Linear Algebra and Data Analysis

Daniel Brice

Department of Mathematics
California State University, Bakersfield
daniel.brice@csub.edu

2 February 2016

Link Analysis

Study of the link structure of a network (WWW hyperlinks, or follows on Twitter, etc...)

- Ignores semantics (such as HTML meta tags)

```
<meta name='description' content='...' />
<meta name='keywords' content='...' />
```

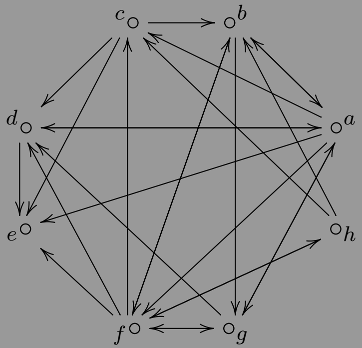
- A link (hyperlink, follow, etc...) confers relevance

$$i \to j$$

i confers some of its relevance onto j

Link Analysis/Example Network

$I =$



$$A = \begin{array}{c} \\ a \\ b \\ c \\ d \\ e \\ f \\ g \\ h \end{array} \begin{array}{c} \begin{array}{cccccccc} a & b & c & d & e & f & g & h \end{array} \\ \begin{pmatrix} 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \end{array}$$

Link Analysis/PageRank

Developed by Brin and Page in 1998, use by Google (Brin and Page 1998).

Let I be the our network

For $i \in I$, let $n_i = \left|\{j \in I; i \to j\}\right|$

$$S_{i,j} = \begin{cases} \frac{1}{n_i} & \text{if } i \to j \\ 0 & \text{otherwise} \end{cases}$$

$$S = \begin{array}{cc} & \begin{array}{cccccccc} a & b & c & d & e & f & g & \end{array} \\ \begin{array}{c} a \\ b \\ c \\ d \\ e \\ f \\ g \\ h \end{array} & \begin{pmatrix} 0.000 & 0.167 & 0.167 & 0.167 & 0.167 & 0.167 & 0.167 & 0.000 \\ 0.333 & 0.000 & 0.000 & 0.000 & 0.000 & 0.333 & 0.333 & 0.000 \\ 0.000 & 0.333 & 0.000 & 0.333 & 0.333 & 0.000 & 0.000 & 0.000 \\ 0.500 & 0.000 & 0.000 & 0.000 & 0.500 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.167 & 0.167 & 0.167 & 0.167 & 0.000 & 0.167 & 0.167 \\ 0.333 & 0.000 & 0.000 & 0.333 & 0.000 & 0.333 & 0.000 & 0.000 \\ 0.000 & 0.333 & 0.333 & 0.000 & 0.000 & 0.333 & 0.000 & 0.000 \end{pmatrix} \end{array}$$

Link Analysis/PageRank/Perron-Frobenius Theorem

M, non-negative matrix, called **irreducible** if M cannot be permuted to block upper triangular form.
spectral radius of M is $r = \max\{|\lambda|; \lambda \text{ an eigenvalue of } M\}$

Theorem (Perron-Frobenius Theorem)

Let $M \in \mathbb{R}^{n \times n}$ be irreducible. Denote by r the spectral radius of M. Then:

1. *r is an eigenvalue of M, and it is the unique eigenvalue of M with maximum absolute value.*
2. *M has a unique (up to scalar) eigenvector p with positive entries, and the eigenvalue corresponding to p is r.*

Link Analysis/PageRank/Dampening Factor

Pick arbitrary *dampening factor* δ (we're using $\delta = 0.9$)

$$S'_{i,j} = \begin{cases} \frac{\delta}{n_i} & \text{if } i \to j \\ \frac{1-\delta}{|I|-n_i} & \text{otherwise} \end{cases}$$

$$S' = \begin{array}{c} \\ a \\ b \\ c \\ d \\ e \\ f \\ g \\ h \end{array} \begin{array}{c} \begin{array}{cccccccc} a & b & c & d & e & f & g & h \end{array} \\ \begin{pmatrix} 0.05 & 0.15 & 0.15 & 0.15 & 0.15 & 0.15 & 0.15 & 0.05 \\ 0.3 & 0.02 & 0.02 & 0.02 & 0.02 & 0.3 & 0.3 & 0.02 \\ 0.02 & 0.3 & 0.02 & 0.3 & 0.3 & 0.02 & 0.02 & 0.02 \\ 0.45 & 0.0167 & 0.0167 & 0.0167 & 0.45 & 0.0167 & 0.0167 & 0.0167 \\ 0.125 & 0.125 & 0.125 & 0.125 & 0.125 & 0.125 & 0.125 & 0.125 \\ 0.05 & 0.15 & 0.15 & 0.15 & 0.15 & 0.05 & 0.15 & 0.15 \\ 0.3 & 0.02 & 0.02 & 0.3 & 0.02 & 0.3 & 0.02 & 0.02 \\ 0.02 & 0.3 & 0.3 & 0.02 & 0.02 & 0.3 & 0.02 & 0.02 \end{pmatrix} \end{array}$$

