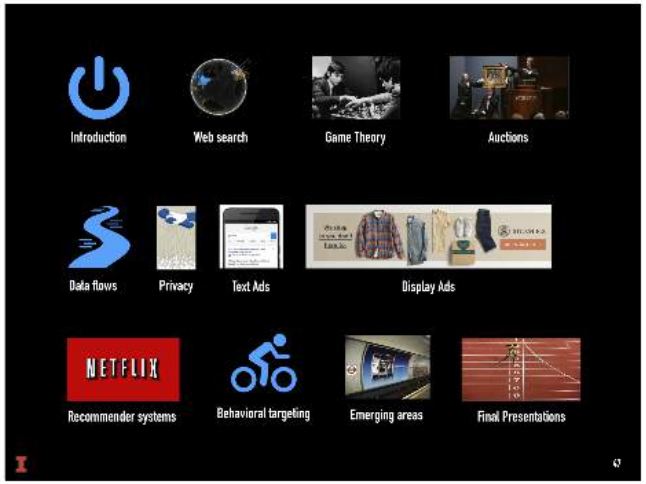


Web Search

hari sundaram
hs1@illinois.edu

Introduction

Web search

Game Theory

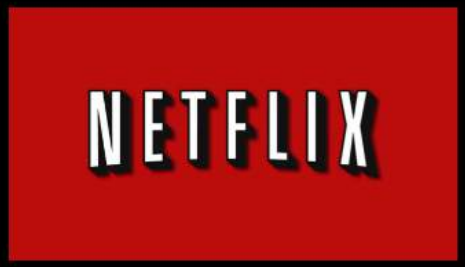
Auctions

Data flows

Privacy

Text Ads

Display Ads



Recommender systems

Behavioral targeting

Emerging areas

Final Presentations

Web Search

hari sundaram

hs1@illinois.edu

some observations


apple

PARLOPHONE

with the beatles

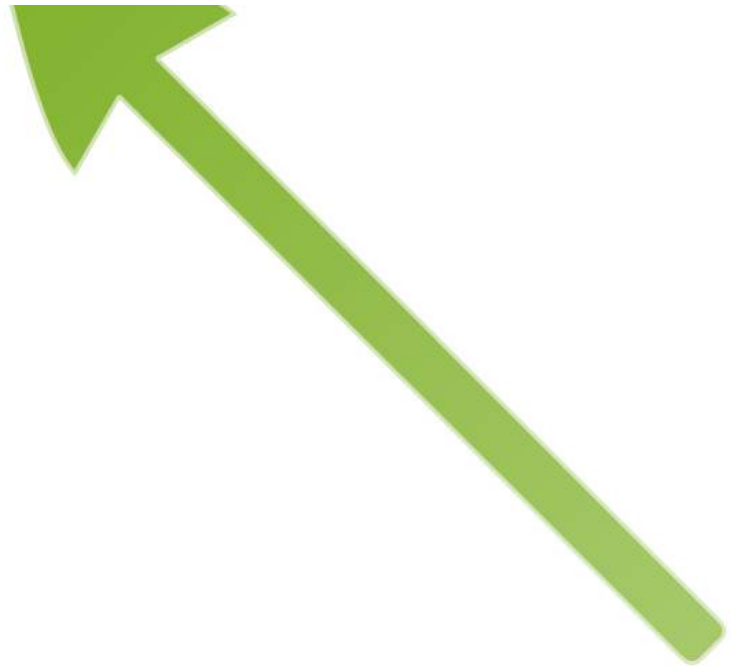
stereo

polysemy

apple

multiple meanings for the same term

"okra"



"ladies' fingers"

synonymy

"okra"

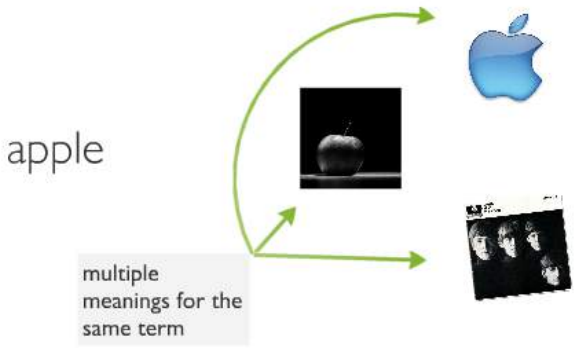
multiple terms mean the same thing

"ladies' fingers"

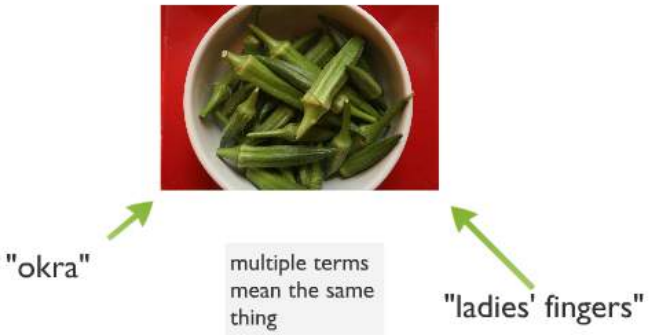
semantics

1

polysemy



synonymy




what happens
when everyone
is an author?

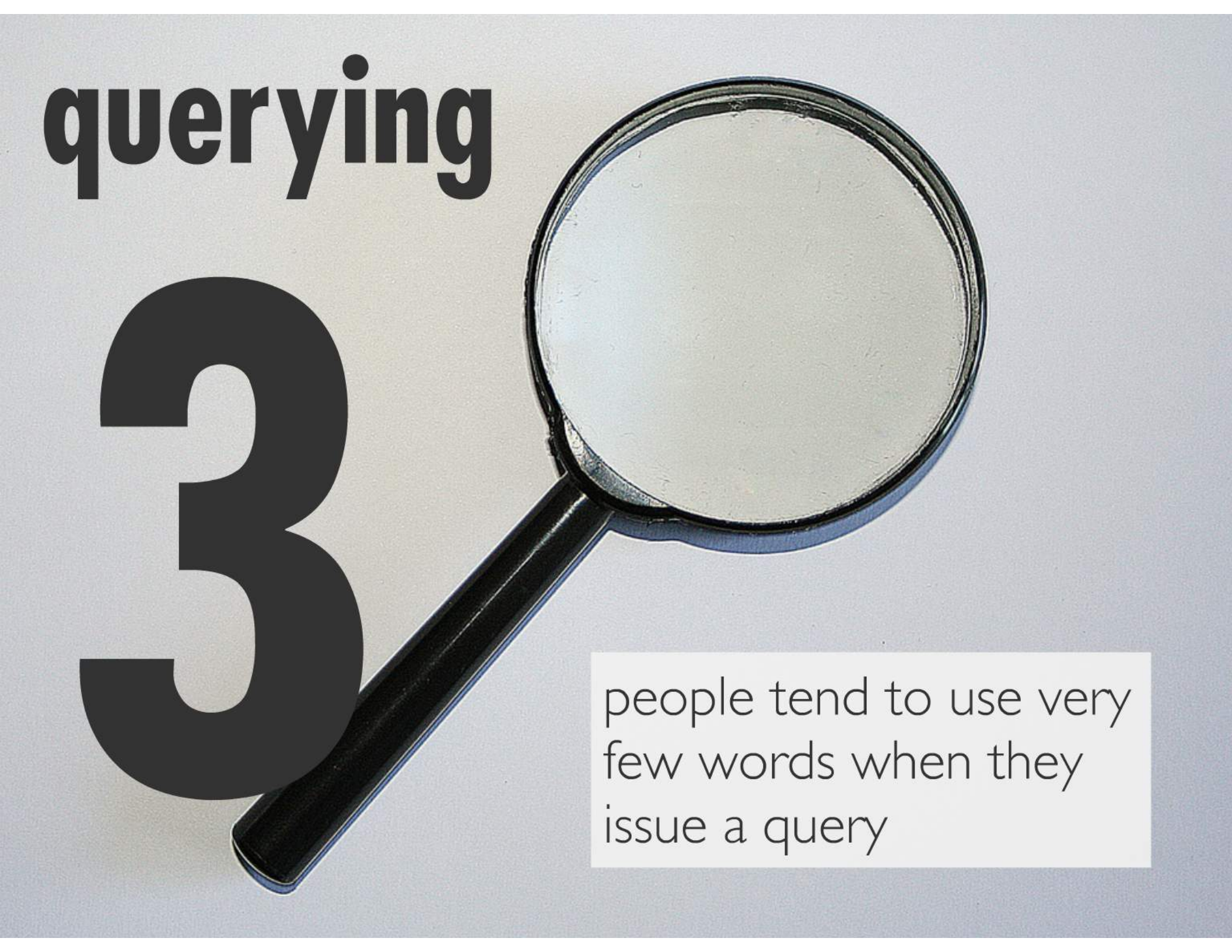
who is
authoritative?

