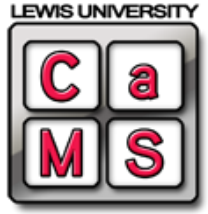
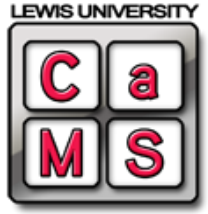
How Web Search Works

Dr. Ray Klump
Professor and Chair
Computer & Mathematical Sciences
Tuesday, September 8, 2015

CaMS Seminar Series - Fall 2015

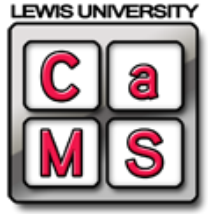


Computer science **is not just** hardware and software.



Computer Science is best when it **deals in ideas**.

Great algorithms come from
great ideas



Web Search

LEWIS UNIVERSITY
CaMS

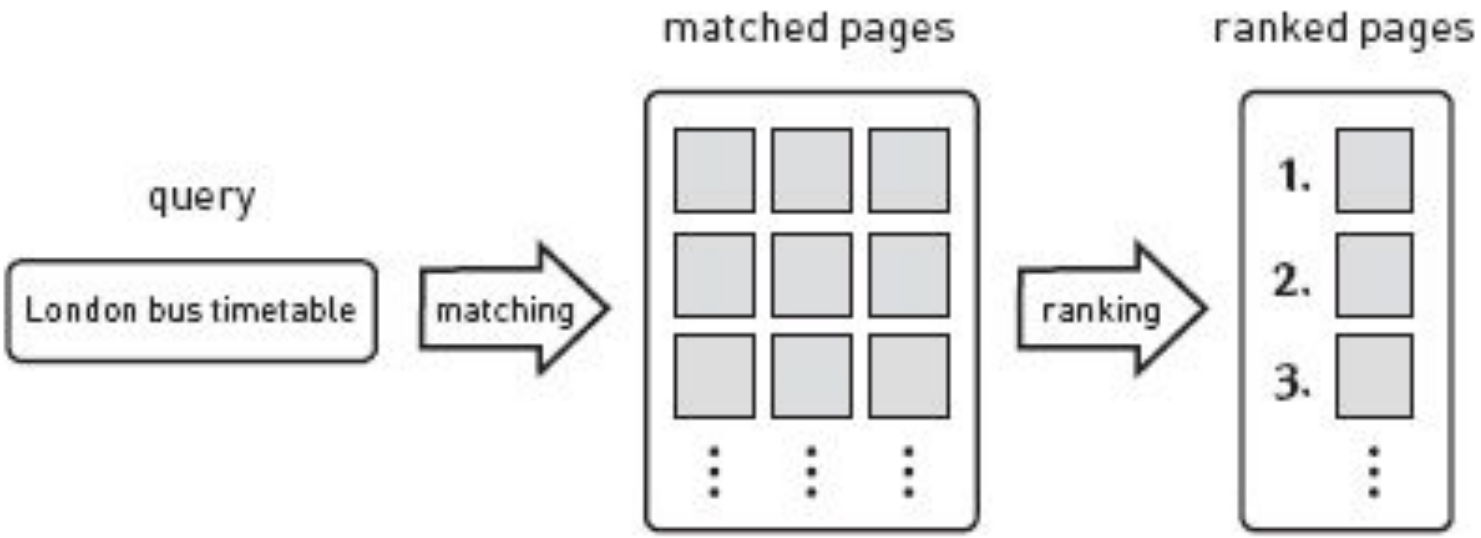
Web search is uses **indexes**.

