Minor in AI

How Search Engines Work

1 The Search Paradox: Finding Needles in Digital Haystacks

Imagine trying to find a specific striped-shirt character in a chaotic "Where's Waldo?" scene. Now picture doing this across billions of web pages. This is the fundamental challenge search engines solve daily. Through two engaging activities, we'll discover how simple concepts evolved into Google's revolutionary PageRank algorithm.

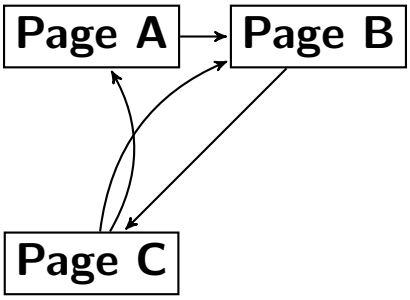
From Homepage to Home

Wikipedia Navigation Exercise:
Students started at Wikipedia's homepage and raced to reach their hometown page through hyperlinks. The winner used just 3 hops:
Main Page → India → New Delhi → Local Landmark
This demonstrated how web pages form interconnected networks.

2 Decoding the Web's Hidden Structure

2.1 The Building Blocks

- **Nodes**: Web pages (A, B, C in our model)
- **Edges**: Hyperlinks between pages
- **In-Degree**: Number of incoming links (popularity metric)
- **Out-Degree**: Number of outgoing links



2.2 The Ranking Challenge

Our 3-page system revealed critical insights:

Spreadsheet Analysis

Page	Links To	In Links	Out-Degree	PageRank
A	B	C	1	0.33
B	C	A	1	0.33
C	A,B	B	2	0.33

Initial equal rankings highlight the need for better metrics - Page C argues: "I link to everyone, shouldn't I be more important?"

2.3 Key Innovations

1. **Random Surfer Model**: Simulates users randomly clicking links
2. **Damping Factor**: Solves infinite loops (e.g., A→B→C→A...)
3. **Popularity Inheritance**: Pages pass ranking power through links

3 From Classroom to Google - The Big Picture

Modern Search Reality

While our model uses 3 pages and 1 parameter, real search engines:

- Process 500-600 ranking factors
- Handle 5.6 billion daily searches
- Update indexes continuously

3.1 Key Takeaways

- **Hyperlinks = Votes**: Each link boosts a page's authority
- **Loops Matter**: Real algorithms need teleportation mechanisms
- **AI Evolution**: Modern systems combine PageRank with machine learning

Notes

The Algorithm That Created a Trillion-Dollar Industry

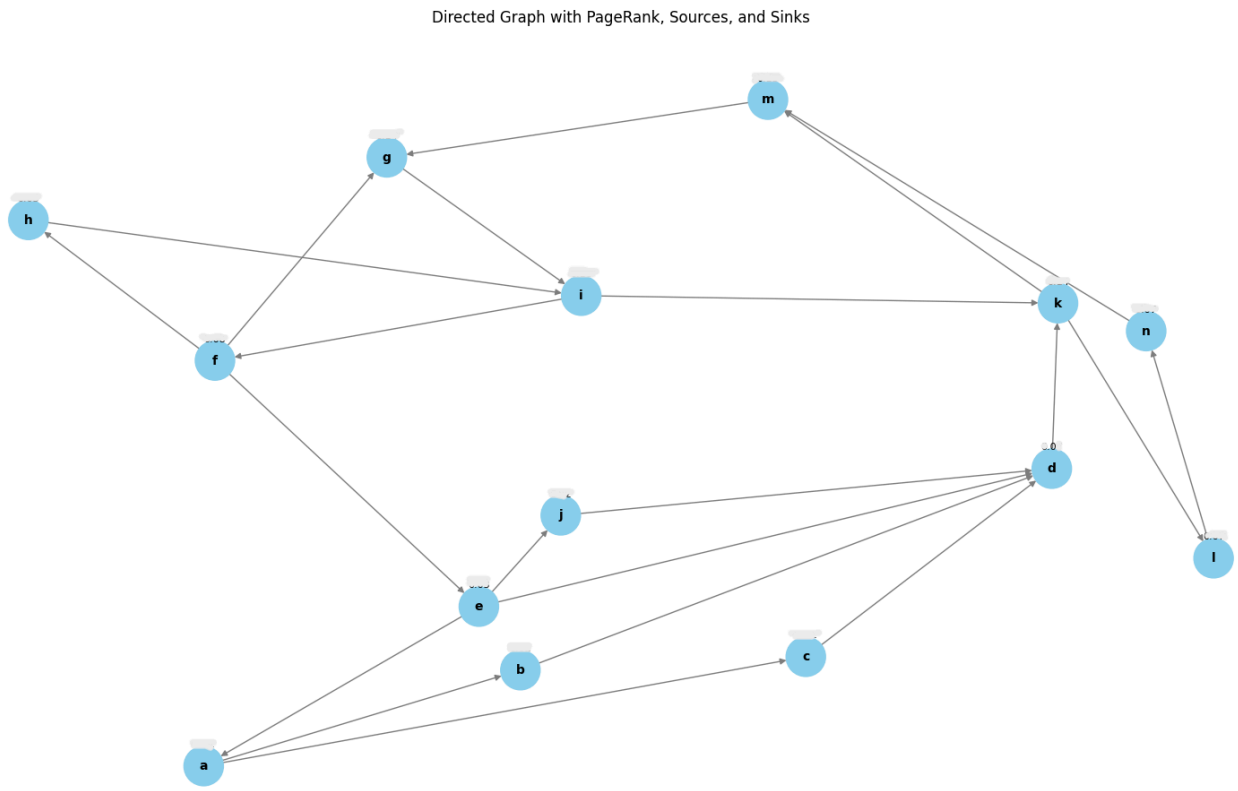
Quote of the class:

> *Data is the new Oil.*

Let us deep dive into a case study.

> I'm a business man and I wanted to set up my coffee shop in one of the point(or node) shown in the following network(or graph). I wanted a high traffic point because only then I can get maximum costumers into my coffee shop. What do you think what it will be?

Network



What do you think, what point should I pick? Will it be a, b, c, d, e, f, g, h, i, j, k, l, m or n?

Just look at the network and tell me what is the most important point for my business?

I think the maximum traffic that I can guess is for point **d** as it has maximum incomming edges. Do you think I am correct?

Let's figure it out together!

Go to Google Colab and follow the following steps:

Step 1: Create the network

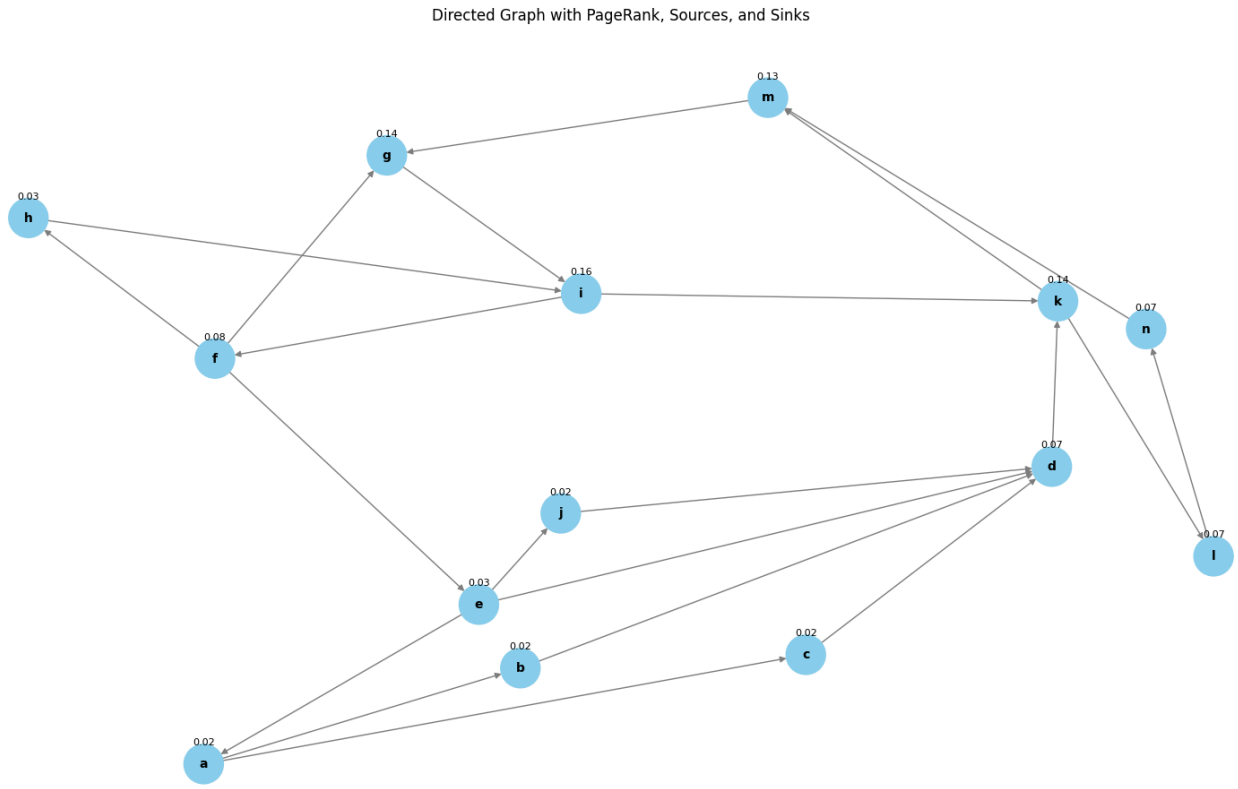
```
L = [
     ['a', 'c'], ['a', 'b'], ['c', 'd'], ['b', 'd'], ['e', 'd'], ['j', 'd'],
     ['f', 'e'], ['f', 'g'], ['f', 'h'], ['h', 'i'], ['g', 'i'], ['d', 'k'],
     ['i', 'k'], ['k', 'l'], ['k', 'm'], ['n', 'm'], ['l', 'n'], ['e', 'j'],
     ['i', 'f'], ['m', 'g'], ['e', 'a']
 ]
```

