Link Analysis, EigenVectors and Stability

- Andrew Ng, Alice Zheng, Michael Jordan - U.C. Berkeley
- Presented by Aditya Pakki in CSci 8363
- 10/25/2017

Presentation Outline

- Introduction
- Experiment Overview
- Algorithm Overview
- HITS analysis under perturbation
- PageRank Analysis under perturbation
- LSI and HITS
- Experiments
- Conclusion

Link Analysis

- Wikipedia definition - Data analysis technique to understand the relationships between nodes & links
- Sample applications include
  - Object classification - Labeling
  - Object ranking - HITS, PageRank
  - Prediction - Recommendation Systems
- Used in Citation analysis, Web page ranking, Social network analysis

EigenVector methods

- A linear transformation which changes the magnitude of vector v, v is eigenvector
- $Av = \lambda v$, λ - eigenvalue, (λ, v) - eigenpair of A
- HITS & PageRank, eigenvector methods, perform Link Analysis ranking

Stability

- **Subjective** Get experts from domain to validate output of algorithms
- **Objective** How consistent are algorithms in a perturbed environment
- Stability is a necessary feature is dynamic & unstable environment as the Internet
- We evaluate objectively in this paper

Stable algorithms are better

1	"Genetic algorithms in search, optimization...", Goldberg	1	3	1	1	1
2	"Adaptation in natural and artificial systems", Holland	2	5	3	3	2
3	"Genetic programming: On the programming of...", Koza	3	12	6	6	3
4	"Analysis of the behavior of a class of genetic...", De Jong	4	52	20	23	4
5	"Uniform crossover in genetic algorithms", Syswerda	5	171	119	99	5
6	"Artificial intelligence through simulated...", Fogel	6	135	56	40	8
7	"A survey of evolution strategies", Back+al	10	179	159	100	7
8	"Optimization of control parameters for genetic...", Grefenstette	8	316	141	170	6
9	"The GENITOR algorithm and selection pressure", Whitley	9	257	107	72	9
10	"Genetic algorithms + Data Structures = ...", Michalewicz	13	170	80	69	18
11	"Genetic programming II: Automatic discovey...", Koza	7	-	-	-	10
2060	"Learning internal representations by error...", Rumelhart+al	-	1	2	2	-
2061	"Learning to predict by the method of temporal...", Sutton	-	9	4	5	-
2063	"Some studies in machine learning using checkers", Samuel	-	-	10	10	-
2065	"Neuronlike elements that can solve difficult...", Barto+Sutton	-	-	8	-	-
2066	"Practical issues in TD learning", Tesauro	-	-	9	9	-
2071	"Pattern classification and scene analysis", Duda+Hart	-	4	7	7	-
2075	"Classification and regression trees", Breiman+al	-	2	5	4	-
2117	"UCI repository of machine learning databases", Murphy+Aha	-	7	-	8	-
2174	"Irrelevant features and the subset selection...", John+al	-	8	-	-	-
2184	"The CN2 induction algorithm", Clark+Niblett	-	6	-	-	-
2222	"Probabilistic reasoning in intelligent systems", Pearl	-	10	-	-	-

Figure 1: HITS under perturbation for 5 datasets

1	"Genetic Algorithms in Search, Optimization and...", Goldberg	1	1	1	1	1
2	"Learning internal representations by error...", Rumelhart+al	2	2	2	2	2
3	"Adaptation in Natural and Artificial Systems", Holland	3	5	6	4	5
4	"Classification and Regression Trees", Breiman+al	4	3	5	5	4
5	"Probabilistic Reasoning in Intelligent Systems", Pearl	5	6	3	6	3
6	"Genetic Programming: On the Programming of ...", Koza	6	4	4	3	6
7	"Learning to Predict by the Methods of Temporal ...", Sutton	7	7	7	7	7
8	"Pattern classification and scene analysis", Duda+Hart	8	8	8	8	9
9	"Maximum likelihood from incomplete data via...", Dempster+al	10	9	9	11	8
10	"UCI repository of machine learning databases", Murphy+Aha	9	11	10	9	10
11	"Parallel Distributed Processing", Rumelhart+McClelland	-	-	-	10	-
12	"Introduction to the Theory of Neural Computation", Hertz+al	-	10	-	-	-