History

Lycos & Infoseek (1994)
Altavista (1996)
Google (1998)

Steps

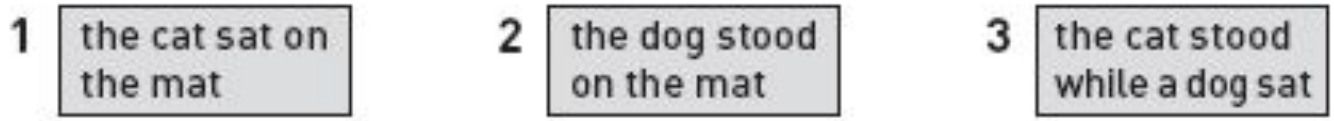
Indexing
Matching
Ranking



Note: illustrations come from *Nine Algorithms that Changed the Future* (MacCormick)

LEWIS UNIVERSITY
C a
M S

Indexing

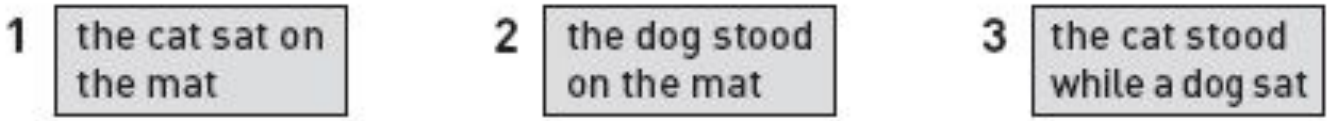


An imaginary World Wide Web that consists of only three pages, numbered 1, 2, and 3.

a	3
cat	1 3
dog	2 3
mat	1 2
on	1 2
sat	1 3
stood	2 3
the	1 2 3
while	3

Matching

Search the “web” for the word “cat”



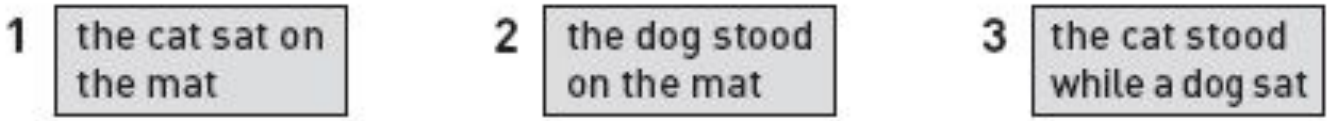
An imaginary World Wide Web that consists of only three pages, numbered 1, 2, and 3.

a	3
cat	1 3
dog	2 3
mat	1 2
on	1 2
sat	1 3
stood	2 3
the	1 2 3
while	3

LEWIS UNIVERSITY

Phrases pose a challenge

Search the “web” for the phrase “cat sat”



An imaginary World Wide Web that consists of only three pages, numbered 1, 2, and 3.

a	3
cat	1 3
dog	2 3
mat	1 2
on	1 2
sat	1 3
stood	2 3
the	1 2 3
while	3

LEWIS UNIVERSITY
C a
M S

Index **words** & **locations**

1	the	cat	sat	on			
	1	*2*	*3*	*4*			
	the	mat					
	5	*6*					

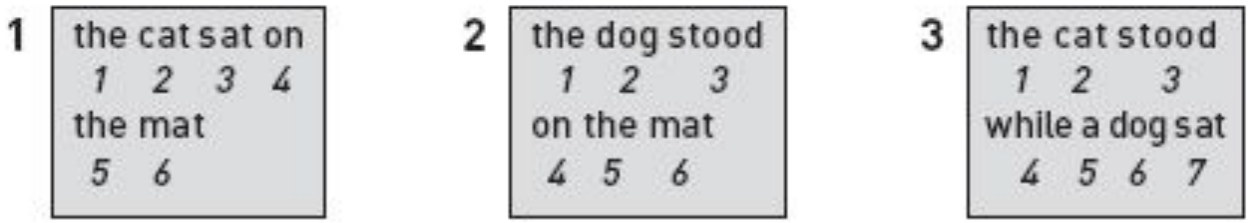
2	the	dog	stood				
	1	*2*	*3*				
	on	the	mat				
	4	*5*	*6*				

3	the	cat	stood				
	1	*2*	*3*				
	while	a	dog	sat			
	4	*5*	*6*	*7*			

a	3-5				
cat	1-2	3-2			
dog	2-2	3-6			
mat	1-6	2-6			
on	1-4	2-4			
sat	1-3	3-7			
stood	2-3	3-3			
the	1-1	1-5	2-1	2-5	3-1
while	3-4				

