The Rank Convergence of HITS Can Be Slow

Enoch Peserico and Luca Pretto

Abstract. We prove that HITS, to “get right” $h_i$ of the top $k$ ranked nodes of an $N \geq 2k$ node graph, can require $h^{\Omega(N^{k/2})}$ iterations (i.e. a substantial $\Omega(N^{k/2} \log k)$ matrix multiplications even with a “squaring trick”). Our proof requires no algebraic tools and is entirely self-contained.

Keywords: Algorithm analysis; Information retrieval; Rank convergence.

1 HITS

Kleinberg’s celebrated HITS algorithm [11] ranks the nodes of a generic graph in order of “importance” based solely on the graph’s topology. Originally proposed to rank web pages in order of authority (and still the basis of some search engines such as Ask [2]), it has been adapted to many different application domains, such as topic distillation [5], word stemming [3], automatic synonym extraction in a dictionary [4], item selection [17], and author ranking in question answer portals [9] (to name just a few - see also [13, 14, 15, 10]).

The original version of HITS works as follows. In response to a query, a search engine 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. Then HITS associates an authority score $a_i$ (as well as a hub score $h_i$) to each node $v_i$ of this base set, and iteratively updates these scores according to the formulas:

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

where $v \to u$ denotes that $v$ points to $u$. At each step the authority and hub vector of scores are normalized in $\| \cdot \|_2$.

Intuitively, HITS places a pebble on each node of the base set graph. At odd timesteps, each pebble on node $v$ sires a pebble on every node $u$ such that $v \to u$, and at even timesteps each pebble on node $v$ sires a pebble on every node $u$ such that $u \to v$ (a pebble is removed upon siring its children). Then, without normalization, $a_i^{(k)}$ equals the number of pebbles on $v_i$ at time $2k - 1$ and $h_i^{(k)}$ that at time $2k$.

*{enoch, pretto}@dei.unipd.it - Dipartimento di Ingegneria dell’Informazione, Università di Padova, Italy. The first author was supported in part by MIUR under PRIN Mainstream and by EU under Integr. Proj. AEOLUS (IP-FP6-015964) The second author was supported in part by Proj. SAPIR within EU Comm. IST Proj. (Contr. IST-045128).
2 Convergence in Score vs. Convergence in Rank

HITS essentially computes a dominant eigenvector of $A^TA$, where $A$ is the adjacency matrix of the base set, using the power method [7] - thus, the convergence rate of the hub and authority score vectors are well known [1]. Nevertheless, what is often really important [16] is the time taken by HITS to converge in rank: intuitively after how many iterations nodes no longer change their relative rank. A formalization of this intuition is more challenging than it might appear [16]. For the purposes of this paper we define convergence in rank as follows:

**Definition 1** Consider an iterative algorithm ALG providing at every iteration $t \geq 0$ a score vector $v^t = [v^t_1, \ldots, v^t_N]$ for the $N$ nodes $v_1, \ldots, v_N$ of a graph; and let the set of the (weakly) top $k$ nodes at step $t$ be $T^k_t = \{v_i : |\{v_j : v^t_j > v^t_i\}| < k\}$. Then ALG converges on $h$ of the top $k$ ranks in $\tau$ steps if $|\bigcap_{t=\tau}^{\infty} T^k_t| \geq h$.

In other words, an algorithm converges on $h$ of the top $k$ ranks in $\tau$ steps if after $\tau$ steps it already “gets right” at least $h$ of the $k$ (eventually) top ranked elements. This definition is closely related to that of convergence in the intersection metric [9] [16] for the top $k$ positions; [16] provides a more thorough discussion of its relationship to other popular metrics such as Kendall’s $\tau$, Cramer-von Mises’ $W^2$, or Kolmogorov-Smirnov’s $D$.

We prove the first non-trivial lower bound on the iterations HITS requires to converge in rank. All previous rank convergence studies save [16] are experimental, and none investigates HITS (focusing instead on PageRank [8, 12, 16]).

