

ASSIGNMENT 1

1. This assignment is due in Feb 10th 11:59pm
2. It requires a certain amount of theory and mathematical knowledge.
3. Let's go over the theory right now.
4. Start early. Like right after I finish talking!

Google PageRank Algorithm (simplified)

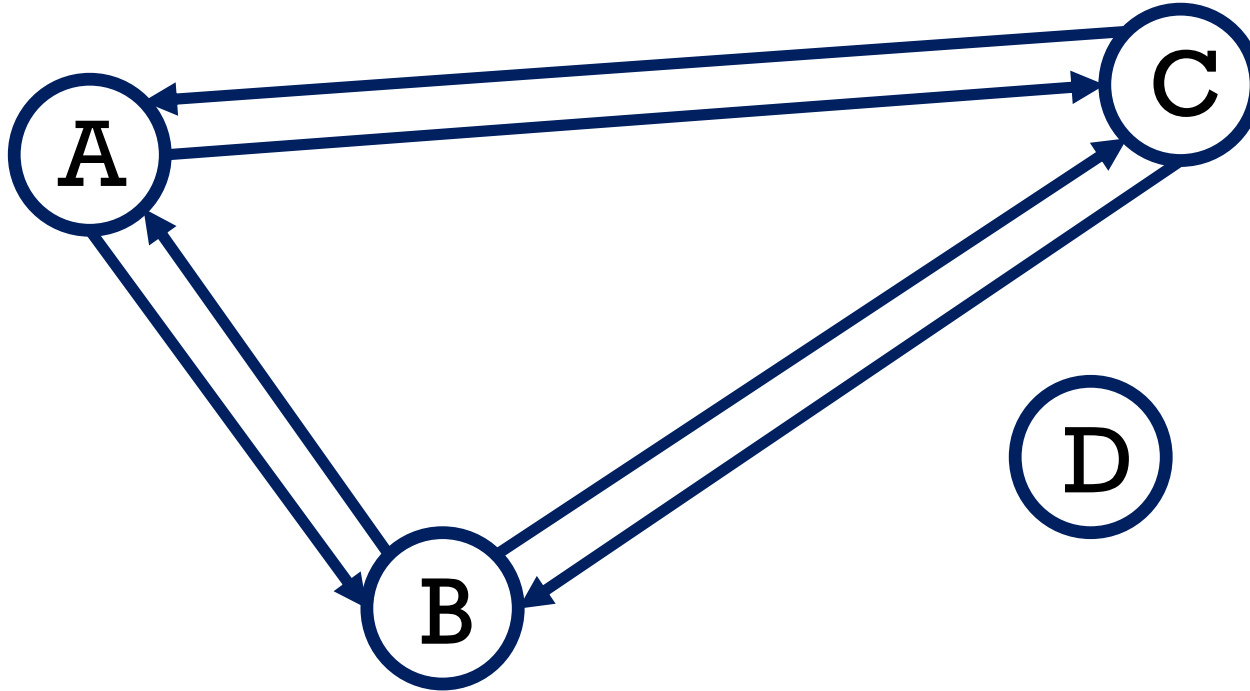
- Great use of the **matrix** and **linear algebra**
- Linear dynamical system with a closed form solution
- **PROBLEM – Lots of web pages on web, how to rank them?**
- Find out how web pages link to each other (connectivity matrix)
- Find out chance to access web pages relative to each other (importance matrix)
- Include non linked pages (stochastic matrix)

Google PageRank Algorithm (simplified)

- Add user randomness into our stochastic matrix
 - User click links with 85% chance
 - User teleports to sites with 15% chance
- This becomes an $n \times n$ **transition matrix**
- Multiply it with a column vector $n \times 1$ repeatedly until the column vector stops changing
- Compare the result with all other sites and get the ranking!

