Inner/Outer Iterations for Computing PageRank

Chen Greif

The University of British Columbia
Vancouver
Canada

Joint work with David Gleich, Andrew Gray, and Tracy Lau

Chinese Academy of Sciences
Beijing
May 15, 2010

PageRank

Links determine the importance and authority of a webpage.

For PageRank, the 'raw' PageRank x_i of page i is defined as

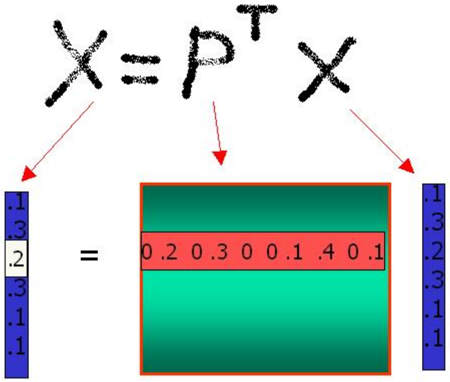
$$x_i = \sum_{j \to i} \frac{x_j}{n_j},$$

where $\{x_j\}$ is the set of pages that link to page i, and n_j is the outdegree of page j.

The problem is thus: find a vector x that satisfies $x = P^T x$, where P is given by

$$P_{ij} = \begin{cases} \frac{1}{n_i} & \text{if } i \to j, \\ 0 & \text{if } i \nrightarrow j. \end{cases}$$

Eigenvalue Problem



Stationary Distribution Vector

The matrix P is row-stochastic; nonnegative entries between 0 and 1, and $Pe = e$, where e is a vector of all-ones.

Definition

Given a Webpage database, the PageRank of the ith Webpage is the ith element π_i of the stationary distribution vector π that satisfies $\pi^T P = \pi^T$, where P is a matrix of *weights* of webpages that indicate their importance.

Theorem

Perron(1907)-Frobenius(1912): A nonnegative irreducible matrix has a simple real eigenvalue equal to its spectral radius, whose associated eigenvector is a vector all of whose entries are nonnegative.

Challenges and Difficulties

1. P is very large (size in the billions): can't just use our favorite decompositions.
2. The existence of dangling nodes (all-zero row in the matrix): could have 'dead-ends'. *Those 'sinks' are not necessarily unimportant!*
3. Cyclic paths, reducibility: a small entity with no entry points and exits to the 'outside world'. *Hurts uniqueness.*
4. The dynamic nature of the Web: new Websites are continually generated, eliminated, changing. The Web link graph is a function of time and so is its dimension. (Will not pursue this in this talk.)

Difficulties

1. The existence of dangling nodes (correspond to an all-zero row in the matrix).
2. Periodicity: a cyclic path in the Webgraph.

PageRank: Scale and Incorporate Rank-One Perturbation

Difficulties

1. The existence of dangling nodes (correspond to an all-zero row in the matrix).
2. Periodicity: a cyclic path in the Webgraph.

Solution

1. *Set $P' = P^T + D$ where $D = dv^T$ is a rank-1 matrix with d being a vector which is 0 if the outdegree is nonzero, and 1 if it is. (In other words, 'perturb' pages without outlinks.)*
2. *Set $\tilde{P} = \alpha P' + (1-\alpha)E$, where E is a rank-1 matrix related to the personalization vector.*

The PageRank Matrix

An interpretation of the convex combination with a low rank matrix: a surfer may jump randomly elsewhere, not necessarily following a link, with probability $1 - \alpha$.

Consider

$$A(\alpha) = \alpha P^T + (1 - \alpha)E,$$

where $E = ev^T$ is a positive rank-1 matrix.

- $A(\alpha)$ is a strictly positive row-stochastic matrix, and for $\alpha < 1$, by Perron-Frobenius there exists a unique (up to scaling) *PageRank vector* $x(\alpha)$ such that $A(\alpha)x(\alpha) = x(\alpha)$.

Many Relevant/Related Models for Other Applications

"For the purpose of evaluating status in a manner free from the deficiencies of popularity contest procedures, this paper presents a new method of computation which takes into account *who* chooses as well as how *many* choose. It is necessary to introduce, in this connection, the concept of attenuation in influence transmitted through intermediaries." (Leo Katz, PSYCHOMETRIKA 1953.)