Figure 2: PageRank under perturbation for the same 5 datasets

Experiment Overview

- Cora Database with thousands of papers & citations in AI
- Left most column is ranking on whole dataset
- Rank papers using HITS & PageRank after randomly deleting 30% of data
- PageRank is stable under perturbation

HITS algorithm Overview

- Article has high "authority" if linked by high weight "hubs"
- Similarly it has high hub score if it links to many authorities
- HITS algorithm
  - Construct a n*n adjacency matrix
  - Initialize the hubs & authorities as $[1, 1, \ldots, 1]^T$
  - Iterate to convergence updating hubs & authority weights
  - $a_i^{t+1} = \Sigma_{j:j \to i} h_j^t$
  - $h_i^{t+1} = \Sigma_{j:i \to j} a_j^{t+1}$

HITS algorithm contd.

- $a^{(t+1)} = A^T h^{(t)} = (A^T A) a^{(t)}$
- $h^{(t+1)} = A a^{(t+1)} = (A A^T) h^{(t)}$
- a^*, h^* are principal eigenvectors of $A^T A, A A^T$ respectively
- This is power iteration to get a principal eigen vector

Page Rank Algorithm overview

- The basis for Google's initial search algorithm
- Given n interlinked pages, rank them in order of importance
- Ordering performed by computing the PR scores for pages
- **Idea:** Use the link structure of the web

Page Rank continued - I

- Start with Adj Matrix A , normalize each row to get M, probability transition matrix
- Equivalent to random surfer jumping linked web pages with probability $1 - \epsilon$, reset web page with probability ϵ
- ϵ typically 0.1 - 0.2
- Markov matrix M - column vectors are transition probabilities
- $x_{k+1} = Mx_k$ gives a Markov Chain for x_k vector.

Page Rank continued - II

- Transition Matrix X $= \epsilon U + (1-\epsilon)M$, $U_{ij} = \frac{1}{n}; \forall i, j$
- PR scores vector p - principal eigen vector of X^T
- $(\epsilon U + (1-\epsilon)M)^T p = p$

Analysis of Algorithms - Example

- Assume algore.com has 100 links, georgebush.com has 103 links, rest are 0. Two eigen vectors, rest are 0.
- Add 5 new links pointing to both the web pages
- Original eigen vectors in Fig 1a, new Eigen Vector in 1(b)
- **Small perturbations causes large change in Eigen vectors**