Step 2: Write the magical prompt given below

> L, as mentioned above, is the directed graph. Visualize this graph and display the page rank of nodes.
> Add code to identify source and sink for all the the nodes in the List
> And sort PageRank of nodes (sorted in descending order)

...and boom see the magic now!

Network



I'm shocked for a moment as the maximum traffic is for point **i** having 16% hits. But how?

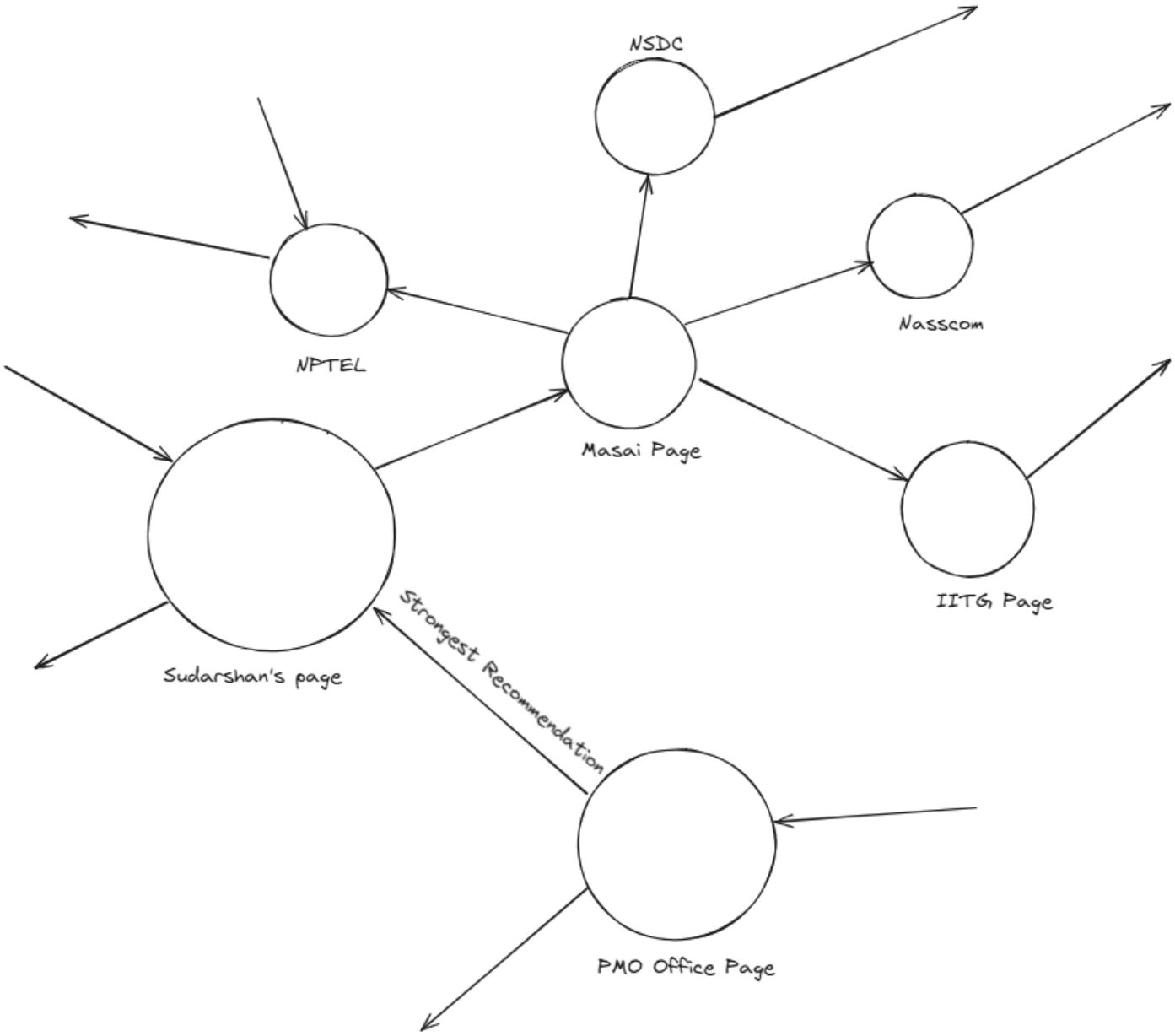
> ## The flow of recommendation is important.

