The structure of the Google search Engine

By Mohamed Ali

Pre Google

- Human Maintained Indices 
  - Subjective
  - Expensive
  - Cannot cover esoteric topics
- Automated Search Engines
  - Relies chiefly on keyword matching
  - Can be misled

How bad Is keyword matching

In 1997, Only one of the top four commercial search engines finds itself.

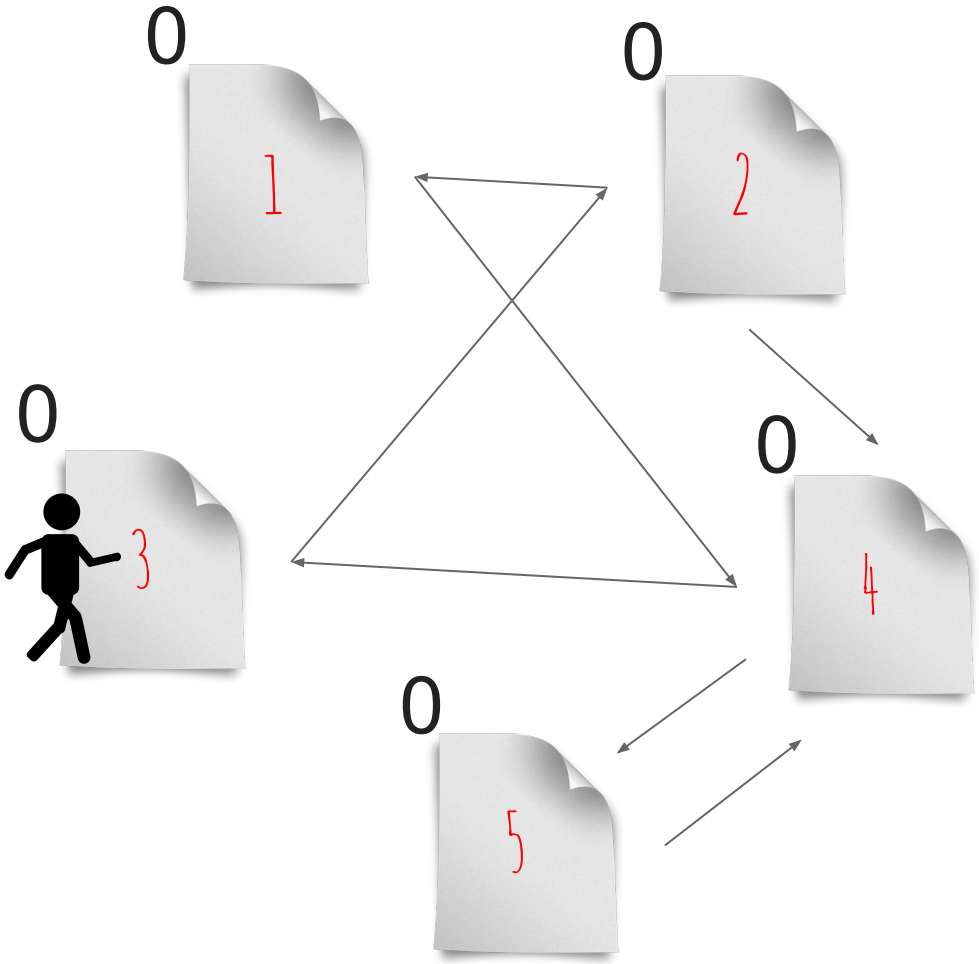
Introducing Google

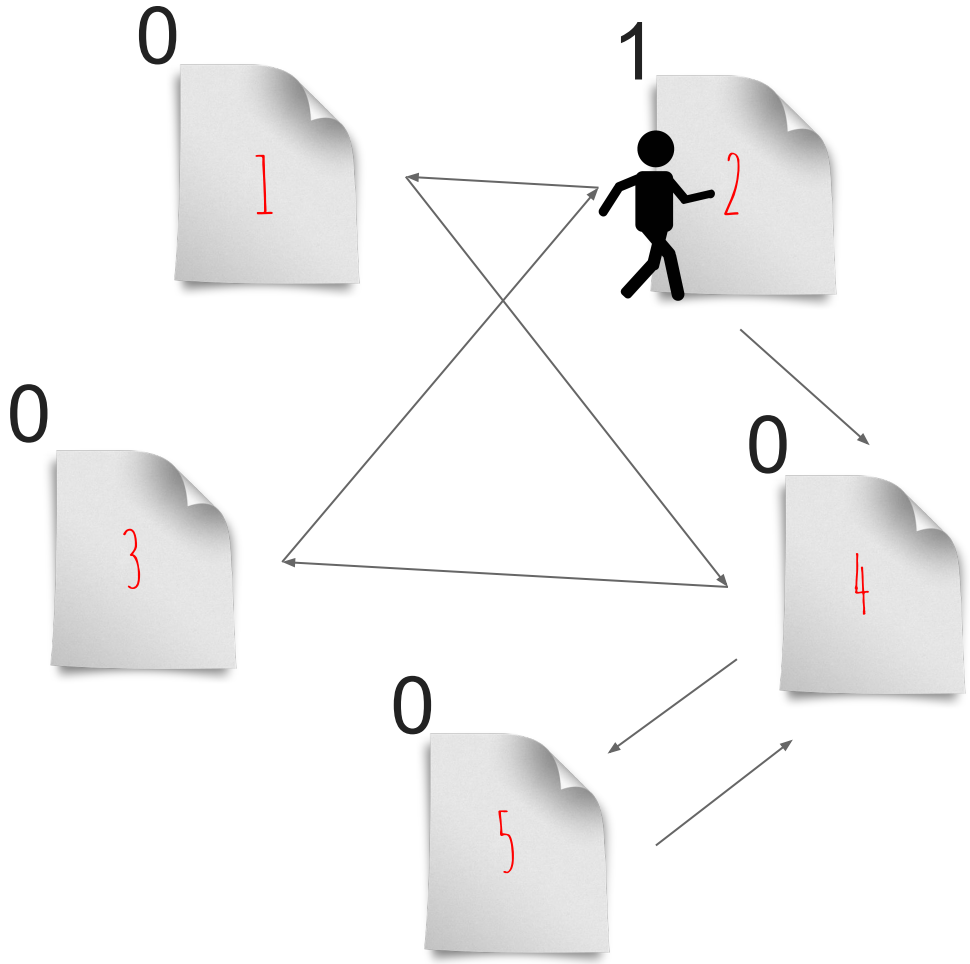
System Features

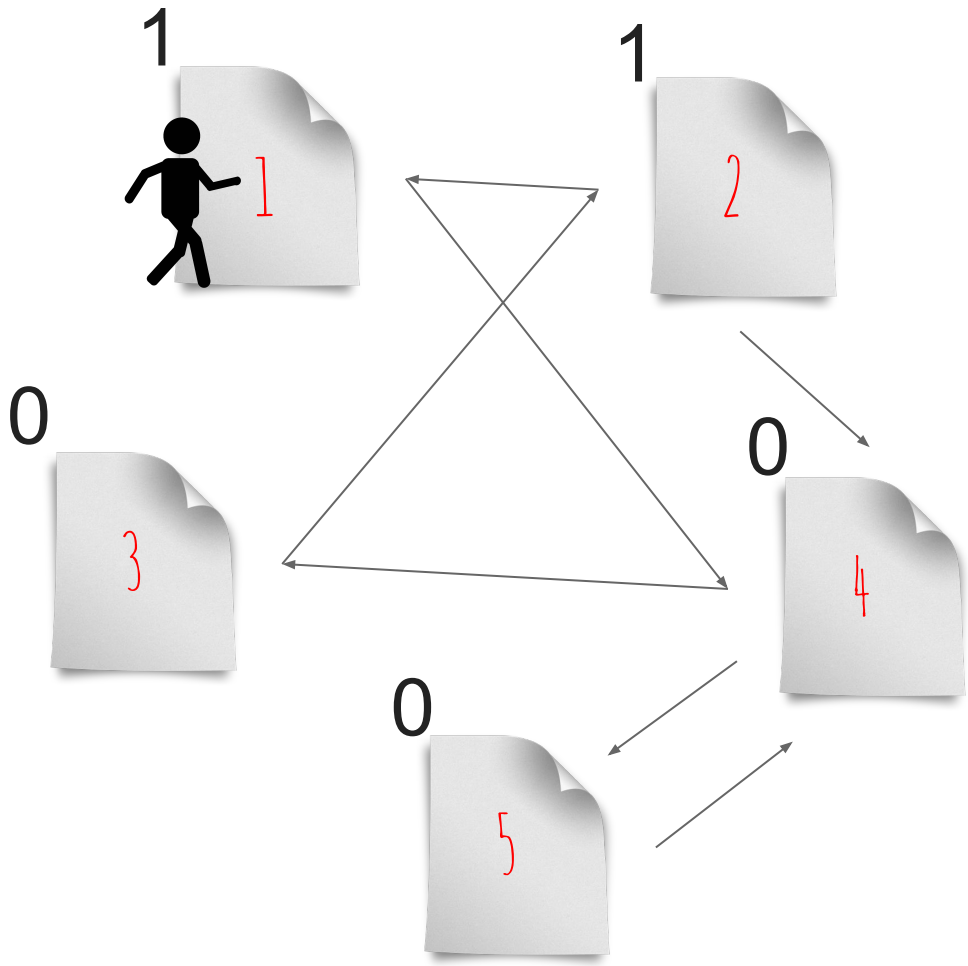
- Making use of the link structure
  - Ordering the web
- Using anchor text
  - Sometimes more descriptive than the website’s title
  - Allows for reaching non-text pages

