SDSC3001 Course Project

PageRank on an evolving graph

Group 12

CHAN Chak Yan 56217054
Choy Ka Hei 56611057
LAW Chun Lok 56628838
YUEN Ho Man 56219047

Introduction

In the era of advanced technology, information updates every second on the internet, and social networks are evolving continuously and constantly. The classical paradigm that reads static datasets is inadequate for the current social network and web graph setting. The accuracy is decreasing due to the high information transmission rate nowadays.

To address the PageRank accuracy declining problem, the research paper “PageRank on an Evolving Graph” by Bahmani, Bahman, et al. proposed it is necessary to design an algorithm that crawls the pages with its decision and computes the PageRank using the obtained information with a lower error. The public is pursuing the internet to understand their needs and be more convenient for receiving information, entertaining, and communicating. Improving the evolving graph is always needed due to social networks and web graphs changing continuously.

Challenges

It encountered some theoretical and technical challenges while reproducing the research problem.

In the real world, using PageRank is an essential technique for measuring a social network graph. The structure of the graph will vary over time due to the high complexity of this idea. Having an efficient algorithm in the Evolving Graph becomes more essential. As we notice, many algorithms for updating PageRank are inefficient, as most of them require complete information of changes in the network. With this assumption, it is unrealistic to apply it in real life. Therefore, we are not able to take previous paper as reference.

Methodologies

The solution of the research paper to compute the PageRank with as little error as possible was implemented in the following steps:

First, set up 2 baseline probing, Random probing and Round Robin probing; these are the traditional probing with an outstanding performance previously in a fixed dataset. Furthermore, set up another 2 improved test probing, Proportional Probing and Priority probing. Second, compare the test probing to the baselines using 3 different datasets. Observe and understand the performance of the above algorithms to choose the algorithm with the lowest error in PageRank computation.

The following are the baseline probing and testing probing algorithms:

Random probing

Intuition:

Probes the chosen node uniformly and randomly at every time step. Every node with the same distribution to probe.

Algorithm:

At each point t, it has a most recent image H^t of the graph, probe the node v chosen randomly.

Round-Robin Probing

Intuition:

Cycling through the nodes and probing them in cycling order. Each Node can only be probed one time in each cycle. Probing without priority and the nodes have equal probability of probing.

Algorithm:

Order the nodes by the nodes PageRank value in descending order. At each point t, choose a node v to probe by following the order.

Proportion probing

Intuition:

Probe the high PageRank nodes with higher frequency, if the high PageRank node's outgoing edges changes, it affects other nodes' PageRank significantly. Nodes in up-to-date graph g^t with higher PageRank should be probed with higher frequency

Algorithm:

At each step t, choose node v to probe and each node has the probability proportional to its PageRank in the current image of the graph, which means a higher PageRank node has a higher probability to probe. The output is the PageRank vector of the current image.

Priority Probing

Intuition:

Probes nodes with frequencies proportional to their current PageRank. Define the priority for each node, and probes nodes with frequencies proportional to their current PageRank in g^t

Algorithm:

All the nodes' priorities are zero initially. In each step t, the node v with the highest priority is probed, and the priority of the probed node v is set to zero. Other nodes' priority is incremented by their PageRank in the current image of the graph.

Theoretical Prove

In this report, we hold the assumption:

Number of nodes is fixed at n

Number of edges remain unchanged at m

Rate of probe = Rate of change

The results shows that the evaluation of the expected pagerank value is given as:

$$(1 - O(1/m))\varphi^t \leq E[\pi^t | G^{t-1}, H^t] \leq (1 + O(1/m))\varphi^t$$

The bounding represents that the expected PageRank values of each algorithm are bounded by the true PageRank values, which is expected.