A random particle starts randomly walking from node i, and moves to its neighbors with probability proportional to their edge weights. The relevance score of node j with respect to note i is defined as the steady state probability that the particle will end up at node j.

- How closely related are two nodes in a weighted graph.
- Social networks, link prediction, ...

A Linear System Formulation

Since $e^T x = 1$ it follows that

$$x = Ax = \alpha P^T x + (1-\alpha) v e^T x \rightarrow \boxed{(\text{I-}\alpha P^T) x = (1-\alpha) v}$$

(See, e.g. Arasu et al. [2002].)

The Neumann expansion is valid for $\alpha < 1$:

$$(I - \alpha P^T)^{-1} = I + \alpha P^T + \alpha^2 (P^T)^2 + \ldots$$

Katz: transmission of information or rumor through a group. The parameter α is an "attenuation factor". ($\alpha = 1$: no attenuation)

Significance of Damping Factor (or Mixing Parameter) α

- In PageRank context, α is the probability of following an outlink (as opposed to jump randomly elsewhere).
- What is 'the right value of α'? A regularization effect: small value (say 0.85) leads to a more stable computation, but further away from true solution. *(This is subject to debate; see, e.g. Boldi, Santini and Vigna [2004].)* See also a nice discussion in *Langville and Meyer [2006]*.
- In general, hard to beat the power method for small values of α. But if α is close to 1, it will not perform well: linear convergence, with asymptotic error constant (recall $\lambda_1 = 1$)

$$\max \frac{|\lambda_j|}{|\lambda_1|}.$$

And we have the following beautiful property...

The Eigenvalues

Theorem

Brauer, 1952; Serra-Capizzano, Eldén - recently
Let P be a column-stochastic matrix with eigenvalues $\{1, \lambda_2, \lambda_3, \ldots, \lambda_n\}$. Then the eigenvalues of $M(\alpha) = \alpha P + (1-\alpha)ve^T$, where $0 < \alpha < 1$ and v is a nonnegative vector with $e^T v = 1$, are

$$\{1, \alpha\lambda_2, \alpha\lambda_3, \ldots, \alpha\lambda_n\}.$$

This implies

$$\frac{|\lambda_j|}{|\lambda_1|} \leq \alpha.$$

Ordering as a Function of α (Rankings in Wikipedia 2005, Size $\equiv 1.1M$, nz $\equiv 18M$)

Entry	$\alpha = 0.85$	$\alpha = 0.90$	$\alpha = 0.95$	$\alpha = 0.99$
United States	1	1	1	1
United Kingdom	3	3	2	2
Canada	8	10	17	17
2005	5	5	11	10
2004	6	6	12	13
2000	7	15	20	29
Category: culture	12	9	8	6
Category: politics	13	7	6	5

Many Acceleration Techniques

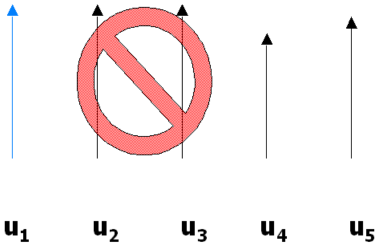
- Quadratic Extrapolation (Kamvar, Haveliwala, Manning, Golub [2003])
- Aggregation/Disaggregation (Langville & Meyer, Stewart [2005])
- Permutations (Del Corso, Gulli and Romani [2007],....)
- Linear system formulation (Arasu, Novak, Tomkins and Tomlin [2002]; Gleich, Zhukov and Berkhin [2004])
- Padé-type acceleration (Brezinski and Redivo-Zaglia [2006])
- Arnoldi-type method (Golub and Greif [2006])
- Power-Arnoldi method (Wu and Wei [2007])
- Ordinal ranking (Wills and Ipsen [2008])

and many many more...
Surveys, properties, etc. : Langville and Meyer [2005-2006], Berkhin [2005], Bryan and Leise [2006], ...

Quadratic Extrapolation (Kamvar, Haveliwala, Manning, Golub [2003])

Slowly convergent series can be replaced by series that converge to the same limit at a much faster rate.

Idea: Estimate components of current iterate in the directions of second and third eigenvectors, and eliminate them.