production

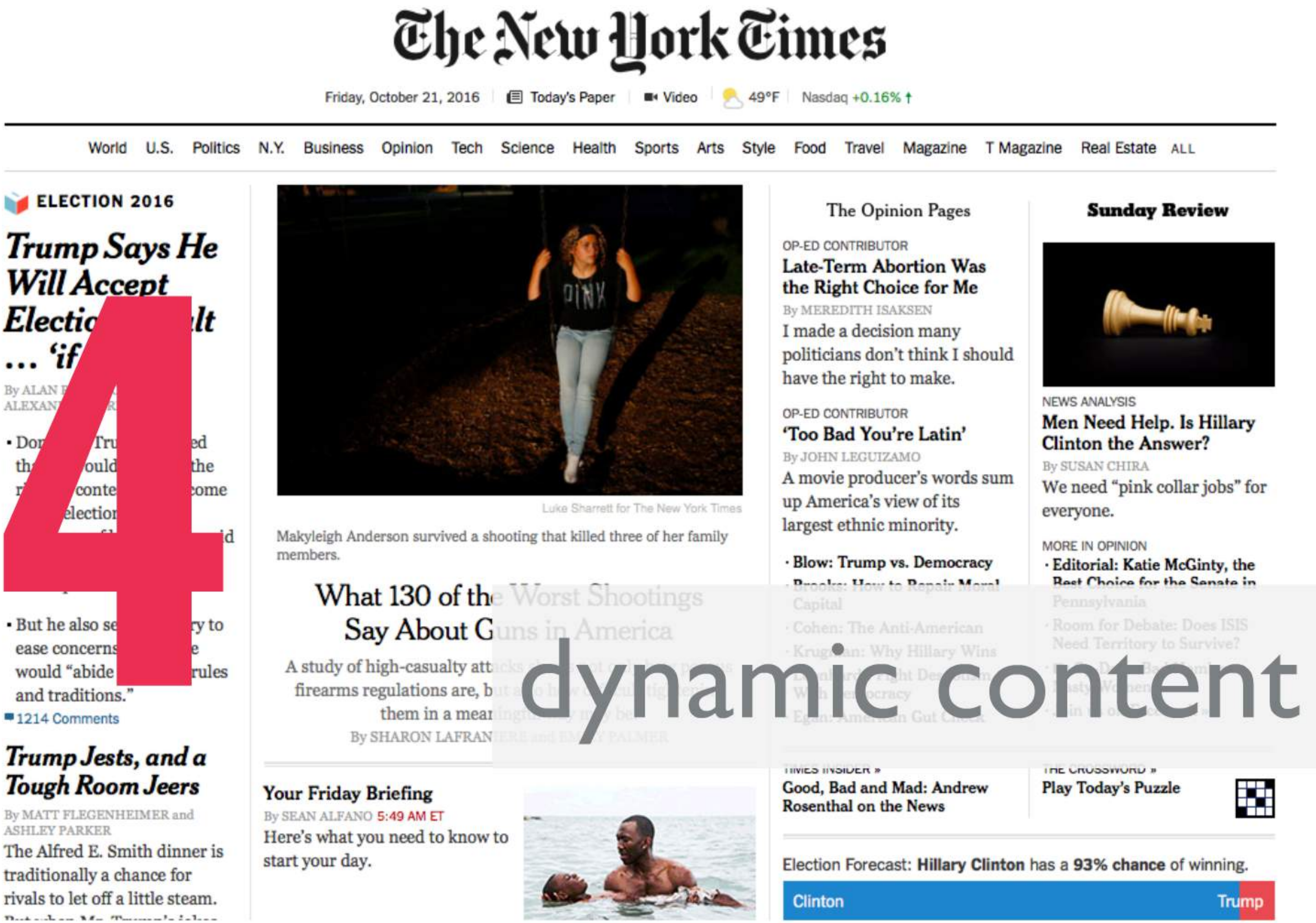
2

querying

3



people tend to use very few words when they issue a query


The New York Times
Friday, October 21, 2016
Today's Paper
Video
49°F
Nasdaq +0.16%
World U.S. Politics N.Y. Business Opinion Tech Science Health Sports Arts Style Food Travel Magazine T Magazine Real Estate ALL
ELECTION 2016
Trump Says He Will Accept Electi… …lt … 'if
4
But he also se… …y to ease concerns… …e would "abide… …rules and traditions."
1214 Comments
Trump Jests, and a Tough Room Jeers
By MATT FLEGENHEIMER and ASHLEY PARKER
The Alfred E. Smith dinner is traditionally a chance for rivals to let off a little steam.
Luke Sharrett for The New York Times
Makyleigh Anderson survived a shooting that killed three of her family members.
What 130 of the Worst Shootings Say About Guns in America
A study of high-casualty att…
firearms regulations are, b…
them in a mea…
By SHARON LAFRAN…
Your Friday Briefing
By SEAN ALFANO 5:49 AM ET
Here's what you need to know to start your day.
The Opinion Pages
OP-ED CONTRIBUTOR
Late-Term Abortion Was the Right Choice for Me
By MEREDITH ISAKSEN
I made a decision many politicians don't think I should have the right to make.
OP-ED CONTRIBUTOR
'Too Bad You're Latin'
By JOHN LEGUIZAMO
A movie producer's words sum up America's view of its largest ethnic minority.
· Blow: Trump vs. Democracy
Sunday Review
NEWS ANALYSIS
Men Need Help. Is Hillary Clinton the Answer?
By SUSAN CHIRA
We need "pink collar jobs" for everyone.
MORE IN OPINION
· Editorial: Katie McGinty, the
dynamic content
Good, Bad and Mad: Andrew Rosenthal on the News
Play Today's Puzzle
Election Forecast: Hillary Clinton has a 93% chance of winning.
Clinton
Trump

A search for "University of Illinois" yields the UIUC page.

There is nothing internal to the UIUC home page that makes it special—it uses UIUC as often as other pages using the term

1 links are essential to ranking; they form implicit endorsements

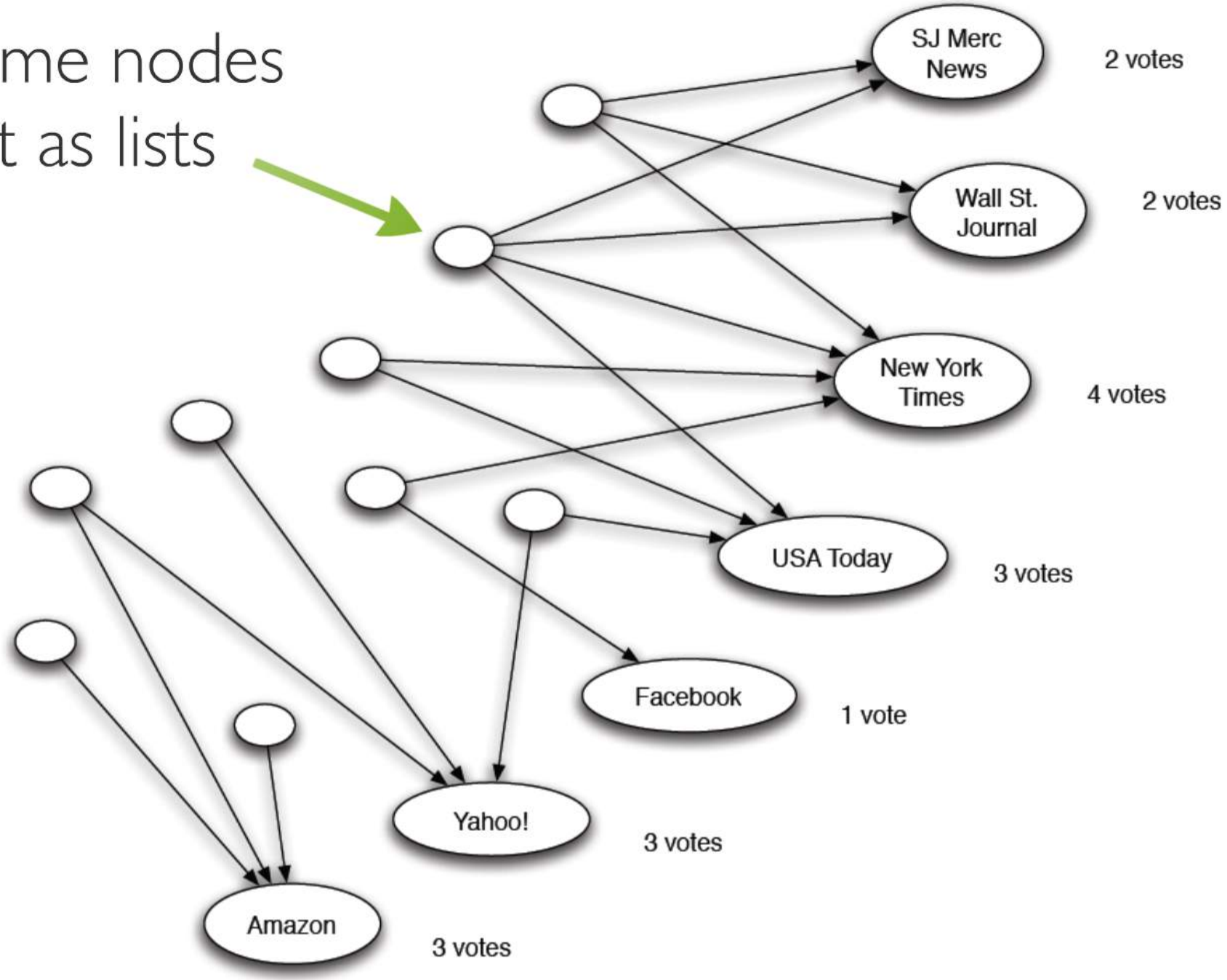
Could we just count the votes from each page linking to UIUC?

2

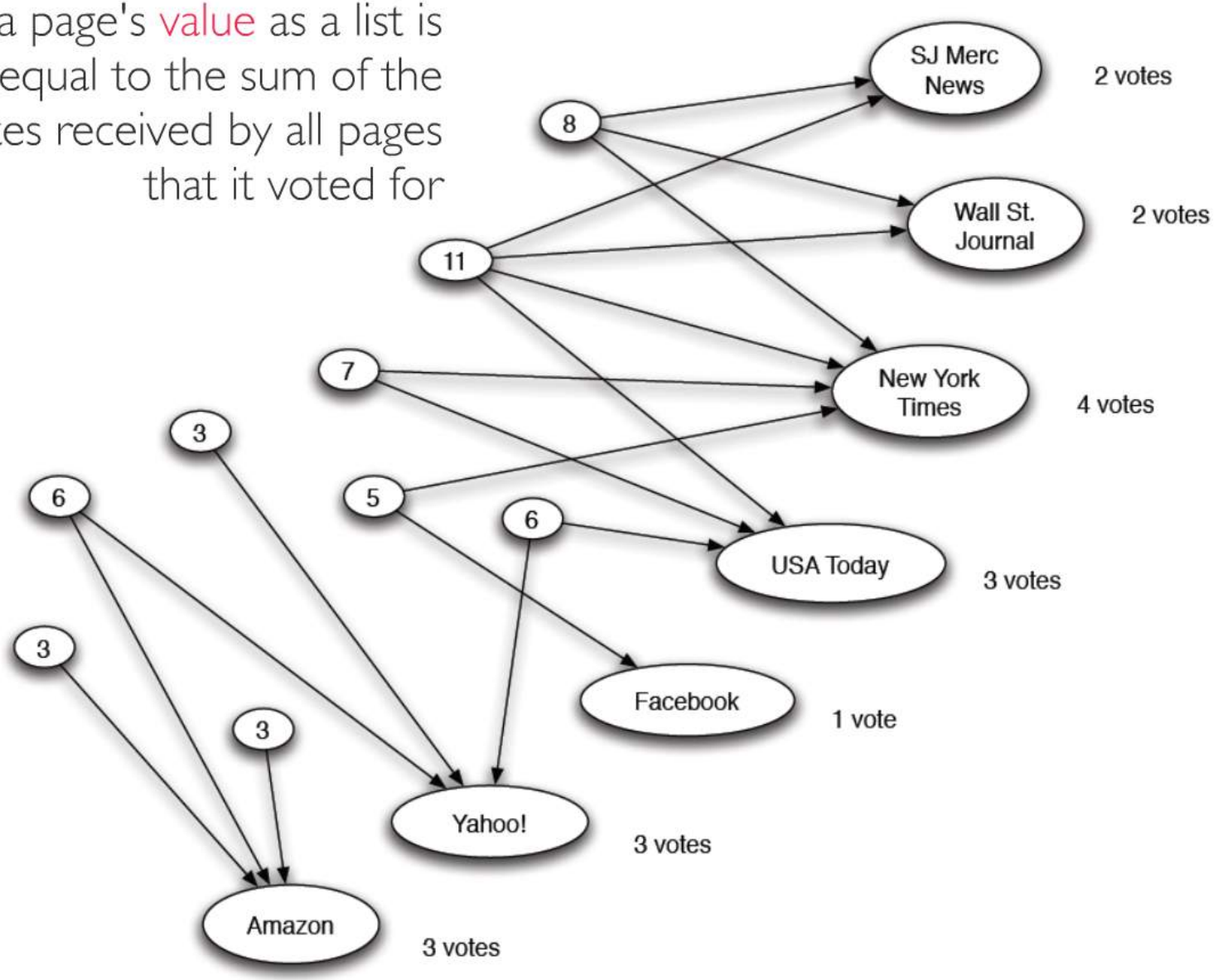
we can exploit the network structure to improve results