The steps of proving the expected pagerank value bounding is given as:

Proof of evaluating the Expected PageRank value:

Step 1: Bounding of total variation distance between πt+1 and πt

$$E[D(\pi^{t+1}, \pi^t)] \leq (1 - \varepsilon)/(m\varepsilon)$$

Step 2: The expected PageRank of any node x

$$\pi_x^t(1 - 1/(\varepsilon^2 m)) \leq E[\pi_x^{t+1} | G^t] \leq \pi_x^t(1 + 1/(\varepsilon^2 m))$$

As Gt+1 is obtained from Gt by changing edge (u, v) to edge(u, v0) only walks that traversed edge (u, v) are affected by this change, and each such walk has expected length 1/ε beyond edge (u, v) or (u, v0).

Step 3: By adding a time difference τ:

$$\pi_x^t(1 - 1/(\varepsilon^2 m))^\tau \leq E[\pi_x^{t+1} | G^t] \leq \pi_x^t(1 + 1/(\varepsilon^2 m))^\tau$$

Step 4: Given $\sum_{k\geq 0}()1/m)(1-\phi_v^t)^k = 2/(\phi_v^t m)$

Bound the expected flowout of v not known to the algorithm:

$$E[\sum_x 1/(\phi_v^t f_{u,x}^t)] = E[\sum_x 1/(\phi_v^t m(1-\varepsilon)\pi_v(1/out_v^t))] \leq \beta$$

Step 5:

Assume we remove the part of the walk from that edge to the end of the random walk for each random walk

$$E[\phi_v^{-t}] = (1-\beta\phi_v^t) \leq E[\pi_v^t | G^{t-1}, H^t]$$

Therefore, in steady state:

$$(1-O(1/m))\varphi^t \leq E[\pi^t | G^{t-1}, H^t] \leq (1+O(1/m))\varphi^t$$

Experiments

Based on the research paper, three different types of dataset are used to examine the algorithm, which will be stated in the next section.

In the experiment testing, we treat Random Probing as a baseline for Proportional Probing and Round-Robin as a baseline for Priority Probing. We assume the probing rate α be 1.

We use two matrices in order to measure the performance of the algorithm. Including:

- L_∞ Matric
  - $L_\infty(\pi^t, \varphi^t) = max_{u \in v}|\pi^t(u) - \varphi^t(u)|$
  - where π is PageRank vector and φ is the vector output
- L_1 Matric
  - $L_1(\pi^t, \varphi^t) = \sum u \in v|\pi^t(u) - \varphi^t(u)|$
  - where π is PageRank vector and φ is the vector output

Autonomous Systems (AS)

Consisting of 733 daily instances from 1997 to 2000. Each AS exchanges traffic flows with some of its peers and one can construct a communication network from the Border Gateway Protocol logs. It is assumed that each link in the communication graph lasts for a day.

With the code reproduced from the research paper, we applied the code with this dataset, the result is obtained below:

Random Probing Result (AS)

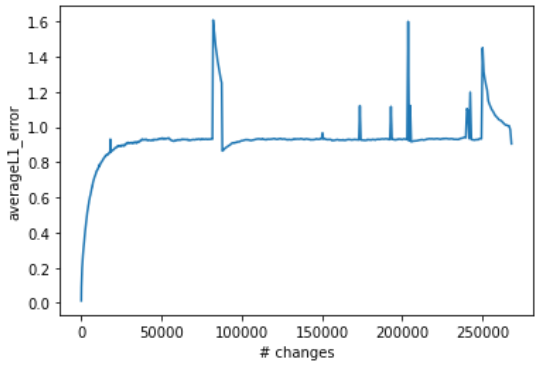

Figure 1: L1 Error of Random Probing under AS

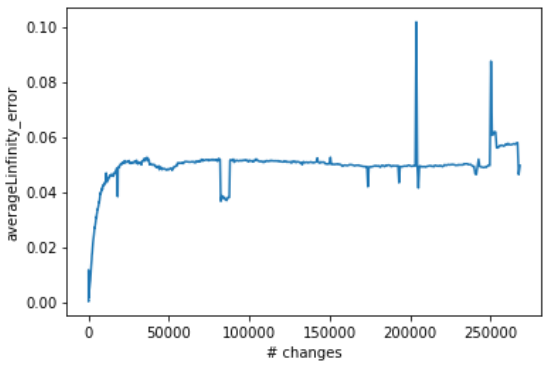

Figure 2: L-infinity Error of Random Probing under AS

Round-Robin Probing Result (AS)

