsilon}$ to use only a carefully-chosen subset of the trusted vertices.

T-Min-PPR has high spam resistance. As noted above, Min-PPR has zero spam resistance by Observation 2. T-Min-PPR, however, is highly spam resistant.

THEOREM 11 (**MAIN RESULT**). *For any $\varepsilon \in (0,1)$ and any positive integer k,* T-Min-PPR$_{k,\varepsilon}$ *is $(\varepsilon/3k)$-spam resistant on n-vertex graphs in $\mathcal{G}$ with worst-case mixing time at most $1/(3\varepsilon(3+\log_2 n))$. A cost function that establishes this spam resistance is the average of the cost functions of the component T-PPRs.*

Recall from Lemma 3 that T-PPR$_\varepsilon$ is at least ε-spam resistant on all graph classes, and recall from Observation 4 that this is close to tight even on fast-mixing graphs — for example, T-PPR$_\varepsilon$ is not $(1.01\varepsilon/(1-\varepsilon))$-spam resistant even on the class of cliques, which is contained in $\mathcal{G}_1$. Since Theorem 11 shows that T-Min-PPR$_{k,\varepsilon}$ is $(\varepsilon/3k)$-spam resistant on suitably fast-mixing graphs, we conclude that in this setting T-Min-PPR$_{k,\varepsilon}$ inherits most of the spam resistance of T-PPR$_\varepsilon$.

Min versus Median. It is natural to ask whether, instead of taking the normalized component-wise minimum of our PPRs, we could take the normalized component-wise median. We show that Min-PPR has a crucial advantage over Median-PPR: The normalized minimum of any set of PageRanks with reset probability ε is itself a PageRank with reset probability ε, whereas the normalized median of any set of PageRanks with reset probability ε is a PageRank with possibly much larger ε.

This preservation of ε is important, as without it the closure condition that we introduce in the next section to show that T-Min-PPR$_{k,\varepsilon}$ is a PageRank would be so weak as to be useless. Notice that, as the reset probability is allowed to become arbitrarily large, the resulting PageRank will approach the reset vector, with little contribution from the underlying graph — indeed, we show that any strictly positive vector whose entries sum to 1 is a PageRank.

In Section 6, we make a distinction between operators that are *strongly closed*, which means they preserve the reset probability, and those that are *weakly closed*, which means that they might not. We give simple necessary and sufficient conditions for a vector to be a PageRank with a given reset probability, and show that min is strongly closed for PageRank (Theorem 18), whereas median is only weakly closed (Lemma 20 and 21). Indeed, the reset probability of Median-PPR may be as high as 1/2 even when ε is arbitrarily small.

Summary. Overall, suppose that our n-vertex pre-spam graph $\mathbf{G}$ lies in $\mathcal{G}_{\mathrm{polylog}(n)}$. Choose our significance threshold $1/n^\delta$ arbitrarily subject to $\delta \geq 1$, and take $k = \Theta(\log n)$ and $\varepsilon = 1/\mathrm{polylog}(n)$. (This requires a rough estimate of n, but this should not be a major obstacle in practice.) Then Theorem 11 implies that T-Min-PPR$_{k,\varepsilon}$ is $(\varepsilon/3k)$-spam resistant, so that for all possible choices of $\mathbf{T}_\mathbf{G} \subseteq \mathbf{V}_\mathbf{G}$ and all possible spam graphs $\mathbf{H}$, T-Min-PPR$_{k,\varepsilon}(\mathbf{H}, \mathbf{T}_\mathbf{G})$ does not award a disproportionate amount of rank to the spammer.

In this setting, T-PPR$_\varepsilon$ is ε-spam resistant but not (2ε)-spam resistant, so we see that T-Min-PPR$_{k,\varepsilon}$ resists spam almost as well as T-PPR$_\varepsilon$. Moreover, Theorem 10 implies that there are many possible choices of $\mathbf{T}_\mathbf{G}$ such that T-Min-PPR$_{k,\varepsilon}(\mathbf{G}, \mathbf{T}_\mathbf{G})$ has $1 + o(1)$ distortion, so that T-Min-PPR$_{k,\varepsilon}$ is accurate on the pre-spam graph. Thus T-Min-PPR$_{k,\varepsilon}$ performs far better than T-PPR$_\varepsilon$, which can have distortion $\Omega(n)$ for all choices of $\mathbf{T}_\mathbf{G}$, or even worst-case UPR$_\varepsilon$, which can have distortion $\Omega(n^\delta)$. Finally, T-Min-PPR is a PageRank by Lemma 8. We therefore believe that T-Min-PPR is a promising new ranking algorithm that warrants significant further study.

6 PAGERANK CLOSURE

In this section, we first introduce notation and make some preliminary observations about PageRank. We then show that PageRank is closed under normalized component-wise min. Finally, we show that PageRank is not closed under all functions, and in particular that it is not closed under median.

6.1 PageRank Preliminaries

Recall that $(\mathbf{G}, \hat{\mathbf{r}}, \varepsilon)$ is the random walk associated with PageRank on $\mathbf{G}$ with reset vector $\hat{\mathbf{r}}$ and reset probability ε. The transition probability matrix, $\mathbf{A}$, of this walk is

$$\mathbf{A} = (1-\varepsilon)\mathbf{M} + \varepsilon\mathbf{R}, \tag{1}$$

where $\mathbf{M}$ and $\mathbf{R}$ denote $n \times n$ matrices as follows:

$$\forall u, v \in \mathbf{V} : \mathbf{M}[u,v] = \begin{cases} 1/|N_{\text{out}}(u)| & \text{if } (u,v) \in \mathbf{E}, \\ 0 & \text{otherwise,} \end{cases}$$
$$\mathbf{R}[u,v] = \hat{\mathbf{r}}[v].$$

For instance, if $\hat{\mathbf{r}}$ is the uniform reset vector $\hat{\mathbf{u}}$, then $\mathbf{R}[u,v] = 1/n$ for all $u, v \in \mathbf{V}$.

Brin and Page noted that $\mathcal{R}(\mathbf{G}, \hat{\mathbf{u}}, \varepsilon)$ is total, that is, it is defined and unique for any graph $\mathbf{G}$ and any $\varepsilon \in (0,1)$. We observe the following more general folklore lemma for PageRanks.

LEMMA 12. *$\mathcal{R}(\mathbf{G}, \hat{\mathbf{r}}, \varepsilon)$ is defined and unique for any graph $\mathbf{G} = (\mathbf{V}, \mathbf{E})$, reset vector $\hat{\mathbf{r}}$, and reset probability $\varepsilon \in (0,1)$.*

PROOF. Let $\mathbf{V}_{\hat{\mathbf{r}}}$ be the support of $\hat{\mathbf{r}}$. Notice that these nodes belong to a single strongly connected component in the walk $(\mathbf{G}, \hat{\mathbf{r}}, \varepsilon)$ consisting of the nodes reachable from $\mathbf{V}_{\hat{\mathbf{r}}}$. These nodes form a unique essential communicating class[5] in the Markov chain of the random walk on $\mathbf{A}$. By Proposition 1.26 in [Levin et al., 2009], such a Markov chain has a unique stationary distribution. □

A similar claim was proved in [Andersen et al., 2006], but for undirected graphs. For directed graphs, as in our case, $\mathcal{R}(\mathbf{G}, \hat{\mathbf{r}}, \varepsilon)$ has weight 0 on all nodes not reachable from a node in $\mathbf{V}_{\hat{\mathbf{r}}}$. We will also need one more ancillary lemma.

LEMMA 13. *Let $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ be an arbitrary graph, let $0 < \varepsilon < 1$, and let $\hat{\mathbf{r}}$ be a reset vector. Let $(Y_t)_{t\geq 0}$ be the uniform random walk on $\mathbf{G}$ with random initial state drawn from $\hat{\mathbf{r}}$. Then for all $v \in \mathbf{V}$, we have*

$$\mathcal{R}(\mathbf{G}, \hat{\mathbf{r}}, \varepsilon)[v] = \varepsilon \sum_{i=0}^{\infty} (1-\varepsilon)^i \mathbb{P}(Y_i = v).$$

PROOF. Follows from Equation 5 in [Jeh and Widom, 2003] and linearity of expectation. □

For all $\varepsilon \in (0,1)$, we denote the set of all possible PageRanks for $\mathbf{G}$ with reset probability ε by $\mathcal{P}_\varepsilon(\mathbf{G}) = \{\mathcal{R}(\mathbf{G}, \hat{\mathbf{r}}, \varepsilon) \colon \hat{\mathbf{r}} \text{ is a ranking vector}\}$. (Recall from Section 3 that a ranking vector on $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ is any function $\hat{\mathbf{x}} \colon \mathbf{V} \to [0,1]$ with $\sum_{v \in \mathbf{V}} \hat{\mathbf{x}}[v] = 1$.) We denote the set of all possible PageRanks for $\mathbf{G}$ with any reset probability by $\mathcal{P}(\mathbf{G}) = \{\mathcal{R}(\mathbf{G}, \hat{\mathbf{r}}, \varepsilon) \colon \hat{\mathbf{r}} \text{ is a ranking vector}, \varepsilon \in (0,1)\}$.

We now set out a necessary and sufficient condition for a ranking vector to be a PageRank with a given reset probability. For all graphs $\mathbf{G} = (\mathbf{V}, \mathbf{E})$, all ranking vectors $\hat{\mathbf{p}}$ on $\mathbf{G}$, and all $\varepsilon \in (0,1)$, define

$$\mathcal{R}^{-1}(\mathbf{G}, \hat{\mathbf{p}}, \varepsilon)[v] := \frac{\hat{\mathbf{p}}[v]}{\varepsilon} - \frac{1-\varepsilon}{\varepsilon} \sum_{w \in N_{\text{in}}(v)} \frac{\hat{\mathbf{p}}[w]}{d_{\text{out}}(w)} \text{ for all } v \in \mathbf{V}.$$

LEMMA 14. *Let $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ be a graph, let $\varepsilon \in (0,1)$, and let $\hat{\mathbf{p}}$ be a ranking vector on $\mathbf{G}$. If $\hat{\mathbf{p}} = \mathcal{R}(\mathbf{G}, \hat{\mathbf{r}}, \varepsilon)$ for some $\hat{\mathbf{r}}$, then $\hat{\mathbf{r}} = \mathcal{R}^{-1}(\mathbf{G}, \hat{\mathbf{p}}, \varepsilon)$. Moreover, $\hat{\mathbf{p}} \in \mathcal{P}_\varepsilon(\mathbf{G})$ if and only if $\mathcal{R}^{-1}(\mathbf{G}, \hat{\mathbf{p}}, \varepsilon) \geq \mathbf{0}$.*

[5] States i and j of a Markov chain belong to the same communicating class if there is a positive probability of moving to state j from state i, and a positive probability of moving to state i from state j.

PROOF. First suppose that $\hat{\mathbf{p}} = \mathcal{R}(\mathbf{G}, \hat{\mathbf{r}}, \varepsilon)$ for some $\hat{\mathbf{r}}$. Let $\mathbf{M}$ be the transition matrix of a uniform random walk on $\mathbf{G}$ whose initial state is given by $\hat{\mathbf{r}}$, and let $\mathbf{R}$ be the $|\mathbf{V}| \times |\mathbf{V}|$ matrix whose rows are given by $\hat{\mathbf{r}}$. By definition, $\hat{\mathbf{p}}$ is the unique (row) vector satisfying $\hat{\mathbf{p}} = \hat{\mathbf{p}}(\varepsilon\mathbf{R} + (1-\varepsilon)\mathbf{M})$. Equivalently, $\hat{\mathbf{p}}$ is the unique vector such that for all $v \in \mathbf{V}$,

$$\hat{\mathbf{p}}[v] = \varepsilon \sum_{w \in \mathbf{V}} \hat{\mathbf{p}}[w]\hat{\mathbf{r}}[v] + (1-\varepsilon) \sum_{w \in \mathbf{V}} \hat{\mathbf{p}}[w]\mathbf{M}[w,v] = \varepsilon\hat{\mathbf{r}}[v] + (1-\varepsilon) \sum_{w \in N_{\text{in}}(v)} \frac{\hat{\mathbf{p}}[w]}{d_{\text{out}}(w)}. \tag{2}$$

Rearranging, we obtain $\hat{\mathbf{r}}[v] = \mathcal{R}^{-1}(\mathbf{G}, \hat{\mathbf{p}}, \varepsilon)[v]$, and so $\hat{\mathbf{r}} = \mathcal{R}^{-1}(\mathbf{G}, \hat{\mathbf{p}}, \varepsilon)$ as required. This also implies that if $\hat{\mathbf{p}} \in \mathcal{P}_\varepsilon$, then $\mathcal{R}^{-1}(\mathbf{G}, \hat{\mathbf{p}}, \varepsilon) \geq \mathbf{0}$.

Suppose now that $\mathcal{R}^{-1}(\mathbf{G}, \hat{\mathbf{p}}, \varepsilon) \geq \mathbf{0}$. We have

$$||\mathcal{R}^{-1}(\mathbf{G}, \hat{\mathbf{p}}, \varepsilon)|| = \frac{1}{\varepsilon} - \frac{1-\varepsilon}{\varepsilon} \sum_{v \in \mathbf{V}} \sum_{w \in N_{\text{in}}(v)} \frac{\hat{\mathbf{p}}[w]}{d_{\text{out}}(w)} = \frac{1}{\varepsilon} - \frac{1-\varepsilon}{\varepsilon} \sum_{w \in \mathbf{V}} \sum_{v \in N_{\text{out}}(w)} \frac{\hat{\mathbf{p}}[w]}{d_{\text{out}}(w)} = \frac{1}{\varepsilon} - \frac{1-\varepsilon}{\varepsilon} = 1,$$

so $\mathcal{R}^{-1}(\mathbf{G}, \hat{\mathbf{p}}, \varepsilon)$ is a ranking vector on $\mathbf{G}$. Moreover, taking $\hat{\mathbf{r}} = \mathcal{R}^{-1}(\mathbf{G}, \hat{\mathbf{p}}, \varepsilon)$, for all $v \in \mathbf{V}$ we have

$$\varepsilon\hat{\mathbf{r}}[v] + (1-\varepsilon) \sum_{w \in N_{\text{in}}(v)} \frac{\hat{\mathbf{p}}[w]}{d_{\text{out}}(w)} = \hat{\mathbf{p}}[v] - (1-\varepsilon) \sum_{w \in N_{\text{in}}(v)} \frac{\hat{\mathbf{p}}[w]}{d_{\text{out}}(w)} + (1-\varepsilon) \sum_{w \in N_{\text{in}}(v)} \frac{\hat{\mathbf{p}}[w]}{d_{\text{out}}(w)} = \hat{\mathbf{p}}[v].$$

Hence by (2), we have $\hat{\mathbf{p}} = \mathcal{R}(\mathbf{G}, \hat{\mathbf{r}}, \varepsilon)$, and in particular $\hat{\mathbf{p}} \in \mathcal{P}_\varepsilon(\mathbf{G})$. □

Next, using Lemma 14, we set out a simple necessary and sufficient condition for a ranking vector to be a PageRank at all.

LEMMA 15. *Let* $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ *be a graph, and let* $\hat{\mathbf{p}}$ *be a ranking vector on* $\mathbf{G}$*. Then* $\hat{\mathbf{p}} \in \mathcal{P}(\mathbf{G})$ *if and only if for all* $(v, w) \in \mathbf{E}$*, if* $\hat{\mathbf{p}}[v] > 0$*, then* $\hat{\mathbf{p}}[w] > 0$.

PROOF. Suppose $\hat{\mathbf{p}} \in \mathcal{P}(\mathbf{G})$ with $\hat{\mathbf{p}} = \mathcal{R}(\mathbf{G}, \hat{\mathbf{r}}, \varepsilon)$, let $v \in \mathbf{V}$, and suppose $\hat{\mathbf{p}}[v] > 0$. Then for all $w \in N_{\text{out}}(v)$, the PageRank random walk associated with $\hat{\mathbf{p}}$ transitions from v to w with probability at least $(1-\varepsilon)/|N_{\text{out}}(v)| > 0$, so we must have $\hat{\mathbf{p}}[w] > 0$.

Conversely, let $\hat{\mathbf{p}}$ be a ranking vector on $\mathbf{G}$, and suppose that $\hat{\mathbf{p}}$ satisfies the condition that for all $(v, w) \in \mathbf{E}$, if $\hat{\mathbf{p}}[v] > 0$, then $\hat{\mathbf{p}}[w] > 0$. For all $v \in \mathbf{V}$, let

$$\Sigma_v := \sum_{w \in N_{\text{in}}(v)} \frac{\hat{\mathbf{p}}[w]}{|N_{\text{out}}(w)|}, \qquad\qquad x_v := \begin{cases} 1 - \hat{\mathbf{p}}[v]/\Sigma_v & \text{if } \Sigma_v \neq 0, \\ 0 & \text{otherwise,} \end{cases}$$

and let $\varepsilon = \max(\{1/2\} \cup \{x_v \colon v \in \mathbf{V}\})$. We now show that $\hat{\mathbf{p}} \in \mathcal{P}_\varepsilon(\mathbf{G})$ by showing that $\varepsilon \in (0,1)$, that $\mathcal{R}^{-1}(\mathbf{G}, \hat{\mathbf{p}}, \varepsilon) \geq \mathbf{0}$, and applying Lemma 14.

By definition, $\varepsilon \geq 1/2 > 0$. For all $v \in \mathbf{V}$ with $\Sigma_v \neq 0$, there must exist $w \in N_{\text{in}}(v)$ with $\hat{\mathbf{p}}[w] > 0$, so by hypothesis we have $\hat{\mathbf{p}}[v] > 0$; hence $x_v < 1$. When $\Sigma_v = 0$ we have $x_v = 0 < 1$ by definition, so it follows that $\varepsilon < 1$; hence $\varepsilon \in (0,1)$.

Now let $v \in \mathbf{V}$. If $\Sigma_v = 0$, then $\mathcal{R}^{-1}(\mathbf{G}, \hat{\mathbf{p}}, \varepsilon)[v] = \hat{\mathbf{p}}[v]/\varepsilon \geq 0$. If instead $\Sigma_v > 0$, then we have

$$\mathcal{R}^{-1}(\mathbf{G}, \hat{\mathbf{p}}, \varepsilon)[v] \geq \frac{1}{\varepsilon}\hat{\mathbf{p}}[v] - \frac{1-x_v}{\varepsilon}\Sigma_v = 0.$$

Thus $\mathcal{R}^{-1}(\mathbf{G}, \hat{\mathbf{p}}, \varepsilon) \geq \mathbf{0}$, so it follows by Lemma 14 that $\hat{\mathbf{p}} \in \mathcal{P}_\varepsilon(\mathbf{G})$. In particular, $\hat{\mathbf{p}} \in \mathcal{P}(\mathbf{G})$ as required. □

In Lemma 17, we enumerate the possible realizations of a given ranking vector as a PageRank; before proving this, we introduce an ancillary lemma.

LEMMA 16. *Let* $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ *be a graph, let* $\hat{\mathbf{p}} \in \mathcal{P}(\mathbf{G})$*, let* $\hat{\mathbf{r}}$ *be a ranking vector on* $\mathbf{G}$*, and suppose there exists* $\varepsilon \in (0,1)$ *such that* $\hat{\mathbf{p}} = \mathcal{R}(\mathbf{G}, \hat{\mathbf{r}}, \varepsilon)$*. Then exactly one of the following holds:*

(1) there exists $v \in \mathbf{V}$ *such that* $\hat{\mathbf{r}}[v] \neq (\hat{\mathbf{p}}\mathbf{M})[v]$*, in which case*

$$\varepsilon = \frac{\hat{\mathbf{p}}[v] - (\hat{\mathbf{p}}\mathbf{M})[v]}{\hat{\mathbf{r}}[v] - (\hat{\mathbf{p}}\mathbf{M})[v]}; \text{ or}$$

(2) $\hat{\mathbf{r}} = \hat{\mathbf{p}} = \hat{\mathbf{p}}\mathbf{M}$*, in which case* $\hat{\mathbf{p}} = \mathcal{R}(\mathbf{G}, \hat{\mathbf{r}}, \eta)$ *for all* $\eta \in (0, 1)$*.*

PROOF. First suppose there exists $v \in \mathbf{V}$ such that $\hat{\mathbf{r}}[v] \neq (\hat{\mathbf{p}}\mathbf{M})[v]$. By Lemma 14, we have

$$\hat{\mathbf{r}}[v] = \mathcal{R}^{-1}(\mathbf{G}, \hat{\mathbf{p}}, \varepsilon)[v] = \frac{\hat{\mathbf{p}}[v]}{\varepsilon} - \frac{1-\varepsilon}{\varepsilon}(\hat{\mathbf{p}}\mathbf{M})[v].$$

Rearranging, we obtain

$$\varepsilon = \frac{\hat{\mathbf{p}}[v] - (\hat{\mathbf{p}}\mathbf{M})[v]}{\hat{\mathbf{r}}[v] - (\hat{\mathbf{p}}\mathbf{M})[v]},$$

as required.