Link Analysis/PageRank/Results

$$p = \begin{pmatrix} 0.12499743671351870589 \\ 0.12499523049915370332 \\ 0.12499967068187360641 \\ 0.12501904438727801105 \\ 0.12499662015324809750 \\ 0.12499716042480689404 \\ 0.12500043967713900250 \\ 0.12499439746298210419 \end{pmatrix}$$

Page	Rank
d	$p_d = 0.12501904438727801105$
g	$p_g = 0.12500043967713900250$
c	$p_c = 0.12499967068187360641$
a	$p_a = 0.12499743671351870589$
f	$p_f = 0.12499716042480689404$
e	$p_e = 0.12499662015324809750$
b	$p_b = 0.12499523049915370332$
h	$p_h = 0.12499439746298210419$

Link Analysis/PageRank/Mathematics

p is an eigenvector of S' with eigenvalue r.

$$p = r^{-1} S' p$$

$$p_k = r^{-1} \left(\delta \sum_{i \to k} \frac{p_i}{n_i} + (1 - \delta) \sum_{i \not\to k} \frac{p_i}{|I| - n_i} \right)$$

for δ close to 1

$$p_k \sim \sum_{i \to k} \frac{p_i}{n_i}$$

Link Analysis/HITS

Developed by Kleinberg in 1999 (Kleinberg 1999; Kolda, Bader, and Kenny 2005).

Runs on I', a *focused subgraph* of I.

Uses singular-value decomposition on A'.

$$A' = U\Sigma V^*$$

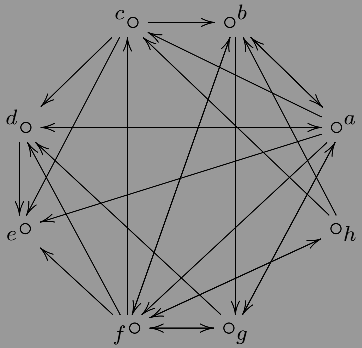
U and V unitary, $\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, ..., \sigma_n)$ with $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_n \geq 0$

σ_1 termed the **principle singular value**
u^1 and v^1 termed the **principle singular vectors**

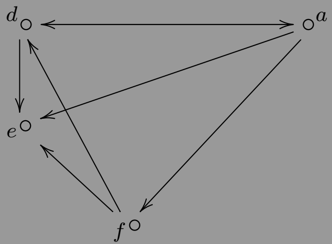
For $i \in I'$,
hub score: $h(i) = u_i^1$ (first column, ith row),
authority score: $a(i) = v_i^1$ (first column, ith row).

Link Analysis/HITS/Example

$I =$



$I' =$



$$A' = \begin{array}{c} \\ a \\ d \\ e \\ f \end{array} \begin{array}{c} \begin{array}{cccc} a & d & e & f \end{array} \\ \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{pmatrix} \end{array}$$

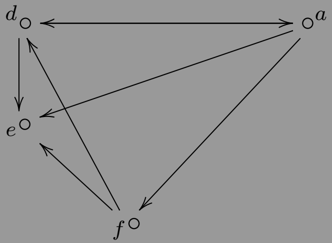
Link Analysis/HITS/Results

$$A' = \begin{array}{c} \\ a \\ d \\ e \\ f \end{array} \begin{array}{c} \begin{array}{cccc} a & d & e & f \end{array} \\ \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{pmatrix} \end{array}$$

$$U = \begin{pmatrix} 0.711 & 0.415 & 0.566 & 0 \\ 0.404 & -0.901 & 0.153 & 0 \\ 0 & 0 & 0 & 1 \\ 0.574 & 0.12 & -0.809 & 0 \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} 2.276 & 0 & 0 & 0 \\ 0 & 1.185 & 0 & 0 \\ 0 & 0 & 0.641 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

$$V = \begin{pmatrix} 0.177 & -0.76 & 0.238 & -0.577 \\ 0.565 & 0.451 & -0.379 & -0.557 \\ 0.742 & -0.309 & -0.14 & 0.557 \\ 0.312 & 0.35 & 0.882 & 0 \end{pmatrix}$$



Page	Hub Rank	Authority Rank
a	0.711	0.177
d	0.404	0.565
e	0	0.742
f	0.574	0.312

Link Analysis/HITS/Mathematics

Singular value decomposition

$$A' = U \Sigma V^*$$

Solve for u^1

$$u^1 = {\sigma_1}^{-1} A' v^1$$

$$h(i) \sim \sum_{i \to j} a(j)$$

Solve for $(v^1)^*$

$$(v^1)^* = {\sigma_1}^{-1} (u^1)^* A'$$

$$a(i) \sim \sum_{j \to i} h(j)$$

Image Analysis

Try to algorithmically recognize features in an image.