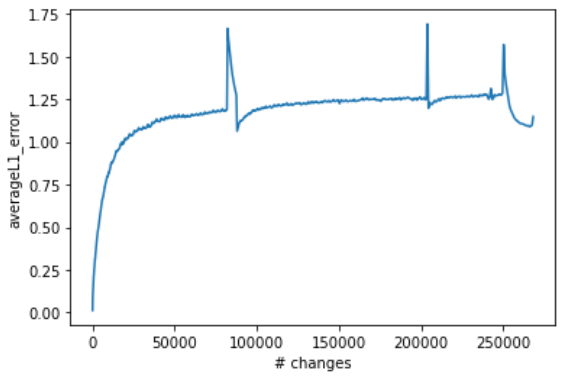

Figure 3: L1 Error of Round-Robin Probing under AS

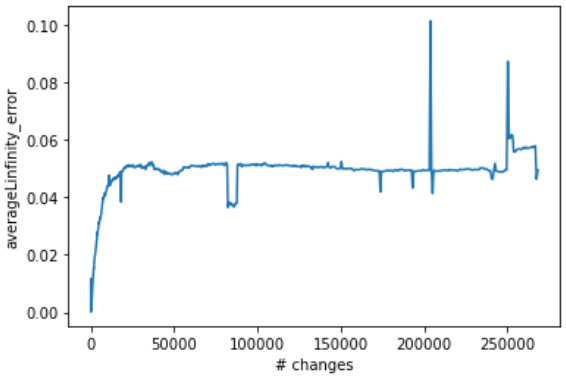

Figure 4: L-infinity Error of Round-Robin Probing under AS

Proportional Probing Result (AS)

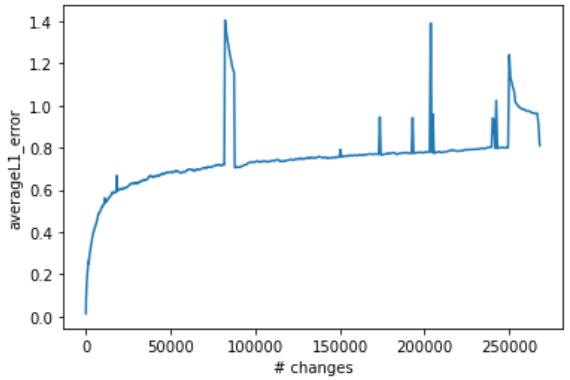

Figure 5: L1 Error of Proportional Probing under AS

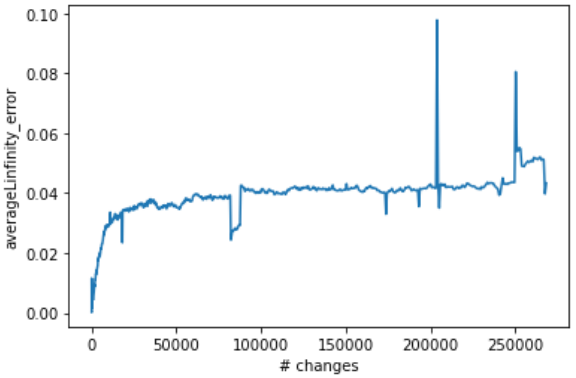

Figure 6: L-infinity Error of Proportional Probing under AS

Priority Probing Result (AS)