If no such $v \in \mathbf{V}$ exists, then we must have $\hat{\mathbf{r}} = \hat{\mathbf{p}}\mathbf{M}$. Since $\hat{\mathbf{p}}$ is a PageRank, it follows that

$$\hat{\mathbf{p}} = \hat{\mathbf{p}}\mathbf{A} = \hat{\mathbf{p}}\big((1-\varepsilon)\mathbf{M} + \varepsilon\mathbf{R}\big) = (1-\varepsilon)\hat{\mathbf{r}} + \varepsilon\hat{\mathbf{r}} = \hat{\mathbf{r}},$$

so $\hat{\mathbf{r}} = \hat{\mathbf{p}}$. Since $\hat{\mathbf{r}} = \hat{\mathbf{p}}\mathbf{M}$, it follows that $\hat{\mathbf{r}} = \hat{\mathbf{p}} = \hat{\mathbf{p}}\mathbf{M}$ as required. Finally, for all $\eta \in (0, 1)$, we have

$$\hat{\mathbf{p}}\big((1-\eta)\mathbf{M} + \eta\mathbf{R}\big) = (1-\eta)\hat{\mathbf{p}}\mathbf{M} + \eta\hat{\mathbf{r}} = \hat{\mathbf{p}},$$

so $\hat{\mathbf{p}} = \mathcal{R}(\mathbf{G}, \hat{\mathbf{r}}, \eta)$ as required. □

LEMMA 17. *Let* $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ *be a graph, and let* $\hat{\mathbf{p}} \in \mathcal{P}(\mathbf{G})$*. Let* $\mathbf{X}$ *be the set of all pairs* $(\hat{\mathbf{r}}, \varepsilon)$ *such that* $\hat{\mathbf{p}} = \mathcal{R}(\mathbf{G}, \hat{\mathbf{r}}, \varepsilon)$*. If* $\hat{\mathbf{p}}\mathbf{M} = \hat{\mathbf{p}}$*, then we have* $\mathbf{X} = \{(\hat{\mathbf{p}}, \varepsilon) \colon \varepsilon \in (0, 1)\}$*; otherwise, we have*

$$\mathbf{X} = \{(\mathcal{R}^{-1}(\mathbf{G}, \hat{\mathbf{p}}, \varepsilon), \varepsilon) \colon \mathcal{R}^{-1}(\mathbf{G}, \hat{\mathbf{p}}, \varepsilon) \geq \mathbf{0}\}.$$

PROOF. Immediate from Lemmas 14 and 16. □

6.2 PageRank Closure Definitions

Let g be a function that takes a finite set of PageRanks on a graph $\mathbf{G}$ and returns a ranking vector for $\mathbf{G}$. We say that g is *weakly closed on PageRanks* if, for any graph $\mathbf{G}$ and any finite set $Q \subset \mathcal{P}(\mathbf{G})$, we have $g(Q) \in \mathcal{P}(\mathbf{G})$.

The reason we call this type of closure "weak" is that by Lemma 15, being a PageRank imposes only a mild condition satisfied by e.g. any vector with no zero entries. The reason this mild condition suffices is that the reset probability, ε, may be arbitrarily close to 1. If PageRank is to capture any graph structure, as opposed to simply approximating the reset vector, then ε needs to be well below, say, 1/2.

Therefore, we define g to be *strongly closed on PageRanks* if, for all graphs $\mathbf{G}$, all $\varepsilon \in (0, 1)$, and all finite sets $Q \subset \mathcal{P}_\varepsilon(\mathbf{G})$, we have $g(Q) \in \mathcal{P}_\varepsilon(\mathbf{G})$. That is, strong closure implies that the operator produces a PageRank without changing ε.

In the following two sections, we demonstrate that min is strongly closed for PageRanks, whereas median is only weakly closed.

6.3 Strongly Closed Operators

In this section, we show that PageRank is strongly closed under the normalized component-wise min operator and prove Theorem 18, which implies Theorem 8.

THEOREM 18. *Let* $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ *be a graph, let* $\varepsilon \in (0, 1)$*, and suppose* $\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_k \in \mathcal{P}_\varepsilon(\mathbf{G})$*. For all* $v \in \mathbf{V}$*, let* $\mathbf{y}[v] = \min\{\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_k\}[v]$*, suppose* $\mathbf{y} \neq \mathbf{0}$*, and let* $\hat{\mathbf{y}} = \mathbf{y}/||\mathbf{y}||$*. Then* $\hat{\mathbf{y}} \in \mathcal{P}_\varepsilon(\mathbf{G})$*.*

PROOF. First note that $\hat{\mathbf{y}}$ is a ranking vector for $\mathbf{G}$. Let $v \in \mathbf{V}$. Then for all $i \in [k]$, since $\hat{\mathbf{x}}_i \in \mathcal{P}_\varepsilon(\mathbf{G})$, by Lemma 14 we have

$$\mathcal{R}^{-1}(\mathbf{G}, \hat{\mathbf{x}}_i, \varepsilon)[v] = \frac{\hat{\mathbf{x}}_i[v]}{\varepsilon} - \frac{1-\varepsilon}{\varepsilon} \sum_{w \in N_{\text{in}}(v)} \frac{\hat{\mathbf{x}}_i[w]}{d_{\text{out}}(w)} \geq \mathbf{0}. \tag{3}$$

For all $v \in \mathbf{V}$, we have $\mathbf{y}[v] = \hat{\mathbf{x}}_j[v]$ for some $j \in [k]$, and $\mathbf{y}[w] \leq \hat{\mathbf{x}}_j[w]$ for all $w \in N_{\text{in}}(v)$. It follows from (3) that

$$\frac{\mathbf{y}[v]}{\varepsilon} - \frac{1-\varepsilon}{\varepsilon} \sum_{w \in N_{\text{in}}(v)} \frac{\mathbf{y}[w]}{d_{\text{out}}(w)} \geq \mathbf{0},$$

and hence $\mathcal{R}^{-1}(\mathbf{G}, \hat{\mathbf{y}}, \varepsilon)[v] \geq 0$. Thus $\mathcal{R}^{-1}(\mathbf{G}, \hat{\mathbf{y}}, \varepsilon) \geq \mathbf{0}$, so by Lemma 14 we have $\hat{\mathbf{y}} \in \mathcal{P}_\varepsilon(\mathbf{G})$. □

Recall from Section 5 that for any graph $\mathbf{G} = (\mathbf{V}, \mathbf{E})$, any $\varepsilon \in (0,1)$, and any coherent set $\mathbf{K} \subseteq \mathbf{V}$, we have $\min\{\mathcal{R}(\mathbf{G}, x, \varepsilon) \colon x \in \mathbf{K}\} \neq \mathbf{0}$. Thus Theorem 18 implies that $\mathcal{R}_{\min}(\mathbf{G}, \mathbf{K}, \varepsilon) \in \mathcal{P}_\varepsilon(\mathbf{G})$, so Lemma 8 follows.

6.4 Weakly Closed Operators

In this section we show that median is weakly closed, but not strongly closed, over PageRanks. We first define the median operator on vectors by component-wise median, so that for vectors $\mathbf{x_1}, \dots, \mathbf{x_k}$,

$$\text{median}\{\mathbf{x_1}, \dots, \mathbf{x_k}\}[i] = \text{median}\{\mathbf{x_1}[i], \dots, \mathbf{x_k}[i]\}.$$

We now formally define the median operator on ranking vectors. The normalized component-wise median of PageRanks may not be well-defined as the median might be identically zero. Therefore, we add a condition to the definition to avoid those cases.

DEFINITION 19. *Let* $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ *be a graph and let* $\mathbf{X} = \{\hat{\mathbf{x}}_1, \dots, \hat{\mathbf{x}}_k\}$ *such that* $\hat{\mathbf{x}}_i \in \mathcal{P}(\mathbf{G})$ *for every* $i = 1, \dots, k$ *and* $||\text{median}\{\hat{\mathbf{x}}_\mathbf{i} \colon \hat{\mathbf{x}}_i \in \mathbf{X}\}|| > 0$. *Then, we define the Median operator as*

$$\mathcal{R}_{\text{med}}(\mathbf{G}, \mathbf{X}) = \llbracket \text{median}\{\hat{\mathbf{x}}_\mathbf{i} \colon \hat{\mathbf{x}}_i \in \mathbf{X}\} \rrbracket$$

The following theorem says that not only is the median operator not strongly closed, but that the property fails badly — in general, we cannot express the normalized component-wise median of even low-reset-probability PPRs as a PageRank without using a reset probability greater than $1/2$.

LEMMA 20. *Let* $\varepsilon \in (0,1)$. *Then there exist infinitely many graphs* $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ *and sets of PPRs* $\mathbf{X} = \{\hat{\mathbf{x}}_1, \dots, \hat{\mathbf{x}}_k\}$ *such that* $\hat{\mathbf{x}}_1, \dots, \hat{\mathbf{x}}_k \in \mathcal{P}_\varepsilon(\mathbf{G})$ *and* $||\text{median}\{\hat{\mathbf{x}}_1, \dots, \hat{\mathbf{x}}_k\}|| > 0$, *but* $\mathcal{R}_{\text{med}}(\mathbf{G}, \mathbf{X}) \notin \mathcal{P}_\varepsilon(\mathbf{G})$. *Moreover,* $\mathcal{R}_{\text{med}}(\mathbf{G}, \mathbf{X}) \notin \mathcal{P}_\eta(\mathbf{G})$ *for any* $\eta \leq 1/2$.

PROOF. Let k be any odd integer satisfying $k \geq 3$ and $k > (1-\varepsilon)/\varepsilon$, and write $k =: 2\ell+1$. Consider the graph in Figure 1, where each node u_i is connected to the $\ell+1$ nodes $v_i, v_{(i+1) \bmod k}, \dots, v_{(i+\ell) \bmod k}$. For all $i \in [k]$, we define $\hat{\mathbf{x}}_i := \mathcal{R}(\mathbf{G}, u_i, \varepsilon)$. Thus the reset vector $\hat{\mathbf{r}}_{\hat{\mathbf{x}}_i}$ of $\hat{\mathbf{x}}_i$ satisfies $\hat{\mathbf{r}}_{\hat{\mathbf{x}}_i}[u_i] = 1$ and $\hat{\mathbf{r}}_{\hat{\mathbf{x}}_i}[v] = 0$ for all $v \neq u_i$. By Lemma 13, we have

$$\hat{\mathbf{x}}_i[v_j] = \begin{cases} \varepsilon(1-\varepsilon)/(\ell+1) & \text{if } u_i \in N_{\text{in}}(v_j), \\ 0 & \text{otherwise}, \end{cases} \qquad \hat{\mathbf{x}}_i[y_1] = \varepsilon(1-\varepsilon)^2. \tag{4}$$

For brevity, write $a = ||\text{median}\{\hat{\mathbf{x}}_1, \dots, \hat{\mathbf{x}}_k\}||$, and note that $a > 0$. Since $\ell + 1 > k/2$, it follows from (4) that

$$\mathcal{R}_{\text{med}}(\mathbf{G}, \mathbf{X})[v_j] = \varepsilon(1-\varepsilon)/(a(\ell+1)) \text{ for all } j \in [k],$$
$$\mathcal{R}_{\text{med}}(\mathbf{G}, \mathbf{X})[y_1] = \varepsilon(1-\varepsilon)^2/a.$$

Now let $\eta \in (0,1)$. By Lemma 14, whenever $\mathcal{R}^{-1}(G, \mathcal{R}_{med}(G,X), \eta)[y_1] < 0$, we have $\mathcal{R}_{med}(G,X) \notin \mathcal{P}_\eta(G)$. By the definition of $\mathcal{R}^{-1}$, we have

$$\mathcal{R}^{-1}\big(G, \mathcal{R}_{med}(G,X), \eta\big)[y_1] = \frac{\varepsilon(1-\varepsilon)^2}{\eta a} - \frac{1-\eta}{\eta} k \frac{\varepsilon(1-\varepsilon)}{a(\ell+1)} = \frac{\varepsilon(1-\varepsilon)}{\eta a}\Big(1 - \varepsilon - \frac{(1-\eta)k}{\ell+1}\Big).$$

Thus $\mathcal{R}_{med}(G,X) \notin \mathcal{P}_\eta(G)$ whenever $1-\eta > (1-\varepsilon)(\ell+1)/k$, which holds if and only if

$$\eta < \frac{k - \ell - 1 + \varepsilon\ell + \varepsilon}{k} = \frac{1+\varepsilon}{2} - \frac{1-\varepsilon}{2k}. \tag{5}$$

Since $k \geq 3$, (5) holds when $\eta = \varepsilon$, and since $k > (1-\varepsilon)/\varepsilon$, (5) holds for all $\eta \leq 1/2$. The result therefore follows. □

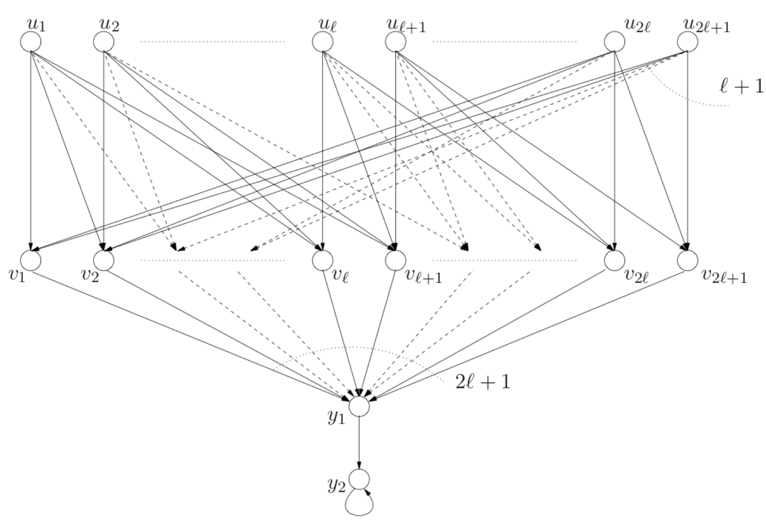


Fig. 1. Illustration of weak closure of median.

Finally we show that Median is weakly closed over PageRanks. Thus taking a median of PageRanks always yields a PageRank, but perhaps one with a much higher reset probability.

LEMMA 21. *[Median is weakly closed over PageRanks] Let* $G = (V, E)$ *be a graph and let* $X = \{\hat{x}_1, \ldots, \hat{x}_{2\ell+1}\}$*, where* $\hat{x}_i \in \mathcal{P}(G)$ *for every* $i = 1, \ldots, 2\ell+1$*. Suppose* $||\text{median}\{\hat{x}_i : \hat{x}_i \in X\}|| > 0$*. Then,* $\mathcal{R}_{med}(G,X) \in \mathcal{P}(G)$.

PROOF. Let $\mathcal{R}_{med}(V,X) = \hat{y}$, and let $v \in V$ with $\hat{y}[v] > 0$. Then $y[v] > 0$, so there exist $i_1, \ldots, i_{\ell+1}$ with $\hat{x}_{i_1}[v], \ldots, \hat{x}_{i_{\ell+1}}[v] > 0$. Since $\hat{x}_1, \ldots, \hat{x}_{2\ell+1} \in \mathcal{P}(G)$, by Lemma 15 it follows that for all $w \in N_{out}(v)$, we have $\hat{x}_{i_1}[w], \ldots, \hat{x}_{i_{\ell+1}}[w] > 0$. Hence by construction we have $y[w] > 0$ and therefore $\hat{y}[w] > 0$ for all $w \in N_{out}(v)$. It follows by Lemma 15 that $\hat{y} \in \mathcal{P}(G)$. □

7 RELATED WORK

Originally, and most famously, PageRank was used by Google as a ranking function for web pages [Google, 2001], but since then, it has been used to analyze networks of neurons [Fletcher and Wennekers, 2016], Twitter recommendation systems [Gupta et al., 2013], protein networks [Iván and Grolmusz, 2010], etc. (See [Gleich, 2015] for a survey of non-web uses).

As noted above, PageRank is susceptible to link spam. Thus, other ranking functions have been proposed [Bhattacharjee and Goel, 2006, Gyöngyi et al., 2004, Kumar et al., 2006]. TrustRank [Gyöngyi et al., 2004] for example is based on assigning higher reputation to a subset of pages curated by an expert, and the assumption that pages linked from these reputable pages are reputable as well. A similar method can be applied for low reputation pages, which is called Anti-Trust Rank [Krishnan and Raj, 2006]. In both, reliability lowers as distance from the reference pages increases.

Other work is geared towards modifications of the PageRank mechanism. For instance, Global Hitting Time [Hopcroft and Sheldon, 2008] was designed as a transformation of PageRank to counter cross-reference link spam, where nodes link each other to increase their rank, but it still suffers if the number of spammers is large. Variants include Personalized Hitting Time [Liu et al., 2016].

Despite the progress on other ranking mechanisms, PageRank still stands as the most popular [StatCounter, [n. d.]] ranking function, and therefore the most attractive for link-spammers. Google discouraged PageRank manipulation through the buying of highly ranked links by social methods: they have announced that pages discovered to participate in such activity will be left out of the PageRank calculation (hence, their rank lowered), they have encouraged the public to notify Google about such pages [Google, [n. d.]].

Other research has focused on link-spam detection [Gyongyi et al., 2006] and quantifying the rank increase obtained by creating Sybil pages [Cheng and Friedman, 2006]. For instance, an algorithm to detect spam analyzing the supporting sets, i.e. the sets of nodes that contribute the most to the PageRank of a given vertex, was presented in [Andersen et al., 2008]. The performance evaluation is experimental. Detection methods for Sybil pages attacks have been surveyed in [Alvisi et al., 2014, Yu, 2011]. Some of those methods [Yu et al., 2010, 2008] are based on detecting abnormal random walk mixing times for what is expected in an "honest" network. Link-spam detection may be useful for excluding pages from the PageRank calculation, but it is better to render an attack futile than to build a fortress. That is, it is better to develop techniques that yield PageRank spam resistant. Towards that end, some work limits or assign reset probability selectively [Fogaras et al., 2005, Gyöngyi et al., 2004]. These approaches are generalizations of Personalized PageRank [Jeh and Widom, 2003].

For graph-theoretic ranking functions, such as Hubs & Authorities (HITS) and PageRank, formalizations of how *stable* they are in face of small perturbations exist [Ng et al., 2001a,b]. Stability refers to how sensitive eigenvector methods such as HITS and PageRank are to small changes in the link structure, and the cost of introducing such perturbations is not considered. Spammability, on the other hand, is a different metric because it relates the cost of perturbations to the gains obtained by those introducing them.

Specifically for HITS on a graph with adjacency matrix A, the authors in [Ng et al., 2001a,b] relate an upper bound on the number of links that may be added (or deleted), given as a function of the maximum out-degree and the eigengap of $A^T A$, to an upper bound on the change of the principal eigenvector of $A^T A$ that those link changes produce. The result characterizes stability because HITS uses the principal eigenvector of $A^T A$ to determine authorities. The authors also show the existence of graphs where a small perturbation (e.g. adding a single link) has a large effect.

For uniform PageRank, in [Ng et al., 2001a,b] they upper bound the aggregated change in rank over all pages (ℓ_1-norm) by a $2/\varepsilon$ factor of the sum of the original rank of the pages whose out-links were changed, where ε is the reset probability. Considering rank as a measure of cost, this result can be seen as relating the overall impact on the system to the cost of introducing changes, but it does not relate to the increase in rank for those nodes. That is, it characterizes stability but not spammability. Moreover, uniform PageRank under this cost measure can be spammed for free by simply creating new nodes, and non-uniform reset vectors are not considered.

As expected, personalized PageRank is biased towards the vicinity of the trusted node. This undesired effect can be compensated for to some extent by concentrating reset probability on a subset of nodes rather than one (as in [Fogaras et al., 2005, Gyöngyi et al., 2004]). Indeed, the approach has been successful for particular areas where the search space is relatively small (e.g. in Linguistic Knowledge Builder graph [Agirre and Soroa, 2009], Social Networks [Bahmani et al., 2010, Jin et al., 2012], and Protein Interaction Networks [Iván and Grolmusz, 2010]). But the scale of the web graph may require a large set of trusted pages for a general purpose PageRank.

REFERENCES

Louigi Addario-Berry and Tao Lei. 2012. The mixing time of the Newman-Watts small world. In *Proceedings of the twenty-third annual ACM-SIAM symposium on Discrete algorithms*. SIAM, 1661–1668.

Eneko Agirre and Aitor Soroa. 2009. Personalizing PageRank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 33–41.

Réka Albert, Hawoong Jeong, and Albert-László Barabási. 1999. Diameter of the world-wide web. *nature* 401, 6749 (1999), 130.