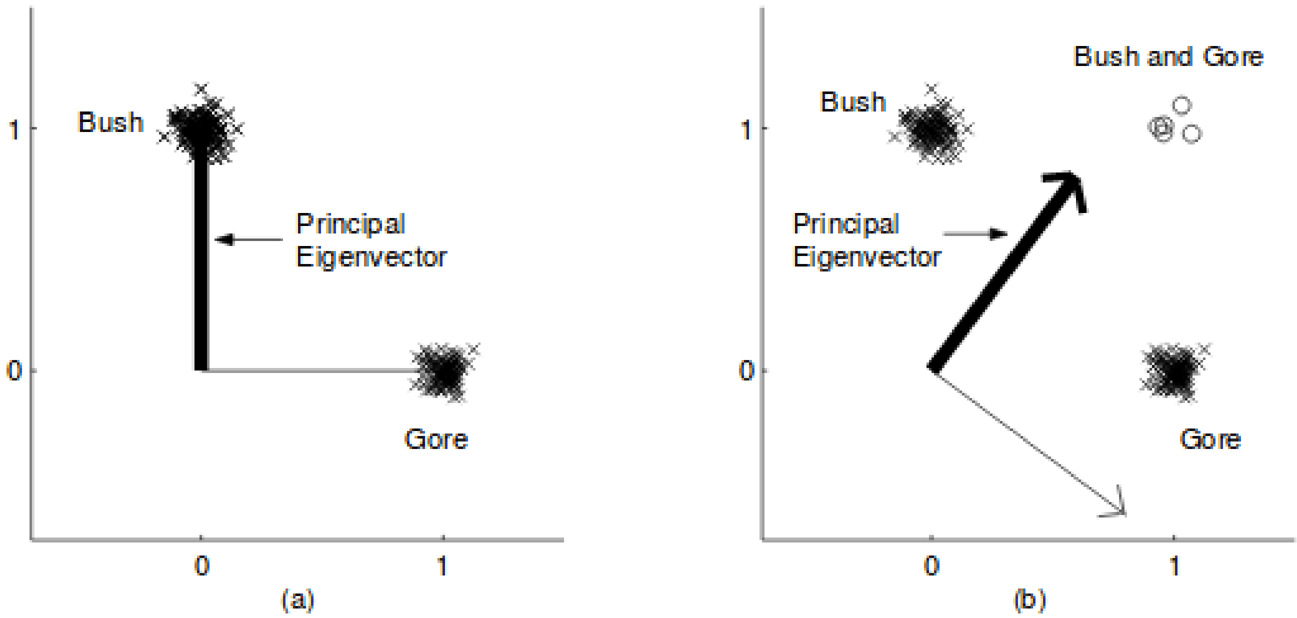


Figure 1: Jittered scatterplot of hyperlink graph.

Analysis of HITS algorithm

- Eigengap $\delta = \lambda_1 - \lambda_2$.
- Matrix S_1 in 2(a)- $\delta_1 \approx 0$, Matrix S_2 in 2(b) $\delta_2 = 2$
- Larger the δ, smaller the impact of perturbations to HITS
- Equivalent to - second or smaller EV can never be principal EV during perturbations

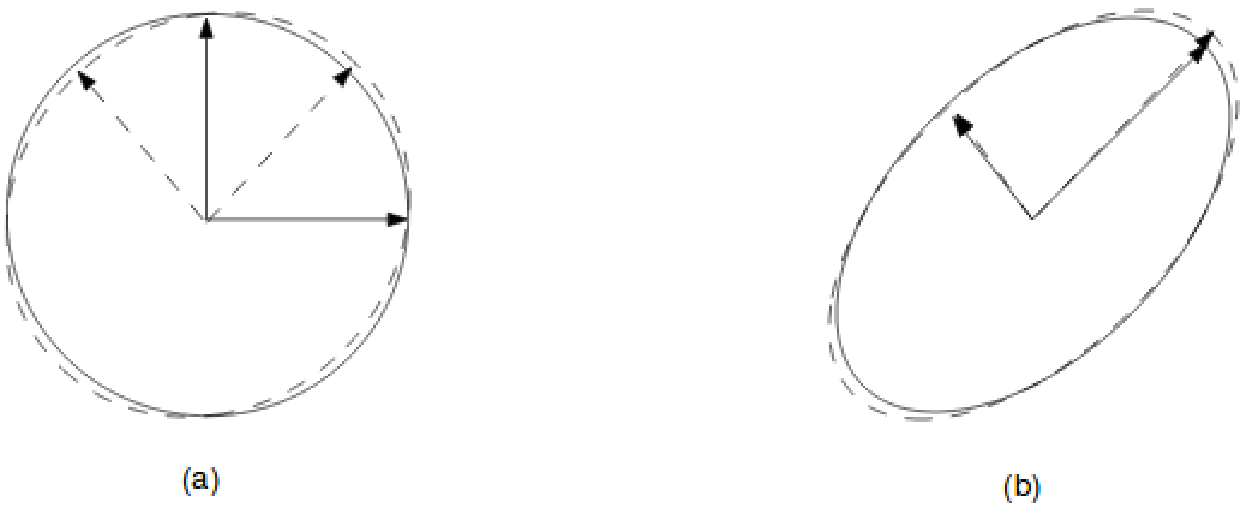


Figure 2: Contours of two matrices with different eigengaps.