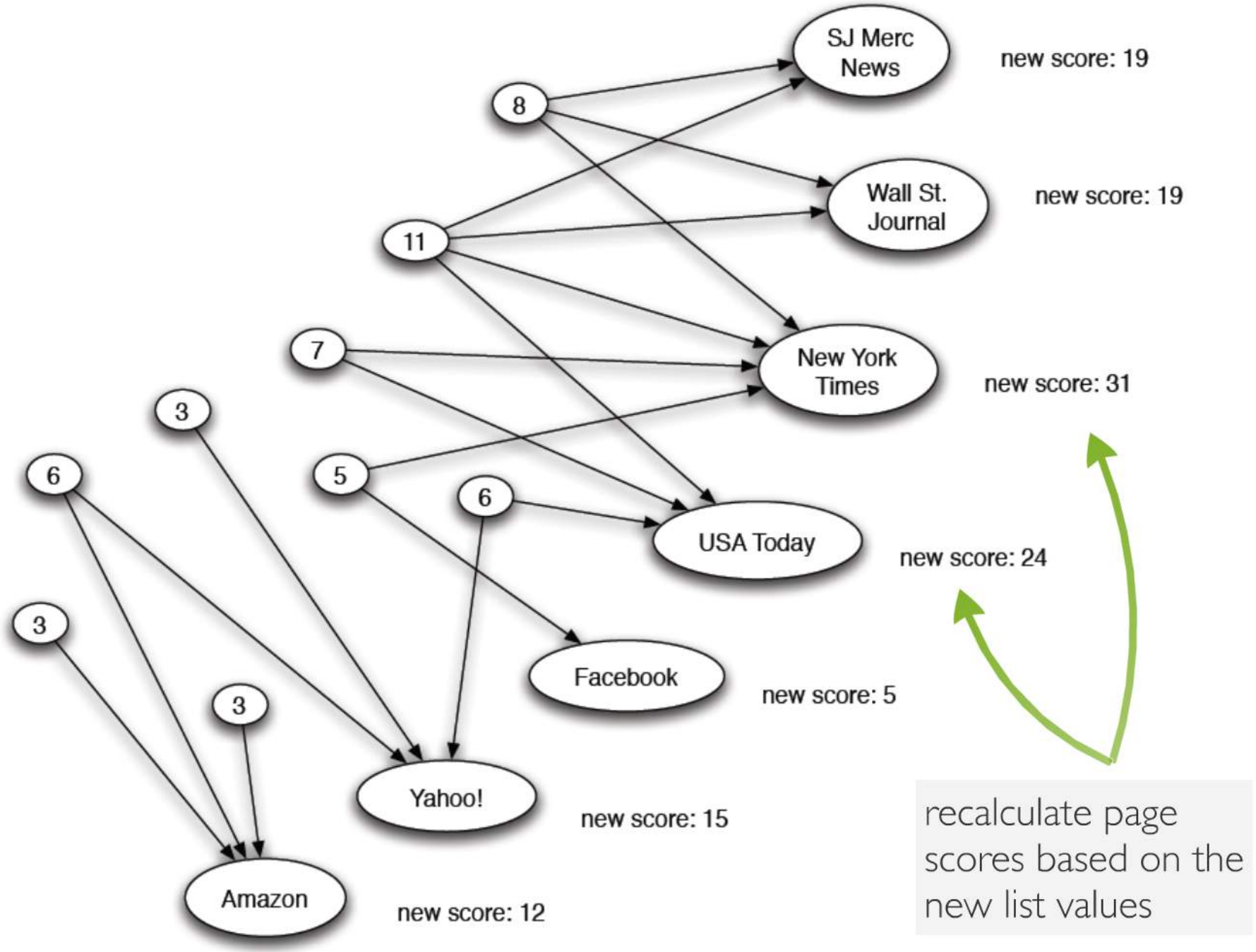
let's query for "newspapers"

some nodes
act as lists



a page's value as a list is equal to the sum of the votes received by all pages that it voted for




SJ Merc News
new score: 19
8
Wall St. Journal
new score: 19
11
7
New York Times
new score: 31
3
6
5
6
USA Today
new score: 24
3
Facebook
new score: 5
3
Yahoo!
new score: 15
Amazon
new score: 12
recalculate page scores based on the new list values

Two kinds of attributes

H & A

hubs

high value lists

every page has a hub score and an authority score

authorities

highly relevant to the query

Two kinds of attributes

H & A

hubs

high value lists

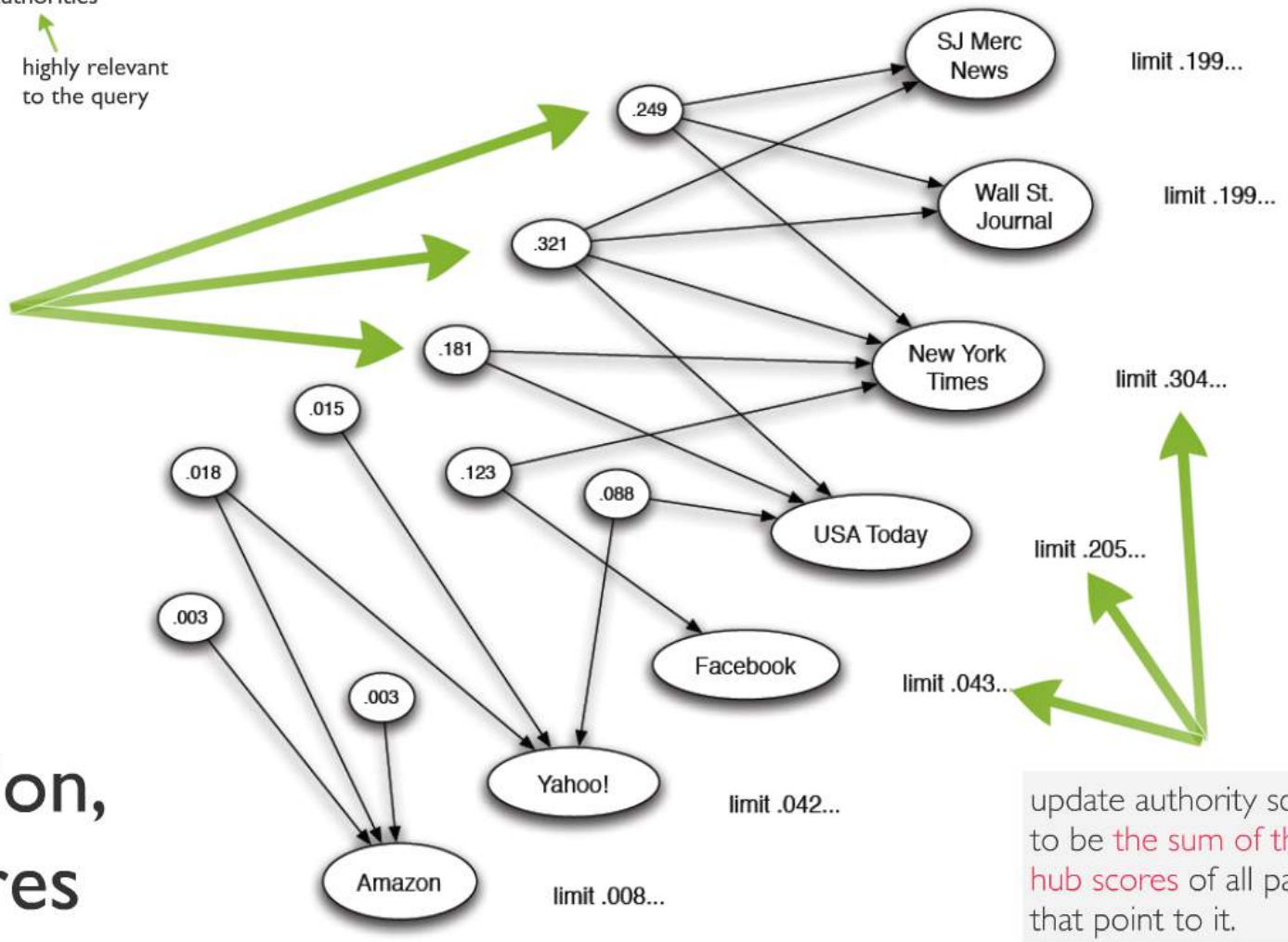
authorities

highly relevant to the query

every page has a hub score and an authority score

for all nodes, set authority and hub scores to be 1

update hub score to be the sum of the authority scores of all pages to which it points.



update authority scores to be the sum of the hub scores of all pages that point to it.