Lorenzo Alvisi, Allen Clement, Alessandro Epasto, Silvio Lattanzi, and Alessandro Panconesi. 2014. Communities, random walks, and social Sybil defense. *Internet Mathematics* 10, 3-4 (2014), 360–420.

Reid Andersen, Christian Borgs, Jennifer T. Chayes, John E. Hopcroft, Kamal Jain, Vahab S. Mirrokni, and Shang-Hua Teng. 2008. Robust PageRank and locally computable spam detection features. In *AIRWeb 2008, Fourth International Workshop on Adversarial Information Retrieval on the Web, Beijing, China, April 22, 2008*. 69–76. https://doi.org/10.1145/1451983.1452000

Reid Andersen, Fan Chung, and Kevin Lang. 2006. Local graph partitioning using PageRank vectors. In *Foundations of Computer Science, 2006. FOCS'06. 47th Annual IEEE Symposium on*. IEEE, 475–486.

Yasuhito Asano, Yu Tezuka, and Takao Nishizeki. 2007. Improvements of HITS algorithms for spam links. In *Advances in Data and Web Management*. Springer, 479–490.

Ricardo A. Baeza-Yates, Carlos Castillo, and Vicente López. 2005. PageRank Increase under Different Collusion Topologies. In *AIRWeb*. 17–24.

Bahman Bahmani, Abdur Chowdhury, and Ashish Goel. 2010. Fast incremental and personalized PageRank. *Proceedings of the VLDB Endowment* 4, 3 (2010), 173–184.

Albert-László Barabási and Réka Albert. 1999. Emergence of scaling in random networks. *science* 286, 5439 (1999), 509–512.

Rajat Bhattacharjee and Ashish Goel. 2006. Incentive based ranking mechanisms. In *First Workshop on the Economics of Networked Systems (Netecon?06)*. 62–68.

Sergey Brin and Lawrence Page. 1998. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks* 30, 1-7 (1998), 107–117. https://doi.org/10.1016/S0169-7552(98)00110-X

Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. 2000. Graph structure in the web. *Computer networks* 33, 1-6 (2000), 309–320.

Alice Cheng and Eric Friedman. 2006. Manipulability of PageRank under Sybil strategies.

Colin Cooper and Alan Frieze. 2007. The cover time of the preferential attachment graph. *Journal of Combinatorial Theory, Series B* 97, 2 (2007), 269–290. https://doi.org/10.1016/j.jctb.2006.05.007

Michalis Faloutsos, Petros Faloutsos, and Christos Faloutsos. 1999. On power-law relationships of the Internet topology. *ACM SIGCOMM computer communication review* 29, 4 (1999), 251–262.

Martín Farach-Colton. [n. d.]. Personal communication from Google employee Number 110.

Jack McKay Fletcher and Thomas Wennekers. 2016. From structure to activity: Using centrality measures to predict neuronal activity. *International journal of neural systems* 28, 02 (2016), 1750013.

Dániel Fogaras, Balázs Rácz, Károly Csalogány, and Tamás Sarlós. 2005. Towards scaling fully personalized PageRank: Algorithms, lower bounds, and experiments. *Internet Mathematics* 2, 3 (2005), 333–358.

Minas Gjoka, Maciej Kurant, Carter T Butts, and Athina Markopoulou. 2010. Walking in Facebook: A case study of unbiased sampling of OSNs. In *2010 Proceedings IEEE Infocom*. Ieee, 1–9.

Christos Gkantsidis, Milena Mihail, and Amin Saberi. 2003. Conductance and congestion in power law graphs. In *Proceedings of the International Conference on Measurements and Modeling of Computer Systems, SIGMETRICS 2003, June 9-14, 2003, San Diego, CA, USA*. 148–159. https://doi.org/10.1145/781027.781046

David F. Gleich. 2015. PageRank beyond the Web. *SIAM Rev.* 57, 3 (2015), 321–363.

Google [n. d.]. Google Online Link Spam Notification Form. https://www.google.com/webmasters/tools/paidlinks?hl=en Accessed: 10/30/2017.

Google 2001. Google Press Center: Fun Facts. https://web.archive.org/web/20010715123343/https://www.google.com/press/funfacts.html. Accessed on 10/28/2017.

Pankaj Gupta, Ashish Goel, Jimmy J. Lin, Aneesh Sharma, Dong Wang, and Reza Zadeh. 2013. WTF: the who to follow service at Twitter. In *22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013*. 505–514. http://dl.acm.org/citation.cfm?id=2488433

Zoltan Gyongyi, Pavel Berkhin, Hector Garcia-Molina, and Jan Pedersen. 2006. Link spam detection based on mass estimation. In *Proceedings of the 32nd international conference on Very large data bases*. VLDB Endowment, 439–450.

Zoltan Gyongyi and Hector Garcia-Molina. 2004. *Web Spam Taxonomy*. Technical Report 2004-25. Stanford InfoLab. http://ilpubs.stanford.edu:8090/646/

Zoltán Gyöngyi and Hector Garcia-Molina. 2005. Link Spam Alliances. In *VLDB*. ACM, 517–528.

Zoltán Gyöngyi, Hector Garcia-Molina, and Jan Pedersen. 2004. Combating web spam with TrustRank. In *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*. VLDB Endowment, 576–587.

Taher H Haveliwala. 2003. Topic-sensitive PageRank: A context-sensitive ranking algorithm for web search. *IEEE transactions on knowledge and data engineering* 15, 4 (2003), 784–796.

John Hopcroft and Daniel Sheldon. 2008. Manipulation-resistant reputations using hitting time. *Internet Mathematics* 5, 1-2 (2008), 71–90.

Gábor Iván and Vince Grolmusz. 2010. When the Web meets the cell: using personalized PageRank for analyzing protein interaction networks. *Bioinformatics* 27, 3 (2010), 405–407.

Glen Jeh and Jennifer Widom. 2003. Scaling personalized web search. In *Proceedings of the Twelfth International World Wide Web Conference, WWW 2003, Budapest, Hungary, May 20-24, 2003*. 271–279. https://doi.org/10.1145/775152.775191

Zhaoyan Jin, Dianxi Shi, Quanyuan Wu, Huining Yan, and Hua Fan. 2012. LBSNRank: personalized PageRank on location-based social networks. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. ACM, 980–987.

Marcos Kiwi and Dieter Mitsche. 2018. Spectral gap of random hyperbolic graphs and related parameters. *Annals of Applied Probability* 28 (2018), 941–989. Issue 2.

Jon M Kleinberg. 1999. Hubs, authorities, and communities. *ACM computing surveys (CSUR)* 31, 4es (1999), 5.

Dimitri Krioukov, Fragkiskos Papadopoulos, Maksim Kitsak, Amin Vahdat, and Marián Bogu ná. 2010. Hyperbolic geometry of complex networks. *Physical Review E* 82 (2010). Issue 3.

Vijay Krishnan and Rashmi Raj. 2006. Web spam detection with anti-trust rank.. In *AIRWeb*, Vol. 6. 37–40.

Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew Tomkins. 2006. Core algorithms in the CLEVER system. *ACM Transactions on Internet Technology (TOIT)* 6, 2 (2006), 131–152.

David Asher Levin, Yuval Peres, and Elizabeth Lee Wilmer. 2009. *Markov chains and mixing times*. American Mathematical Soc.

Brandon K Liu, David C Parkes, and Sven Seuken. 2016. Personalized hitting time for informative trust mechanisms despite Sybils. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 1124–1132.

Milena Mihail, Christos Papadimitriou, and Amin Saberi. 2006. On certain connectivity properties of the internet topology. *J. Comput. System Sci.* 72, 2 (2006), 239–251.

Michael Mitzenmacher and Eli Upfal. 2005. *Probability and Computing*. Cambridge University Press.

Abedelaziz Mohaisen, Aaram Yun, and Yongdae Kim. 2010. Measuring the mixing time of social graphs. In *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*. ACM, 383–389.

Mark EJ Newman and Duncan J Watts. 1999. Renormalization group analysis of the small-world network model. *Physics Letters A* 263, 4-6 (1999), 341–346.

Andrew Y. Ng, Alice X. Zheng, and Michael I. Jordan. 2001a. Link Analysis, Eigenvectors and Stability. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence, IJCAI 2001, Seattle, Washington, USA, August 4-10, 2001*. 903–910.

Andrew Y. Ng, Alice X. Zheng, and Michael I. Jordan. 2001b. Stable Algorithms for Link Analysis. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (New Orleans, Louisiana, USA) *(SIGIR '01)*. ACM, New York, NY, USA, 258–266. https://doi.org/10.1145/383952.384003

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. *The PageRank citation ranking: Bringing order to the web.* Technical Report. Stanford InfoLab.

StatCounter. [n. d.]. Search Engine Market Share. http://gs.statcounter.com/search-engine-market-share. Accessed: 10/30/2017.

Danny Sullivan. [n. d.]. Google Has New Head Of Web Spam But Won't Be The "New Matt Cutts". https://searchengineland.com/google-head-web-spam-221482. Accessed on 5/17/2019.

Johan Ugander, Brian Karrer, Lars Backstrom, and Cameron Marlow. 2011. The anatomy of the Facebook social graph. *arXiv preprint arXiv:1111.4503* (2011).

University of Milan Laboratory of Web Algorithmics. 2007. Web Spam Collections. http://chato.cl/webspam/datasets/ URLs retrieved 04/2020.

Bimal Viswanath, Ansley Post, Krishna P Gummadi, and Alan Mislove. 2011. An analysis of social network-based Sybil defenses. *ACM SIGCOMM Computer Communication Review* 41, 4 (2011), 363–374.

Haifeng Yu. 2011. Sybil Defenses via Social Networks: A Tutorial and Survey. *SIGACT News* 42, 3 (Oct. 2011), 80–101. https://doi.org/10.1145/2034575.2034593

Haifeng Yu, Phillip B. Gibbons, Michael Kaminsky, and Feng Xiao. 2010. SybilLimit: A Near-Optimal Social Network Defense Against Sybil Attacks. *IEEE/ACM Trans. Netw.* 18, 3 (2010), 885–898. https://doi.org/10.1109/TNET.2009.2034047

Haifeng Yu, Michael Kaminsky, Phillip B Gibbons, and Abraham D Flaxman. 2008. SybilGuard: defending against Sybil attacks via social networks. *IEEE/ACM Transactions on networking* 16, 3 (2008), 576–589.

APPENDIX

A DISTORTION OF MIN-PPR

We first set out notation for mixing times from specific initial states. Let $\mathrm{G} \in \mathcal{G}$, let $\hat{\mathbf{r}}$ be a probability distribution on $\mathbf{V}$, let $X \sim \hat{\mathbf{r}}$, and let $\hat{\mathbf{p}}_{i,\hat{\mathbf{r}}}$ be the distribution of the uniform random walk on G at time $i \geq 0$ from initial state X. Then for all $\rho > 0$, we define

$$\tau_{\mathrm{G}}(\rho, \hat{\mathbf{r}}) := \min\Big\{i \geq 0 \colon d_{\mathrm{TV}}\big(\hat{\mathbf{p}}_{i,\hat{\mathbf{r}}}, \mathcal{R}(\mathrm{G})\big) \leq \rho\Big\}.$$

In the special case where $\hat{\mathbf{r}}$ is deterministic, i.e. there exists $x \in \mathbf{V}$ such that $\hat{\mathbf{r}}[x] = 1$, we write $\tau_{\mathrm{G}}(\rho, x) := \tau_{\mathrm{G}}(\rho, \hat{\mathbf{r}})$. We take the default value of ρ to be $1/4$, so that $\tau_{\mathrm{G}} := \tau_{\mathrm{G}}(1/4)$, $\tau_{\mathrm{G}}(x) := \tau_{\mathrm{G}}(1/4, x)$ and $\tau_{\mathrm{G}}(\hat{\mathbf{r}}) := \tau_{\mathrm{G}}(1/4, \hat{\mathbf{r}})$. We also define $\mathrm{emt}_{\mathrm{G}}(\hat{\mathbf{r}}) := \mathbb{E}_{X \sim \hat{\mathbf{r}}}(\tau_{\mathrm{G}}(X))$, where "emt" stands for "expected mixing time". For all positive integers k we will write $[k] = \{1, \ldots, k\}$.

We now state some well-known preliminary lemmas.

Lemma 22 ([Haveliwala, 2003, Eq 10]). *Let* G *be an arbitrary graph, and let* $\hat{\mathbf{r}}$ *be a reset vector on* G. *Then for all* $y \in \mathbf{V}_{\mathrm{G}}$, $\mathcal{R}(\mathrm{G}, \hat{\mathbf{r}}, \varepsilon)[y] = \sum_{x \in \mathbf{V}_{\mathrm{G}}} \hat{\mathbf{r}}[x] \mathcal{R}(\mathrm{G}, x, \varepsilon)[y]$.

Lemma 23. *Let* $\mathrm{G} \in \mathcal{G}$, *let* $\rho \in (0,1)$, *and let* $\hat{\mathbf{r}}$ *be a reset vector on* G. *Then* $\tau_{\mathrm{G}}(\rho, \hat{\mathbf{r}}) \leq \lceil \log_2(1/\rho) \rceil \tau_{\mathrm{G}}(\hat{\mathbf{r}})$.

Proof. This is immediate from [Mitzenmacher and Upfal, 2005, Theorem 11.6], taking P to be the transition matrix of the random walk associated with $\mathcal{R}(\mathrm{G}, \hat{\mathbf{r}}, \varepsilon)$ and c to be $1/4$. □

Lemma 24 ([Mitzenmacher and Upfal, 2005, Theorem 4.4]). *Let* X *be a binomial random variable with mean* μ, *and let* $0 < \eta \leq 1$. *Then*

$$\mathbb{P}(X \geq (1+\eta)\mu) \leq e^{-\eta^2 \mu/3}.$$ □

A.1 All PageRanks are close in total variation distance

In this section, we prove Lemma 9 (which we will use later in the proof of Theorem 10) and bound the total variation distance between any PageRank and the corresponding reference rank.

Lemma 25. *Let* $\mathrm{G} = (\mathbf{V}, \mathbf{E})$ *be a graph in* $\mathcal{G}$, *let* $\hat{\mathbf{r}}$ *be a reset vector on* G, *and let* $0 < \varepsilon < 1$. *Then for all* $y \in \mathbf{V}$ *with* $\mathcal{R}(\mathrm{G})[y] \neq 0$,

$$\frac{\mathcal{R}(\mathrm{G})[y] - \mathcal{R}(\mathrm{G}, \hat{\mathbf{r}}, \varepsilon)[y]}{\mathcal{R}(\mathrm{G})[y]} \leq \varepsilon \tau_{\mathrm{G}}(\hat{\mathbf{r}}) \big(3 - \log_2 \mathcal{R}(\mathrm{G})[y]\big).$$

Proof. Fix $y \in \mathbf{V}$, and for brevity define $\tau := \tau_{\mathrm{G}}(\hat{\mathbf{r}})$ and $I := \lfloor -\log_2 \mathcal{R}(\mathrm{G})[y] \rfloor + 2$. Let $(M(t))_{t \geq 0}$ be the uniform random walk on G with initial state drawn from $\hat{\mathbf{r}}$. Then by Lemma 13, we have

$$\mathcal{R}(\mathrm{G}, \hat{\mathbf{r}}, \varepsilon)[y] = \varepsilon \sum_{t=0}^{\infty} (1-\varepsilon)^t \mathbb{P}\big(M(t) = y\big) \geq \varepsilon \sum_{i=I}^{\infty} \sum_{t=0}^{\tau-1} (1-\varepsilon)^{i\tau+t} \mathbb{P}\big(M(i\tau+t) = y\big).$$

By Lemma 23, for all $i \geq 1$ we have $\tau_{\mathrm{G}}(2^{-i}, \hat{\mathbf{r}}) \leq i\tau$. Thus by the definition of a mixing time, it follows that

$$\mathcal{R}(\mathrm{G}, \hat{\mathbf{r}}, \varepsilon)[y] \geq \varepsilon \sum_{i=I}^{\infty} \sum_{t=0}^{\tau-1} (1-\varepsilon)^{i\tau+t} \big(\mathcal{R}(\mathrm{G})[y] - 2^{-i}\big). \tag{6}$$

We now split (6) into two terms and bound each separately. We have

$$\varepsilon \sum_{i=I}^{\infty} \sum_{t=0}^{\tau-1} (1-\varepsilon)^{i\tau+t} \mathcal{R}(\mathrm{G})[y] = \varepsilon(1-\varepsilon)^{I\tau} \mathcal{R}(\mathrm{G})[y] \sum_{t=0}^{\infty} (1-\varepsilon)^t = (1-\varepsilon)^{I\tau} \mathcal{R}(\mathrm{G})[y] \geq (1-\varepsilon I\tau)\mathcal{R}(\mathrm{G})[y]. \tag{7}$$

Moreover,

$$\begin{aligned}\varepsilon \sum_{i=I}^{\infty} \sum_{t=0}^{\tau-1} (1-\varepsilon)^{i\tau+t} 2^{-i} &= \varepsilon \cdot \sum_{i=I}^{\infty} \big((1-\varepsilon)^{\tau}/2\big)^i \cdot \sum_{t=0}^{\tau-1} (1-\varepsilon)^t \leq \varepsilon \cdot \sum_{i=I}^{\infty} 2^{-i} \cdot \frac{1-(1-\varepsilon)^{\tau}}{\varepsilon} \\ &\leq \varepsilon \cdot 2^{-I+1} \cdot \frac{1-(1-\varepsilon\tau)}{\varepsilon} \leq \varepsilon \cdot \mathcal{R}(\mathrm{G})[y] \cdot \tau.\end{aligned}$$

(Here the last inequality relies on the definition of I.) It follows from (6) and (7) that

$$\mathcal{R}(\mathrm{G}, \hat{\mathbf{r}}, \varepsilon)[y] \geq \big(1 - \varepsilon(I+1)\tau\big)\mathcal{R}(\mathrm{G})[y] \geq \big(1 - \varepsilon\tau(3 - \log_2 \mathcal{R}(\mathrm{G})[y])\big)\mathcal{R}(\mathrm{G})[y],$$

so the result follows. □

LEMMA 9 (RESTATED). *Let $\delta > 0$, let $\varepsilon \in (0,1)$, and let $T(n)$ be any function such that, for all n,*

$$0 \leq T(n) \leq 1/(2\varepsilon(3 + \delta \log_2 n)).$$

Then for all n-vertex graphs $\mathrm{G} \in \mathcal{G}_T$, all reset vectors $\hat{\mathbf{r}}$ on G, and all $y \in \mathrm{V_G}$, we have $\mathrm{Cont}_{\delta}(\mathcal{R}(\mathrm{G}, \hat{\mathbf{r}}, \varepsilon), \mathrm{G}, y) \leq 1 + 2\varepsilon T(n)(3 + \delta \log_2 n)$.

PROOF. Let $y \in \mathrm{V_G}$. If $\mathcal{R}(\mathrm{G})[y] \leq 1/n^{\delta}$, then $\mathrm{Cont}_{\delta}(\mathcal{R}(\mathrm{G}, \hat{\mathbf{r}}, \varepsilon), y) \leq 1$, so suppose $\mathcal{R}(\mathrm{G})[y] > 1/n^{\delta}$; thus we have $\mathrm{Cont}_{\delta}(\mathcal{R}(\mathrm{G}, \hat{\mathbf{r}}, \varepsilon), y) \leq \mathcal{R}(\mathrm{G})[y]/\mathcal{R}(\mathrm{G}, \hat{\mathbf{r}}, \varepsilon)[y]$. By Lemma 25, using the fact that $\mathrm{G} \in \mathcal{G}_T$ (as defined in Section 4), it follows that

$$\mathrm{Cont}_{\delta}(\mathcal{R}(\mathrm{G}, \hat{\mathbf{r}}, \varepsilon), y) \leq \frac{1}{1 - \varepsilon\tau_{\mathrm{G}}(\hat{\mathbf{r}})\big(3 - \log_2 \mathcal{R}(\mathrm{G})[y]\big)} \leq \frac{1}{1 - \varepsilon T(n)(3 + \delta \log_2 n)}.$$

Since $T(n) \leq 1/(2\varepsilon(3 + \delta \log_2 n))$, it follows that $\mathrm{Cont}_{\delta}(\mathcal{R}(\mathrm{G}, \hat{\mathbf{r}}, \varepsilon), y) \leq 1 + 2\varepsilon T(n)(3 + \delta \log_2 n)$ as required. □

For any vector $\hat{\mathbf{x}}\colon V \to [0,1]$ let $H(\hat{\mathbf{x}})$ be its Shannon entropy, namely $H(\hat{\mathbf{x}}) = -\sum_{v \in V} \hat{\mathbf{x}}[v] \log_2 \hat{\mathbf{x}}[v]$. The following theorem bounds the total variation distance between $\mathcal{R}(\mathrm{G}, \hat{\mathbf{r}}, \varepsilon)$ and $\mathcal{R}(\mathrm{G})$ in terms of the Shannon entropy of $\mathcal{R}(\mathrm{G})$.