Quadratic Extrapolation

Suppose A has three distinct eigenvalues. The minimal polynomial is given by $P_A(\lambda) = \gamma_0 + \gamma_1\lambda + \gamma_2\lambda^2 + \gamma_3\lambda^3$. By the Cayley-Hamilton theorem, $P_A(A) = 0$. Hence for any vector z,

$$P_A(A)z = (\gamma_0 + \gamma_1 A + \gamma_2 A^2 + \gamma_3 A^3)z = 0.$$

In particular, set $z = x^{(k-3)}$ and use the fact that $x^{(k-2)} = Ax^{(k-3)}$ and so on.
In the end, since A has more than three eigenvalues, solve a least squares problem.

An Arnoldi-type Method for PageRank (Golub and Greif, 2006)

Similar to computing *refined Ritz vectors* (Jia) , but pretend largest eigenvalue stays 1 in smaller space, i.e., we do not compute any Ritz values, and *avoid complex arithmetic*.

Set initial guess q and k, the Arnoldi steps number
Repeat
.....$[Q, H] = Arnoldi(A, q, k)$
.....Compute $H - [I; 0] = U\Sigma V^T$
.....Set $v = V(:, k)$
.....Set $q = Qv$
Until $\sigma_{\min}(H - [I; 0]) < \varepsilon$

The Arnoldi Algorithm

```
q_1 = q/||q||_2
for j=1 to k
.....z = A q_j
.....for i = 1 to j
.........h_{i,j} = q_i^T z
.........z = z - h_{i,j} q_i
.....end for
.....h_{j+1,j} = ||z||_2
.....if h_{j+1,j} = 0, quit
.....q_{j+1} = z/h_{j+1,j}
end for
```

$q_1 = q/\|q\|_2$
for j=1 to k
.....$z = Aq_j$
.....for $i = 1$ to j
.........$h_{i,j} = q_i^T z$
.........$z = z - h_{i,j} q_i$
.....end for
.....$h_{j+1,j} = \|z\|_2$
.....if $h_{j+1,j} = 0$, quit
.....$q_{j+1} = z/h_{j+1,j}$
end for

Arnoldi(A, q, k).

Pros and Cons

Good:

- Orthogonalization allows for going up in dimension.
- Shift by identity: exploit knowledge of the largest eigenvalue.
- No Ritz value estimates are involved and no complex arithmetic, however can't expect H to really be singular.
- It can be shown that $\|Aq - q\|_2 = \sigma_{\min}(H - [I; 0])$.
- Smooth convergence behavior.

Bad:

- Not as easy to implement as power method and other techniques.
- Iterations more computationally expensive.
- Even for small k the required storage quickly becomes as large as original matrix.

An Inner/Outer Stationary Iteration: Gleich, Gray, Greif, and Lau (SISC, 2010)

Back to linear system formulation:

$$(I - \alpha P^T)x = (1 - \alpha)v.$$

A simple preconditioning/splitting algorithm:

Outer iteration:

$$(I - \beta P^T)x_{k+1} = (\alpha - \beta)P^T x_k + (1 - \alpha)v, \quad k = 1, 2, \ldots$$

Inner iteration:

$$y_{j+1} = \beta P^T y_j + (\alpha - \beta)P^T x_k + (1 - \alpha)v,$$

Iteration Matrices

The outer iteration is associated with the splitting

$$I - \alpha P = M_O - N_O \;; \quad M_O = I - \beta P \;; \quad N_O = (\alpha - \beta) P,$$

and the corresponding outer iteration matrix is given by

$$T_O = M_O^{-1} N_O = (\alpha - \beta)(I - \beta P)^{-1} P.$$

The inner iteration corresponds to the splitting

$$M_O = I - \beta P = M_I - N_I \;; \quad M_I = I \;; \quad N_I = \beta P.$$

The corresponding inner iteration matrix is

$$T_I = M_I^{-1} N_I = \beta P.$$

Note: end cases $\beta = 0, \alpha$ yield power method.

Algorithm 1 basic inner-outer iteration

Input: $P, \alpha, \beta, \tau, \eta, v$

Output: x

$x \leftarrow v$

$y \leftarrow Px$

while $\|\alpha y + (1-\alpha)v - x\|_1 \geq \tau$

 $f \leftarrow (\alpha - \beta)y + (1-\alpha)v$

 repeat

  $x \leftarrow f + \beta y$

  $y \leftarrow Px$

 until $\|f + \beta y - x\|_1 < \eta$

end while

$x \leftarrow \alpha y + (1-\alpha)v$

Advantages (if We Can Make It Fast Enough)

- Attractively simple
- No projections, orthogonalizations, permutations, ...
- Minimal storage requirements
- Easily parallelizable: based almost exclusively on matrix-vector products.