at every iteration, normalize scores

Two kinds of attributes

H & A

hubs

high value lists

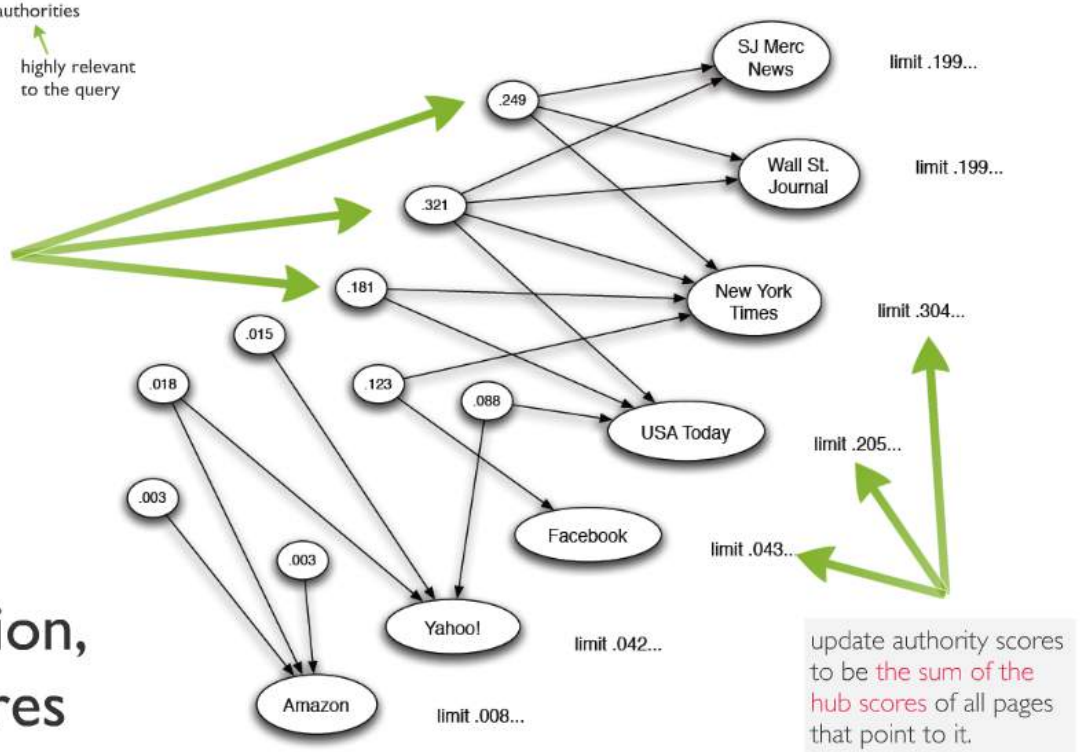
every page has a hub score and an authority score

authorities

highly relevant to the query

for all nodes, set authority and hub scores to be 1

update hub score to be the sum of the authority scores of all pages to which it points.



update authority scores to be the sum of the hub scores of all pages that point to it.

at every iteration, normalize scores

regardless of the initial values

2 The final values are always the same*

1 The HITS algorithm converges

Ranking Schemes

1 links are essential to ranking; they form implicit endorsements

2 we can exploit the network structure to improve results

A search for "University of Illinois" yields the UIUC page.

There is nothing internal to the UIUC home page that makes it special—it uses UIUC as often as other pages using the term

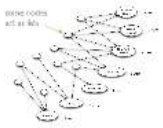
A page is important, if it endorsed by other important people

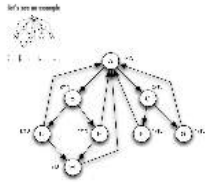
The PageRank algorithm

In a network with n nodes, we assign all nodes the same initial PageRank, set to be 1/n

iterate k times

let's query for "newspapers"





How can we improve upon HITS and PageRank?

Let's make a small change

A page is important, if it endorsed by other important people

The PageRank algorithm

In a network with n nodes, we assign all nodes the same initial PageRank, set to be 1/n

iterate
k times

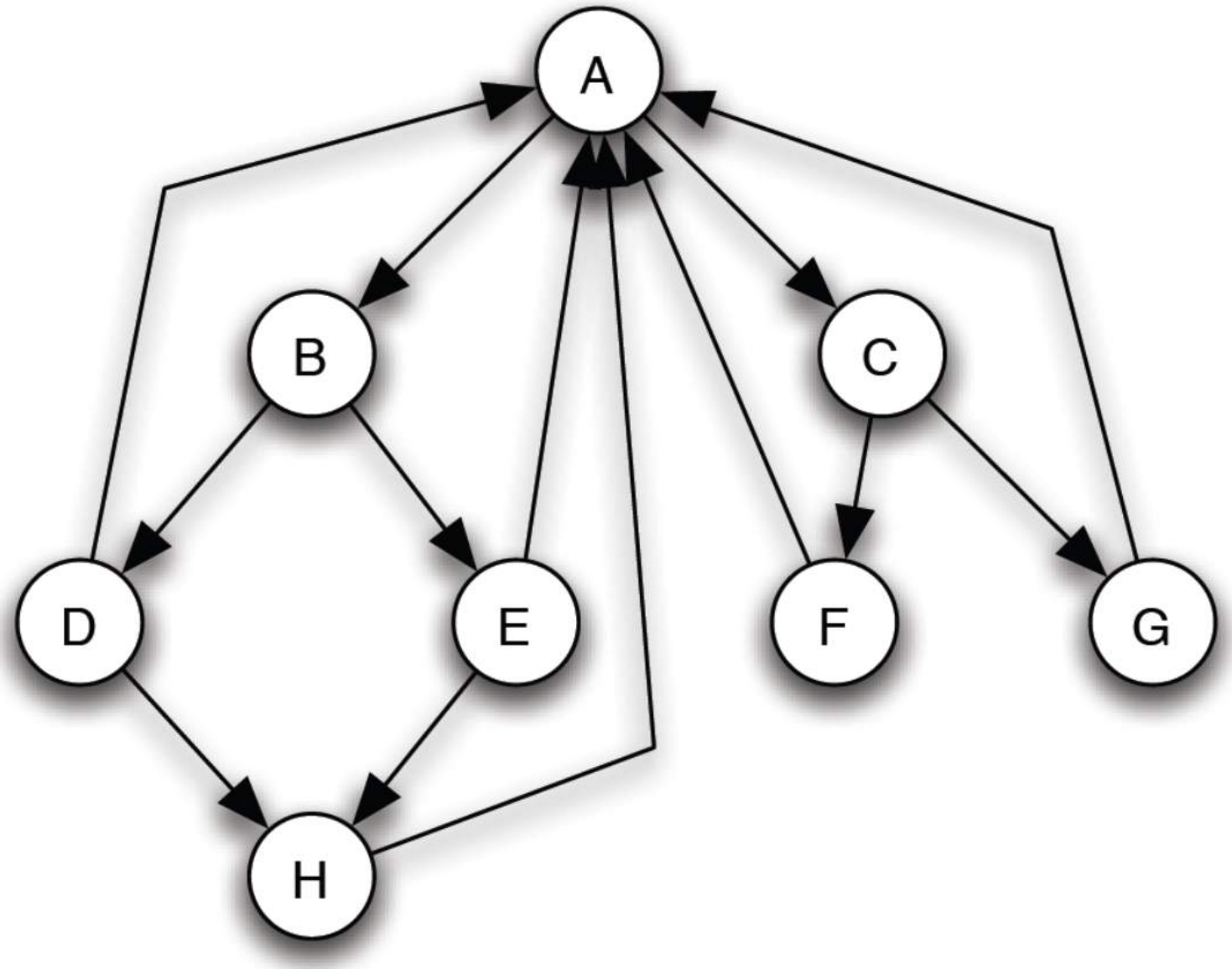
Each page divides its current PageRank equally across its out-going links

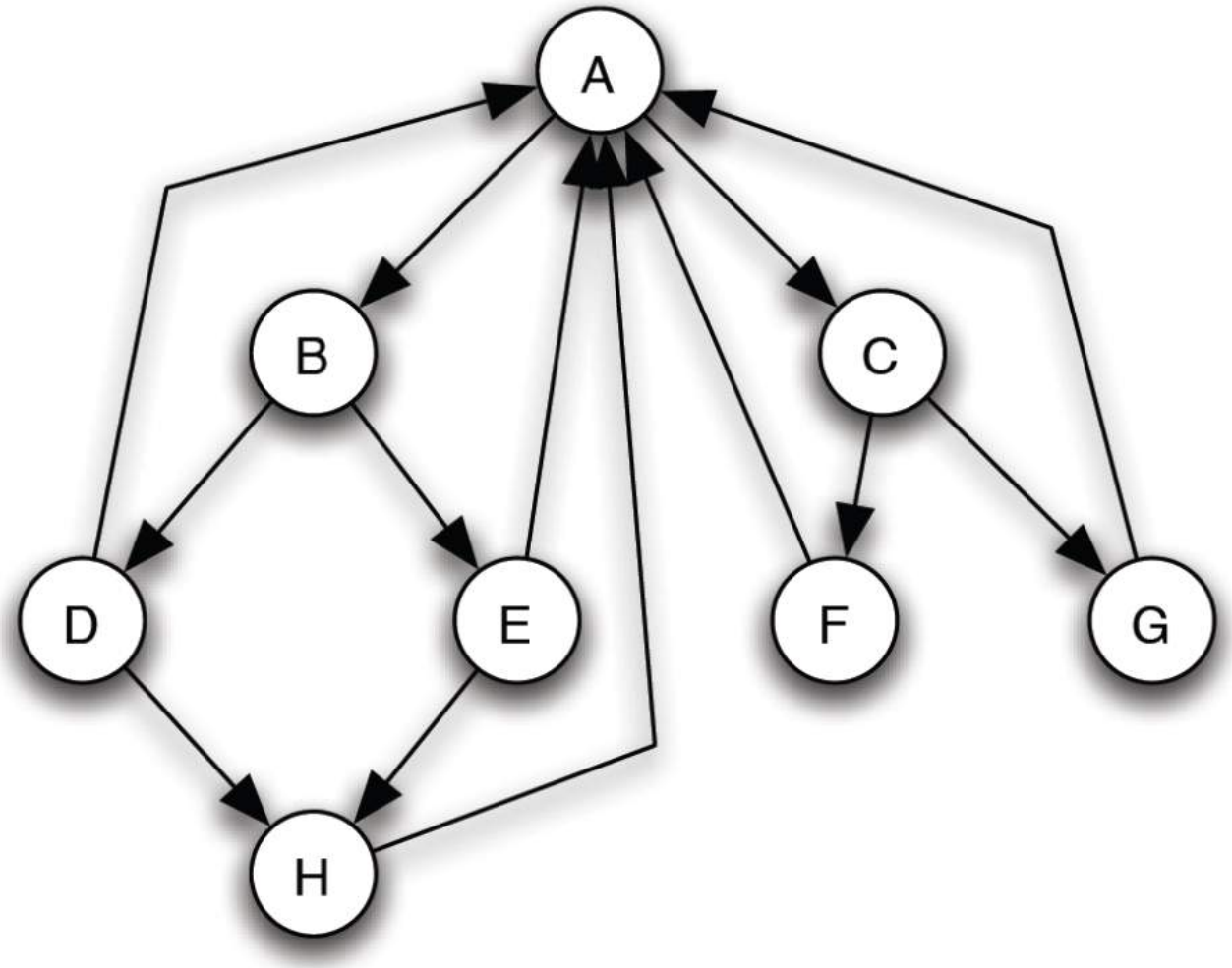
Each page updates its new PageRank to be the sum of the shares it receives.

basic update

If a page has no out-going links, it passes all its current PageRank to itself.

