# Complexity of Classical Acceleration for $\ell_1$-Regularized PageRank

**Kimon Fountoulakis***
University of Waterloo, Canada
kimon.fountoulakis@uwaterloo.ca

**David Martínez-Rubio***
IMDEA Software Institute, Madrid, Spain
david.martinezrubio@imdea.org

## Abstract

We study the degree-weighted work required to compute $\ell_1$-regularized PageRank using the standard one-gradient-per-iteration accelerated proximal-gradient method (FISTA) (Bec17). For non-accelerated local methods (ISTA) (Bec17), the best known worst-case work scales as $\widetilde{O}((\alpha\rho)^{-1})$, where $\alpha$ is the teleportation parameter and $\rho$ is the $\ell_1$-regularization parameter. A natural question is whether FISTA can improve the dependence on $\alpha$ from $1/\alpha$ to $1/\sqrt{\alpha}$ while preserving the $1/\rho$ locality scaling. The challenge is that acceleration can break locality by transiently activating nodes that are zero at optimality, thereby increasing the cost of gradient evaluations. We analyze FISTA with respect to a less regularized objective and show that, under a checkable confinement condition, all spurious activations remain inside a boundary set $\mathcal{B}$. This yields a bound consisting of an accelerated $(\rho\sqrt{\alpha})^{-1}\log(\alpha/\varepsilon)$ term plus a boundary overhead $\sqrt{\mathrm{vol}(\mathcal{B})}/(\rho\alpha^{3/2})$. We provide graph-structural conditions that imply such confinement. Experiments on synthetic and real graphs show the resulting speedup and slowdown regimes under the degree-weighted work model.

## 1 Introduction

Personalized PageRank (PPR) is a diffusion primitive that, from a seed node or distribution $s$, produces a nonnegative score vector concentrated near $s$, with applications to local graph clustering and ranking (ACL06; Gle15). A key requirement is *locality*: the running time to compute the vector should scale with the size of the target set of nodes, not the full graph. $\ell_1$ regularization is useful here because it induces sparsity. In the $\ell_1$-regularized PageRank formulation (GM14; FRS+19), one solves a strongly convex problem whose minimizer is sparse and nonnegative[1]. Concretely, for teleportation parameter $\alpha \in (0, 1]$ and sparsity parameter $\rho > 0$, we consider problems of the form

$$\min_{x\in\mathbb{R}^n} \underbrace{\frac{1}{2}x^\top Qx - \alpha\langle D^{-1/2}s, x\rangle}_{\text{smooth PageRank quadratic}} + \underbrace{\alpha\rho\|D^{1/2}x\|_1}_{\ell_1 \text{ sparsity penalty}},$$

where $D$ is the degree matrix and $Q$ is a symmetric, scaled and shifted, Laplacian matrix, see Section 3. Let $x^\star$ denote the unique minimizer and let $S^\star := \mathrm{supp}(x^\star)$ be its support.

For the above problem, the primitives of first-order methods can be implemented locally: if an iterate is supported on a set $S$, evaluating its gradient and performing a proximal gradient step only requires accessing edges incident to $S$. This motivates the degree-weighted work model (FRS+19), in which scanning the neighborhood of a vertex $i$ costs $d_i$ work, and the cost of a set $S$ of non-zero nodes is $\mathrm{vol}(S) := \sum_{i\in S} d_i$. The total work of an algorithm is the cumulative number of neighbor accesses with repetition performed over its execution.

*Equal contribution.

[1] A simple corollary of FRS+19: proximal-gradient iterates started at zero are nondecreasing coordinatewise.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

**Motivation.** Accelerated first-order methods are worst-case optimal in gradient evaluations for smooth convex problems (Nes04; BT09). For $\ell_1$-regularized PageRank, however, the relevant measure is degree-weighted work, so the cost of an iteration depends on which coordinates are active. On undirected graphs, ISTA reaches a prescribed accuracy with worst-case total work $\widetilde{O}((\alpha\rho)^{-1})$ (FRS+19). It is not known whether classical acceleration methods can improve $1/\alpha$ to $1/\sqrt{\alpha}$ while preserving the $1/\rho$ locality scaling, or whether they can be asymptotically worse. We study this question for the standard one-gradient-per-iteration FISTA method. The challenge is that extrapolation can create transient activations outside $S^\star$, and even a few such activations can touch high-degree nodes and dominate the total work. We provide a negative worst-case result and a conditional upper bound on the total work.

**Total work bound and sufficient conditions.** For FISTA, under an explicit confinement condition ensuring that all spurious activations remain within a boundary set $\mathcal{B}$, we obtain a work bound of the form

$$\widetilde{O}\left(\frac{1}{\rho\sqrt{\alpha}}\log\left(\frac{\alpha}{\varepsilon}\right) \;+\; \frac{\sqrt{\mathrm{vol}(\mathcal{B})}}{\rho\alpha^{3/2}}\right).$$

The first term is the accelerated cost of converging on a slightly less regularized problem; the second term is an explicit overhead capturing the cumulative cost for exploring spurious nodes. We also give graph-structural sufficient conditions: a no-percolation criterion that makes the confinement hypothesis explicit once a candidate core set is specified. When this criterion holds for a set $S$ containing the relevant optimal support, it guarantees that momentum-induced activations cannot percolate arbitrarily far into the graph: for all iterations $k$, the iterates remain supported in $S \cup \partial S$. In particular, any activation outside $S$ is confined to the vertex boundary $\mathcal{B} := \partial S$, so the locality overhead in our work bound is governed by the boundary volume $\mathrm{vol}(\mathcal{B}) = \mathrm{vol}(\partial S)$. This makes the second term interpretable as the cost of probing only the immediate neighborhood of the core region.

**Worst-case negative result.** We show that, on star graph instances with center degree $m$, ISTA remains supported on the seed leaf and therefore has graph-size-independent work. In contrast, standard FISTA activates the high-degree center after two extrapolated steps, and incurs $\Omega(m)$ total degree-weighted work before reaching a fixed target accuracy. Thus, standard FISTA can be asymptotically worse than ISTA in the worst case.

**Contributions.** Our main contributions can be summarized as follows.

- *From KKT slack to cumulative spurious work, via over-regularization.* We show that activating an inactive coordinate forces a quantitative jump in per-iteration work controlled by its KKT slack, which together with FISTA's geometric contraction bounds cumulative spurious work. To avoid dependence on arbitrarily small slacks, we analyze the algorithm with respect to a slightly *less-regularized problem* and use regularization-path monotonicity (HFM21) to absorb nearly active nodes into the true support, charging only clearly inactive ones.
- *A conditional work bound for classic FISTA on the regularized objective.* Under a boundary confinement condition (spurious activations stay within a boundary set $\mathcal{B}$), we obtain an explicit work bound with an accelerated term $\widetilde{O}((\rho\sqrt{\alpha})^{-1}\log(\alpha/\varepsilon))$ plus a boundary overhead quantified by $\sqrt{\mathrm{vol}(\mathcal{B})}/(\rho\alpha^{3/2})$ (cf. Theorem 4.3).
- *Graph-structural confinement guarantees and degree-based non-activation.* We give a sufficient no-percolation condition for boundary confinement, and in Section B we give a sufficient degree condition under which high-degree inactive nodes provably never activate.
- *A negative worst-case result for standard FISTA.* We construct seed-at-leaf star instances for which standard FISTA activates a high-degree center and incurs $\Omega(m)$ total degree-weighted work to reach a fixed target accuracy, while ISTA remains supported on the seed and reaches the same target with $O\big(\frac{1}{\alpha}\log\frac{1}{\varepsilon}\big)$ work independent of $m$ (cf. Proposition D.4). Thus standard FISTA can be asymptotically worse than ISTA in the degree-weighted work model.

## 2 Related work

Personalized PageRank (PPR) is widely used for ranking and network analysis (Gle15). A foundational locality result of Andersen, Chung, and Lang [ACL06] shows that an $\varepsilon$-approximate PPR vector can be computed in time $\tilde{O}(1/(\alpha\varepsilon))$ independent of graph size, enabling local graph partitioning.

**Variational formulations and worst-case locality for non-accelerated methods.** The variational perspective of Gleich and Mahoney [GM14]; Fountoulakis et al. [FRS+19] shows that local clustering guarantees can be obtained by solving an $\ell_1$-regularized PageRank objective. FRS+19 show that ISTA can be implemented locally with total work $\widetilde{O}((\alpha\rho)^{-1})$, giving a worst-case graph-size-independent bound. An analogous result for standard accelerated methods, such as FISTA is an open problem. A related line studies statistical and path properties of these objectives; for instance, Ha, Fountoulakis, and Mahoney [HFM21] analyze the $\ell_1$-regularized PageRank solution path, which we leverage when reasoning about our regularized problem and a slightly less regularized one.

**The COLT'22 open problem on acceleration and its solutions/attempts.** (FY22) posed the COLT'22 open problem of whether one can obtain a provable accelerated algorithm for $\ell_1$-regularized PageRank with work $\widetilde{O}((\rho\sqrt{\alpha})^{-1})$, improving the $\alpha$-dependence by a factor of $1/\sqrt{\alpha}$ over ISTA while preserving locality. They emphasized that existing ISTA analyses do not cover acceleration and that it was unclear whether worst-case work might even degrade under acceleration. The first affirmative solution is due to Martínez-Rubio, Wirth, and Pokutta [MWP23], who design accelerated algorithms that retain sparse updates. Their method *ASPR* uses an expanding-subspace (outer-inner) scheme: it grows a set of "good" coordinates and runs an accelerated projected gradient subroutine on the restricted feasible set. This yields a worst-case bound of $O(|S^\star|\widetilde{\mathrm{vol}}(S^\star)\alpha^{-1/2}\log(1/\varepsilon) + |S^\star|\,\mathrm{vol}(S^\star))$, where $S^\star$ is the support of the optimal solution and $\widetilde{\mathrm{vol}}(S^\star)$ is the number of edges of the subgraph formed only by nodes in $S^\star$. Compared to $O(\mathrm{vol}(S^\star)\alpha^{-1}\log(1/\varepsilon)) = \widetilde{O}(1/(\rho\alpha))$ of ISTA, the solution improves the $\alpha$-dependence with a different sparsity dependence than ISTA. In this work, we provide a support-sensitive, degree-weighted work analysis of the classic one-gradient-per-iteration FISTA method. Our contribution is algorithmically quite different, and the upper bound establishes a new trade-off under explicit confinement conditions on a candidate core set. Zhou et al. [ZSB+24] study locality for accelerated linear-system solvers and obtain an accelerated guarantee under an additional run-dependent residual-reduction assumption. In contrast, our bounds for standard FISTA are explicit and quantify when acceleration helps or hurts total work.

**Support identification, strict complementarity.** Our complementarity-gap viewpoint connects to the constraint-identification literature: under strict-complementarity-type conditions, proximal and proximal-gradient methods identify the optimal support in finitely many iterations (BM94; NSH19; SJN+19), and acceleration can delay identification via oscillations (BI20) (see also (Wol70; GM86; Gar20)). These results are iteration-complexity statements under unit-cost steps and do not quantify locality-aware total work for accelerated methods.

## 3 Preliminaries and notation

We assume undirected and unweighted graphs, and we use $[n] := \{1, \ldots, n\}$. $\|\cdot\|_2$ denotes the Euclidean norm and $\|\cdot\|_1$ denotes the $\ell_1$ norm. For a set $S \subseteq [n]$ we write $|S|$ for its cardinality. If the indices in $S$ represent node indices of a graph, we use $\mathrm{vol}(S) := \sum_{i\in S} d_i$ for the graph volume, where $d_i$ is the number of neighbors of node $i$, that is, its degree. We assume $d_i > 0$ for all vertices.

We say a differentiable function $f$ is $L$-smooth if $\nabla f$ is $L$-Lipschitz with respect to $\|\cdot\|_2$, that is $\|\nabla f(x) - \nabla f(y)\|_2 \le L\|x - y\|_2$. We denote by $\mu > 0$ the strong-convexity parameter of a strongly-convex function $F$ with respect to $\|\cdot\|_2$. In such a case $F$ has a unique minimizer $x^\star$. For one such problem, define the optimal support and its complement as $S^\star := \mathrm{supp}(x^\star)$ and $I^\star := [n] \setminus S^\star$.

The main objective that we consider in this work is the personalized PageRank quadratic objective, and its $\ell_1$ regularized version. For a parameter $\alpha > 0$, called the teleportation parameter, and an initial distribution of nodes $s$ (i.e., $\langle \mathbb{1}, s\rangle = 1$, $s \ge 0$), the unregularized PageRank objective is

$$f(x) := \frac{1}{2}\langle x, Qx\rangle - \alpha\langle D^{-1/2}s, x\rangle \;\text{ for }\; Q = \alpha I + \frac{1-\alpha}{2}\mathcal{L},$$

where $\mathcal{L} := I - D^{-1/2}AD^{-1/2}$ is the symmetric normalized Laplacian matrix, which is known to satisfy $0 \preccurlyeq \mathcal{L} \preccurlyeq 2I$ (BC14). Thus, $\alpha I \preccurlyeq Q \preccurlyeq I$, which implies the objective is $\alpha$-strongly convex and 1-smooth. We will assume the seed is a single node $v$, that is $s = e_v$. This is the case for clustering applications, where one seeks to find a cluster of nodes near $v$ that have high intraconnectivity and low connectivity to the rest of the graph (ACL06; Gle15).

A common objective for obtaining sparse PageRank solutions is the $\ell_1$-Regularized Personalized PageRank problem (RPPR), which comes with the sparsity guarantee $\mathrm{vol}(S^\star) \le 1/\rho$, cf. (FRS+19,

Theorem 2), where $\rho > 0$ is a regularization weight on the objective:

$$\min_{x\in\mathbb{R}^n} F_\rho(x), \qquad \text{where} \qquad F_\rho(x) := f(x) + g(x). \tag{RPPR}$$

where $g(x) := \alpha\rho\|D^{1/2}x\|_1$. This is the central problem we study in this work.

It is worth noticing some properties of (RPPR). The initial gap from $x_0 = 0$ is $\Delta_0 := F(0) - F(x^\star) \le \alpha/2$, cf. Lemma A.1, and so by strong convexity, the initial distance to $x^\star$ satisfies $\|x^\star\|_2 \le \sqrt{2\Delta_0/\mu} \le 1$. Finally, the minimizer $x^\star(\rho)$ of $F_\rho$ is coordinatewise nonnegative and the optimality conditions are, cf. (FRS+19):

$$\nabla_i f\big(x^\star(\rho)\big) \in \begin{cases} \{-\alpha\rho\sqrt{d_i}\}, & x_i^\star(\rho) > 0, \\ [-\alpha\rho\sqrt{d_i}, 0], & x_i^\star(\rho) = 0. \end{cases} \tag{1}$$

### 3.1 The FISTA Algorithm

We introduce here the classical accelerated proximal-gradient method (FISTA) (BT09) and the properties we use later. We present the method for a composite objective $F(x) := f(x) + g(x)$ where $f$ is $L$-smooth and $F$ is $\mu$-strongly convex with respect to $\|\cdot\|_2$. For (RPPR), we have $L = 1$ and $\mu = \alpha$ (since $\alpha I \preccurlyeq Q \preccurlyeq I$), so the standard choice is step size $\eta = 1/L = 1$ and momentum parameter $\beta = \frac{\sqrt{L/\mu}-1}{\sqrt{L/\mu}+1} = \frac{1-\sqrt{\alpha}}{1+\sqrt{\alpha}}$. The iterates of the FISTA algorithm initialized with $x_{-1} = x_0 = 0$ are, for $k \ge 0$:

$$y_k = x_k + \beta(x_k - x_{k-1}), \qquad x_{k+1} = \operatorname{prox}_{\eta g}\big(y_k - \eta\nabla f(y_k)\big). \tag{FISTA}$$

The proximal operator is defined as $\operatorname{prox}_{\eta g}(x) := \arg\min_y\{\eta g(y) + \frac{1}{2}\|y - x\|_2^2\}$. For the RPPR regularizer $g(x) = \alpha\rho\|D^{1/2}x\|_1$ the prox is separable and yields:

$$x_{k+1,i} = \operatorname{sign}(y_{k,i} - \eta\nabla_i f(y_k))\ \max\Big\{\big|y_{k,i} - \eta\nabla_i f(y_k)\big| - \eta\alpha\rho\sqrt{d_i},\ 0\Big\}. \tag{2}$$

**Definition 3.1** *We measure runtime via a degree-weighted work model. For an iterate pair $(y_k, x_{k+1})$ we define the per-iteration work as*

$$\mathrm{work}_k \ := \ \mathrm{vol}(\mathrm{supp}(y_k)) + \mathrm{vol}(\mathrm{supp}(x_{k+1})). \tag{3}$$

*For ISTA, $y_k = x_k$; for FISTA, $y_k = x_k + \beta(x_k - x_{k-1})$. The total work to reach the stopping target is the sum of $\mathrm{work}_k$ over the iterations taken*[2].

## 4 FISTA's work analysis in RPPR

We provide a lower bound and a conditional upper bound on the total work of (FISTA) on (RPPR). First, the upper bound is proved by splitting the total work into a core cost and a spurious-exploration overhead. We run FISTA on the regularized objective $F_\rho$, while taking as core set $S = \mathrm{supp}(x^\star(\rho/2))$, so that $\mathrm{vol}(S) \le 2/\rho$ by the RPPR sparsity guarantee (cf. Section 3). That is, we analyze the work by looking at the sparsity of a slightly less regularized solution. The main task is then to bound the cumulative overhead from transient activations outside $S$, using complementarity slacks, the confinement condition, and FISTA's iteration complexity; this leads to the work bound proved in Theorem 4.3. We then complement this with a worst-case negative result showing that standard FISTA can be asymptotically worse than ISTA in Section D.