THEOREM 26. *Let $\varepsilon \in (0,1)$. Then for all n-vertex graphs $\mathrm{G} \in \mathcal{G}$ and all reset vectors $\hat{\mathbf{r}}$ on G, the total variation distance between $\mathcal{R}(\mathrm{G}, \hat{\mathbf{r}}, \varepsilon)$ and $\mathcal{R}(\mathrm{G})$ is at most $\varepsilon\tau_{\mathrm{G}}(\hat{\mathbf{r}})(3 + H(\mathcal{R}(\mathrm{G})))$.*

For any $\mathrm{G} \in \mathcal{G}$, τ_{G} is the mixing time of G from an arbitrary vertex and $H(\mathcal{R}(\mathrm{G})) \leq \log_2 n$, so Theorem 26 implies that on an n-vertex graph $\mathrm{G} \in \mathcal{G}$, any PageRank with reset probability ε has total variation distance at most $\tau_{\mathrm{G}}\varepsilon(3 + \log_2 n)$ to the reference rank of G.

PROOF. We have

$$d_{\mathrm{TV}}\big(\mathcal{R}(\mathrm{G}, \hat{\mathbf{r}}, \varepsilon),\, \mathcal{R}(\mathrm{G})\big) = \frac{1}{2} \sum_{y \in \mathrm{V_G}} \big|\mathcal{R}(\mathrm{G}, \hat{\mathbf{r}}, \varepsilon)[y] - \mathcal{R}(\mathrm{G})[y]\big|. \tag{8}$$

For all $y \in \mathbf{V}_\mathrm{G}$, let

$$\delta_y := \begin{cases} \big(3 - \log_2 \mathcal{R}(\mathrm{G})[y]\big)\mathcal{R}(\mathrm{G})[y] & \text{if } \mathcal{R}(\mathrm{G})[y] \neq 0, \\ 0 & \text{otherwise.} \end{cases}$$

By Lemma 25, each vertex $y \in \mathbf{V}_\mathrm{G}$ with $\mathcal{R}(\mathrm{G}, \hat{\mathbf{r}}, \varepsilon)[y] < \mathcal{R}(\mathrm{G})[y]$ contributes at most $\varepsilon\tau_\mathrm{G}(\hat{\mathbf{r}})\delta_y/2$ to the sum in (8); thus in total such vertices contribute at most $\varepsilon\tau_\mathrm{G}(\hat{\mathbf{r}}) \sum_y \delta_y/2$. Moreover, the total contribution of all vertices $y \in \mathbf{V}_\mathrm{G}$ with $\mathcal{R}(\mathrm{G}, \hat{\mathbf{r}}, \varepsilon)[y] > \mathcal{R}(\mathrm{G})[y]$ is exactly the same. Thus in total,

$$d_{\mathrm{TV}}\big(\mathcal{R}(\mathrm{G}, \hat{\mathbf{r}}, \varepsilon),\, \mathcal{R}(\mathrm{G})\big) \leq \varepsilon\tau_\mathrm{G}(\hat{\mathbf{r}}) \sum_{y \in \mathbf{V}_\mathrm{G}} \delta_y = \varepsilon\tau_\mathrm{G}(\hat{\mathbf{r}})\big(3 + H\big(\mathcal{R}(\mathrm{G})\big)\big),$$

as required. □

A.2 Min-PPR can approximate $\mathcal{R}(\mathrm{G})$ well everywhere

We first prove probabilistic bounds on the relative error of PPR when its center vertex is chosen randomly according to the reference rank of the graph; these are Lemmas 27 and 28.

LEMMA 27. *Let $\varepsilon \in (0,1)$. Let G be a graph in $\mathcal{G}$, and let X be a vertex chosen randomly from $\mathbf{V}_\mathrm{G}$ according to $\mathcal{R}(\mathrm{G})$. Then for all $y \in \mathbf{V}_\mathrm{G}$ with $\mathcal{R}(\mathrm{G})[y] \neq 0$, with probability at least $7/8$,*

$$\mathcal{R}(\mathrm{G}, X, \varepsilon)[y] \leq \Big(1 + 8\varepsilon\,\mathrm{emt}_\mathrm{G}(\mathcal{R}(\mathrm{G}))\big(3 - \log_2 \mathcal{R}(\mathrm{G})[y]\big)\Big)\mathcal{R}(\mathrm{G})[y].$$

PROOF. Let $\delta := 8\varepsilon\,\mathrm{emt}_\mathrm{G}(\mathcal{R}(\mathrm{G}))(3 - \log_2 \mathcal{R}(\mathrm{G})[y])$ for brevity, and let S be the set of all vertices X satisfying $\mathcal{R}(\mathrm{G}, X, \varepsilon)[y] \geq (1+\delta)\mathcal{R}(\mathrm{G})[y]$. We will prove the lemma by showing that $\mathcal{R}(\mathrm{G})[\mathrm{S}] \leq 1/8$.

Observe that $\mathcal{R}(\mathrm{G}, \mathcal{R}(\mathrm{G}), \varepsilon) = \mathcal{R}(\mathrm{G})$, since $\mathcal{R}(\mathrm{G})$ remains invariant under both resetting to $\mathcal{R}(\mathrm{G})$ and uniformly random steps on G. By Lemma 22, applied with $\hat{\mathbf{r}} = \mathcal{R}(\mathrm{G})$, it follows that

$$\mathcal{R}(\mathrm{G})[y] = \sum_{X \in \mathbf{V}_\mathrm{G}} \mathcal{R}(\mathrm{G})[X] \cdot \mathcal{R}(\mathrm{G}, X, \varepsilon)[y]. \tag{9}$$

We split the sum in (9) into two parts. By the definition of S, we have

$$\sum_{X \in \mathrm{S}} \mathcal{R}(\mathrm{G})[X] \cdot \mathcal{R}(\mathrm{G}, X, \varepsilon)[y] \geq \mathcal{R}(\mathrm{G})[\mathrm{S}] \cdot (1+\delta)\mathcal{R}(\mathrm{G})[y]. \tag{10}$$

Moreover, by Lemma 25 (applied with $\hat{\mathbf{r}} = X$) we have

$$\begin{aligned}\sum_{X \in \mathbf{V}_\mathrm{G} \setminus \mathrm{S}} \mathcal{R}(\mathrm{G})[X] \cdot \mathcal{R}(\mathrm{G}, X, \varepsilon)[y] &\geq \sum_{X \in \mathbf{V}_\mathrm{G} \setminus \mathrm{S}} \mathcal{R}(\mathrm{G})[X] \Big(1 - \varepsilon\tau_\mathrm{G}(X)\big(3 - \log_2 \mathcal{R}(\mathrm{G})[y]\big)\Big) \mathcal{R}(\mathrm{G})[y] \\ &\geq \Big(\mathcal{R}(\mathrm{G})[\mathbf{V}_\mathrm{G} \setminus \mathrm{S}] - \varepsilon\,\mathrm{emt}_\mathrm{G}(\mathcal{R}(\mathrm{G}))\big(3 - \log_2 \mathcal{R}(\mathrm{G})[y]\big)\Big)\mathcal{R}(\mathrm{G})[y].\end{aligned}$$

It follows by (9) and (10) that

$$\mathcal{R}(\mathrm{G})[y] \geq \Big(1 + \delta\mathcal{R}(\mathrm{G})[\mathrm{S}] - \varepsilon\,\mathrm{emt}_\mathrm{G}(\mathcal{R}(\mathrm{G}))\big(3 - \log_2 \mathcal{R}(\mathrm{G})[y]\big)\Big)\mathcal{R}(\mathrm{G})[y],$$

so by rearranging we obtain

$$\mathcal{R}(\mathrm{G})[\mathrm{S}] \leq \frac{1}{\delta} \cdot \varepsilon\,\mathrm{emt}_\mathrm{G}(\mathcal{R}(\mathrm{G}))\big(3 - \log_2 \mathcal{R}(\mathrm{G})[y]\big) = \frac{1}{8}$$

as required. □

LEMMA 28. *Let $\varepsilon \in (0,1)$. Let G be a graph in $\mathcal{G}$, and let X be a vertex chosen randomly from $\mathbf{V}_\mathrm{G}$ according to $\mathcal{R}(\mathrm{G})$. Then with probability at least $7/8$, for all $y \in \mathbf{V}_\mathrm{G}$ with $\mathcal{R}(\mathrm{G})[y] \neq 0$,*

$$\mathcal{R}(\mathrm{G}, X, \varepsilon)[y] \geq \Big(1 - 8\varepsilon\,\mathrm{emt}_\mathrm{G}(\mathcal{R}(\mathrm{G}))\big(3 - \log_2 \mathcal{R}(\mathrm{G})[y]\big)\Big)\mathcal{R}(\mathrm{G})[y].$$

Note that while Lemma 27 applies to only a single vertex $y \in \mathbf{V}_\mathrm{G}$, Lemma 28 applies collectively to all such vertices.

PROOF. By Markov's inequality, we have

$$\mathbb{P}\big(\tau_\mathrm{G}(X) \geq 8\mathrm{emt}_\mathrm{G}(\mathcal{R}(\mathrm{G}))\big) \leq 1/8. \tag{11}$$

Suppose $\tau_\mathrm{G}(X) < 8\mathrm{emt}_\mathrm{G}(\mathcal{R}(\mathrm{G}))$. Now consider a vertex $y \in \mathbf{V}_\mathrm{G}$ with $\mathcal{R}(\mathrm{G})[y] \neq 0$. By Lemma 25, applied with $\hat{\mathbf{r}} = X$, we have

$$\frac{\mathcal{R}(\mathrm{G})[y] - \mathcal{R}(\mathrm{G}, X, \varepsilon)[y]}{\mathcal{R}(\mathrm{G})[y]} \leq \varepsilon\tau_\mathrm{G}(X)\big(3 - \log_2 \mathcal{R}(\mathrm{G})[y]\big) < 8\varepsilon\,\mathrm{emt}_\mathrm{G}(\mathcal{R}(\mathrm{G}))\big(3 - \log_2 \mathcal{R}(\mathrm{G})[y]\big).$$

The result therefore follows from (11). □

We are now in a position to prove a general error bound for Min-PPR with a randomly-chosen set, from which Theorem 10 will follow easily.

LEMMA 29. *Let $\varepsilon \in (0,1)$, and let* G *be an n-vertex graph in $\mathcal{G}$. Suppose $\tau_\mathrm{G} \leq 1/(2\varepsilon(3 + H(\mathcal{R}(\mathrm{G}))))$. Let $\hat{\mathbf{p}}$ be a probability distribution on $\mathbf{V}_\mathrm{G}$ with $d_\mathrm{TV}(\hat{\mathbf{p}}, \mathcal{R}(\mathrm{G})) \leq 1/8$. Let $X_1, \dots, X_k \sim \hat{\mathbf{p}}$ be independent and identically distributed, and let $\mathbf{K} = \{X_1, \dots, X_k\}$. Then with probability at least $1 - 4^{-k}n$, for all $y \in \mathbf{V}_\mathrm{G}$:*

(1) if $\mathcal{R}(\mathrm{G})[y] = 0$, then $\mathcal{R}_{\min}(\mathrm{G}, \mathbf{K}, \varepsilon)[y] = 0$, and
(2) if $\mathcal{R}(\mathrm{G})[y] > 0$, then

$$\Big|\mathcal{R}_{\min}(\mathrm{G}, \mathbf{K}, \varepsilon)[y] - \mathcal{R}(\mathrm{G})[y]\Big| \leq 35\tau_\mathrm{G}\varepsilon\Big(1 + H(\mathcal{R}(\mathrm{G})) - \log_2 \mathcal{R}(\mathrm{G})[y]\Big)\mathcal{R}(\mathrm{G})[y]. \tag{12}$$

PROOF. For all $y \in \mathbf{V}_\mathrm{G}$ with $\mathcal{R}(\mathrm{G})[y] = 0$ we say that y is *good* if $\mathcal{R}_{\min}(\mathrm{G}, \mathbf{K}, \varepsilon)[y] = 0$. For all y with $\mathcal{R}(\mathrm{G})[y] > 0$, we say that y is good if (12) holds. We will prove that each vertex y is good with probability at least $1 - 4^{-k}$, splitting the proof into two cases according to $\mathcal{R}(\mathrm{G})[y]$. The result then follows by a union bound over all $y \in \mathbf{V}_\mathrm{G}$.

Case 1: $\mathcal{R}(\mathrm{G})[y] = 0$. In this case, y is good if and only if for some $X_i \in \mathbf{K}$, $\mathcal{R}(\mathrm{G}, X_i, \varepsilon)[y] = 0$. Since $\mathcal{R}(\mathrm{G})[y] = 0$, no vertex with positive reference rank has a path to y in G, so for all vertices x with $\mathcal{R}(\mathrm{G})[x] \neq 0$, we have $\mathcal{R}(\mathrm{G}, x, \varepsilon)[y] = 0$. Since $d_\mathrm{TV}(\hat{\mathbf{p}}, \mathcal{R}(\mathrm{G})) \leq 1/8$, it follows that for all $i \in [k]$, $\mathbb{P}(\mathcal{R}(\mathrm{G}, X_i, \varepsilon)[y] \neq 0) \leq 1/8$. Since $X_1, \dots, X_k$ are independent, it follows that

$$\mathbb{P}(y \text{ is good}) = \mathbb{P}(\mathcal{R}_{\min}(\mathrm{G}, \mathbf{K}, \varepsilon)[y] = 0) \geq 1 - 8^{-k} > 1 - 4^{-k},$$

as claimed.

Case 2: $\mathcal{R}(\mathrm{G})[y] \neq 0$. For brevity, let $\gamma(y) := \tau_\mathrm{G}\varepsilon(3 - \log_2 \mathcal{R}(\mathrm{G})[y])$. We will show that, for all $y \in \mathbf{V}$,

$$\mathcal{R}_{\min}(\mathrm{G}, \mathbf{K}, \varepsilon)[y] \geq (1 - \gamma(y))\mathcal{R}(\mathrm{G})[y] \tag{13}$$

and that, for all $y \in \mathbf{V}$, with probability at least $1 - 4^{-k}$,

$$\mathcal{R}_{\min}(\mathrm{G}, \mathbf{K}, \varepsilon)[y] \leq \Big(1 + 35\tau_\mathrm{G}\varepsilon\big(1 + H(\mathcal{R}(\mathrm{G})) - \log_2 \mathcal{R}(\mathrm{G})[y]\big)\Big)\mathcal{R}(\mathrm{G})[y]. \tag{14}$$

Since $\gamma(y) \leq 35\tau_\mathrm{G}\varepsilon(1 + H(\mathcal{R}(\mathrm{G})) - \log_2 \mathcal{R}(\mathrm{G})[y])$, Equation (13) and (14) imply condition (ii).

Recall that $\mathcal{R}_{\min}(\mathrm{G}, \mathbf{K}, \varepsilon)[y] = \min\{\mathcal{R}(\mathrm{G}, X, \varepsilon)[y] : X \in \mathbf{K}\}/\Upsilon$ where

$$\Upsilon = \sum_{v \in \mathbf{V}} \min\{\mathcal{R}(\mathrm{G}, X, \varepsilon)[v] : X \in \mathbf{K}\}.$$

First, we note that Equation (13) follows from two observations.

- By Lemma 25, for all $i \in [k]$ we have $\mathcal{R}(\mathrm{G})[y] - \mathcal{R}(\mathrm{G}, X_i, \varepsilon)[y] \leq \gamma(y)\mathcal{R}(\mathrm{G})[y]$, and
- $\Upsilon \leq 1$.

Next, we prove that for all $y \in \mathbf{V}$, with probability at least $1 - 4^{-k}$, (14) holds.

Lemma 27 implies that if $X \sim \mathcal{R}(\mathrm{G})$, then with probability at least $7/8$, $\mathcal{R}(\mathrm{G}, X, \varepsilon)[y] - \mathcal{R}(\mathrm{G})[y] \leq 8\gamma(y)\mathcal{R}(\mathrm{G})[y]$. Since $d_{\mathrm{TV}}(\hat{\mathbf{p}}, \mathcal{R}(\mathrm{G})) \leq 1/8$, it follows that for all $i \in [k]$, with probability at least $3/4$, $\mathcal{R}(\mathrm{G}, X_i, \varepsilon)[y] - \mathcal{R}(\mathrm{G})[y] \leq 8\gamma(y)\mathcal{R}(\mathrm{G})[y]$. Thus,

$$\mathbb{P}\Big(\min \big\{ \mathcal{R}(\mathrm{G}, X, \varepsilon)[y] \colon X \in \mathbf{K} \big\} - \mathcal{R}(\mathrm{G})[y] \leq 8\gamma(y)\mathcal{R}(\mathrm{G})[y] \Big) \geq 1 - 4^{-k}. \tag{15}$$

To derive (14) from (15), we next derive a lower bound for Υ. By (13), for all $i \in [k]$, $\mathcal{R}(\mathrm{G}, X_i, \varepsilon)[y] \geq (1 - \gamma(y))\mathcal{R}(\mathrm{G})[y]$. Thus,

$$\Upsilon \geq \sum_{y \in \mathbf{V} \colon \mathcal{R}(\mathrm{G})[y] > 0} (1 - \gamma(y))\mathcal{R}(\mathrm{G})[y] = 1 - \tau_{\mathrm{G}}\varepsilon\big(3 + H(\mathcal{R}(\mathrm{G}))\big).$$

Since $\tau_{\mathrm{G}} < 1/(2\varepsilon(3 + H(\mathcal{R}(\mathrm{G}))))$, it follows that

$$\Upsilon^{-1} \leq 1 + 2\tau_{\mathrm{G}}\varepsilon\big(3 + H(\mathcal{R}(\mathrm{G}))\big).$$

Thus, if the event of (15) occurs for some $y \in \mathbf{V}_{\mathrm{G}}$, then we have

$$\begin{aligned}
\mathcal{R}_{\min}(\mathrm{G}, \mathbf{K}, \varepsilon)[y] &= \min \big\{ \mathcal{R}(\mathrm{G}, X, \varepsilon)[y] \colon X \in \mathbf{K} \big\} \cdot \Upsilon^{-1} \\
&\leq \Big(1 + 9\gamma(y) + 2\tau_{\mathrm{G}}\varepsilon\big(3 + H(\mathcal{R}(\mathrm{G}))\big) \Big) \mathcal{R}(\mathrm{G})[y] \\
&= \Big(1 + \tau_{\mathrm{G}}\varepsilon\big(33 - 9\log_2 \mathcal{R}(\mathrm{G})[y] + 2H(\mathcal{R}(\mathrm{G}))\big) \Big) \mathcal{R}(\mathrm{G})[y].
\end{aligned}$$

□

Since the event of (15) occurs with probability at least $1 - 4^{-k}$, by (15), it follows that (14) holds for y with probability at least $1 - 4^{-k}$ as required.

THEOREM 10 (RESTATED). *Let $\delta \geq 1$. Let $\varepsilon \in (0, 1)$ and let $T(n) \leq 1/(210\varepsilon\delta \log_2 n)$. Let $\mathrm{G} \in \mathcal{G}$ be an n-vertex graph with $n \geq 3$, and suppose that the worst-case mixing time of G is at most $T(n)$. Let $k \geq 1$, let $\hat{\mathbf{r}}$ be an arbitrary reset vector, let $X_1, \ldots, X_k$ be drawn independently from $\mathbf{V}_{\mathrm{G}}$ with probabilities given by $\mathcal{R}(\mathrm{G}, \hat{\mathbf{r}}, \varepsilon)$, and let $\mathbf{K} = \{X_1, \ldots, X_k\}$. Then with probability at least $1 - 4^{-k}n$, the distortion of* T-Min-PPR$_{k,\varepsilon}(\mathrm{G}, \mathbf{K})$ *satisfies $\mathbf{D}_\delta(\mathcal{R}_{\min}(\mathrm{G}, \mathbf{K}, \varepsilon), \mathrm{G}) \leq 1 + 210\varepsilon\delta T(n) \log_2 n$.*

PROOF. Since $T(n) \leq 1/(32\varepsilon\delta \log_2 n)$, $\mathrm{G} \in \mathcal{G}_T$, and $\delta \geq 1$, Theorem 26 implies that

$$d_{\mathrm{TV}}\big(\mathcal{R}(\mathrm{G}, \hat{\mathbf{r}}, \varepsilon), \mathcal{R}(\mathrm{G})\big) \leq \tau_{\mathrm{G}}\varepsilon\big(3 + H(\mathcal{R}(\mathrm{G}))\big) \leq 4\varepsilon T(n) \log_2 n \leq 1/8.$$

Thus by Lemma 29, with probability at least $1 - 4^{-k}n$:

(1) for all $y \in \mathbf{V}_{\mathrm{G}}$ with $\mathcal{R}(\mathrm{G})[y] = 0$, we have $\mathcal{R}_{\min}(\mathrm{G}, \mathbf{K}, \varepsilon)[y] = 0$ also; and
(2) for all $y \in \mathbf{V}_{\mathrm{G}}$ with $\mathcal{R}(\mathrm{G})[y] \neq 0$, (12) holds for y.

