Hubs and Authority

Link Analysis

There are two famous link analysis methods:

1.PageRank Algorithm
2.HITS Algorithm

Ranking

- Today's search engines may return millions of pages for a certain query
- It is not possible for a user to preview all the returned results
- So, ranking is helpful

Rankers

Rankers are classified into two groups :

1.Content-based rankers

–number of matched terms

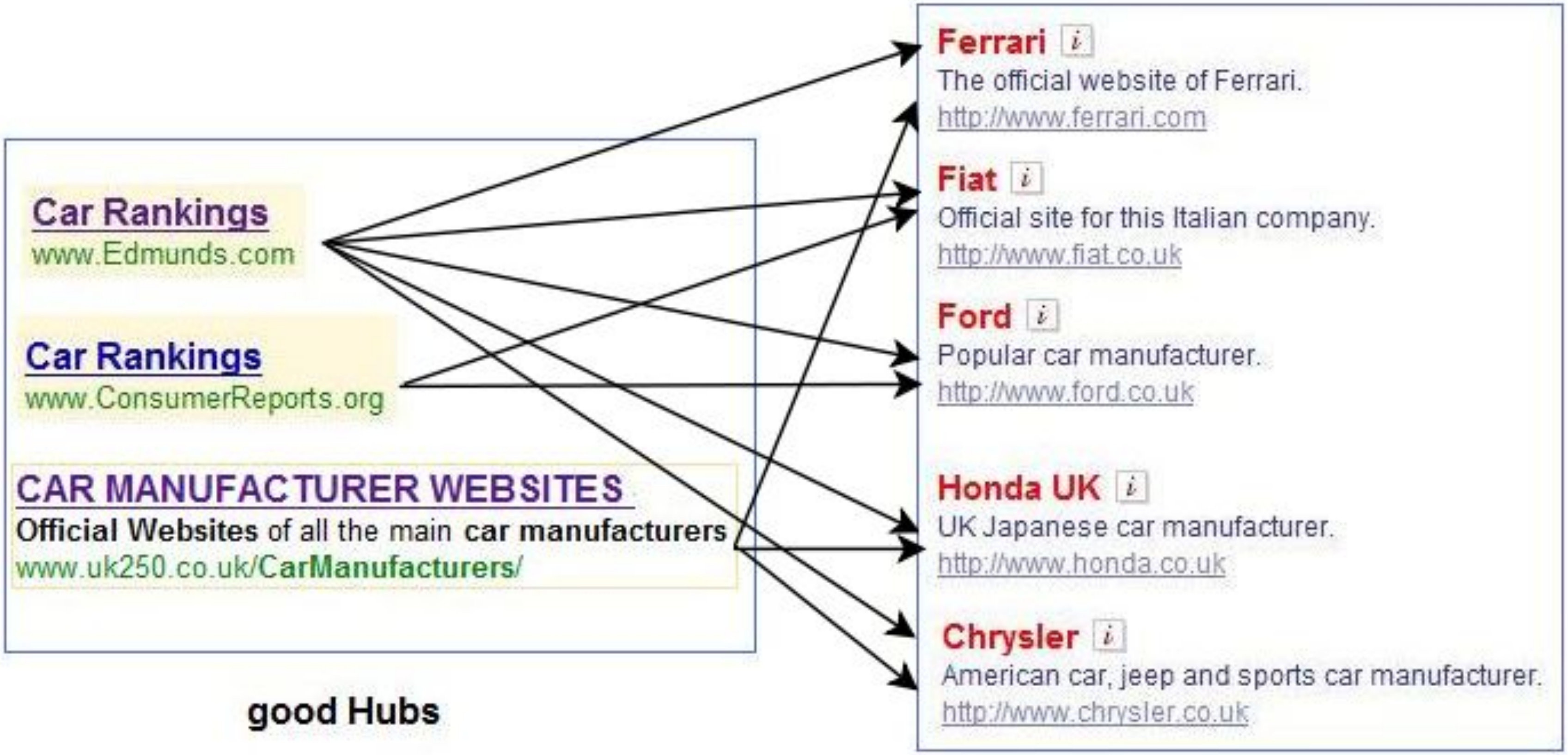
–frequency of terms

–location of terms

2.Connectivity-based rankers

–links that point to them

Hubs and Authority



Car Rankings
www.Edmunds.com

Car Rankings
www.ConsumerReports.org

CAR MANUFACTURER WEBSITES
Official Websites of all the main **car manufacturers**
www.uk250.co.uk/**CarManufacturers**/

Ferrari
The official website of Ferrari.
http://www.ferrari.com

Fiat
Official site for this Italian company.
http://www.fiat.co.uk

Ford
Popular car manufacturer.
http://www.ford.co.uk

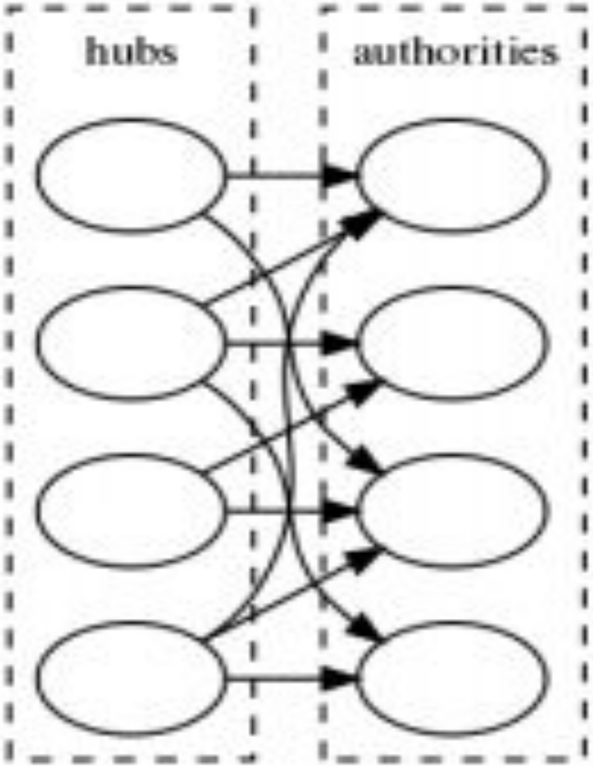
Honda UK
UK Japanese car manufacturer.
http://www.honda.co.uk

Chrysler
American car, jeep and sports car manufacturer.
http://www.chrysler.co.uk

good Hubs

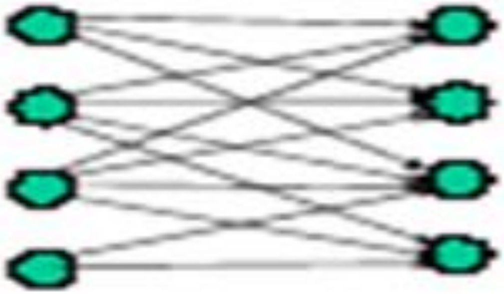
good Authorities

- Hypertext Induced Topics Search (HITS) developed by Jon Kleinberg
- HITS is applied on a subgraph after a search is done on the complete graph
- An authority is a page that many hubs link to
- A hub is a page that links to many authorities



HITS Algorithm

- Hubs point to lots of authorities.
- Authorities are pointed to by lots of hubs.
- Together they form a bipartite graph:
- Hubs Authorities



- In real life, when you buy a car, you are more inclined to purchase it from a certain dealer that your friend recommends.
- Following the analogy, the authority in this case would be the car dealer.
- And the hub would be your friend. You trust your friend, therefore you trust what your friend recommends.
- In the world wide web, hubs for our query about automobiles might be pages that contain rankings of the cars, blogs where people discuss about the cars that they purchased, and so on.
- www.bmw.com is a authority.

Authority

- Page *i* is called an ***authority*** for the query "automobile makers" if it contains valuable information on the subject. Official web sites of car manufacturers, such as www.bmw.com, HyundaiUSA.com, www.mercedes-benz.com would be authorities for this search. These are the ones truly relevant to the given query. These are the ones that the user expects back from the query engine.