If we are beginning from any point in our network and travelling ramdomely, then the point **i** will be the most likely the one through which we will cross.

Google Colab link for the reference.

Let us understand it with a great analogy.

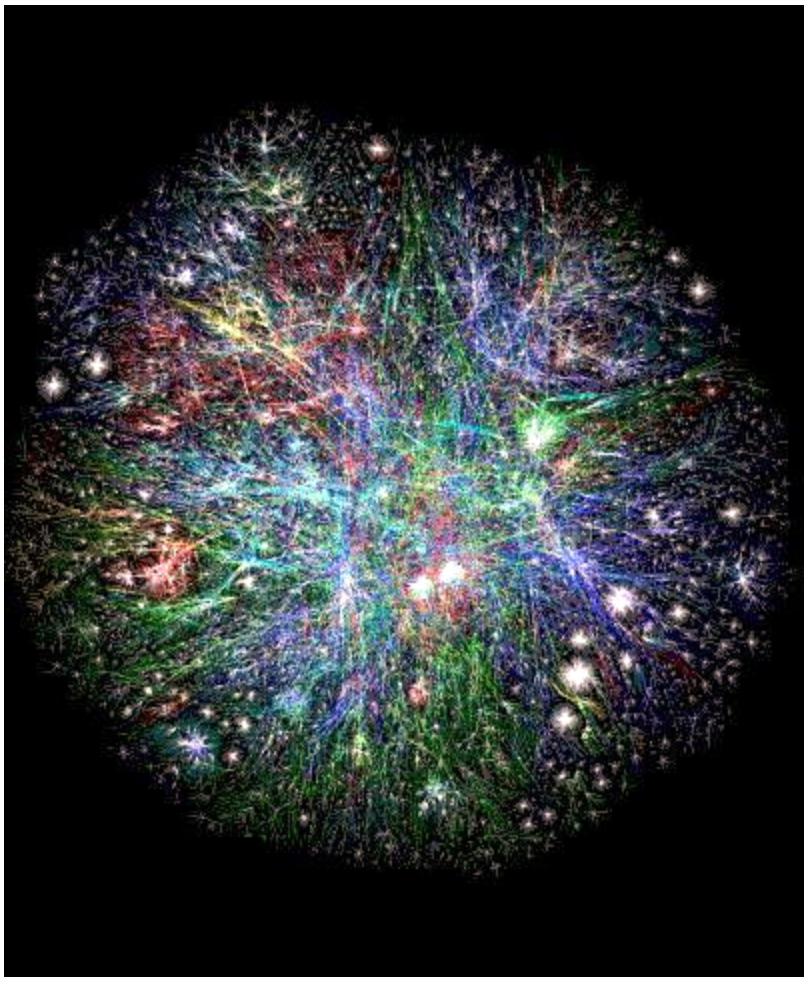
Following network shows multiple pages, it has Sudarshan's page, Masai page, PMO page, NPTEL, etc. Among all the pages, the maximum hits will happen on Sudarshan's page as this page has the strongest recommendation. This page has direct recommendation from Prime Minister. PMO page already had millions in traffic. Therefore, it's recommendation is super strong.



...coming to the conclusion

We hope that via this class you must be able to get an idea around how the webpages are ranked. How Search engines like Google, Bing, Yahoo actually rank them.

Following image is a network of our World Wide Web.



Now guess, what does it look like?

> *Yes, the Neural Network of our brain.*

Moreover, we would like to recommended an amazing book on Randomness.

> *The The Drunkard's Walk: How Randomness Rules Our Lives*

We hope that you enjoyed today's session. Keep learning, keep growing!

Thank you.