1. Start with a web

Let W be a set of webpages of size n

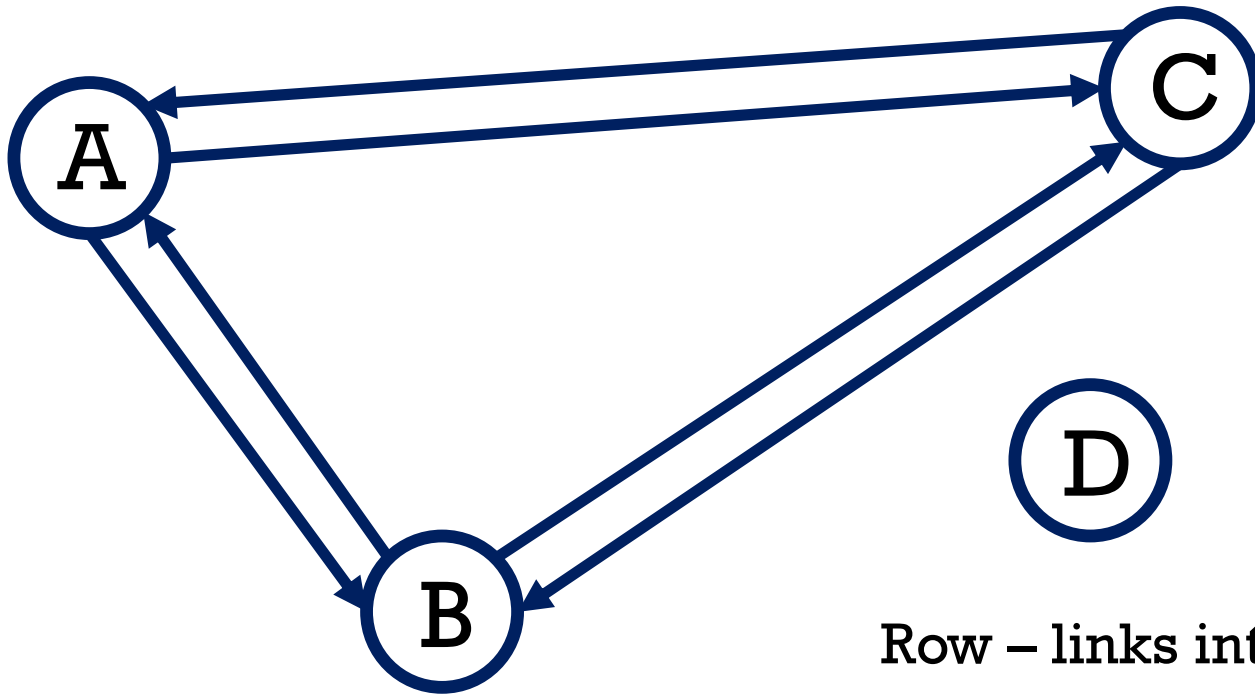


$W =$
{
Apple,
Bell,
Cisco,
Dropbox
}

$\text{Sizeof}(W) = 4$

2. Our connectivity matrix **G**

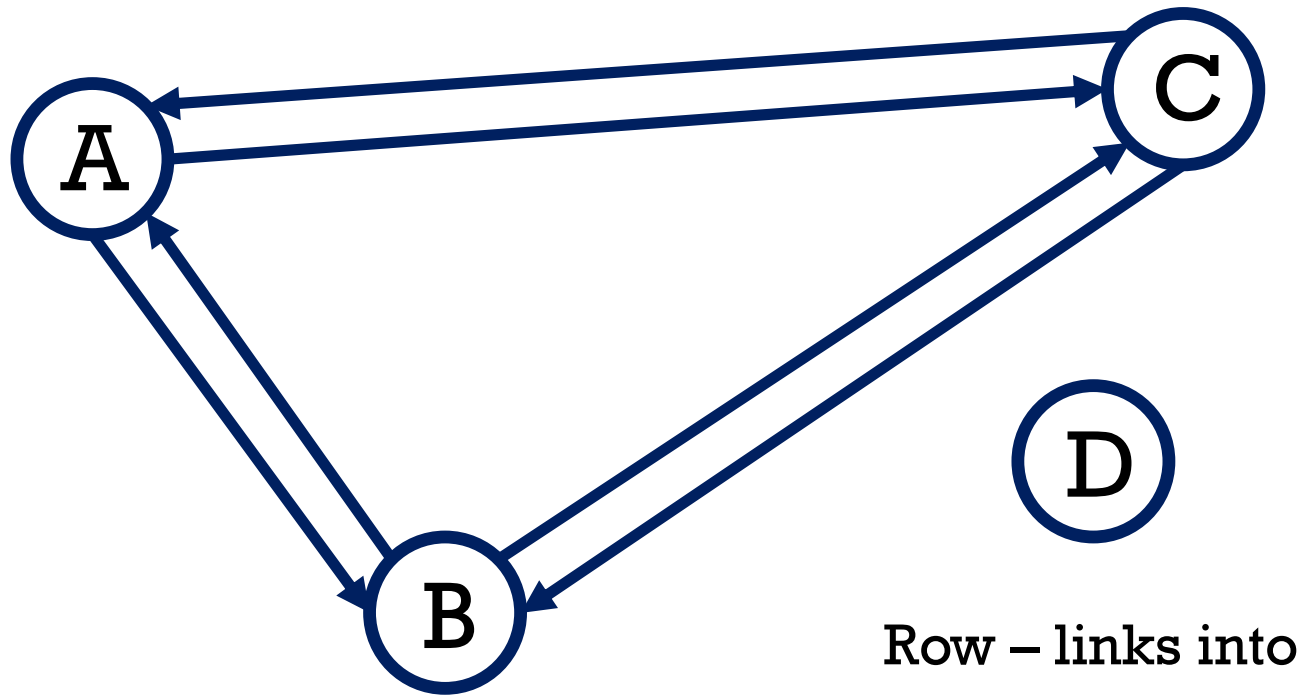
Column – links out from



Row – links into

	A	B	C	D
A				
B				
C				
D				

2. Our connectivity matrix **G**



Column – links out from

	A	B	C	D
A	0	1	1	0
B	1	0	1	0
C	1	1	0	0
D	0	0	0	0

3. Degree

	A	B	C	D
A	0	1	1	0
B	1	0	1	0
C	1	1	0	0
D	0	0	0	0

In-degree \mathbf{r}_i is the number of 1s in row i .

Out-degree \mathbf{c}_j is the number of 1s in column j .

4. Importance matrix S

We can modify our connectivity matrix to show us this "importance" if we divide each value in each column by the sum of each column.

	A	B	C	D
A	0	0.5	0.5	0
B	0.5	0	0.5	0
C	0.5	0.5	0	0
D	0	0	0	0

5. Importance matrix **S** (what about site D)

We can modify our connectivity matrix to show us this ``importance" if we divide each value in each column by the sum of each column.

Total 4 web pages.
D equal random chance to go to all. $1 / 4 = 0.25$

	A	B	C	D
A	0	0.5	0.5	0.25
B	0.5	0	0.5	0.25
C	0.5	0.5	0	0.25
D	0	0	0	0.25