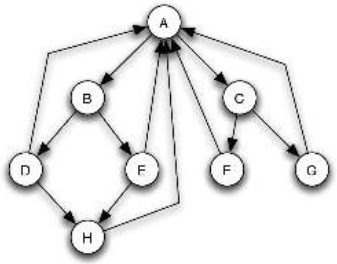
let's see an example



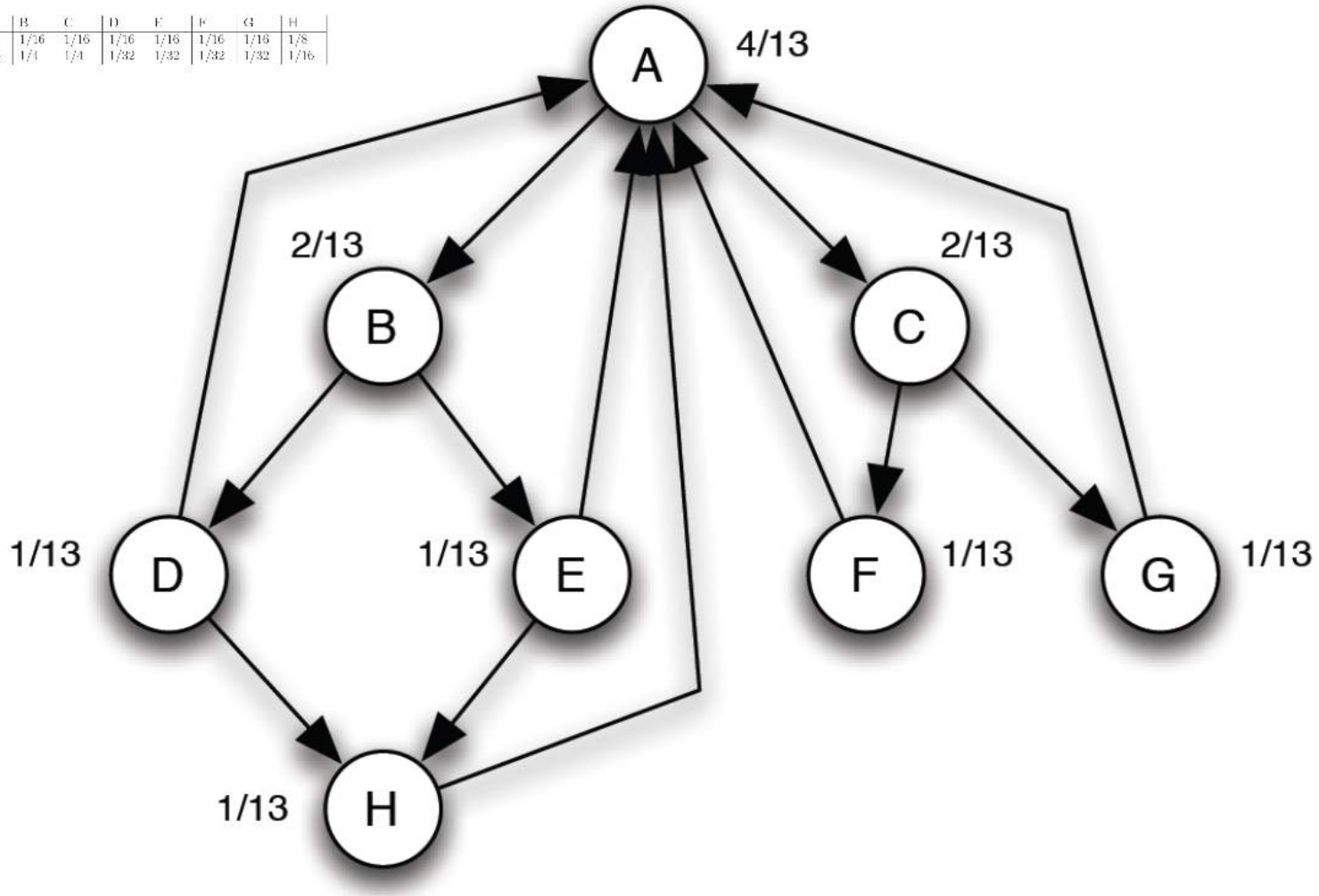


Step	A	B	C	D	E	F	G	H
1	1/2	1/16	1/16	1/16	1/16	1/16	1/16	1/8
2	3/16	1/4	1/4	1/32	1/32	1/32	1/32	1/16

let's see an example



Step	A	B	C	D	E	F	G	H
1	1/2	1/16	1/16	1/16	1/16	1/16	1/16	1/8
2	3/16	1/1	1/4	1/32	1/32	1/32	1/32	1/16