### 4.1 Linking work with different regularization levels

A direct upper bound analysis of FISTA naturally runs into a margin issue. In the arguments that follow, spurious activations will be controlled by KKT slacks at the optimum. For RPPR, however, the smallest slack over inactive coordinates can be arbitrarily small (see Section C), so any bound that

[2] Since each FISTA iteration computes a single gradient at $y_k$, one could alternatively take $\mathrm{work}_k := \mathrm{vol}(\mathrm{supp}(y_k))$. Our definition (3) is a convenient symmetric upper bound (it also covers evaluations at $x_{k+1}$, e.g., for stopping diagnostics), and it matches the quantities controlled in our proofs up to an absolute constant.

depends on the minimum slack would be vacuous. To obtain a work bound that remains meaningful, we link the analysis to a slightly less regularized objective, and we will relate the support of the solutions for $F_{\rho/2}$ and $F_\rho$. For these two problems, we introduce the notation:

$$g_A(x) := \alpha\rho\|D^{1/2}x\|_1 \qquad \text{and} \qquad g_B(x) := \frac{\alpha\rho}{2}\|D^{1/2}x\|_1,$$

and the corresponding minimizers

$$x_A^\star \in \arg\min_x \big(f(x) + g_A(x)\big), \qquad x_B^\star \in \arg\min_x \big(f(x) + g_B(x)\big),$$

with supports $S_A := \operatorname{supp}(x_A^\star),\ S_B := \operatorname{supp}(x_B^\star),\ I_A := [n] \setminus S_A$. We run standard (FISTA) on the (A) problem $\min_x f(x) + g_A(x)$, and we treat $S_B$ as a region where coordinates are potentially active at every iteration, even if some are inactive for $x_A^\star$. The work for the optimal nodes for both is similar, since $\operatorname{vol}(S_A) \le 1/\rho$ and $\operatorname{vol}(S_B) \le 2/\rho$, cf. Section 3. We note that the factor $1/2$ in the (B) problem is arbitrary: we could link the analysis to the problem with a $c$-factor instead $c \in (0, 1)$.

We also have to account for the work of nodes that are active outside $S_B$. Define the spurious active set at step $k$ by $\widetilde{A}_k := \operatorname{supp}(x_{k+1}) \cap S_B^c$. Such activations are the only mechanism by which FISTA can incur additional locality overhead beyond the cost of working inside $S_B$. Then, after $N$ iterations, the total degree-weighted work is bounded up to an absolute constant by

$$O\left(N \operatorname{vol}(S_B) + \sum_{k=0}^{N-1} \operatorname{vol}(\widetilde{A}_k)\right). \tag{4}$$

The first term corresponds to the cost of running $N$ proximal-gradient steps while remaining in $S_B$, since computing the gradient and applying the prox map costs work proportional to the volume of the active set. The second term is the cumulative overhead from transient activations outside $S_B$.

Our goal is to bound (4). The next subsection controls the second term. The complementarity-slack Lemma 4.1 links spurious activations to deviations in the forward-gradient map, while Lemma 4.2 ensures a uniform slack bound outside $S_B$. Together, these remove dependence on tiny margins and allow summation of spurious volumes via FISTA's geometric contraction.

### 4.2 Complementarity slack and spurious activations

We formalize how momentum-induced activations outside the optimal support translate into a quantitative cost. For a coordinate that is zero at the optimum of the (A) problem, the KKT conditions define an interval for its gradient, and the distance to its boundary measures how safely inactive it is. If FISTA activates such a coordinate, the forward step must deviate by at least this margin, which allows us to bound the cumulative work on spurious supports.

Fix the (A) problem and let $x^\star := x_A^\star$. For every inactive coordinate $i \in I_A$, define its degree-normalized complementarity (KKT) slack by

$$\gamma_i := \frac{\lambda_i - |\nabla_i f(x^\star)|}{\sqrt{d_i}} = \frac{\lambda_i + \nabla_i f(x^\star)}{\sqrt{d_i}} \ge 0, \qquad \text{where } \lambda_i := \alpha\rho\sqrt{d_i}. \tag{5}$$

The quantity $\gamma_i$ is the (degree-normalized) gap between the soft-threshold level $\lambda_i$ and the magnitude of the optimal gradient at coordinate $i$. In other words, it measures how far coordinate $i$ is from becoming active at the optimum. Define the gradient map and the set of spurious nodes by

$$u(x) := x - \eta\nabla f(x), \qquad A(y) := \operatorname{supp}\big(\operatorname{prox}_{\eta g_A}(u(y))\big) \cap I_A \quad \text{for any } x \in \mathbb{R}^n, y \in \mathbb{R}^n.$$

Here $u(\cdot)$ is the standard forward step used by proximal-gradient methods, and $A(y)$ is the subset of (A)-inactive indices that become nonzero after applying the prox map at the point $y$. The set $A(y_k)$ is exactly what creates the extra per-iteration cost due to spurious exploration with respect to an ideal local acceleration complexity of $\tilde{O}(1/(\rho\sqrt{\alpha}))$.

We now formalize the connection between a spurious activation and a nontrivial deviation in the forward step. This is the basic bridge from optimality structure to the work bound.

**Lemma 4.1** [↓] *Fix $y \in \mathbb{R}^n$. For every $i \in A(y)$, $|u(y)_i - u(x^\star)_i| > \eta\gamma_i\sqrt{d_i}$.*

Lemma 4.1 is what allows us to turn spurious activations into a summable error budget that is compatible with FISTA's convergence rate, given that the distance to optimizer bounds $\|u(y_k) - u(x^\star)\|$ and contracts with time. Recall that the margin for the (A) problem can be written as

$$\gamma_i^{(A)} := \rho\alpha + \frac{\nabla_i f(x_A^\star)}{\sqrt{d_i}} \qquad \text{for } i \in I_A. \tag{6}$$

We now show that the margin of coordinates that are not in $S_B$ is large enough.

**Lemma 4.2** [↓] *Let $I_A := [n] \setminus S_A$ be the inactive set for problem (A), and define*

$$I_A^{\text{small}} := \left\{ i \in I_A : \gamma_i^{(A)} < \rho\alpha/2 \right\}, \qquad I_A^{\text{large}} := \left\{ i \in I_A : \gamma_i^{(A)} \geq \rho\alpha/2 \right\}.$$

*Then $I_A^{\text{small}} \subseteq S_B$.*

Next, we compute a bound on the work $\text{vol}(\widetilde{A}_k)$ which is proportional to the inverse minimum margin of the coordinates involved, and this quantity is no more than $2/(\rho\alpha)$ by the lemma above.

### 4.3 Work bound and sufficient conditions

We now derive a conditional upper bound on the work. Recalling the decomposition (4), the uniform bound for the margin of coordinates in $\widetilde{A}_k$ in the previous section, together with Cauchy-Schwarz and the distance contraction of $\|y_k - x_A^\star\|_2$ that we show in Corollary A.3, makes the series $\sum_k \text{vol}(\widetilde{A}_k)$ summable and leads to the overhead term in the theorem below.

**Theorem 4.3** [↓] *For the (A) problem with objective $F_A(x) := f(x) + g_A(x)$, run (FISTA). Let $\mathcal{B}$ be a set such that $\widetilde{A}_k \subseteq \mathcal{B}$ for all $k \geq 0$, Then, we reach $F_A(x_N) - F_A(x_A^\star) \leq \varepsilon$ after a total degree-weighted work of at most*

$$\text{Work}(N_\varepsilon) \;\leq\; O\left( \frac{1}{\rho\sqrt{\alpha}} \log\left(\frac{\alpha}{\varepsilon}\right) \;+\; \frac{\sqrt{\text{vol}(\mathcal{B})}}{\rho\alpha^{3/2}} \right).$$

The bound in Theorem 4.3 separates the cost of converging on $S_B$, from the extra cost of transient exploration. The first term is the baseline accelerated contribution: FISTA needs $N_\varepsilon = O\big((\sqrt{\alpha})^{-1} \log(\alpha/\varepsilon)\big)$ iterations, and each iteration costs at most a constant times $\text{vol}(S_B) \leq 2/\rho$, yielding $O((\rho\sqrt{\alpha})^{-1} \log(\alpha/\varepsilon))$. The second term bounds the entire cumulative volume of the spurious sets $\widetilde{A}_k$. The factor $\rho^{-1}\alpha^{-3/2}$ reflects the combination of (i) the uniform margin $\rho\alpha/2$ obtained from Lemma 4.2 and (ii) the geometric contraction of the iterates, which sums to $O(\alpha^{-1/2})$.

We used hypothesis $\widetilde{A}_k \subseteq \mathcal{B}$ as an explicit locality requirement. We now give a graph-structural sufficient condition that implies such confinement, with $\mathcal{B}$ identified as a vertex boundary.

**Theorem 4.4** [↓] *Consider the (A) problem objective $F_\rho$ in (RPPR), with $\alpha \in (0, 1)$, and let $S$ be a set such that $S_A \subseteq S$. Define $\partial S$ as the vertex boundary of $S$ and $\text{Ext}(S) := V \setminus (S \cup \partial S)$. Assume that for all $i \in \text{Ext}(S)$,*

$$\frac{|\mathcal{N}(\{i\}) \cap \partial S|}{d_i} \leq \left( \frac{\alpha\rho}{4(1-\alpha)} \right)^2 d_i d_{\min \partial S}, \tag{7}$$

*where $d_{\min \partial S} := \min_{j \in \partial S} d_j$. Then the iterates of (FISTA) satisfy $\text{supp}(x_k) \subseteq S \cup \partial S$ for all $k \geq 0$:*

$$\text{supp}(x_k) \subseteq S \cup \partial S.$$

Theorem 4.4 gives a mechanism for ruling out spurious activations outside a candidate region. Condition (7) upper-bounds the fraction of an exterior node's incident edges that connect to the boundary $\partial S$, preventing extrapolated FISTA iterates from percolating into the exterior.

**Remark 4.5** *If $S = S_B$ in Theorem 4.4, then $\text{supp}(x_k) \subseteq S_B \cup \partial S_B$ for all $k \geq 0$. So any spurious activation happens in $\mathcal{B} := \partial S_B$, that is, the confinement hypothesis in Theorem 4.3 holds with $\mathcal{B} = \partial S_B$, and the overhead term in the work bound is governed by the boundary volume $\text{vol}(\partial S_B)$.*

**Remark 4.6** *In Section B, we prove a complementary property showing that nodes of high-enough degree do not ever get activated, which implies that the $\text{vol}(\mathcal{B})$ term in Theorem 4.3 can be reduced to the volume of nodes in $\mathcal{B}$ with degree below the threshold given there.*

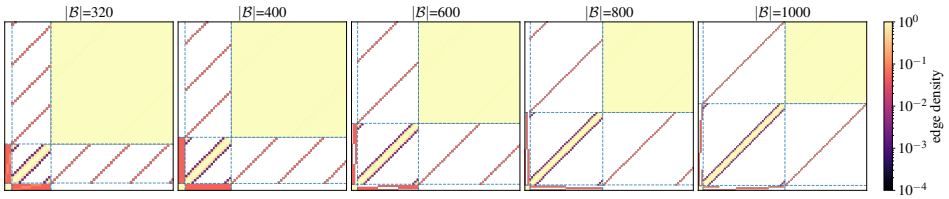



Figure 1: *Adjacency density.* We visualize the adjacency density for our synthetic block-structured graphs as a function of the size of the boundary. For readability, we show a display-only binned edge-density heatmap (bin size 20) with a contracted exterior block. Dashed lines mark the core | boundary | exterior block boundaries. The plots show the clique (upper-right block), the boundary circulant band, the contracted exterior block (lower-left), and the sparse cross-region interfaces.

### 4.4 FISTA can be worse than ISTA

We use a seed-at-leaf star instance for providing a FISTA lower bound: the optimum is supported only on the seed leaf, so ISTA stays local, but FISTA activates the high-degree center after two extrapolated steps. Since the center has degree $m$, this creates $\Omega(m)$ degree-weighted work before the method can reach the target accuracy. The full construction and proof are deferred to Section D and Proposition D.4.

**Proposition 4.7 (Informal)** *Fix $\alpha \in (0,1)$. There exists a seed-at-leaf star instance whose center has degree $m$, and a threshold $\varepsilon_0(\alpha) > 0$, such that standard FISTA needs at least $2m$ total work to reach any accuracy $0 < \varepsilon \le \varepsilon_0(\alpha)$. In contrast, ISTA needs $O\big(\frac{1}{\alpha}\log\frac{1}{\varepsilon}\big)$ work, independent of $m$.*

## 5 Experiments

This section evaluates the total work of FISTA in comparison to ISTA, reflecting the tradeoff in Theorem 4.3. We present two sets of experiments. First, we consider a controlled synthetic core-boundary-exterior graph family. Details on parameter tuning are given in Section E. Second, we compare ISTA and FISTA on real data: SNAP (LK14) graphs.

For synthetic experiments the no-percolation condition is satisfied by construction. Numerical validation margins are reported in Section H.1. We use a three-block node partition $V = S \cup \mathcal{B} \cup \mathrm{Ext}$, where $S$ contains the seed. The induced subgraph on $S$ is a clique, $\mathcal{B}$ is an internally connected boundary band, and Ext is a complete exterior block represented implicitly. Cross-region connectivity is sparse: each core node has a fixed boundary fan-out, and only a prescribed number of exterior nodes have one boundary neighbor. This yields the block structure in Figure 1; details are in Section E.[3]

### 5.1 Synthetic boundary-volume sweep experiment

This section provides synthetic experiments illustrating how the boundary volume can dominate the running time of accelerated proximal-gradient methods. The experiment illustrates how transient boundary exploration can make the accelerated method slower as $\mathrm{vol}(\mathcal{B})$ increases.

We use the core-boundary-exterior synthetic construction from Section 5, and vary only the boundary size $|\mathcal{B}|$ (and hence $\mathrm{vol}(\mathcal{B})$), keeping the core, exterior, and all degree/connectivity parameters fixed. On each instance of the sweep we solve the $\ell_1$-regularized PageRank objective (RPPR) with $\alpha = 0.20$ and $\rho = 10^{-4}$, comparing ISTA and FISTA under the common initialization, parameter choices, stopping protocol, and work accounting described in Section 5. We set $\varepsilon = 10^{-6}$.

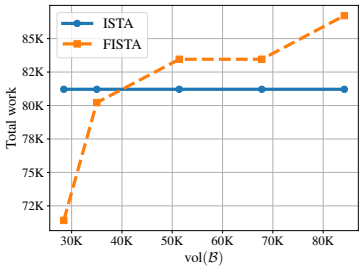



Figure 2: *Work vs.* $\mathrm{vol}(\mathcal{B})$. Work by ISTA and FISTA against $\mathrm{vol}(\mathcal{B})$.

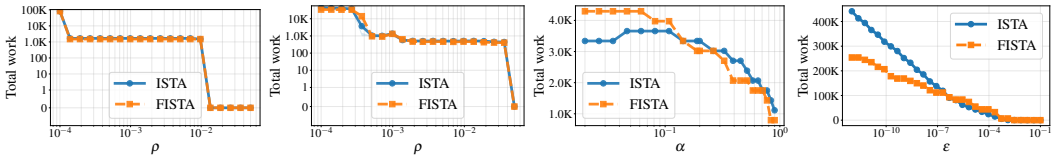
[3]The supplementary package includes the experiment scripts, saved diagnostic results, and reproduction instructions.

Figure 2 plots the work to reach the common stopping target as a function of $\mathrm{vol}(\mathcal{B})$. The key trend is that FISTA becomes increasingly expensive as $\mathrm{vol}(\mathcal{B})$ grows. For sufficiently large $\mathrm{vol}(\mathcal{B})$ it becomes slower than ISTA. The exact work decomposition in Section H.2 shows that transient activity outside the target-support reference accounts for the entire increase in FISTA's work.

### 5.2 Sweeps in $\rho$, $\alpha$ and $\varepsilon$ at fixed boundary size

We fix $|\mathcal{B}| = 600$, and we run three sweeps (summarized in Figure 3): (i) a $\rho$-sweep, reported for both a dense-core (clique) instance and a sparse-core variant (connected, $3\%$ of clique edges). For both $\rho$ sweeps, we average over 9 randomized realizations per $\rho$ and plot interquartile ranges (25%–75%) as shaded bands; in the dense-core case the randomization is over boundary/exterior permutations, while in the sparse-core case it also includes the connected sparse-core draw. We use these two variants to confirm that the observed $\rho$-dependence is not an artifact of the symmetric clique core; (ii) an $\alpha$-sweep with $\rho = 10^{-4}$; and (iii) an $\varepsilon$-sweep at fixed $\alpha = 0.20$ (complete details in Section F).

The $\rho$ sweeps (Figures 3a and 3b) show that work decreases as $\rho$ increases and collapses to 0 beyond the trivial-solution threshold; across the sweep, ISTA and FISTA exhibit qualitatively similar $1/\rho$-type scaling, consistent with the $\rho$-dependence in Theorem 4.3 for fixed $\alpha$ and fixed boundary size. The $\alpha$ sweep (Figure 3c) shows increasing work as $\alpha$ decreases, and FISTA can be slower than ISTA over a small-$\alpha$ range, consistent with the interpretation of Theorem 4.3. Finally, the $\varepsilon$ sweep (Figure 3d) shows increasing work as the tolerance decreases. FISTA is faster for small $\varepsilon$. See Section H.4 for the extended dense-core sweep.