Hub

- There is a second category of pages called ***hubs.***
- ***Hubs** help to find authority pages.*
- Their role is to advertise the authoritative pages.
- They contain useful links towards the authoritative pages.
- In other words, hubs point the search engine in the "right direction".

HITS vs PageRank

- Both HITS and PageRank correspond to matrix computations.
- Both can be unstable: changing a few links can lead to quite different rankings.
- PageRank doesn't handle pages with no outedges very well, because they decrease the PageRank overall

Step By Step HITS

- For each node initiliaze the ap and hp to 1/n
- In each iteration calculate the authority weight for each node in S

$$a_p = \sum_{q:q\rightarrow p} h_q$$

Step By Step HITS

- In each iteration calculate the hub weight for each node in S

$$h_p = \sum_{q:p \rightarrow q} a_q$$

- **Note:** The hub weights are computed from the current authority weights, which were computed from the previous hub weights.

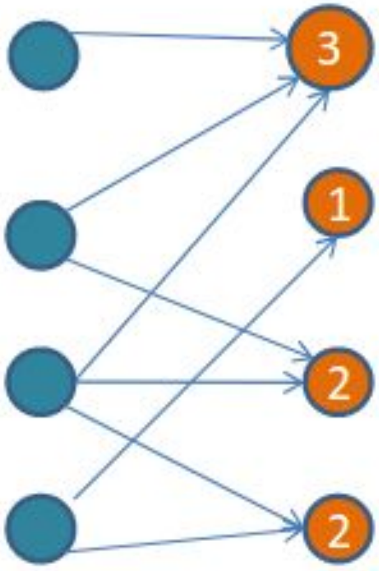
Step By Step HITS

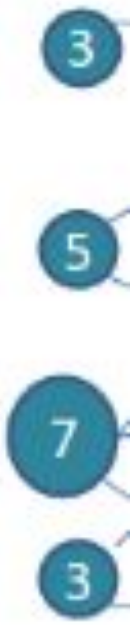
- After new weights are computed for all nodes, the weights are normalized:

$$\sum_{p \in S} (a_p)^2 = 1 \qquad \sum_{p \in S} (h_p)^2 = 1$$

The Iterative Algorithm no normalization

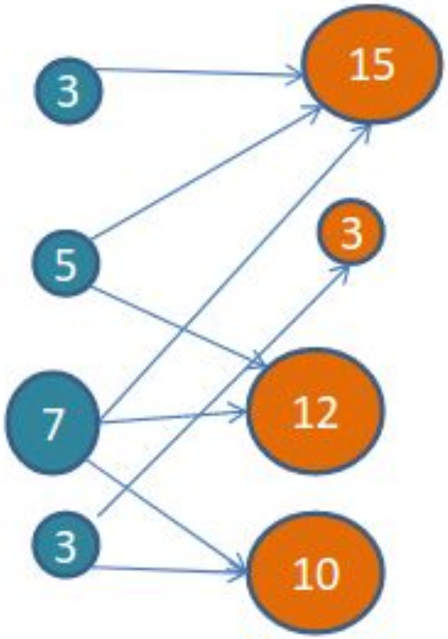
- 1st Iteration
- I Step





The Iterative Algorithm no normalization

- 2nd Iteration
- I Step



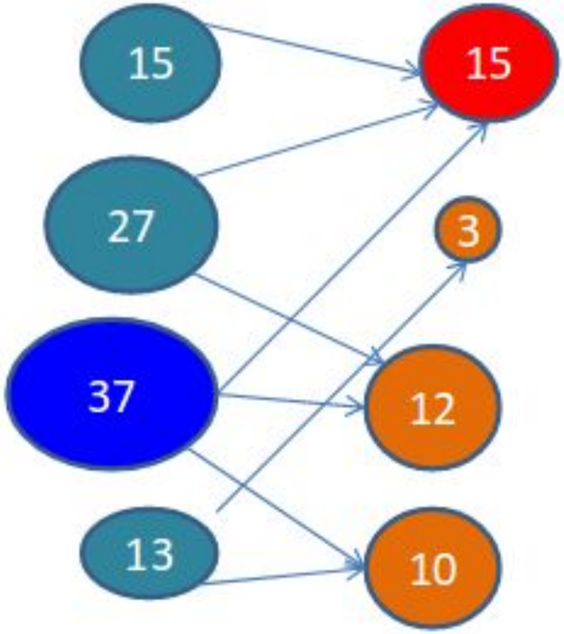
The Iterative Algorithm no normalization

