# Mathematical Data Science

## Numerical Linear Algebra for Big Data

Martin van Gijzen

Delft University of Technology

March 29 2018

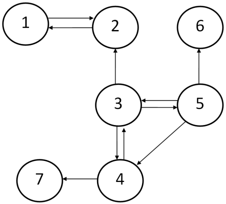
# Outline

# The PageRank and Linkspamming

- The PageRank of a webpage determines the importance of that page
- A high PageRank makes that page easy to find
- Linkspamming is the name for a collection of techniques to increase the PageRank
- In this lecture we will discuss
    - The PageRank model
    - How to manipulate it
    - How to detect that it has been manipulated

# A model of the websurfer

Webpage's are connected through hyperlinks.




Figuur: Model of part of the web

To model this mathematically we introduce

- The binary matrix $\mathbf{G}$, with $\mathbf{G}_{i,j} = 1$ if there is a link from page $j$ to page $i$. This matrix is the representation of a directed graph. Pages are nodes of this graph.
- The row-stochastic transition matrix $\mathbf{P}$, that gives the probability of going from one state to the next.

# The matrices $\mathbf{G}$ and $\mathbf{P}$

- We assume that the outlinks have equal probability to be followed. So if the number of outlinks for page $j$ is equal to $d_j$ then $\mathbf{P}_{j,i} = \mathbf{G}_{i,j}/d_j$ if $d_j \neq 0$.
- A *dangling node* is a webpage without outlinks, i.e. $d_j = 0$. We assume that these are connected to every webpage in the web with equal probability. In that case we get that $\mathbf{P}_{j,i} = \frac{1}{N}$, with $N$ the total number of pages.

# The matrices for the example

$$\mathbf{G} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

# The matrices for the example

$$\mathbf{G} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

$$\mathbf{P}^T = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 1/7 & 1/7 \\ 1 & 0 & 1/3 & 0 & 0 & 1/7 & 1/7 \\ 0 & 0 & 0 & 1/2 & 1/3 & 1/7 & 1/7 \\ 0 & 0 & 1/3 & 0 & 1/3 & 1/7 & 1/7 \\ 0 & 0 & 1/3 & 0 & 0 & 1/7 & 1/7 \\ 0 & 0 & 0 & 0 & 1/3 & 1/7 & 1/7 \\ 0 & 0 & 0 & 1/2 & 0 & 1/7 & 1/7 \end{pmatrix}$$

# Teleportation

Teleportation is the name for jumping to a webpage without following a link.
To account for this behaviour the transition matrix is modified as follows

$$\mathbf{A} = p\mathbf{P}^T + \frac{1-p}{N}\mathbf{e}\mathbf{e}^T$$

Here the vector $\mathbf{e}$ contains all ones.
The first term says that the surfer follows an outlink with probability $p$ and the second term that the surfers teleports with probability $1-p$.

# The PageRank

- The PageRank vector is the probability vector after infinitely long surfing.
- The PageRank vector is the dominant eigenvector of the Google matrix $\mathbf{A}$.
- By the Perron-Frobenius theorem for positive matrices, $\mathbf{A}$ has a unique largest eigenvalue. Since $\mathbf{A}$ is column-stochastic this eigenvalue is $\lambda_1 = 1$.
- The corresponding eigenvector is positive and, if properly scaled, stochastic.
- The rank of a page is the index in the ordered Pagerank vector.

# Computation of the PageRank vector

- The standard way to compute the PageRank vector is by the Power method.

# Computation of the PageRank vector

- The standard way to compute the PageRank vector is by the Power method.
- Since we know that $\lambda_1 = 1$, the PageRank vector can also be computed by solving a linear system (proposed by Moler):

$$(p\mathbf{P}^T + \frac{1-p}{N}\mathbf{e}\mathbf{e}^T)\mathbf{x}^{(\mathbf{1})} = \mathbf{x}^{(\mathbf{1})} \quad \Rightarrow$$

$$(\mathbf{I} - p\mathbf{P}^T)\mathbf{x}^{(\mathbf{1})} = \frac{1-p}{N}\mathbf{e}$$

The solution $\mathbf{x}$ must be scaled to make it stochastic.