Inexact Iterations

- For $0 \leq \beta \leq \alpha < 1$ can show convergence of exact inner iterations (use M-matrix theory, regular splittings, etc.).
- Practicality dictates using inexact inner/outer iterations.
- Central question: how to determine the parameters.
- Inner/outer iterations: interesting previous work for stationary methods: Golub and Overton [1988]; Elman and Golub [1994]; Giladi, Golub and Keller [1998], ... and also for Krylov solvers: Simoncini and Szyld [2003], ...

Outer Iteration – Correction Interpretation

Write

$$A(\alpha) = \alpha P + (1-\alpha) v e^T \ ; \quad A(\beta) = \beta P + (1-\beta) v e^T .$$

Then

$$A'(\alpha) = A'(\beta) = P - v e^T .$$

(By linearity derivative does not depend on the damping factor.)
Thus

$$A(\alpha) = A(\beta) + (\alpha - \beta) A' .$$

Thus, can interpret the outer iteration as a procedure in which we locally solve the PageRank problem for β rather than for α, and correct for the remainder, which is a scalar multiple of the derivative of A.

Convergence Analysis for Outer Iterations (Cont.)

$M_O = I - \beta P$ is a diagonally dominant M-matrix. If λ_i is an eigenvalue of P then

$$\mu_i = \frac{(\alpha - \beta)\lambda_i}{1 - \beta\lambda_i}$$

is an eigenvalue of T_O. Can show $\rho(T_O) = \frac{\alpha-\beta}{1-\beta} < 1$ (equality!).

Lemma Given $0 < \alpha < 1$, if the inner iterations are solved exactly, the scheme converges for any $0 < \beta < \alpha$. Furthermore,

$$\|x_{k+1} - x\|_1 \leq \frac{\alpha - \beta}{1 - \beta} \, \|x_k - x\|_1$$

and

$$\|x_{k+1} - x_k\|_1 \leq \frac{\alpha - \beta}{1 - \beta} \, \|x_k - x_{k-1}\|_1,$$

and hence the contraction factor $\frac{\alpha-\beta}{1-\beta}$ indicates that the closer β is to α, the faster the outer iterations converge.

Convergence Analysis for Inner Iterations

Difference inequalities:

$$Mx_{k+1} = Nx_k + g + \delta_k \ ; \quad e_k = x_k - x.$$

Suppose $\|\delta_k\| \le \eta\|e_k - e_{k-1}\| = \eta\|x_k - x_{k-1}\|$ for some η. Then

$$\|e_{k+1}\| \le \|T\| \, \|e_k\| + \eta\|M^{-1}\| \, \|e_k - e_{k-1}\|.$$

Defining $\rho = \|T\|$ and $\sigma = \eta\|M^{-1}\|$, can show that

$$\|e_k\| \le a_1\xi_+^k + a_2\xi_-^k$$

where

$$\xi_\pm = \frac{\rho + \sigma}{2}\left(1 \pm \sqrt{1 + \frac{4\sigma}{(\rho+\sigma)^2}}\right).$$

The coefficients are given by

$$a_1 = \frac{2(\beta_2 - \beta_1 \xi_-)}{(\rho+\sigma)^2 s(s+1)} \; ; \qquad a_2 = \frac{2(\beta_2 - \beta_1 \xi_+)}{(\rho+\sigma)^2 s(s-1)} \; ,$$

with $s = \sqrt{1 + 4\sigma/(\rho+\sigma)^2}$, $\beta_1 = \|e_1\|$, $\beta_2 = \|e_2\|$.
After some further simplifications, a sufficient (but really and truly not necessary) condition for convergence is

$$0 < \eta < \frac{1-\alpha}{2}.$$

The closer α is to 1, the smaller η should be chosen to be, which makes sense intuitively.