3 HITS Can Converge Slowly in Rank

Informally, we prove that HITS on an $N$ node graph can take $h^{\Omega(N^{1/2})}$ steps to “get right” $h$ of the top $k$ elements. This effectively means $\Omega\left(\frac{Nk\log h}{k}\right)$ matrix multiplications using the standard “squaring trick” that computes the $p^{th}$ power of a matrix $M$ by first computing the matrices $M^2, M^4, \ldots, M^{2\left\lfloor \log_2 p \right\rfloor}$. More formally, we devote the rest of this section to the proof of:

**Theorem 1** For all $h$ and $k$ such that $k > h > 5$, and all odd $n \geq \max(3, \frac{k-h+2}{2})$, there is an undirected graph $\Gamma_{h,k,n}$ of $N = \left\lceil \frac{k-2}{h-3} \right\rceil (2n + h - 3) + 1 \approx \frac{2nk}{h} + k$ vertices on which HITS requires more than $\bar{\ell} = \frac{3\ln(7/6)}{4e} (\frac{h-3}{2})^{n-1} = h^{\Omega(n)}$ steps to converge on $h$ of the top $k$ ranks (and the last term is $h^{\Omega(N^{1/2})}$ for $N \geq 2k$).

$\Gamma_{h,k,n}$ (Fig. 1) is formed by a subgraph $\tilde{\Gamma}_{m,n}$ (with $m = h - 3$) and $\ell = \left\lceil \frac{k-2}{h-3} \right\rceil$ isomorphic subgraphs $\Gamma^1_{m,n}, \ldots, \Gamma^n_{m,n}$. $\tilde{\Gamma}_{m,n}$ has $2n + m + 1$ vertices, $v_{-n}, \ldots, v_0, \ldots, v_n$. The $2n + 1$ vertices $v_{-n}, \ldots, v_0, \ldots, v_n$ form a chain, with $v_i$ connected to $v_{i-1}$. The first and last vertices of the chain, $v_{-n}$ and $v_n$, are also connected to each of the $m$ vertices $v_{n+1}, \ldots, v_{n+m}$. $\Gamma^1_{m,n}$ has $2n + m$ vertices, $u_{-n}, \ldots, u_{-1}, u_1, \ldots, u_{n+m}$, and is almost isomorphic to $\tilde{\Gamma}_{m,n}$: $u_i$ is connected to $u_j$ if and only if $v_i$ is connected to $v_j$. The only difference is that $u_0$ is missing.
The proof of the theorem proceeds as follows. After introducing some notation, Lemma 1 bounds the growth rate of the number of pebbles on $v_i$ and $u_i$ as a function of $i$. Lemma 2 allows us prove, in Lemma 3 that $v_{n+1}$ acquires pebbles only minimally faster than $u_{n+1}$. We then prove the theorem showing that $\Gamma_{m,n}$ eventually holds all the top $k$ nodes, but $u_{n+1}, \ldots, u_{n+m}$ and the corresponding nodes in $\Gamma_1^{m,n}, \ldots, \Gamma_{\ell}^{m,n}$ for $t \leq \bar{t}$ still outrank $v_{n-1}, \ldots, v_0, \ldots, v_{-n+1}$ (and thus at least $\ell m > k - h$ of the top $k$ nodes lie outside $\Gamma_{m,n}$).

Denote by $v^t$ the number of descendants at time $t$ of a pebble present at time 0 on $v$ - which is also equal to the number of pebbles present on $v$ after a total of $t$ timesteps, since both quantities are described by the recursive equation $v^{t+1} = \sum_{u \leftrightarrow v} u^t$ with $v^0 = 1$.

Also, mark with a timestamp $\tau$ any pebble present at time $\tau$ on $v_0$ and any pebble not on $v_0$ whose most recent ancestor on $v_0$ was present at time $\tau$. Note that the number of unmarked pebbles present at any given time on $v_i$ (for any $i$) is equal to the total number of pebbles present at that time on $u_i$; and, more generally, it is straightforward to verify by induction on $t - \tau$ that any pebble present on a vertex $u_i$ at time $\tau$ has, on any vertex $u_j$ and at any time $t \geq \tau$, a number of descendants equal to the number of descendants not marked after $\tau$ that any pebble present on a vertex $v_i$ at time $\tau$ has on $v_j$ at time $t$.