Theorem 1. *Let $S = A^T A$ be given. Let a^* be the principal eigenvector and δ the eigengap of S. Assume the maximum out-degree of every web page is bounded by d. For any $\varepsilon > 0$, suppose we perturb the web/citation graph by adding or deleting at most k links from one page, where $k < (\sqrt{d+\alpha} - \sqrt{d})^2$, where $\alpha = \varepsilon\delta/(4 + \sqrt{2}\varepsilon)$. Then the perturbed principal eigenvector $\tilde{a}^*$ of the perturbed matrix $\tilde{S}$ satisfies:*

$$||a^* - \tilde{a}^*||_2 \leq \varepsilon \tag{2}$$

For the eigenpair (λ^*, a^*) and perturbed eigenpair$(\widetilde{\lambda}, \widetilde{a})$,we have the following two properties

$||a^* - \widetilde{a}||_2 \leq \frac{4||E||_F}{\delta - \sqrt{(2)}||E||_F}$ and

$|\lambda^* - \widetilde{\lambda}| \leq \sqrt{2}||E||_F$

Let (L_2, X_2) be eigen space where X_2 is orthonormal containing eigenvectors other than a^* & L_2 the diagonal matrix of those eigen vectors; $SX_2 = X_2L_2$. Similarly

$$||L_2 - \widetilde{L_2}||_F \leq \sqrt{2}||E||_F$$
$$\implies \widetilde{\lambda}_2 \leq \lambda_2 + \sqrt{2}||E||_F$$

We can bound the norm of the perturbation to S by

$||E||_F = ||\widetilde{S} - S||_F \leq k + 2\sqrt{dk}$

Theorem 1 Proof Contd.

- Substituting $||E||_F$ in $||a^* - \widetilde{a}||_2$ eqn, we get a bound $k \leq (\sqrt{d+\alpha} - \sqrt{d})^2, \alpha = \frac{\epsilon\delta}{(4+\sqrt{2}\epsilon)}$
- In Fig 3, we see small sub-community with links in solid arrows; dashed arrows are after perturbation
- Principal EV is 20, by addition of new link, $\widetilde{a^*}$ is now 25.
- If a larger community exists with $20 < \lambda_1 < 25$, with the addition of below community, $\widetilde{\lambda_1}$ is now from this sub-community.
- **Principal EV $\widetilde{a^*}$ now has values only for those nodes and zeros elsewhere**

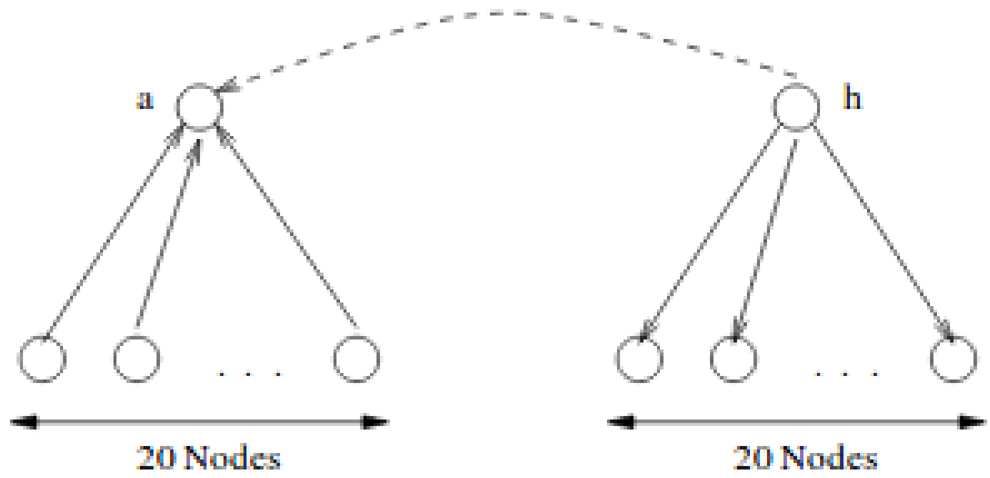


Figure 3: Picture of a web community.