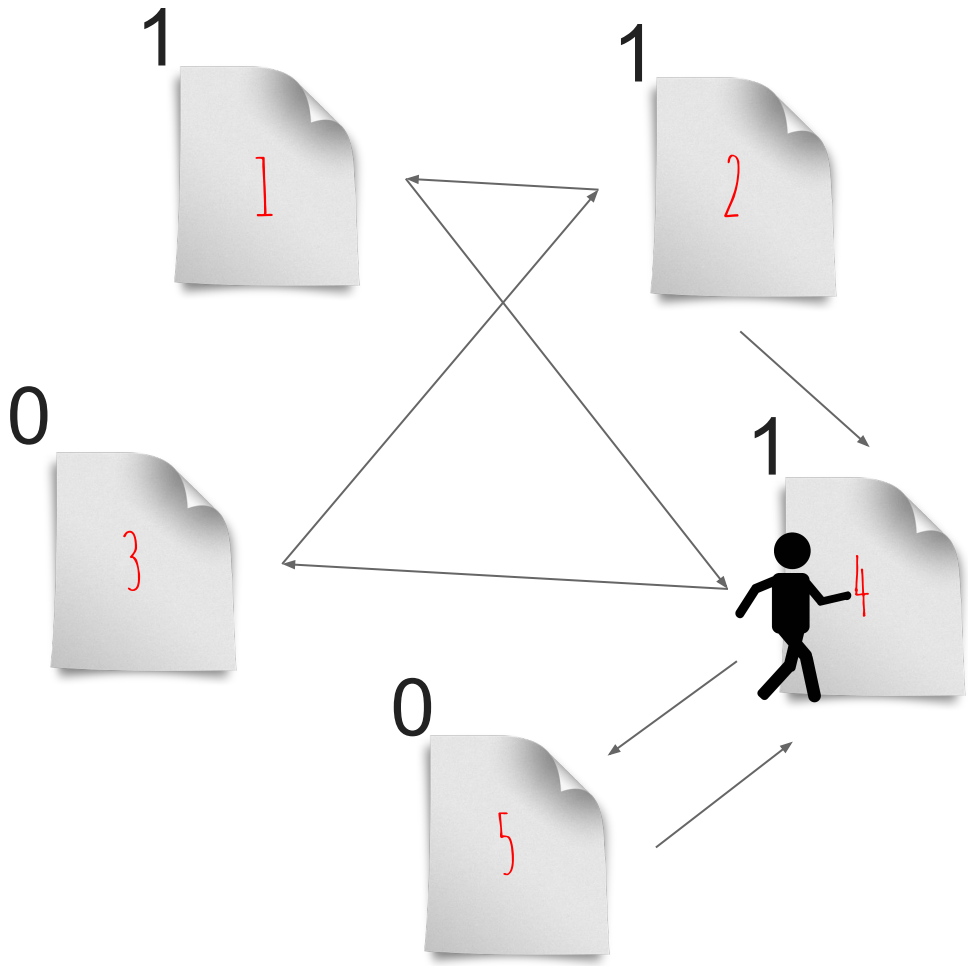
Crawling and indexing

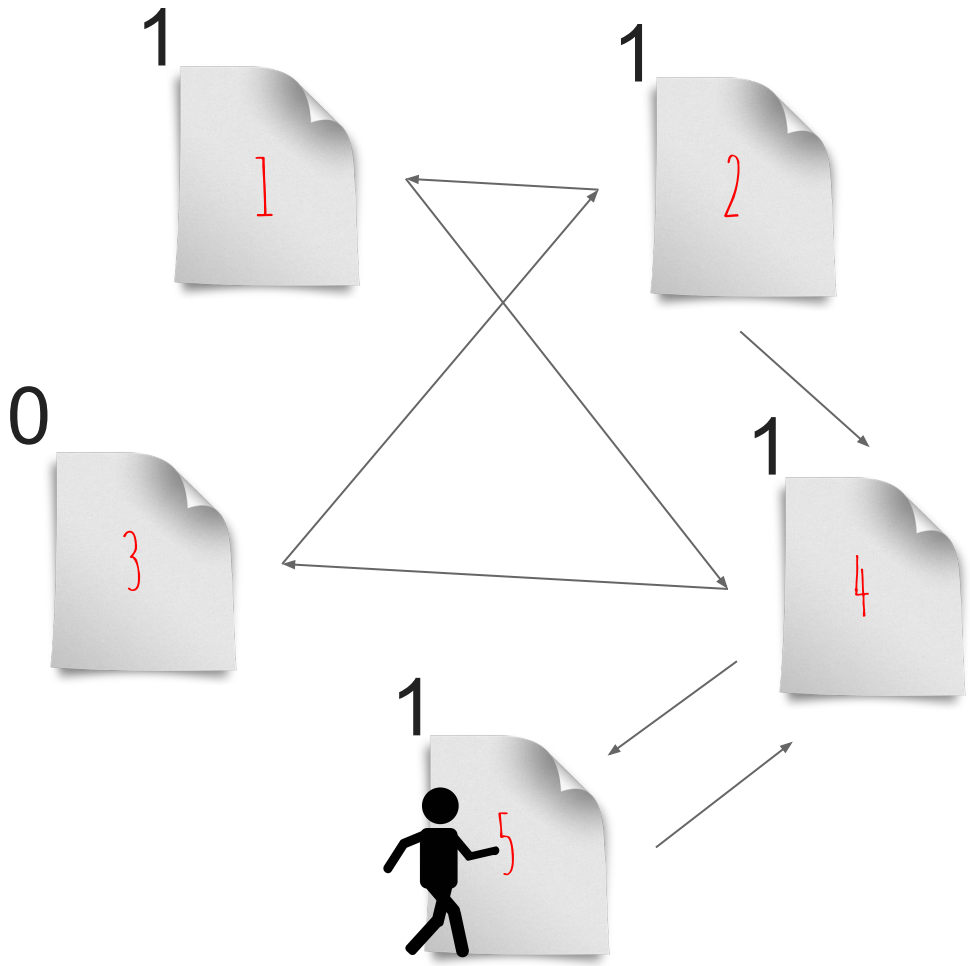
- The Random Surfer Model (PageRank)