**Lemma 1** For any $t \geq 0$, and any $i, j$ such that $(i \equiv j) \mod 2$ and $0 \leq i < j \leq n + 1$, we have $1 \leq \frac{v_i^{t+1}}{v_i^t} \leq \frac{v_j^{t+1}}{v_j^t} \leq m + 1$ and similarly (if $i > 0$) $1 \leq \frac{u_i^{t+1}}{u_i^t} \leq m + 1$.

**Proof.** We prove that $1 \leq \frac{v_i^{t+1}}{v_i^t} \leq \frac{v_j^{t+1}}{v_j^t} \leq m + 1$ by induction on $t$. The base case $t = 0$ is easily verified. $\frac{v_i^{t+1}}{v_i^t} = \frac{\sum_{v_{i,j} \rightarrow v_i} v_{i,j}^{t+1}}{\sum_{v_{i,j} \rightarrow v_i} v_{i,j}^t}$ is a weighted average (with positive weights) of all ratios $\frac{v_{i,j}^{t+1}}{v_{i,j}^t}$. By inductive hypothesis, $1 \leq \min_{v_{i,j} \rightarrow v_i} \frac{v_{i,j}^{t+1}}{v_{i,j}^t} \leq \frac{v_i^{t+1}}{v_i^t} \leq m + 1$.
max_{v_j \to v_i} \frac{v_i^t}{v_j^t} = \frac{v_i^{t+1}}{v_i^{t+1}} = \min_{v_j \to v_i} \frac{v_i^t}{v_j^t} \leq \frac{v_i^{t+1}}{v_j^{t+1}} = \max_{v_j \to v_i} \frac{v_j^t}{v_i^t} \leq m + 1. The proof that $1 \leq \frac{u_{i+1}^t}{u_i^t} \leq \frac{u_{j+1}^t}{u_j^t} \leq m + 1$ is identical. □

Lemma 2. For all $i \geq 0$ consider an unmarked pebble $p_i$ present at time $t$ on vertex $v_i$. At some time $t + \tau$, with $\tau \leq 2n + 2$, $v_{n+1}$ holds at least one marked descendant $p_i'$ of $p_i$; by virtue of Lemma 1 $p_i'$ thereafter always has at least as many descendants as any of the other at most $(m + 1)^{\tau}$ descendants of $p_i$ present at time $t + \tau$. Then, every $2n + 2$ timesteps, the fraction of unmarked descendants of an unmarked pebble drops by a factor at least $1 - (m + 1)^{-(2n+2)}$. □

Lemma 3. For all $i \geq 0$ consider an unmarked pebble $p_i$ present at time $t$ on vertex $v_i$. At some time $t + \tau$, with $\tau \leq 2n + 2$, $v_{n+1}$ holds at least one marked descendant $p_i'$ of $p_i$; by virtue of Lemma 1 $p_i'$ thereafter always has at least as many descendants as any of the other at most $(m + 1)^{\tau}$ descendants of $p_i$ present at time $t + \tau$. Then, every $2n + 2$ timesteps, the fraction of unmarked descendants of an unmarked pebble drops by a factor at least $1 - (m + 1)^{-(2n+2)}$. □

We can now prove Theorem 1. If $\Gamma_{m,n}$ is the number of descendants at time $t$ of a pebble initially on $v_{n+1}$ whose timestamp is $\tau$, and with $D_{\Gamma}^t$ the number of those descendants that are unmarked. Since all pebbles whose timestamp is $\tau$ descend from pebbles present in $v_0$ at time $\tau$, Lemma 1 guarantees that the growth rates of $D_{\Gamma}^t$ satisfy $D_{\Gamma}^{t+1} \leq \frac{D_{\Gamma}^t}{m}$ for any $\tau$ for which $D_{\Gamma}^t \neq 0$. Thus, $u_{n+1}^{t+1} = \frac{D_{\Gamma}^{t+1} + D_{\Gamma}^{t+1} + \ldots + D_{\Gamma}^{t+1}}{D_{\Gamma}^{t} + D_{\Gamma}^{t} + \ldots + D_{\Gamma}^{t}} \leq \frac{D_{\Gamma}^{t+1} + v_{n+1}^{1/m}}{u_{n+1}^{t+1} + 1} = \frac{u_{n+1}^{t+1} + \frac{1}{m} u_{n+1}^{1/m}}{u_{n+1}^{t+1} + 1}$. Therefore, $u_{n+1}^{t+1}/u_{n+1}^t \leq 1 + \frac{1}{m} v_{n+1}^{1/m} \leq 1 + 1/v_{n+1}^t \leq 1 + 1/m + v_{n+1}^{1/m} \leq □$