Converse to Theorem 1

Theorem 2. Suppose S is a symmetric matrix with eigengap δ. Then there exists a $O(\delta)$ perturbation to S that causes a large ($\Omega(1)$) change in the principal eigenvector.
Proof:

- Since $S = A^T A$, using SVD decomposition
$S = U \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \Sigma \end{bmatrix} V^T$
- For an orthonormal col u_i in U, we have $\widetilde{S} = S + 2\delta u_2 u_2^T$.
$||2\delta u_2 u_2^T||_F = 2\delta$
- $\widetilde{S} = U \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 + 2\delta & 0 \\ 0 & 0 & \Sigma \end{bmatrix} V^T$
- $\widetilde{\lambda_2} = \lambda_2 + 2\delta > \lambda_1$,
$\implies (\widetilde{\lambda_2}, u_2)$ is the perturbed principal eigenpair.
- u_2, u_1 are orthonormal, so $||u_2 - u_1||_2 = \Omega(1)$

Page Rank Perturbation Analysis

Theorem 3. *Let M be given, and let p be the principal right eigenvector of $(\epsilon U + (1-\epsilon)M)^T$. Let articles/pages $i_1, i_2, \ldots, i_k$ be changed in any way, and $\tilde{M}$ be the corresponding (new) transition matrix. Then the new PageRank scores $\tilde{p}$ satisfies:*

$$||\tilde{p} - p||_1 \leq \frac{2\sum_{j=1}^{k} p_{i_j}}{\epsilon} \tag{8}$$

- $(X_t, Y_t) : t \geq 0$ be two coupled Markov Chains, $X_0 = Y_0$
- At time t, reset $X_t = Y_t$ to same page with probability ϵ, or if $X_{t-1} = Y_{t-1}, \& X_{t-1}$ is an unperturbed page, $X_t = Y_t$
- Otherwise $X_{t-1} \to X_t, Y_{t-1} \to Y_t$ independently at random
- $X_t = (\epsilon U + (1-\epsilon)M)^T; Y_t = (\epsilon U + (1-\epsilon)\widetilde{M})^T$
- Resets are in lock steps to both the Markov chains but distribution of $X_t = p, Y_t = \widetilde{p}$

Page Rank Analysis Contd.

- $d_t = P(X_t \neq Y_t); d_0 = 0$, With $\mathcal{P}$ be set of perturbed pages
- To get a dissimilar page at $t+1$, possible only when $X_t \in \mathcal{P}$
- $P(X_\infty \neq Y_\infty)$ is the upper bound $d_\infty \leq \frac{\Sigma_{i \in \mathcal{P}} p_i}{\epsilon}$
- Two random variables have d_∞ chance of diverging $\implies$ $\frac{1}{2}\Sigma_i ||p_i - \widetilde{p}_i||_1 < d_\infty$

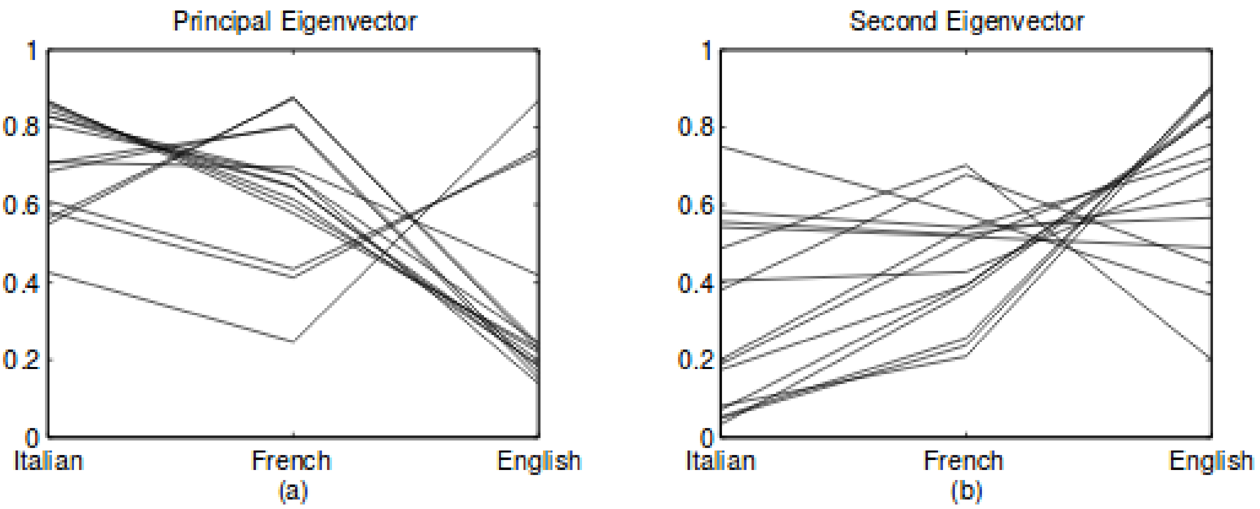
$$
\begin{aligned}
d_{t+1} &= P(X_{t+1} \neq Y_{t+1}) \\
&= P(X_{t+1} \neq Y_{t+1} | \text{reset at } t+1) P(\text{reset}) \\
&\quad + P(X_{t+1} \neq Y_{t+1} | \text{no reset at } t+1) P(\text{no reset}) \\
&= 0 \cdot \epsilon + (1-\epsilon) P(X_{t+1} \neq Y_{t+1} | \text{no reset at } t+1) \\
&= (1-\epsilon)[P(X_{t+1} \neq Y_{t+1}, X_t \neq Y_t | \text{no reset at } t+1) \\
&\quad + P(X_{t+1} \neq Y_{t+1}, X_t = Y_t | \text{no reset at } t+1)] \\
&\leq (1-\epsilon)[P(X_t \neq Y_t | \text{no reset at } t+1) \\
&\quad + P(X_{t+1} \neq Y_{t+1}, X_t = Y_t, X_t \in \mathcal{P} | \text{no reset at } t+1)] \\
&\leq (1-\epsilon)(P(X_t \neq Y_t) + P(X_t \in \mathcal{P} | \text{no reset at } t+1)) \\
&\leq (1-\epsilon)(d_t + \textstyle\sum_{i \in \mathcal{P}} p_i)
\end{aligned}
$$

