A PageRank Algorithm based on Asynchronous Gauss-Seidel Iterations

D. Silvestre, J. Hespanha and C. Silvestre

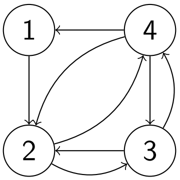
2018 American Control Conference
Milwaukee

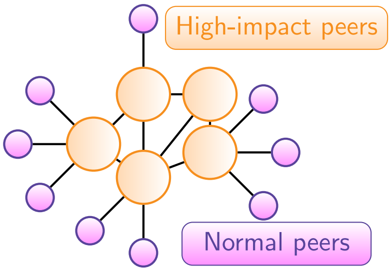
June 27-29 2018

Outline

Motivation

- Traditional PageRank - Compute relative importance of each webpage in the network.
- Ranking of Researchers - PageRank has also been proposed to be an alternative to citation count.
- General Ranking - Since PageRank is equivalent to the relative time spent in a node for a random walk on the graph many more applications are possible.





Intuition behind PageRank

- The ranks should translate to the relative time one would spent on that node for a random walk.
- It should be a steady-state vector.
- Dangling nodes must not appear observant and the random walk must be able to jump to different partitions.
- Problem: $x^\star \in [0,1]^n, \quad \sum_{i=1}^n x_i^\star = 1$

Intuition behind PageRank

- The ranks should translate to the relative time one would spent on that node for a random walk.
- It should be a steady-state vector.
- Dangling nodes must not appear observant and the random walk must be able to jump to different partitions.
- Problem: $x^\star = Ax^\star, \quad x^\star \in [0,1]^n, \quad \sum_{i=1}^{n} x_i^\star = 1$

Intuition behind PageRank

- The ranks should translate to the relative time one would spent on that node for a random walk.
- It should be a steady-state vector.
- Dangling nodes must not appear observant and the random walk must be able to jump to different partitions.
- Problem:

$$x^\star = Mx^\star, \quad x^\star \in [0,1]^n, \quad \sum_{i=1}^{n} x_i^\star = 1$$

$$M = (1-m)A + \frac{m}{n}S, \quad m = 0.15, \quad S := \mathbb{1}_n \mathbb{1}_n^\intercal$$

Some related work

- Defining distributed randomized versions of the power method: C. Ravazzi et al. (2015). "Ergodic Randomized Algorithms and Dynamics Over Networks". In: *IEEE Transactions on Control of Network Systems*
- Use of sequential Gauss-Seidel with no projection: A. Arasu et al. (2002). "PageRank computation and the structure of the Web: experiments and algorithms". In: *In The Eleventh International WWW Conference*. ACM Press
- Using stochastic approximation of the gradient: J. Lei and H. F. Chen (2015). "Distributed Randomized PageRank Algorithm Based on Stochastic Approximation". In: *IEEE Transactions on Automatic Control*

Equivalent formulations for PageRank

- Using the power method it becomes:

$$x(k+1) = Mx(k) = (1-m)Ax(k) + \frac{m}{n}1_n$$

- Intuitively, the Power Method searches for the eigenvector;
- It can be solved as an optimization problem;
- The rank vector is also a eigenvector:

$$(I_n - (1-m)A)x = \frac{m}{n}1_n.$$

Problem Statement

- Find an algorithm that can exploit the fact that subsets of pages are stored in the same processor;
- The iteration should involve A and not M to keep sparsity;
- Solve the equation:

$$(I_n - (1-m)A)x = \frac{m}{n}1_n \tag{1}$$

with constraints:

$$x^\star \in [0,1]^n, \quad \sum_{i=1}^{n} x_i^\star = 1 \tag{2}$$

PageRank Algorithm Problem

How to find a faster algorithm satisfying the above restrictions?

Power Method as the Jacobi Method

- For a linear equation $Ax = b$ partition $A = D + R$ where $D = \text{diag}(A)$ and $R = A - \text{diag}(A)$;
- The Jacobi Method is given by the iteration:

$$x(k+1) = D^{-1}(b - Rx(k)) \tag{3}$$

- If we take $D = I_n$, $R = -(1-m)A$ and $b = \frac{m}{n}1_n$ the Jacobi Method is equivalent to the Power Method;
- Jacobi Method is typically used for solving diagonally dominant equations.

PageRank using Gauss-Seidel

- Partitioning $A = L + D + U$, for lower and upper triangular matrices L and U and diagonal D;
- The Gauss-Seidel Method becomes:

$$x(k+1) = (L+D)^{-1}(b - Ux(k)) \tag{4}$$

- The method can be written to take advantage of updated values

$$x_i(k+1) = \frac{1}{A_{ii}} \left(b_i - \sum_{j=1}^{i-1} A_{ij} x_j(k+1) - \sum_{j=i+1}^{n} A_{ij} x_j(k) \right). \tag{5}$$

- The Gauss-Seidel is also an iterative algorithm for linear equations but corresponding to a different partitioning of the matrix.