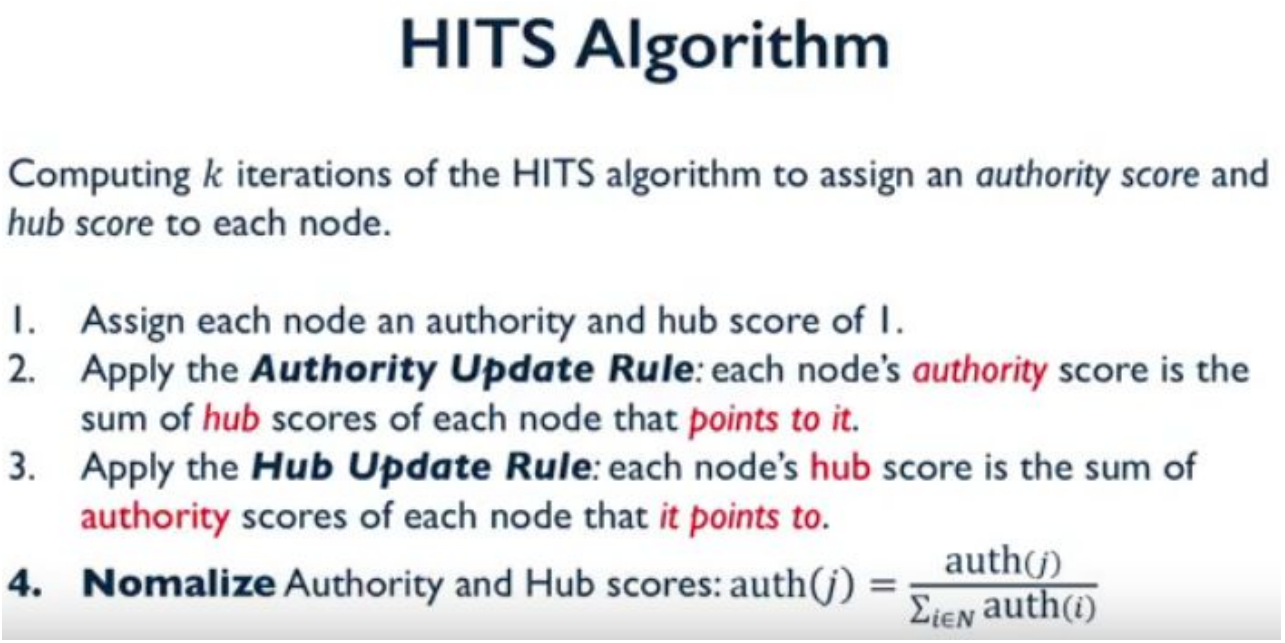
- 2nd Iteration
- I Step
- O Step

…

…

…



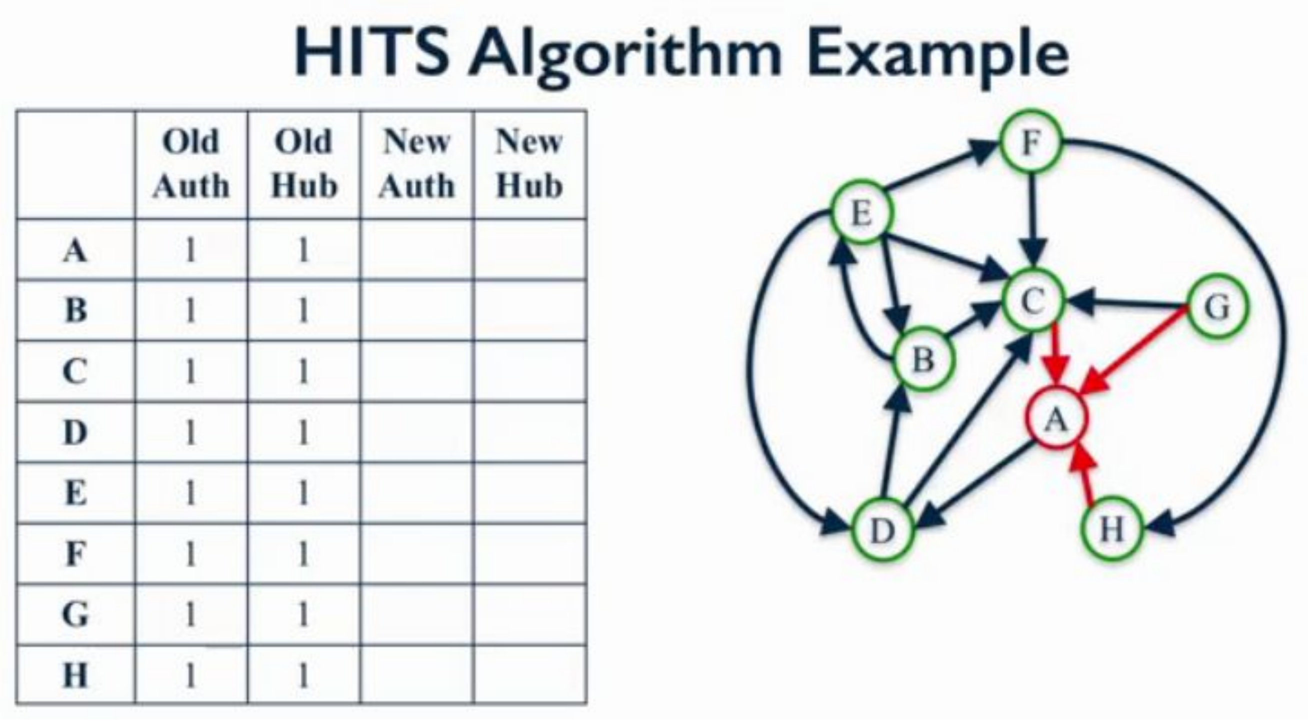


HITS Algorithm

Computing k iterations of the HITS algorithm to assign an *authority score* and *hub score* to each node.

1. Assign each node an authority and hub score of 1.
2. Apply the ***Authority Update Rule***: each node's *authority* score is the sum of *hub* scores of each node that *points to it*.
3. Apply the ***Hub Update Rule***: each node's hub score is the sum of authority scores of each node that *it points to*.
4. **Nomalize** Authority and Hub scores: $\text{auth}(j) = \frac{\text{auth}(j)}{\sum_{i \in N} \text{auth}(i)}$

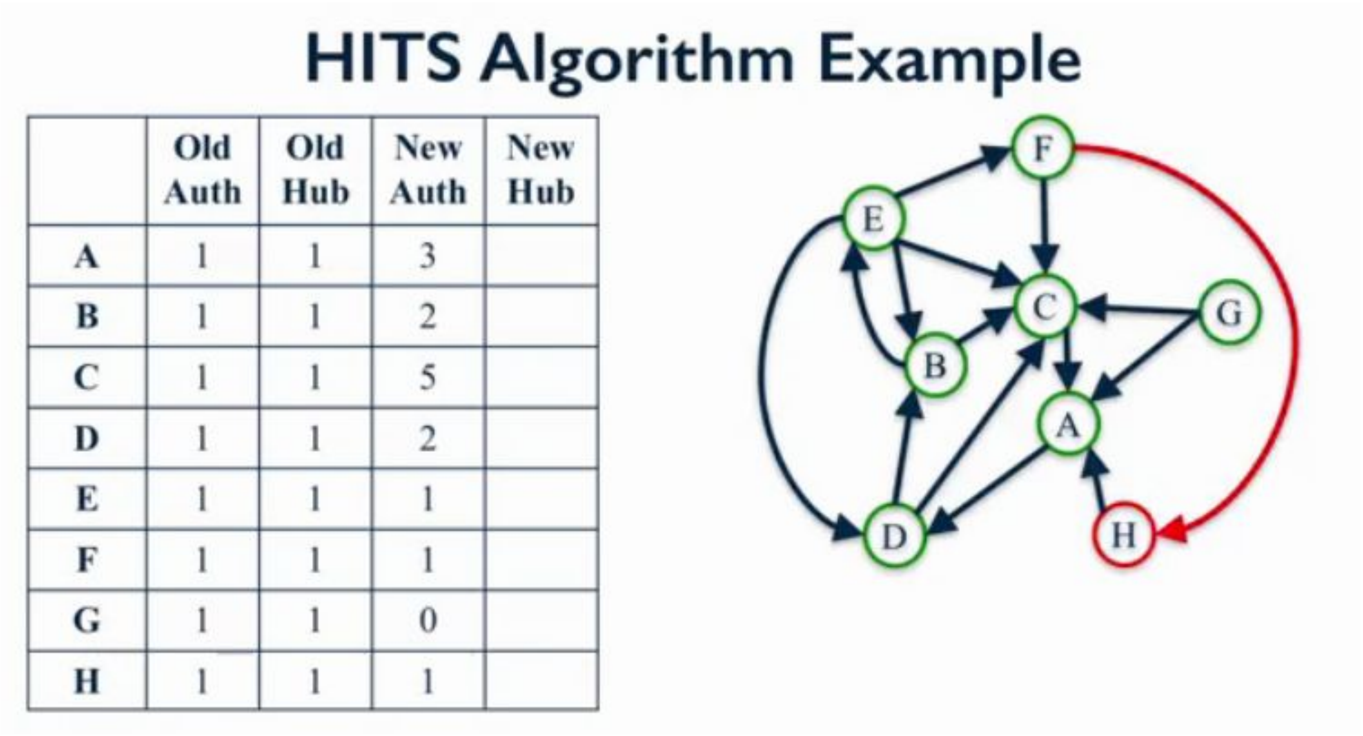
Authority Score - Consider inlinks (Addition of old hub score)
Hub Score - Consider outlinks (Addition of old auth score)



HITS Algorithm Example

	Old Auth	Old Hub	New Auth	New Hub
A	1	1		
B	1	1		
C	1	1		
D	1	1		
E	1	1		
F	1	1		
G	1	1		
H	1	1		