LSI and HITS

- LSI - represent a document set and word frequency per document in a matrix
- Group synonyms and in turn reduce subspace during Info retrieval
- Represent doc set & words as nodes, with link from node to doc it appears
- Apply HITS, word nodes have positive hub weights, docs have positive authority weights
- Recall - hubs have out links , authority have in links
- Left singular vector of LSI is hub weights

Lessons from LSI to HITS

- Corpora of English, French, Italian sets to test HITS EV direction
- Principal EV in high dimensional space and 4(a),4(b) show them in each language direction
- We see no order for the Eigen Vector for 15 runs even in presence of clusters



Experiments

- Use Cora database containing AI papers
- Choose a subset from Cora and perturb by deleting 30% of data
- Perform 5 such runs on HITS & PageRank. Page Rank is stable and HITS authority scores changes drastically
- Similar results on web pages

Cora Dataset perturbations for HITS & PageRank

1	"Classification and Regression Trees", Brieman+al	1	1	1	1	1
2	"Pattern classification and scene analysis", Duda+Hart	2	2	3	2	2
3	"UCI repository of machine learning databases", Murphy+Aha	4	3	7	3	3
4	"Learning internal representations by error...", Rumelhart+al	3	13	2	28	20
5	"Irrelevant Features and the Subset Selection Problem", John+al	7	4	12	4	4
6	"Very simple classification rules perform well on...", Holte	8	5	15	5	5
7	"C4.5: Programs for Machine Learning", Quinlan	11	10	14	10	6
8	"Probabilistic Reasoning in Intelligent Systems", Pearl	6	459	4	462	461
9	"The CN2 induction algorithm", Clark+Niblett	9	54	11	78	105
10	"Learning Boolean Concepts in the ...", Almuallim+Dietterich	14	11	34	9	13
11	"The MONK's problems: A performance comparison...", Thrun	-	9	-	6	7
12	"Inferring decision trees using the MDL Principle", Quinlan	-	8	-	7	8
13	"Multi-interval discretization of continuous...", Fayyad+Irani	-	-	-	-	10
14	"Learning Relations by Pathfinding", Richards+Moon	-	6	-	-	-
15	"A conservation law for generalization performance", Schaffer	-	7	-	8	-
20	"The Feature Selection Problem: Traditional..." Kira+Randall	-	-	-	-	9
21	"Maximum likelihood from incomplete data via..." Dempster+al	10	-	5	-	-
23	"Learning to Predict by the Method of Temporal...", Sutton	5	-	6	-	-
36	"Introduction to the Theory of Neural Computation", Hertz+al	-	-	8	-	-
49	"Explanation-based generalization: a unifying view", Mitchell	-	-	10	-	-
282	"A robust layered control system for a mobile robot", Brooks	-	-	9	-	-