Suppose this event occurs, so that (i) and (ii) hold, and let $y \in \mathbf{V}_{\mathrm{G}}$; we will use (i) and (ii) to bound the distortion of $\mathcal{R}_{\min}(\mathrm{G}, \mathbf{K}, \varepsilon)$ on y. We split into cases depending on $\mathcal{R}(\mathrm{G})[y]$.

Case 1: $\mathcal{R}(\mathrm{G})[y] = 0$. By (i), this implies $\mathcal{R}_{\min}(\mathrm{G}, \mathbf{K}, \varepsilon)[y] = 0$, so $\mathbf{D}_\delta(\mathcal{R}_{\min}(\mathrm{G}, \mathbf{K}, \varepsilon), \mathrm{G}) = 1$.

Case 2: $0 < \mathcal{R}(\mathrm{G})[y] < 1/n^\delta$. In this case we have $\mathrm{Cont}_\delta(\mathcal{R}_{\min}(\mathrm{G}, \mathbf{K}, \varepsilon), \mathrm{G}, y) \leq 1$, and

$$\mathrm{Stretch}_\delta\big(\mathcal{R}_{\min}(\mathrm{G}, \mathbf{K}, \varepsilon), \mathrm{G}, y\big) = \max \big\{ 1,\ n^\delta\, \mathcal{R}_{\min}(\mathrm{G}, \mathbf{K}, \varepsilon)[y] \big\}. \tag{16}$$

Since (ii) holds, by (12) we have

$$n^\delta\, \mathcal{R}_{\min}(\mathrm{G}, \mathbf{K}, \varepsilon)[y] \leq n^\delta \Big(1 + 35\tau_{\mathrm{G}}\varepsilon\big(1 + H(\mathcal{R}(\mathrm{G})) - \log_2 \mathcal{R}(\mathrm{G})[y]\big) \Big) \mathcal{R}(\mathrm{G})[y]. \tag{17}$$

The function $x \mapsto -(\log_2 x)x$ is increasing over $x \in (0, 1/3]$, and $\mathcal{R}(\mathbf{G})[y] \le 1/n^\delta \le 1/3$, so $-(\log_2 \mathcal{R}(\mathbf{G})[y])\mathcal{R}(\mathbf{G}[y]) \le (\delta \log_2 n)/n^\delta$. We also have $H(\mathcal{R}(\mathbf{G})) \le \log_2 n$ by a standard bound on Shannon entropy. It follows from (17) that

$$n^\delta \mathcal{R}_{\min}(\mathbf{G}, \mathbf{K}, \varepsilon)[y] \le 1 + 35\tau_{\mathbf{G}}\varepsilon\big(1 + \log_2 n + \delta \log_2 n\big) \le 1 + 105\tau_{\mathbf{G}}\varepsilon\delta \log_2 n.$$

Since $\mathbf{G} \in \mathcal{G}_T$, it follows by (16) that $\mathbf{D}_\delta(\mathcal{R}(\mathbf{G}, \mathbf{K}, \varepsilon), \mathbf{G}, y) \le 1 + 105\varepsilon\delta T(n) \log_2 n$.

Case 3: $\mathcal{R}(\mathbf{G})[y] \ge 1/n^\delta$. In this case, for brevity, let

$$\Gamma = 35\tau_{\mathbf{G}}\varepsilon\big(1 + H(\mathcal{R}(\mathbf{G})) - \log_2 \mathcal{R}(\mathbf{G})[y]\big).$$

Following Case 2, observe that $\Gamma \le 105\tau_{\mathbf{G}}\varepsilon\delta \log_2 n$. Since $\mathbf{G} \in \mathcal{G}_T$, it follows that $\Gamma \le 105\varepsilon\delta T(n) \log_2 n \le 1/2$.

Since (ii) holds, (12) holds for y, so

$$\mathrm{Cont}_\delta\big(\mathcal{R}_{\min}(\mathbf{G}, \mathbf{K}, \varepsilon), \mathbf{G}, y\big) \le \frac{\mathcal{R}(\mathbf{G})[y]}{\mathcal{R}_{\min}(\mathbf{G}, \mathbf{K}, \varepsilon)[y]} \le \frac{1}{1 - \Gamma}.$$

Since $\Gamma \le 1/2$, it follows that

$$\mathrm{Cont}_\delta\big(\mathcal{R}_{\min}(\mathbf{G}, \mathbf{K}, \varepsilon), \mathbf{G}, y\big) \le 1 + 2\Gamma \le 1 + 210\varepsilon\delta T(n) \log_2 n.$$

Moreover, we have

$$\mathrm{Stretch}_\delta\big(\mathcal{R}_{\min}(\mathbf{G}, \mathbf{K}, \varepsilon), \mathbf{G}, y\big) = \max\left\{\frac{1/n^\delta}{\mathcal{R}(\mathbf{G})[y]}, \frac{\mathcal{R}_{\min}(\mathbf{G}, \mathbf{K}, \varepsilon)[y]}{\mathcal{R}(\mathbf{G})[y]}\right\} \le \max\left\{1, \frac{\mathcal{R}_{\min}(\mathbf{G}, \mathbf{K}, \varepsilon)[y]}{\mathcal{R}(\mathbf{G})[y]}\right\},$$

where

$$\frac{\mathcal{R}_{\min}(\mathbf{G}, \mathbf{K}, \varepsilon)[y]}{\mathcal{R}(\mathbf{G})[y]} \le 1 + \Gamma < 1 + 210\varepsilon T(n) \log_2 n.$$

The result therefore follows. □

We now prove a version of Lemma 29 which gives error bounds in terms of $\mathrm{emt}_{\mathbf{G}}(\mathcal{R}(\mathbf{G}))$ rather than in terms of $\tau_{\mathbf{G}}$. To ensure a reasonable lower bound, we will need to discard some of our centers; we do this in a graph-agnostic way, to facilitate turning the lemma into a ranking algorithm.

LEMMA 30. *Let $\delta > 1$. Let $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ be an n-vertex graph in $\mathcal{G}$ where $n \ge 2$ is sufficiently large that $n^{\delta-1} \ge 32$. Let $\varepsilon \in (0, 1)$ be sufficiently small that $8\varepsilon\, \mathrm{emt}_{\mathbf{G}}(\mathcal{R}(\mathbf{G})) \le 1/(20\delta \log_2(n))$. Let γ be any real number satisfying $8\varepsilon\, \mathrm{emt}_{\mathbf{G}}(\mathcal{R}(\mathbf{G})) \le \gamma \le 1/(20\delta \log_2(n))$.*

Let $\hat{\mathbf{p}}$ be a probability distribution on $\mathbf{V}$ with $d_{\mathrm{TV}}(\hat{\mathbf{p}}, \mathcal{R}(\mathbf{G})) \le 1/8$. Let $X_1, \dots, X_{2k+1} \sim \hat{\mathbf{p}}$ be independent and identically distributed. For each $y \in \mathbf{V}$, let $M(y)$ be the median of $\{\mathcal{R}(\mathbf{G}, X_i, \varepsilon)[y] : i \in [2k+1]\}$. For each $i \in [2k+1]$, let

$$\xi_i := \max\left\{\frac{M(y) - \mathcal{R}(\mathbf{G}, X_i, \varepsilon)[y]}{M(y)} : y \in \mathbf{V},\, M(y) \ge 1/(2n^\delta)\right\}.$$

Let $f : [2k+1] \to [2k+1]$ be an arbitrary permutation such that $\xi_{f(1)} \le \dots \le \xi_{f(2k+1)}$, and let $\mathbf{K} = \{X_{f(i)} : i \in [k+1]\}$. Then with probability at least $1 - (n+1)e^{-k/6}$, for all $y \in \mathbf{V}$,

$$\mathcal{R}_{\min}(\mathbf{G}, \mathbf{K}, \varepsilon)[y] \le \begin{cases} \big(1 + 2\gamma(3 - \log_2 \mathcal{R}(\mathbf{G})[y]) + 12\delta\gamma \log_2 n + 2n^{1-\delta}\big)\mathcal{R}(\mathbf{G})[y] & \textit{if } \mathcal{R}(\mathbf{G})[y] \ne 0, \\ 0 & \textit{otherwise.} \end{cases} \tag{18}$$

Moreover, for all $y \in \mathbf{V}$ with $\mathcal{R}(\mathbf{G})[y] \ge n^{-\delta}$,

$$\mathcal{R}_{\min}(\mathbf{G}, \mathbf{K}, \varepsilon)[y] \ge (1 - 6\delta\gamma \log_2 n)\mathcal{R}(\mathbf{G})[y]. \tag{19}$$

PROOF. We use the following definitions throughout the proof. For all $y \in \mathbf{V}$ with $\mathcal{R}(\mathbf{G})[y] \neq 0$, we let $L(y) = 3 - \log_2 \mathcal{R}(\mathbf{G})[y]$. Let $\mathbf{A}$ be the set of all indices $i \in [2k+1]$ such that for some $y \in \mathbf{V}$ with $\mathcal{R}(\mathbf{G})[y] \geq 1/(4n^\delta)$, $\mathcal{R}(\mathbf{G}, X_i, \varepsilon)[y] < (1 - \gamma L(y))\mathcal{R}(\mathbf{G})[y]$. Finally, for all $y \in \mathbf{V}$, let

$$\mathbf{B}_y = \begin{cases} \big\{i \in [2k+1] : \mathcal{R}(\mathbf{G}, X_i, \varepsilon)[y] > (1 + \gamma L(y))\mathcal{R}(\mathbf{G})[y]\big\} & \text{if } \mathcal{R}(\mathbf{G})[y] \neq 0, \\ \big\{i \in [2k+1] : \mathcal{R}(\mathbf{G}, X_i, \varepsilon)[y] \neq 0\big\} & \text{otherwise.} \end{cases}$$

We will first bound $|\mathbf{A}|$ and each $|\mathbf{B}_y|$ above with high probability, then use these bounds to prove the lemma.

Claim 1: $\mathbb{P}\big(|\mathbf{A}| \leq k \wedge (\forall y \in \mathbf{V}, |\mathbf{B}_y| \leq k)\big) \geq 1 - (n+1)e^{-k/6}$.

Proof of Claim 1: We first bound $|\mathbf{A}|$. Since $\gamma \geq 8\varepsilon\, \mathrm{emt}_{\mathbf{G}}(\mathcal{R}(\mathbf{G}))$, by Lemma 28, if $X \sim \mathcal{R}(\mathbf{G})$ then

$$\mathbb{P}\Big(\mathcal{R}(\mathbf{G}, X, \varepsilon)[y] < (1 - \gamma L(y))\mathcal{R}(\mathbf{G})[y] \text{ for some } y \in \mathbf{V} \text{ with } \mathcal{R}(\mathbf{G})[y] > 0\Big) \leq 1/8.$$

Since $d_{\mathrm{TV}}(\mathcal{R}(\mathbf{G}), \hat{\mathbf{p}}) \leq 1/8$, it follows that for any $i \in [2k+1]$, we have $\mathbb{P}(i \in \mathbf{A}) \leq 1/4$, so $|\mathbf{A}|$ is a binomial variable with mean at most $(2k+1)/4$. Thus by a Chernoff bound (Lemma 24 applied with $\eta = 1$), it follows that

$$\mathbb{P}(|\mathbf{A}| \geq k+1) \leq \mathbb{P}\Big(|\mathbf{A}| \geq 2 \cdot \frac{2k+1}{4}\Big) \leq e^{-(2k+1)/12} \leq e^{-k/6}. \tag{20}$$

We now bound each $|\mathbf{B}_y|$.

- First, suppose $y \in \mathbf{V}$ with $\mathcal{R}(\mathbf{G})[y] = 0$.
  As in the proof of Lemma 29, no vertex with positive reference rank has a path to y in $\mathbf{G}$, so if $X \sim \mathcal{R}(\mathbf{G})$ then $\mathbb{P}(\mathcal{R}(\mathbf{G}, X, \varepsilon)[y] \neq 0) = 0$. Since $d_{\mathrm{TV}}(\hat{\mathbf{p}}, \mathcal{R}(\mathbf{G})) \leq 1/8$, for each $i \in [2k+1]$ it follows that $\mathbb{P}(\mathcal{R}(\mathbf{G}, X_i, \varepsilon)[y] \neq 0) \leq 1/8$. Thus $|\mathbf{B}_\mathrm{y}|$ is a binomial variable with mean at most $(2k+1)/8$, so by Lemma 24 we have
  $$\mathbb{P}(|\mathbf{B}_y| \geq k+1) \leq e^{-k/6} \text{ whenever } \mathcal{R}(\mathbf{G})[y] = 0. \tag{21}$$
- Next suppose $y \in \mathbf{V}$ with $\mathcal{R}(\mathbf{G})[y] \neq 0$.
  Since $\gamma \geq 8\varepsilon\, \mathrm{emt}_{\mathbf{G}}(\mathcal{R}(\mathbf{G}))$, by Lemma 27, if $X \sim \mathcal{R}(\mathbf{G})$ then
  $$\mathbb{P}\Big(\mathcal{R}(\mathbf{G}, X, \varepsilon)[y] > (1 + \gamma L(y))\mathcal{R}(\mathbf{G})[y]\Big) \leq 1/8.$$
  Since $d_{\mathrm{TV}}(\mathcal{R}(\mathbf{G}), \hat{\mathbf{p}}) \leq 1/8$, it follows that $\mathbb{P}(i \in \mathbf{B}_y) \leq 1/4$. Thus once again by Lemma 24, we have
  $$\mathbb{P}(|\mathbf{B}_y| \geq k+1) \leq e^{-k/6} \text{ whenever } \mathcal{R}(\mathbf{G})[y] \neq 0. \tag{22}$$

Combining (20)–(22) with a union bound, the claim follows. **(End of Proof of Claim 1.)**

The lemma follows from Claim 1, together with the following claim.

Claim 2: If $|\mathbf{A}| \leq k$ and, for all $y \in \mathbf{V}$, $|\mathbf{B}_y| \leq k$, then:

- for all $y \in \mathbf{V}$, inequality (18) holds, and
- for all $y \in \mathbf{V}$ with $\mathcal{R}(\mathbf{G})[y] \geq n^{-\delta}$, inequality (19) holds.

Proof of Claim 2: From now on we will assume that $|\mathbf{A}| \leq k$ and, for all $y \in \mathbf{V}$, that $|\mathbf{B}_y| \leq k$.

Equations (18) and (19) give upper and lower bounds on $\mathcal{R}_{\min}(\mathbf{G}, \mathbf{K}, \varepsilon)[y]$ for certain $y \in \mathbf{V}$. Recall that $\mathcal{R}_{\min}(\mathbf{G}, \mathbf{K}, \varepsilon)[y] = \min\{\mathcal{R}(\mathbf{G}, X, \varepsilon)[y] : X \in \mathbf{K}\}/\Upsilon$ where

$$\Upsilon = \sum_{v \in \mathbf{V}} \min\big\{\mathcal{R}(\mathbf{G}, X, \varepsilon)[v] : X \in \mathbf{K}\big\}.$$

We first observe that for all $y \in \mathbf{V}$ and all $(k+1)$-element subsets $\mathbf{I}$ of $[2k+1]$, since $|\mathbf{B}_y| \le k$,

$$\min\left\{\mathcal{R}(\mathbf{G}, X_i, \varepsilon)[y] : i \in \mathbf{I}\right\} \le \begin{cases} (1+\gamma L(y))\mathcal{R}(\mathbf{G})[y] & \text{if } \mathcal{R}(\mathbf{G})[y] \ne 0, \\ 0 & \text{otherwise.} \end{cases} \tag{23}$$

It will be useful to upper-bound the right-hand-side of (23) as follows. Suppose $0 < \mathcal{R}(\mathbf{G})[y] \le r \le 1/3$. In this case,

$$(1+\gamma L(y))\mathcal{R}(\mathbf{G})[y] \le (1+\gamma(3-\log_2 r))r. \tag{24}$$

To see this, note that the function $f(x) = x\log_2(1/x)$ is increasing for $x \in (0, 1/3]$, so

$$\begin{aligned}(1+\gamma(3-\log_2 \mathcal{R}(\mathbf{G})[y]))\mathcal{R}(\mathbf{G})[y] &= 1 + 3\gamma\mathcal{R}(\mathbf{G})[y] + \gamma\mathcal{R}(\mathbf{G})[y]\log_2(1/\mathcal{R}(\mathbf{G})[y]) \\ &\le 1 + 3\gamma r + \gamma r\log_2(1/r) \\ &= \big(1+\gamma(3-\log_2 r)\big)r.\end{aligned}$$

The proof of the claim will proceed as follows. In Step 1, we prove an upper bound on ξ_i for all $i \notin \mathbf{A}$, and hence (as we will see) for all i such that $X_i \in \mathbf{K}$. In Step 2, we will turn this into a lower bound on $\min\{\mathcal{R}(\mathbf{G}, X, \varepsilon)[y] : X \in \mathbf{K}\}$ whenever $\mathcal{R}(\mathbf{G})[y] \ge 1/n^\delta$. This will suffice in Step 3 to bound the normalizing factor Υ below. In Step 4, we use this, together with (23) and the lower bound of Step 2, to prove (18) and (19).

Step 1: Consider any $v \in \mathbf{V}$ with $M(v) \ge 1/(2n^\delta)$. Each time we apply (23) in this step we will take $\mathbf{I}$ to be a set of $k+1$ indices i with $\mathcal{R}(\mathbf{G}, X_i, \varepsilon)[v]$ as large as possible.

We first prove $\mathcal{R}(\mathbf{G})[v] > 0$. For contradiction, suppose $\mathcal{R}(\mathbf{G})[v] = 0$. By (23) we have $M(v) = 0$, contradicting our choice of v.

Given that $\mathcal{R}(\mathbf{G})[v] > 0$ we now prove $\mathcal{R}(\mathbf{G})[v] \ge 1/(4n^\delta)$. For contradiction, suppose $\mathcal{R}(\mathbf{G})[v] < 1/(4n^\delta)$. By (23) we have $M(v) \le (1+\gamma L(y))\mathcal{R}(\mathbf{G})[v]$. Using (24) with $r = 1/(4n^\delta)$, we have

$$M(v) \le \big(1+\gamma(3+\log_2(4n^\delta))\big)\frac{1}{4n^\delta} = \big(1+\gamma(5+\delta\log_2(n))\big)\frac{1}{4n^\delta}.$$

Since (from the statement) $\delta\log_2 n \ge 5$, this is at most $(1+2\gamma\delta\log_2(n))\frac{1}{4n^\delta}$. Using the upper bound on γ from the statement of the lemma, this is less than $1/(2n^\delta)$. This contradicts our choice of v, so we must have $\mathcal{R}(\mathbf{G})[v] \ge 1/(4n^\delta)$ as claimed.

Now consider any $i \in [2k+1] \setminus \mathbf{A}$. It follows from the definition of $\mathbf{A}$ that

$$\mathcal{R}(\mathbf{G}, X_i, \varepsilon)[v] \ge (1-\gamma L(v))\mathcal{R}(\mathbf{G})[v]. \tag{25}$$

Once again, by (23) we have $M(v) \le (1+\gamma L(v))\mathcal{R}(\mathbf{G})[v]$, so

$$\mathcal{R}(\mathbf{G})[v] \ge (1+\gamma L(v))^{-1}M(v) \ge (1-\gamma L(v))M(v).$$

It follows from (25) that $\mathcal{R}(\mathbf{G}, X_i, \varepsilon)[v] \ge (1-2\gamma L(v))M(v)$. Since $\mathcal{R}(\mathbf{G})[v] \ge 1/(4n^\delta)$, it follows using the definition of $L(v)$ that

$$\mathcal{R}(\mathbf{G}, X_i, \varepsilon)[v] \ge \big(1-2\gamma(3+\log_2(4n^\delta))\big)M(v) = \big(1-2\gamma(5+\delta\log_2 n)\big)M(v).$$

Since (from the statement) $\delta\log_2 n \ge 5$, this is at least $(1-4\gamma\delta\log_2(n))M(v)$. We have shown that for every $v \in \mathbf{V}$ with $M(v) \ge 1/(2n^\delta)$, we have $\mathcal{R}(\mathbf{G}, X_i, \varepsilon)[v] \ge (1-4\gamma\delta\log_2(n))M(v)$.