(a) *Dense core (clique).* (b) *Sparse core.* (c) $\alpha$ *sweep* (d) $\varepsilon$ *sweep*

Figure 3: *Sweeps at fixed* $|\mathcal{B}| = 600$. Figure 3a shows the $\rho$-sweep with a clique core, averaged over 9 randomized boundary/exterior orderings per $\rho$; Figure 3b shows the corresponding sparse-core sweep with a connected $3\%$ core sparsification, averaged over 9 independent realizations per $\rho$. In both $\rho$ panels, curves show means and shaded bands show interquartile ranges (25%–75%). Figure 3c sweeps $\alpha \in [2 \cdot 10^{-2}, 0.9]$ at fixed residual tolerance $\varepsilon = 10^{-6}$. Figure 3d sweeps the tolerance $\varepsilon$ at fixed $\alpha = 0.20$.

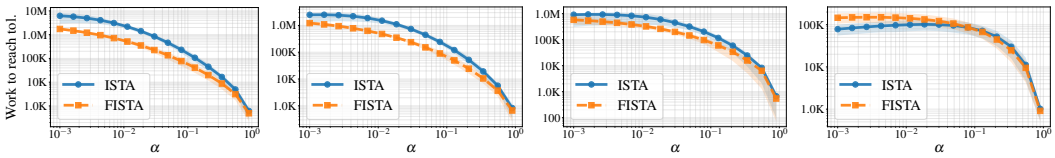
### 5.3 Real-data benchmarks on SNAP graphs

We benchmark heterogeneous SNAP (LK14) networks to examine work reductions and slowdowns without an engineered core-boundary-exterior structure. On `com-Amazon`, `com-DBLP`, and `com-Youtube` we typically observe a consistent work reduction with FISTA, whereas `com-Orkut` exhibits a setting where FISTA can be slower due to costly exploration beyond the optimal support. [4]. We track FISTA iterates and extrapolated points against an ISTA reference support at $\rho/2$. All 1,200 runs at $(\alpha, \rho) = (0.2, 10^{-4})$ remain inside it through termination; Section H.3 gives details and the $\alpha = 10^{-3}$ stress test.

**Work vs. parameter $\alpha$.** We sweep $\alpha$ over a log-spaced grid in $[10^{-3}, 0.9]$ while fixing $\rho = 10^{-4}$ and $\varepsilon = 10^{-8}$; results are in Figure 4.[5] On `com-Amazon`, `com-DBLP`, and `com-Youtube`, FISTA consistently reduces work relative to ISTA across the full $\alpha$ range. On `com-Orkut`, however, FISTA can be slower than ISTA for small $\alpha$ before becoming competitive again at moderate and large $\alpha$, illustrating that acceleration can lose under our work metric.

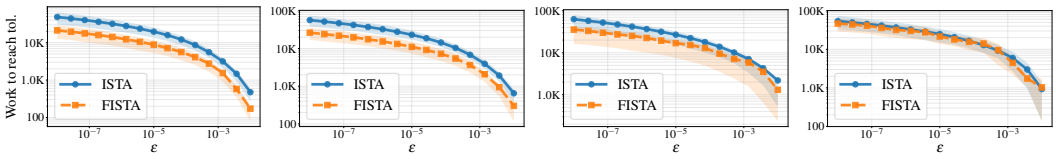
[4] For each dataset we sample 300 seed nodes uniformly at random from the non-isolated vertices; the same seed set is reused for both ISTA and FISTA and across all sweeps. In the sweep plots below, solid lines are means over the 300 seeds and the shaded bands show the interquartile range (25%-75%).

[5] Note that $\varepsilon = 10^{-8}$ is different from the value used in the synthetic experiments, which was $\varepsilon = 10^{-6}$. This is because we observed that the latter setting was too large for the real data to produce meaningful plots.



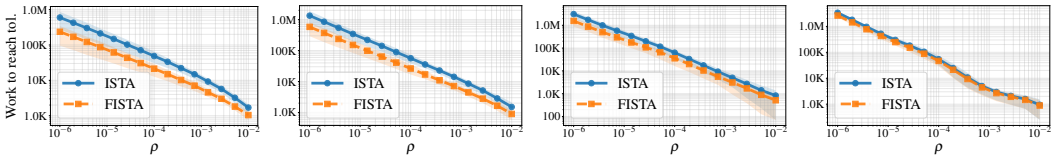

(a) com-Amazon. (b) com-DBLP. (c) com-Youtube. (d) com-Orkut.

Figure 4: *Real graphs: work vs. $\alpha$.* Work to reach tolerance $10^{-8}$ as a function of $\alpha$, with $\rho = 10^{-4}$ fixed. Curves show mean over 300 random seeds; shaded bands are interquartile ranges.




(a) com-Amazon. (b) com-DBLP. (c) com-Youtube. (d) com-Orkut.

Figure 5: *Real graphs: work vs. KKT tolerance.* Work to reach $\varepsilon$, with $\alpha = 0.20$ and $\rho = 10^{-4}$ fixed. Curves show mean over 300 random seeds; shaded bands are interquartile ranges.




(a) com-Amazon. (b) com-DBLP. (c) com-Youtube. (d) com-Orkut.

Figure 6: *Real graphs: work vs. $\rho$.* Work to reach $10^{-8}$ as a function of $\rho$, with $\alpha = 0.20$ fixed. Curves show mean over 300 random seeds; shaded bands are interquartile ranges.

**Work vs. KKT tolerance $\varepsilon$.** We next fix $\alpha = 0.20$ and $\rho = 10^{-4}$ and sweep the tolerance $\varepsilon$ over a log-spaced grid in $[10^{-8}, 10^{-2}]$; results are in Figure 5. Tightening the tolerance (smaller $\varepsilon$) increases work for both methods, and FISTA typically achieves the same tolerance with less total work on these datasets, though the gap can be small (notably on com-Orkut) for intermediate tolerances.

**Work vs. sparsity parameter $\rho$.** Finally, we fix $\alpha = 0.20$ and $\varepsilon = 10^{-8}$ and sweep $\rho$ over a log-spaced grid in $[10^{-6}, 10^{-2}]$; results are shown in Figure 6. As $\rho$ increases (stronger regularization), the solutions become more localized and the work decreases sharply. Across all four graphs, FISTA generally improves upon ISTA by a modest constant factor for these settings. We provide extension up to $\rho = 10$ in Section H.4, which yield zero work at every tested $\rho \geq 1$.

**Additional diagnostics.** The diagnostics in Section G separate iteration count from per-iteration cost. On com-Orkut, costly transient exploration can offset iteration savings (Figures 7 and 8).

## 6 Conclusion, limitations and future work

We analyzed classical FISTA for $\ell_1$-regularized PageRank under a degree-weighted locality work model. Under an explicit confinement assumption, the resulting complexity decomposes into acceleration together with an explicit overhead that quantifies momentum-induced transient exploration. We also provide a lower bound for the total work of standard FISTA on the original objective, based on a family of bad instances. Overall, we provide a comprehensive understanding of the behavior of FISTA on PageRank and regimes where it yields advantages. Natural next steps are to test whether the same slack-and-confinement framework extends to restarted or monotone variants of accelerated proximal gradient and to locality-preserving randomized/greedy coordinate proximal-gradient methods.

## Acknowledgements

K. Fountoulakis would like to acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC). Cette recherche a été financée par le Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG), [RGPIN-2019-04067, DGECR-2019-00147].

D. Martínez-Rubio was partially funded by the Spanish Ministry of Science, Innovation, and Universities and by the State Research Agency (MICIU/AEI/10.13039/501100011033/) under grant PID2024-160448NA-I00. He was also supported by a fellowship of the Bank Foundation "la Caixa" (ID 100010434) with code LCF/BQ/PI25/12100033

The authors declare no competing interests.

## References

[ACL06] Reid Andersen, Fan R. K. Chung, and Kevin J. Lang. Local Graph Partitioning using PageRank Vectors. In: *47th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2006), 21-24 October 2006, Berkeley, California, USA, Proceedings*. IEEE Computer Society, 2006, pp. 475–486 (cit. on pp. 1–3).

[BC14] Steve E. Butler and Fan R. K. Chung. Spectral Graph Theory. In: *Handbook of Linear Algebra*. Ed. by Leslie Hogben. 2nd. Boca Raton, FL, USA: CRC Press, 2014, pp. 47-1–47-14. ISBN: 9781466507289 (cit. on p. 3).

[Bec17] Amir Beck. *First-Order Methods in Optimization*. Vol. 25. MOS-SIAM Series on Optimization. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics (SIAM), 2017. ISBN: 9781611974980 (cit. on pp. 1, 12, 23).

[BI20] Gilles Bareilles and Franck Iutzeler. On the interplay between acceleration and identification for the proximal gradient algorithm. In: *Computational Optimization and Applications* 77 (2020), pp. 351–378 (cit. on p. 3).

[BM94] James V. Burke and Jorge J. Moré. Exposing Constraints. In: *SIAM J. Optim.* 4.3 (1994) (cit. on p. 3).

[BT09] Amir Beck and Marc Teboulle. A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. In: *SIAM J. Imaging Sci.* 2.1 (2009) (cit. on pp. 2, 4).

[FRS+19] Kimon Fountoulakis, Farbod Roosta-Khorasani, Julian Shun, Xiang Cheng, and Michael W Mahoney. Variational perspective on local graph clustering. In: *Mathematical Programming* 174.1 (2019), pp. 553–573 (cit. on pp. 1–4, 13, 14, 23).

[FY22] Kimon Fountoulakis and Shenghao Yang. Open Problem: Running time complexity of accelerated $\ell_1$-regularized PageRank. In: *Proceedings of Thirty Fifth Conference on Learning Theory*. Vol. 178. Proceedings of Machine Learning Research. PMLR, 2022, pp. 5630–5632 (cit. on p. 3).

[Gar20] Dan Garber. Revisiting Frank-Wolfe for Polytopes: Strict Complementarity and Sparsity. In: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*. Ed. by Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin. 2020 (cit. on p. 3).

[Gle15] David F. Gleich. PageRank Beyond the Web. In: *SIAM Review* 57.3 (2015), pp. 321–363 (cit. on pp. 1–3).

[GM14] David Gleich and Michael Mahoney. Anti-differentiating approximation algorithms:A case study with min-cuts, spectral, and flow. In: *Proceedings of the 31st International Conference on Machine Learning*. Vol. 32. Proceedings of Machine Learning Research 2. PMLR, 2014, pp. 1018–1025 (cit. on pp. 1, 3).

[GM86] Jacques Guélat and Patrice Marcotte. Some comments on Wolfe's 'away step'. In: *Math. Program.* 35 (1986), pp. 110–119 (cit. on p. 3).

[HFM21] Wooseok Ha, Kimon Fountoulakis, and Michael W. Mahoney. Statistical guarantees for local graph clustering. In: *Journal of Machine Learning Research* 22.148 (2021), pp. 1–54 (cit. on pp. 2, 3, 13).

[LK14] Jure Leskovec and Andrej Krevl. *SNAP Datasets: Stanford Large Network Dataset Collection*. *http://snap.stanford.edu/data*. June 2014 (cit. on pp. 7, 8).

[MWP23] David Martínez-Rubio, Elias Wirth, and Sebastian Pokutta. Accelerated and Sparse Algorithms for Approximate Personalized PageRank and Beyond. In: *Proceedings of Thirty Sixth Conference on Learning Theory*. Vol. 195. Proceedings of Machine Learning Research. PMLR, 2023, pp. 2852–2876 (cit. on p. 3).

[Nes04] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Vol. 87. Applied Optimization. Springer, 2004. ISBN: 978-1-4020-7553-7 (cit. on p. 2).

[NSH19] Julie Nutini, Mark Schmidt, and Warren Hare. Active-Set Complexity of Proximal Gradient: How Long Does It Take to Find the Sparsity Pattern? In: *Optimization Letters* 13 (2019), pp. 645–655 (cit. on p. 3).

[SJN+19] Yifan Sun, Halyun Jeong, Julie Nutini, and Mark Schmidt. Are we there yet? Manifold identification of gradient-related proximal methods. In: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*. Vol. 89. Proceedings of Machine Learning Research. PMLR, 2019, pp. 1110–1119 (cit. on p. 3).

[Wol70] Philip Wolfe. Convergence Theory in Nonlinear Programming. In: *Integer and Nonlinear Programming*. Ed. by J. Abadie. Amsterdam, Netherlands: North-Holland Publishing Company, 1970, pp. 1–36. ISBN: 9780444100009 (cit. on p. 3).

[ZSB+24] Baojian Zhou, Yifan Sun, Reza Babanezhad Harikandeh, Xingzhi Guo, Deqing Yang, and Yanghua Xiao. Iterative Methods via Locally Evolving Set Process. In: *Advances in Neural Information Processing Systems*. Vol. 37. Curran Associates, Inc., 2024, pp. 141528–141586 (cit. on p. 3).

# A Proofs

**Lemma A.1 (Initial gap)** *Assume the seed is a single node $v$ so $s = e_v$ and initialize $x_0 = 0$. Then $F_\rho(0) = 0$ and*

$$\Delta_0 = F_\rho(0) - F_\rho(x^\star) = -F_\rho(x^\star) \le \frac{\alpha}{2d_v} \le \frac{\alpha}{2}.$$

**Proof** We have $F_\rho(x) \ge f(x)$ and thus $F_\rho(x^\star) \ge \min_x f(x)$. The unconstrained minimizer of the quadratic $f$ is $x_f^\star = Q^{-1}b$ and $\min f = -\frac{1}{2}b^\top Q^{-1}b$. Because $Q \succeq \alpha I$, we have $Q^{-1} \preceq \frac{1}{\alpha}I$, and therefore

$$b^\top Q^{-1}b \;\le\; \frac{1}{\alpha}b^\top b \;=\; \frac{1}{\alpha}\alpha^2\|D^{-1/2}s\|_2^2 \;=\; \alpha s^\top D^{-1}s.$$

For $s = e_v$, we have $s^\top D^{-1}s = 1/d_v$, which yields

$$\min f \ge -\frac{\alpha}{2d_v}, \qquad \text{and hence} \qquad -F_\rho(x^\star) \le -\min f \le \frac{\alpha}{2d_v}.$$

■

We now state the classical guarantee on the function-value convergence on FISTA.

**Fact A.2 (FISTA convergence rate)** *Assume $f$ is $L$-smooth and $F$ is $\mu$-strongly convex with respect to $\|\cdot\|_2$. Run (FISTA) with $\eta = 1/L$ and $\beta := \frac{\sqrt{L/\mu}-1}{\sqrt{L/\mu}+1} \in [0,1)$. Then*

$$F(x_k) - F(x^\star) \le 2\,(F(x_0) - F(x^\star))\left(1 - \sqrt{\frac{\mu}{L}}\right)^k \qquad \text{for all } k \ge 1.$$

See (Bec17, Section 10.7.7) and take into account that $\frac{\mu}{2}\|x_0 - x^\star\|_2^2 \le F(x_0) - F(x^\star)$ by strong convexity. We note that the convergence guarantee above, along with strong convexity, yields bounds on the distance to the minimizer $x^\star$.

As a corollary, we can bound the distance to optimizer of the iterates along the whole computation path.

**Corollary A.3 (FISTA iterates)** *In the setting of Fact A.2, we have:*

$$\|y_0 - x^\star\|_2^2 \le \frac{4\Delta_0}{\mu} \quad \text{and} \quad \|y_k - x^\star\|_2^2 \le M\left(1 - \sqrt{\frac{\mu}{L}}\right)^{k-1} \qquad \text{for all } k \ge 1,$$

*for $M := \frac{8\Delta_0}{\mu}\Big((1+\beta)^2(1-\sqrt{\mu/L}) + \beta^2\Big)$.*

Using the bounds on $\Delta_0$ and $\mu$ in Section 3, one obtains that for (RPPR) it is $M \le 20$ and $\Delta_0/\mu \le 1/2$. Thus $\|y_k - x^\star\|_2 \le \sqrt{20}$ for all $k \ge 0$.

**Proof** Let $q := 1 - \sqrt{\mu/L}$. Strong convexity gives $F(x) - F(x^\star) \ge \frac{\mu}{2}\|x - x^\star\|_2^2$, hence, by Fact A.2:

$$\|x_k - x^\star\|_2^2 \le \frac{2}{\mu}(F(x_k) - F(x^\star)) \le \frac{4\Delta_0}{\mu}q^k,$$

which already yields the result for $k = 0$, using $y_0 = x_0$. For $k \ge 1$, write $y_k - x^\star = (1+\beta)(x_k - x^\star) - \beta(x_{k-1} - x^\star)$ and use $\|a-b\|_2^2 \le 2\|a\|_2^2 + 2\|b\|_2^2$ to obtain

$$\|y_k - x^\star\|_2^2 \le 2(1+\beta)^2\|x_k - x^\star\|_2^2 + 2\beta^2\|x_{k-1} - x^\star\|_2^2.$$

Substitute the bounds on $\|x_k - x^\star\|_2^2$ and $\|x_{k-1} - x^\star\|_2^2$. ■

**Proof of Lemma 4.1.** Fix $i \in A(y) \subseteq I^\star$. Then $x_i^\star = 0$, and we have

$$|u(y)_i - u(x^\star)_i| \overset{①}{\ge} ||u(y)_i| - |u(x^\star)_i|| \ge |u(y)_i| - |u(x^\star)_i| \overset{②}{>} \eta\lambda_i - \eta(\lambda_i - \gamma_i\sqrt{d_i}) = \eta\gamma_i\sqrt{d_i}.$$

where we used the reverse triangle inequality in ①. In ② we used that $i \in A(y)$ means $\operatorname{prox}_{\eta g}(u(y))_i \neq 0$ and so $|u(y)_i| > \eta\lambda_i$. And also that by definition, we have $|u(x^\star)_i| = \eta|\nabla f(x^\star)_i| = \eta(\lambda_i - \gamma_i\sqrt{d_i})$ by the definition of $\gamma_i$. ■

The following Lemma A.4 is proved, for example, as Lemma 4 in Ha, Fountoulakis, and Mahoney [HFM21] (see also the discussion of regularization paths in Fountoulakis et al. [FRS+19]).