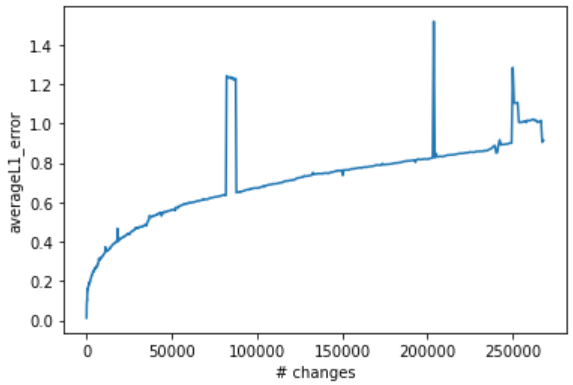

Figure 7: L1 Error of Priority Probing under AS

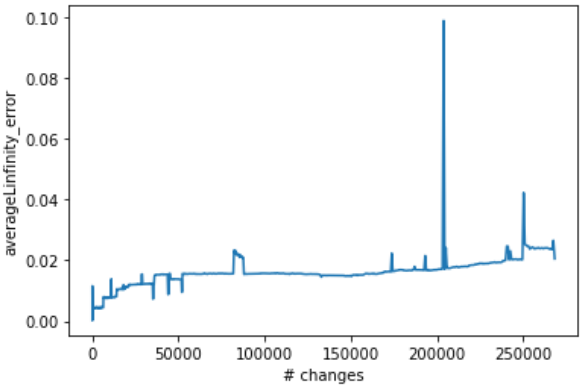

Figure 8: L-infinity Error of Priority Probing under AS

Result Evaluation (AS)

	L1 Error	L-infinity Error
Random	Stable around 0.9	Stable around 0.05
Round-Robin	Stable around 1.1, Keep increasing	Stable around 0.05
Proportional	Stable around 0.7	Stable around 0.04
Priority	Starting from 0.2 to 0.8	Stable around 0.01

The final result demonstrate the result:

- Proportional Probing have lower error than Random Probing relatively
- Priority Probing have lower error than Round-Robin Probing relatively

It is the same as the result in the research paper, which is expected.

Center for Applied Internet Data Analysis (CAIDA)

Contains 122 CAIDA Autonomous Systems graphs, which denotes the communication patterns of the routers. It is assumed that each link exists until the next snapshot.

With the code reproduced from the research paper, we applied the code with this dataset, the result is obtained below:

Random Probing Result (CAIDA)

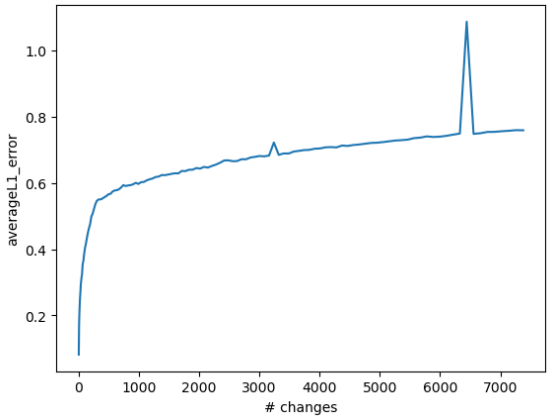

Figure 9: L1 Error of Random Probing under CAIDA

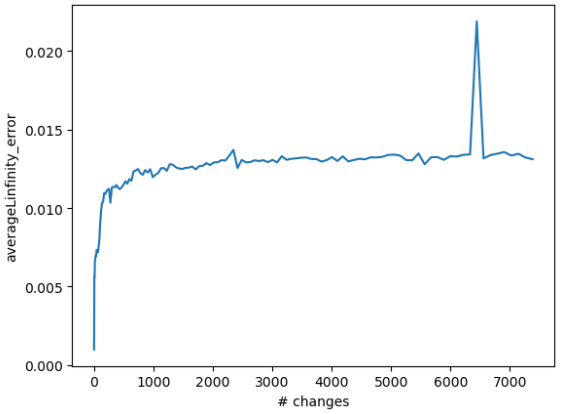

Figure 10: L-infinity Error of Random Probing under CAIDA

Round-Robin Probing Result (CAIDA)