LEWIS UNIVERSITY

Find “cat sat”

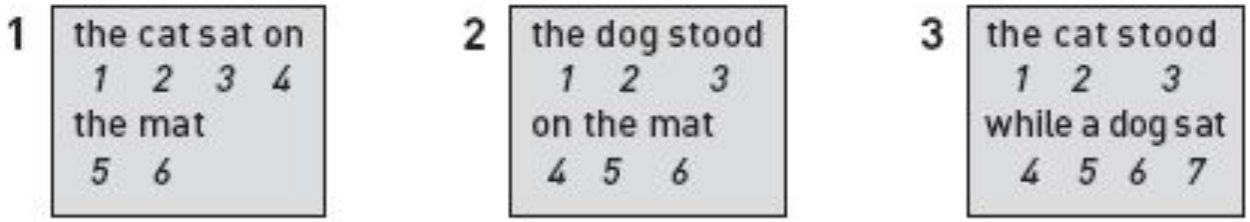


a	3-5				
cat	1-2	3-2			
dog	2-2	3-6			
mat	1-6	2-6			
on	1-4	2-4			
sat	1-3	3-7			
stood	2-3	3-3			
the	1-1	1-5	2-1	2-5	3-1
while	3-4				

LEWIS UNIVERSITY

Near-To Queries

Find pages where cat & dog are within 5 words of each other



a	3-5				
cat	1-2	3-2			
dog	2-2	3-6			
mat	1-6	2-6			
on	1-4	2-4			
sat	1-3	3-7			
stood	2-3	3-3			
the	1-1	1-5	2-1	2-5	3-1
while	3-4				

LEWIS UNIVERSITY

Near-to queries help **discern relevance**.

Which page discusses the causes of malaria?

1 By far the most common cause of malaria is being bitten by an infected mosquito, but there are also other ways to contract the disease.

2 Our cause was not helped by the poor health of the troops, many of whom were suffering from malaria and other tropical diseases.

also	1-19	
...		
cause	1-6	2-2
...		
malaria	1-8	2-19
...		
whom	2-15	

LEWIS UNIVERSITY
CaMS

Word-location indexing
maps **content** to **structure**

Web pages have **parts**

```
<html>
<head>

<link rel="icon" type="image/
<title>Lewis University :: Co
<link rel="stylesheet" type="
href="styles.css" />
<STYLE>
<!--
a {text-decoration: none}
//-->
img.left {
        float:left;
        margin-right:15px;
        margin-bottom:15px;
}
img.right {
        float:right;
        margin-left:15px;
        margin-bottom:15px;
}
</STYLE>
</head>
<body>
```

```
<head>
<title>Not a Meta Tag, but required anyway </title>
<meta name="description" content="Awesome Description
Here">
<meta http-equiv="content-type"
content="text/html;charset=UTF-8">
</head>
```

LEWIS UNIVERSITY
C a
M S

Searching with **structure**

1 <titleStart> my cat <titleEnd> <bodyStart> the cat sat on the mat <bodyEnd>

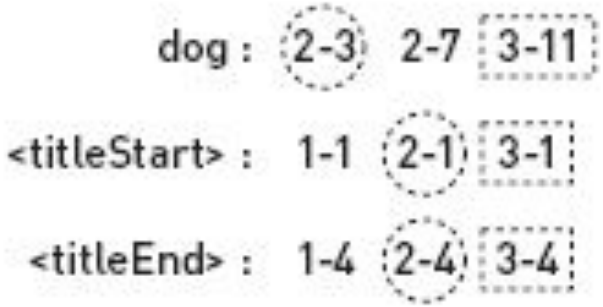
2 <titleStart> my dog <titleEnd> <bodyStart> the dog stood on the mat <bodyEnd>

3 <titleStart> my pets <titleEnd> <bodyStart> the cat stood while a dog sat <bodyEnd>