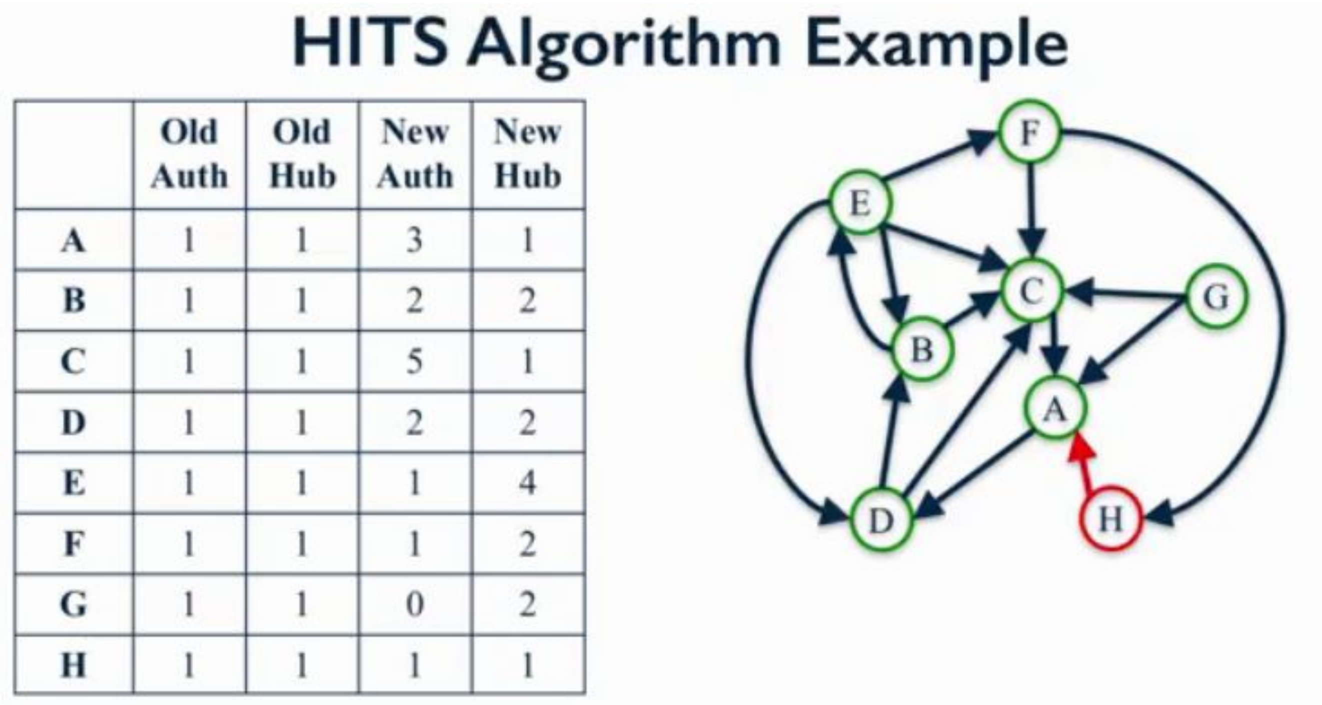
Authority Score - Consider inlinks (Addition of old hub score)
Hub Score - Consider outlinks (Addition of old auth score)



HITS Algorithm Example

	Old Auth	Old Hub	New Auth	New Hub
A	1	1	3	
B	1	1	2	
C	1	1	5	
D	1	1	2	
E	1	1	1	
F	1	1	1	
G	1	1	0	
H	1	1	1	

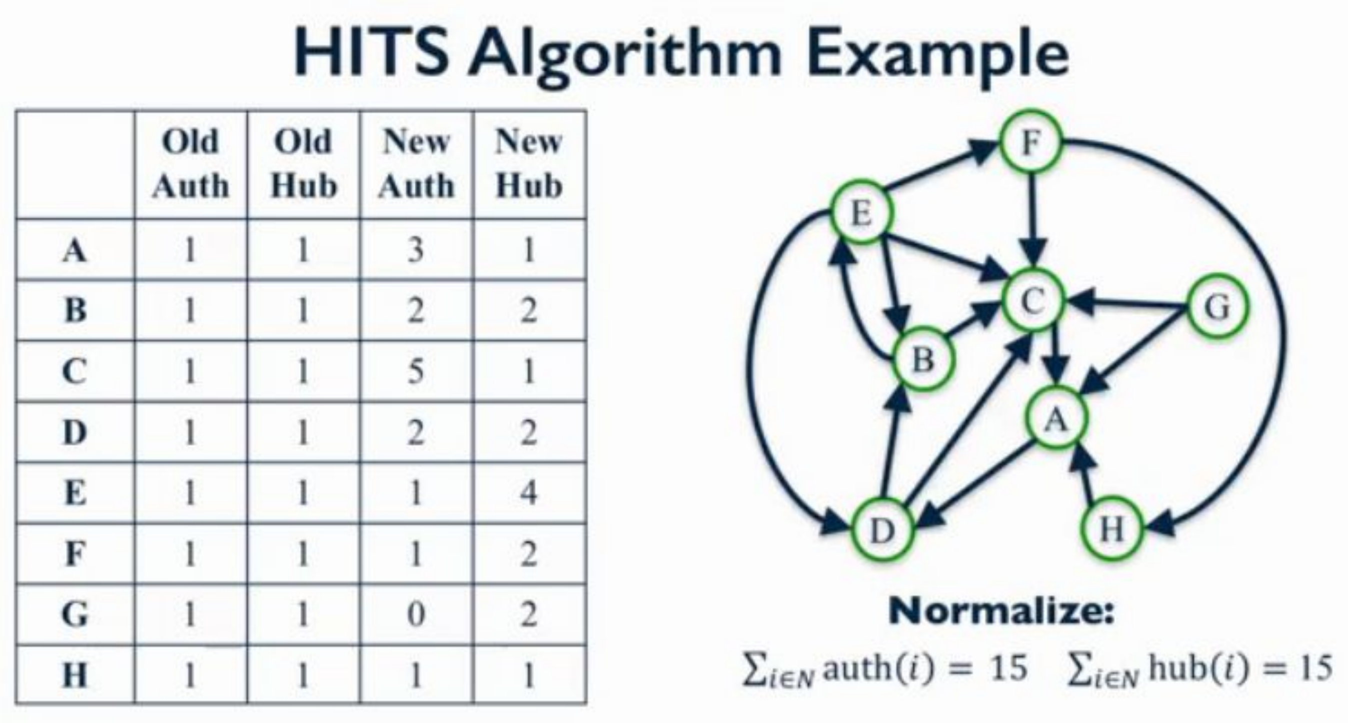
Authority Score - Consider inlinks (Addition of old hub score)
Hub Score - Consider outlinks (Addition of old auth score)



HITS Algorithm Example

	Old Auth	Old Hub	New Auth	New Hub
A	1	1	3	1
B	1	1	2	2
C	1	1	5	1
D	1	1	2	2
E	1	1	1	4
F	1	1	1	2
G	1	1	0	2
H	1	1	1	1

Authority Score - Consider inlinks (Addition of old hub score)
Hub Score - Consider outlinks (Addition of old auth score)