**Lemma A.4 (Monotonicity of the $\ell_1$-regularized PageRank path (HFM21) )** *For the family (RPPR), let $x^\star(\rho) := \arg\min_x F_\rho(x)$, for any $\rho > 0$. The solution path is monotone: if $\rho' > \rho \geq 0$, then*

$$x^\star(\rho') \leq x^\star(\rho) \quad \textit{coordinatewise.}$$

**Lemma A.5 (Monotonicity of proximal gradient steps for PageRank)** *If $z \geq z' \geq 0$, then*

$$\operatorname{prox}_{g_c}\left(z - \frac{1}{L}\nabla f(z)\right) \geq \operatorname{prox}_{g_c}\left(z' - \frac{1}{L}\nabla f(z')\right) \qquad \textit{(componentwise).}$$

**Proof** Using the definition of the forward-gradient map $u(z) := z - \frac{1}{L}\nabla f(z)$, and that, for PageRank, $\nabla f(z) = Qz - b$ with $b = \alpha D^{-1/2}s$, we have

$$u(z) = z - \frac{1}{L}(Qz - b) = \left(I - \frac{1}{L}Q\right)z + \frac{1}{L}b.$$

Since $Q$ is an $M$-matrix (positive semidefinite and off-diagonal entries are nonpositive) and $Q \preccurlyeq LI$ (by $L$-smoothness), the matrix $I - \frac{1}{L}Q$ is entrywise nonnegative. Therefore $u(\cdot)$ is monotone componentwise: $z \geq z' \Rightarrow u(z) \geq u(z')$.

Next, for $g_c(x) := c\alpha\|D^{1/2}x\|_1$, the proximal map is separable and monotone componentwise, since

$$\left(\operatorname{prox}_{g_c}(w)\right)_i = \operatorname{sign}(w_i)\max\{|w_i| - c\alpha\sqrt{d_i}, 0\}.$$

Composing these two monotone maps yields the claim. ■

**Proof of Lemma 4.2.** Fix any $i \in I_A^{\text{small}} \subseteq I_A$. Then $x^\star_{A,i} = 0$. Recall that $u(z) := z - \nabla f(z)$ (here $\eta = 1$).

By the definition (6) of the (A)-margin at coordinate $i$ and the fact that $x^\star_{A,i} = 0$, we have

$$\begin{aligned}
\gamma_i^{(A)} < \frac{\rho\alpha}{2} &\iff \rho\alpha + \frac{\nabla_i f(x^\star_A)}{\sqrt{d_i}} < \frac{\rho\alpha}{2} \\
&\iff \frac{\nabla_i f(x^\star_A)}{\sqrt{d_i}} < -\frac{\rho\alpha}{2} \\
&\iff \nabla_i f(x^\star_A) < -\frac{\rho\alpha}{2}\sqrt{d_i} \\
&\iff -\nabla_i f(x^\star_A) > \frac{\rho\alpha}{2}\sqrt{d_i} \\
&\iff u(x^\star_A)_i > \frac{\rho\alpha}{2}\sqrt{d_i}.
\end{aligned}$$

Therefore, applying the coordinate formula for the prox of $g_A$ (cf. (2)) gives

$$\left(\operatorname{prox}_{g_B}(u(x^\star_A))\right)_i = \operatorname{sign}(u(x^\star_A)_i)\max\{|u(x^\star_A)_i| - \rho\alpha\sqrt{d_i}/2,\, 0\} > 0.$$

Next, since $\rho_A = \rho > \rho_B = \rho/2$, path monotonicity Lemma A.4 yields $x^\star_B \geq x^\star_A$ componentwise. Applying Lemma A.5 with $c = \rho/2$ (i.e., $g_B$), $z = x^\star_B$, and $z' = x^\star_A$, we obtain

$$\operatorname{prox}_{g_B}(u(x^\star_B)) \geq \operatorname{prox}_{g_B}(u(x^\star_A)) \qquad \text{(componentwise).}$$

Finally, $x^\star_B$ is a fixed point of the proximal-gradient map for the (B) problem, so $x^\star_B = \operatorname{prox}_{g_B}(u(x^\star_B))$. Hence

$$x^\star_{B,i} = \left(\operatorname{prox}_{g_B}(u(x^\star_B))\right)_i \geq \left(\operatorname{prox}_{g_B}(u(x^\star_A))\right)_i > 0,$$

which implies $i \in \operatorname{supp}(x_B^\star) = S_B$. Therefore $I_A^{\text{small}} \subseteq S_B$. ∎

**Proof of Theorem 4.3.** Recall that $\widetilde{A}_k = \operatorname{supp}(x_{k+1}) \cap S_B^c$ and that we assume $\widetilde{A}_k \subseteq \mathcal{B}$ for all $k \geq 0$. By Lemma 4.2, any index in $\widetilde{A}_k$ is (A)-inactive with margin at least $\rho\alpha/2$; concretely, $\gamma_i^{(A)} \geq \rho\alpha/2$ for all $i \in \widetilde{A}_k$. Let $\mathcal{B}'$ be the subset of $\mathcal{B}$ such that $\gamma_i^{(A)} \geq \rho\alpha/2$ for all $i \in \mathcal{B}'$. Thus, $\widetilde{A}_k \subseteq \mathcal{B}'$.

We first bound the total spurious work:

$$\begin{aligned}
\sum_{k=0}^{\infty} \sum_{i \in \widetilde{A}_k} d_i &= \sum_{k=0}^{\infty} \sum_{i \in \widetilde{A}_k} \left( \frac{\sqrt{d_i}}{\gamma_i^{(A)}} \right) \left( \gamma_i^{(A)} \sqrt{d_i} \right) \\
&\leq \sum_{k=0}^{\infty} \sqrt{\sum_{i \in \mathcal{B}'} \frac{d_i}{(\gamma_i^{(A)})^2}} \sqrt{\sum_{i \in \widetilde{A}_k} (\gamma_i^{(A)})^2 d_i} \qquad \text{(Cauchy-Schwarz, and } \widetilde{A}_k \subseteq \mathcal{B}') \\
&\leq \sum_{k=0}^{\infty} \sqrt{\sum_{i \in \mathcal{B}'} \frac{4d_i}{\rho^2\alpha^2}} \sqrt{\sum_{i \in \widetilde{A}_k} |u(y_k)_i - u(x_A^\star)_i|^2} \qquad \text{(since } \gamma_i^{(A)} \geq \rho\alpha/2, \textit{ Lemma 4.1)} \\
&\leq \sum_{k=0}^{\infty} \frac{2\sqrt{\operatorname{vol}(\mathcal{B})}}{\rho\alpha} \|u(y_k) - u(x_A^\star)\|_2 \qquad \text{(because } \mathcal{B}' \subseteq \mathcal{B}) \\
&\leq \sum_{k=0}^{\infty} \frac{2\sqrt{\operatorname{vol}(\mathcal{B})}}{\rho\alpha} \left( \|y_k - x_A^\star\|_2 + \eta \|\nabla f(y_k) - \nabla f(x_A^\star)\|_2 \right) \\
&\leq \sum_{k=0}^{\infty} \frac{2\sqrt{\operatorname{vol}(\mathcal{B})}}{\rho\alpha} (1 + \eta L) \|y_k - x_A^\star\|_2. \qquad \text{(by smoothness)}
\end{aligned} \tag{8}$$

For RPPR we use $\eta = 1$ and $L = 1$, hence $(1 + \eta L) = 2$. Using Corollary A.3 and writing $q := 1 - \sqrt{\mu/L} = 1 - \sqrt{\alpha}$, we have $\|y_0 - x_A^\star\|_2 \leq \sqrt{2}$ and $\|y_k - x_A^\star\|_2 \leq \sqrt{20}\, q^{(k-1)/2}$ for $k \geq 1$. The geometric series satisfies

$$\sum_{k=1}^{\infty} q^{(k-1)/2} = \frac{1}{1 - \sqrt{q}} = \frac{1 + \sqrt{q}}{1 - q} \leq \frac{2}{1 - q} = \frac{2}{\sqrt{\alpha}}.$$

Consequently,

$$\sum_{k=0}^{\infty} \|y_k - x_A^\star\|_2 \leq \sqrt{2} + \frac{2\sqrt{20}}{\sqrt{\alpha}} = O(\alpha^{-1/2}),$$

which bounds the distance series in (8). Therefore,

$$\sum_{k=0}^{\infty} \operatorname{vol}(\widetilde{A}_k) \;=\; O\left( \frac{\sqrt{\operatorname{vol}(\mathcal{B})}}{\rho\alpha^{3/2}} \right).$$

Next, we bound the core work. By the sparsity guarantee for RPPR (FRS+19, Theorem 2), $\operatorname{vol}(S_B) \leq 2/\rho$. Thus, after $N$ iterations, the total work is bounded by

$$\text{Work}(N) = O\left( N \operatorname{vol}(S_B) + \sum_{k=0}^{N-1} \operatorname{vol}(\widetilde{A}_k) \right) = O\left( \frac{N}{\rho} + \frac{\sqrt{\operatorname{vol}(\mathcal{B})}}{\rho\alpha^{3/2}} \right).$$

Finally, by Fact A.2, to ensure $F_A(x_N) - F_A(x_A^\star) \leq \varepsilon$ it suffices to take

$$N \geq N_\varepsilon := \left\lceil \frac{\log(2\Delta_0/\varepsilon)}{\log\left(1/(1 - \sqrt{\mu/L})\right)} \right\rceil = O\left( \frac{1}{\sqrt{\alpha}} \log\left( \frac{\alpha}{\varepsilon} \right) \right),$$

using $L = 1$, $\mu = \alpha$, and $\Delta_0 \leq \alpha/2$ from Section 3. Substituting $N = N_\varepsilon$ yields the stated bound. ∎

**Proof of Theorem 4.4.** We prove by induction that $\operatorname{supp}(x_k) \subseteq S \cup \partial S$ for all $k \geq 0$. The base case is trivial, since $x_0 = 0$, so $\operatorname{supp}(x_0) = \emptyset \subseteq S \cup \partial S$.

Now assume $\operatorname{supp}(x_{k-1}) \subseteq S \cup \partial S$ and $\operatorname{supp}(x_k) \subseteq S \cup \partial S$ for some $k \geq 0$ (with $x_{-1} = x_0 = 0$ covering $k = 0$). Then $\operatorname{supp}(y_k) \subseteq \operatorname{supp}(x_k) \cup \operatorname{supp}(x_{k-1}) \subseteq S \cup \partial S$.

Fix any $i \in \operatorname{Ext}(S)$. Since $i \notin S \cup \partial S$ and $\operatorname{supp}(y_k) \subseteq S \cup \partial S$, we have $y_{k,i} = 0$. Also $i \neq v$ (the seed lies in $S$), hence $(D^{-1/2}s)_i = 0$.

For RPPR, $\nabla f(y) = Qy - \alpha D^{-1/2}s$, so with $\eta = 1$ we have

$$u(y) = y - \nabla f(y) = (I - Q)y + \alpha D^{-1/2}s.$$

Using $Q = \frac{1+\alpha}{2} I - \frac{1-\alpha}{2} D^{-1/2} A D^{-1/2}$, we get $(I - Q) = \frac{1-\alpha}{2}\left(I + D^{-1/2} A D^{-1/2}\right)$. Therefore, for our fixed $i \in \operatorname{Ext}(S)$,

$$u(y_k)_i = \frac{1-\alpha}{2} \sum_{j \sim i} \frac{y_{k,j}}{\sqrt{d_i d_j}} = \frac{1-\alpha}{2} \sum_{j \in \mathcal{N}(\{i\}) \cap \partial S} \frac{y_{k,j}}{\sqrt{d_i d_j}},$$

because $i$ has no neighbors in $S$ (by definition of $\operatorname{Ext}(S)$) and $y_k$ has no support outside $S \cup \partial S$. Taking absolute values and using $d_j \geq d_{\min \partial S}$ for $j \in \partial S$ gives

$$\begin{aligned} |u(y_k)_i| &\leq \frac{1-\alpha}{2\sqrt{d_i}} \sum_{j \in \mathcal{N}(\{i\}) \cap \partial S} \frac{|y_{k,j}|}{\sqrt{d_j}} \overset{①}{\leq} \frac{1-\alpha}{2\sqrt{d_i}} \cdot \frac{\sqrt{|\mathcal{N}(\{i\}) \cap \partial S|}}{\sqrt{d_{\min \partial S}}} \cdot \|y_k\|_2 \\ &\overset{②}{\leq} \frac{\alpha}{8} \rho \sqrt{d_i} (\|y_k - x^\star\|_2 + \|x^\star\|_2) \overset{③}{\leq} \alpha \rho \sqrt{d_i}. \end{aligned}$$

where ① uses Cauchy-Schwarz and ② uses the triangular inequality and our no-percolation assumption (7). Finally ③ uses the bound $\|y_k - x^\star\|_2 \leq \sqrt{20}$ from Corollary A.3 and $\|x^\star\|_2 \leq 1$ from the preliminaries Section 3 and bound the resulting constants to an integer number.

For the (A) problem, the shrinkage threshold is $\lambda_i = \alpha\rho\sqrt{d_i}$, so we showed $|u(y_k)_i| \leq \lambda_i$ and thus the proximal update keeps $x_{k+1,i} = 0$ (cf. the coordinate formula (2)). Since this holds for every $i \in \operatorname{Ext}(S)$, we conclude $\operatorname{supp}(x_{k+1}) \subseteq S \cup \partial S$. This completes the induction. ■

## B Out-of-solution high-degree nodes do not activate

We provide an extra property of (FISTA). Spurious nodes of high-enough degree can never be activated by the accelerated iterates. The proof argues that (FISTA) on problem (A) with a large margin prevents high-degree nodes, and large margins are guaranteed for coordinates outside $S_B$.

**Proposition B.1 (Large-degree nodes do not activate)** *Run (FISTA) on problem (A) $F_\rho$, and let $R$ be any uniform bound such that $\|y_k - x_A^\star\|_2 \leq R$ for all $k \geq 0$. Let $i$ be such that the minimizer of problem (B) $F_{\rho/2}$ satisfies $x^\star_{B,i} = 0$. If*

$$d_i \geq \left(\frac{2LR}{\alpha\rho}\right)^2,$$

*then (FISTA) does not activate node $i$, that is $x_{k,i} = y_{k,i} = 0$ for all $k \geq 0$.*

For our PageRank problem (RPPR), we have $L \leq 1$, and Corollary A.3 allows us to take $R \leq \sqrt{20}$. Therefore nodes satisfying $d_i \geq 80\alpha^{-2}\rho^{-2}$ will not get activated, unless they are part of the $(B)$ solution.

**Proof** Fix $i$ such that $x^\star_{B,i} = 0$. By path monotonicity Lemma A.4 and nonnegativity of the minimizers,

$$\Delta x := x_B^\star - x_A^\star \geq 0 \qquad \text{and} \qquad \Delta x_i = 0.$$

Recall $\nabla f(x) = Qx - b$, where $b \geq 0$ and $Q_{ij} \leq 0$ for $i \neq j$. Since $x_B^\star, x_A^\star \geq 0$ and $x_{B,i}^\star = x_{A,i}^\star = 0$, we have

$$\nabla_i f(x_B^\star) \leq 0, \qquad \nabla_i f(x_A^\star) \leq 0,$$

and

$$\nabla_i f(x_A^\star) - \nabla_i f(x_B^\star) = -(Q\Delta x)_i = -\sum_{j \neq i} Q_{ij} \Delta x_j \geq 0.$$

Hence

$$|\nabla_i f(x_A^\star)| \leq |\nabla_i f(x_B^\star)| \leq \alpha\rho\sqrt{d_i}/2,$$

where the last inequality holds by the KKT conditions for problem (B), that is, $(\rho/2)$-regularization.

We now prove $x_{k,i} = 0$ for all $k$ by induction. The base case holds since $x_{-1} = x_0 = 0$. Assume $x_{k,i} = x_{k-1,i} = 0$. Then $y_{k,i} = 0$ and by $L$-smoothness,

$$|\nabla_i f(y_k)| \leq |\nabla_i f(x_A^\star)| + \|\nabla f(y_k) - \nabla f(x_A^\star)\|_2 \leq \frac{\alpha\rho}{2}\sqrt{d_i} + L\|y_k - x_A^\star\|_2.$$

Using $\|y_k - x_A^\star\|_2 \leq R$ and $d_i \geq \left(\frac{2LR}{\alpha\rho}\right)^2$, we obtain

$$|\nabla_i f(y_k)| \leq \alpha\rho\sqrt{d_i}. \tag{9}$$

Let $w_k = y_k - \eta\nabla f(y_k)$. Since $y_{k,i} = 0$, $|w_{k,i}| = \eta|\nabla_i f(y_k)|$. The proximal map of $g(x) = \alpha\rho\|D^{1/2}x\|_1$ is weighted soft-thresholding, so by (9),

$$x_{k+1,i} = \operatorname{sign}(w_{k,i}) \max\{|w_{k,i}| - \eta\alpha\rho\sqrt{d_i}, 0\} = 0.$$

This closes the induction and proves $x_{k,i} = 0$ for all $k \geq 0$. ■

## C Bad instances where the margin $\gamma$ can be very small

We now record two explicit graph families showing that the degree-normalized strict-complementarity margin (the one that naturally interfaces with our degree-weighted work model in (4)) can be made arbitrarily small (and even $\gamma = 0$) by choosing $\rho$ near a breakpoint of the regularization path where an inactive KKT inequality becomes tight. This motivates our theory for linking a problem (A) with the sparsity pattern of a slightly less regularized one (B), so that no requirement in the minimum margin is made (since we split the coordinates into low-margin ones which are included in the support of the (B) solution and then the high-margin ones). Concretely, let $x^\star$ be the minimizer and $I^\star := \{i : \ x_i^\star = 0\}$. Define the coordinatewise degree-normalized KKT slack

$$\gamma_i := \frac{\lambda_i - |\nabla_i f(x^\star)|}{\sqrt{d_i}} \qquad (i \in I^\star), \qquad \lambda_i := \rho\alpha\sqrt{d_i},$$

and the global margin

$$\gamma := \min_{i \in I^\star} \gamma_i.$$

### C.1 Star graph (seed at the center)

Let $G$ be a star on $m + 1$ nodes with center node $c$ of degree $d_c = m$ and leaves $\ell$ of degree 1. Let the seed be $s = e_c$.

