HITS can converge slowly, but not too slowly, in score and rank

Enoch Peserico, Luca Pretto - D.E.I., Univ. Padova, Italy
{enoch,pretto}@dei.unipd.it

Abstract

This paper explores the fundamental question of how many iterations the celebrated HITS algorithm requires on a general graph to converge in score and, perhaps more importantly, in rank (i.e. to "get right" the order of the nodes). We prove upper and almost matching lower bounds. We also extend our results to weighted graphs.

1 Introduction

How many iterations does HITS require on a general graph to converge in score and, perhaps more importantly, in rank? This introduction briefly motivates the question in the context of existing literature (Subsection 1.1), and provides a brief description of our results and of the organization of the rest of the paper (Subsection 1.2).

1.1 Motivation and related work

Kleinberg's celebrated HITS (Hypertext Induced Topic Search) algorithm [17] is one of the most famous [27, 14] link analysis algorithms and probably *the* most widely used outside the context of Web search - making it a reference algorithm for today's link analysis, much like quicksort or heapsort for sorting. HITS was originally proposed to rank Web pages, and is the basis of search engines such as Ask (previously Teoma). It has been subsequently employed, sometimes with small variations, to rank graph nodes in a vast (and growing!) number of application domains, often with little or no connection to Web search: among others topic distillation [5], word stemming [2], automatic synonym extraction in a dictionary [3], item selection [31] and author ranking in question answer portals [15] (see also [7, 20, 28, 26, 16]).

HITS is an iterative algorithm computing for each node of a generic graph an *authority* score at every iteration. Thus, any analysis of its computational requirements must consider the number of iterations required to converge within a sufficiently small distance of a limit score vector (which always exists, [17]). In fact, most applications employ the score vector directly to *rank* the nodes of the target graph. In all these cases it is even more important to understand the number of iterations required by HITS to converge *in rank* - informally to assign scores to nodes, that could be potentially quite different from the limit scores, but that still place all or almost all nodes in the "correct" order. The issue of rank convergence as opposed to score convergence is indeed widely regarded as one of the major theoretical challenges of link analysis [12, 19, 22, 23, 29].

HITS effectively computes the dominant eigenvector of the matrix $\mathbf{A}^T\mathbf{A}$ (where $\mathbf{A}$ is the adjacency matrix of the input graph) using the Power Method [11] and thus it is well known how its speed of convergence in score is tied to the separation of the first and second eigenvalues of that matrix. However, no bounds on this separation are known *for the matrix derived from a graph* - for arbitrary matrices of any fixed size the separation can be arbitrarily small and the convergence rate arbitrarily slow. Perhaps even more importantly, no bounds are known on the convergence of HITS *in rank*: only a few experimental results are available, and only for the Web Graph [17]. Being heavily application-dependent, these provide little information to guide the researcher who would port the famous algorithm to new application domains.

1.2 Our results

This is the first paper providing non-trivial bounds on the convergence rate of HITS, both in score and in rank (some weaker lower bounds can be found in our technical report [30]). In a nutshell, we show that HITS can converge slowly, but not too slowly, both in score and in rank. On the one hand, we show that an exponential number of iterations might be necessary to converge to a ranking (and a score) that is even remotely accurate. On the other hand, we show that an exponential number of iterations is also sufficient - and since the iterative process can be accelerated through a well-known "repeated squaring trick" [18] this entails that the complexity of converging to a result extremely close to the limit ranking (and score) is at most polynomial in the number of nodes n of the graph ($O(n^{4+\mu})\lg(n)$, where $\Theta(n^{2+\mu})$ is the complexity of n by n matrix multiplication). Thus, acceleration by repeated squaring seems both sufficient and, in general, necessary to provide convergence guarantees for HITS on graphs of moderate size. The rest of the paper is organized as follows.

Section 2 briefly reviews HITS and some well-known results tying the convergence of the Power Method to the eigengap of a matrix.

Section 3 formally defines the apparently natural, but extremely slippery notion of convergence in rank (for a more detailed discussion of the topic, see [29]).

Section 4 exploits the structure of the matrix $\mathbf{A}^T\mathbf{A}$ to provide bounds on the separation of its eigenvalues, and translates those bounds into upper bounds on the convergence of the score vector provided by HITS. In particular, we leverage a recent result from polynomial algebra [4] to show that HITS never requires more than $(\lg(\frac{1}{\epsilon})+\lg(n))(wg)^{\Theta(m^2)}$ iterations to converge, both in score and in rank, to a vector within distance ϵ of the limit score vector, on any n-node graph of maximum degree g whose links have integer weights bounded by w, and whose authority connected components [25] have at most $m \leq n$ nodes. We tighten this bound to $(\lg(\frac{1}{\epsilon})+\lg(n))(wg)^{\Theta(m)}$ iterations (with an entirely self-contained proof requiring only some basic grounding in linear algebra) when the dominant eigenvalues of $\mathbf{A}^T\mathbf{A}$ belong to the same irreducible block - this includes the important class of authority connected graphs [25]. In this case, the integrality condition can also be relaxed into one simply requiring the minimum weight to be 1.

These bounds are not as weak as they might appear, for two reasons. First of all, note that one might compute the p^{th} power of an n by n matrix $\mathbf{M}$ with at most $2\lfloor\lg(p)\rfloor$ matrix multiplications using a "repeated squaring trick" - first computing $\mathbf{M}, \mathbf{M}^2, \mathbf{M}^4, \ldots, \mathbf{M}^{2^{\lfloor\lg(p)\rfloor}}$ and then multiplying an appropriate subset of those $\lfloor\lg(p)\rfloor$ matrices [18]. Thus, our upper bounds show that, on any n node graph, the complexity of computing the HITS score vector to any precision up to $\frac{1}{2^{2^{\Theta(n)}}}$ is $O(n^{4+\mu}\lg(n))$ - where $n^{2+\mu}$ with $0 \leq \mu \leq 1$ is the complexity of n by n matrix multiplication - and $O(n^{3+\mu}\lg(n))$ in the case of authority connected graphs. This holds even if the graph's arcs have integer weights bounded by $poly(n)$.

Furthermore, Section 5 almost matches the upper bounds of Section 4 by exhibiting, for all

$s \geq 3$, unweighted authority connected graphs of maximum degree k and $\approx k^3 + 3ks$ nodes that, even after $k^{\Theta(s)}$ iterations, fail to "get right" more than $k+1$ of the top k^2+k ranks, even if one accepts as "correct" the ranking provided not just by the limit score vector, but by any score vector at distance less than $\bar{\epsilon} = \Theta(\frac{1}{k\sqrt{k}})$ from it. This also implies lack of convergence to a distance less than $\bar{\epsilon}$ of the limit score vector. In other words, HITS fails not only to get the score error below a relatively large value (since the score vector of HITS is normalized in $\|\cdot\|_2$, the $(k^2)^{th}$ largest component itself can be no larger than $1/k$), but fails also to "get right" more than a small fraction of the top ranks unless allowed to run for exponentially many iterations.

Section 6 summarizes our results, discusses their significance, and briefly reviews some important problems this paper leaves open - before concluding with the bibliography and an appendix with the full proof of the results of Section 5.

2 HITS

This section briefly describes the original HITS algorithm designed for the Web graph [17] (subsection 2.1) summarizing the most important mathematical results in the literature (subsection 2.2, see [13, 1, 10] for more details). Many preprocessing heuristics, typically dependent on the application domain, have been proposed to modify the target graph (e.g. the removal of all intradomain links as biased conferrals of authority [17]). One should then interpret our analysis as applying to the resulting graph.

2.1 The algorithm

The original version of HITS works as follows. In response to a query, a search engine first retrieves a set of nodes of the Web graph on the basis of pure textual analysis; for each such node it also retrieves all nodes pointed by it, and up to d nodes pointing to it.

HITS operates on the subgraph induced by this *base set* (which obviously depends on the query) associating to each node v_i an *authority* score a_i that summarizes both its quality and its relevance to the query, as well as an ancillary *hub* score h_i, according to the iterative formulas:

$$h_i^{(0)} = 1 \qquad a_i^{(t)} = \sum_{v_j \to v_i} h_j^{(t-1)} \qquad h_i^{(t)} = \sum_{v_i \to v_j} a_j^{(t)}, \qquad t = 1, 2, \ldots$$

where $v \to u$ denotes that v points to u.