6. Stochastic matrix \mathbf{S} = Probability matrix \mathbf{S}

- Called a “left stochastic matrix” because
 - All columns add to 1
 - All elements are $[0, 1]$

	A	B	C	D
A	0	0.5	0.5	0.25
B	0.5	0	0.5	0.25
C	0.5	0.5	0	0.25
D	0	0	0	0.25
	=1	=1	=1	=1

7. Introduce concept of randomness

We need to introduce the notion of a random walk

We need to multiply our probability matrix by a **random walk probability factor**

For our assignment, we will designate this variable **p**, and set **p = 0.85**.

```
double p{0.85};
```

7. Introduce concept of randomness + teleport

$p = 0.85$ //probability we'll follow the previous matrix

$1 - 0.85 = 0.15$ //probability we won't follow the matrix

0.15 chance we'll **teleport** to another site

- Don't follow link
- Enter address in address bar

8. Create our transition matrix M

Equal chance to go to any page with **teleportation**

Q is an $n \times n$ matrix in which each element is $1/n$

We have 4 web pages so $1 / 4 = 0.25$

$$Q = \begin{matrix} & \begin{matrix} 0.25 & 0.25 & 0.25 & 0.25 \end{matrix} \\ \begin{matrix} 0.25 \\ 0.25 \\ 0.25 \\ 0.25 \end{matrix} & \begin{matrix} 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \end{matrix} \end{matrix}$$