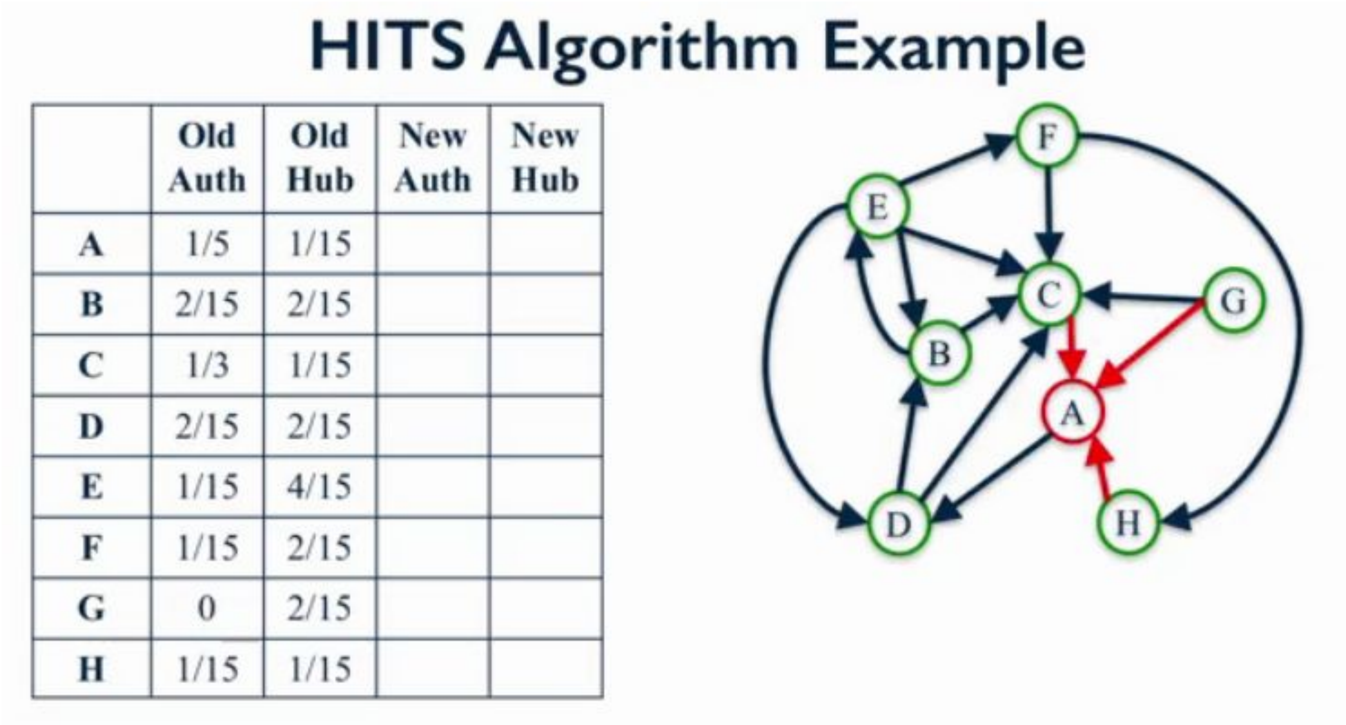
HITS Algorithm Example

	Old Auth	Old Hub	New Auth	New Hub
A	1	1	3	1
B	1	1	2	2
C	1	1	5	1
D	1	1	2	2
E	1	1	1	4
F	1	1	1	2
G	1	1	0	2
H	1	1	1	1

Normalize:

$\sum_{i \in N} \text{auth}(i) = 15 \quad \sum_{i \in N} \text{hub}(i) = 15$

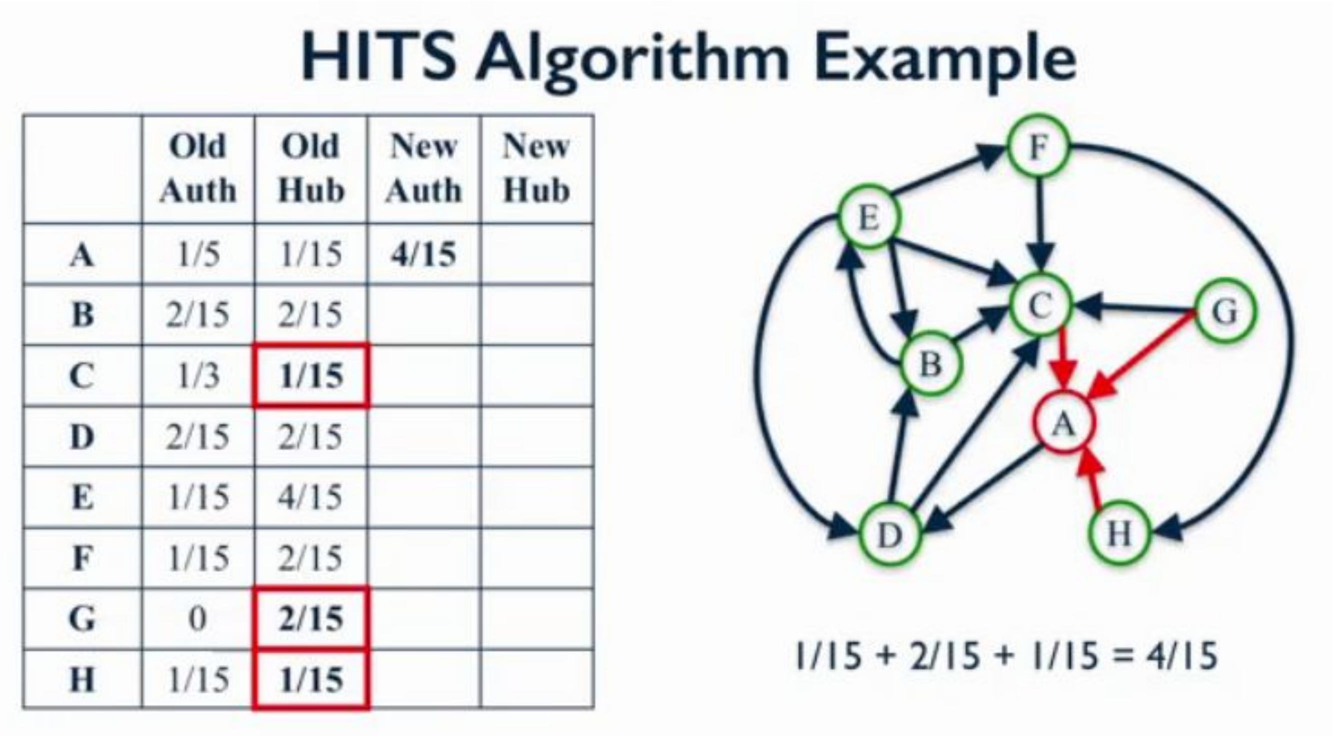
Authority Score - Consider inlinks (Addition of old hub score)
Hub Score - Consider outlinks (Addition of old auth score)



HITS Algorithm Example

	Old Auth	Old Hub	New Auth	New Hub
A	1/5	1/15		
B	2/15	2/15		
C	1/3	1/15		
D	2/15	2/15		
E	1/15	4/15		
F	1/15	2/15		
G	0	2/15		
H	1/15	1/15		

Authority Score - Consider inlinks (Addition of old hub score)
Hub Score - Consider outlinks (Addition of old auth score)