**Lemma C.1 (Star graph breakpoint: $\gamma$ can be 0)** *Fix $\alpha \in (0, 1)$ and $m \geq 1$ and define*

$$\rho_0 := \frac{1 - \alpha}{2m}.$$

*For any $\rho \in [\rho_0, 1/m)$, let*

$$x^\star := x_c^\star e_c, \qquad x_c^\star = \frac{2\alpha(1 - \rho m)}{(1 + \alpha)\sqrt{m}}.$$

*Then $x^\star$ is a minimizer of $F_\rho$, hence the unique minimizer by $\alpha$-strong convexity. In particular, $S^\star = \{c\}$. Moreover, for any leaf $\ell$ (recall $d_\ell = 1$), with $\lambda_\ell = \alpha\rho\sqrt{d_\ell} = \alpha\rho$, the degree-normalized slack equals*

$$\gamma_\ell \;:=\; \frac{\lambda_\ell - |\nabla_\ell f(x^\star)|}{\sqrt{d_\ell}} \;=\; \lambda_\ell - |\nabla_\ell f(x^\star)| \;=\; \frac{2\alpha}{1+\alpha}\,(\rho - \rho_0),$$

*and thus at the breakpoint $\rho = \rho_0$ one has $\gamma = 0$.*

**Proof** Recall that $f(x) = \frac{1}{2}x^\top Qx - \alpha\langle D^{-1/2}s, x\rangle$, hence

$$\nabla f(x) = Qx - \alpha D^{-1/2}s,$$

and

$$Q = \alpha I + \frac{1-\alpha}{2}\big(I - D^{-1/2}AD^{-1/2}\big) = \frac{1+\alpha}{2}I - \frac{1-\alpha}{2}D^{-1/2}AD^{-1/2}.$$

On the star with seed $s = e_c$, we have $D^{-1/2}s = e_c/\sqrt{m}$, and for each leaf $\ell$, $(D^{-1/2}AD^{-1/2})_{\ell c} = 1/\sqrt{m}$. Therefore

$$Q_{cc} = \frac{1+\alpha}{2}, \qquad Q_{\ell c} = Q_{c\ell} = -\frac{1-\alpha}{2\sqrt{m}}.$$

Assume $x^\star = x_c^\star e_c$ with $x_c^\star > 0$. For the $\ell_1$-regularized PageRank objective $F_\rho(x) = f(x) + \alpha\rho\|D^{1/2}x\|_1$, the coordinatewise KKT conditions (cf. (1)) give, at an active coordinate, $\nabla_c f(x^\star) = -\alpha\rho\sqrt{d_c} = -\alpha\rho\sqrt{m}$. Using $\nabla_c f(x^\star) = (Qx^\star)_c - \alpha/\sqrt{m} = Q_{cc}x_c^\star - \alpha/\sqrt{m}$, we obtain

$$0 = \nabla_c f(x^\star) + \alpha\rho\sqrt{m} = Q_{cc}x_c^\star - \frac{\alpha}{\sqrt{m}} + \alpha\rho\sqrt{m} = \frac{1+\alpha}{2}x_c^\star - \frac{\alpha}{\sqrt{m}} + \alpha\rho\sqrt{m},$$

which yields

$$x_c^\star = \frac{2\alpha(1-\rho m)}{(1+\alpha)\sqrt{m}},$$

and this is positive exactly when $\rho < 1/m$. Now fix any leaf $\ell$. Since $x_\ell^\star = 0$, the KKT condition requires $\nabla_\ell f(x^\star) \in [-\alpha\rho\sqrt{d_\ell},\, 0] = [-\alpha\rho,\, 0]$. Here

$$\nabla_\ell f(x^\star) = (Qx^\star)_\ell = Q_{\ell c}x_c^\star = -\frac{1-\alpha}{2\sqrt{m}}\,x_c^\star \le 0,$$

so the inactive condition is equivalent to

$$|\nabla_\ell f(x^\star)| = \frac{1-\alpha}{2\sqrt{m}}\,x_c^\star \le \alpha\rho.$$

Substituting the expression for $x_c^\star$ and cancelling $\alpha > 0$ gives

$$\frac{1-\alpha}{2\sqrt{m}} \cdot \frac{2\alpha(1-\rho m)}{(1+\alpha)\sqrt{m}} \le \alpha\rho \quad\Longleftrightarrow\quad \frac{(1-\alpha)(1-\rho m)}{(1+\alpha)m} \le \rho \quad\Longleftrightarrow\quad \rho \ge \frac{1-\alpha}{2m} = \rho_0.$$

Hence for $\rho \in [\rho_0, 1/m)$, the point $x^\star = x_c^\star e_c$ satisfies all KKT conditions. Since $F_\rho$ is $\alpha$-strongly convex, these KKT conditions certify that $x^\star$ is the unique minimizer and $S^\star = \{c\}$. Finally, for any leaf $\ell$ (with $d_\ell = 1$) the degree-normalized slack is

$$\gamma_\ell = \frac{\lambda_\ell - |\nabla_\ell f(x^\star)|}{\sqrt{d_\ell}} = \alpha\rho - \frac{1-\alpha}{2\sqrt{m}}\,x_c^\star = \alpha\rho - \frac{1-\alpha}{2\sqrt{m}} \cdot \frac{2\alpha(1-\rho m)}{(1+\alpha)\sqrt{m}} = \frac{2\alpha}{1+\alpha}\,(\rho - \rho_0),$$

so at $\rho = \rho_0$ we indeed have $\gamma = 0$. ∎

### C.2 Path graph (seed at an endpoint)

Let $G = P_{m+1}$ be the path on nodes $1, 2, \ldots, m+1$ with edges $(i, i+1)$. Let $s = e_1$ (seed at endpoint 1). Assume $m \ge 2$, so that $d_1 = d_{m+1} = 1$ and $d_i = 2$ for $2 \le i \le m$. Consider candidates of the form $x = x_1 e_1$.

**Lemma C.2 (Path graph breakpoint: $\gamma$ can be 0)** *Fix $\alpha \in (0,1)$ and $m \geq 2$ and define*

$$\rho_0 := \frac{1-\alpha}{3+\alpha}.$$

*For any $\rho \in [\rho_0, 1)$, let*

$$x^\star := x_1^\star e_1, \qquad x_1^\star = \frac{2\alpha(1-\rho)}{1+\alpha}.$$

*Then $x^\star$ is a minimizer of $F_\rho$, hence the unique minimizer by $\alpha$-strong convexity. In particular, $S^\star = \{1\}$. Moreover, the* degree-normalized *KKT slack at node 2 (where $d_2 = 2$), with $\lambda_2 = \alpha\rho\sqrt{d_2} = \alpha\rho\sqrt{2}$, equals*

$$\gamma_2 \;:=\; \frac{\lambda_2 - |\nabla_2 f(x^\star)|}{\sqrt{d_2}} \;=\; \frac{\alpha(3+\alpha)}{2(1+\alpha)}\,(\rho - \rho_0).$$

*In particular, at the breakpoint $\rho = \rho_0$ one has $\gamma = 0$.*

**Proof** Recall that $f(x) = \frac{1}{2}x^\top Q x - \alpha\langle D^{-1/2}s, x\rangle$, hence

$$\nabla f(x) = Qx - \alpha D^{-1/2}s, \qquad Q = \frac{1+\alpha}{2}I - \frac{1-\alpha}{2}D^{-1/2}AD^{-1/2}.$$

Since $s = e_1$ and $d_1 = 1$, we have $D^{-1/2}s = e_1$. Also, for the edge $(1,2)$ we have $(D^{-1/2}AD^{-1/2})_{21} = 1/\sqrt{d_2 d_1} = 1/\sqrt{2}$, so

$$Q_{11} = \frac{1+\alpha}{2}, \qquad Q_{21} = Q_{12} = -\frac{1-\alpha}{2\sqrt{2}}.$$

Assume $x^\star = x_1^\star e_1$ with $x_1^\star > 0$. For the $\ell_1$-regularized PageRank objective $F_\rho(x) = f(x) + \alpha\rho\|D^{1/2}x\|_1$, the active KKT condition at node 1 is

$$\nabla_1 f(x^\star) = -\alpha\rho\sqrt{d_1} = -\alpha\rho.$$

But $\nabla_1 f(x^\star) = (Qx^\star)_1 - \alpha = Q_{11}x_1^\star - \alpha$, hence

$$0 = \nabla_1 f(x^\star) + \alpha\rho = Q_{11}x_1^\star - \alpha + \alpha\rho = \frac{1+\alpha}{2}\,x_1^\star - \alpha + \alpha\rho,$$

which yields

$$x_1^\star = \frac{2\alpha(1-\rho)}{1+\alpha},$$

and this is positive iff $\rho < 1$. Now consider node 2 (which is inactive under our candidate). Since $(D^{-1/2}s)_2 = 0$ and $x^\star$ is supported only on node 1,

$$\nabla_2 f(x^\star) = (Qx^\star)_2 = Q_{21}x_1^\star = -\frac{1-\alpha}{2\sqrt{2}}\,x_1^\star \leq 0,$$

so

$$|\nabla_2 f(x^\star)| = \frac{1-\alpha}{2\sqrt{2}}\,x_1^\star.$$

The inactive KKT condition at node 2 requires $|\nabla_2 f(x^\star)| \leq \alpha\rho\sqrt{d_2} = \alpha\rho\sqrt{2}$. Substituting $x_1^\star$ gives

$$\frac{1-\alpha}{2\sqrt{2}} \cdot \frac{2\alpha(1-\rho)}{1+\alpha} \leq \alpha\rho\sqrt{2} \quad\Longleftrightarrow\quad \frac{(1-\alpha)(1-\rho)}{1+\alpha} \leq 2\rho \quad\Longleftrightarrow\quad \rho \geq \frac{1-\alpha}{3+\alpha} = \rho_0.$$

For nodes $i \geq 3$, we have $(Qx^\star)_i = Q_{i1}x_1^\star = 0$ because node 1 is adjacent only to node 2, and also $(D^{-1/2}s)_i = 0$, hence $\nabla_i f(x^\star) = 0$, which satisfies the inactive KKT condition $|\nabla_i f(x^\star)| \leq \alpha\rho\sqrt{d_i}$. Therefore, for any $\rho \in [\rho_0, 1)$, the point $x^\star = x_1^\star e_1$ satisfies all KKT conditions. Since $F_\rho$ is $\alpha$-strongly convex, this certifies that $x^\star$ is the unique minimizer and $S^\star = \{1\}$. Finally, the degree-normalized slack at node 2 is

$$\begin{aligned}\gamma_2 &= \frac{\lambda_2 - |\nabla_2 f(x^\star)|}{\sqrt{d_2}} = \frac{\alpha\rho\sqrt{2} - \frac{1-\alpha}{2\sqrt{2}}x_1^\star}{\sqrt{2}} = \alpha\rho - \frac{1-\alpha}{4}x_1^\star \\ &= \alpha\rho - \frac{1-\alpha}{4}\cdot\frac{2\alpha(1-\rho)}{1+\alpha} = \frac{\alpha}{2(1+\alpha)}\Big((3+\alpha)\rho - (1-\alpha)\Big) = \frac{\alpha(3+\alpha)}{2(1+\alpha)}\,(\rho - \rho_0),\end{aligned}$$

and at $\rho = \rho_0$ this slack is 0, so $\gamma = 0$. ∎

## D FISTA can be worse than ISTA: a lower bound

We exhibit a family of star instances for which ISTA remains supported on the seed leaf and therefore has graph-size-independent work, whereas standard FISTA activates the high-degree center after two extrapolated steps, and incurs $\Omega(m)$ degree-weighted work, where $m+1$ is the number of nodes in the star graph. Consequently, for a fixed target accuracy depending only on $\alpha$, FISTA can be asymptotically worse than ISTA by a factor linear in the graph size. The proof is as follows: we identify the regularization level for which the center stays inactive at optimality, show that activation for FISTA reduces to a seed-coordinate condition, prove that FISTA satisfies the condition after two steps, and then show that the resulting cost is incurred before the target accuracy is reached.

### D.1 Construction

Fix an integer $m \geq 2$. Let $G(m)$ be the star graph on $n = m+1$ vertices with vertex set $\{w, v, u_1, \ldots, u_{m-1}\}$, where $w$ is the center and $v, u_1, \ldots, u_{m-1}$ are the leaves. The edge set is $\big\{\{w,v\}\big\} \cup \big\{\{w,u_i\} : i = 1, \ldots, m-1\big\}$. Thus every leaf is adjacent only to the center $w$, and there are no edges between distinct leaves. In particular, the center has degree $d_w = m$, while each leaf has degree 1, that is, $d_v = d_{u_i} = 1$ for all $i = 1, \ldots, m-1$. We distinguish $v$ as the seed node. For the regularization regime in this section, the optimal solution is supported only on the seed leaf, so $S^\star = \{v\}$, $\partial S^\star = \{w\}$, and $\mathrm{Ext}(S^\star) = \{u_1, \ldots, u_{m-1}\}$. Hence $\mathrm{vol}(S^\star) = 1$ and $\mathrm{vol}(\partial S^\star) = m$.

### D.2 Results

The following lemma pins down the optimal solution and the critical regularization breakpoint.

**Lemma D.1** *For the graph $G(m)$ with seed $s = e_v$ and any $\rho \in [\rho_0, 1)$ where*

$$\rho_0 \;:=\; \frac{1-\alpha}{m(1+\alpha)+(1-\alpha)},$$

*let*

$$x^\star := x^\star_v e_v, \qquad x^\star_v = \frac{2\alpha(1-\rho)}{1+\alpha}.$$

*Then $x^\star$ is a minimizer of $F_\rho$ (cf. (RPPR)), hence the unique minimizer by $\alpha$-strong convexity. In particular, $S^\star = \{v\}$. The degree-normalized complementarity margin at the center $w$ is*

$$\gamma_w \;=\; \frac{\alpha\big(m(1+\alpha)+(1-\alpha)\big)}{(1+\alpha)m}\,(\rho - \rho_0).$$

*In particular, at $\rho = \rho_0$ the margin is $\gamma_w = 0$.*

**Proof** Since $d_v = 1$ and $s = e_v$, we have $D^{-1/2}s = e_v$. The PageRank matrix satisfies

$$Q_{vv} = \frac{1+\alpha}{2}, \qquad Q_{wv} = Q_{vw} = -\frac{1-\alpha}{2\sqrt{m}}, \qquad Q_{u_i v} = 0 \quad \text{for all } i,$$

because there are no edges between leaves. Assume $x^\star = x^\star_v e_v$ with $x^\star_v > 0$. The active KKT condition at $v$ gives

$$Q_{vv}x^\star_v - \alpha = -\alpha\rho,$$

so

$$x^\star_v = \frac{2\alpha(1-\rho)}{1+\alpha},$$

which is positive for $\rho < 1$. For the center $w$ (with $d_w = m$),

$$\nabla_w f(x^\star) = Q_{wv}x^\star_v = -\frac{1-\alpha}{2\sqrt{m}} \cdot \frac{2\alpha(1-\rho)}{1+\alpha} = -\frac{\alpha(1-\alpha)(1-\rho)}{(1+\alpha)\sqrt{m}}.$$

The inactive KKT condition at $w$ requires $|\nabla_w f(x^\star)| \leq \alpha\rho\sqrt{m}$, i.e.,

$$\frac{(1-\alpha)(1-\rho)}{1+\alpha} \leq \rho m.$$

Solving the equality case yields

$$\rho_0 = \frac{1-\alpha}{m(1+\alpha)+(1-\alpha)},$$

and thus the condition holds for all $\rho \geq \rho_0$. Each pendant leaf $u_i$ is neither the seed nor adjacent to $v$, so

$$|\nabla_{u_i} f(x^\star)| = 0 \leq \alpha\rho,$$

and its KKT condition is satisfied. By $\alpha$-strong convexity, $x^\star$ is the unique minimizer and $S^\star = \{v\}$. Finally, the degree-normalized margin at $w$ is

$$\gamma_w = \alpha\rho - \frac{|\nabla_w f(x^\star)|}{\sqrt{m}} = \alpha\rho - \frac{\alpha(1-\alpha)(1-\rho)}{(1+\alpha)m} = \frac{\alpha\big(m(1+\alpha)+(1-\alpha)\big)}{(1+\alpha)m}\,(\rho-\rho_0),$$

which vanishes exactly at $\rho = \rho_0$. ■

We next derive the exact criterion for activation of the center $w$. The FISTA update decides activation through the forward point $u(y) = y - \nabla f(y)$, since the proximal step makes the coordinate $w$ nonzero exactly when $|u(y)_w|$ exceeds the weighted soft-threshold $\alpha\rho_0\sqrt{m}$. On the star graph, if $y$ is supported on $\{v, w\}$, then a direct calculation shows that $u(y)_w = \frac{1-\alpha}{2}(y_w + y_v/\sqrt{m})$, so the center is influenced only by its own value and by the seed value transmitted through the unique edge $(v, w)$. The breakpoint $\rho = \rho_0$ is chosen so that, at the optimum supported only on the seed leaf, the center $w$ is exactly at the point where the proximal update changes from keeping it zero to making it nonzero, namely $\frac{1-\alpha}{2\sqrt{m}}x_v^\star = \alpha\rho_0\sqrt{m}$. Rewriting the threshold test using this identity yields the criterion below. In particular, before the first activation, when $y_w = 0$ and the iterates are nonnegative, the condition reduces to $y_v > x_v^\star$. Thus the lemma is the bridge to Lemma D.3: once we prove that FISTA overshoots the seed coordinate beyond $x_v^\star$, activation of the center follows immediately.