More precisely, at each step the authority and hub vector of scores are normalized in $\|\cdot\|_2$ (this is well defined assuming, as we shall do throughout the rest of the paper, that the subgraph induced by the base set has at least one arc). Then the authority vector $\mathbf{a}^{(t+1)}$, whose i^{th} element $a_i^{(t+1)}$ corresponds to the authority of node i at timestep $t+1$, can be computed from the adjacency matrix $\mathbf{A}$ of the base set subgraph:

$$\mathbf{a}^{(t+1)} = \frac{(\mathbf{A}^T\mathbf{A})^t\mathbf{A}^T\mathbf{1}}{\|(\mathbf{A}^T\mathbf{A})^t\mathbf{A}^T\mathbf{1}\|_2} \tag{1}$$

where $\mathbf{1}$ is the vector $[1 \ldots 1]^T$.

2.2 Convergence of the authority vector

Equation (1) shows that HITS is essentially computing a dominant eigenvector of $\mathbf{A}^T\mathbf{A}$ using the iterative Power Method ([11]) starting from the initial vector $\mathbf{A}^T\mathbf{1}$; since $\mathbf{A}^T\mathbf{A}$ is symmetric and positive semidefinite the convergence to a limit vector is guaranteed ([11]).

It is well known, and easy to verify, that the error of approximating the limit vector with $\mathbf{a}^{(t)}$ is tied both to the gap between the largest and second largest eigenvalues of $\mathbf{A}^T\mathbf{A}$, and to the modulus of the projection of the initial vector $\mathbf{A}^T\mathbf{1}$ on the dominant eigenspace.

Since $\mathbf{A}^T\mathbf{A}$ is symmetric and positive semidefinite, its eigenvectors form an orthonormal base and its eigenvalues are all real and non-negative. Denote by $\lambda_1, \ldots, \lambda_m$ the m distinct eigenvalues of $\mathbf{A}^T\mathbf{A}$, with $\lambda_i > \lambda_{i+1}$. Then one can write:

$$\mathbf{A}^T\mathbf{1} = \alpha_1\mathbf{v_1} + \cdots + \alpha_m\mathbf{v_m}$$

where $\mathbf{v_i}$ is a normalized eigenvector relative to λ_i, and $\alpha_1 > 0$ ([1]). Therefore

$$\frac{(\mathbf{A}^T\mathbf{A})^t\mathbf{A}^T\mathbf{1}}{\|(\mathbf{A}^T\mathbf{A})^t\mathbf{A}^T\mathbf{1}\|_2} = \frac{\mathbf{v_1} + \frac{\alpha_2\lambda_2^t}{\alpha_1\lambda_1^t}\mathbf{v_2} + \cdots + \frac{\alpha_m\lambda_m^t}{\alpha_1\lambda_1^t}\mathbf{v_m}}{\|\mathbf{v_1} + \frac{\alpha_2\lambda_2^t}{\alpha_1\lambda_1^t}\mathbf{v_2} + \cdots + \frac{\alpha_m\lambda_m^t}{\alpha_1\lambda_1^t}\mathbf{v_m}\|_2}$$

and the last term obviously converges to $\mathbf{v_1}$ since $\lim_{t\to\infty} \frac{\alpha_i\lambda_i^t}{\alpha_1\lambda_1^t} = 0$ for $i > 1$.

The 2-norm of the error vector $\bar{\mathbf{v}}^t$ after t steps (assuming $\mathbf{A}$ has order $n \geq m$ and none of its columns adds up to more than r) is then equal to:

$$\begin{aligned}\|\bar{\mathbf{v}}^t\|_2 &= \left\| \frac{\mathbf{v_1} + \frac{\alpha_2\lambda_2^t}{\alpha_1\lambda_1^t}\mathbf{v_2} + \cdots + \frac{\alpha_m\lambda_m^t}{\alpha_1\lambda_1^t}\mathbf{v_m}}{\|\mathbf{v_1} + \frac{\alpha_2\lambda_2^t}{\alpha_1\lambda_1^t}\mathbf{v_2} + \cdots + \frac{\alpha_m\lambda_m^t}{\alpha_1\lambda_1^t}\mathbf{v_m}\|_2} - \mathbf{v}_1 \right\|_2 \\ &= \left\| \frac{\mathbf{v_1} + \sum_{i=2}^m \frac{\alpha_i}{\alpha_1}(\frac{\lambda_i}{\lambda_1})^t\mathbf{v_i} - (1 + \sum_{i=2}^m(\frac{\alpha_i}{\alpha_1})^2(\frac{\lambda_i}{\lambda_1})^{2t})^{\frac{1}{2}}\mathbf{v_1}}{(1 + \sum_{i=2}^m(\frac{\alpha_i}{\alpha_1})^2(\frac{\lambda_i}{\lambda_1})^{2t})^{\frac{1}{2}}} \right\|_2 \\ &\leq (2\sum_{i=2}^m(\frac{\alpha_i}{\alpha_1})^2(\frac{\lambda_i}{\lambda_1})^{2t})^{\frac{1}{2}} \leq \frac{(2n)^{\frac{1}{2}}}{\alpha_1}(\frac{\lambda_2}{\lambda_1})^t r \qquad (2)\end{aligned}$$

where the last inequality follows from the fact that $(\sum_{i=1}^m \alpha_i^2)^{\frac{1}{2}} = \|\mathbf{A}^T\mathbf{1}\|_2 \leq r\|\mathbf{1}\|_2 = rn^{1/2}$. It may then seem that the convergence of HITS is well understood; this is not the case, as we shall see in the next sections.

3 Convergence in Score vs. Convergence in Rank

All the most popular link analysis algorithms iteratively compute a score for each node of the input graph. In many applications (e.g. [6, 8, 2, 17, 24]) this score vector is used alone to rank the nodes - whether because no other scores might be reasonably combined with it (e.g. in web crawling, word stemming, automatic construction of summaries), or because the algorithm operates on a query dependent graph already capturing the relevance of each node to the query at hand (as in the case of HITS as opposed to e.g. PageRank). In these cases the speed of convergence in score of the algorithm is less crucial than the speed of *convergence in rank* - informally how many iterations are required to rank the nodes in the "correct" order (that induced by the limit score vector).

This informal definition suffers from at least two major flaws. First, it fails to explicitly deal with ties or "almost ties" in score: if the difference between the limit scores of two nodes is negligible, an algorithm effectively converges to a "correct" ranking even if it keeps switching the relative positions of the two nodes. Second, the definition above fails to distinguish between an algorithm that takes a long time to reach the ultimate ranking, but quickly reaches a ranking "close" to it (e.g. with all elements correctly ranked, save the last few), and an algorithm that simply fails for a long time to produce a ranking even remotely close to the ultimate ranking. In this regard, the top ranks are typically much more important than the last ones: failing to converge on the last 20 ranks is almost always far less of a problem than failing to converge on the top 20.

To address these two issues we introduce a more general and formal definition of convergence in rank. We begin by formalizing the notion of ranking induced by a score vector:

Definition 1. *Given a score vector* $\mathbf{v} = [v_1, \ldots, v_n]$, a ranking ρ compatible with $\mathbf{v}$ *is an ordered n-tuple* $[i_1, \ldots, i_n]$ *containing each integer between* 1 *and n exactly once, such that* $\forall j \; v_{i_j} \geq v_{i_{j+1}}$.

Informally, a ranking is compatible with a score vector if no node that has a higher score is ranked worse than one with a lower score (ties can be broken arbitrarily).

Definition 2. *Consider an iterative algorithm ALG producing at each iteration t a score vector* $\mathbf{v}(t)$ *and converging to a score vector* $\mathbf{v}(\infty)$*. Then ALG* ϵ*-converges on h of the top k ranks in (at most)* τ *steps if, for all iterations* $t \geq \tau$*, at least h of the top k items in a ranking compatible with* $\mathbf{v}(t)$ *are also among the top k items in a ranking compatible with* $\mathbf{v}(\infty)$*, or compatible with some vector* $\mathbf{w}(t)$ *at distance at most* ϵ *from* $\mathbf{v}(\infty)$*.*

In other words, we assume an algorithm has converged in ranking as soon as it "gets right" (and keeps getting right) at least h of the top k items of any ranking compatible with the limit score vector, or with any score vector "sufficiently close" to it (note that the definition above implicitly assumes some distance function between score vectors - e.g. $\|\cdot\|_2$).

Our definition is related to the notion of intersection metric [9]. The distance in the intersection metric of two rankings that share $h(k)$ of the top k items is the average over k of $1 - h(k)/k$. In our definition we do not "summarize" the size of the intersection between the top k ranks of the current and limit rankings, instead leaving k as a parameter. Furthermore, we consider "acceptable" a whole set of limit rankings induced by "sufficiently close" score vectors.

It is important to note that, with $\epsilon = 0$ and $h = k \forall k$ our definition collapses back to the stricter, "naive" definition of convergence in rank; and that if an algorithm ϵ-converges in score in t iterations (i.e. if the score vector after t iterations always remains within distance ϵ of the limit vector) then that algorithm also ϵ-converges on all its ranks (i.e. on k of its top k ranks $\forall k$) in t steps - but the reverse is not necessarily true.