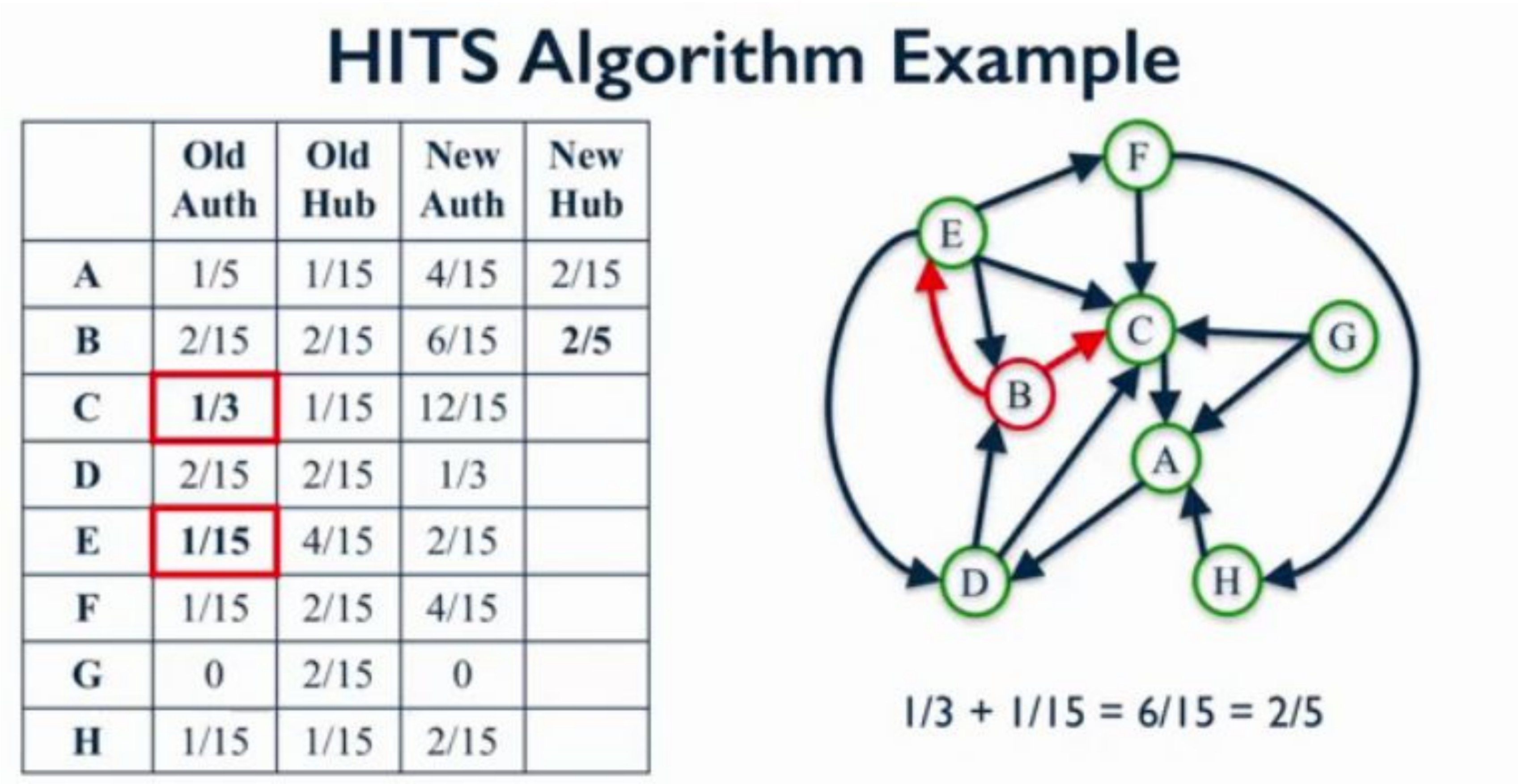


HITS Algorithm Example

	Old Auth	Old Hub	New Auth	New Hub
A	1/5	1/15	**4/15**	
B	2/15	2/15		
C	1/3	**1/15**		
D	2/15	2/15		
E	1/15	4/15		
F	1/15	2/15		
G	0	**2/15**		
H	1/15	**1/15**		

1/15 + 2/15 + 1/15 = 4/15

Authority Score - Consider inlinks (Addition of old hub score)
Hub Score - Consider outlinks (Addition of old auth score)

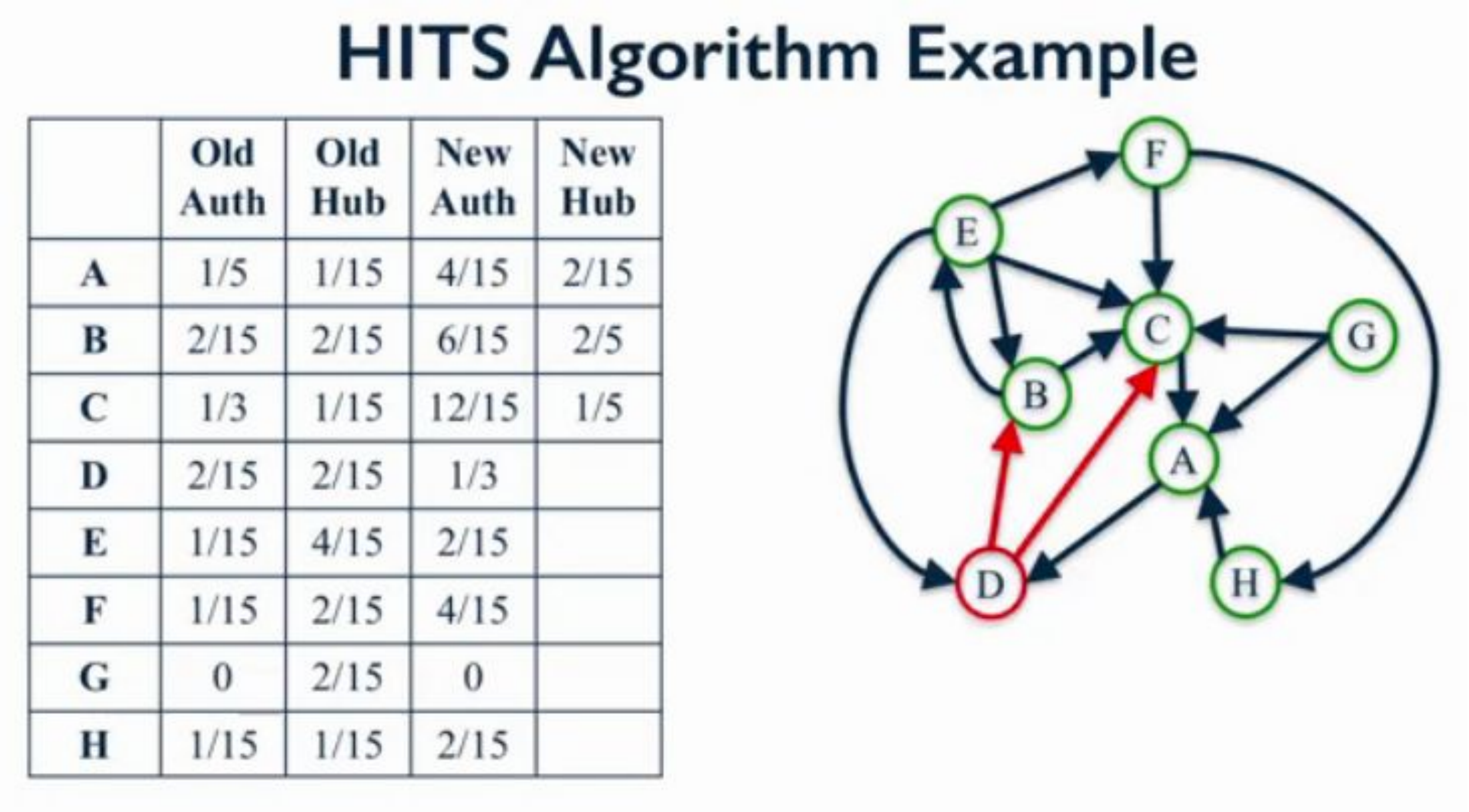


HITS Algorithm Example

	Old Auth	Old Hub	New Auth	New Hub
A	1/5	1/15	4/15	2/15
B	2/15	2/15	6/15	**2/5**
C	**1/3**	1/15	12/15	
D	2/15	2/15	1/3	
E	**1/15**	4/15	2/15	
F	1/15	2/15	4/15	
G	0	2/15	0	
H	1/15	1/15	2/15	