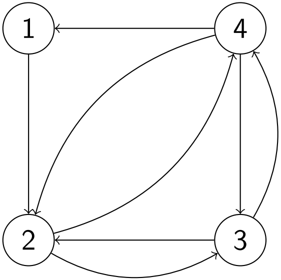
Results

- Theoretical result: $\rho(T_{gs}) < \rho(T_p)$
- Main challenge: computed $x^\star$ using the Gauss-Seidel might not sum to one.
- It is simulated:
  - Projection on the n-simplex;
  - Normalization;
  - Randomized and Asynchronous versions.

Simulation Results (1/2)

Setup: Simple case of four pages.

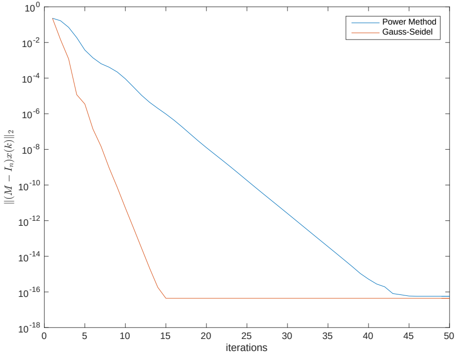
- Gauss-Seidel is faster than the Power Method.
- Performance is affected with no projection.
- Normalization and projection are similar.
- Asynchronous version has an intermediary performance.



Simulation Results (1/2)

Setup: Simple case of four pages.

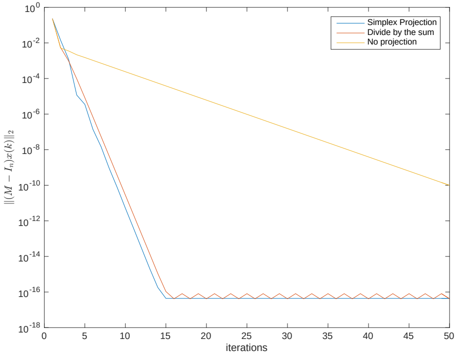
- Gauss-Seidel is faster than the Power Method.
- Performance is affected with no projection.
- Normalization and projection are similar.
- Asynchronous version has an intermediary performance.



Simulation Results (1/2)

Setup: Simple case of four pages.

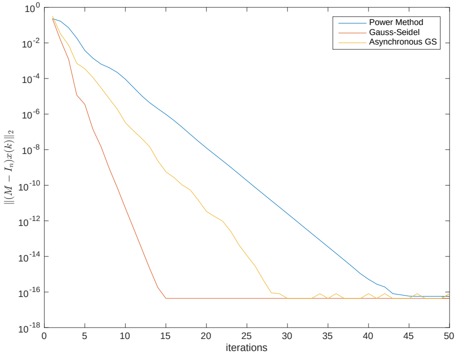
- Gauss-Seidel is faster than the Power Method.
- Performance is affected with no projection.
- Normalization and projection are similar.
- Asynchronous version has an intermediary performance.



Simulation Results (1/2)

Setup: Simple case of four pages.

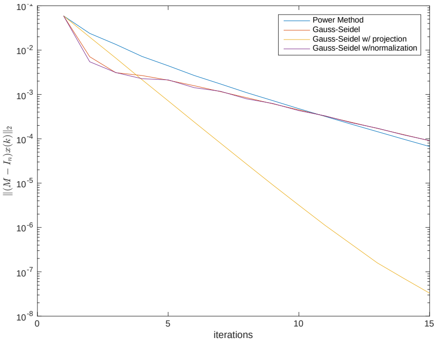
- Gauss-Seidel is faster than the Power Method.
- Performance is affected with no projection.
- Normalization and projection are similar.
- Asynchronous version has an intermediary performance.



Simulation Results (2/2)

Setup: 1000 Monte Carlo experiment for a 500-page network generated by Barabási-Albert algorithm to simulate a power law network as the World-Wide Web.

- In a more realistic scenario the projection is key to get very small errors;
- Fast initial convergence with normalization;
- The problem of having Gauss-Seidel going to zero can be avoided by normalization.



Concluding Remarks

Contributions:

- We have shown that Gauss-Seidel is faster than the Power Method
- Gauss-Seidel does not keep the sum of the state
- Two approaches:
  - projection - implies two communication rounds;
  - normalization - no added complexity.
- Shown how it can be made Asynchronous and Randomized with performance between the Power Method and the Synchronous Gauss-Seidel.

The end

- Thank you for your time.

A PageRank Algorithm based on Asynchronous Gauss-Seidel Iterations

D. Silvestre, J. Hespanha and C. Silvestre

2018 American Control Conference
Milwaukee

June 27-29 2018