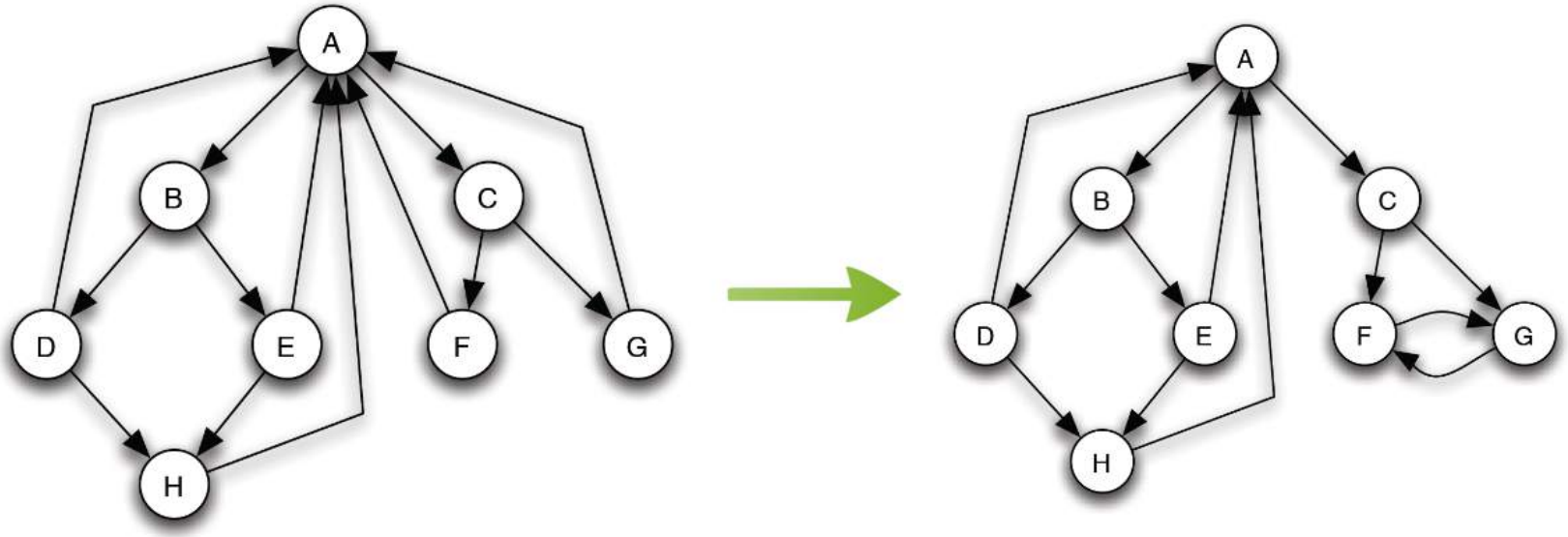


the algorithm is guaranteed to converge

Some things to keep in mind

if the graph is strongly connected, then the equilibrium values are unique

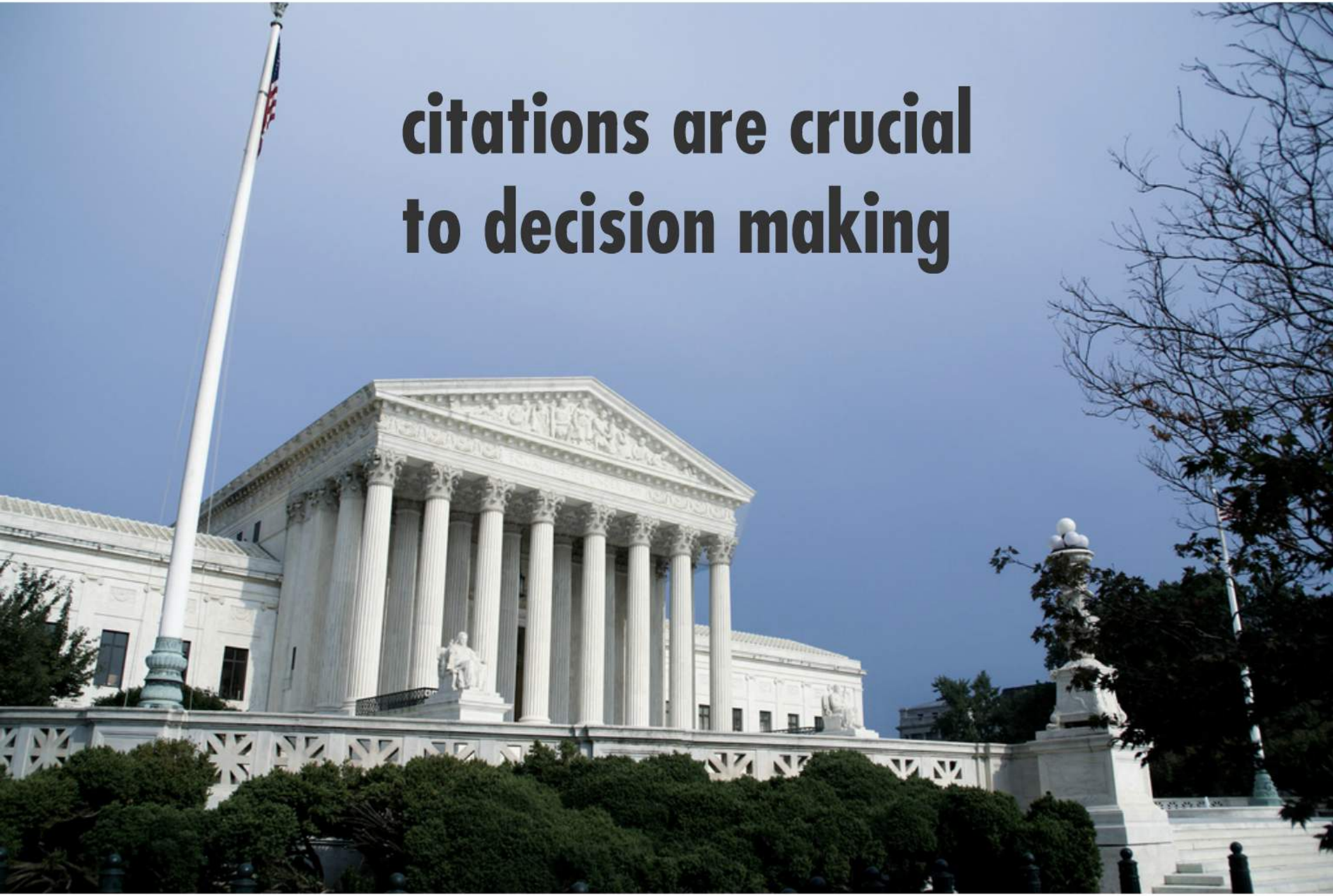
Let's make a small change

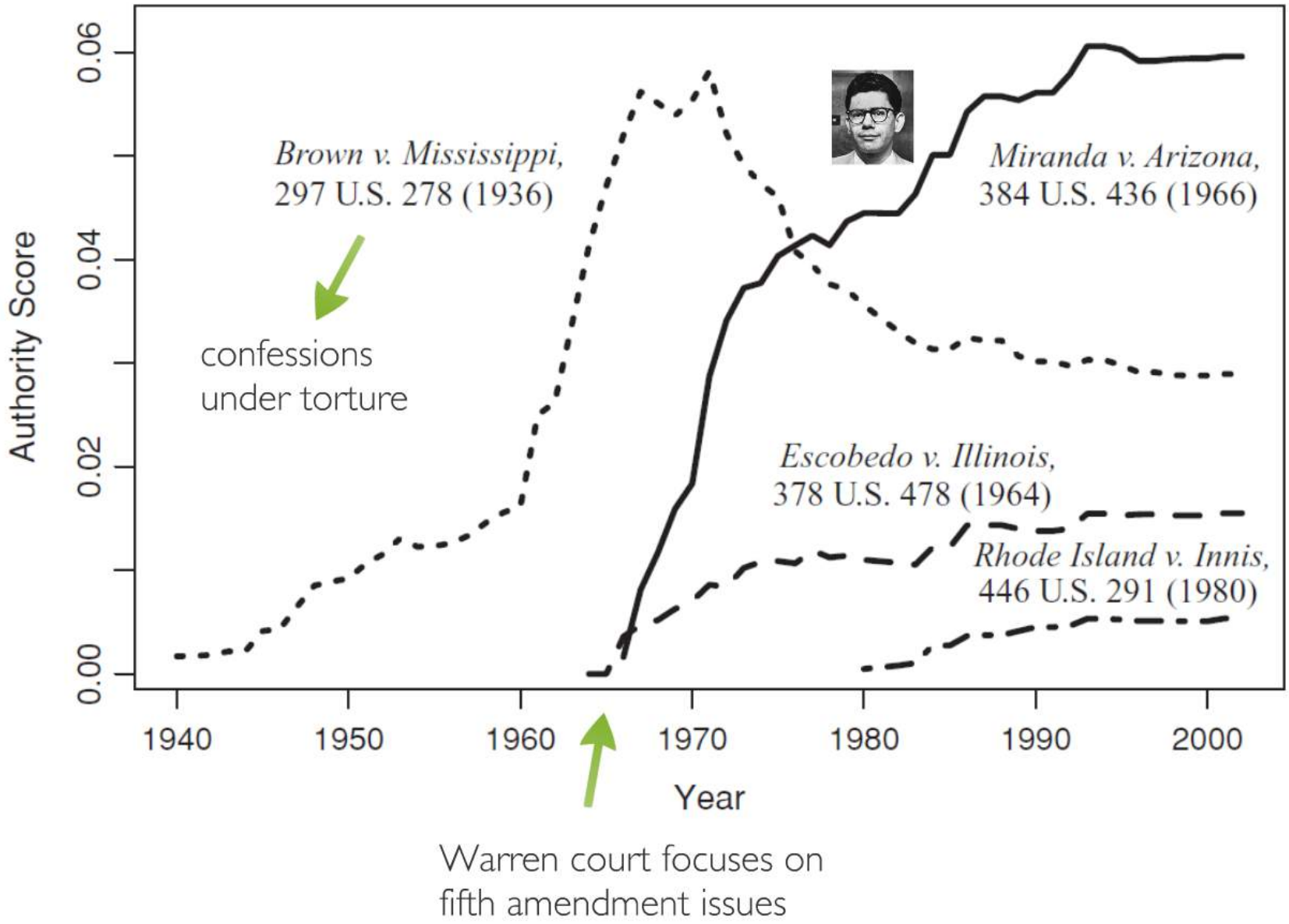


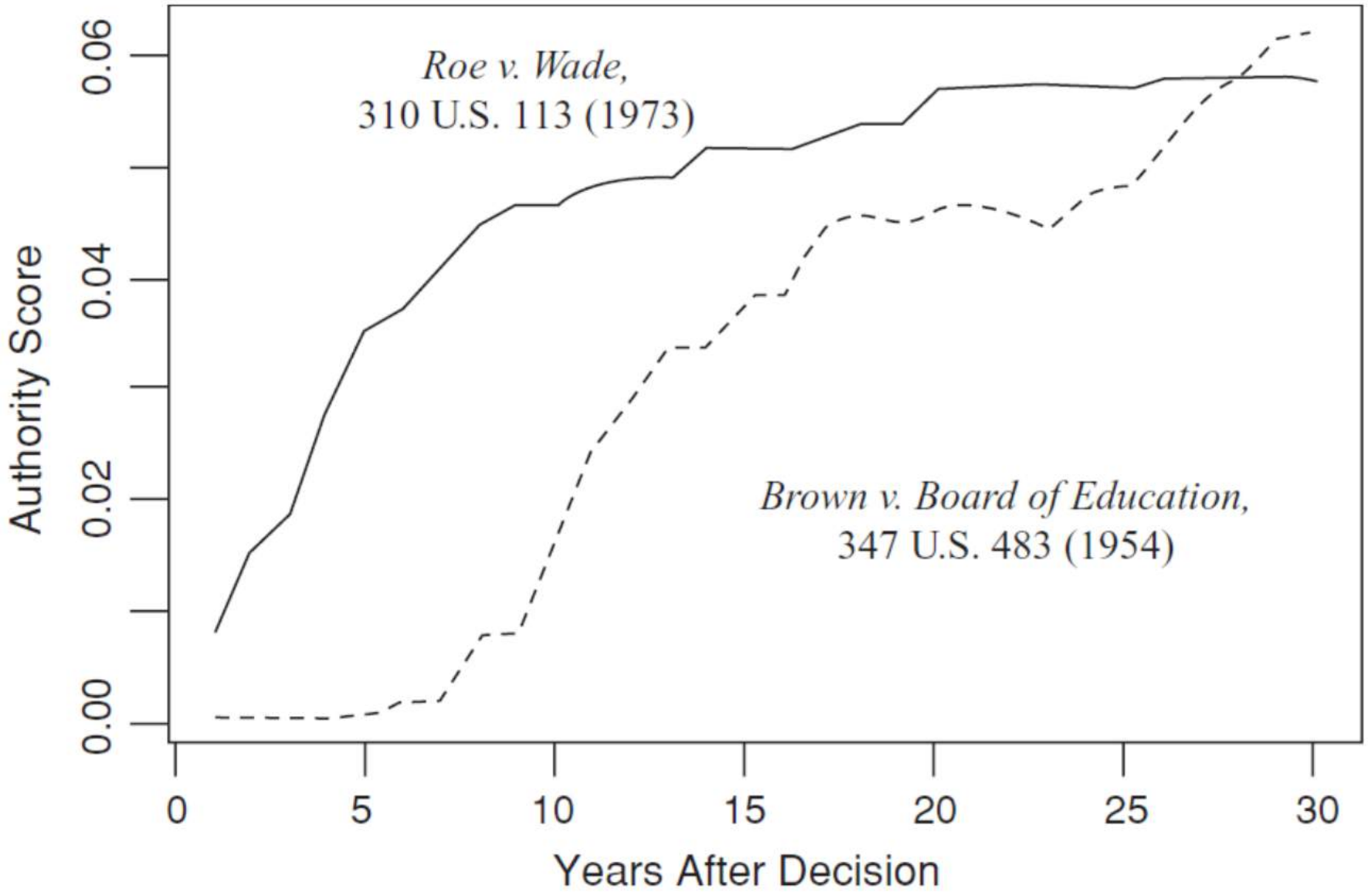
pick a scaling factor s

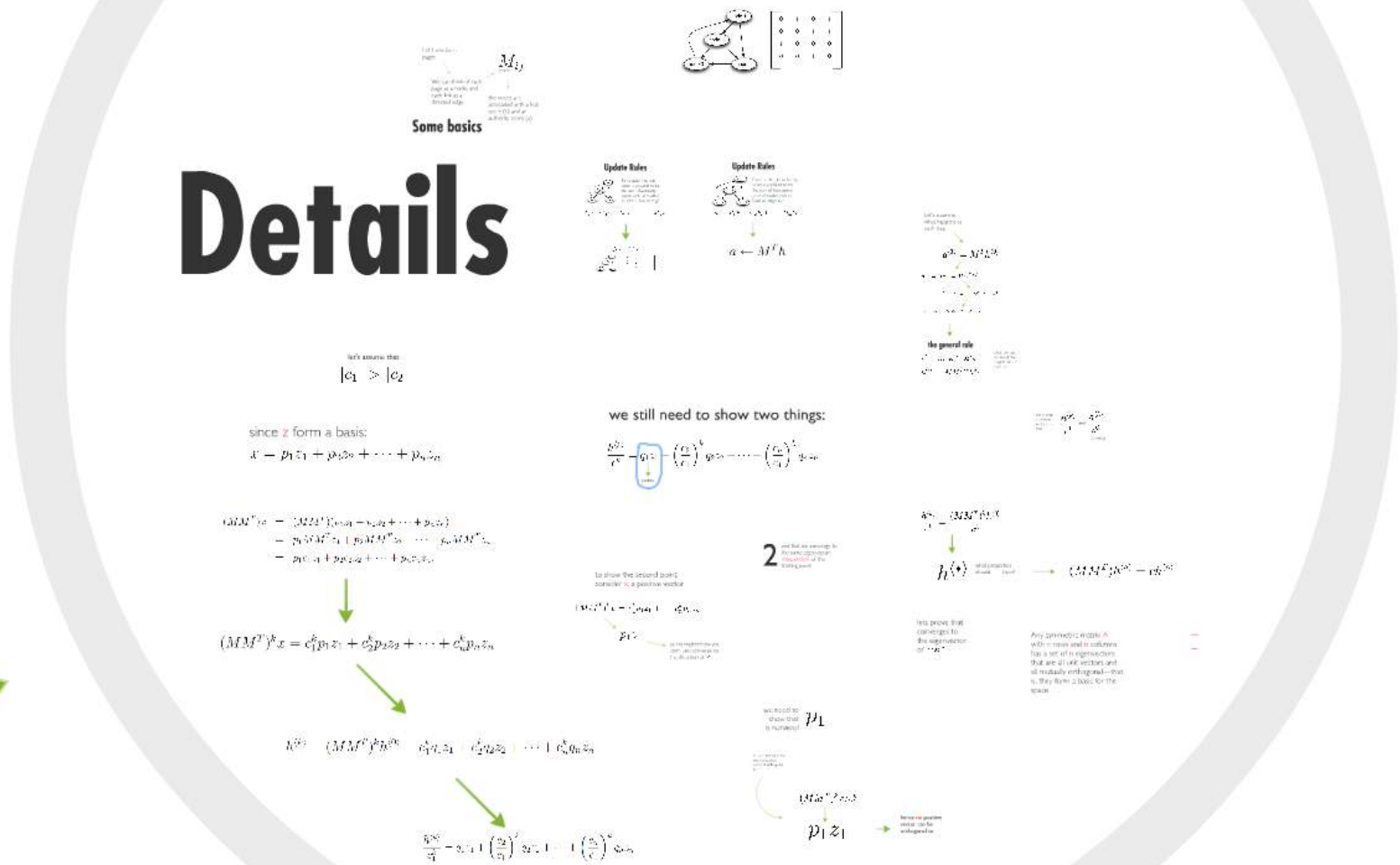
First apply the Basic PageRank Update Rule. Then scale down all PageRank values by a factor of s. We divide the residual 1-s units of PageRank equally over all nodes, giving (1-s)/n to each.

How can we improve upon HITS and PageRank?


citations are crucial
to decision making


Authority Score
0.00
0.02
0.04
0.06
1940
1950
1960
1970
1980
1990
2000
Year
Brown v. Mississippi,
297 U.S. 278 (1936)
confessions
under torture
Miranda v. Arizona,
384 U.S. 436 (1966)
Escobedo v. Illinois,
378 U.S. 478 (1964)
Rhode Island v. Innis,
446 U.S. 291 (1980)
Warren court focuses on
fifth amendment issues


Roe v. Wade,
310 U.S. 113 (1973)
Brown v. Board of Education,
347 U.S. 483 (1954)
Authority Score
0.00
0.02
0.04
0.06
0
5
10
15
20
25
30
Years After Decision


Some basics