**Lemma D.2** *At $\rho = \rho_0$, consider any point $y$ with $\mathrm{supp}(y) \subseteq \{v, w\}$. Let*

$$x^+ := \mathrm{prox}_{\alpha\rho_0\|D^{1/2}\cdot\|_1}\big(u(y)\big), \qquad u(y) := y - \nabla f(y).$$

*Then*

$$x_w^+ \neq 0 \quad \Longleftrightarrow \quad |u(y)_w| > \alpha\rho_0\sqrt{m} \quad \Longleftrightarrow \quad |y_w\sqrt{m} + y_v| > x_v^\star.$$

*In particular, if $y_w, y_v \geq 0$, then*

$$x_w^+ > 0 \quad \Longleftrightarrow \quad y_w\sqrt{m} + y_v > x_v^\star.$$

**Proof** By the coordinate formula for the proximal operator,

$$x_w^+ \neq 0 \quad \Longleftrightarrow \quad |u(y)_w| > \alpha\rho_0\sqrt{m}.$$

Since $\mathrm{supp}(y) \subseteq \{v, w\}$ and $w$ is not the seed,

$$u(y)_w = \big((I-Q)y\big)_w = \frac{1-\alpha}{2}\left(y_w + \frac{y_v}{\sqrt{m}}\right).$$

At the breakpoint,

$$\frac{1-\alpha}{2\sqrt{m}}\,x_v^\star = \alpha\rho_0\sqrt{m}.$$

Therefore,

$$|u(y)_w| > \alpha\rho_0\sqrt{m} \quad \Longleftrightarrow \quad \frac{1-\alpha}{2}\left|y_w + \frac{y_v}{\sqrt{m}}\right| > \frac{1-\alpha}{2\sqrt{m}}x_v^\star \quad \Longleftrightarrow \quad |y_w\sqrt{m} + y_v| > x_v^\star.$$

If $y_w, y_v \geq 0$, then $u(y)_w \geq 0$, so $x_w^+ > 0$ iff $x_w^+ \neq 0$, giving the last claim. ■

We now show that FISTA generates exactly the condition required by Lemma D.2. The key point is that, before the center becomes active, every iterate remains supported on the seed leaf $v$, so the dynamics reduce to a one-dimensional accelerated proximal-gradient iteration on the seed coordinate alone. In this regime, the update at $v$ is affine, and the error relative to the optimum, $e_k := x_{k,v} - x_v^\star$, satisfies an explicit scalar recurrence. This allows us to compute the first few iterates exactly. We verify first that the extrapolated points at $k = 0$ and $k = 1$ do not cross the activation threshold, so the center is still inactive. At $k = 2$, however, the momentum term pushes the extrapolated seed coordinate past the critical value $x_v^\star$, that is, $y_{2,v} > x_v^\star$. By Lemma D.2, this activates the center $w$.

**Lemma D.3** *Run (FISTA) on $F_{\rho_0}$ (cf. (RPPR)) for the graph $G(m)$ with seed $s = e_v$, starting from $x_{-1} = x_0 = 0$. Then*

$$y_{2,v} - x^\star_v = \frac{(1-\alpha)\beta^2}{2}\, x^\star_v > 0. \tag{10}$$

*Consequently, FISTA activates the center $w$ at iteration $k = 2$.*

**Proof** At $k = 0$, we have $y_0 = 0$, so only the seed coordinate can become active. Thus

$$x_{1,v} = \alpha(1-\rho_0) > 0, \qquad x_{1,w} = x_{1,u_i} = 0.$$

Hence $x_1$ is supported on $\{v\}$. Define the errors

$$e_k := x_{k,v} - x^\star_v, \qquad \tilde{e}_k := y_{k,v} - x^\star_v.$$

As long as $w$ has not been activated, both $x_k$ and $y_k$ are supported on $\{v\}$. On such steps,

$$u(y_k)_v = (1 - Q_{vv})y_{k,v} + \alpha = \frac{1-\alpha}{2} y_{k,v} + \alpha.$$

Since $y_{k,v} \geq 0$ on the steps we consider, the soft-threshold at $v$ acts affinely, and therefore

$$x_{k+1,v} = u(y_k)_v - \alpha\rho_0 = \frac{1-\alpha}{2} y_{k,v} + \alpha(1-\rho_0).$$

Writing

$$a := \frac{1-\alpha}{2},$$

and subtracting the fixed-point identity

$$x^\star_v = a\, x^\star_v + \alpha(1-\rho_0),$$

we get

$$e_{k+1} = a\,\tilde{e}_k = a\big((1+\beta)e_k - \beta e_{k-1}\big).$$

Equivalently,

$$e_{k+1} = a(1+\beta)e_k - a\beta e_{k-1}. \tag{11}$$

*Initial values.* We have

$$e_0 = -x^\star_v.$$

The first FISTA step gives

$$x_{1,v} = \alpha(1-\rho_0),$$

hence

$$e_1 = \alpha(1-\rho_0) - \frac{2\alpha(1-\rho_0)}{1+\alpha} = \frac{1-\alpha}{2}\, e_0 = a e_0.$$

*The center is not activated at $k = 0$ or $k = 1$.* At $k = 0$, $y_{0,v} = 0 < x^\star_v$, so Lemma D.2 shows that $w$ is not activated. At $k = 1$, since $x_{1,w} = x_{0,w} = 0$, we have $y_{1,w} = 0$, and

$$\tilde{e}_1 = (1+\beta)e_1 - \beta e_0 = e_0\big((1+\beta)a - \beta\big).$$

Using

$$(1+\beta)a = 1 - \sqrt{\alpha} \qquad \text{and} \qquad (1+\beta)a - \beta = \sqrt{\alpha}\beta > 0,$$

we get

$$\tilde{e}_1 = \sqrt{\alpha}\beta\, e_0 < 0$$

because $e_0 < 0$. Thus $y_{1,v} < x^\star_v$, and Lemma D.2 again implies that $w$ is not activated at $k = 1$.

*Computing $\tilde{e}_2$.* Since $w$ is not activated at $k = 0$ or $k = 1$, the recurrence (11) applies up to $e_2$:

$$e_2 = a(1+\beta)e_1 - a\beta e_0 = a e_0\big(a(1+\beta) - \beta\big) = a\sqrt{\alpha}\beta\, e_0.$$

Therefore,

$$\begin{aligned}
\tilde{e}_2 &= (1+\beta)e_2 - \beta e_1 \\
&= (1+\beta)\, a\sqrt{\alpha}\beta\, e_0 - \beta a e_0 \\
&= a\beta e_0\big((1+\beta)\sqrt{\alpha} - 1\big).
\end{aligned}$$

Now

$$(1+\beta)\sqrt{\alpha} = \frac{2\sqrt{\alpha}}{1+\sqrt{\alpha}},$$

so

$$(1+\beta)\sqrt{\alpha} - 1 = \frac{\sqrt{\alpha}-1}{1+\sqrt{\alpha}} = -\beta.$$

Hence

$$\tilde{e}_2 = -a\beta^2 e_0 = a\beta^2 x_v^\star = \frac{(1-\alpha)\beta^2}{2}\, x_v^\star > 0.$$

Thus $y_{2,v} > x_v^\star$. Also, since $x_{2,w} = x_{1,w} = 0$, we have $y_{2,w} = 0$. Applying Lemma D.2 with $y = y_2$ therefore shows that FISTA activates $w$ at iteration $k = 2$. ■

We now convert the activation of the center into a lower bound on the total degree-weighted work. Once Lemma D.3 shows that the center becomes active at iteration $k = 2$, the next iterate $x_3$ already contains the high-degree node $w$, and the following extrapolated point $y_3$ contains it as well. Under our work model, this immediately creates work of order $m$ in two successive iterations. To obtain a lower bound for reaching a prescribed accuracy, it remains to show that this expensive activation occurs before FISTA can terminate. We therefore bound the objective gap explicitly along the first few iterates and prove that, for every target $\varepsilon \le \varepsilon_0(\alpha)$, none of $x_0, x_1, x_2, x_3$ is yet $\varepsilon$-accurate. Hence any successful run must execute at least four iterations and must incur at least $2m$ total work. By contrast, ISTA remains supported on the seed leaf throughout, so its work stays independent of $m$.

**Proposition D.4** *Fix $\alpha \in (0,1)$ and define*

$$\varepsilon_0(\alpha) \;:=\; \frac{\alpha^3(1-\alpha)^4\beta^4}{2(3+\alpha)^2} \;>\; 0, \qquad \beta = \frac{1-\sqrt{\alpha}}{1+\sqrt{\alpha}}.$$

*On the graph $G(m)$ with seed $s = e_v$ and $\rho = \rho_0$, for every target accuracy*

$$0 < \varepsilon \le \varepsilon_0(\alpha),$$

*standard FISTA requires total degree-weighted work at least $2m$ to reach*

$$F_{\rho_0}(x_N) - F_{\rho_0}(x^\star) \le \varepsilon.$$

*By contrast, ISTA reaches the same target with total work*

$$O\left(\frac{1}{\alpha}\log\frac{1}{\varepsilon}\right),$$

*independent of $m$.*

**Proof** Let

$$a := \frac{1-\alpha}{2}.$$

*FISTA lower bound.* By Lemma D.3, FISTA activates $w$ at iteration $k = 2$, i.e., $x_{3,w} > 0$. Hence

$$w \in \mathrm{supp}(x_3) \qquad \text{and} \qquad \mathrm{vol}(\mathrm{supp}(x_3)) \ge \deg(w) = m.$$

Also, $x_{2,w} = 0$, so

$$y_3 = x_3 + \beta(x_3 - x_2)$$

satisfies

$$y_{3,w} = (1+\beta)x_{3,w} > 0.$$

Therefore

$$w \in \mathrm{supp}(y_3) \qquad \text{and} \qquad \mathrm{vol}(\mathrm{supp}(y_3)) \ge m.$$

Thus iterations $k = 2$ and $k = 3$ each incur per-iteration work at least $m$:

$$\mathrm{work}_2 \ge \mathrm{vol}(\mathrm{supp}(x_3)) \ge m, \qquad \mathrm{work}_3 \ge \mathrm{vol}(\mathrm{supp}(y_3)) \ge m.$$

It remains to show that, for every target accuracy $0 < \varepsilon \le \varepsilon_0(\alpha)$, the algorithm must execute at least four iterations. For $k = 0, 1, 2$, the center has not yet been activated, so $x_k$ is supported on $\{v\}$. Moreover these iterates are nonnegative. Hence, on the ray $\{xe_v : \ x \ge 0\}$,

$$F_{\rho_0}(xe_v) = \frac{1+\alpha}{4}x^2 - \alpha(1-\rho_0)x,$$

and therefore

$$F_{\rho_0}(x_k) - F_{\rho_0}(x^\star) = \frac{1+\alpha}{4}(x_{k,v} - x_v^\star)^2.$$

From the proof of Lemma D.3, the corresponding errors satisfy

$$e_0 := x_{0,v} - x_v^\star = -x_v^\star, \qquad e_1 = -a x_v^\star, \qquad e_2 = -a\sqrt{\alpha}\beta\, x_v^\star.$$

Since $a, \sqrt{\alpha}\beta \in (0,1)$, the smallest of the first three gaps occurs at $k = 2$, and therefore

$$F_{\rho_0}(x_k) - F_{\rho_0}(x^\star) \ \ge \ F_{\rho_0}(x_2) - F_{\rho_0}(x^\star) = \frac{1+\alpha}{4}a^2\alpha\beta^2(x_v^\star)^2 \qquad \text{for } k = 0, 1, 2.$$

At $\rho = \rho_0$,

$$x_v^\star = \frac{2\alpha(1-\rho_0)}{1+\alpha} = \frac{2\alpha m}{m(1+\alpha)+(1-\alpha)} \ge \frac{4\alpha}{3+\alpha},$$

where the last inequality uses $m \ge 2$. Substituting this bound yields

$$F_{\rho_0}(x_k) - F_{\rho_0}(x^\star) \ge \frac{1+\alpha}{4}a^2\alpha\beta^2\left(\frac{4\alpha}{3+\alpha}\right)^2 = \frac{\alpha^3(1+\alpha)(1-\alpha)^2\beta^2}{(3+\alpha)^2} > \varepsilon_0(\alpha)$$

for $k = 0, 1, 2$, because

$$2(1+\alpha) > (1-\alpha)^2\beta^2.$$

For $k = 3$, using Lemma D.3 and the $v$-update,

$$x_{3,v} - x_v^\star = a(y_{2,v} - x_v^\star) = a^2\beta^2 x_v^\star.$$

By $\alpha$-strong convexity,

$$F_{\rho_0}(x_3) - F_{\rho_0}(x^\star) \ge \frac{\alpha}{2}\|x_3 - x^\star\|_2^2 > \frac{\alpha}{2}(x_{3,v} - x_v^\star)^2,$$

where the inequality is strict because $x_{3,w} > 0$ while $x_w^\star = 0$. Using the bound on $x_v^\star$ above,

$$F_{\rho_0}(x_3) - F_{\rho_0}(x^\star) > \frac{\alpha}{2}\left(a^2\beta^2 \cdot \frac{4\alpha}{3+\alpha}\right)^2 = \varepsilon_0(\alpha).$$

Hence

$$F_{\rho_0}(x_N) - F_{\rho_0}(x^\star) > \varepsilon_0(\alpha) \ge \varepsilon \qquad \text{for every } N \le 3.$$

So any run that reaches

$$F_{\rho_0}(x_N) - F_{\rho_0}(x^\star) \le \varepsilon \qquad \text{with } 0 < \varepsilon \le \varepsilon_0(\alpha)$$

must have $N \ge 4$. Since total work sums over iterations,

$$\mathrm{Work}(N) \ge \mathrm{work}_2 + \mathrm{work}_3 \ge 2m.$$

*ISTA upper bound on the same instance.* By Theorem 1(ii) in FRS+19, the support of each ISTA iterate is contained in the optimal support. Furthermore, Lemma D.1 shows that the optimal support is $S^\star = \{v\}$, so $|S^\star| = 1$. Hence, the per-iteration work of ISTA is $\mathcal{O}(1)$ with respect to $m$. Moreover, Theorem 10.30 in Bec17 states that ISTA requires $\mathcal{O}\big(\frac{1}{\alpha}\log\frac{1}{\varepsilon}\big)$ iterations to obtain a solution whose objective value is within $\varepsilon$ of the optimum. Therefore, the total work of ISTA is $\mathcal{O}\big(\frac{1}{\alpha}\log\frac{1}{\varepsilon}\big)$, which is independent of $m$. ■

## E Experimental setting details

This section collects the common experimental ingredients used throughout the synthetic experiments in Sections 5.1 and 5.2. All experiments solve the $\ell_1$-regularized PageRank objective (RPPR) and report runtime using the degree-weighted work metric in (3). When we refer to the no-percolation diagnostic, we mean the inequality from Theorem 4.4.

**Synthetic graph family: core-boundary-exterior construction.** Each synthetic instance is an undirected graph with a three-way partition of the node set $V = S \cup \mathcal{B} \cup \mathrm{Ext}$, where $S$ is a core (containing the seed), $\mathcal{B}$ is a boundary region, and $\mathrm{Ext}$ is an exterior. The base construction is deterministic once the node orderings are fixed. For the deterministic sweeps we use the canonical order. For the $\rho$-sweeps, we draw independent boundary/exterior permutations for repeated realizations at each $\rho$; in the sparse-core $\rho$-sweep we additionally sample the connected sparse core in each realization. Given sizes $|S|$, $|\mathcal{B}|$, and $|\mathrm{Ext}|$, edges are added according to the following rules:

- *Core clique.* The induced subgraph on $S$ is a complete graph (a clique).
- *Core-boundary connectivity.* Let the core nodes be ordered as $S = \{0, 1, \ldots, |S| - 1\}$ and let the boundary nodes be stored in an ordered list $(b_0, b_1, \ldots, b_{|\mathcal{B}|-1})$. Each core node has $c_{\mathrm{bnd}}$ boundary per core neighbors in $\mathcal{B}$. For each core node $u \in S$ and each $j \in \{0, 1, \ldots, c_{\mathrm{bnd}} - 1\}$ we add the edge $(u,\ b_{(u \cdot c_{\mathrm{bnd}} + j) \bmod |\mathcal{B}|})$. When $|\mathcal{B}| \geq c_{\mathrm{bnd}}$ (as in our sweeps), this gives $c_{\mathrm{bnd}}$ distinct boundary neighbors per core node. Each core node has fixed degree $d_u = (|S| - 1) + c_{\mathrm{bnd}}$ for $|\mathcal{B}| > 0$.
- *Boundary internal connectivity.* The boundary induces a circulant graph with an even degree parameter $\deg_{\mathcal{B}}$, capped at $|\mathcal{B}| - 1$, and adjusted to be even.
- *Exterior internal connectivity.* In the numerical synthetic experiments, the exterior induces a complete graph on $|\mathrm{Ext}|$ nodes. The implementation does not materialize this dense block: matrix-vector products with the exterior are evaluated implicitly via $A_{\mathrm{Ext}} z = (\mathbf{1}^\top z)\mathbf{1} - z$.
- *Boundary-exterior connectivity.* A prescribed number $m_{\text{Ext-}\mathcal{B}}$ of exterior nodes have exactly one neighbor in $\mathcal{B}$, assigned round-robin over the boundary nodes. The remaining exterior nodes have no boundary neighbor. Thus an attached exterior node has degree $|\mathrm{Ext}|$, while an unattached exterior node has zero boundary exposure.