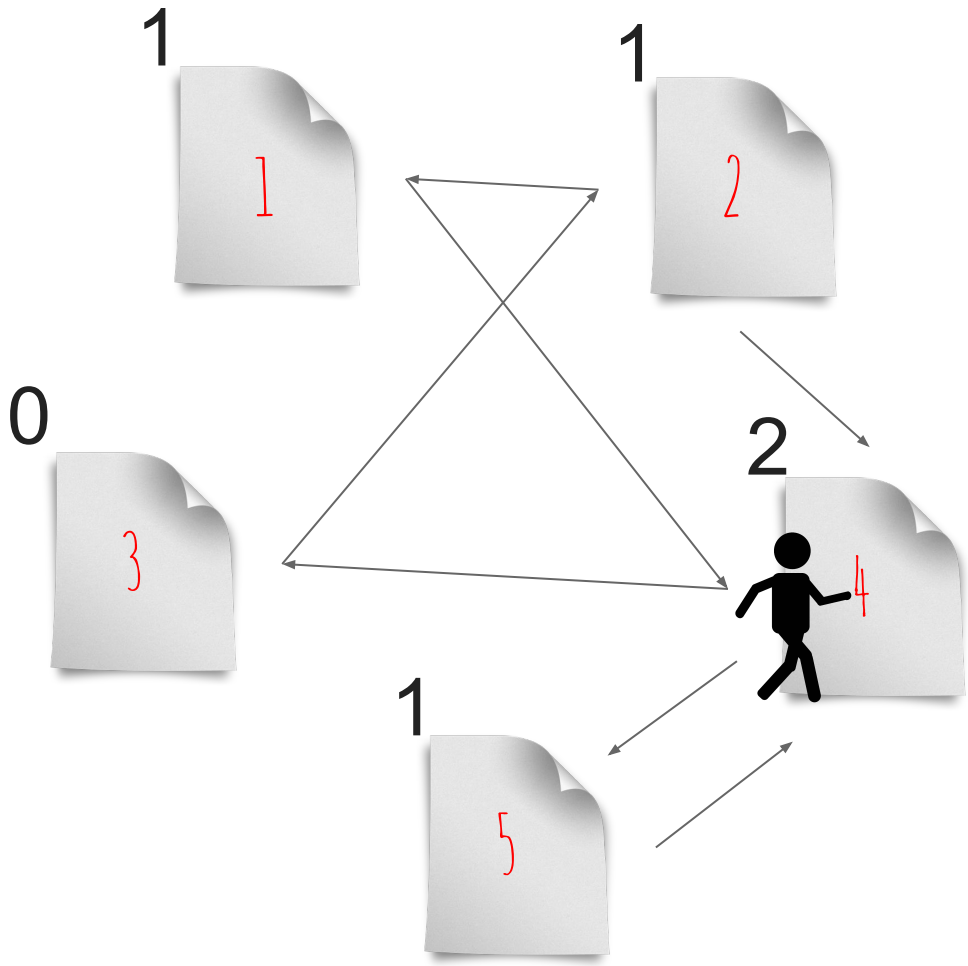

0
1
0
2
0
3
0
4
0
5

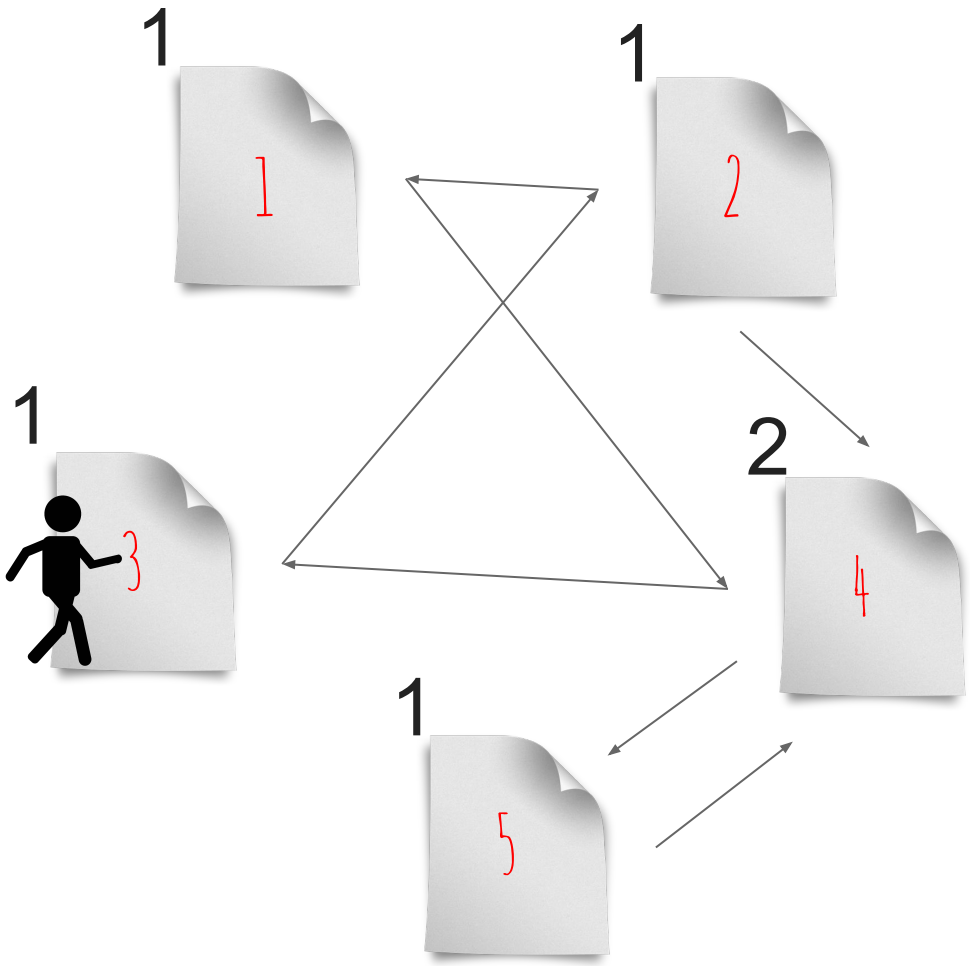

0
1
1
2
0
3
0
4
0
5

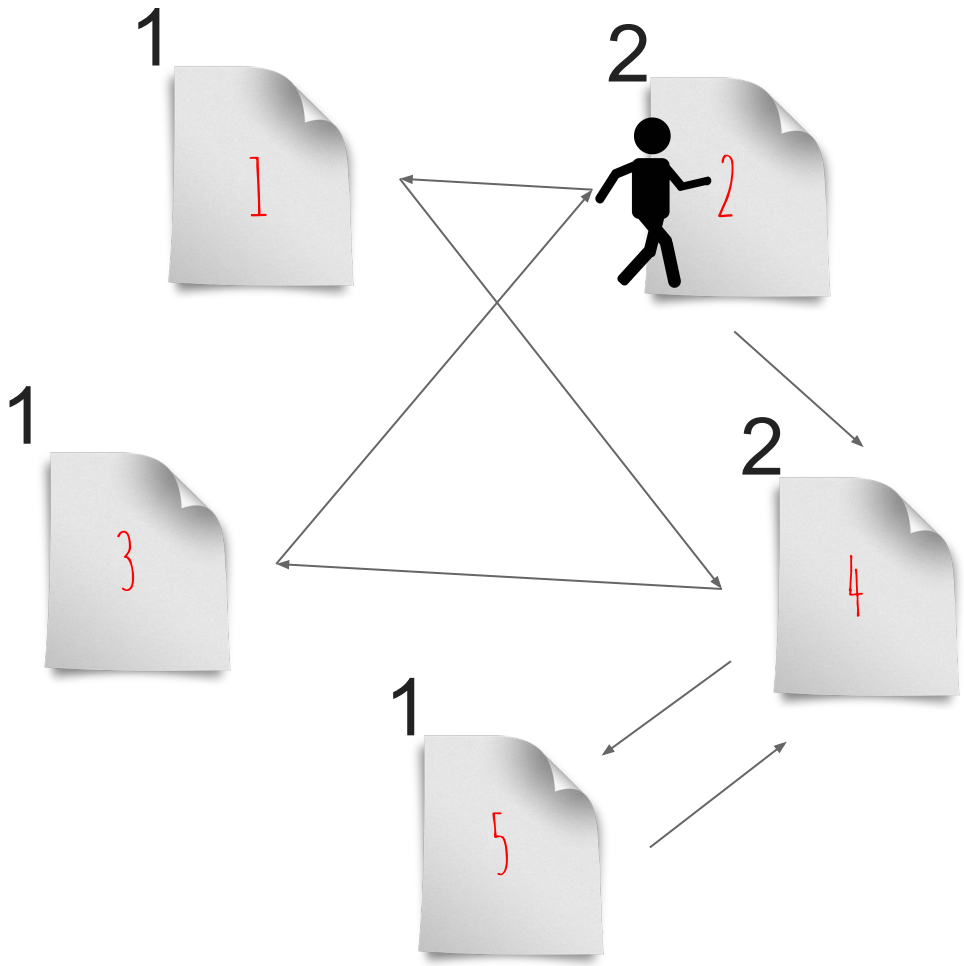

1
1
2
0
3
0
4
0
5

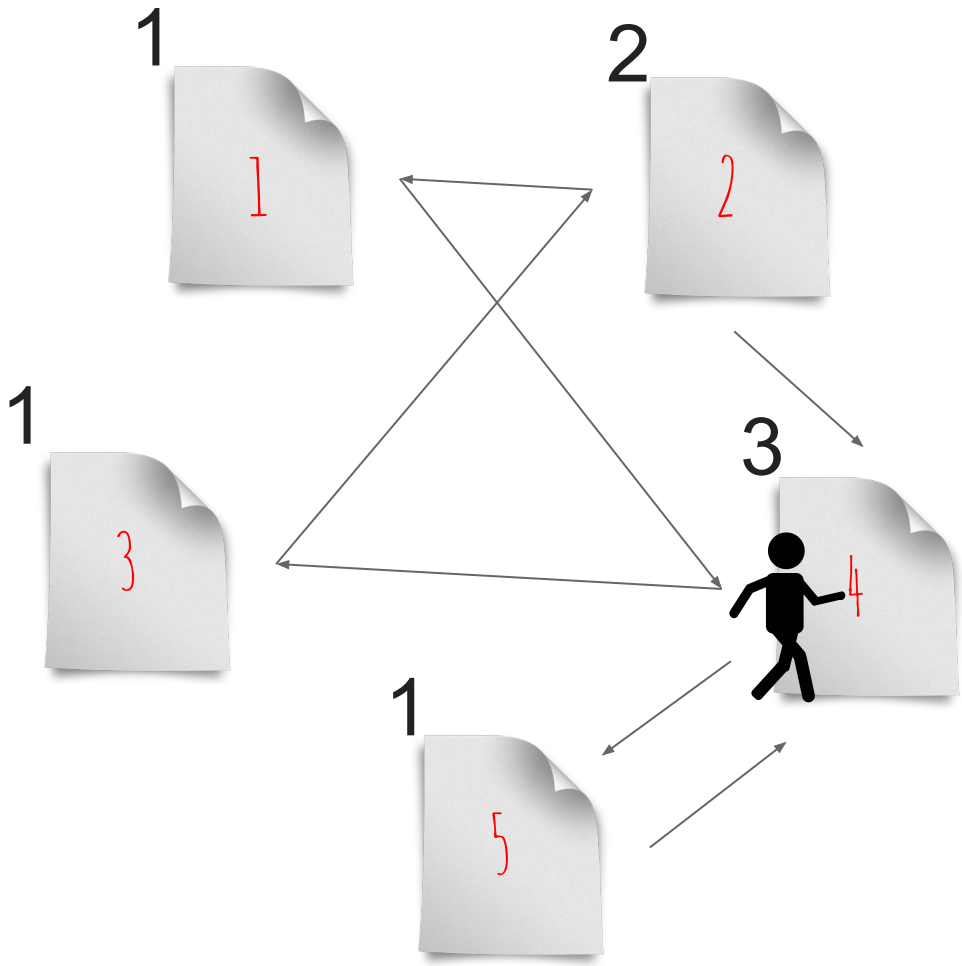

1
1
1
2
0
3
1
4
0
5