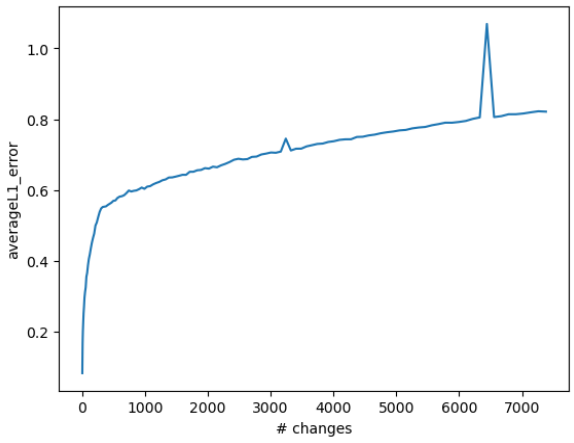

Figure 11: L1 Error of Round-Robin Probing under CAIDA

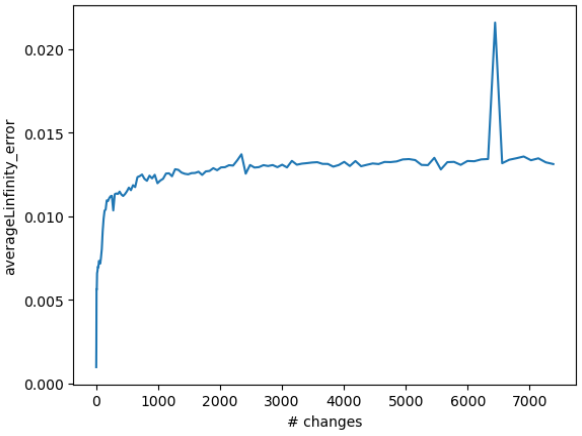

Figure 12: L-infinity Error of Round-Robin Probing under CAIDA

Proportional Probing Result (CAIDA)

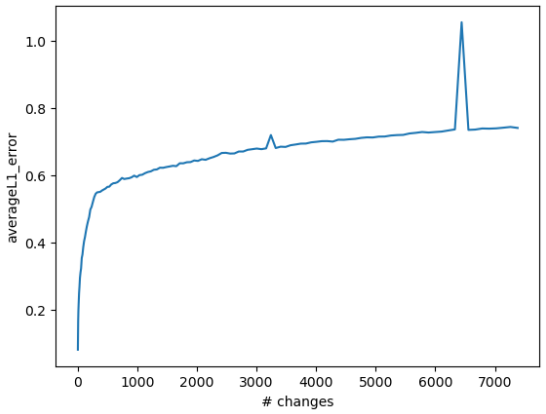

Figure 13: L1 Error of Proportional Probing under CAIDA

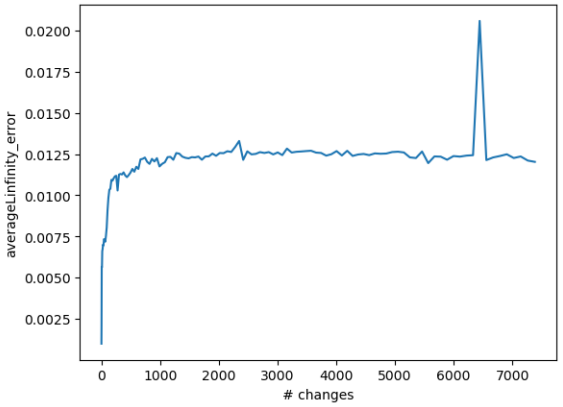

Figure 14: L-infinity Error of Proportional Probing under CAIDA

Priority Probing Result (CAIDA)