This construction yields a dense core, an internally connected boundary band, and a highly connected exterior, with sparse cross-region interfaces. When we visualize adjacency matrices, this produces a clear block structure (core | boundary | exterior) and a boundary region whose size/volume can be varied independently of the core neighborhood.

*Adjacency-density visualization.* The adjacency-density figure in Figure 1 is a display-only visualization. To keep the core and boundary visible in the figure, the visualization uses a contracted display exterior with $|\mathrm{Ext}|_{\mathrm{disp}} = 1000$ and display exterior degree $\deg_{\mathrm{Ext,disp}} = 998$, while preserving the same boundary-side interface size $m_{\text{Ext-}\mathcal{B}} = 1000$. This display graph is not used our experiments.

**Optimization objective and parameters.** On each graph instance we solve the $\ell_1$-regularized PageRank objective (RPPR) with a single-node seed $s = e_v$. Unless otherwise specified, the seed node $v$ is a fixed core vertex (in the code, $v = 0$). Each experiment specifies a teleportation parameter $\alpha \in (0, 1]$ and a sparsity parameter $\rho > 0$. When using FISTA we set the momentum parameter to the standard strongly-convex choice $\beta := \frac{1-\sqrt{\alpha}}{1+\sqrt{\alpha}}$ (for PageRank, $L = 1$ and $\mu = \alpha$). Both ISTA and FISTA are initialized at $x_{-1} = x_0 = 0$.

**Stopping criterion.** All experiments compare ISTA and FISTA under the same KKT surrogate based on the proximal-gradient fixed point. With unit step size, define the prox-gradient map

$$T_{\alpha,\rho}(x) := \mathrm{prox}_g\big(x - \nabla f(x)\big), \qquad r(x) := \|x - T_{\alpha,\rho}(x)\|_\infty.$$

A point $x^\star$ is optimal for (RPPR) if and only if $x^\star = T_{\alpha,\rho}(x^\star)$, i.e., $r(x^\star) = 0$. We therefore declare convergence when the fixed-point residual satisfies $r(x_k) \leq \varepsilon$, where $\varepsilon > 0$ is the prescribed tolerance. This termination rule is applied identically to ISTA and FISTA. In the work-vs-$\varepsilon$ sweeps, the $x$-axis parameter *is* this residual tolerance $\varepsilon$; for the other sweeps, $\varepsilon$ is held fixed (and we impose a single large global iteration cap, e.g. 50,000, for all runs). We terminate the algorithm based on the fixed-point residual rather than the objective gap $F_\rho(x_k) - F_\rho(x^\star)$, since evaluating this gap requires

the unknown optimal value $F_\rho(x^\star)$. The objective value $F_\rho(x_k)$ itself is computable, but it does not by itself give a stopping certificate without a reference value.

**Degree-weighted work model.** We measure runtime via a degree-weighted work model (3). For an iterate pair $(y_k, x_{k+1})$ we define the per-iteration work as $\mathrm{work}_k := \mathrm{vol}(\mathrm{supp}(y_k)) + \mathrm{vol}(\mathrm{supp}(x_{k+1}))$. For ISTA, $y_k = x_k$; for FISTA, $y_k = x_k + \beta(x_k - x_{k-1})$. The work to reach the stopping target is the sum of $\mathrm{work}_k$ over the iterations taken.

**No-percolation diagnostic.** The no-percolation Assumption (7) is satisfied for all our synthetic experiments. Conceptually, this condition is favorable for accelerated methods: it rules out "percolation" of extrapolated iterates into the exterior, so FISTA is not penalized by activating a large, highly connected ambient region. Nevertheless, our sweeps still exhibit regimes where FISTA does not improve work (and can be slower than ISTA), showing that even when exterior exploration is provably suppressed, acceleration can lose due to transient boundary activations.

**Default synthetic parameters.** Unless a sweep varies them, the fixed-$\alpha$ synthetic numerical experiments use $|S| = 60$, $|\mathrm{Ext}| = 25000$, $c_{\mathrm{bnd}} = 20$, $\deg_{\mathcal{B}} = 82$, $m_{\text{Ext-}\mathcal{B}} = 1000$, with seed node $v = 0 \in S$, $\alpha = 0.20$, and $\rho = 10^{-4}$. The exterior is complete and represented implicitly.

**Per-point graph generation and how to read sweep plots.** The boundary-volume and fixed-$\varepsilon$ synthetic sweeps are deterministic given their parameters. In the $\rho$-sweeps, each plotted point is the mean over 9 controlled randomized realizations from the same specified family, and the shaded band is the interquartile range (25%–75%). In the dense-core $\rho$-sweep, the realizations differ by independent boundary/exterior permutations while keeping the same block sizes and degree parameters. In the sparse-core $\rho$-sweep, each realization also samples an independent connected sparse-core graph. The no-percolation assumption is checked for every generated realization.

## F Full details for the fixed-boundary sweeps experiments

We provide full details for the experiments in Section 5.2. All use the construction and work metric from Section E and fix $|\mathcal{B}| = 600$. For the $\rho$- and $\varepsilon$-sweeps we use the fixed Assumption-(7)-compliant baseline $|\mathrm{Ext}| = 25{,}000$, $m_{\text{Ext-}\mathcal{B}} = 1000$, $c_{\mathrm{bnd}} = 20$, $\deg_{\mathcal{B}} = 82$.

This experiment complements the boundary-volume sweep of Section 5.1 by holding the boundary size fixed ($|\mathcal{B}| = 600$) and varying only the regularization strength $\rho$. The aim is to isolate the $\rho$-dependence suggested by Theorem 4.3 (both terms scale as $1/\rho$ when $\alpha$ and the boundary are fixed), and to check whether ISTA and FISTA respond similarly as $\rho$ increases, since their worst-case theoretical running time depends on $\rho$ in the same way. We run two versions of the $\rho$-sweep, both averaging over 9 randomized realizations per $\rho$ and plotting interquartile ranges:

- *Dense-core sweep.* The core subgraph is a clique. For each $\rho$, we generate 9 randomized variants by independently permuting the boundary/exterior order. We sweep

  $$\rho \in \mathrm{logspace}(10^{-4}, 5 \cdot 10^{-2}, 20).$$

  Curves report the mean over the 9 realizations and shaded bands report the interquartile range (25%–75%). Assumption (7) is checked for every realization.

- *Sparse-core sweep.* The core subgraph is replaced by a connected sparsification: the implementation samples a random spanning tree and then adds random core-core edges up to 3% of the clique edge count. We use the same $\rho$-grid as the dense sweep and average over 9 independent connected sparse-core realizations per $\rho$. Curves report the mean over the 9 realizations and shaded bands report the interquartile range (25%–75%). Assumption (7) is checked for every realization.

For the $\alpha$-sweep, we keep $\rho = 10^{-4}$ and $\varepsilon = 10^{-6}$ fixed and vary only $\alpha$. Since Assumption (7) becomes more stringent as $\alpha$ decreases, we use a single unweighted $\alpha$-sweep-specific instance sized to satisfy the no-percolation inequality at the smallest swept value $\alpha_{\min} = 2 \cdot 10^{-2}$. The exterior is a complete graph on $|\mathrm{Ext}|$ nodes; exactly $m_{\text{Ext-}\mathcal{B}} = 1000$ exterior nodes have one boundary neighbor, and the remaining exterior nodes have no boundary neighbor. We set $|\mathcal{B}| = 600$, $c_{\mathrm{bnd}} = 100$, $\deg_{\mathcal{B}} = 82$, $m_{\text{Ext-}\mathcal{B}} = 1000$. For this construction, $|\mathrm{Ext}|$ is chosen from the closed-form lower bound implied by Assumption (7), with safety factor 1.05. Equivalently, using $d_{\min \partial S} = 93$, the

Table 1: Basic statistics of the SNAP graphs used in the real-data experiments.

| Dataset | Number of nodes | Number of edges |
|---|---|---|
| com-Amazon | 334,863 | 925,872 |
| com-DBLP | 317,080 | 1,049,866 |
| com-Youtube | 1,134,890 | 2,987,624 |
| com-Orkut | 3,072,441 | 117,185,083 |

code takes

$$|\mathrm{Ext}| = \left\lceil 1.05\,\frac{4(1-\alpha_{\min})}{\alpha_{\min}\rho\sqrt{d_{\min\,\partial S}}} \right\rceil = 213{,}405.$$

The large complete exterior is a certification device: it raises the degree of the few exterior vertices that touch the boundary, thereby making their boundary-exposure ratio small enough to satisfy the sufficient no-percolation condition at $\alpha_{\min}$. The complete exterior is represented implicitly and is not materialized as a dense adjacency block. At $\alpha_{\min}$, the left-hand side of Assumption (7) is $4.69 \times 10^{-6}$, while the right-hand side is $5.17 \times 10^{-6}$, giving margin about 1.10. Since the right-hand side is increasing in $\alpha$ for fixed graph parameters and $\alpha \in (0, 1)$, this verifies no-percolation for every $\alpha \geq \alpha_{\min}$ in the sweep. The sampled grid is

$$\mathrm{geomspace}(2 \cdot 10^{-2}, 0.19, 9) \;\cup\; \mathrm{linspace}(0.20, 0.90, 12).$$

For the $\varepsilon$-sweep, we keep $\alpha = 0.20$ fixed and vary the fixed-point residual tolerance over a log-spaced grid $\varepsilon \in [10^{-12}, 10^{-1}]$. We use the fixed Assumption (7)-compliant baseline instance with $|\mathrm{Ext}| = 25{,}000$, $m_{\text{Ext-}\mathcal{B}} = 1000$, $c_{\mathrm{bnd}} = 20$, $\deg_{\mathcal{B}} = 82$.

## G Additional real-data statistics and diagnostics

In this section we interpret the results of the experiments on real data from Section 5.3. First, we mention a few details about the real data in Table 1.

**Diagnosing slowdowns: iterations vs. per-iteration work.** The work metric counts degree-weighted support volumes touched by both the extrapolated point $y_k$ and the proximal update, so FISTA can lose either by taking more iterations than ISTA or by having a larger per-iteration locality cost. To separate these effects, for each seed (at $\alpha = 10^{-3}$, $\rho = 10^{-4}$, $\varepsilon = 10^{-8}$) we plot

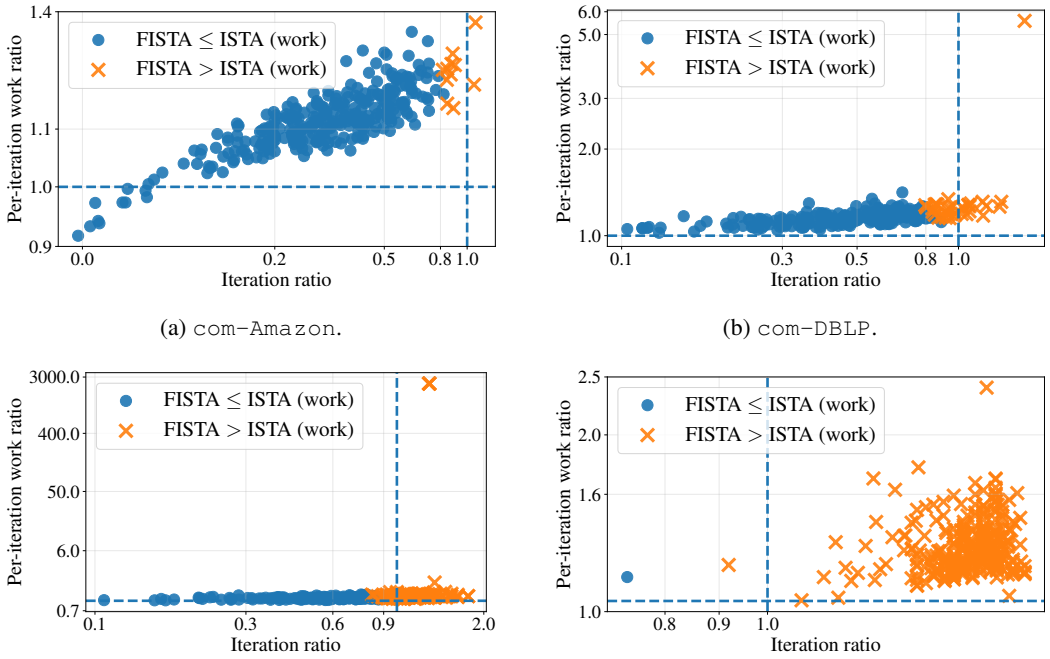
$$\text{iteration ratio} \;=\; \frac{N_{\mathrm{F}}}{N_{\mathrm{I}}} \qquad \text{vs.} \qquad \text{per-iter ratio} \;=\; \frac{(W_{\mathrm{F}}/N_{\mathrm{F}})}{(W_{\mathrm{I}}/N_{\mathrm{I}})},$$

where $N_{\mathrm{I}}, N_{\mathrm{F}}$ are iteration counts and $W_{\mathrm{I}}, W_{\mathrm{F}}$ are total works for ISTA and FISTA, respectively. Since $\frac{W_{\mathrm{F}}}{W_{\mathrm{I}}} = \frac{N_{\mathrm{F}}}{N_{\mathrm{I}}} \cdot \frac{(W_{\mathrm{F}}/N_{\mathrm{F}})}{(W_{\mathrm{I}}/N_{\mathrm{I}})}$, points with both ratios above 1 correspond to clear slowdowns. Figure 7 shows that on com-Orkut at $\alpha = 10^{-3}$, FISTA is frequently slower because it often incurs both a larger iteration count and a larger per-iteration work cost, whereas on the other datasets FISTA typically reduces iterations while paying a moderate per-iteration locality overhead.

**Degree heterogeneity.** Because our work metric is degree-weighted, transient activations of even a small number of high-degree nodes can dominate the locality cost. Figure 8 plots the empirical degree complementary cumulative distribution function for the four datasets and highlights the substantially heavier tail of com-Orkut (and, to a lesser extent, com-Youtube), which is consistent with the larger variability and the small-$\alpha$ slowdowns observed in Figures 4 and 7.

## H Additional experimental results and diagnostics

We extend the experiments in Section 5 with additional regularization values and per-iteration diagnostics. Method runs use the same solver updates, degree-weighted work metric, fixed-point residual stopping rule, iteration cap of 50,000, and seed nodes. Residual tolerances are $10^{-8}$ on real data and $10^{-6}$ on synthetic data. The real-data experiments use the same 300 seeds per dataset; the $\alpha = 10^{-3}$ stress test uses the first 100 of these seeds. Hats denote reference supports formed from nonzero computed entries. Activation measurements cover the executed iterations through termination. Auxiliary reference solves are excluded from reported method work.




(a) com-Amazon. (b) com-DBLP.

(c) com-Youtube. (d) com-Orkut.

Figure 7: *Iterations vs. per-iteration work tradeoff.* Each point is a seed node (same seeds as in the sweep experiments), at $\alpha = 10^{-3}$, $\rho = 10^{-4}$, and $\varepsilon = 10^{-8}$. The $x$-axis is the iteration ratio $N_{\mathrm{F}}/N_{\mathrm{I}}$ and the $y$-axis is the per-iteration work ratio $(W_{\mathrm{F}}/N_{\mathrm{F}})/(W_{\mathrm{I}}/N_{\mathrm{I}})$. Markers distinguish seeds where FISTA is faster/slower in total work.

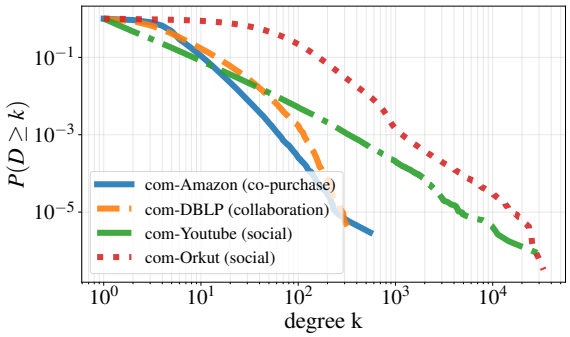



Figure 8: *Degree distributions on the real datasets.* The heavier-tailed degree profiles (notably com-Orkut) amplify the impact of transient exploration.

### H.1 Numerical validation of the synthetic construction

The synthetic experiments check (7) for every generated realization. At $(\alpha, \rho) = (0.2, 10^{-4})$, the multiplicative margins, defined as the right-hand side divided by the left-hand side, are 2.05–2.15 across the boundary-volume family and 2.08 for the nine boundary/exterior realizations of the regularization sweeps. At the hardest point of the $\alpha$-sweep, $\alpha = 0.02$, the margin is 1.10. This last instance is sized using a 1.05 degree safety factor, giving a margin of $1.05^2 \simeq 1.10$; the construction is described in Section F. These checks cover the synthetic operating point $\rho = 10^{-4}$ and the $\alpha$-sweep over $[0.02, 0.9]$.