# Link spamming

Next we will discuss how to increase the PageRank by changing the link structure.

# Irreducible closed subchains

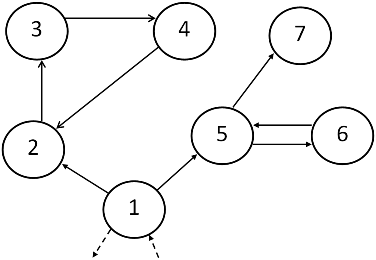
An irreducible closed subchain (subgraph):

- Once you enter a closed subchain you cannot leave;
- Every node in an irreducible subchain can be reached.
- Nodes in an irreducible closed subchain receive a high PageRank value.

# Irreducible closed subchains

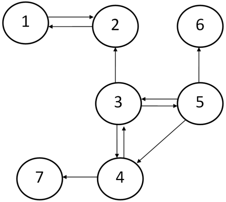
An irreducible closed subchain (subgraph):

- Once you enter a closed subchain you cannot leave;
- Every node in an irreducible subchain can be reached.
- Nodes in an irreducible closed subchain receive a high PageRank value.

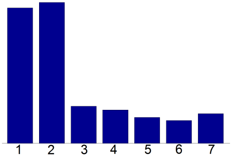



Figuur: Irreducible closed subchains?

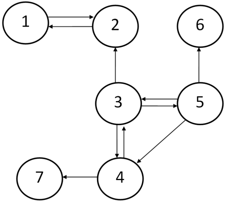
# Linkspamming: node 4



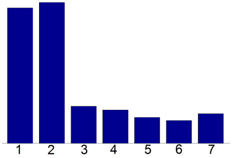

met $\mathbf{x}^{(1)} =$



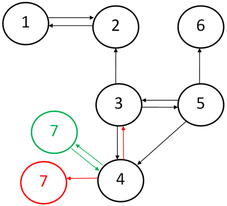
# Linkspamming: node 4



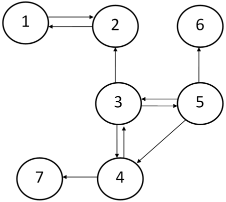

met $\mathbf{x}^{(1)} =$

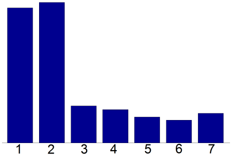



$\Downarrow$

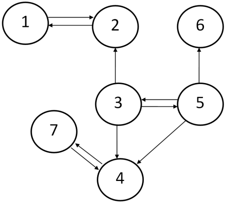


# Linkspamming: node 4

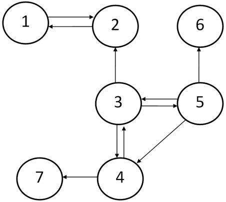



met $\mathbf{x}^{(1)} =$ 

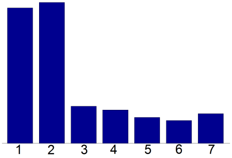

$$\Downarrow$$



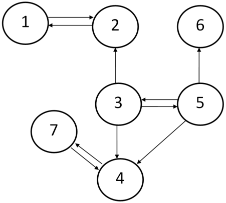
# Linkspamming: node 4



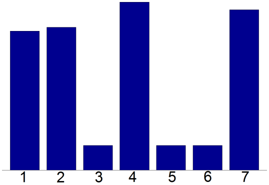

met $\mathbf{x}^{(1)} =$




$\Downarrow$




met $\mathbf{x}^{(1)} =$



# Tarjan's algorithm

This type of link spamming can be detected by analyzing the graph of the (original) web:

- Find the strongly connected components of the graph (SCC: every node in an SSC can be reached form any of the other nodes).
- Strongly connected components that contain one node are dangling nodes.
- A group of strongly connected components without outlinks is an irreducible closed subchain.

An efficient algorithm for computing the strongly connected components of a graph is *Tarjan's algorithm*. Its complexity is linear in the number of nodes.