8. Create our transition matrix M

$M = (\text{probability click links}) + (\text{probability teleport})$

$$\mathbf{M} = 0.85 * \mathbf{S} + (1 - 0.85) * \mathbf{Q}$$

$$\mathbf{M} = 0.85 * \begin{array}{|c|c|c|c|} \hline 0 & 0.5 & 0.5 & 0.25 \\ \hline 0.5 & 0 & 0.5 & 0.25 \\ \hline 0.5 & 0.5 & 0 & 0.25 \\ \hline 0 & 0 & 0 & 0.25 \\ \hline \end{array} + 0.15 * \begin{array}{|c|c|c|c|} \hline 0.25 & 0.25 & 0.25 & 0.25 \\ \hline 0.25 & 0.25 & 0.25 & 0.25 \\ \hline 0.25 & 0.25 & 0.25 & 0.25 \\ \hline 0.25 & 0.25 & 0.25 & 0.25 \\ \hline \end{array}$$

8. Create our transition matrix M

M =

0.0375	0.4625	0.4625	0.25
0.4625	0.0375	0.4625	0.25
0.4625	0.4625	0.0375	0.25
0.0375	0.0375	0.0375	0.25

9. Create a column matrix **rank** of size $n \times 1$

Column matrix **rank**

1.0

1.0

1.0

1.0

10. The Markov Process

Multiply the transition matrix **M** by our matrix **rank**, and then multiply **M** by the result and then keep doing this until the rank stops changing (**result converges**), e.g., $M * \text{rank} = \text{rank}$.

$M * \text{rank} = \text{rank}$

The diagram illustrates the equation $M * \text{rank} = \text{rank}$. Three blue arrows originate from the text: one points to the transition matrix **M** (a 4x4 grid), one points to the rank matrix (a 4x1 column), and one points to the resulting rank matrix (a 4x1 column). The multiplication is represented by an asterisk, and the equality by an equals sign.

0.0375	0.4625	0.4625	0.25
0.4625	0.0375	0.4625	0.25
0.4625	0.4625	0.0375	0.25
0.0375	0.0375	0.0375	0.25

*

1.0
1.0
1.0
1.0

=

1.2125
1.2125
1.2125
0.3625

10. The Markov Process

Multiply the transition matrix **M** by our matrix **rank**, and then multiply **M** by the result and then keep doing this until the rank stops changing (**result converges**), e.g., $M * \text{rank} = \text{rank}$.

$M * \text{rank} = \text{rank}$ //repeat until results converge

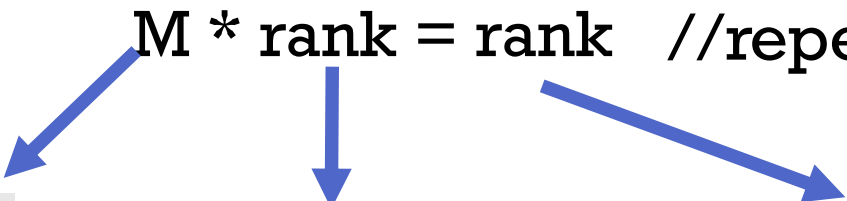
0.0375	0.4625	0.4625	0.25
0.4625	0.0375	0.4625	0.25
0.4625	0.4625	0.0375	0.25
0.0375	0.0375	0.0375	0.25

*

1.2125
1.2125
1.2125
0.3625

=

???
???
???
???



10. The Markov Process

Multiply the transition matrix **M** by our matrix **rank**, and then multiply **M** by the result and then keep doing this until the rank stops changing (**result converges**), e.g., $M * \text{rank} = \text{rank}$. In this case, we get:

1.2698

1.2698

1.2698

0.1905

1 1. And finally

Divide each element in rank by the sum of the values in rank
(scale rank so its elements sum to 1):

rank =	1.2698	/	3.999	=	0.3175	A
	1.2698	/	3.999		0.3175	B
	1.2698	/	3.999		0.3175	C
	0.1905	/	3.999		0.0476	D

And that's that!

- The result makes intuitive sense
- Each of pages A, B, and C has a rank of about 32%, and page D ranks fourth with about 5%
- Keeping in mind that we haven't considered how a user's query will affect the rank, **you now understand how Google's PageRank* works.**

*a simplified version