### H.2 Where work is spent in the boundary-volume experiment

For the five boundary-volume instances of Section 5.1, we measure work relative to $\widehat{S}_A$, the support returned by an auxiliary ISTA solve of $F_\rho$. Table 2 gives the exact work accounting at $(\alpha, \rho) = (0.2, 10^{-4})$.

| $\mathrm{vol}(\mathcal{B}) = \mathrm{vol}(\partial S)$ | Work ISTA | Work FISTA | FISTA outside $\widehat{S}_A$ |
|---|---|---|---|
| $28,440$ | $81,212$ | $71,416$ | $0$ |
| $35,000$ | $81,212$ | $80,216$ | $8,800$ |
| $51,400$ | $81,212$ | $83,456$ | $12,040$ |
| $67,800$ | $81,212$ | $83,456$ | $12,040$ |
| $84,200$ | $81,212$ | $86,716$ | $15,300$ |

Table 2: Work decomposition for the five boundary-volume instances. The last column sums the degree-weighted contributions outside $\widehat{S}_A$ from both $y_k$ and $x_{k+1}$ over the executed iterations, using the same accounting as total work.

In the five instances in Table 2, $\widehat{S}_A$ is the 60-vertex core. ISTA remains within this set and requires total work $81{,}212$ in every instance. For FISTA, the table reports total work in the column headed "Work FISTA" and the contribution from vertices outside $\widehat{S}_A$ in the last column. Subtracting the latter from the former gives the work inside $\widehat{S}_A$, which equals $71{,}416$ in all five instances and is about $12\%$ below ISTA's total. FISTA also transiently activates up to 20 boundary vertices outside $\widehat{S}_A$. The work contributed by these vertices increases from zero to $15{,}300$ as the boundary volume grows from $28{,}440$ to $84{,}200$. Consequently, FISTA's total work increases from $71{,}416$ to $86{,}716$, while its contribution from vertices inside the core remains constant. This additional boundary work accounts for the entire increase in FISTA's total work and makes it more expensive than ISTA on the instances with larger boundary volumes. All recorded activations remain within the core or its boundary.

### H.3 Confinement measurements on SNAP graphs

For each seed, we compute the reference support $\widehat{S}_B$ by running ISTA on $F_{\rho/2}$ with residual tolerance $10^{-8}$. We then run FISTA on $F_\rho$, recording $\widehat{A}_k = \mathrm{supp}(x_{k+1}) \setminus \widehat{S}_B$ and its volume at each iteration, and tracking the extrapolated points $y_k$ likewise. The table reports whether all $\widehat{A}_k$ are empty and the volume of $\bigcup_k \widehat{A}_k$, which records iterate activations. The work share sums outside-reference support volumes from both $x_{k+1}$ and $y_k$ over iterations, divided by total FISTA work.

| Dataset | $\alpha$ | Median $\mathrm{vol}(\widehat{S}_B)$ | Seeds with $\widehat{A}_k = \emptyset\ \forall k$ | Median $\mathrm{vol}(\bigcup_k \widehat{A}_k)$* | Median work share* |
|---|---|---|---|---|---|
| Amazon | $0.2$ | $761$ | $100\%$ | n/a | n/a |
| DBLP | $0.2$ | $1,067$ | $100\%$ | n/a | n/a |
| Youtube | $0.2$ | $1,646$ | $100\%$ | n/a | n/a |
| Orkut | $0.2$ | $3,051$ | $100\%$ | n/a | n/a |
| Amazon | $10^{-3}$ | $12,568$ | $18\%$ | $315$ | $0.16\%$ |
| DBLP | $10^{-3}$ | $10,070$ | $9\%$ | $402$ | $0.28\%$ |
| Youtube | $10^{-3}$ | $6,518$ | $20\%$ | $243$ | $0.39\%$ |
| Orkut | $10^{-3}$ | $5,102$ | $86\%$ | $40$ | $0.14\%$ |

Table 3: Confinement diagnostics at $\rho = 10^{-4}$ and residual tolerance $10^{-8}$. Dataset names abbreviate the corresponding `com-` graphs. The baseline uses 300 seeds per graph; the $\alpha = 10^{-3}$ stress test uses the first 100 of these seeds. Columns marked * give medians among seeds with at least one nonempty $\widehat{A}_k$. The entry n/a indicates that there are no such seeds.

At $(\alpha, \rho) = (0.2, 10^{-4})$, no recorded iterate or extrapolated point activates a vertex outside $\widehat{S}_B$ in any of the $1{,}200$ runs. At this operating point, FISTA reduces mean work by factors of 2.3, 2.2, 1.8, and 1.2 on Amazon, DBLP, Youtube, and Orkut, respectively. In the $\alpha = 10^{-3}$ stress test, outside-reference activations occur on $82\%$, $91\%$, $80\%$, and $14\%$ of seeds, respectively. Among

those seeds, the median union volume is 40–402; the maximum is 3,559. The measured work share remains below 3.5%, with medians of 0.1%–0.4% among seeds with activations.

### H.4 Extended regularization sweeps

We extend the real-data grid to $\rho \in [10^{-4}, 10]$ on all four SNAP graphs, giving 19,200 runs per method. We use $\alpha = 0.2$, residual tolerance $10^{-8}$, and the same 300 sampled non-isolated seed vertices per dataset. Table 4 summarizes the large-$\rho$ results.

| Dataset | Zero solutions | Median support volume | Mean work ISTA/FISTA | Zero at tested $\rho \geq 1$ |
|---|---|---|---|---|
| com-Amazon | 8.7% | 4 | 137/125 | 100% |
| com-DBLP | 18.3% | 3 | 125/110 | 100% |
| com-Youtube | 7.0% | 1 | 99/83 | 100% |
| com-Orkut | 87.3% | 0 | 20/19 | 100% |

Table 4: Extended regularization sweep on the SNAP graphs. The first three numerical columns are at $\rho = 0.1$; the last gives the zero-solution fraction at each tested $\rho \geq 1$. Support volumes use the terminated ISTA iterates. They are medians over seeds; work values are means. All summaries include zero solutions.

For $\rho \in \{1, 2.15, 4.64, 10\}$, with the interior grid values rounded here, every seed gives a computed zero solution and both methods report zero iterations and zero work. At $\rho = 0.1$, the computed solution is zero exactly for seeds of degree at least 10. Throughout the sweep, computed zero solutions coincide with $\rho d_{\text{seed}} \geq 1$. No computed support at $\rho = 0.1$ has volume above 9, every nonzero support contains one to three vertices, and the mean work of either method is at most 137 on every graph. Consequently, this regime provides little information about large-scale work complexity.

The original dense-core synthetic regularization sweep already crosses its zero-solution transition: the seed degree is 79, giving a transition at $1/79 \simeq 1.27 \cdot 10^{-2}$. The extension records zero work at tested points beyond it in all nine realizations. Together with the original real-data sweep, the additional runs cover $\rho$ from $10^{-6}$ to 10.

### H.5 Prefix-support and local checking-cost diagnostics

Coverage and excess are the checkpoint support volumes inside and outside the final FISTA support, respectively, each divided by final support volume. Work fractions divide cumulative checkpoint work by completed-run work. At the same baseline parameters and seeds, after 10 iterations the iterate support covers a median 100% of the final FISTA iterate's support volume on all four graphs. The lower quartile is at least 98.9%, median excess is zero, and the median work consumed is 32%–59% of the completed run's work. The median stopping iteration is 16–23. After 5 iterations, median support-volume coverage is 76%–100%, at 7.7%–24% of the work.

Using the tolerance-$10^{-8}$ ISTA reference support $\widehat{S}_B$ for $F_{\rho/2}$, the estimated local scan cost is $\text{vol}(\widehat{S}_B) + \text{vol}(\partial \widehat{S}_B)$. The median estimates are $2.6 \cdot 10^3$, $8.9 \cdot 10^3$, $2.0 \cdot 10^5$, and $3.9 \cdot 10^5$ adjacency accesses on Amazon, DBLP, Youtube, and Orkut, respectively. These are 0.28%, 0.85%, 6.6%, and 0.34% of the respective graph edge counts. The scan excludes candidate construction and the ISTA reference solve.

## I AI-assisted development and prompt traceability

This paper was developed with the assistance of an interactive large-language-model (LLM) workflow. The LLM was used as a proof-synthesis and rewriting aid: it generated candidate lemmas, algebraic manipulations, and LaTeX skeletons, while the human author(s) provided the research direction, imposed algorithmic constraints, requested specific locality-aware bounds, identified missing assumptions, and validated (or rejected) intermediate arguments. The final statements and proofs appearing in the paper were human-checked and edited for correctness and presentation.

### I.1 Prompt clusters and how they map to results in the paper

The interactive prompting that led to the final results naturally grouped into a small number of "prompt clusters." Below we summarize each cluster, the key human supervision intervention(s), and the resulting manuscript artifacts (with cross-references). We use GPT-5.2 Pro (extended thinking) for all results and experiments.

**(P1) "Standard accelerated algorithm only; avoid expensive subproblems."** The initial constraint was to analyze classic one-gradient-per-iteration acceleration (FISTA) rather than outer-inner schemes or methods that solve expensive auxiliary subproblems. This constraint fixed the algorithmic object of study and ruled out approaches akin to expanding-subspace or repeated restricted solves. It directly shaped the scope of the main runtime result Theorem 4.3 and the fact that all bounds are expressed in the degree-weighted work model.

**(P2) "Use the margin/KKT slack idea."** This idea was suggested by GPT, but we found it useful and, therefore, retained it in the final results. A key prompt requested a self-contained argument based on a margin parameter. This produced the degree-normalized slack definition (5) and its operational meaning: an inactive coordinate can become active at an extrapolated point only if its forward-gradient map deviates from the optimum by an amount proportional to its slack. The corresponding quantitative statement is Lemma 4.1, which is the main bridge from optimality structure to spurious activations.

**(P3) "Transient support is the bottleneck; bound the sum of supports/volumes."** A crucial human intervention was to point out that it is not enough to argue eventual identification: one must control the cumulative degree-weighted work over the entire transient. This prompted the transition from pointwise identification to a global summation argument: Cauchy-Schwarz converts "activation implies a jump" (from Lemma 4.1) into a bound on $\sum_k \mathrm{vol}(A_k)$, and geometric contraction of FISTA controls the resulting series. This is the backbone of the spurious-work bound in the proof of Theorem 4.3 (see in particular the derivation around (8)).

**(P4) "Avoid vacuous bounds when the minimum margin is tiny; use over-regularization."** Another human-directed prompt asked how to proceed when the minimum slack can be arbitrarily small, which would make any bound that depends on $\min_{i \in I^\star} \gamma_i$ meaningless. Thus the idea for analyzing a more regularized problem ("(A)") but treating nearly-active nodes as part of the target support of the less-regularized problem ("(B)") was suggested to the LLM. Concretely, this yielded the split in Lemma 4.2, which uses regularization-path monotonicity (cf. Lemma A.4) to show that "small (A)-margin" nodes must lie in $S_B$ and should not be charged as spurious. This is a key input to the work bound Theorem 4.3.

**(P5) "Turn the work bound into a running-time bound using $\mathrm{vol}(S^\star) \le 1/\rho$."** A prompt explicitly requested that the final complexity be stated in the degree-weighted work model and use the known sparsity guarantee $\mathrm{vol}(S^\star) \le 1/\rho$. This guided the decomposition (4) into "work on the target support" plus "spurious work," and it is the reason the first term in Theorem 4.3 scales as $\widetilde{O}((\rho\sqrt{\alpha})^{-1})$ (up to logarithms).

**(P6) "Give an explicit confinement condition so spurious activations stay local."** After the spurious-work summation bound was obtained, a prompt requested a graph-explicit assumption guaranteeing that all spurious activations remain confined to a boundary set. This produced the exposure/no-percolation-style sufficient condition formalized as Theorem 4.4, which is referenced immediately after Theorem 4.3 to justify the boundary-set hypothesis $\widetilde{A}_k \subseteq \mathcal{B}$.

**(P7) "Identify explicit bad instances where $\gamma$ can be very small (or $0$)."** To stress-test the margin-based reasoning, a sequence of prompts asked for explicit graphs where the slack is smaller than $\sqrt{\rho}$ and even $o(\sqrt{\rho})$. This led to the breakpoint constructions recorded in Section C, including the star graph (Lemma C.1) and the path graph (Lemma C.2). These examples motivate why the paper avoids global dependence on $\gamma$ and instead relies on the over-regularization/two-tier strategy (Lemma 4.2) together with confinement (Theorem 4.4).

**(P8) "High-degree non-activation under over-regularization."** A later prompt suggested to use the overregularization idea to rule out spurious activations of very high-degree nodes. This yielded the explicit degree cutoff condition in Proposition B.1, which provides an additional structural non-activation guarantee that complements the boundary-confinement approach.

**(P9) "Experiments."** All code was generated by the LLM. However, the authors heavily supervised the process.

### I.2 How much human supervision was required?

The development required human-in-the-loop supervision. Across roughly two dozen interactive turns, the human prompts performed tasks that the LLM did not do reliably on its own:

- **Problem framing and constraints.** The human author fixed the algorithmic scope (standard FISTA; no expensive subproblems) and demanded a locality-aware work bound rather than a standard iteration bound (driving Theorem 4.3).
- **Identifying the real bottleneck.** A key correction was the insistence that bounding eventual identification is insufficient; one must bound the sum of transient supports/volumes (leading to the summation argument in the proof of Theorem 4.3).
- **Stress-testing with counterexamples.** The human prompts requested explicit worst cases (star and path) and used them to diagnose when naive $\gamma$-based bounds become vacuous (motivating Section C and the different regularizations strategy used in Lemma 4.2).
- **Assumption checking and proof repair.** When an intermediate proof relied on an unproven positivity/sign assumption, the human author demanded either a proof or a repair; this resulted in a revised subgradient/KKT-based certificate (ultimately not needed for the core theorems, but an important correctness checkpoint).
- **LaTeX integration/debugging.** Compile errors and presentation issues (e.g., list/itemization mistakes) were identified via human compilation and corrected in subsequent iterations.

Overall, the LLM contributed most effectively as a rapid generator of candidate proofs and algebraic manipulations, while the human supervision was essential for (i) setting the right target statement, (ii) insisting on the correct work metric, (iii) enforcing locality constraints, (iv) catching missing assumptions, and (v) selecting which generated material belonged in the final paper.

## NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

**The checklist answers are an integral part of your paper submission.** They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes].

   Justification: The claims made exactly reflect our theoretical contributions.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We mention several times that the final upper bound is conditional on the confinement assumption. We also mention in our conclusion that this work is limited to FISTA, and that other algorithms remain an open problem.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.

- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: We do exactly what the question asks.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The main text and experimental appendices specify the protocols, seeds, tolerances, and work accounting. The supplementary package includes the original experiments, additional diagnostic scripts, saved measurements, and reproduction instructions.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.

- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The supplementary ZIP includes the updated experiment code, saved diagnostic results, and a README with dependencies, commands, experiment/table mappings, and access instructions for the public SNAP datasets.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (*https://neurips.cc/public/guides/CodeSubmissionPolicy*) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (*https://neurips.cc/public/guides/CodeSubmissionPolicy*) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The main text and appendix specify the synthetic graph construction, parameter grids, stopping criterion, fixed-point residual tolerance, iteration cap, momentum parameter, seed choice, degree-weighted work metric, and the number of randomized realizations used for the synthetic $\rho$-sweeps. The real-data section specifies the SNAP datasets, number of random seed nodes, parameter grids, and aggregation convention. Section H specifies the additional sweeps, numerical reference supports, and the stress test using the first 100 of the original 300 seeds.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   For the real-data sweeps, curves report the mean over 300 randomly sampled seed nodes and shaded bands show the interquartile range (25%–75%). For the synthetic $\rho$-sweeps at fixed $|\mathcal{B}| = 600$, both the dense-core and sparse-core panels report the mean over 9 randomized realizations per $\rho$, with shaded bands showing the interquartile range (25%–75%). The remaining synthetic sweeps are deterministic controlled sweeps, so stochastic error bars are not applicable there. The confinement table identifies medians conditional on excursions. Prefix diagnostics report medians and lower quartiles across the same 300 seeds per graph; these are descriptive sample summaries.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: The reported metric is degree-weighted work and is independent of hardware. For reproducibility, the original experiments were run on a MacBook Pro with Apple M3 Max, with 48GB of RAM, using Python 3.11.5. The original synthetic experiments took approximately 30 minutes, and the original real-data sweeps took approximately 2 hours. No GPU was used. The environment used to reproduce the prefix diagnostics is recorded in the supplementary package.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics *https://neurips.cc/public/EthicsGuidelines*?

Answer: [Yes]

Justification: We do exactly what the question states.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [No]

Justification: Our work is theoretical and we don't believe it has a direct path to negative societal impact.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.

- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Our paper is theoretical. It poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We use datasets from the SNAP repository, which is cited in the main paper. The real datasets are the public SNAP com-Amazon, com-DBLP, com-Youtube, and com-Orkut graphs; we cite SNAP and use the data according to the SNAP dataset terms. We do not redistribute modified datasets in the paper PDF.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, *paperswithcode.com/datasets* has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

Justification: We attach a supplementary code package containing the updated synthetic and real-data scripts, the additional experimental diagnostics and saved results, and reproduction instructions with dependency, environment, and data-access information. The submission does not introduce a new dataset or model.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing, nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: Yes, an LLM played a significant role in this paper. We report all details in the appendix.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.