4 Upper Bounds

While HITS can take many iterations to converge, either in score or in rank (see Section 5), in general it can not take *too many.* This section provides upper bounds on this number (in subsection 4.2), by first proving lower bounds on the separation between the eigenvalues of $\mathbf{A}^T\mathbf{A}$ (in subsection 4.1). These separation bounds are yielded by two completely different proof techniques (both potentially of independent interest); which technique is applicable depends on subtle hypotheses both on the structure of the graphs and on the nature of the link weights. Subsection 4.3 sheds some light on this issue.

4.1 Some novel bounds on eigenvalue separation

Assume without loss of generality that $\mathbf{A}^T\mathbf{A}$ is a block matrix - this can always be achieved through appropriate row and column transpositions that correspond to a simple renaming of the nodes. The eigenvalue separation bounds we prove in this subsection depend on whether the two largest eigenvalues are dominant eigenvalues of two different blocks of $\mathbf{A}^T\mathbf{A}$, or if one isn't. In the latter situation (Lemma 2), the bounds are tighter than in the former (Lemma 1). Note that the latter situation includes the important class of *authority connected graphs* [25]: informally, these are graphs where, for every pair of nodes v and v' with positive indegree, one can reach v from v' by first following a link backwards, then following a link forward, then again a link backwards - and so on.

Lemma 1. *Let $\mathbf{B_1}$ and $\mathbf{B_2}$ be two integer, symmetric, non-negative and positive semidefinite m_1 by m_1 and m_2 by m_2 matrices, with no row adding up to more than (respectively) r_1 and r_2, whose dominant eigenvalues are (respectively) λ_1 and $\lambda_2 < \lambda_1$. Then $\frac{\lambda_1}{\lambda_2} \geq 1 + 2^{1-m_2}(m_2+1)^{\frac{1}{2}-m_1}(m_1+1)^{-\frac{m_2}{2}} r_1^{-m_1 m_2} r_2^{-m_1 m_2 - 1} = 1 + r_1 r_2^{-\Theta(m_1 m_2)}$.*

Proof. The eigenvalues of $\mathbf{B_1}$ and $\mathbf{B_2}$ are the roots of their characteristic (integer) polynomials $P_1(\lambda)$ and $P_2(\lambda)$. The coefficients of $P_i(\lambda)$ are bounded by $r_i^{m_i}$, since the determinant (of order m_i) of $\mathbf{B_i}$ is computed as the weighted sum of m_i determinants of order $m_i - 1$, with weights adding up to at most r_i (and order 1 determinants also bound by r_i). Since λ_1 is not a root of $P_2(\lambda)$, then $\lambda_1 - \lambda_2 \geq 2^{1-m_2}(m_2+1)^{\frac{1}{2}-m_1}(m_1+1)^{-\frac{m_2}{2}} r_1^{-m_1 m_2} r_2^{-m_1 m_2}$ ([4]) and $\frac{\lambda_1}{\lambda_2} = 1 + \frac{\lambda_1 - \lambda_2}{\lambda_2} \geq 1 + \frac{\lambda_1 - \lambda_2}{r_2} = 1 + 2^{1-m_2}(m_2+1)^{\frac{1}{2}-m_1}(m_1+1)^{-\frac{m_2}{2}} r_1^{-m_1 m_2} r_2^{-m_1 m_2 - 1}$. □

Lemma 2. *Let $\mathbf{B}$ be a symmetric, irreducible, positive semidefinite m by m matrix, whose non-zero elements are all at least 1. Denote by λ_1 and λ_2, respectively, the first and second eigenvalue of $\mathbf{B}$. If $\lambda_2 \neq 0$, then $\lambda_1 \geq \lambda_2(1 + 2\lambda_2^{-2m})^{\frac{1}{2m}}$.*

Proof. In a nutshell, any non-dominant eigenvector $\mathbf{v}$ has both negative and positive components, which partially cancel out when $\mathbf{v}$ is multiplied by $\mathbf{A}^T\mathbf{A}$; this reduces the corresponding (non-dominant) eigenvalue by an amount that we bound away from 0 to obtain the thesis.

To formalize this intuition, we first need some notation. Given a d dimensional column vector $\mathbf{w} = [w_1, \ldots, w_d]^T$, let $\mathbf{w}^+ = [w_1^+, \ldots, w_d^+]^T$ with $w_i^+ = w_i$ if $w_i > 0$ and $w_i^+ = 0$ otherwise; let $\mathbf{w}^- = [w_1^-, \ldots, w_d^-]^T$ with $w_i^- = |w_i|$ if $w_i < 0$ and $w_i^- = 0$ otherwise (so that $\mathbf{w} = \mathbf{w}^+ - \mathbf{w}^-$); and let $\mathbf{w}^\pm = \mathbf{w}^+ + \mathbf{w}^-$. Given two vectors $\mathbf{u} = [u_1, \ldots, u_d]^T$ and $\mathbf{v} = [v_1, \ldots, v_d]^T$ write $\mathbf{u} \geq \mathbf{v}$ if $u_i \geq v_i \geq 0$ for every i.

Let $\mathbf{v}$ and $\mathbf{w}$ be two eigenvectors of $\mathbf{B}$ corresponding, respectively, to λ_1 and λ_2; assume without loss of generality that $\|\mathbf{v}\| = \|\mathbf{w}\| = 1$ and that all elements of $\mathbf{v}$ are positive (this is possible by the theorem of Perron-Frobenius, since $\mathbf{B}$ is non-negative and irreducible). By the same theorem $\mathbf{w}$ has at least one positive and at least one negative element. Then $\mathbf{w}^\pm \cdot \mathbf{v} > 0$ and $\lim_{t\to\infty}(\|\mathbf{B}^t\mathbf{w}^\pm\|)^{\frac{1}{t}} = \lim_{t\to\infty}(\|\mathbf{B}^t\mathbf{v}\|)^{\frac{1}{t}} = \lambda_1$.

Let us now consider the difference $\mathbf{u} = [u_1, \ldots, u_m]^T = \mathbf{B}^m\mathbf{w}^\pm - (\mathbf{B}^m\mathbf{w})^\pm$. Denote by $\mathbf{b_i}$ the i^{th} row vector of $\mathbf{B}^m$ (note that $\mathbf{b_i} = \mathbf{b_i}^+$). Then $\mathbf{b_i} \cdot \mathbf{w}^\pm = \mathbf{b_i} \cdot \mathbf{w}^+ + \mathbf{b_i} \cdot \mathbf{w}^-$ whereas

$(\mathbf{b_i} \cdot \mathbf{w})^{\pm} = (\mathbf{b_i} \cdot \mathbf{w}^{+} - \mathbf{b_i} \cdot \mathbf{w}^{-})^{\pm}$. Thus $u_i = \mathbf{b_i} \cdot \mathbf{w}^{\pm} - (\mathbf{b_i} \cdot \mathbf{w})^{\pm} \geq 0$ and u_i is respectively equal to $2\mathbf{b_i} \cdot \mathbf{w}^{-}$ if $\mathbf{b_i} \cdot \mathbf{w}^{+} \geq \mathbf{b_i} \cdot \mathbf{w}^{-} \Leftrightarrow \mathbf{b_i} \cdot \mathbf{w} \geq 0 \Leftrightarrow \lambda_2^m w_i \geq 0$ and to $2\mathbf{b_i} \cdot \mathbf{w}^{+}$ otherwise. Since at least one element of $\mathbf{w}$ is positive and at least one is negative, and every element of $\mathbf{b_i}$ is at least 1 (by the irreducibility of $\mathbf{B}$), then every element of $\mathbf{B}^m\mathbf{u}$ is at least equal to $2\sum_{i=1}^{m} |w_i|$ and $\mathbf{B}^{2m}\mathbf{w}^{\pm} - (\mathbf{B}^{2m}\mathbf{w})^{\pm} \geq \mathbf{B}^{2m}\mathbf{w}^{\pm} - \mathbf{B}^m(\mathbf{B}^m\mathbf{w})^{\pm} = \mathbf{B}^m\mathbf{u} \geq 2\mathbf{w}^{\pm}$.