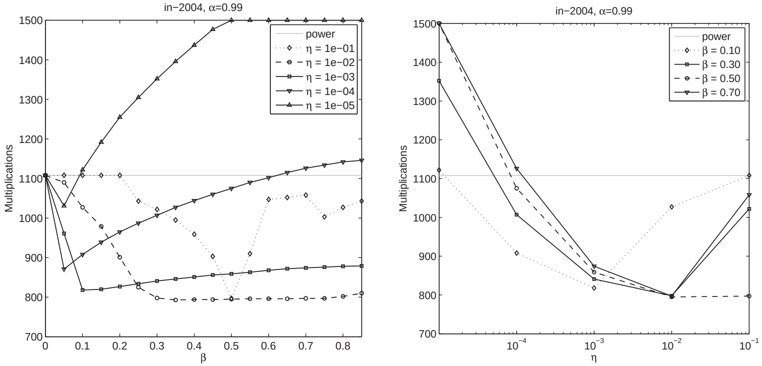


Figure: Total number of matrix-vector products required for convergence of the inner-outer scheme, for the in-2004 matrix. (Outer tolerance 10^{-7}, $\alpha = 0.99$, β, and η varied. The iteration limit was 1500 and causes the ceiling on the left figure.)

A Few Examples

Table: Dimensions and number of nonzeros of a few test matrices

name	size	nonzeros	avg nz per row
ubc-cs-2006	51,681	673,010	13.0
ubc-2006	339,147	4,203,811	12.4
eu-2005	862,664	19,235,140	22.3
in-2004	1,382,908	16,917,053	12.2
wb-edu	9,845,725	57,156,537	5.8
arabic-2005	22,744,080	639,999,458	28.1
sk-2005	50,636,154	1,949,412,601	38.5
uk-2007	105,896,555	3,738,733,648	35.3

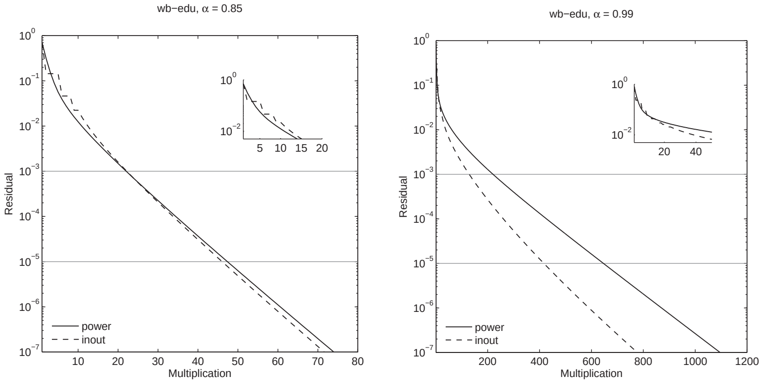


Figure: Convergence of the computation for the $9,845,725 \times 9,845,725$ wb-edu matrix ($\tau = 10^{-7}$, $\beta = 0.5$ and $\eta = 10^{-2}$ in the inner-outer method.) The interior figures highlight the first few iterations.

Table: $\alpha = 0.99$, $\beta = 0.5$, $\eta = 10^{-2}$; 8-core parallel code.

tol.	graph	work (mults.)			time (secs.)		
		power	in/out	gain	power	in/out	gain
10^{-3}	ubc-cs-2006	226	141	37.6%	1.9	1.2	35.2%
	ubc	242	141	41.7%	13.6	8.3	38.4%
	in-2004	232	129	44.4%	51.1	30.4	40.5%
	eu-2005	149	150	-0.7%	26.9	28.3	-5.3%
	wb-edu	221	130	41.2%	291.2	184.6	36.6%
	arabic-2005	213	139	34.7%	779.2	502.5	35.5%
	sk-2005	156	144	7.7%	1718.2	1595.9	7.1%
	uk-2007	145	125	13.8%	2802.0	2359.3	15.8%
10^{-5}	ubc-cs-2006	574	432	24.7%	4.7	3.6	22.9%
	ubc	676	484	28.4%	37.7	27.8	26.2%
	in-2004	657	428	34.9%	144.3	97.5	32.4%
	eu-2005	499	476	4.6%	89.3	87.4	2.1%
	wb-edu	647	417	35.5%	850.6	572.0	32.8%
	arabic-2005	638	466	27.0%	2333.5	1670.0	28.4%
	sk-2005	523	460	12.0%	5729.0	5077.1	11.4%
	uk-2007	531	463	12.8%	10225.8	8661.9	15.3%
10^{-7}	ubc-cs-2006	986	815	17.3%	8.0	6.8	15.4%
	ubc	1121	856	23.6%	62.5	49.0	21.6%
	in-2004	1108	795	28.2%	243.1	179.8	26.0%
	eu-2005	896	814	9.2%	159.9	148.6	7.1%
	wb-edu	1096	777	29.1%	1442.9	1059.0	26.6%
	arabic-2005	1083	843	22.2%	3958.8	3012.9	23.9%
	sk-2005	951	828	12.9%	10393.3	9122.9	12.2%
	uk-2007	964	857	11.1%	18559.2	16016.7	13.7%