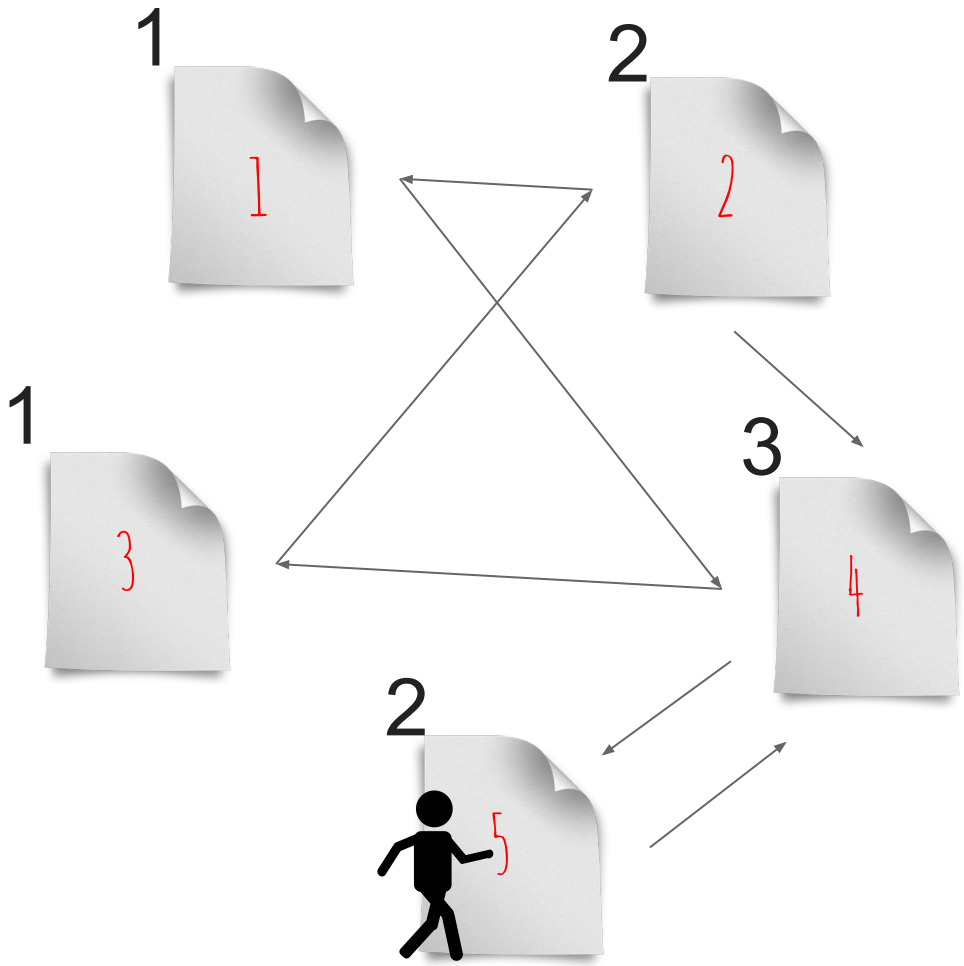

1
1
1
2
0
3
1
4
1
5

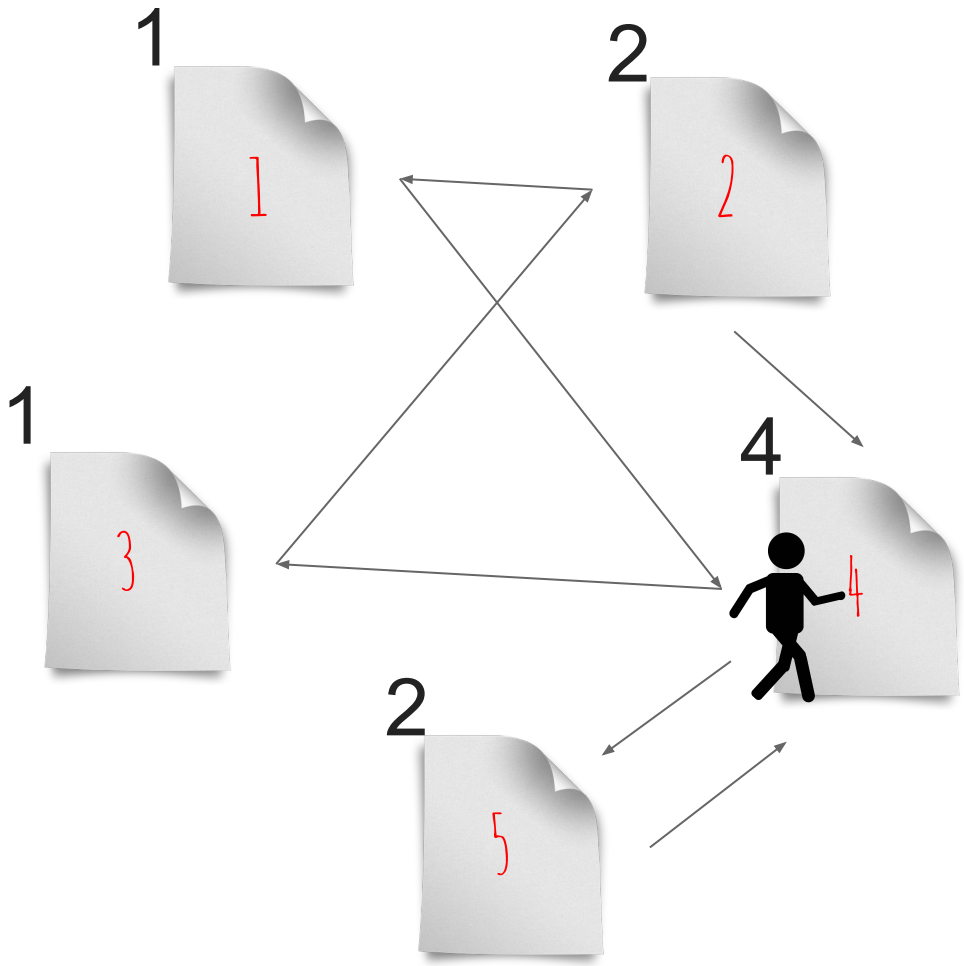

1
1
1
2
0
3
2
4
1
5

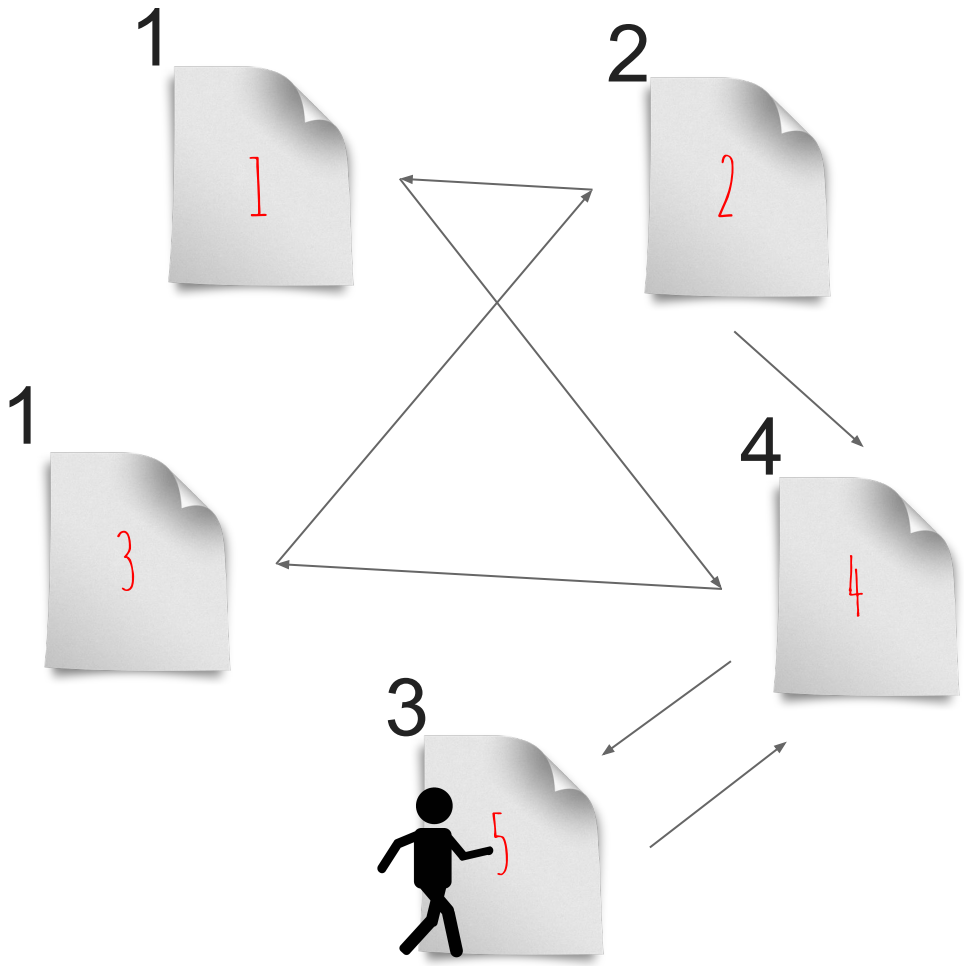

1
1
1
2
1
3
2
4
1
5

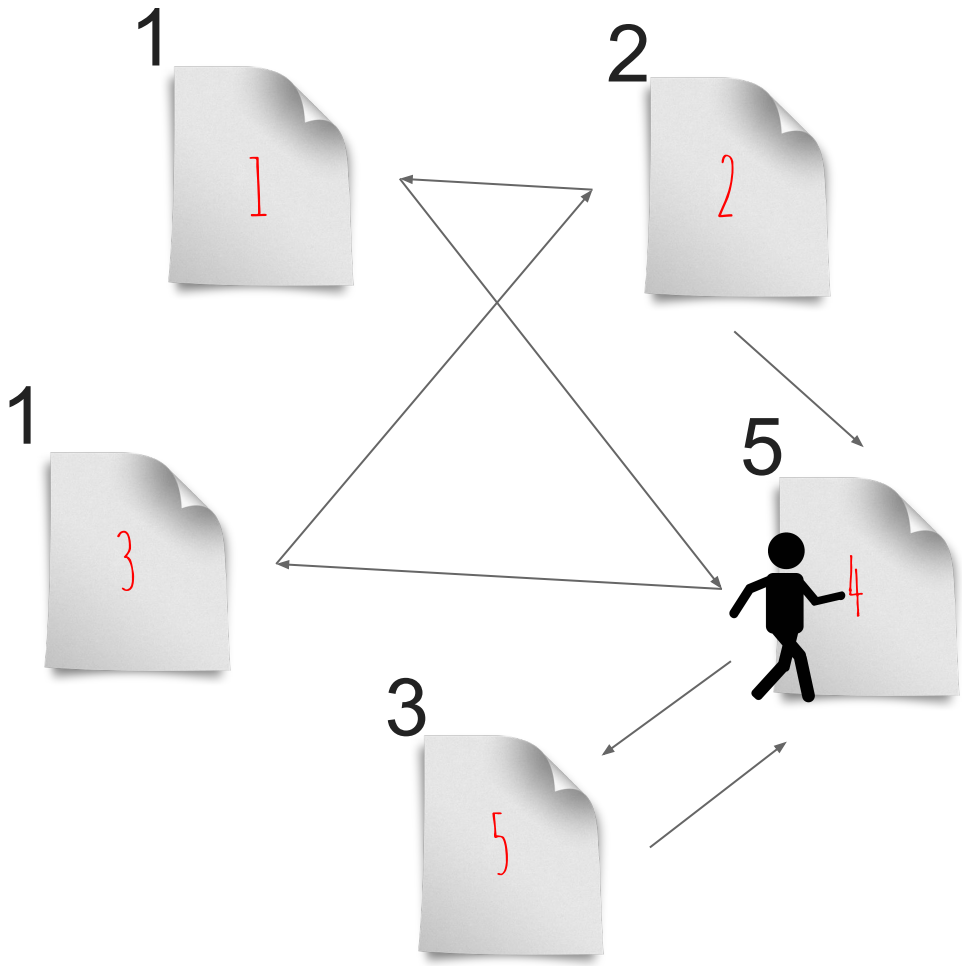

1
1
2
2
1
3
2
4
1