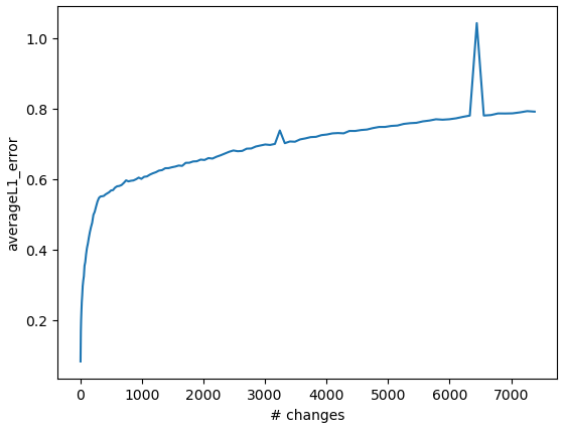

Figure 15: L1 Error of PriorityProbing under CAIDA

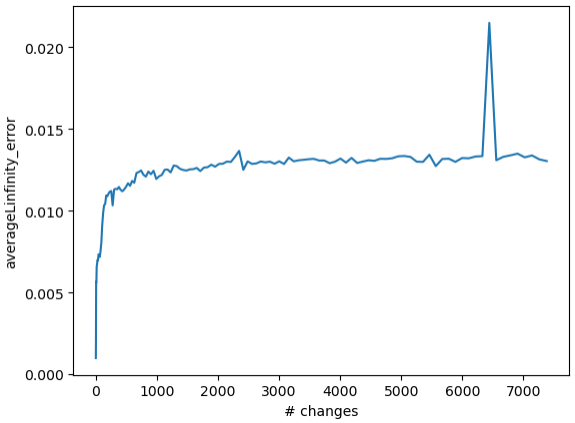

Figure 16: L-infinity Error of PriorityProbing under CAIDA

Result Evaluation (CAIDA)

	L1 Error	L-infinity Error
Random	Stable around 0.7	Stable around 0.013
Round-Robin	Stable around 0.7-0.8, Keep increasing	Stable around 0.013
Proportional	Stable around 0.7	Stable around 0.0125
Priority	Stable around 0.6-0.8, Keep increasing	Stable around 0.013

The final result demonstrate the result:

- Proportional Probing have lower error than Random Probing relatively
- Priority Probing have lower error than Round-Robin Probing relatively
- Proportional Probing Perform the best within all algorithm

Limitation

For the algorithm implementation, executing the probing algorithm has already taken an extremely high time complexity due to an enormous dataset being used. On average, it takes around 24 hours to finish executing all 4 probing algorithms in 2 datasets. With the long computational time, it is more challenging to test the algorithm's accuracy and cannot achieve the hybrid algorithm.

In the research paper, it evaluates how the error varies to the change of α. The results show that the higher the α, the lower the error. When the algorithms probe in a higher frequency, the error will be lower and perform better. Yet due to the time consuming problem, we can only fix α as 1. Therefore, we can’t get the best value for α.

Besides α, for $\beta \in [0, 1]$ which is used for parametrizing the hybrid algorithm between Proportional Probing and Round-Robin Probing. Due to the extremely high time complexity of the hybrid algorithm(10 hours per algorithm), the report lacks the hybrid algorithm implementation.

Conclusion

During reproducing the research paper, we have processed and demonstrated the effective algorithms that have a great performance to solve real-world problems regarding the evolving web page. We are able to use non-trivial analysis to get provable guarantees under a model of the world.

In this report, we have achieved to derive algorithms that work well in practice. Besides, we are able to analyze the theoretical error bounds of the algorithm for a particular model of graph evolution. However, it is always necessary to extend the theoretical analysis and algorithms to pursue higher accuracy of PageRank in evolving graphs.

Reference:

Bahmani, Bahman, et al. "Pagerank on an evolving graph." *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2012.