From the definition of ξ_i, there is some $v \in \mathbf{V}$ with $M(v) \ge 1/(2n^\delta)$ such that $\xi_i = 1 - \mathcal{R}(\mathbf{G}, X_i, \varepsilon)[v]/M(v)$. Thus $\xi_i \le 4\gamma\delta\log_2 n$ for all $i \notin \mathbf{A}$. Since $|\mathbf{A}| \le k$ and $|\mathbf{K}| = 2k+1$, it follows that $\xi_{f(1)}, \dots, \xi_{f(k+1)} \le 4\gamma\delta\log_2 n$.

Step 2: Consider any $y \in \mathbf{V}$ with $\mathcal{R}(\mathbf{G})[y] \ge 1/n^\delta$. By the definition of $\mathbf{A}$, since $|\mathbf{A}| \le k$, we have $M(y) \ge (1-\gamma L(y))\mathcal{R}(\mathbf{G})[y]$. Since $\delta\log_2 n \ge 3$,

$$L(y) = 3 + \log_2(1/\mathcal{R}(\mathbf{G})[y]) \le 3 + \delta\log_2 n \le 2\delta\log_2 n,$$

so we conclude that $M(y) \geq (1-\gamma L(y))\mathcal{R}(\mathbf{G})[y] \geq (1-2\gamma\delta\log_2 n)\mathcal{R}(\mathbf{G})[y]$. Since the upper bound on γ in the statement guarantees that $\gamma\delta\log_2 n \leq 1/4$, we have

$$M(y) \geq (1-2\gamma\delta\log_2 n)\mathcal{R}(\mathbf{G})[y] \geq \mathcal{R}(\mathbf{G})[y]/2 \geq 1/(2n^{\delta}). \tag{26}$$

We proved in Step 1 that $\xi_{f(1)},\dots,\xi_{f(k+1)} \leq 4\gamma\delta\log_2 n$. Thus, by the definition of $\mathbf{K}$ in the statement of the lemma, for each $X_i \in \mathbf{K}$, $\xi_i \leq 4\gamma\delta\log_2 n$. Since (26) guarantees that $M(y) \geq 1/(2n^{\delta})$, the definition of ξ_i ensures that $\xi_i \geq 1 - \mathcal{R}(\mathbf{G}, X_i, \varepsilon)[y]/M(y)$ so

$$\min\{\mathcal{R}(\mathbf{G}, X, \varepsilon)[y] : X \in \mathbf{K}\} \geq (1-4\gamma\delta\log_2 n)M(y).$$

Using the first inequality in (26), we have

$$\min\left\{\mathcal{R}(\mathbf{G}, X, \varepsilon)[y] : X \in \mathbf{K}\right\} \geq (1-6\gamma\delta\log_2 n)\mathcal{R}(\mathbf{G})[y]. \tag{27}$$

Note that we have proved (27) for all $y \in \mathbf{V}$ with $\mathcal{R}(\mathbf{G})[y] \geq 1/n^{\delta}$.

Step 3: We next bound the normalizing factor Υ. By (27), we have

$$\Upsilon = \sum_{v\in\mathbf{V}} \min\left\{\mathcal{R}(\mathbf{G}, X, \varepsilon)[v] : X \in \mathbf{K}\right\} \geq \sum_{\substack{v\in\mathbf{V} \\ \mathcal{R}(\mathbf{G})[v]\geq 1/n^{\delta}}} (1-6\gamma\delta\log_2 n)\mathcal{R}(\mathbf{G})[v]$$
$$\geq 1-6\gamma\delta\log_2 n - n(1/n^{\delta}) = 1-6\gamma\delta\log_2 n - n^{1-\delta}.$$

The upper bound on γ in the statement of the lemma guarantees that $6\gamma\delta\log_2 n \leq 6/20$. Since $n^{\delta-1} \geq 5$, $n^{1-\delta} \leq 4/20$ so the sum of these is at most $1/2$. It follows that

$$\frac{1}{\Upsilon} \leq 1 + 12\gamma\delta\log_2 n + 2n^{1-\delta}. \tag{28}$$

Step 4: We are now ready to prove (18) and (19). For all $y \in \mathbf{V}$ with $\mathcal{R}(\mathbf{G})[y] \geq n^{-\delta}$, since $\mathcal{R}_{\min}(\mathbf{G}, \mathbf{K}, \varepsilon)[y] = \min\{\mathcal{R}(\mathbf{G}, X, \varepsilon)[y] : X \in \mathbf{K}\}/\Upsilon$ and $\Upsilon \leq 1$, (27) implies that (19) holds. By (23) applied with $\mathbf{I} = \{f(1),\dots,f(k+1)\}$, for all $y \in \mathbf{V}$ with $\mathcal{R}(\mathbf{G})[y] = 0$, we have $\mathcal{R}_{\min}(\mathcal{R}(\mathbf{G}), \mathbf{K}, \varepsilon)[y] = 0$ (as required by (18)). Finally, again by (23), for all $y \in \mathbf{V}$ with $\mathcal{R}(\mathbf{G})[y] \neq 0$, we have

$$\min\left\{\mathcal{R}(\mathbf{G}, X, \varepsilon)[y] : X \in \mathbf{K}\right\} \leq (1+\gamma L(y))\mathcal{R}(\mathbf{G})[y].$$

The upper bound on γ in the statement of the lemma ensures that $12\gamma\delta\log_2 n \leq 3/5$. Since $n^{\delta-1} \geq 5$, $2n^{1-\delta} \leq 2/5$. Thus, their sum (in the right-hand side of (28)) is at most 1. It follows by (28) that

$$\mathcal{R}_{\min}(\mathcal{R}(\mathbf{G}), \mathbf{K}, \varepsilon) \leq \left(1 + 2\gamma L(y) + 12\delta\gamma\log_2 n + 2n^{1-\delta}\right)\mathcal{R}(\mathbf{G})[y].$$

Hence (18) follows. **(End of Proof of Claim 2.)** □

We now turn Lemma 30 into a ranking algorithm T-Min-PPR$_{\gamma,\delta,k,\varepsilon}$ as follows. The parameter k is a positive integer and the other parameters are real numbers satisfying $\gamma, \varepsilon \in (0,1)$ and $\delta > 1$. Given a graph $\mathbf{G}$ and a trusted set $\mathbf{T} \subseteq \mathbf{V}_{\mathbf{G}}$, T-Min-PPR$_{\gamma,\delta,k,\varepsilon}(\mathbf{G}, \mathbf{T})$ chooses a set $\mathbf{K} \subseteq \mathbf{T}$ of size $\min\{2k-1, |\mathbf{T}|\}$. Then the algorithm calculates, for each $y \in \mathbf{V}$, the median $M(y)$ of $\{\mathcal{R}(\mathbf{G}, c, \varepsilon) : c \in \mathbf{K}\}$, and the observed divergences

$$\xi_c := \max\left\{\frac{M(y) - \mathcal{R}(\mathbf{G}, c, \varepsilon)[y]}{M(y)} : y \in \mathbf{V},\, M(y) \geq \frac{1}{2n^{\delta}}\right\}.$$

The algorithm then forms a set $\mathbf{K}' \subseteq \mathbf{K}$ by discarding the $k-1$ vertices in $\mathbf{K}$ with the highest values of ξ_c, then taking a coherent subset that is as large as possible. Finally, the algorithm outputs $\mathcal{R}_{\min}(\mathbf{G}, \mathbf{K}', \varepsilon)$.

Essentially, T-Min-PPR$_{\gamma,\delta,k,\varepsilon}$ is similar to T-Min-PPR$_{k,\varepsilon}$, except that, rather than choosing k centers arbitrarily from $\mathbf{T}$, the algorithm chooses them according to which of their PPRs agrees most closely with the median PPR. As the following theorem shows, when $\mathbf{G} \in \mathcal{G}$, $n^{-\delta}$ acts as a significance threshold, γ acts as an accuracy parameter for vertices with reference rank above this threshold, and the algorithm gives good results when ε is small relative to $n^{-\delta}$, γ, and $\tau_{\mathbf{G}}$.

THEOREM 31. *Let k be a positive integer. Let $\delta > 1$. Let $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ be an n-vertex graph in $\mathcal{G}$ where $n \geq 2$ is sufficiently large that $n^{\delta-1} \geq 32$. Let $\varepsilon \in (0,1)$ be sufficiently small that $8\varepsilon\, \mathrm{emt}_{\mathbf{G}}(\mathcal{R}(\mathbf{G})) \leq 1/(20\delta \log_2(n))$. Let γ be any real number satisfying $8\varepsilon\, \mathrm{emt}_{\mathbf{G}}(\mathcal{R}(\mathbf{G})) \leq \gamma \leq 1/(20(10 + \delta \log_2(n)))$. Let $\hat{\mathbf{p}}$ be a probability distribution on $\mathbf{V}$ with $d_{\mathrm{TV}}(\hat{\mathbf{p}}, \mathcal{R}(\mathbf{G})) \leq 1/8$. Let $X_1, \ldots, X_{2k-1} \sim \hat{\mathbf{p}}$ be independent and identically distributed. Let $\mathbf{T} = \{X_1, \ldots, X_{2k-1}\}$. Then with probability at least $1 - (n+1)e^{-(k-1)/6}$,*

$$\mathbf{D}_\delta\big(\text{T-Min-PPR}_{\gamma,\delta,k,\varepsilon}(\mathbf{G}, \mathbf{T}), \mathbf{G}\big) \leq 1 + 40\gamma\delta \log_2 n + 2n^{1-\delta}. \tag{29}$$

Theorem 31 says that when $\mathbf{G}$ is ergodic, and the parameters are chosen appropriately, then T-Min-PPR$_{\gamma,\delta,k,\varepsilon}$ performs essentially at least as well as the simpler algorithm T-Min-PPR$_{k,\varepsilon}$. The key difference is that T-Min-PPR$_{k,\varepsilon}$ requires an upper bound on the worst-case mixing time $\tau_{\mathbf{G}}$, while T-Min-PPR$_{\gamma,\delta,k,\varepsilon}$ requires an analogous upper bound on the (potentially much smaller) average-case mixing time $\mathrm{emt}_{\mathbf{G}}(\mathcal{R}(\mathbf{G}))$.

PROOF. Recall from the definition that in computing a ranking function, T-Min-PPR$_{\gamma,\delta,k,\varepsilon}(\mathbf{G}, \mathbf{T})$ first chooses a subset of $\mathbf{T}$ of size $\min\{2k-1, |\mathbf{T}|\}$; since $|\mathbf{T}| = 2k-1$, this must be $\mathbf{T}$ itself. It then chooses a subset $\mathbf{K} \subseteq \mathbf{T}$ by discarding $k-1$ vertices as in the statement of Lemma 30. (Note that in this proof, we will take the k of Lemma 30 to be our present $k-1$.) Since $\mathbf{G} \in \mathcal{G}$, any non-empty subset of $\mathbf{V}$ is coherent; thus $\mathbf{K}$ is coherent, so we have T-Min-PPR$_{\gamma,\delta,k,\varepsilon}(\mathbf{G}, \mathbf{T}) = \mathcal{R}_{\min}(\mathbf{G}, \mathbf{K}, \varepsilon)$.

By Lemma 30, with probability at least $1 - (n+1)e^{-(k-1)/6}$, (18) holds for all $y \in \mathbf{V}$ and (19) holds for all $y \in \mathbf{V}$ with $\mathcal{R}(\mathbf{G})[y] \geq n^{-\delta}$. Suppose this event occurs; then we will show that (29) holds. To bound the distortion at each vertex $y \in \mathbf{V}$, we split into cases depending on $\mathcal{R}(\mathbf{G})[y]$.

Case 1: $\mathcal{R}(\mathbf{G})[y] = 0$. By (18), this implies $\mathcal{R}_{\min}(\mathbf{G}, \mathbf{K}, \varepsilon)[y] = 0$. Thus $\mathcal{R}_{\min}(\mathbf{G}, \mathbf{K}, \varepsilon)$, and hence T-Min-PPR$_{\gamma,\delta,k,\varepsilon}(\mathbf{G}, \mathbf{T})$, has distortion exactly 1 at y.

Case 2: $0 < \mathcal{R}(\mathbf{G})[y] < 1/n^\delta$. In this case, $\mathcal{R}_{\min}(\mathbf{G}, \mathbf{K}, \varepsilon)$ has contraction at most 1 at y, and

$$\mathrm{Stretch}_\delta\big(\mathcal{R}_{\min}(\mathbf{G}, \mathbf{K}, \varepsilon), \mathbf{G}, y\big) = \max\big\{1, n^\delta \mathcal{R}_{\min}(\mathbf{G}, \mathbf{K}, \varepsilon)[y]\big\}.$$

Since (18) holds for y, we have

$$n^\delta \mathcal{R}_{\min}(\mathbf{G}, \mathbf{K}, \varepsilon)[y] \leq n^\delta\big(1 + 2\gamma(3 - \log_2 \mathcal{R}(\mathbf{G})[y]) + 12\gamma\delta \log_2 n + 2n^{1-\delta}\big)\mathcal{R}(\mathbf{G}).$$

The function $x \mapsto -(\log_2 x)x$ is increasing over $x \in (0, 1/3]$, and $\mathcal{R}(\mathbf{G})[y] \leq 1/n^\delta \leq 1/3$, so $-(\log_2 \mathcal{R}(\mathbf{G})[y])\mathcal{R}(\mathbf{G})[y] \leq (\delta \log_2 n)/n^\delta$. It follows that

$$n^\delta \mathcal{R}_{\min}(\mathbf{G}, \mathbf{K}, \varepsilon)[y] \leq 1 + 2\gamma(3 + \delta \log_2 n) + 12\gamma\delta \log_2 n + 2n^{1-\delta} \leq 1 + 20\gamma\delta \log_2 n + 2n^{1-\delta}.$$

It follows that $\mathbf{D}_\delta(\mathcal{R}_{\min}(\mathbf{G}, \mathbf{K}, \varepsilon), \mathbf{G}, y) \leq 1 + 40\gamma\delta \log_2 n + 2n^{1-\delta}$, as required.

Case 3: $\mathcal{R}(\mathbf{G})[y] \geq 1/n^\delta$. Since (19) holds for all $y \in \mathbf{V}$ with $\mathcal{R}(\mathbf{G})[y] \geq n^{-\delta}$, for all such y we have

$$\mathrm{Cont}_\delta\big(\mathcal{R}_{\min}(\mathbf{G}, \mathbf{K}, \varepsilon), \mathbf{G}, y\big) \leq \frac{\mathcal{R}(\mathbf{G})[y]}{\mathcal{R}_{\min}(\mathbf{G}, \mathbf{K}, \varepsilon)[y]} \leq \frac{1}{1 - 6\delta\gamma \log_2 n}.$$

Since $\gamma \leq 1/(20(10 + \delta \log_2 n))$, we have $6\delta\gamma \log_2 n \leq 1/2$; hence

$$\mathrm{Cont}_\delta\big(\mathcal{R}_{\min}(\mathbf{G}, \mathbf{K}, \varepsilon), \mathbf{G}, y\big) \leq 1 + 12\delta\gamma \log_2 n < 1 + 40\gamma\delta \log_2 n + 2n^{1-\delta}. \tag{30}$$

Moreover, we have

$$\mathrm{Stretch}_\delta\big(\mathcal{R}_{\min}(\mathbf{G}, \mathbf{K}, \varepsilon), \mathbf{G}, y\big) = \max\left\{\frac{1/n^\delta}{\mathcal{R}(\mathbf{G})[y]}, \frac{\mathcal{R}_{\min}(\mathbf{G}, \mathbf{K}, \varepsilon)[y]}{\mathcal{R}(\mathbf{G})[y]}\right\} \leq \max\left\{1, \frac{\mathcal{R}_{\min}(\mathbf{G}, \mathbf{K}, \varepsilon)[y]}{\mathcal{R}(\mathbf{G})[y]}\right\},$$

where (18) implies that

$$\frac{\mathcal{R}_{\min}(\mathbf{G},\mathbf{K},\varepsilon)[y]}{\mathcal{R}(\mathbf{G})[y]} \leq 1 + 40\gamma\delta\log_2 n + 2n^{1-\delta}.$$

The result therefore follows from (30). □

B SPAM RESISTANCE OF MIN-PPR

In this section, we bound the spam resistance of Min-PPR$_{k,\varepsilon}$, proving Theorem 11. We first prove a technical lemma which bounds the effect a spammer can have on PPR.

LEMMA 32. *Let* $\mathbf{G} = (\mathbf{V}, \mathbf{E_G})$ *be a graph. Let* $\mathbf{P} \subseteq \mathbf{V}$*, and let* $\mathbf{H} = (\mathbf{V} \cup \mathbf{S}, \mathbf{E_H}) \in \mathbf{G_P}$ *(where* $\mathbf{S}$ *and* $\mathbf{V}$ *are disjoint). Let* $\hat{\mathbf{r}}$ *be a reset vector on* $\mathbf{G}$ *with* $\hat{\mathbf{r}}[v] = 0$ *for all* $v \in \mathbf{P}$*, and let* $\hat{\mathbf{r}}'$ *be the corresponding reset vector on* $\mathbf{H}$ *with* $\hat{\mathbf{r}}'[v] = \hat{\mathbf{r}}[v]$ *for all* $v \in \mathbf{V}$ *and* $\hat{\mathbf{r}}'[v] = 0$ *for all* $v \in \mathbf{S}$*. Then for all* $0 < \varepsilon < 1$ *and all* $\mathbf{A} \subseteq \mathbf{V}$*, we have*

$$\mathcal{R}(\mathbf{H}, \hat{\mathbf{r}}', \varepsilon)[\mathbf{A} \cup \mathbf{S}] \leq \mathcal{R}(\mathbf{G}, \hat{\mathbf{r}}, \varepsilon)[\mathbf{A} \setminus \mathbf{P}] + \varepsilon^{-1}\mathcal{R}(\mathbf{G}, \hat{\mathbf{r}}, \varepsilon)[\mathbf{P}]$$

Remark: By taking $\mathbf{A} = \mathbf{P}$, Lemma 32 implies that T-PPR$_\varepsilon$ is ε-spam resistant (see the proof of Lemma 3 below). In fact, the same holds for any ranking algorithm which carries out some form of PageRank that only resets to trusted vertices. However, Lemma 32 does not immediately imply ε-spam resistance for Min-PPR$_{k,\varepsilon}$, since the reset vector on $\mathbf{H}$ does not in general match the reset vector on $\mathbf{G}$ as required by the lemma. Nevertheless, we will use the full strength of the lemma (with a more subtle choice of $\mathbf{A}$) to demonstrate spam resistance of Min-PPR$_{k,\varepsilon}$ in the proof of Theorem 11 below.

PROOF. Let $(X_i)_{i\geq 0}$ be a uniform random walk on $\mathbf{G}$ with initial state drawn from $\hat{\mathbf{r}}$, and let $(Y_i)_{i\geq 0}$ be a uniform random walk on $\mathbf{H}$ with initial state drawn from $\hat{\mathbf{r}}'$. By Lemma 13 applied to $\mathbf{H}$ and $\hat{\mathbf{r}}'$, we have

$$\mathcal{R}(\mathbf{H}, \hat{\mathbf{r}}', \varepsilon)[\mathbf{A} \cup \mathbf{S}] = \varepsilon \sum_{i=0}^{\infty} (1-\varepsilon)^i \mathbb{P}(Y_i \in \mathbf{A} \cup \mathbf{S}). \tag{31}$$

If $Y_i \in \mathbf{A} \cup \mathbf{S}$, then either Y passed through $\mathbf{P}$ at some time in $[0, i]$ or it did not, and in the latter case we must have $Y_i \in \mathbf{A} \setminus \mathbf{P}$. For all $j \geq 0$, let $\mathcal{E}_j$ be the event that $Y_0, \ldots, Y_{j-1} \notin \mathbf{P}$. Then by (31) and a union bound, we have

$$\mathcal{R}(\mathbf{H}, \hat{\mathbf{r}}', \varepsilon)[\mathbf{A} \cup \mathbf{S}] \leq \varepsilon \sum_{i=0}^{\infty} (1-\varepsilon)^i \Big(\sum_{j=0}^{i} \mathbb{P}\big(Y_j \in \mathbf{P} \text{ and } \mathcal{E}_j\big) + \mathbb{P}\big(Y_i \in \mathbf{A} \setminus \mathbf{P} \text{ and } \mathcal{E}_i\big)\Big)$$

Before hitting $\mathbf{P}$, Y behaves exactly like X; more formally, the two chains can be coupled until this stopping time. Thus

$$\begin{aligned}\mathcal{R}(\mathbf{H}, \hat{\mathbf{r}}', \varepsilon)[\mathbf{A} \cup \mathbf{S}] &\leq \varepsilon \sum_{i=0}^{\infty} (1-\varepsilon)^i \Big(\sum_{j=0}^{i} \mathbb{P}\big(X_j \in \mathbf{P} \text{ and } \mathcal{E}_j\big) + \mathbb{P}\big(X_i \in \mathbf{A} \setminus \mathbf{P} \text{ and } \mathcal{E}_i\big)\Big) \\ &\leq \varepsilon \sum_{i=0}^{\infty} (1-\varepsilon)^i \Big(\sum_{j=0}^{i} \mathbb{P}\big(X_j \in \mathbf{P}\big) + \mathbb{P}\big(X_i \in \mathbf{A} \setminus \mathbf{P}\big)\Big) \end{aligned} \tag{32}$$

We now simplify each part of the right-hand side of (32). By Lemma 13 applied to $\mathbf{G}$ and $\hat{\mathbf{r}}$, we have

$$\varepsilon \sum_{i=0}^{\infty} (1-\varepsilon)^i \mathbb{P}(X_i \in \mathbf{A} \setminus \mathbf{P}) = \mathcal{R}(\mathbf{G}, \hat{\mathbf{r}}, \varepsilon)[\mathbf{A} \setminus \mathbf{P}]. \tag{33}$$

Moreover, by reordering the summation, we have

$$\varepsilon\sum_{i=0}^{\infty}(1-\varepsilon)^i\sum_{j=0}^{i}\mathbb{P}\big(X_j\in\mathbf{P}\big)=\varepsilon\sum_{j=0}^{\infty}\mathbb{P}(X_j\in\mathbf{P})\sum_{i=j}^{\infty}(1-\varepsilon)^i=\sum_{j=0}^{\infty}\mathbb{P}(X_j\in\mathbf{P})(1-\varepsilon)^j$$

By Lemma 13 applied to $\mathbf{G}$ and $\hat{\mathbf{r}}$, it follows that

$$\varepsilon\sum_{i=0}^{\infty}(1-\varepsilon)^i\sum_{j=0}^{i}\mathbb{P}\big(X_j\in\mathbf{P}\big)=\frac{1}{\varepsilon}\mathcal{R}(\mathbf{G},\hat{\mathbf{r}},\varepsilon)[\mathbf{P}].$$

The result therefore follows from (32) and (33). □

We now use Lemma 32 to prove our spam resistance results for T-PPR and T-Min-PPR.