Since $(\mathbf{B}^{2m}\mathbf{w})^{\pm} = \lambda_2^{2m}\mathbf{w}^{\pm}$, then $\mathbf{B}^{2m}\mathbf{w}^{\pm} \geq (\mathbf{B}^{2m}\mathbf{w})^{\pm} + 2\mathbf{w}^{\pm} = (1 + 2\lambda_2^{-2m})\lambda_2^{2m}\mathbf{w}^{\pm}$. Thus $\lambda_1 = \lim_{t\to\infty}(\|\mathbf{B}^{2mt}\mathbf{w}^{\pm}\|)^{\frac{1}{2mt}} \geq ((1 + 2\lambda_2^{-2m})\lambda_2^{2m})^{t\cdot\frac{1}{2mt}} = \lambda_2(1 + 2\lambda_2^{-2m})^{\frac{1}{2m}}$. □

4.2 Upper bounds on the ϵ-convergence of HITS

Lemmas 1 and 2, as well as Equation (2), allow us to derive upper bounds on the number of iterations required for HITS ϵ-converge in score (in $\|\cdot\|_2$) - and thus on all ranks. In fact, our results are also applicable to weighted graphs, allowing us to deal with modifications of HITS that assign weights to links (e.g. [21, 26]) as long as these weights satisfy some mild conditions.

Theorem 1. *Let G be a graph of n nodes and maximum degree g whose edges have weights at least 1 and at most w. Denoting by $\mathbf{A}$ the weighted adjacency matrix of G, if $\mathbf{A}^T\mathbf{A}$ is a block matrix such that all its non-zero blocks have size at most m and if the largest and the second largest eigenvalues of $\mathbf{A}^T\mathbf{A}$ are relative to the same block (including the case of just one non-zero block, i.e. if G is authority connected), then HITS ϵ-converges in score (in $\|\cdot\|_2$), and therefore on all ranks, on the nodes of G in at most $m(wg)^{4m}(\lg(\frac{1}{\epsilon}) + \frac{1}{2}\lg(2n) + \lg(wg)) = (wg)^{O(m)}(\lg(\frac{1}{\epsilon}) + \lg(n))$ iterations.*

Proof. By Equation (2), if no row of $\mathbf{A}^T$ adds up to more than r, then after t iterations the error is bounded by $\frac{(2n)^{\frac{1}{2}}}{\alpha_1}(\frac{\lambda_2}{\lambda_1})^t r$; thus, unless $\lambda_2 = 0$ (in which case HITS converges after the first iteration) the number of iterations required to converge within distance at most ϵ of the limit score vector is no more than $\frac{\lg(2n)/2+\lg(1/\epsilon)+\lg(1/\alpha_1)+\lg r}{\lg(\lambda_1/\lambda_2)}$. Each block of $\mathbf{A}^T\mathbf{A}$ is symmetric, irreducible and positive semidefinite, every non-zero element is at least 1, and no eigenvalue surpasses $(wg)^2$ (a value bounding the sum of the elements of any row, since each node has incoming links from at most g other nodes, each in turn linking at most g other nodes). Also, λ_1 and λ_2 (unless 0) are relative to the *same* irreducible block. By Lemma 2, then $\lg(\frac{\lambda_1}{\lambda_2}) \geq \frac{1}{2m}\lg(1+2(wg)^{-4m}) \geq \frac{1}{m(wg)^{4m}}$ (since $\lg(1 + x) \geq x$ for $x \leq 1$). Moreover, wg is an upper bound for r. All is left to prove is that $\alpha_1 \geq 1$ and thus $\lg(1/\alpha_1) \leq 0$. The i^{th} element of $\mathbf{A^T1}$ is 0 only if all elements of the i^{th} row of $\mathbf{A}^T$ are 0, and is at least 1 otherwise; and in the former case the i^{th} element of $\mathbf{v_1} = \frac{1}{\lambda_1}\mathbf{A}^T\mathbf{A}\mathbf{v_1}$ is also 0. Then, $\alpha_1 = \mathbf{v_1} \cdot \mathbf{A}^T\mathbf{1} \geq \mathbf{v_1} \cdot \mathbf{1} = \|\mathbf{v_1}\|_1 \geq \|\mathbf{v_1}\|_2 = 1$. □

Theorem 2. *Let G be a graph of n nodes and maximum degree g whose edges have* integer *weights at least 1 and at most w. Denote by $\mathbf{A}$ the weighted adjacency matrix of G. If $\mathbf{A}^T\mathbf{A}$ is a block matrix with at least two blocks of size m_1 and m_2 whose dominant, positive eigenvalues λ_1 and $\lambda_2 < \lambda_1$ are respectively the largest and second largest eigenvalue of $\mathbf{A}^T\mathbf{A}$, then HITS ϵ-converges in score (in $\|\cdot\|_2$), and therefore on all ranks, on the nodes of G in at most $2^{m_2-1}(m_2+1)^{m_1-\frac{1}{2}}(m_1+1)^{\frac{m_2}{2}}(wg)^{2m_1m_2}(wg)^{(2m_1m_2+2)}(\lg(\frac{1}{\epsilon}) + \frac{1}{2}\lg(2n) + \lg(wg)) = (wg)^{O(m_1m_2)}(\lg(\frac{1}{\epsilon}) + \lg(n))$ iterations.*

Proof. The proof is almost identical to that of Theorem 1, the only difference being the looser bound one must use for λ_1/λ_2 (derived from Lemma 1 rather than from Lemma 2). □

4.3 Graph structure and link weights

It is interesting to compare the conditions that Theorems 1 and 2 place on the link weights. Both require links to fall between 1 and some maximum weight w, which is equivalent to requiring a maximum ratio of w between link weights. In addition, Theorem 2 requires the weights to be integers: this enforces a separation between different weights.

It is easy to prove that *the bound on link weights is essential to ensure bounds on convergence, even if a graph is authority connected.* Consider an undirected, unweighted graph G formed of 6 nodes $v_1, \ldots, v_6$, with v_1, v_2 and v_3 forming a 3-clique, v_4 linked (only) to v_1, v_5 to v_2, and v_6 to v_3 (see Figure 1). Denote by $v_i(t)$ the score of v_i at time t (with $v_i(0) = 1/\sqrt{6}$). It is easy to verify by induction on t that, for all $t \geq 1$ we have $3v_4(t) \geq v_1(t) \geq 2v_4(t)$, since $v_4(t+1) = v_1(t)$ and $v_1(t+1) = v_2(t) + v_3(t) + v_4(t) = 2v_1(t) + v_4(t)$.

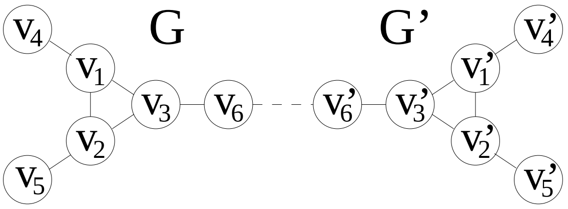

Figure 1: With a weight of ab on all edges of G, a weight of $a(b+1)$ on all edges of G', and a weight 1 on the edge connecting them, sufficiently large integers a and b can make the convergence of HITS arbitrarily slow. Similarly, if G and G' are disconnected, and their edges have weights respectively 1 and $1+\epsilon$, a sufficiently small $\epsilon > 0$ can make convergence arbitrarily slow.

Weight all the links of G with a weight ab, (with a and b positive integers) and all the links of an isomorphic graph G' (with $v'_1, \ldots, v'_6$ respectively corresponding to $v_1, \ldots, v_6$) with a weight $a(b+1)$. Then $v'_4(t) = (\frac{b+1}{b})^t v_4(t)$ and obviously the ratio $\frac{v_i(t)}{v_j(t)}$ remains identical to the unweighted case (and to $\frac{v'_i(t)}{v'_j(t)}$). Thus $v'_4(t) < v_1(t)$ for $(\frac{b+1}{b})^t < 2$ and $v'_4(t) > v_1(t)$ for $(\frac{b+1}{b})^t > 3$; and, while eventually $v'_4(t) > v_1(t)$, a sufficiently large b can delay this for arbitrarily long time. And a sufficiently large a ensures that, even if G and G' are linked by an edge of weight 1, the effects of this link are negligible for an arbitrarily long time.

Note that *if we allow for disconnected graphs, some separation between the weights (e.g. that guaranteed by the integrality constraint) is also necessary*; otherwise one can simply assign weights 1 and $1+\epsilon$ to G and G' (keeping the two graphs disconnected) and, although $v'_4(t)$ eventually surpasses $v_1(t)$, a sufficiently small ϵ guarantees that $v'_4(t)$ remains smaller than $2v_4(t) \leq v_1(t)$ for an arbitrarily long time.