We first prove that, for $n - 1 \leq t \leq \bar{t}$, $v_0^t \leq e(2/m)^{n-1}$. It is straightforward that $v_0^0 = 2n-1$ and $v_0^{n-1} v_0^{n-1} \geq (2m)^{m-1}$. Then, $v_0^t \leq e(2/m)^{n-1}$. It is straightforward that $v_0^0 = 2n-1$ and $v_0^{n-1} v_0^{n-1} \geq (2m)^{m-1}$. Then, $v_0^t \leq e(2/m)^{n-1}$. It is straightforward that $v_0^0 = 2n-1$ and $v_0^{n-1} v_0^{n-1} \geq (2m)^{m-1}$. Then, $v_0^t \leq e(2/m)^{n-1}$. It is straightforward that $v_0^0 = 2n-1$ and $v_0^{n-1} v_0^{n-1} \geq (2m)^{m-1}$. Then, $v_0^t \leq e(2/m)^{n-1}$.
4 Conclusions and Open Problems

This paper presents a self-contained proof that HITS might require $h^\Omega(N^{1/h})$ iterations to "get right" $h$ of the top $k$ nodes of an $N \geq 2k$ node graph. This translates into $\Omega(N^{h \log h/k})$ matrix multiplications even using a "squaring trick" - a substantial load when HITS must be used on-line on large graphs (e.g. in web search engines).

We conjecture that $g^\Omega(N^{1/g})$ is a tight worst case bound on the iterations required by HITS to converge in rank on $h$ of the top $k$ ranked nodes of an $N \geq 2k$ node graph of maximum degree $g$. This is slightly more (for $h$ subpolynomial in $N$) than the lower bound presented here.

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.
[2] [http://www.ask.com](http://www.ask.com)
[3] M. Bacchin, N. Ferro, and M. Melucci. The effectiveness of a graph-based algorithm for stemming. In Proc. of ICADL, 2002.
[4] M. W. Berry, editor. Survey of Text Mining: Clustering, Classification, and Retrieval. Springer, 2004.
[5] S. Chakrabarti, B. E. Dom, D. Gibson, R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Experiments in topic distillation. In Proc. of the ACM SIGIR Workshop on Hypertext IR on the Web, 1998.
[6] R. Fagin, R. Kumar, and D. Sivakumar. Comparing top $k$ lists. In Proc. of ACM-SIAM SODA, 2003.
[7] G. H. Golub and C. F. Van Loan. Matrix Computations, 3rd ed. Johns Hopkins Univ. Press, 1996.
[8] T. H. Haveliwala. Efficient computation of PageRank, Stanford U. Tech. R., 1999.
[9] P. Jurczyk and E. Agichtein. HITS on question answer portals: Exploration of link analysis for author ranking. In Proc. of ACM SIGIR, 2007.
[10] B. Kimelfeld, E. Kovacs, Y. Sagiv, and D. Yahav. Using language models and the HITS algorithm for XML retrieval. In Comput. Eval. of XML IR Systems, 2007.
[11] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. Journal of the ACM, 46(5):604–632, 1999.
[12] S. D. Kumar, T. H. Haveliwala, C. D. Manning, and G. H. Golub. Extrapolation methods for accelerating PageRank computations. In Proc. of WWW, 2003.
[13] O. Kurian and L. Lee. Respect my authority! HITS without hyperlinks, utilizing cluster-based language models. In Proc. of ACM SIGIR, 2006.
[14] S. Mizzaro and S. Robertson. HITS hits TREC - exploring IR evaluation results with network analysis. In Proc. of ACM SIGIR, 2007.
[15] N. Oyama, Y. Masunaga, and K. Tachi. A diachronic analysis of gender-related web communities using a HITS-based mining tool. In APWeb, 2006.
[16] E. Peserico and L. Pretto. What does it mean to converge in rank? In Proc. of ICTIR, 2007.
[17] K. Wang and M.-Y. T. Su. Item selection by "hub-authority" profit ranking. In Proc. of ACM SIGKDD, 2002.