Details
Update Rules
Update Rules
since z form a basis:
we still need to show two things:
2

Let there be n pages

We can think of each page as a node, and each link as a directed edge

$$M_{ij}$$

There is an adjacency matrix associated with this graph

the nodes are associated with a hub score (h) and an authority score (a)

Some basics

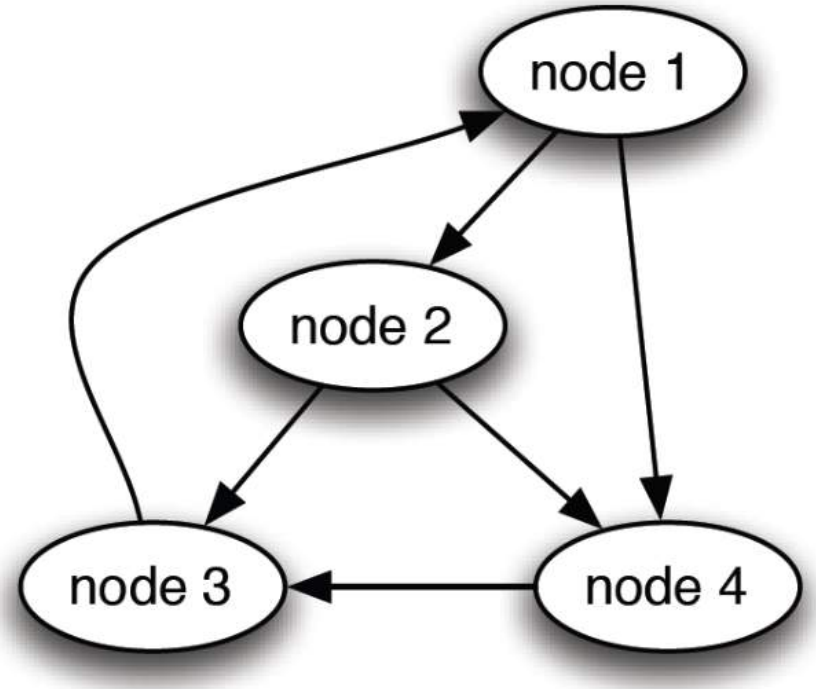
Let there be n pages

We can think of each page as a node, and each link as a directed edge

M_{ij}

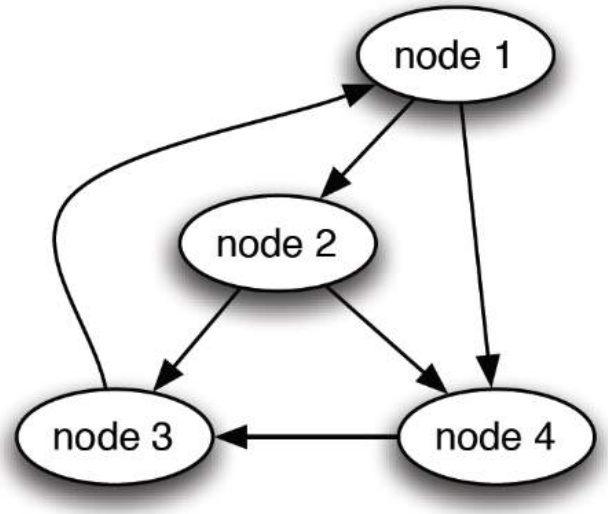
There is an adjacency matrix associated with this graph

the nodes are associated with a hub score (h) and an authority score (a)



$$\begin{bmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

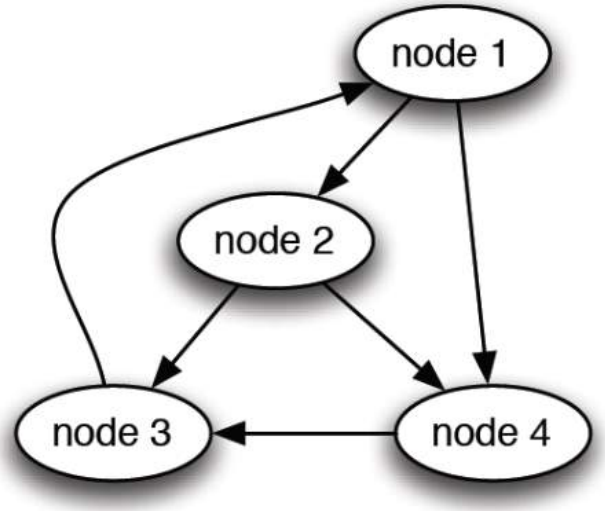
Update Rules



For a node i, its hub score is updated to be the sum of authority scores over all nodes j to which i has an edge.

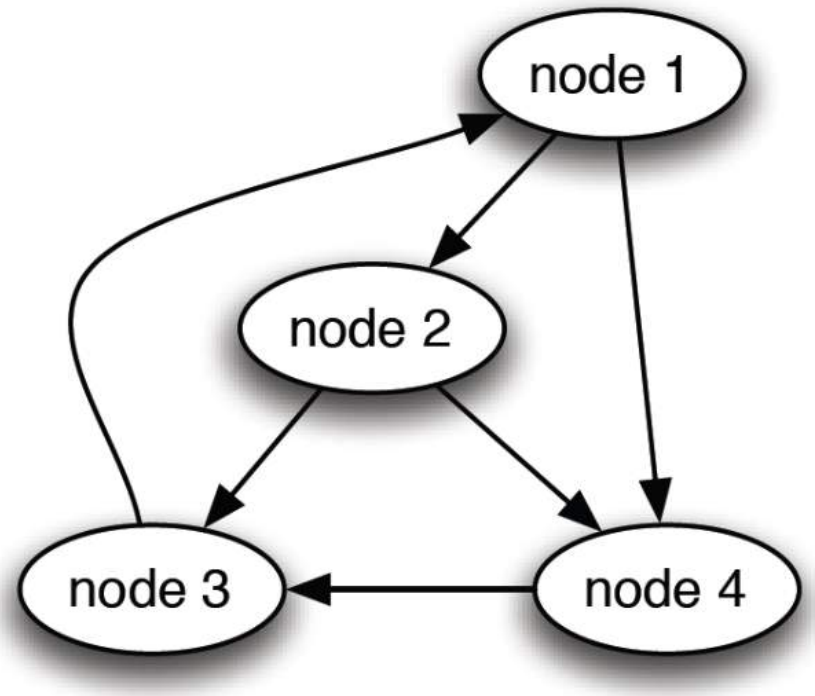
$$h_i \leftarrow M_{i1}a_1 + M_{i2}a_2 + \cdots + M_{in}a_n$$

$$h \leftarrow Ma$$



$$\begin{bmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 2 \\ 6 \\ 4 \\ 3 \end{bmatrix} = \begin{bmatrix} 9 \\ 7 \\ 2 \\ 4 \end{bmatrix}$$