# The matrix P

- The graph will in general contain many irreducible closed subchains;

# The matrix P

- The graph will in general contain many irreducible closed subchains;
- The matrix $\mathbf{P}$ can therefore be permuted to:

$$\mathbf{P} = \left(\begin{array}{cccc|cccc}
\mathbf{P_{11}} & \mathbf{P_{12}} & \cdots & \mathbf{P_{1r}} & \mathbf{P_{1,r+1}} & \mathbf{P_{1,r+2}} & \cdots & \mathbf{P_{1m}} \\
\mathbf{0} & \mathbf{P_{22}} & \cdots & \mathbf{P_{2r}} & \mathbf{P_{2,r+1}} & \mathbf{P_{2,r+2}} & \cdots & \mathbf{P_{2m}} \\
\vdots & & \ddots & \vdots & \vdots & \vdots & \cdots & \vdots \\
\mathbf{0} & \mathbf{0} & \cdots & \mathbf{P_{rr}} & \mathbf{P_{r,r+1}} & \mathbf{P_{r,r+2}} & \cdots & \mathbf{P_{rm}} \\
\hline
\mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{P_{r+1,r+1}} & \mathbf{0} & \cdots & \mathbf{0} \\
\mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{P_{r+2,r+2}} & \cdots & \mathbf{0} \\
\vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
\mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{P_{mm}}
\end{array}\right)$$

- The submatrices $\mathbf{P_{ii}}$ correspond to strongly connected components, $\mathbf{P_{ii}}, i = r+1, \cdots, m$ to irreducible closed subchains.

# The largest eigenvalue of P

- Each of the submatrices $\mathbf{P_{ii}}, i = r+1, \cdots, m$ is row-stochastic and non-negative

# The largest eigenvalue of P

- Each of the submatrices $\mathbf{P_{ii}}, i = r+1, \cdots, m$ is row-stochastic and non-negative
- Therefore, by the theorem of Perron each of these submatrices has an eigenvalue $1$, with a corresponding positive eigenvector.

# The largest eigenvalue of P

- Each of the submatrices $\mathbf{P_{ii}}, i = r+1, \cdots, m$ is row-stochastic and non-negative
- Therefore, by the theorem of Perron each of these submatrices has an eigenvalue $1$, with a corresponding positive eigenvector.
- This means that $\mathbf{P}$ has $m-r$ eigenvalues 1. Since $\mathbf{P}$ is row stochastic, this is also the in modulus largest eigenvalue (complex eigenvalues with modulus 1 may exist).

# A property of the second eigenvector of $\mathbf{A}$

Because $\mathbf{A}$ is column stochastic, we have

$$\mathbf{e}^T \mathbf{A} = \mathbf{e}^T$$

this is, $\mathbf{e}$ is the left-eigenvector corresponding to the eigenvalue $\lambda_1 = 1$.

Since left and right eigenvectors are bi-orthogonal we have that

$$\mathbf{e}^T \mathbf{x}^{(\mathbf{2})} = 0$$

so the coefficients of the second eigenvector(s) add up to zero.

# The second eigenvectors of $\mathbf{A}$

The (second) eigenvectors of $\mathbf{A}$ satisfy

$$(p\mathbf{P}^T + \frac{(1-p)}{N}\mathbf{e}\mathbf{e}^T)\mathbf{x}^{(\mathbf{2})} = \lambda \mathbf{x}^{(\mathbf{2})}$$

and

$$\mathbf{e}^T \mathbf{x}^{(\mathbf{2})} = 0 \quad .$$

This means that

$$p\mathbf{P}^T \mathbf{x}^{(\mathbf{2})} = \lambda \mathbf{x}^{(\mathbf{2})}$$

If $\mathbf{P}^T$ contains at least two irreducible submatrices $\mathbf{P_{ii}}$ we can construct an eigenvector $\mathbf{x}$ for the eigenvalue 1 of $\mathbf{P}^T$ that satisfies $\mathbf{e}^T\mathbf{x}$.