LEWIS UNIVERSITY

Find “dog” **in the title**

a	3-10
cat	1-3 1-7 3-7
dog	2-3 2-7 3-11
mat	1-11 2-11
my	1-2 2-2 3-2
on	1-9 2-9
pets	3-3
sat	1-8 3-12
stood	2-8 3-8
the	1-6 1-10 2-6 2-10 3-6
while	3-9
<bodyEnd>	1-12 2-12 3-13
<bodyStart>	1-5 2-5 3-5
<titleEnd>	1-4 2-4 3-4
<titleStart>	1-1 2-1 3-1



LEWIS UNIVERSITY
C a
M S

But there's more than **indexing** and **matching**.

PageRank

(Google's MoneyMaker)

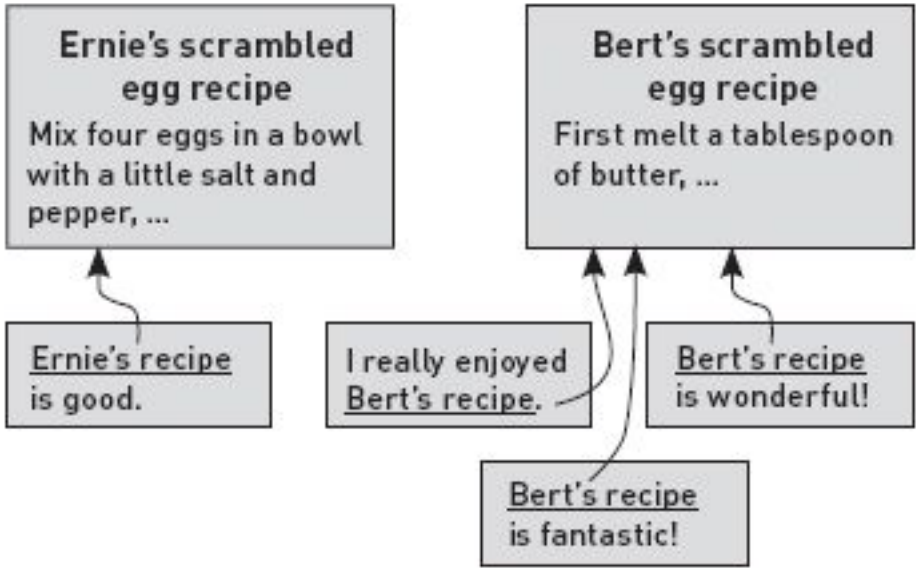
Strategies

Hyperlink Count

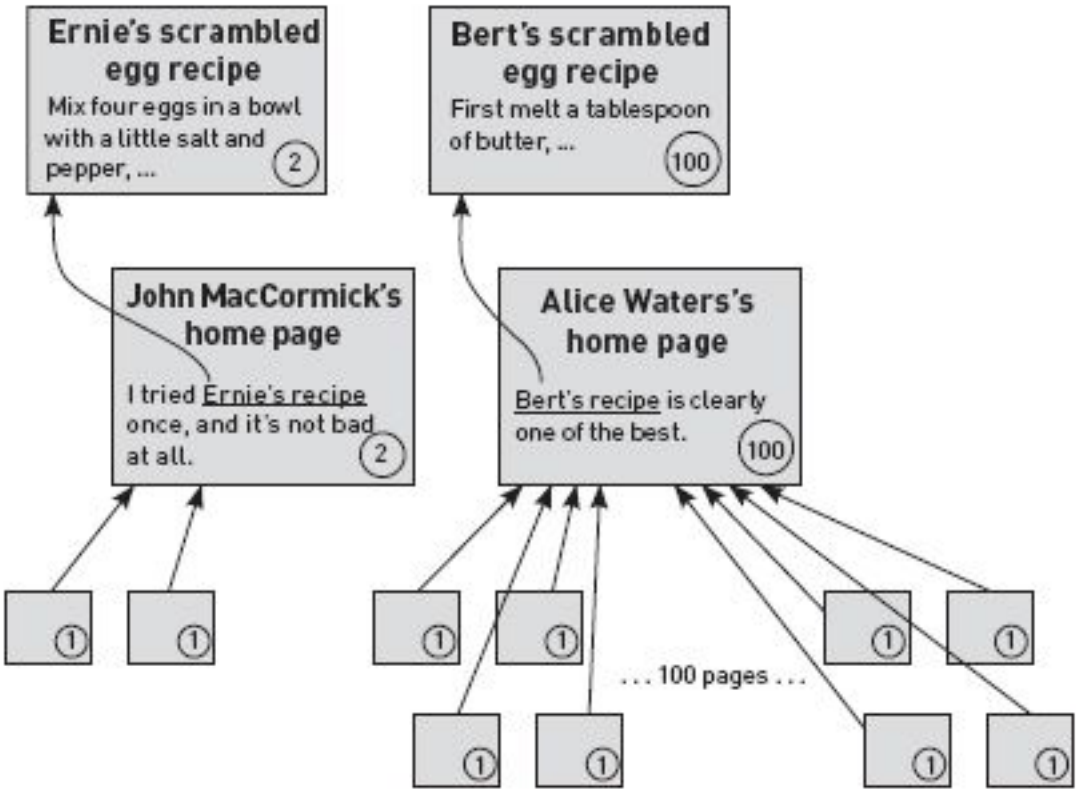
Weighted Hyperlink Count

Weighted Hyperlink Count with Randomness

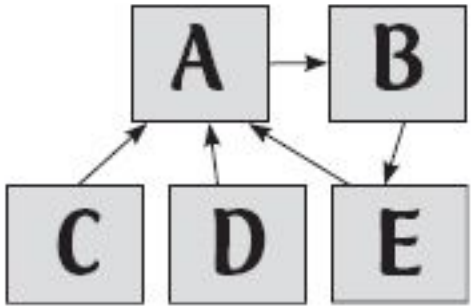
Hyperlink Count