5

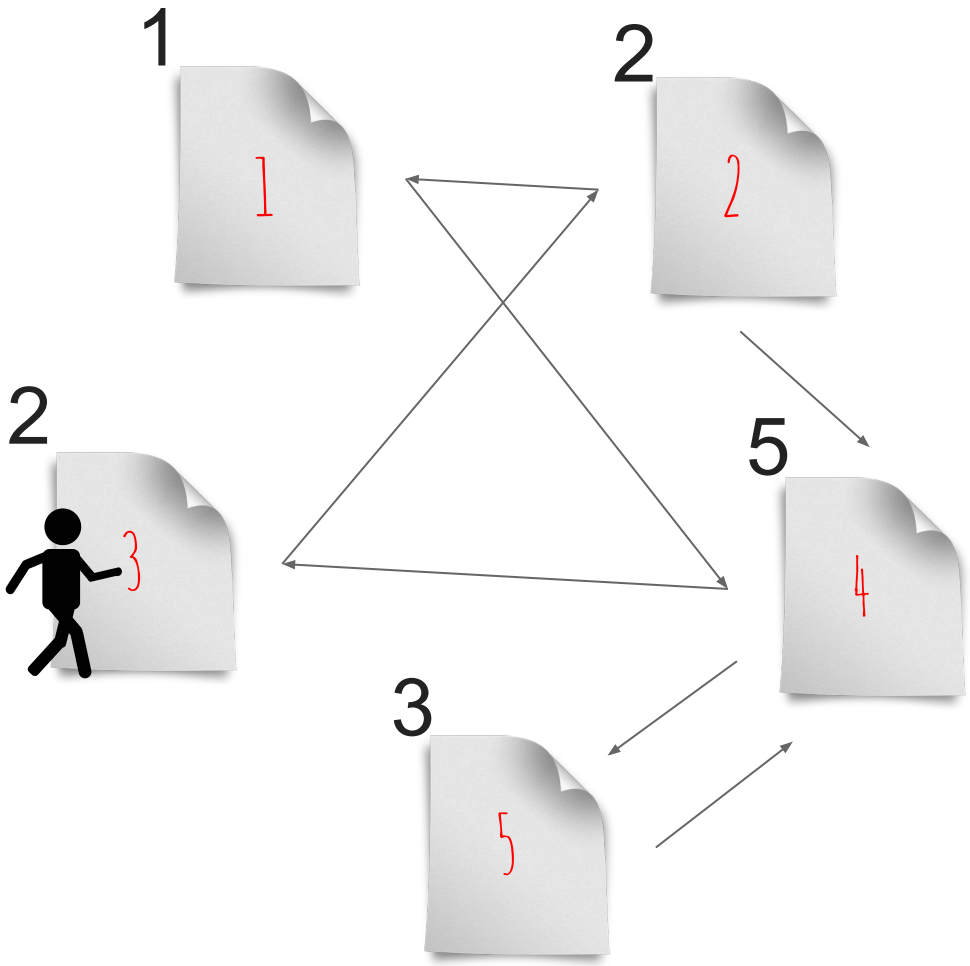

1
1
2
2
1
3
3
4
1
5

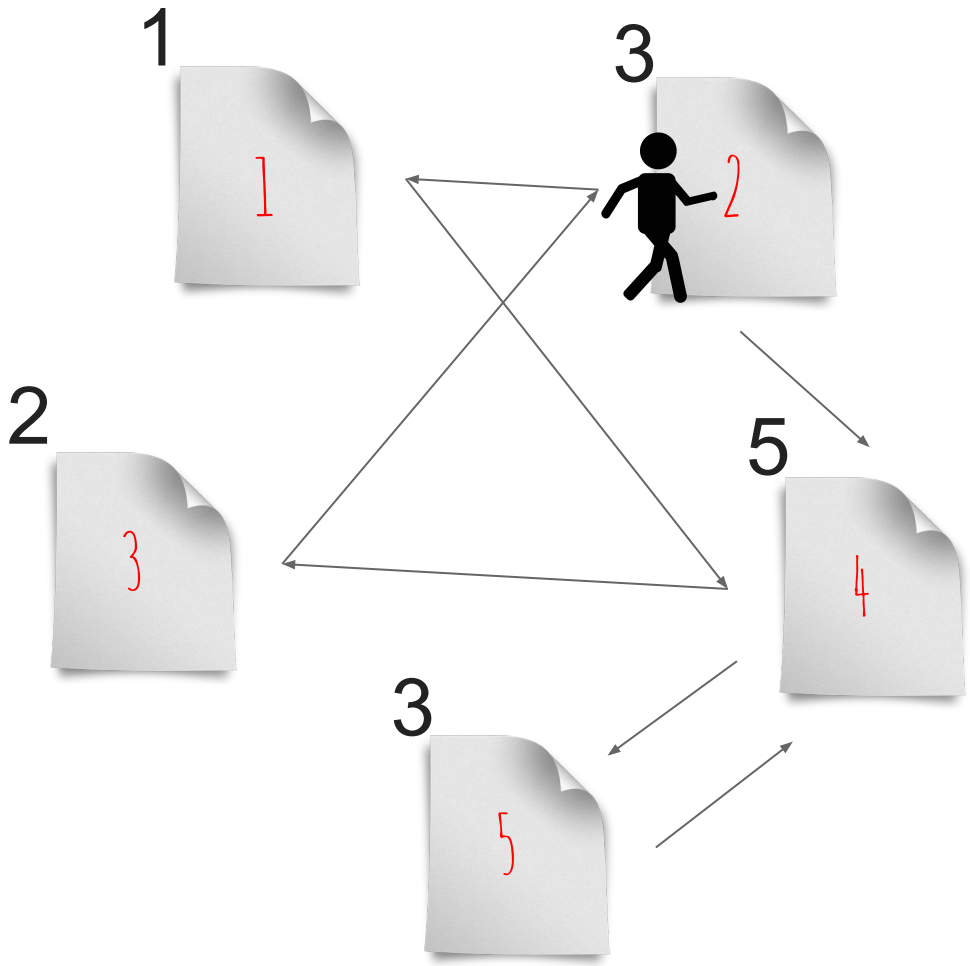

1
1
2
2
1
3
3
4
2
5

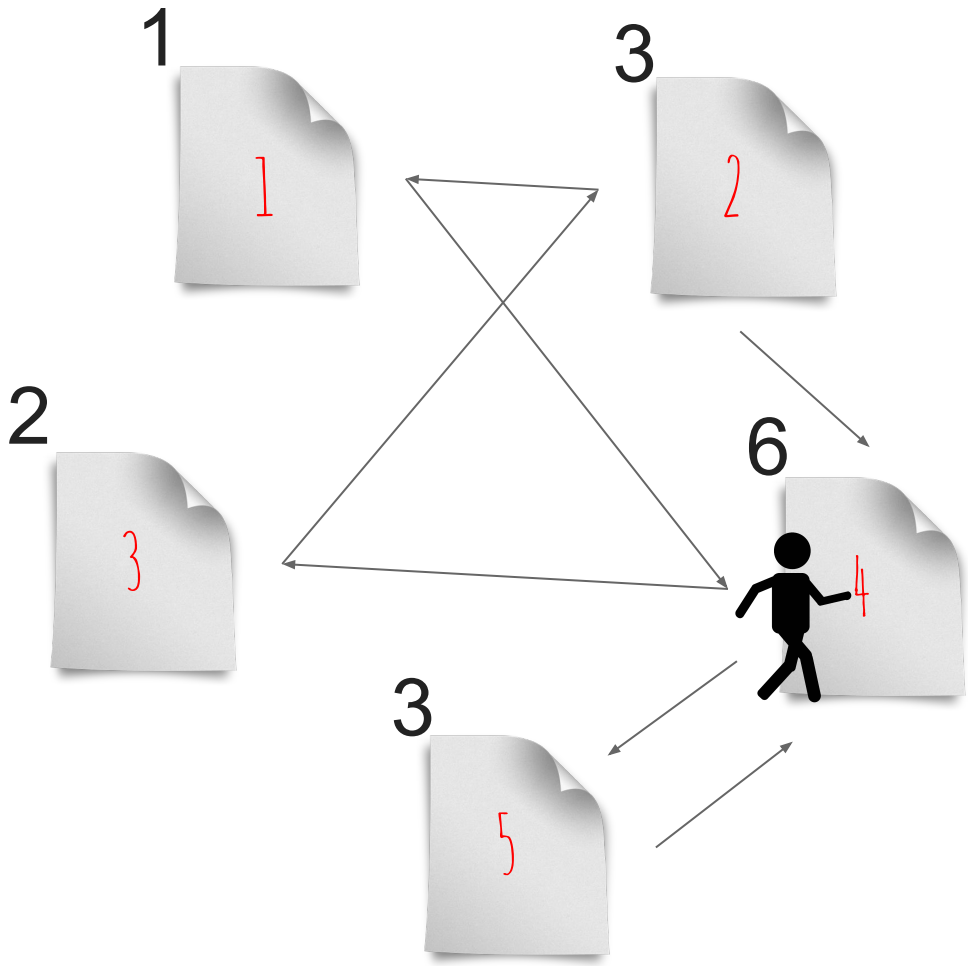

1
1
2
2
1
3
4
4
2
5

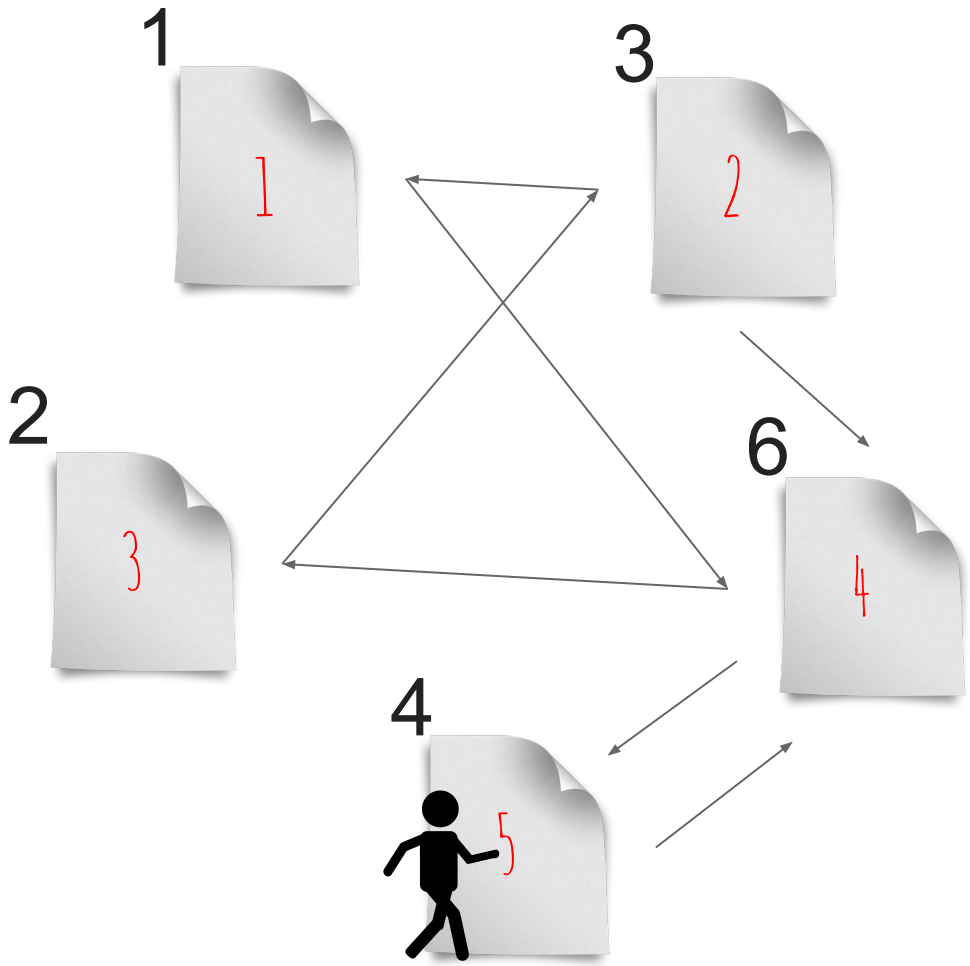

1
1
2
2
1
3
4
4