(http://www.eoas.ubc.ca/research/cdsst/Tad_home/)

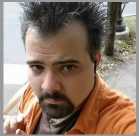
Usually involves machine learning, using *training data* as a basis of comparison.
The results obtained from the training data are compared to *test data* coming in.

Image Analysis/Eigenfaces

Introduced by Sirovich and Kirby in 1987 and subsequently developed by Turk and Pentland (Turk and Pentland 1991; Barrett et al. 1997).

Every image is a vector.

A 480x640 greyscale image lives in a 307,200 dimension space.
For RGB, multiply by 3. For HD images, multiply by 4.



However, not all possible images are a face.

How can we isolate the subspace consisting only of faces?

Image Analysis/Eigenfaces/Mathematics

Use singular value decomposition (called *principal component analysis* in image analysis circles.)

Theorem (Singular Value Decomposition)

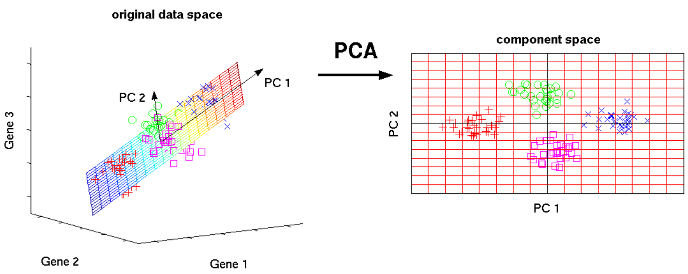
Let M *be an* $n \times m$ *matrix over* $\mathbb{C}$ *with rank* k*. There is an* $n \times n$ *unitary matrix* U*, an* $m \times m$ *unitary matrix* V*, and an* $n \times m$ *rectangular diagonal matrix* $\Sigma = \mathrm{diag}(\sigma_1, ..., \sigma_k, 0, ..., 0)$ *where each* σ_i *is real and* $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_k$ *such that*

$$M = U\Sigma V^*.$$

SVD decomposes M as a sum of rank-one matrices.

$$M = \sigma_1 u^1 (v^1)^* + \sigma_2 u^2 (v^2)^* + ... + \sigma_k u^k (v^k)^*$$

Image Analysis/Eigenfaces/Principal Component Analysis



- Each image is a datapoint in a 300k+ dimension space.
- Want to find best-fit hyperplane for data.
- Organize all datapoints into one large matrix.
- SVD gives orthonormal basis for best-fit hyperplane *eigenfaces*.
- Existing datapoints compactly stored as linear combination of eigenfaces.
- Future datapoints compared to eigenfaces instead of to entire data set.

Image Analysis/Eigenfaces/Examples

(http://mikedusenberry.com/)

$\mapsto$

(http://mikedusenberry.com/)

Is Toomer's Corner Being Rolled Right Now (.com)

Live feed:

http://www.auburnalabama.org/mvc/cams/City-Cameras/Toomer%27s-Corner

References I

- Brin, Sergey and Lawrence Page (1998). "The anatomy of a large-scale hypertextual Web search engine". In: *Computer networks and ISDN systems* 30.1, pp. 107–117.
- Kleinberg, Jon M (1999). "Authoritative sources in a hyperlinked environment". In: *Journal of the ACM (JACM)* 46.5, pp. 604–632.
- Kolda, Tamara G, Brett W Bader, and Joseph P Kenny (2005). "Higher-order web link analysis using multilinear algebra". In: *Data Mining, Fifth IEEE International Conference on*. IEEE, 8–pp.
- Turk, Matthew and Alex Pentland (1991). "Eigenfaces for recognition". In: *Journal of cognitive neuroscience* 3.1, pp. 71–86.
- Barrett, William et al. (1997). "A survey of face recognition algorithms and testing results". In: *Signals, Systems & Computers, 1997. Conference Record of the Thirty-First Asilomar Conference on*. Vol. 1. IEEE, pp. 301–305.
- Horn, Roger A. and Charles R. Johnson (1985). *Matrix Analysis*. Cambridge University Press, Cambridge, pp. xiii+561. ISBN: 0-521-30586-1.