5 HITS Can Converge Slowly in Rank

This section almost matches the upper bounds of the previous one. It is entirely devoted to:

Theorem 3. *For all $k \geq 3$ and $s \geq 3$ there exists an authority connected graph Γ of maximum degree $2k$ and $3(k+1)s + k^3 + 2k^2 + 2k + 2 \approx k^3 + 3ks$ vertices on which HITS fails to ϵ-converge on more than $k+1$ of the top $k^2 + k$ ranks in less than $k^{\Theta(s)}$ iterations for all $\epsilon \leq \bar{\epsilon} = \Theta(\frac{1}{k\sqrt{k}})$.*

Proof. Here we only sketch the proof of the theorem. The details can be found in the Appendix. The graph Γ (see Figure 2) is formed by $k+1$ "flower" subgraphs $G_0, \ldots, G_k$. G_0 is formed by a clique "corolla" of $k+1$ vertices $v_0, \ldots, v_k$, each connected to the other k, as well as to other k "petal" vertices $v_{i,1} \ldots, v_{i,k}$ except in the case of v_0. v_0 is not connected to any petal but instead to the first vertex v_{-1} of a "stem" of $s+1$ vertices $v_{-1}, \ldots, v_{-(s+1)}$ (with v_j connected to v_{j+1} and to v_{j-1}). G_1 is almost isomorphic to G_0, with a vertex v'_i corresponding to each vertex v_i and a vertex $v'_{i,j}$ corresponding to each vertex $v_{i,j}$, the only difference being that $v'_{-(s+1)}$ is missing (i.e. the stem only has s, rather than $s+1$, vertices). G_i for $i > 1$ is isomorphic to G_1. All flowers are strung in a "garland" by connecting the vertices corresponding to v_0 in G_i and $G_{(i+1)\%(k+1)}$ with a "string" of $2s$ vertices. Note that Γ is symmetric and unweighted, strongly connected, and since it certainly holds for $k \geq 3$ a cycle of odd length, it is authority connected.

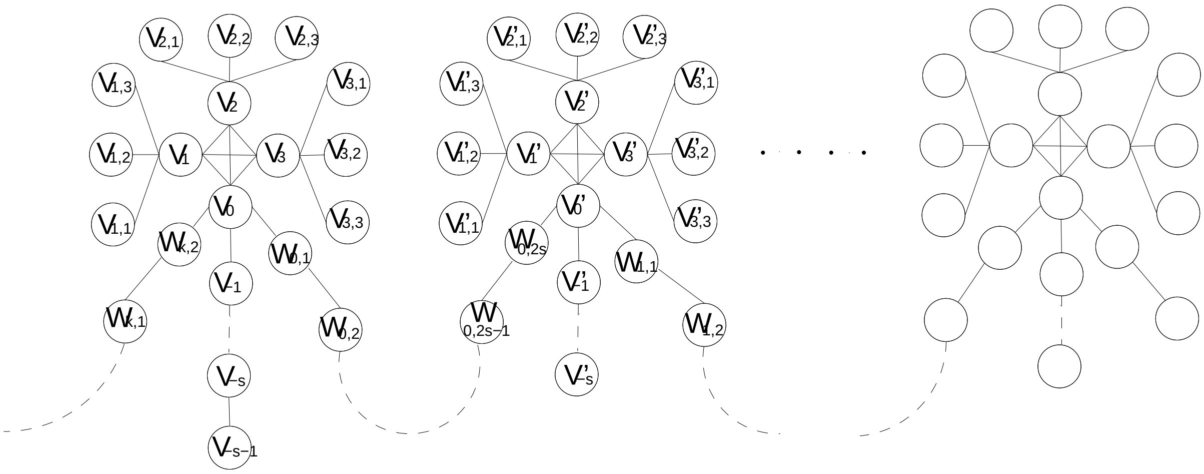

Figure 2: The undirected, unweighted, authority-connected graph Γ for $k=3$. Γ is formed by k "flower" subgraphs $G_1, \ldots G_k$ (right) almost isomorphic to the subgraph G_0 (left, with one more vertex in the stem). Each flower has a corolla of $k+1$ vertices, all but one with k petals, attached to a stem of $\approx s$ vertices. Strings of $2s$ vertices string the flowers into a circular "garland".

After introducing some notation, the proof proceeds as follows. We first consider the simpler graph where no strings are present and each flower stands disconnected from the others. We show that in $\Theta(s)$ iterations the score of the vertices of every flower stabilizes into a configuration where each vertex of the stem has a score $\Theta(k)$ times larger than the score of the vertex below it, and each vertex of the corolla has a score $\Theta(k)$ times larger than the score of each petal and of the top vertex of the stem. Then, we show that, eventually, the contribution of $v_{-(s+1)}$ is sufficient to bring the score of every vertex of G_0 to be arbitrarily larger than that of any vertex of G_1 (and thus of G_i with $i > 0$); but this contribution is sufficiently small that for $k^{\Omega(s)}$ iterations the difference in the scores of any vertex of G_0 and of the corresponding vertex of G_1 remains negligible. Thus, for $k^{\Omega(s)}$ iterations the corollas of the flowers remain the top k^2+k vertices - their scores outstripping that of every other vertex by a factor $\Omega(k)$ - but eventually, almost all the score is concentrated in the corolla and the petals of G_0. A simple calculation shows that any score vector with a (corolla) vertex outside of G_0 in the top k^2+k positions (where for $k^{\Theta(s)}$ steps all corolla vertices belong) differs, in $\|\cdot\|_2$, from the limit score vector by at least $\Theta(\frac{1}{k\sqrt{k}})$. Thus, HITS does not ϵ-converge in score for $k^{\Theta(s)}$ iterations for any ϵ less than some $\bar{\epsilon} = \Theta(\frac{1}{k\sqrt{k}})$. We complete the proof by showing that the strings connecting the different flowers are sufficiently long to bring a negligible contribution to the relative scores of the corollas and petals of different flowers, at least until the score of all vertices in flowers other than G_0 becomes negligibly small - indeed the only purpose of the strings is to show that the theorem also holds in the case of authority-connected graphs, for which the upper bounds provided in Section 4 are tighter. □

6 Conclusions and Open Problems

The vast and growing number of applications of HITS (many with little or no connection to Web search) motivates this paper - the first paper to provide non-trivial upper and lower bounds on the convergence of the celebrated algorithm, both in score and, perhaps more importantly, in rank.

We prove that HITS never requires more than $(\lg(\frac{1}{\epsilon}) + \lg(n))(wg)^{\Theta(m^2)}$ iterations to ϵ-converge both in score and on all ranks on any n node graph of maximum degree g whose authority connected components sport at most m nodes and whose links are weighted with integral weights bounded by w. This bound can be tightened somewhat, to $(\lg(\frac{1}{\epsilon}) + \lg(n))(wg)^{\Theta(m)}$ iterations, if the dominant eigenvalues of $\mathbf{A}^T\mathbf{A}$ (where $\mathbf{A}$ is the adjacency matrix of the graph) belong to the same irreducible block - this includes the important class of authority connected graphs. In this case, the integrality condition can also be relaxed into one simply requiring the minimum weight to be 1.

While these bounds might seem weak, "repeated squaring" acceleration translates them into polynomial upper bounds on the time complexity of reaching up to $\frac{1}{2^{2^{\Theta(n)}}}$ precision, even with $poly(n)$ arc weights: more precisely $O(n^{4+\mu}\lg(n))$ - where $\Theta(n^{2+\mu})$ is the complexity of n by n matrix multiplication - and $O(n^{3+\mu}\lg(n))$ for a large class of graphs, including the important case of authority connected graphs.

Also, we almost match the upper bounds above by exhibiting unweighted authority connected graphs of $\approx k^3+3ks$ nodes and maximum degree $2k$ that fail to ϵ-converge on more than $k+1$ of the top k^2+k ranks (and thus to ϵ-converge in score) even after $k^{\Theta(s)}$ iterations, for all $\epsilon \leq \bar{\epsilon} = \Theta(\frac{1}{k\sqrt{k}})$ - in other words, HITS fails not only to get the score error below than a not-so-small constant, but fails also to "get right" more than a small fraction of the top ranks unless allowed to run for exponentially many iterations.

Thus, employing repeated squaring acceleration seems absolutely necessary to ensure that one can always reach a satisfactory result in a reasonable time. Graphs of up to a few thousand nodes (like those used in web-search - unlike PageRank, HITS typically operates on small sets of pages preselected through pure textual analysis) seem then certainly tractable. Scaling to beyond a million nodes with convergence guarantees is a challenge probably hard to match even for today's most powerful computational platforms - unless the application domain ensures the graphs involved meet some "favorable" structural conditions (e.g. in an n node graph with authority connected components of $polylog(n)$ nodes HITS requires complexity $O(n \cdot polylog(n))$ even without sacrificing high accuracy).

Exploring these conditions is certainly a promising direction for future research. It would also be interesting to understand more in depth what conditions one must enforce on link weights to guarantee convergence. Finally, it would be interesting to understand whether the gap in the convergence upper bounds between authority connected and general graphs is indeed fundamental or simply a weakness of our proof techniques.

References

[1] M. Agosti and L. Pretto. A theoretical study of a generalized version of Kleinberg's HITS algorithm. *Information Retrieval*, 8:219–243, 2005. Special topic issue: Advances in Mathematical/Formal Methods in Information Retrieval.