3
5

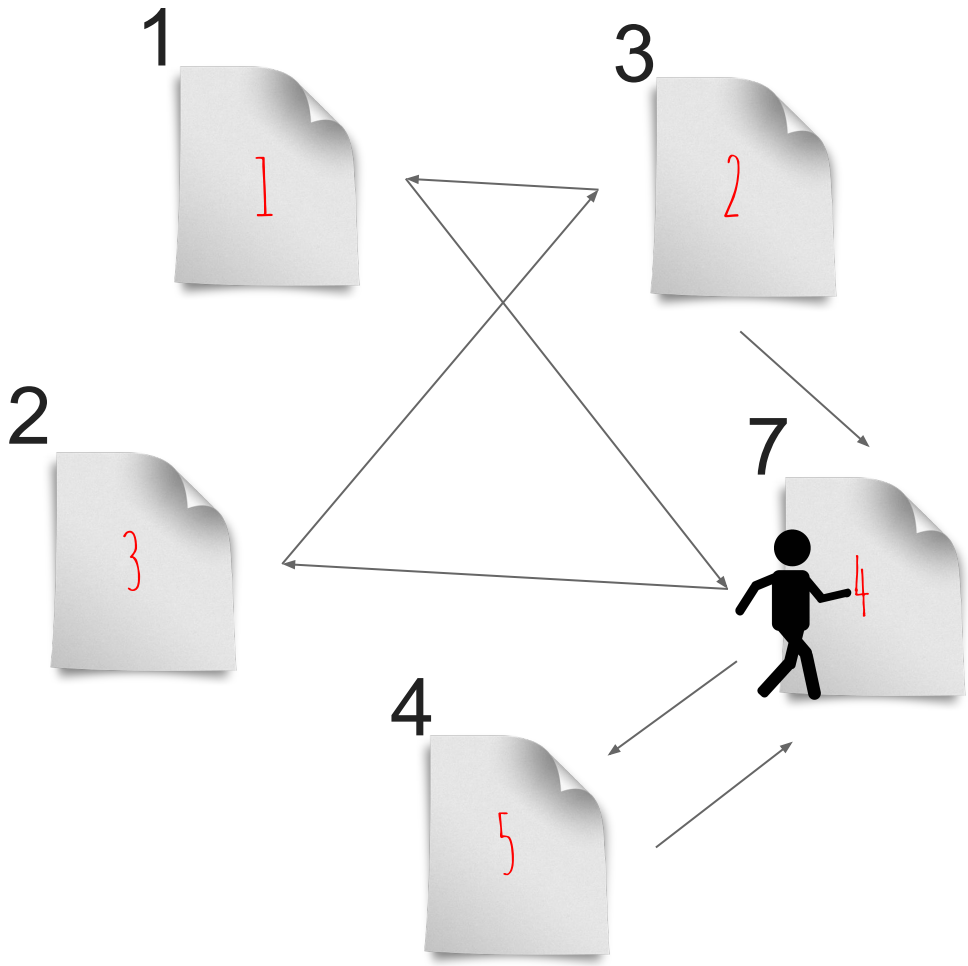

1
1
2
2
1
3
5
4
3
5

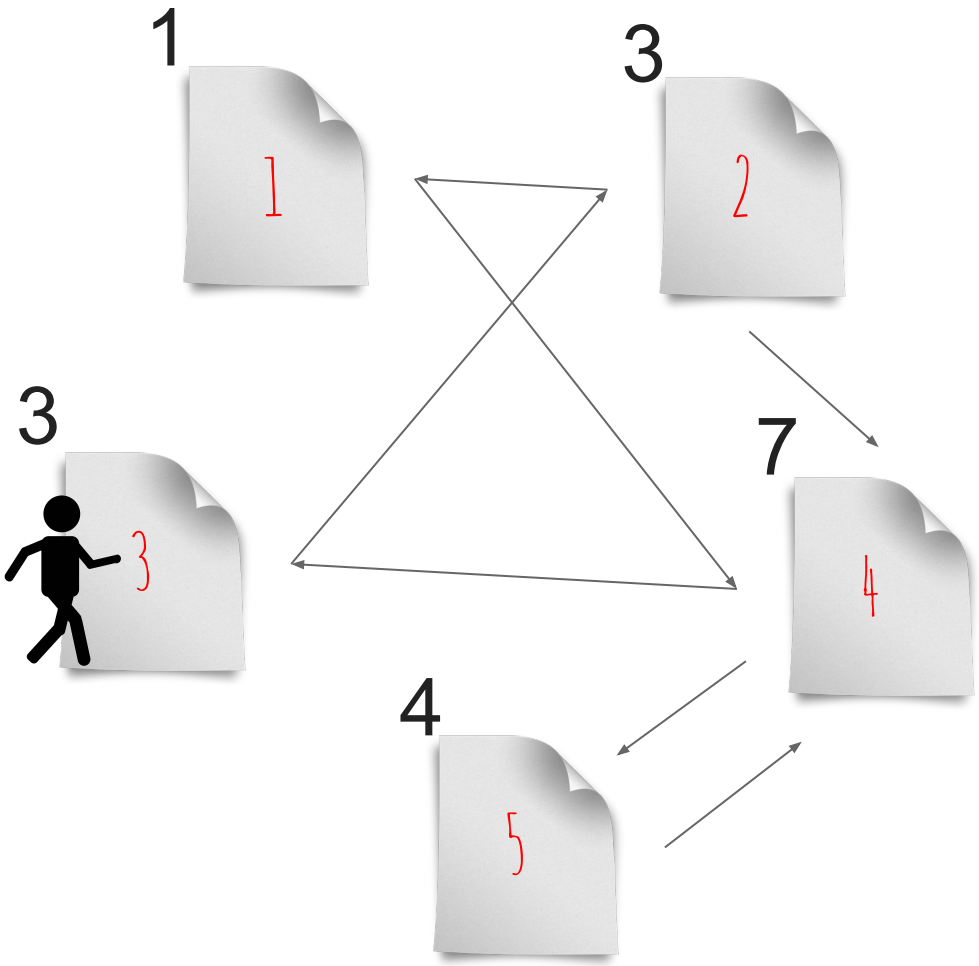

1
1
2
2
2
3
5
4
3
5


1
1
3
2
2
3
5
4
3
5

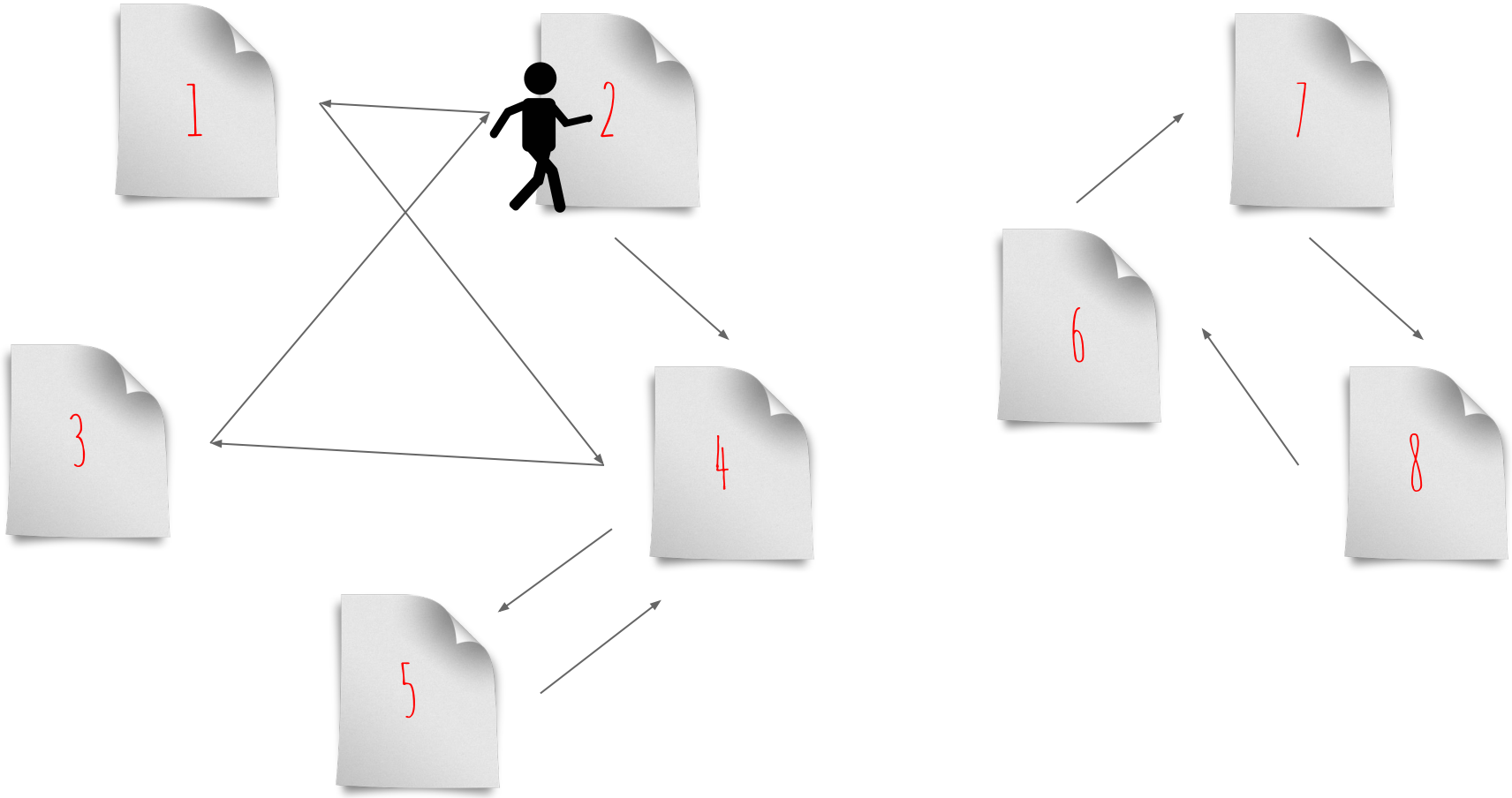

1
1
3
2
2
3
6
4
3
5

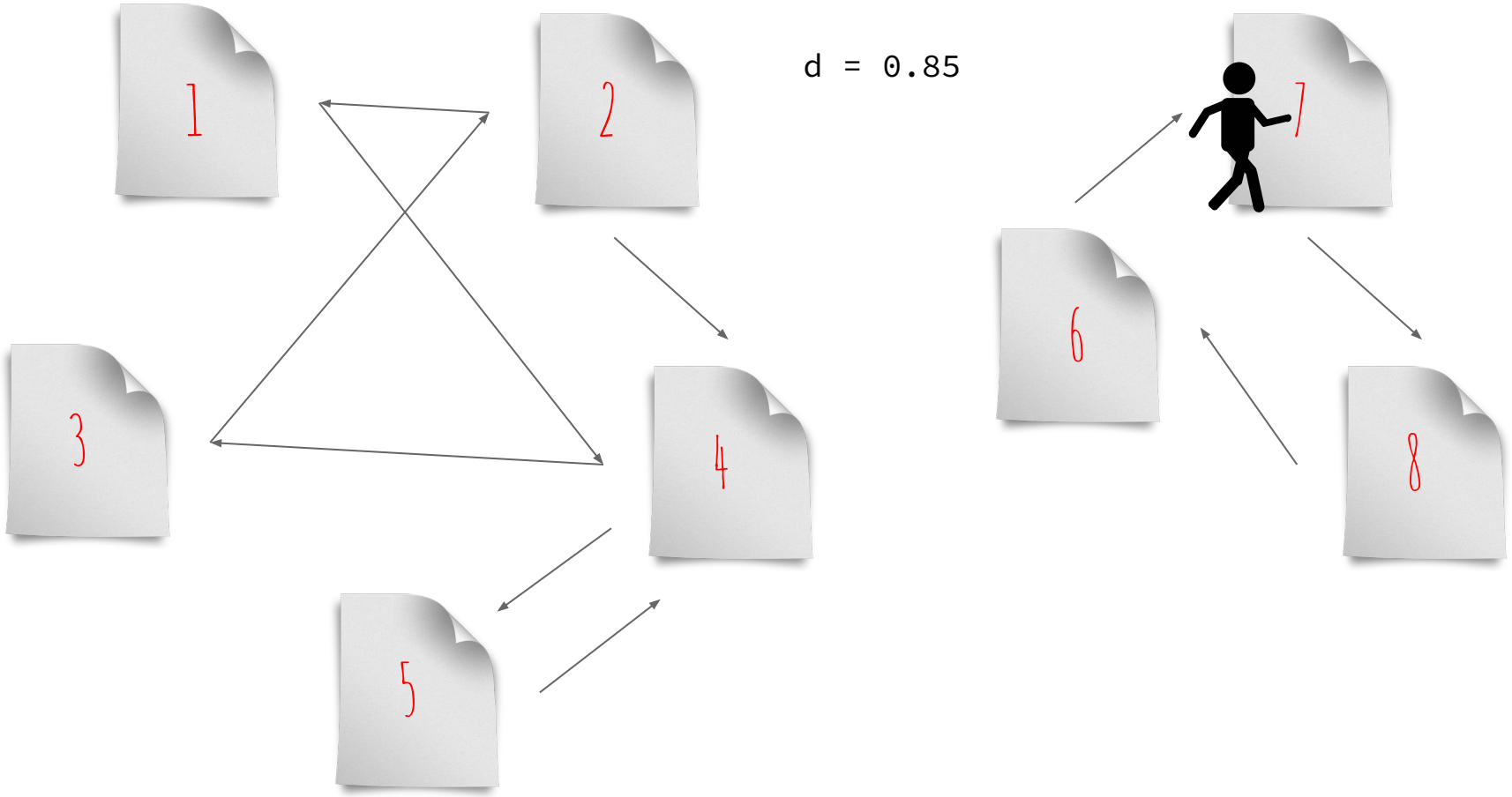

1
1
3
2
2
3
6
4
4
5


1
1
3
2
2
3
7
4
4
5


1
1
3
2
3
3
7
4
4
5

Is the web connected?


1
2
3
4
5
6
7
8


d = 0.85
1
2
3
4
5
6
7
8

What is being saved?

- PageRank
- Number of links per page
- Word occurrences
- Word positions
- Capitalization

DOCID

docID				
1	cat	home	ball	…
2	ball	park	home	…
3	home	paint	people	…

WORDID

wordID				
cat	doc1			…
ball	doc2	doc1		…
home	doc3	doc1	doc2	…

Finally,
Search Queries

- Hits
- Position
- Capitalization
- PageRank
- Anchor text

Thanks!