LEMMA 3 (RESTATED). *For all $\varepsilon\in(0,1)$, T-PPR$_\varepsilon$ is ε-spam resistant on all graph classes. The cost function that establishes this spam resistance is the T-PPR$_\varepsilon$ itself, normalized over untrusted nodes.*

PROOF. Let $\mathbf{G}=(\mathbf{V},\mathbf{E_G})$ be a graph, let $\mathbf{T_G}\subseteq\mathbf{V}$ be non-empty, let $c\in\mathbf{T_G}$ be the center chosen by T-PPR$_\varepsilon$, and let $\varepsilon\in(0,1)$. We define our cost function by $C[v]:=\llbracket\mathcal{R}(\mathbf{G},c,\varepsilon)\rrbracket[v]$. (Recall that we write $\llbracket\mathbf{x}\rrbracket:=\mathbf{x}/||\mathbf{x}||$; here the normalization is necessary since C is only defined on $\mathbf{V}\setminus\mathbf{T_G}$.) Let $\mathbf{P}\subseteq\mathbf{V}\setminus\mathbf{T_G}$, and let $\mathbf{H}=(\mathbf{V}\cup\mathbf{S},\mathbf{E_H})\in\mathbf{G_P}$, where $\mathbf{V}$ and $\mathbf{S}$ are disjoint. Then T-PPR$_\varepsilon(\mathbf{H},\mathbf{T_G})=\mathcal{R}(\mathbf{H},c,\varepsilon)$, so by Lemma 32 (applied with $\mathbf{A}=\mathbf{P}$ and $\hat{\mathbf{r}}=c$) we have

$$\text{T-PPR}_\varepsilon(\mathbf{H},\mathbf{T_G})[\mathbf{S}\cup\mathbf{P}]\le\varepsilon^{-1}\mathcal{R}(\mathbf{G},c,\varepsilon)[\mathbf{P}]\le C[\mathbf{P}]/\varepsilon,$$

as required by the definition of spam resistance. □

THEOREM 11 (RESTATED). *For any $\varepsilon\in(0,1)$ and any positive integer k,* T-Min-PPR$_{k,\varepsilon}$ *is $(\varepsilon/3k)$-spam resistant on n-vertex graphs in $\mathcal{G}$ with worst-case mixing time at most* $1/(3\varepsilon(3+\log_2 n))$. *A cost function that establishes this spam resistance is the average of the cost functions of the component T-PPRs.*

PROOF. Let $\mathbf{G}=(\mathbf{V},\mathbf{E_G})\in\mathcal{G}$ have n vertices and satisfy $\tau_\mathbf{G}\le 1/(3\varepsilon(2+\log_2 n))$, and let $\mathbf{T_G}\subseteq\mathbf{V}$ be non-empty. Let $\mathbf{K}$ be the subset of $\mathbf{T_G}$ chosen by T-Min-PPR$_{k,\varepsilon}$; thus $|\mathbf{K}|\le k$. (Recall that $\mathbf{K}$ depends only on $\mathbf{T_G}$, not on $\mathbf{H}$.) We will then define our cost function by

$$C[v]:=\left\llbracket\sum_{c\in\mathbf{K}}\mathcal{R}(\mathbf{G},c,\varepsilon)\Big)\right\rrbracket[v]\text{ for all }v\in\mathbf{V}\setminus\mathbf{T_G}.$$

In order to bound the cost function later, it helps to do the normalization explicitly. Let $\gamma(v)=\frac{1}{|\mathbf{K}|}\sum_{c\in\mathbf{K}}\mathcal{R}(\mathbf{G},c,\varepsilon)[v]$ and $Z=\sum_{v\in\mathbf{V}\setminus\mathbf{T_G}}\gamma(v)$. For $v\in\mathbf{V}\setminus\mathbf{T_G}$, $C[v]=\gamma(v)/Z$. We will use the fact (which we will prove shortly) that

$$C[\mathbf{P}]\ge\frac{1}{|\mathbf{K}|}\sum_{c\in\mathbf{K}}\mathcal{R}(\mathbf{G},c,\varepsilon)[\mathbf{P}].\tag{34}$$

To establish (34), note that

$$C[\mathbf{P}]=\frac{1}{Z}\sum_{v\in\mathbf{P}}\gamma(v)=\frac{1}{|\mathbf{K}|Z}\sum_{v\in\mathbf{P}}\sum_{c\in\mathbf{K}}\mathcal{R}(\mathbf{G},c,\varepsilon)[v]=\frac{1}{|\mathbf{K}|Z}\sum_{c\in\mathbf{K}}\mathcal{R}(\mathbf{G},c,\varepsilon)[\mathbf{P}].$$

So (34) follows from $Z\le 1$, which follows from the following calculation.

$$Z=\sum_{v\in\mathbf{V}\setminus\mathbf{T_G}}\gamma(v)\le\sum_{v\in\mathbf{V}}\gamma(v)=\frac{1}{|\mathbf{K}|}\sum_{c\in\mathbf{K}}\mathcal{R}(\mathbf{G},c,\varepsilon)[\mathbf{V}]=\frac{|\mathbf{K}|}{|\mathbf{K}|}=1$$

Let $\mathbf{P} \subseteq \mathbf{V} \setminus \mathbf{T_G}$, let $\mathbf{H} = (\mathbf{V} \cup \mathbf{S}, \mathbf{E_H}) \in \mathbf{G_P}$, where $\mathbf{V}$ and $\mathbf{S}$ are disjoint, and let $\mathbf{K}'$ be the maximal coherent subset of $\mathbf{K}$ chosen by T-Min-PPR$_{k,\varepsilon}$. Then to prove the result, it suffices to show that

$$\mathcal{R}_{\min}(\mathbf{H}, \mathbf{K}', \varepsilon)[\mathbf{S} \cup \mathbf{P}] \leq \frac{3k}{\varepsilon} C[\mathbf{P}]. \tag{35}$$

For convenience, we define

$$M(\mathbf{G})[v] := \min\{\mathcal{R}(\mathbf{G}, c, \varepsilon)[v] \colon c \in \mathbf{K}\} \text{ for all } v \in \mathbf{V},$$
$$M(\mathbf{H})[v] := \min\{\mathcal{R}(\mathbf{H}, c, \varepsilon)[v] \colon c \in \mathbf{K}\} \text{ for all } v \in \mathbf{V} \cup \mathbf{S}.$$

(Note that $M(\mathbf{H})[v]$ is defined in terms of $\mathbf{K}$, not $\mathbf{K}'$.) We now split into two cases depending on the value of $M(\mathbf{H})[\mathbf{V}]$.

Case 1: $M(\mathbf{H})[\mathbf{V}] \leq 1/3$. In this case, we will argue that the spammer has had to pay so high a price that the behavior of T-Min-PPR$_{k,\varepsilon}$ is irrelevant. We first use the assumption that $\tau_\mathbf{G} \leq 1/(3\varepsilon(3 + \log_2 n))$ to prove that $M(\mathbf{G})[\mathbf{V}] \geq 2/3$. By Lemma 25, for all $c \in \mathbf{T}$ and all $v \in \mathbf{V}$ with $\mathcal{R}(\mathbf{G})[v] \neq 0$,

$$\mathcal{R}(\mathbf{G}, c, \varepsilon)[v] \geq \Big(1 - \varepsilon\tau_\mathbf{G}\big(3 - \log_2 \mathcal{R}(\mathbf{G})[v]\big)\Big)\mathcal{R}(\mathbf{G})[v].$$

Summing over all such $v \in \mathbf{V}$ and using the fact that $G \in \mathcal{G}_{1/3\varepsilon(3+\log_2 n)}$, we obtain

$$M(\mathbf{G})[\mathbf{V}] \geq 1 - \varepsilon\tau_\mathbf{G}\big(3 + H(\mathcal{R}(\mathbf{G}))\big) \geq 1 - \varepsilon\tau_\mathbf{G}(3 + \log_2 n) \geq 2/3. \tag{36}$$

Now, for each $v \in \mathbf{V} \cup \mathbf{S}$, let $\chi(v)$ be an arbitrary vertex $c \in \mathbf{K}$ such that $M(\mathbf{H})[v] = \mathcal{R}(\mathbf{H}, c, \varepsilon)[v]$. For all $c \in \mathbf{K}$, let $\mathbf{B_c} := \{v \in \mathbf{V} \colon \chi(v) = c\}$. Then we have

$$\begin{aligned}
M(\mathbf{G})[\mathbf{V}] - M(\mathbf{H})[\mathbf{V}] &\leq \sum_{c \in \mathbf{K}} \big(\mathcal{R}(\mathbf{G}, c, \varepsilon)[\mathbf{B_c}] - \mathcal{R}(\mathbf{H}, c, \varepsilon)[\mathbf{B_c}]\big) \\
&= \sum_{c \in \mathbf{K}} \Big(\big(1 - \mathcal{R}(\mathbf{G}, c, \varepsilon)[\mathbf{V} \setminus \mathbf{B_c}]\big) - \big(1 - \mathcal{R}(\mathbf{H}, c, \varepsilon)[(\mathbf{S} \cup \mathbf{V}) \setminus \mathbf{B_c}]\big)\Big) \\
&\leq \sum_{c \in \mathbf{K}} \Big(\mathcal{R}(\mathbf{H}, c, \varepsilon)[(\mathbf{V} \setminus \mathbf{B_c}) \cup \mathbf{S}] - \mathcal{R}(\mathbf{G}, c, \varepsilon)[(\mathbf{V} \setminus \mathbf{B_c}) \setminus \mathbf{P}]\Big).
\end{aligned}$$

Using Lemma 32, applied to each term in the sum with $\mathbf{A} = \mathbf{V} \setminus \mathbf{B_c}$, it follows that

$$M(\mathbf{G})[\mathbf{V}] - M(\mathbf{H})[\mathbf{V}] \leq \sum_{c \in \mathbf{K}} \frac{1}{\varepsilon} \mathcal{R}(\mathbf{G}, c, \varepsilon)[\mathbf{P}].$$

Hence by (34), we have

$$M(\mathbf{G})[\mathbf{V}] - M(\mathbf{H})[\mathbf{V}] \leq \frac{|\mathbf{K}|}{\varepsilon} C[\mathbf{P}] \leq \frac{k}{\varepsilon} C[\mathbf{P}].$$

Recall that $M(\mathbf{H})[\mathbf{V}] \leq 1/3$ by hypothesis, and $M(\mathbf{G})[\mathbf{V}] \geq 2/3$ by (36). Thus

$$C[\mathbf{P}] \geq \frac{\varepsilon}{3k} \geq \frac{\varepsilon}{3k} \mathcal{R}_{\min}(\mathbf{H}, \mathbf{K}', \varepsilon)[\mathbf{S} \cup \mathbf{P}],$$

so (35) holds as required.

Case 2: $M(\mathbf{H})[\mathbf{V}] > 1/3$. In this case, there exists $v \in \mathbf{V}$ such that $M(\mathbf{H})[v] > 0$, so $\mathbf{K}$ is coherent. Hence $\mathbf{K}' = \mathbf{K}$. By the definition of $\mathcal{R}_{\min}$, it follows that

$$\begin{aligned}
\mathcal{R}_{\min}(\mathbf{H}, \mathbf{K}', \varepsilon)[\mathbf{S} \cup \mathbf{P}] = \frac{M(\mathbf{H})[\mathbf{S} \cup \mathbf{P}]}{M(\mathbf{H})[\mathbf{S} \cup \mathbf{V}]} &\leq \frac{M(\mathbf{H})[\mathbf{S} \cup \mathbf{P}]}{M(\mathbf{H})[\mathbf{V}]} \leq 3M(\mathbf{H})[\mathbf{S} \cup \mathbf{P}] \\
&\leq 3 \sum_{c \in \mathbf{K}} \mathcal{R}(\mathbf{H}, c, \varepsilon)[\mathbf{S} \cup \mathbf{P}].
\end{aligned}$$

By Lemma 32, applied with $\mathbf{A} = \mathbf{P}$, and again using (34) it follows that

$$\mathcal{R}_{\min}(\mathbf{H}, \mathbf{K}', \varepsilon)[\mathbf{S} \cup \mathbf{P}] \leq \frac{3}{\varepsilon} \sum_{c \in \mathbf{K}} \mathcal{R}(\mathbf{G}, c, \varepsilon)[\mathbf{P}] \leq \frac{3|\mathbf{K}|}{\varepsilon} C[\mathbf{P}] \leq \frac{3k}{\varepsilon} C[\mathbf{P}].$$

Thus (35) holds in all cases, as required. □

C EXPERIMENTAL EVALUATION OF PAGERANKS

In this section, we evaluate the performance of four ranking functions on an actual web graph: Min-PPR, Median-PPR, UPR, and Mean-PPR, the componentwise mean of PPRs. The graph is WEBSPAM-UK2007 [University of Milan Laboratory of Web Algorithmics, 2007], which consists of 114,529 hosts, 5,709 of which have been labeled as *trusted* and 344 as *spam*. We evaluate Min-PPR, Median-PPR and Mean-PPR on constituent PPRs based on picking centers at random from among the trusted sites. We consider an empirical evaluation of spam resistance and distortion on a graph that has already been spammed, so we must modify the definitions to suit our needs, as described below. For each ranking function we evaluate:

Spam Resistance. We cannot evaluate spam resistance directly using the definition, since we do not know what the pre-spam graph is, but we can evaluate the ***spam rank***, or the sum of the ranks assigned to all nodes labeled spam. In addition to the total ranking for all spammers, we evaluate spam ranks for each ranking function by ordering each node in the graph according to rank and then counting the number of spam sites in each decile. This view of the spam rankings ensures that spam-resistance benefits are well distributed across the ranking vector, rather than potentially reflecting a difference for only a few bad sites.

Ideally, an optimal spam-resistant ranking function would not inadvertently penalize non-spam sites in an effort to assign low rank to spammers. In order to assess this potential trade-off, we measure the ***trusted rank***, or the sum of the ranks of trusted sites, in addition to the spam rank. Analogously to spam rank, we also compute the distribution of trusted sites in each decile of rank.

Distortion. Since our definition of distortion applies only to strongly connected graphs, for our experiments, we first restrict to the largest strongly connected component of the graph and renormalize each ranking vector. We then compute distortion for UPR, Min-PPR, Mean-PPR and Median-PPR using the stationary distribution on the largest strongly connected component as the reference ranking. We use a value of $\delta = 2$, and we also tested δ values of 2.5, 3 and 4, but the differences were minimal.

Stability. Since the ranking functions in this paper rely on selecting a set of trusted centers, we would like to examine how sensitive the results are to an accurate assessment of which sites are indeed trusted. Ideally, the spam resistance of a ranking would not swing wildly if a "wrong" center made its way into the trusted set. Thus, we evaluate the *stability* of the ranking functions by measuring the changes in rank produced by using maximally bad sites as centers. Specifically, we compute the multiplicative changes in spam rank and trusted rank for a set of experiments run entirely with known spam sites as centers, relative to the same measurements on a set of trusted centers. This metric is not a feature of our theoretical analysis, but provides interesting additional insight in the experimental results. In particular, stability will help distinguish between Min-PPR and Median-PPR, which both perform well under the other three metrics.

Results. We find:

Spam Rank and Trusted Rank: We find that UPR has spam rank that is 192% higher than Mean-, Median- and Min-PPR, on average, as expected from the theoretical analysis. We find that

Mean-, Median- and Min-PPR have similar trusted rank once k is at least 3, and that Min-PPR has 26% higher spam rank than Mean- and Median-PPR on average.

Distortion: We find that UPR has distortion that is on average 47 times higher than Min- and Median-PPR, and the distortion of Mean-PPR is 15 times higher.

Stability: Although Min-PPR and Median-PPR are roughly comparable in terms of spam rank, trusted rank and distortion, when we compute both functions using randomly selected spam nodes as centers, Min- PPR shows greater stability. The Median-PPR spam ranks jump by an average of 68% and trusted ranks fall by 19%, while for Min-PPR, the spam rank and trusted rank change by 20% and -2%, respectively. We posit that the instability of Median-PPR is due to the weak closure of the median operator and present evidence to support this hypothesis.

Furthermore, we find that in all cases, picking three centers is enough to achieve most of the benefits of these ranking functions, so the computational cost is only a factor of three higher than that of UPR.

We conclude that Min-PPR has an attractive combination of spam rank, trusted rank, distortion, stability and computational cost, in accordance with the theoretical results in this paper.

C.1 Experimental Setup

C.1.1 The Web Graph. We employed the web graph dataset WEBSPAM-UK2007 [University of Milan Laboratory of Web Algorithmics, 2007], which is based on a crawl of web pages in the .UK domain that were labeled by volunteers for research purposes. The dataset contains 114,529 hosts with directed edges between them denoting webpage outlinks. Some of the hosts are labeled as *spam* (344 hosts), or *normal* (5,709 hosts, which we refer to as *trusted*). The largest strongly connected component of the graph has 59,160 nodes, with 134 spam nodes and 3,167 trusted nodes.

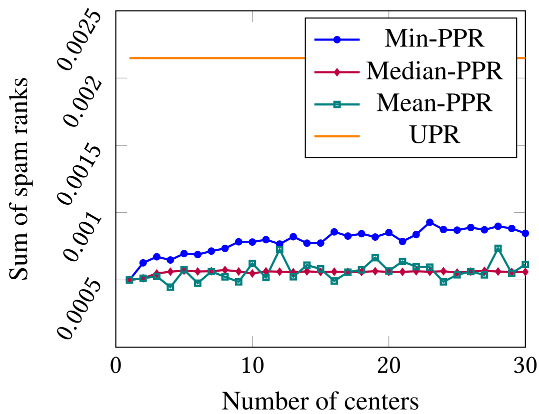
C.1.2 Parameters of the Experiments. We compute UPR, Mean-PPR, Min-PPR and Median-PPR for various choices of the reset probability ε, and where applicable, various numbers of trusted centers k and subsets of the trusted sites used as centers.

We ran 50 independent trials for each $k \in \{1, 2, \ldots, 30\}$ and each $\varepsilon \in \{0.15, 0.05, 0.01, 0.001\}$. For the value of ε, 0.15 was selected based on the historical precedent from [Page et al., 1999] of using this reset probability for PageRank, and the smaller values were chosen based on our theoretical results implying that smaller ε yields better outcomes. Of these four, we find that $\varepsilon = 0.15$ and $\varepsilon = 0.01$ were representative, so these are the ones reported.