Figure 3: HITS experiment runs

1	"Classification and Regression Trees", Breiman+al	1	1	1	1	2
2	"Probabilistic Reasoning in Intelligent Systems", Pearl	3	2	2	2	1
3	"Learning internal representations by error...", Rumelhart+al	2	3	3	3	3
4	"Pattern classification and scene analysis", Duda+Hart	4	4	4	4	4
5	"A robust layered control system for a mobile robot", Brooks	5	6	7	5	5
6	"Maximum likelihood from incomplete data via...' Dempster+al	6	7	6	6	6
7	"Learning to Predict by the Method of Temporal...", Sutton	7	5	5	7	7
8	"UCI repository of machine learning databases", Murphy+Aha	8	9	9	9	11
9	"Numerical Recipes in C", Press+al	10	12	8	11	8
10	"Parallel Distributed Processing", Rumelhart+al	9	14	13	10	9
12	"An implementation of a theory of activity", Agre+Chapmanre	-	8	10	8	-
13	"Introduction to the Theory of Neural Computation", Hertz+al	-	10	-	-	-
22	"A Representation and Library for Objectives in...", Valente+al	-	-	-	-	10

Figure 4: Page Rank experiment runs

Web page perturbations for HITS & PageRank

1	http://www.freecode.com/	82	1	1	1	82
2	http://www.htmlworks.com/	85	2	2	2	83
3	http://www.internettrafficreport.com/	86	3	4	3	85
4	http://slashdot.org/	88	4	5	5	86
5	http://windows.davecentral.com/	87	5	3	4	84
6	http://www.gifworks.com/	84	6	6	6	87
7	http://www.thinkgeek.com/	91	7	7	7	88
8	http://www.animfactory.com/	89	9	8	8	89
9	http://freshmeat.net/	90	8	9	9	90
10	http://subscribe.andover.net/membership.htm	92	10	10	10	91
1385	http://ourstory.about.com/index.htm	1	-	-	-	1
1386	http://home.about.com/index.htm	2	-	-	-	2
1387	http://home.about.com/musicperform/index.htm	3	-	-	-	3
1388	http://home.about.com/teens/index.htm	4	-	-	-	4
1389	http://home.about.com/sports/index.htm	5	-	-	-	5
1390	http://home.about.com/autos/index.htm	6	-	-	-	6
1391	http://home.about.com/style/index.htm	7	-	-	-	7
1392	http://home.about.com/careers/index.htm	8	-	-	-	8
1393	http://home.about.com/citiestowns/index.htm	9	-	-	-	9
1394	http://home.about.com/travel/index.htm	10	-	-	-	10

In contrast, PageRank returned:

1	http://www.team-mp3.com/	*	1	1	1	1
2	http://click.linksynergy.com/fs-bin/click	1	3	2	4	9
3	http://www.elizandra.com/	2	2	3	2	2
4	http://stores.yahoo.com/help.html	4	14	5	10	11
5	http://shopping.yahoo.com/	3	10	4	12	13
6	http://www.netins.net/showcase/phdss/	*	8	6	3	3
7	http://www.thecounter.com/	13	6	9	8	7
8	http://ourstory.about.com/index.htm	5	4	7	5	4
9	http://a-zlist.about.com/index.htm	6	5	10	6	6
10	http://www.netins.net/showcase/phdss/getm	*	9	8	7	5
11	http://software.mp3.com/software/	7	7	-	-	8
12	http://www.winamp.com/	8	-	-	-	-
13	http://www.nullsoft.com/	10	-	-	-	-
14	http://www.consumerspot.com/redirect/1cac	9	-	-	9	10