Update Rules



For a node i, its authority score is updated to be the sum of hub scores over all nodes j which have an edge to i

$$a_i \leftarrow M_{1i}h_1 + M_{2i}h_2 + \cdots + M_{ni}h_n$$

$$a \longleftarrow M^T h$$

Let's examine what happens at each step.

step.

$$a^{\langle 1 \rangle} = M^T h^{\langle 0 \rangle}$$

$$Ma^{\langle 1 \rangle} = MM^T h^{\langle 0 \rangle}$$

$$a^{\langle 1 \rangle} = M^T$$

$$h^{\langle 1 \rangle} = M a^{\langle 1 \rangle} = M M^T h^{\langle 0 \rangle}$$

$$a^{\langle 2 \rangle} = M^T h^{\langle 1 \rangle} = M^T M$$

$$Ma^{\langle 1 \rangle} = MM^T h^{\langle 0 \rangle}$$

$$a^{\langle 2 \rangle} = M^T h^{\langle 1 \rangle} = M^T M M^T h^{\langle 0 \rangle}$$

$$^{\langle 2 \rangle} = MM^T MM^T h^{\langle 0 \rangle} = (MM^T)^2 h^{\langle 0 \rangle}$$

$$a^{\langle 2 \rangle} = M^T h^{\langle 1 \rangle} = M^T M M$$

$$h^{\langle 2 \rangle} = M a^{\langle 2 \rangle} = M M^T M M^T h^{\langle 0 \rangle} = (M M^T)^2 h^{\langle 0 \rangle}$$

the general rule

$$a^{\langle k \rangle} = (M^T M)^{k-1} M^T h^{\langle 0 \rangle}$$

$$h^{\langle k \rangle} = (M M^T)^k h^{\langle 0 \rangle}$$

the general rule

$$a^{\langle k \rangle} = (M^T M)^{k-1} M^T h^{\langle 0 \rangle}$$

$$h^{\langle k \rangle} = (M M^T)^k h^{\langle 0 \rangle}$$

what can we say about the magnitudes of h and a?

there exist constants c and d such that

$$\frac{h^{\langle k \rangle}}{c^k} \quad \text{and} \quad \frac{a^{\langle k \rangle}}{d^k}$$

converge

$$\frac{h^{\langle k \rangle}}{c^k} = \frac{(MM^T)^k h^{\langle 0 \rangle}}{c^k}$$

$$h^{\langle * \rangle}$$

what properties should $h^{\langle * \rangle}$ have?

$$(MM^T)h^{\langle * \rangle} = ch^{\langle * \rangle}$$

lets prove that $\frac{h^{\langle k \rangle}}{c^k}$ converges to the eigenvector of MM^T

Any symmetric matrix A with n rows and n columns has a set of n eigenvectors that are all unit vectors and all mutually orthogonal—that is, they form a basis for the space $\mathbf{R}^n$.

eigenvectors

$$z_1, z_2, \cdots, z_n$$

$$|c_1| \geq |c_2| \geq \cdots \geq |c_n|$$

eigenvalues

let's assume that

$$|c_1| > |c_2|$$

$$|c_1| > |c_2|$$

since z form a basis:

$$x = p_1 z_1 + p_2 z_2 + \cdots + p_n z_n$$

$$MM^T)x = (MM^T)(p_1 z_1 + p_2 z_2 + \cdots + p_n z_n)$$

since z form a basis:

$$x = p_1 z_1 + p_2 z_2 + \cdots + p_n z_n$$

$$\begin{aligned}
(MM^T)x &= (MM^T)(p_1 z_1 + p_2 z_2 + \cdots + p_n z_n) \\
&= p_1 MM^T z_1 + p_2 MM^T z_2 + \cdots + p_n MM^T z_n \\
&= p_1 c_1 z_1 + p_2 c_2 z_2 + \cdots + p_n c_n z_n,
\end{aligned}$$

$$(MM^T)^k x = c_1^k p_1 z_1 + c_2^k p_2 z_2 + \cdots + c_n^k p_n z_n$$

$$
\begin{aligned}
&= p_1 M M^T z_1 + p_2 M M^T z_2 + \cdots + p_n M M^T z_n \\
&= p_1 c_1 z_1 + p_2 c_2 z_2 + \cdots + p_n c_n z_n,
\end{aligned}
$$

$$
(M M^T)^k x = c_1^k p_1 z_1 + c_2^k p_2 z_2 + \cdots + c_n^k p_n z_n
$$

$$
h^{\langle k \rangle} = (M M^T)^k h^{\langle 0 \rangle} = c_1^k q_1 z_1 + c_2^k q_2 z_2
$$

$$^{k}x = c_1^k p_1 z_1 + c_2^k p_2 z_2 + \cdots + c_n^k p_n z_n$$

$$h^{\langle k \rangle} = (MM^T)^k h^{\langle 0 \rangle} = c_1^k q_1 z_1 + c_2^k q_2 z_2 + \cdots + c_n^k q_n z_n$$

$$\frac{h^{\langle k \rangle}}{c_1^k} = q_1 z_1 + \left(\frac{c_2}{c_1}\right)^k q_2 z_2 + \cdots + \left(\frac{c_n}{c_1}\right)$$

$$\frac{h^{\langle k \rangle}}{c_1^k} = q_1 z_1 + \left(\frac{c_2}{c_1}\right)^k q_2 z_2 + \cdots + \left(\frac{c_n}{c_1}\right)^k q_n z_n$$

we still need to show two things:

$$\frac{h^{\langle k \rangle}}{c_1^k} = q_1 z_1 + \left(\frac{c_2}{c_1}\right)^k q_2 z_2 + \cdots + \left(\frac{c_n}{c_1}\right)^k q_n z_n$$

positive

2

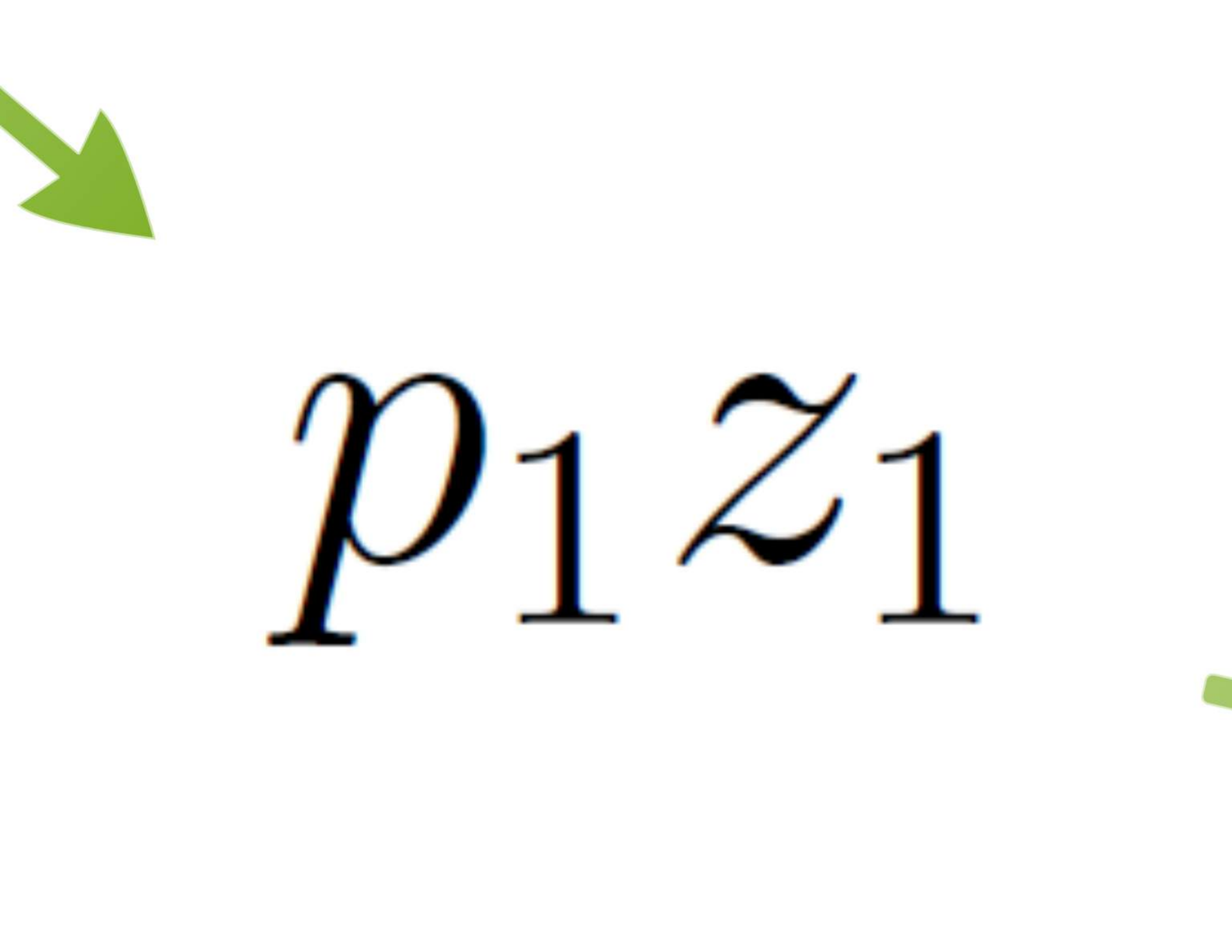
and that we converge to the same eigenvector, independent of the starting point.

to show the second point,
consider x, a positive vector

$$(MM^T)^k x = c_1^k p_1 z_1 + \cdots c_n^k p_n z_n$$

$$p_1 z_1$$

so no matter how


$p_1 z_1$

so no matter how you start, you converge to the direction of z_1

we need to show that p_1 is nonzero!

or, we need to show that no positive vector is orthogonal to z_1

$$(MM^T)^k x / c_1^k$$

$$p_1 z_1$$

hence no positive vector can be orthogonal to z_1

Web Search

hari sundaram

hs1@illinois.edu