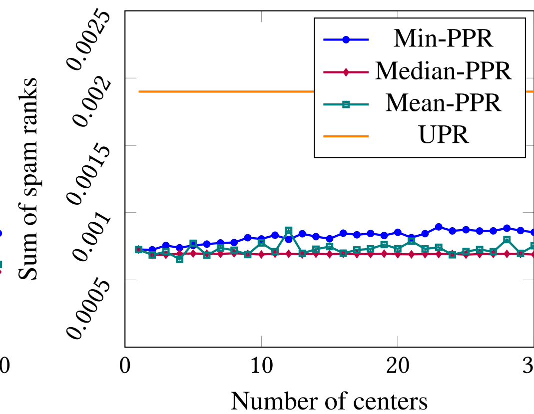
For a given k, the trusted centers were selected from the trusted sites in the largest strongly connected component. Sites were selected independently at random, according to a distribution of weights corresponding to their relative ranks in the stationary distribution. This reference rank-weighted selection is consistent with Theorem 10.

For Min-PPR, Median-PPR and Mean-PPR, each point in Figures 2, 4 and 6 reflects the average rank/distortion over all 50 trials for each value of k. In Figure 3 and Figure 5, each decile bar reflects the average count over all trials and all values of k. UPR does not select specific centers, and therefore does not vary with k.

To run the experiments, we first computed a uniform PageRank ranking vector and PPRs for all centers for each value of the reset probability ε using the python networkx library *pagerank*() function. We selected a tolerance for convergence of 10^{-12} after testing various tolerance levels, and determining that smaller tolerances did not significantly impact the results. For rankings with multiple trusted centers, the pointwise averages, minimums and medians are then computed from the PPRs to produce the Mean-, Min- and Median-PPR ranking vectors, respectively. Min- and Median-PPR are normalized so that the total rank of the graph is 1.0.



(a) $\varepsilon = 0.15$



(b) $\varepsilon = 0.01$

Fig. 2. Sum of ranks of all labeled spam sites for each ranking function; lower is better. The x-axis indicates the number of trusted centers, k, used for Min-PPR, Median-PPR and Mean-PPR. Each point represents the average sum of spam ranks over 50 independent trials. Min-PPR, Median-PPR, and Mean-PPR exhibit substantially lower spam ranks than UPR across all k.

For the stability experiment, we repeated the spam resistance experiments using centers selected entirely from the set of spam sites. We ran 10 independent trials for $k = 3$ centers. In Section C.2, we observe that the spam resistance and distortion benefits for Min- and Median-PPR are achievable with as low as three centers, which informed the selection of k used here. As with the trusted sites, the spam centers were randomly selected with probabilities determined by their ranks in the stationary distribution. We then compared the spam resistance metrics – total spam ranks and total trusted ranks – to the same metrics for $k = 3$ trusted centers, as described in Section C.2.4.

C.2 Experimental Results

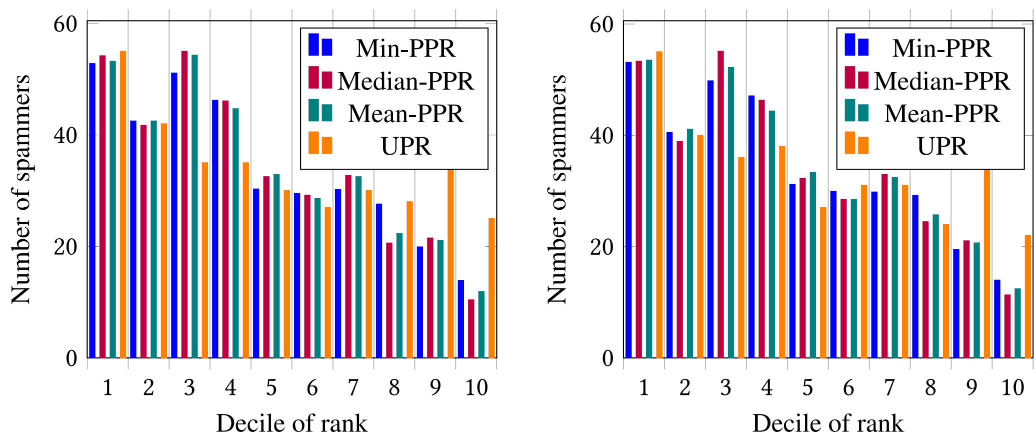
In this section, we discuss the results of the experiments described in Section C. The results are also presented for each metric and $\varepsilon = 0.15, 0.01$ in Figure 2 through Figure 7.

C.2.1 Spam Rank Results. Based on our theoretical results, we would expect each of Mean-, Min- and Median-PPR to perform better than UPR in terms of spam resistance. Indeed, as shown in Figure 2, all three of these ranking functions assign significantly lower rank to the spam sites than UPR across all k.

For $\varepsilon = 0.15$, the total rank assigned to spammers by UPR is 0.0021, while the average spammer ranks over all trials and k were 0.0008 for Min-PPR, 0.0006 for Median-PPR, and 0.0004 for Mean-PPR, for an overall reduction of between 63% and 74%. Similarly, for $\varepsilon = 0.01$, the total spam rank for UPR is 0.0019, while the total spam ranks average 0.0008 for Min-PPR, and 0.0007 for Median- and Mean-PPR, for a reduction of between 57% and 62%.

Of the three PPR ranking functions, Min-PPR shows the highest spam rank, which is also expected. Averaged over all k and over $\varepsilon \in \{0.15, 0.01\}$, the total spam rank for Min-PPR is 24% higher than for Mean-PPR. Median-PPR performs the best overall, with 3% lower average spam rank than Mean-PPR, with an even better improvement of 15% seen for $\varepsilon = 0.01$.

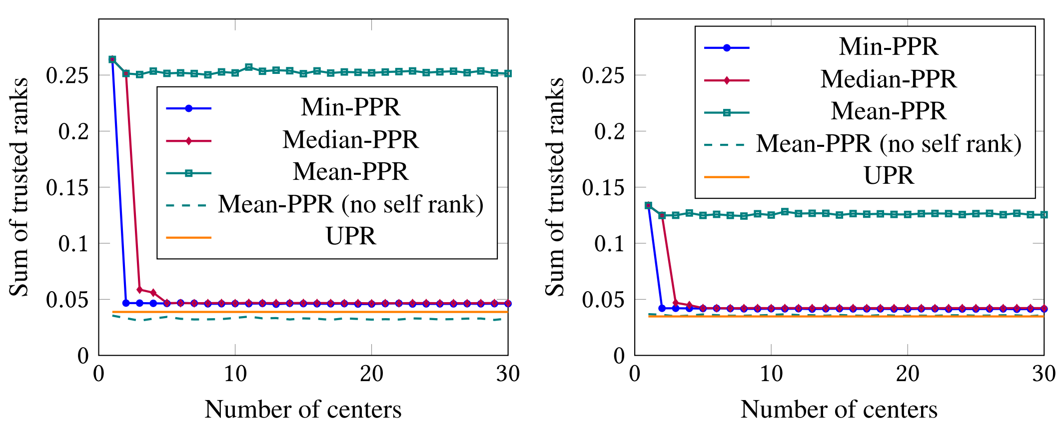
Figure 3 demonstrates that the improvements in spam resistance over UPR are seen across the ranking vector, and not just in total spam rank. When we order all nodes of the graph by rank and



(a) $\varepsilon = 0.15$

(b) $\varepsilon = 0.01$

Fig. 3. Count of spam sites in each decile of rank for each ranking function. The x-axis indicates the decile when each node in the web graph is ordered according to rank, so 10 is the highest decile and a lower count in high deciles is better. Each bar represents the average count of spammers over each $k \in \{1, 2, \ldots 30\}$ and 50 independent trials for each k. Min-PPR, Median-PPR and Mean-PPR exhibit a general shift of spammers to deciles 3-5 from deciles 8-10 ($\varepsilon = 0.15$) or 9-10 ($\varepsilon = 0.01$) compared to UPR.

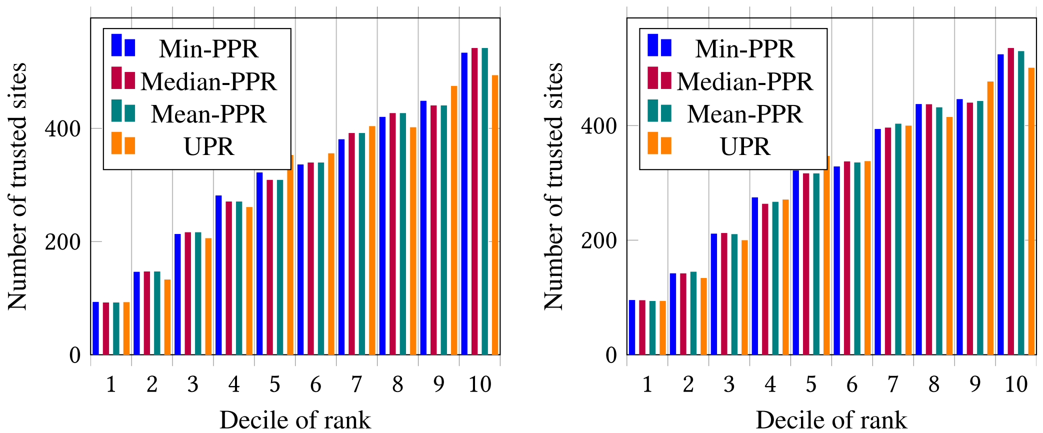


(a) $\varepsilon = 0.15$

(b) $\varepsilon = 0.01$

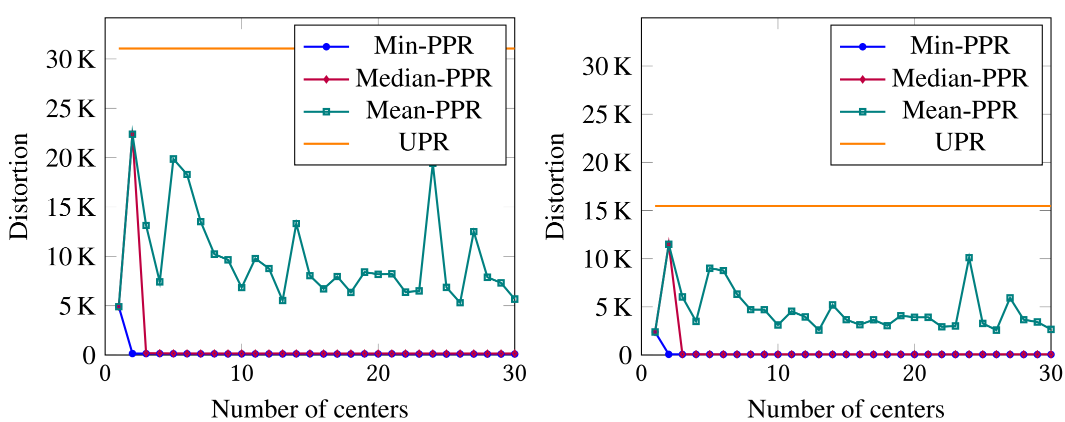
Fig. 4. Sum of ranks of all labeled trusted sites for each ranking function; higher is better. The x-axis indicates the number of trusted centers, k, used for Min-PPR, Median-PPR and Mean-PPR. Each point represents the average sum of trusted ranks over 50 independent trials. For Mean-PPR, we also plot the trusted ranks less the contributions from the "self-ranks" of each center (the rank of the center in its PPR over k). Min-PPR, Median-PPR and the no-self-ranks version of Mean-PPR all exhibit roughly similar trusted rank levels to UPR.

divide them into deciles, UPR assigns many more spammers to the highest deciles (8-10) than the other three ranking functions, and fewer spammers to most of the deciles of lower rank.



(a) $\varepsilon = 0.15$ (b) $\varepsilon = 0.01$

Fig. 5. Count of trusted sites in each decile of rank for each ranking function. The x-axis indicates the decile when each node in the web graph is ordered according to rank, so 10 is the highest decile and a higher count in high deciles is better. Each bar represents the average count of trusted sites over each $k \in \{1, 2, \ldots 30\}$ and 50 independent trials for each k. Min-PPR, Median-PPR and Mean-PPR exhibit a roughly similar distribution to UPR, with a slight increase in trusted sites in the highest decile.

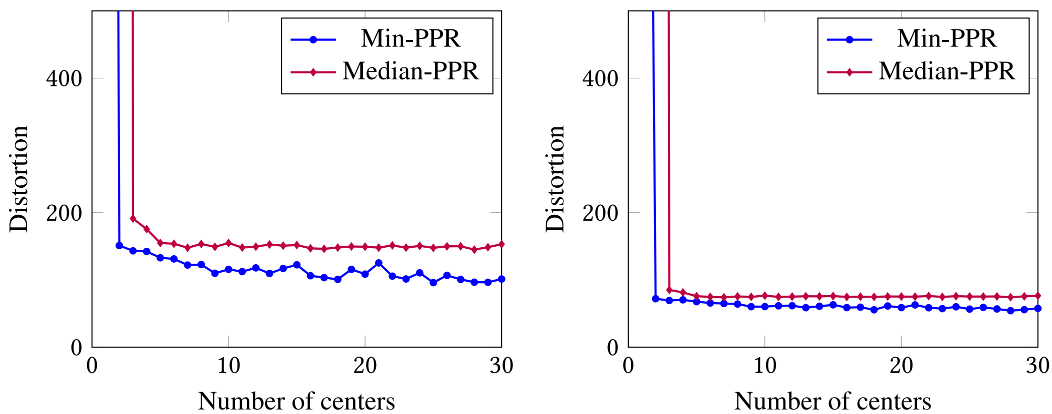


(a) $\varepsilon = 0.15$ (b) $\varepsilon = 0.01$

Fig. 6. Distortion by number of trusted centers k for each ranking function; lower is better. Each point for Min-PPR, Median-PPR and Mean-PPR represents the average distortion over 50 independent trials for each k. Min- and Median-PPR exhibit substantially lower distortion than Mean-PPR, which exhibits substantially lower distortion than UPR.

Additionally, our experimental results demonstrate that these benefits are achievable with even a very low number of trusted centers – 2 for Min-PPR and 3 for Median-PPR – and increasing the centers generally does not significantly increase the spam resistance.

C.2.2 Trusted Rank Results. Examining the trusted rank diagnostics, we see that all four ranking functions are relatively close in trusted site ranking, implying that not much trusted rank is traded



(a) $\varepsilon = 0.15$ (b) $\varepsilon = 0.01$

Fig. 7. Distortion zoomed in on Min-PPR and Median-PPR. The data is identical to Figure 6, but the scale of the y-axis is more granular in the lower distortion values, allowing greater visibility of Min-PPR and Median-PPR.

for the increased spam resistance. In fact, the plots of trusted sites in each decile of rank in Figure 5 show a modest shift of trusted sites towards the highest decile (10) for Min-, Median- and Mean-PPR compared to UPR. For the sum of trusted ranks shown in Figure 4, we note that since a PPR assigns at least ε rank to its center, and since the centers used are trusted sites, the trusted rank statistic will be inflated for Mean-PPR. Therefore, examining the rank of trusted sites apart from the contribution from the center nodes is more informative, and this statistic is plotted in the green dashed line. This adjustment does not need to be made to Min- and Median-PPR, since the min and median operations prevent large ranks at the centers. After applying this correction for Mean-PPR, we also conclude from this plot that all four ranking functions show similar overall trusted rank, with Min- and Median-PPR improving by 16-18% over UPR.

C.2.3 Distortion Results. For distortion, once again all three PPR ranking functions exhibit a significant improvement over UPR, as shown in Figure 6. This is particularly interesting with respect to the relative performance of Mean-PPR, which we would expect to have high distortion in the neighborhood of each trusted center due to the concentration of the reset vector. We find that Mean-PPR's unexpectedly low distortion when compared to UPR is explained by the observation that, for Mean-PPR, the distortion is typically maximized at the centers, which already tend to have high reference rank because they are trusted. For UPR, the distortion is maximized at nodes with virtually no reference rank, since all nodes receive an equal share of reset in UPR and therefore have a fairly high minimum rank.

Although better than UPR, the distortion of Mean-PPR is still higher than those of Min- and Median-PPR. This aligns as well with the theoretical results in this paper. Note that for one center, each of Min-, Median- and Mean-PPR is just a PPR, so the three functions exhibit equal distortion, and similarly, Median- and Mean-PPR are equal for $k = 2$ since the median of two numbers is their mean.

For $\varepsilon = 0.15$, the distortion of UPR is 31,059, while the average distortion over all trials and all k is 274 for Min-PPR, 1,051 for Median-PPR, and 9,482 for Mean-PPR. For $\varepsilon = 0.01$, the distortion of UPR is 15,480, while the average distortion over k is 139 for Min-PPR, 533 for Median-PPR, and 4,647 for Mean-PPR. A zoomed-in view in Figure 7 shows that while the distortion is quite low for

Table 1. Stability comparison for 3 centers and $\varepsilon = 0.15$

	spam ranks			trusted ranks		
	spam centers	trusted centers	difference	spam centers	trusted centers	difference
Min-PPR	0.0009	0.0007	28.5%	0.0453	0.0466	-2.8%
Median-PPR	0.0011	0.0005	105.1%	0.0450	0.0587	-23.2%

Table 2. Stability comparison for 3 centers and $\varepsilon = 0.01$

	spam ranks			trusted ranks		
	spam centers	trusted centers	difference	spam centers	trusted centers	difference
Min-PPR	0.0008	0.0008	12.2%	0.0412	0.0420	-2.0%
Median-PPR	0.0010	0.0007	38.5%	0.0409	0.0469	-12.8%

both Min-PPR and Median-PPR when k is at at least 3, Min-PPR dominates under this metric for all k. Again, we see that increasing the trusted centers beyond $k = 3$ does not lower the distortion significantly.

C.2.4 Stability Results. We compared the spam resistance of each ranking function using three spam centers to our results for three trusted centers. The results in Tables 1 and 2 summarize average spam and trusted ranks over all trials, along with the overall percentage change produced by running each function with spam centers instead of trusted centers. Min-PPR exhibits greater stability under this variation in centers, with a 12-28% increase in spam ranks and a 2-3% decrease in trusted ranks, versus spam rank increases of 39-105% and trusted rank decreases of 13-23% for Median-PPR.

We hypothesize that the relative instability of Median-PPR is due to the weak closure of the median operator: if the effective ε of the PageRank random walk represented by Median-PPR is large, then random paths will be short, and the PageRank will concentrate around the nodes that receive reset. Note that nodes throughout the graph may receive reset in a Median-PPR, not just the k PPR centers. This is because the closure of median only guarantees that there exists *some* (reset vector, ε) pair that will yield the Median-PPR as a PageRank. If the (induced) reset vector for Median-PPR with spam centers is relatively rich in spammers and the ε is large, then this would explain the instability of Median-PPR.

We selected a representative triple of trusted sites and a triple of spam sites to compute the blowup in ε. To compute the effective ε, we found the minimum ε with a non-negative reset vector. As expected, since the min operator is strongly closed, the effective ε for Min-PPR was the original ε: either 0.15 or 0.01, respectively. For Median-PPR, the input $\varepsilon = 0.15$ blows up to an effective ε of 0.40 with trusted centers and 0.47 for spam centers. When the input $\varepsilon = 0.01$, the effective ε for Median-PPR blows up to 0.22 for trusted centers and 0.36 for spam centers. Thus, the effective ε for Median-PPR is quite large.

Table 3 summarizes the results of measuring the amount of reset that spam nodes receive, for both Median-PPR and Min-PPR, when trusted or spam centers are selected. Median-PPR gives spam nodes higher reset when spam nodes are selected as centers, which, when combined with a much higher ε, explains why Median-PPR is less stable than Min-PPR.

Table 3. Sum of spam node reset probabilities

	original $\varepsilon = 0.15$			original $\varepsilon = 0.01$		
	3 spam centers	3 trusted centers	difference	3 spam centers	3 trusted centers	difference
Min-PPR	0.0007	0.0004	57.2%	0.0007	0.0004	52.6%
Median-PPR	0.0011	0.0004	165.3%	0.0010	0.0006	70.5%

C.3 Conclusion

We tested UPR, Min-PPR, Median-PPR, and Mean-PPR on a real-world web graph, in order to evaluate their performance under our formalized notions of spam resistance and distortion. The experimental outcomes conform closely with the theoretical results. Namely, UPR has low spam resistance and also suffers from high local distortion. Additionally, while Mean-PPR is significantly more spam resistant than UPR, the distortion around the trusted centers is still high.

Min-PPR and Median-PPR both exhibit promising performance under both metrics, with Median-PPR performing the best in the spam resistance trials and Min-PPR performing the best with respect to distortion. Finally, we analyzed the stability of both ranking functions under the choice of PPR centers, using a particularly extreme example of a "wrong" choice involving only labeled spam centers, and found Min-PPR to be far more stable than Median-PPR. We show evidence that suggests that the difference in stability may be related to the weaker PageRank closure properties of the median operator compared to min.

Our experiments also showed that the benefits of Min-PPR are achievable even when using as few as two trusted centers. Our results support our theoretical conclusion that Min-PPR enjoys a strong combination of high spam resistance and low distortion with low computational cost.