# The second eigenvectors of $\mathbf{A}$

The (second) eigenvectors of $\mathbf{A}$ satisfy

$$(p\mathbf{P}^T + \frac{(1-p)}{N}\mathbf{e}\mathbf{e}^T)\mathbf{x}^{(\mathbf{2})} = \lambda\mathbf{x}^{(\mathbf{2})}$$

and

$$\mathbf{e}^T\mathbf{x}^{(\mathbf{2})} = 0 \quad .$$

This means that

$$p\mathbf{P}^T\mathbf{x}^{(\mathbf{2})} = \lambda\mathbf{x}^{(\mathbf{2})}$$

If $\mathbf{P}^T$ contains at least two irreducible submatrices $\mathbf{P_{ii}}$ we can construct an eigenvector $\mathbf{x}$ for the eigenvalue 1 of $\mathbf{P}^T$ that satisfies $\mathbf{e}^T\mathbf{x}$.
This is an eigenvector of $\mathbf{A}$ for the eigenvalue $\lambda = p$.

# Computation of all the second eigenvectors

To compute a second eigenvector of $\mathbf{A}$ we can solve the homogeneous equation

$$(\mathbf{I} - \mathbf{P}^T)\mathbf{x} = 0 \quad .$$

# Computation of all the second eigenvectors

To compute a second eigenvector of $\mathbf{A}$ we can solve the homogeneous equation

$$(\mathbf{I} - \mathbf{P}^T)\mathbf{x} = 0 \quad .$$

The nonzero-elements in $\mathbf{x}$ correspond to nodes in an irreducible subchain. To compute the different irreducible subchains, we use again Tarjan's algorithm.

# Computation of all the second eigenvectors

To compute a second eigenvector of $\mathbf{A}$ we can solve the homogeneous equation

$$(\mathbf{I} - \mathbf{P}^T)\mathbf{x} = 0 \quad .$$

The nonzero-elements in $\mathbf{x}$ correspond to nodes in an irreducible subchain. To compute the different irreducible subchains, we use again Tarjan's algorithm.

From the indices of the nodes in the same subchain we can form the submatrix $\mathbf{P_{ii}}$. The eigenvectors of two such submatrices can be combined to one second eigenvector of $\mathbf{A}$.

# Computation of all the second eigenvectors

To compute a second eigenvector of $\mathbf{A}$ we can solve the homogeneous equation

$$(\mathbf{I} - \mathbf{P}^T)\mathbf{x} = 0 \quad .$$

The nonzero-elements in $\mathbf{x}$ correspond to nodes in an irreducible subchain. To compute the different irreducible subchains, we use again Tarjan's algorithm.

From the indices of the nodes in the same subchain we can form the submatrix $\mathbf{P_{ii}}$. The eigenvectors of two such submatrices can be combined to one second eigenvector of $\mathbf{A}$.

In total there are $m - r$ second eigenvectors. Since these eigenvectors are pairwise combinations of nonoverlapping vectors they have in total not more than $2N$ nonzero entries.

# Detection of linkspamming

Now there are two obvious algorithms to detect linkspamming:

1. Search the web for irreducible subchains (Tarjan's algorithm)
2. Compute a first eigenvector of $\mathbf{P}^{\mathbf{T}}$, determine the nonzero entries and use Tarjan's algorithm only on the nodes corresponding to nonzero entries.

# Results

| Test problem | Size | Closed subchains | CPU-time Tarjan | CPU-time Eigvec |
|---|---|---|---|---|
| wb-cs-stanford | 9914 | 113 | 0.3 | 1.4 |
| flickr | 820878 | 5394 | 399.3 | 160.8 |
| wikipedia-20051105 | 1634989 | 68 | 1515.3 | 140.2 |
| wikipedia-20060925 | 2983494 | 63 | 5077.1 | 166.6 |
| wikipedia-20061104 | 3148440 | 59 | 5696.9 | 155.1 |
| wikipedia-20070206 | 3566907 | 58 | 7462.7 | 313.6 |
| wb-edu | 9845725 | 49573 | 75703.2 | 2825.6* |

Computing time for web crawls by Gleich

Note: For `wb-edu` the eigenvector algorithm found 41606 subchains.

# Concluding remarks

- We have discussed the PageRank algorithm and the relation between linkspamming and the second eigenvector of the Google matrix:
- The PageRank algorithm is one example of a mathematical method to order importance of nodes in a graph
- Many other applications exist.
- We also discussed some mathematical properties of the transition matrix and developed ideas that can be used for 'Big Data'.