[2] M. Bacchin, N. Ferro, and M. Melucci. A probabilistic model for stemmer generation. *Information Processing and Management*, 41(1):121–137, Jan. 2005.

[3] M. W. Berry, editor. *Survey of Text Mining: Clustering, Classification, and Retrieval.* Springer, New York, 2004.

[4] Y. Bugeaud and M. Mignotte. On the distance between roots of integer polynomials. *Proceedings of the Edinburgh Mathematical Society*, 47:553–556, Oct. 2004.

[5] S. Chakrabarti, B. E. Dom, D. Gibson, R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Experiments in topic distillation. In *Proceedings of the ACM SIGIR Workshop on Hypertext Information Retrieval on the Web*, pages 117–128, New York, 1998. ACM.

[6] J. Cho, H. Garcia-Molina, and L. Page. Efficient crawling through URL ordering. *Computer Networks*, 30(1–7):161–172, 1998.

[7] Y. Duan, J. Wang, M. Kam, and J. Canny. Privacy preserving link analysis on dynamic weighted graph. *Computational & Mathematical Organization Theory*, 11:141–159, 2005.

[8] G. Erkan and D. R. Radev. LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479, Dec. 2004.

[9] R. Fagin, R. Kumar, and D. Sivakumar. Comparing top k lists. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 28–36. ACM/SIAM, Jan. 2003.

[10] A. Farahat, T. Lofaro, J. C. Miller, G. Rae, and L. A. Ward. Authority rankings from HITS, PageRank and SALSA: Existence, uniqueness and effect of initialization. *SIAM Journal on Scientific Computing*, 27(4):1181–1201, 2006.

[11] G. H. Golub and C. F. Van Loan. *Matrix Computations.* Johns Hopkins Studies in the Mathematical Sciences. The Johns Hopkins University Press, Baltimore, third edition, 1996.

[12] T. H. Haveliwala. Efficient computation of PageRank. Technical report, Stanford University, 1999.

[13] D. Hong and S. Man. Analysis of Web search algorithm HITS. *International Journal of Foundations of Computer Science*, 15(4):649–662, Aug. 2004.

[14] `http://www.cs.uvm.edu/~icdm/` Links 'Top 10 Algorithms' and then '18 Candidates for the Top 10 Algorithms in Data Mining'.

[15] P. Jurczyk and E. Agichtein. HITS on question answer portals: Exploration of link analysis for author ranking. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 845–846. ACM, July 2007.

[16] B. Kimelfeld, E. Kovacs, Y. Sagiv, and D. Yahav. Using language models and the HITS algorithm for XML retrieval. In *Comparative Evaluation of XML Information Retrieval Systems*, pages 253–260. Springer Berlin/Heidelberg, 2007.

[17] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, Sept. 1999.

[18] D. E. Knuth. *The Art of Computer Programming*, volume 2. Addison-Wesley, third edition, 1998. Section 4.6.3.

[19] S. D. Kumar, T. H. Haveliwala, C. D. Manning, and G. H. Golub. Extrapolation methods for accelerating PageRank computations. In *Proceedings of the Twelfth International World Wide Web Conference*, pages 261–270. ACM, May 2003.

[20] O. Kurland and L. Lee. PageRank without hyperlinks: Structural re-ranking using links induced by language models. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 306–313. ACM, Aug. 2005.

[21] O. Kurland and L. Lee. Respect my authority! HITS without hyperlinks, utilizing cluster-based language models. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 83–90. ACM, Aug. 2006.

[22] A. N. Langville and C. D. Meyer. Deeper inside PageRank. *Internet Mathematics*, 1(3):335–380, 2004.

[23] A. N. Langville and C. D. Meyer. *Google's PageRank and Beyond: The Science of Search Engine Rankings.* Princeton University Press, Princeton, 2006.

[24] R. Lempel and S. Moran. SALSA: The stochastic approach for link-structure analysis. *ACM Transactions on Information Systems*, 19(2):131–160, Apr. 2001.

[25] R. Lempel and S. Moran. Rank-stability and rank-similarity of link-based Web ranking algorithms in authority-connected graphs. *Information Retrieval*, 8:219–243, 2005. Special topic issue: Advances in Mathematical/Formal Methods in Information Retrieval.

[26] S. Mizzaro and S. Robertson. HITS hits TREC - exploring IR evaluation results with network analysis. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 479–486. ACM, July 2007.

[27] `http://www.mathunion.org/Prizes/Nevanlinna/Prizewinners.html`.

[28] N. Oyama, Y. Masunaga, and K. Tachi. A diachronic analysis of gender-related Web communities using a HITS-based mining tool. In *Frontiers of WWW Research and Development - APWeb 2006*, pages 355–366. Springer Berlin/Heidelberg, 2006.

[29] E. Peserico and L. Pretto. What does it mean to converge in rank? In S. Dominich and F. Kiss, editors, *Studies in Theory of Information Retrieval*, Alma Mater, pages 239–245, Budapest, Oct. 2007. Foundation for Information Society. Available at: `http://www.dei.unipd.it/~pretto/ICTIR/ictir2007-peserico-pretto.pdf`.

[30] E. Peserico and L. Pretto. The rank convergence of HITS can be slow. *CoRR*, abs/0807.3006, 2008.

[31] K. Wang and M.-Y. T. Su. Item selection by "hub-authority" profit ranking. In O. R. Zaïane and R. Goebel, editors, *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 652–657, July 2002.

Appendix

This appendix is devoted to the proof of Theorem 3. To lighten the burden of the reader, we first prove a simpler version of the theorem yielding a slightly stronger bound on disconnected graphs. We then extend this weaker theorem, slightly weakening the constants involved, to provide bounds for authority connected graphs. In particular, we begin by proving the following:

Theorem 3′. *For all $k \geq 3$ and $s \geq 3$ there exists an unweighted undirected graph G of $k^3 + 2k^2 + 2k + (k+1)s + 2 \approx k^3 + ks$ vertices on which HITS fails to ϵ-converge on more than $k+1$ of the top $k^2 + k$ ranks in less than $k^{\Theta(s)}$ iterations for all $\epsilon \leq \bar{\epsilon} = \Theta(\frac{1}{k\sqrt{k}})$.*

Proof. Consider the unweighted undirected graph G formed by $k+1$ "flower" subgraphs $G_0, \ldots, G_k$ (see Figure 3). G_0 is formed by a clique "corolla" of $k+1$ vertices $v_0, \ldots, v_k$, each connected to the other k, as well as to other k "petal" vertices $v_{i,1} \ldots, v_{i,k}$ except in the case of v_0. v_0 is not connected to any petal but instead to the first vertex v_{-1} of a "stem" of $s+1$ vertices $v_{-1}, \ldots, v_{-(s+1)}$ (with v_j connected to v_{j+1} and to v_{j-1}). G_1 is almost isomorphic to G_0, with a vertex v'_i corresponding to each vertex v_i and a vertex $v'_{i,j}$ corresponding to each vertex $v_{i,j}$, the only difference being that $v'_{-(s+1)}$ is missing (i.e. the stem has only s, rather than $s+1$, vertices). G_i for $i > 1$ is isomorphic to G_1.

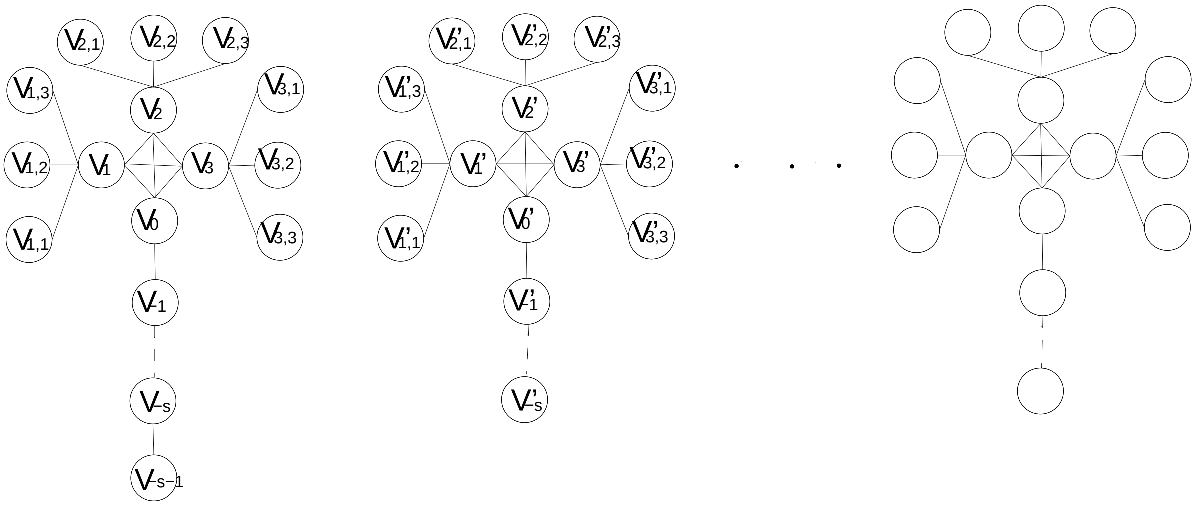

Figure 3: The disconnected graph G for $k = 3$ is formed by a number of "flower" subgraphs almost isomorphic to G_0 (left).