Spectral Structure of Preconditioned Matrix

For `wb-cs.stanford` with $\alpha = 0.99$, plot the eigenvalues of the preconditioned matrix

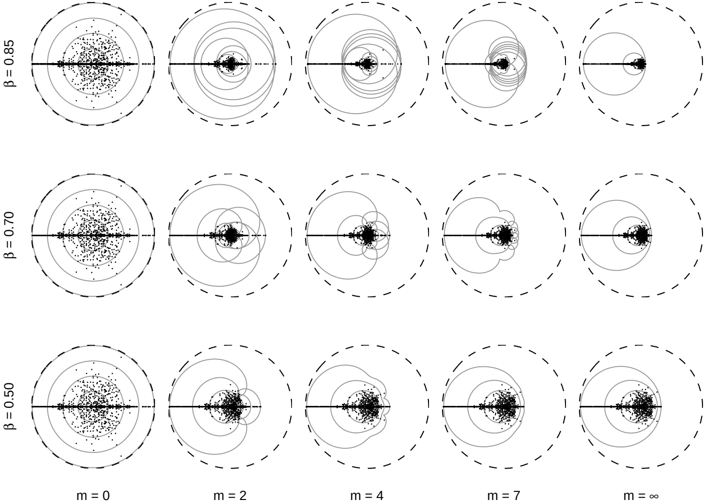
$$(I - \beta P)^{-1}(I - \alpha P)$$

and approximations based on Neumann series. Each dashed circle encloses a circle of radius 1 in the complex plane centered at $\lambda = 1$, and hence the scale is the same in each small figure. Gray lines are contours of the function

$$p_m(\lambda) = (1 - \alpha\lambda)(1 + \beta\lambda + \cdots + (\beta\lambda)^m),$$

which is the identity matrix for $m = 0$ (i.e. no preconditioning) and the exact inverse when $m = \infty$. The dots are the eigenvalues of the preconditioned system, $p_m(\lambda_i)$.

wb-cs.stanford



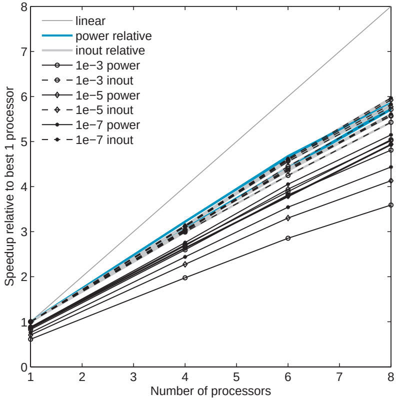


Figure: Parallel scalability for the three large graphs arabic-2005, sk-2005, and uk-2007. Light gray or teal lines are the relative speedup compared with the same algorithm on one processor. Black lines are the true speedup compared with the best single processor code. The parameters were $\alpha = 0.99, \beta = 0.5$, and $\eta = 10^{-2}$.

Conclusions

- Large scale computation of stationary distribution vectors of parameterized Markov chains requires fast and simple methods, with minimal memory requirements.
- Presented an inner/outer stationary iteration which indeed is simple, fast and requires minimal overhead: no permutations, projections, orthogonalizations or decompositions of any sort are involved.
- Programming the method is easy, and it is straightforward parallelizable.
- The algorithm is parameter-dependent, but an effective choice of the parameters can be made.
- Detailed convergence analysis.

Conclusions (Cont.)

- The proposed technique is effective for a large range of inner tolerances, particularly for large values of the damping factor.
- The gains are often made in the initial iterates.
- Any iterative solver that is effective for the linear system formulation of PageRank computations can have inner-outer iterations incorporated to possibly accelerate convergence.
- Future work: how to dynamically determine the parameters β and η, and explore the performance of the algorithm as an acceleration technique for a variety of methods, and for other problems of a similar flavour to PageRank.
- Code is available to download and test at

  `http://www.stanford.edu/~dgleich/publications/2009/innout/`.
- Paper is now published in SISC (2010).

THANK YOU!