Conclusions

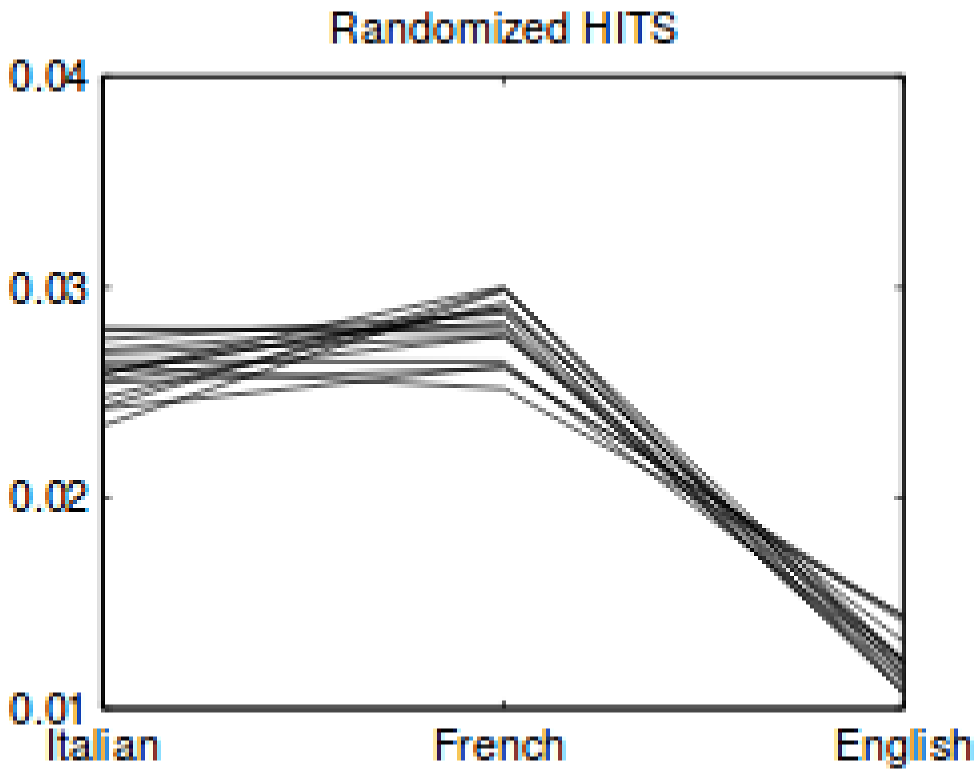
- Subspace spanned by several EV is stable under perturbation but not individually
- LSI projects data to lower subspace, stability not a priority
- Eigenvector methods sensitive to perturbation, HITS is sensitive PageRank is not
- Suggest a variation of HITS - **Randomized HITS**

Randomized HITS results on subset of Cora AI papers ($\epsilon = 0.2$):

1	"Learning internal representations by error...", Rumelhart+al	1	3	3	2	1
2	"Probabilistic Reasoning in Intelligent Systems", Pearl	4	1	1	1	2
3	"Classification and Regression Trees", Breiman+al	2	2	2	3	4
4	"Pattern classification and scene analysis", Duda+Hart	3	4	4	4	3
5	"Maximum likelihood from incomplete data via...", Dempster+al	5	6	6	6	5
6	"A robust layered control system for a mobile robot", Brook+al	6	5	5	5	6
7	"Numerical Recipes in C", Press+al	7	7	7	7	7
8	"Learning to Predict by the Method of Temporal...", Sutton	8	8	8	8	8
9	"STRIPS: A New Approach to ... Theorem Proving", Fikes+al	9	10	10	10	15
10	"Introduction To The Theory Of Neural Computation", Hertz+al	11	11	9	9	9
11	"Stochastic relaxation, gibbs distributions, ...", Geman+al	10	9	-	-	-
12	"Introduction to Algorithms", Cormen+al	-	-	-	-	10

Randomized HITS

- Combination of Markov Chain from PR & hubs, authority score from HITS
- Equivalent to coin toss with bias ϵ. If heads - go to a random webpage chosen uniformly.
- If tails, odd time step go to out-link, even timestep go to back-link
- Random walk on web pages - odd time steps give hub score, authority scores on even time step
- Below figure is for 3 language corpora set to see EV directions



Acknowledgments

The images in the presentation all belong to the original paper - Link Analysis, Eigenvectors and Stability