The proof proceeds as follows. After introducing some notation, we show that in $\Theta(s)$ iterations the score of the vertices of every flower stabilizes into a configuration where each vertex of the stem has a score $\Theta(k)$ times larger than the score of the vertex below it, and each vertex of the corolla has a score $\Theta(k)$ times larger than the score of each petal and of the top vertex of the stem. Then, we show that, eventually, the contribution of $v_{-(s+1)}$ is sufficient to bring the score of every vertex of G_0 to be arbitrarily larger than that of any vertex of G_1 (and thus also of G_i with $i > 1$); but this contribution is sufficiently small that for $k^{\Omega(s)}$ iterations the scores of any vertex of G_0 and of the corresponding vertex of G_1 remain within a factor 2 of each other. Thus, for $k^{\Omega(s)}$ iterations the corollas of the flowers remain the top $\approx k^2$ vertices - their scores outstripping that of every other vertex by a factor $\Omega(k)$ - but eventually, almost all the score is concentrated in the corolla and the petals of G_0.

Throughout the theorem, we shall resort to the following *pebble interpretation* of HITS. Initially, every vertex holds one pebble. Then, at every step, a pebble on a vertex v sires a pebble on each neighbour u of v (itself being removed in the process). It is immediate to verify that the number of pebbles $v(2t-1)$ on a vertex after the t^{th} odd step is proportional to its authority score at the t^{th} iteration (it would be equal if we did not normalize the score vector at every iteration), whereas the number of pebbles $v(2t)$ on that vertex after the t^{th} even step is proportional to the hub score at the t^{th} iteration. It is worth noting that the number of descendants after t steps of a pebble initially on a vertex v is actually equal to the number of pebbles on that vertex after t steps, $v(t)$, since both quantities obey the recursive equation $v(0) = 1$ and $v(t) = \sum_{u \in N(v)} u(t-1)$ (denoting by $N(v)$ the set of neighbours of v). We shall use this descendants/pebbles dual interpretation repeatedly in the proof of the theorem.

Let us now focus on G_0. Let a pebble be *positive* if either that pebble or one of its ancestors was located in a vertex v_i with $i > 0$. Denote by $v^+(t)$ and $v^-(t)$ respectively the number of positive and non-positive descendants after t steps of a pebble π initially on vertex v, and denote by $v_+(t)$ the minimum number of pebbles on any vertex in the corolla (i.e. $v_+(t) = \min(v_0(t), v_1(t))$).

We can then prove that:

Lemma 3. *For any $i < 0$, $v_i^+(t) \leq (k-1)^{i-1} v_+(t)$.*

Proof. Let us focus on the case of $i < 0$. Consider a pebble π_i initially in v_i and a pebble π_+ initially either in v_0 or in v_1, and tie to each descendant π of π_i a disjoint set $\Pi(\pi)$ of descendants of π_+ *all in the corolla* (ignore every descendant of π_+ outside the corolla, and all its descendants), as follows. If π is on v_i with $i \leq -2$, tie to its child on v_{i+1} a total of $k-1$ children on $v_1, \ldots, v_k$ of each pebble of $\Pi(\pi)$, and to its child on v_{i-1} the remaining children of $\Pi(\pi)$ (at least $|\Pi(\pi)|$). If π is on v_{-1}, tie to its child in v_0 the children of $\Pi(\pi)$ also in v_0 and to its child on v_{-2} the remaining children of $\Pi(\pi)$. Then, if π is on v_0, all the pebbles in $\Pi(\pi)$ are also all on v_0, and one can tie to each descendant of π a total of $|\Pi(\pi)|$ descendants of π_+ on exactly the same vertex - the same obviously holds if π is on v_1. Thus, each descendant of π_i on a vertex v_j has at least as many pebbles in the corolla tied to it as its parent, and at least $(k-1)$ as many if it is closer to the corolla and $j \leq -1$. Therefore, each of the positive descendants of π_i (with $i < 0$), having had its ancestors take at least $|i| - 1$ steps "up" before reaching v_{-1}, has at least $(k-1)^{|i|-1}$ unique descendants of π_+ tied to it. □

Since $v_+(t)$ obviously grows by a factor at least k every step (the corolla being a $k+1$ vertex clique) whereas $v_i^-(t)$ can obviously grow at most by a factor 2 each step (a vertex in the stem has at most 2 neighbours, and v_0 has only one neighbour in the stem), we have that:

Corollary 1. $\forall i < 0 \; v_+(t) \geq \frac{1}{2}(v_i(t) \cdot \min((\frac{k}{2})^t, (k-1)^{|i|-1})$.

Thus, (a lower bound on) the ratio between $v_+(t)$ and $v_i(t)$ grows by a factor at least $\frac{k}{2}$ at each step, until it plateaus at $(k-1)^{|i|-1}$ within time $2|i|$.

Note that no vertex v has more than $2k$ neighbours, and thus, since $\frac{v(t)}{v(t-1)} = \frac{\sum_{u \in N(v)} u(t-1)}{\sum_{u \in N(v)} u(t-2)}$ (which is a weighted average of ratios $\frac{u(t-1)}{u(t-2)}$), by induction the pebble population of any vertex can grow at most by a factor $2k$ at each step. We can then easily prove the following:

Lemma 4. *For $t \geq 1$:*

1. $k - 1 \leq \frac{v_1(t)}{v_{1,1}(t)} \leq \frac{107}{54}k$.
2. $\frac{54}{125} \leq \frac{v_0(t)}{v_1(t)} \leq \frac{22}{9}$.
3. $\frac{3}{17}k \leq \frac{v_+(t)}{v_{-1}(t)}$

Proof. The proof proceeds by simultaneous induction on three points, with the base case $t = 1$ being trivially verified. Point 1 follows from the fact that $v_{1,1}(t+1) = v_1(t)$ and $(k-1)v_1(t) \leq v_1(t+1) = (k-1)v_1(t) + v_0(t) + kv_{1,1}(t) \leq ((k-1) + \frac{22}{9} + \frac{k}{k-1})v_1(t) \leq \frac{107}{54}kv_1(t)$. Point 2 follows from the fact that $\frac{54}{125} \leq \frac{k}{(k-1)+22/9+k/(k-1)} \leq \frac{kv_1(t)+v_{-1}(t)}{(k-1)v_1(t)+v_0(t)+kv_{1,1}(t)} = \frac{v_0(t+1)}{v_1(t+1)} \leq \frac{(k+17/9)v_1(t)}{(k-1)v_1(t)}$. Point 3 follows from the fact that $\frac{\min(v_0(t+1),v_1(t+1))}{v_{-1}(t+1)} \geq \frac{(k-1)v_1(t)}{v_0(t)+v_{-2}(t)}$ and the last term is $\frac{k-1}{2}$ for $t = 0$ (which is greater than $\frac{3}{17}k$ for all $k \geq 3$) and (by Corollary 1) at least $\frac{kv_1(t)}{(22/9)v_1(t)+(4/k)v_1(t)} \geq \frac{3}{17}k$ for $t \geq 1$. □

Then, after s iterations, the majority of the pebble population of G_0 is concentrated between the petals and the corolla, with each petal and v_{-1} holding $\Theta(1/k^2)$ of all the pebbles, and each vertex of the corolla holding $\Theta(1/k)$.

Let us now focus on the impact of the pebbles generated by v_{-s-1} on the pebble population of G_0. To this end, mark each pebble on v_{-s-1}, or with an ancestor on v_{-s-1}, at time t, with a timestamp t. It is immediate to verify that the set of unmarked pebbles on any vertex of G_0 at any given time coincides with the total pebble population of the corresponding vertex of G_1 at that same time.