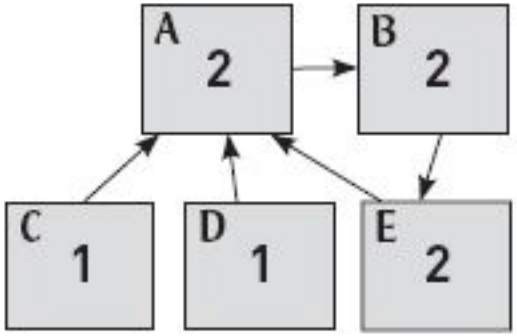


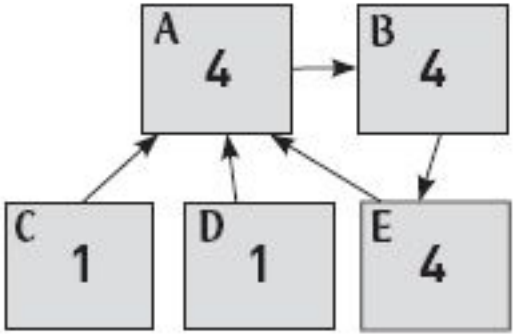
Weighted Hyperlink Count



Cycles pose problems

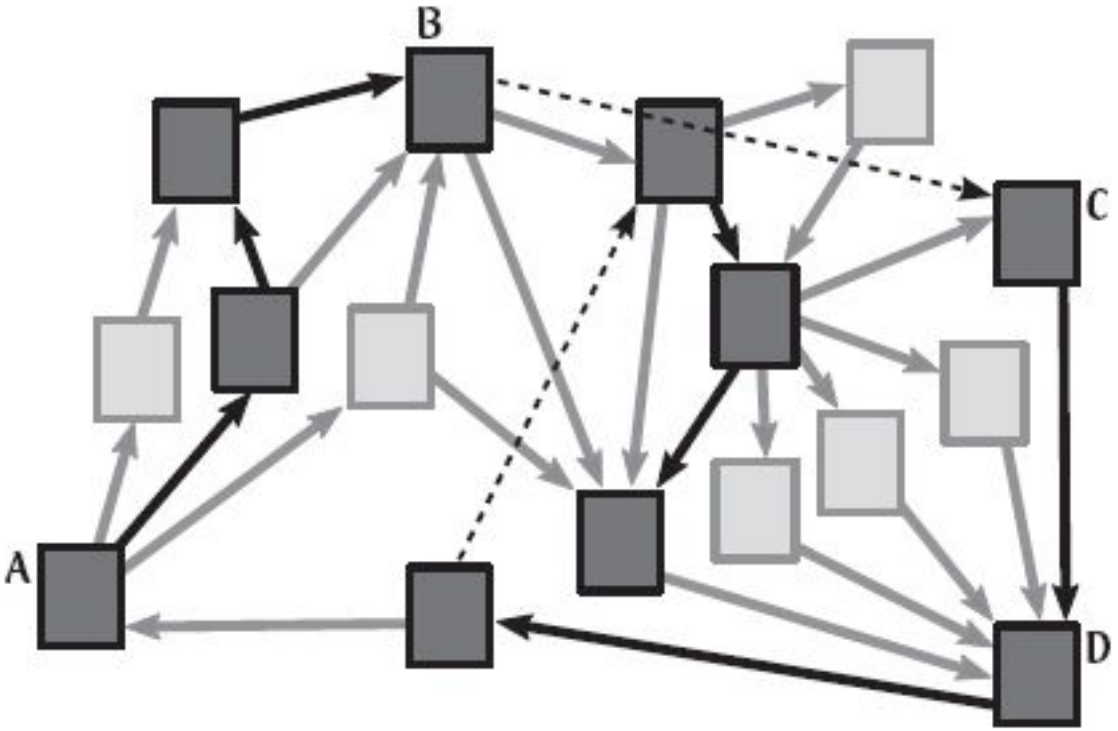






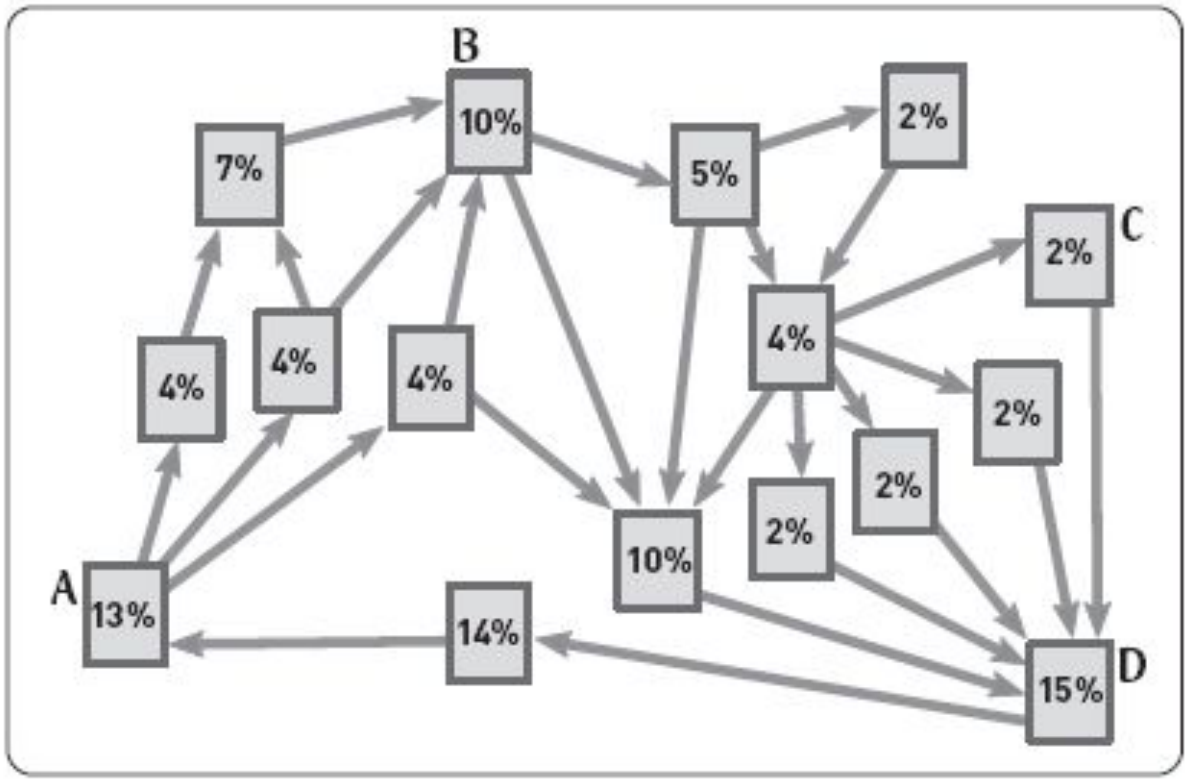
LEWIS UNIVERSITY
CaMS

Derive authority score **using randomness**



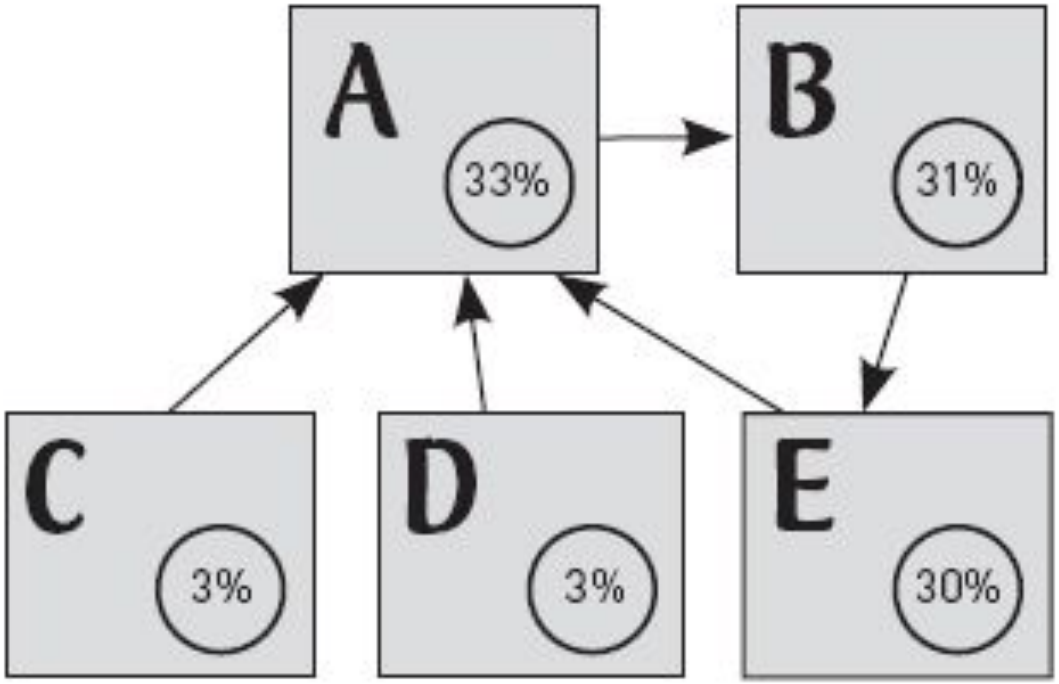
LEWIS UNIVERSITY
CaMS

Weighted Hyperlink Count with **Randomness**



LEWIS UNIVERSITY
CaMS

Pages can then be ranked **even when there are cycles**.



LEWIS UNIVERSITY
CaMS

PageRank today includes up to **200 different factors**.

Search =

Indexing + Matching + Ranking

LEWIS UNIVERSITY
CaMS

Thank you.

LEWIS UNIVERSITY
C a
M S