Note that, for every integer $i \leq 1$, a pebble resting on a node v_i at time t has at least one descendant on v_{-s-1} at some time $t' \leq t + s + 2$; this descendant has, in turn, one descendant on v_1 at some time $t'' \leq t + 2s + 4$; and this descendant has, in turn, one descendant on *each* node of G_0 at time $t''' = t + 3s + 6$. Then, since a pebble has at most $(2k)^\tau$ descendants after τ timesteps, after time $t + 3s + 6$ on any node the pebble's descendants whose latest timestamp is at least t but less than $t + 3s + 6$ are at least a fraction $(2k)^{-(3s+6)}$ of those descendants with no timestamps greater or equal to $t + 3s + 6$. Thus, after τ timesteps, the number of unmarked pebbles on any vertex of G_0 is no more than a fraction $(1 - (2k)^{-(3s+6)})^{\lfloor \frac{\tau}{3s+6} \rfloor}$ of the total number of pebbles on that vertex and, since for any pair of adjacent vertices v and u we have that $v(t) \leq (2k)v(t-1) \leq (2k)u(t)$, eventually the ratio between the pebble population of *any* vertex of G_1 and of any vertex of G_0 (not necessarily the corresponding one) becomes arbitrarily small.

Yet, it is easy to verify that this cannot happen too quickly. Denote by $M_\tau(t)$ the number of pebbles at time t whose latest timestamp is τ, and by $G_0(t)$ the total number of pebbles on G_0 at time t. $\frac{M_\tau(\tau)}{G_0(\tau)} < \frac{M_\tau(\tau)}{v_0(\tau)} = \frac{v_{-s-1}(\tau)}{v_0(\tau)}$; for $t > \tau$, remembering that $v(t-\tau)$ represents the number of descendants after $t - \tau$ timesteps of a pebble in v, $\frac{M_\tau(t)}{G_0(t)} < \frac{v_{-s-1}(\tau)v_{-s-1}(t-\tau)}{v_0(\tau)v_0(t-\tau)}$. By Corollary 1,

$\frac{v_{-s-1}(\tau)}{v_0(\tau)} \leq 1$ for all τ and $\frac{v_{-s-1}(\tau)}{v_0(\tau)} \leq 2(\frac{2}{k})^s$ for $\tau \geq s$; thus, for $t \geq 2s$, $\frac{M_\tau(t)}{G_0(t)} < 2(\frac{2}{k})^s$ and $k^{\Theta(s)}$ steps are required to mark even a polynomially (in k) small fraction - e.g. $\Theta(\frac{1}{k^3})$ - of all pebbles of G_0.

Thus, for $k^{\Theta(s)}$ steps, the $(k+1)^2$ vertices of the corollas of the different flowers will each hold a fraction of all the pebbles of G that is at least $\Theta(\frac{1}{k^2})$ and at least $\Theta(k)$ larger than the number of pebbles on any petal or stem; but eventually, any flower other than G_0 will hold only a negligible fraction of all the pebbles, whereas each vertex of the corolla and each petal of G_0, as well as v_{-1}, (a total of (k^2+k+2) vertices) will each hold at least a fraction $\Omega(\frac{1}{k^2})$ of all the pebbles of G, and (an additive) $\Omega(\frac{1}{k^2})$ more pebbles than any other vertex, save at most v_{-2}.

This means that for $k^{\Theta(s)}$ steps, the normalized (in $\|\cdot\|_2$) score vector has $\Theta(k^2)$ components (corresponding to the vertices of the corollas) each with a value $\Theta(k^{-1})$, and all remaining components with a value $O(k^{-2})$; but the normalized limit score vector has $\Theta(k)$ components (corresponding to the vertices of the corolla of G_0) with a value of $\Theta(k^{-\frac{1}{2}})$, another $\Theta(k^2)$ (corresponding to the vertices adjacent to the corolla of G_0) each with a value of $\Theta(k^{-\frac{3}{2}})$, and all remaining ones save at most v_{-2} with a value of $O(k^{-\frac{5}{2}})$.

Then any score vector with a (corolla) vertex outside of G_0 in the top k^2+k positions (where all corolla vertices belong for $k^{\Theta(s)}$ steps) differs, in $\|\cdot\|_2$, from the limit score vector by at least $\Theta(\frac{1}{k\sqrt{k}})$, and HITS does not ϵ-converge in rank for any $\epsilon \leq \bar{\epsilon} = \Theta(\frac{1}{k\sqrt{k}})$ for at least $k^{\Theta(s)}$ iterations on more than $k+1$ of the top k^2+k vertices of G. □

We can easily adapt the lower bound to deal with graphs that are both strongly connected and authority connected. Let the gate γ_i of G_i be the vertex corresponding to v_0 in G_i. The basic idea is to string all the flowers of G in a "garland", linking $\forall i$ the gate of G_i to the gate of $G_{(i+1)\%(k+1)}$ through a "string" of $2n$ vertices $w_{i,1}, \ldots, w_{i,2n}$ (with $w_{i,j}$ connected to $w_{i,(j+1)}$, $w_{i,1}$ connected to the gate of G_i and $w_{i,2n}$ connected to the gate of $G_{(i+1)\%(k+1)}$ - see Figure 4). Note that the new graph is still symmetric and unweighted, it is strongly connected, and since it certainly holds for $k \geq 3$ a cycle of odd length, it is authority connected.

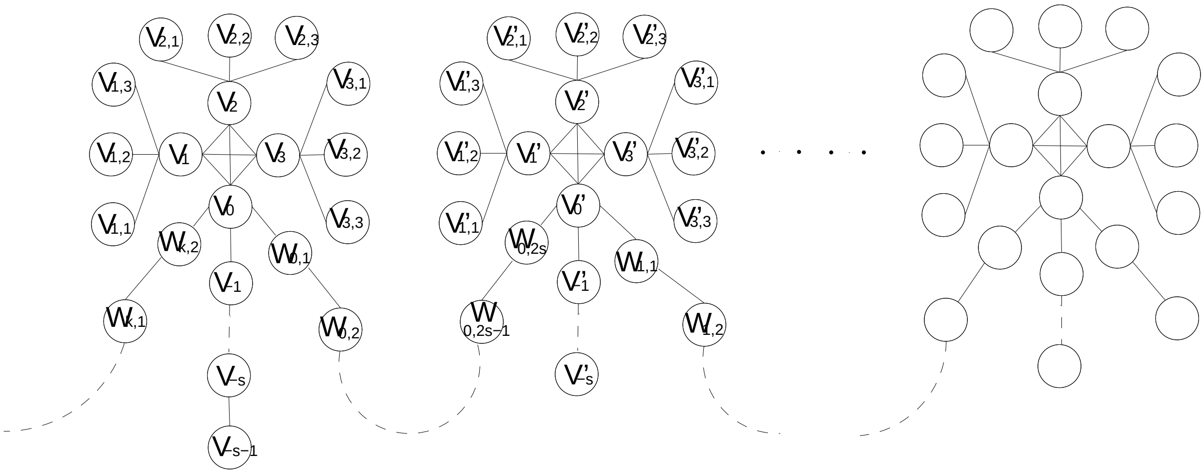

Figure 4: One can connect the multiple flower subgraphs in a "garland" using strings of length $2s$.

Let the pebble population of G_i be the set of all pebbles on vertices of G_i, or on the two strings connecting G_i to the two neighboring flowers whose latest ancestor on a gate was actually on the gate of G_i. Note that there are pebbles that do not belong to the population of any flower (those on a string whose ancestors were always on that same string), but since each such pebble can sire at most two other such pebbles, the total number of such pebbles dwindles to a negligible fraction $k^{-\Omega(s)}$ of the population of any vertex of G within time $O(s)$ and can be disregarded. It is immediate to repeat the proof above to show that, in the absence of v_{-s-1}, the total population of all flowers would remain identical, with corollas rapidly growing to each hold a(n additive) fraction $\Theta(1/k^2)$ of all the pebbles more than any other vertex. It is also immediate to show that the effects of v_{-s-1} remain negligible for $k^{\Theta(s)}$ iterations. All we have to prove is that, despite the strings, eventually the corolla and petals of G_0 hold the top k^2+k positions, each holding at least a fraction $\Omega(1/k^2)$ of all the pebbles more than any other vertex, save at most v_{-1} and v_{-2}.

Note that each flower now receives a stream of pebbles (directly at its gate) from the two neighboring flowers. All we have to prove is that the number of pebbles reaching the gate of G_0 from v_{-s-1} remains a substantially larger fraction of the population of that gate (e.g. by a factor 2) than the fraction of the population of the gate of G_i represented by the stream of pebbles reaching it from the gates of the neighboring flowers. But since each string is $2s$ vertices long, whereas v_{-s-1} is only $s+1$ vertices away from v_0, this is true as long as the ratio of the population of the two gates remains bounded by some $\rho = k^{\Theta(s)}$. Then, eventually the corolla and petals of G_0 hold at least a fraction $\Omega(1/k^2)$ of all the pebbles more than any other vertex, save at most v_{-1} and v_{-2}, and HITS still fails to ϵ-converge on more than $k+1$ of the top k^2+k ranks in less than $k^{\Theta(s)}$ iterations for all $\epsilon \leq \bar{\epsilon} = \Theta(\frac{1}{k\sqrt{k}})$.