1/3 + 1/15 = 6/15 = 2/5

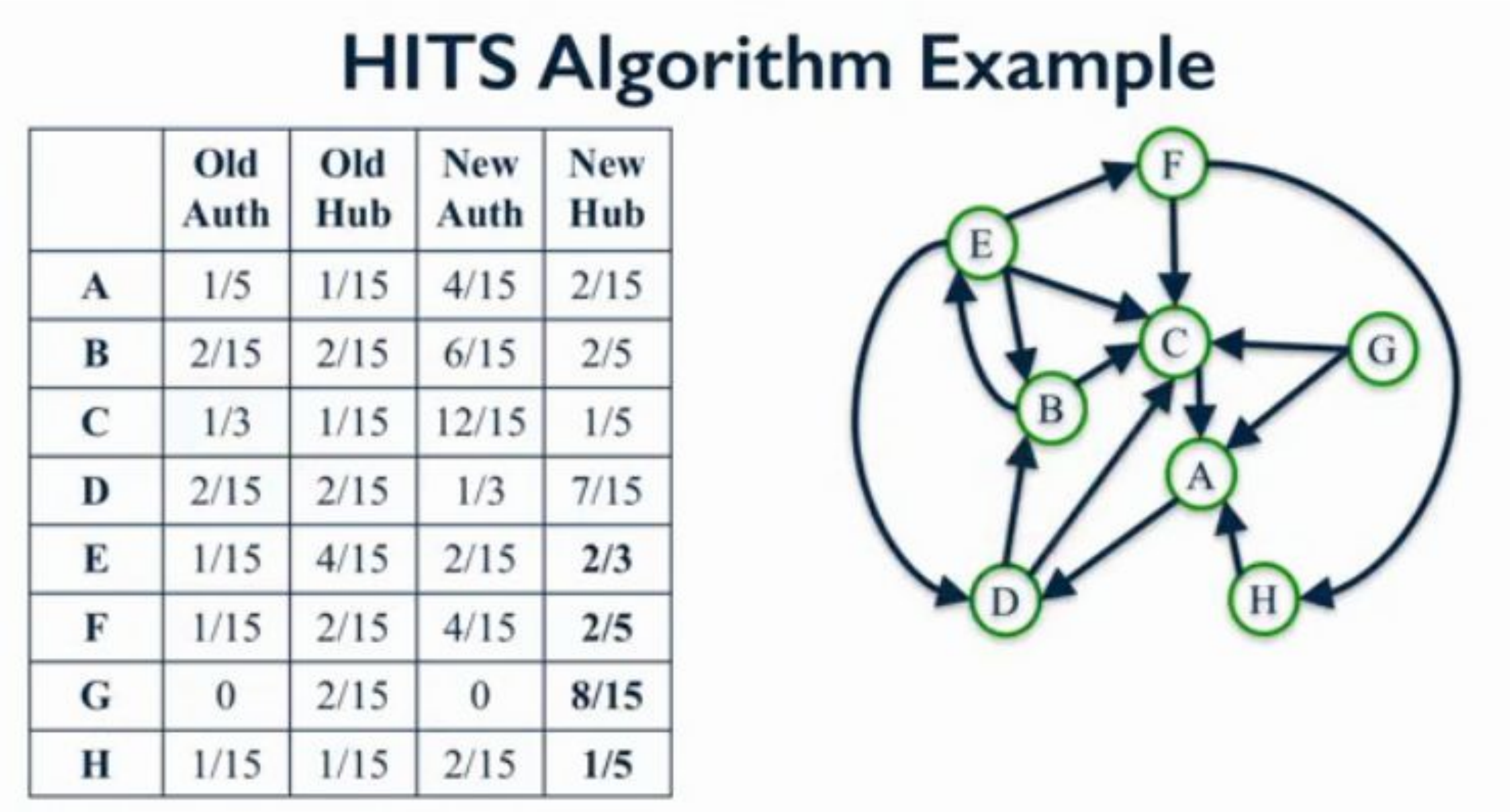
Authority Score - Consider inlinks (Addition of old hub score)
Hub Score - Consider outlinks (Addition of old auth score)



HITS Algorithm Example

	Old Auth	Old Hub	New Auth	New Hub
A	1/5	1/15	4/15	2/15
B	2/15	2/15	6/15	2/5
C	1/3	1/15	12/15	1/5
D	2/15	2/15	1/3	
E	1/15	4/15	2/15	
F	1/15	2/15	4/15	
G	0	2/15	0	
H	1/15	1/15	2/15	

Authority Score - Consider inlinks (Addition of old hub score)
Hub Score - Consider outlinks (Addition of old auth score)

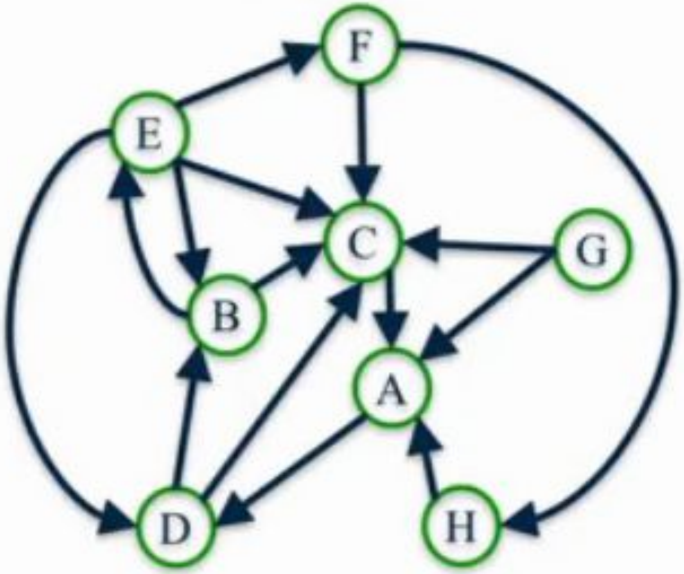


HITS Algorithm Example

	Old Auth	Old Hub	New Auth	New Hub
A	1/5	1/15	4/15	2/15
B	2/15	2/15	6/15	2/5
C	1/3	1/15	12/15	1/5
D	2/15	2/15	1/3	7/15
E	1/15	4/15	2/15	**2/3**
F	1/15	2/15	4/15	**2/5**
G	0	2/15	0	**8/15**
H	1/15	1/15	2/15	**1/5**

HITS Algorithm Example

	Old Auth	Old Hub	New Auth	New Hub
A	1/5	1/15	4/15	2/15
B	2/15	2/15	6/15	2/5
C	1/3	1/15	12/15	1/5
D	2/15	2/15	1/3	7/15
E	1/15	4/15	2/15	2/3
F	1/15	2/15	4/15	2/5
G	0	2/15	0	8/15
H	1/15	1/15	2/15	1/5

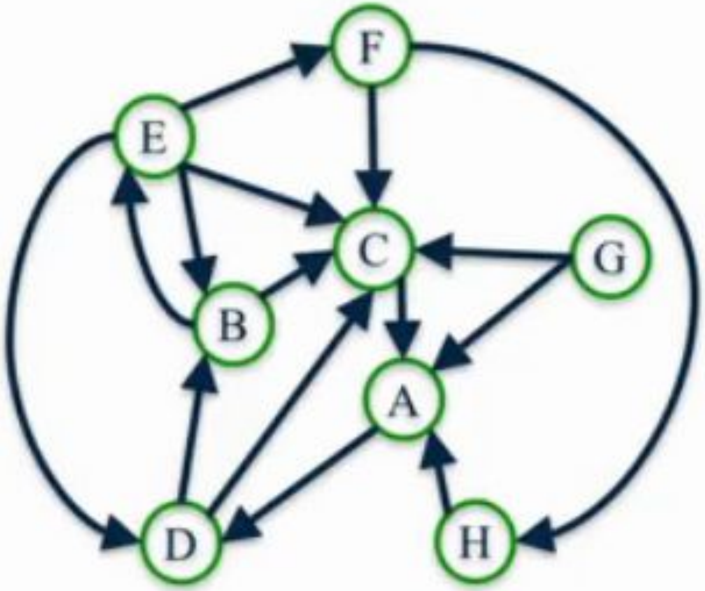


Normalize:

$$\sum_{i \in N} \text{auth}(i) = {}^{35}/_{15}$$

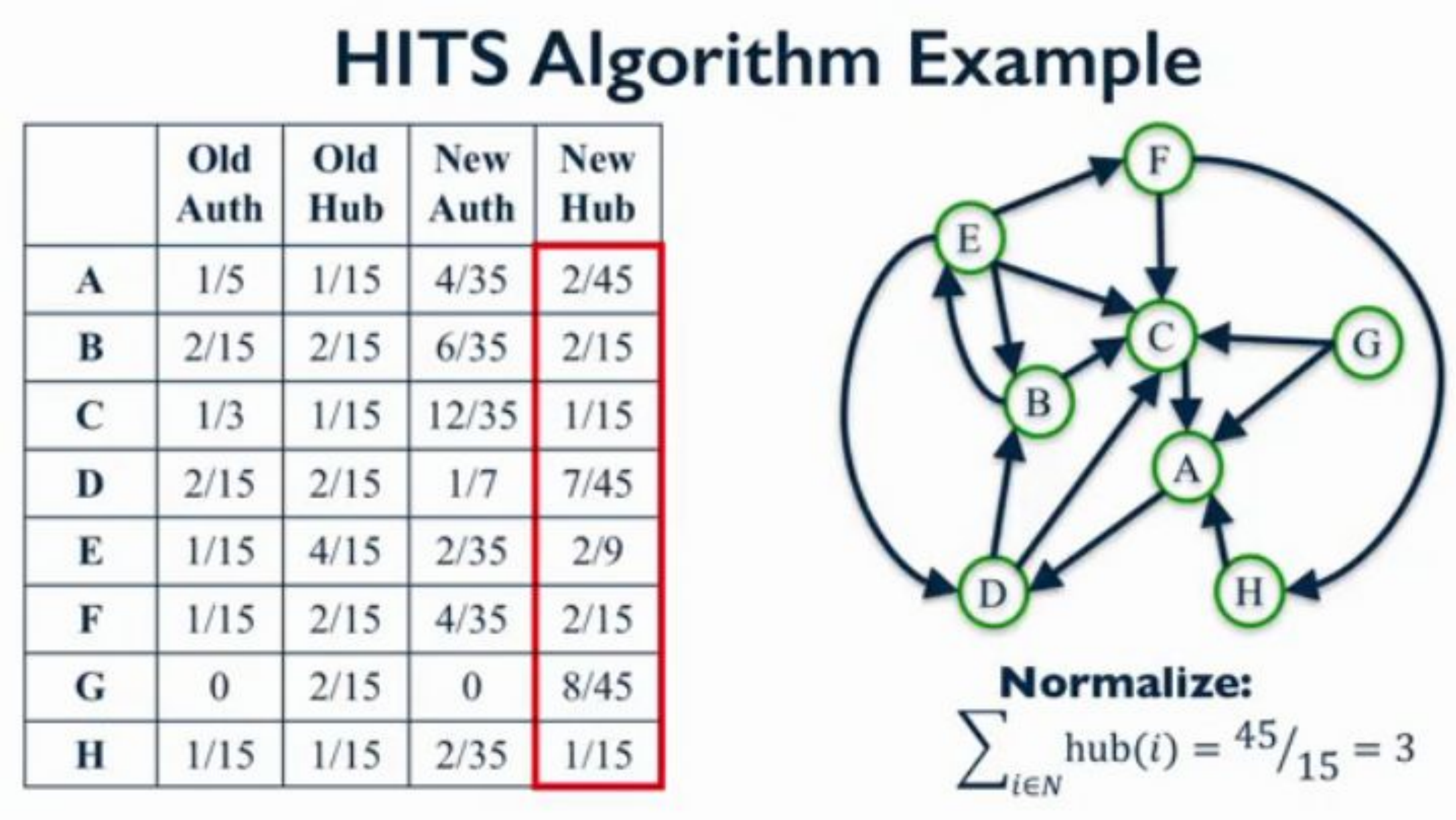
HITS Algorithm Example

	Old Auth	Old Hub	New Auth	New Hub
A	1/5	1/15	4/35	2/15
B	2/15	2/15	6/35	2/5
C	1/3	1/15	12/35	1/5
D	2/15	2/15	1/7	7/15
E	1/15	4/15	2/35	2/3
F	1/15	2/15	4/35	2/5
G	0	2/15	0	8/15
H	1/15	1/15	2/35	1/5



Normalize:

$$\sum_{i\in N} \text{hub}(i) = {}^{45}/_{15} = 3$$



HITS Algorithm Example

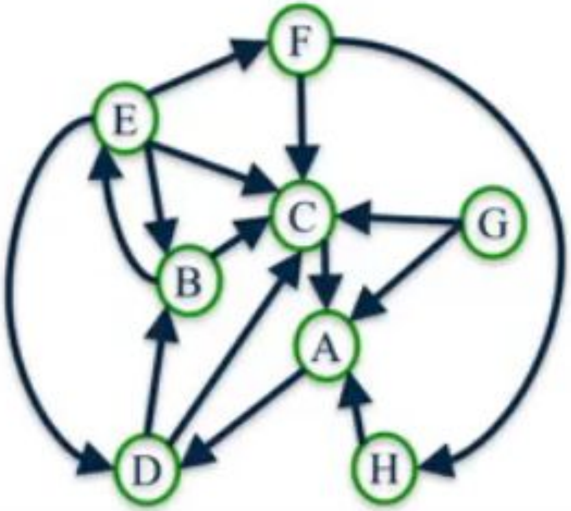
	Old Auth	Old Hub	New Auth	New Hub
A	1/5	1/15	4/35	2/45
B	2/15	2/15	6/35	2/15
C	1/3	1/15	12/35	1/15
D	2/15	2/15	1/7	7/45
E	1/15	4/15	2/35	2/9
F	1/15	2/15	4/35	2/15
G	0	2/15	0	8/45
H	1/15	1/15	2/35	1/15

Normalize:

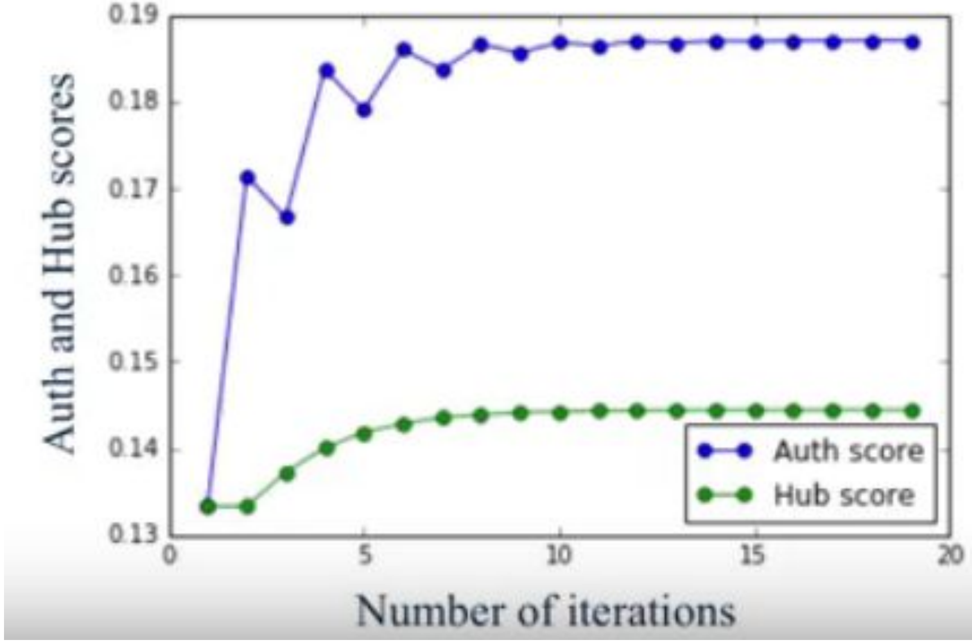
$$\sum_{i \in N} \text{hub}(i) = {}^{45}/_{15} = 3$$

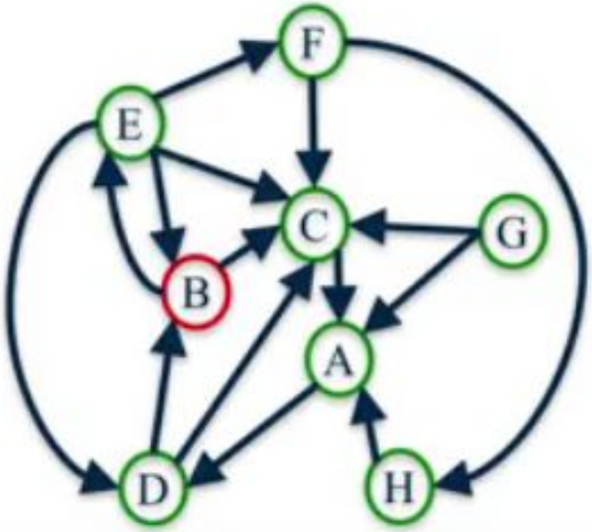
HITS Algorithm Convergence

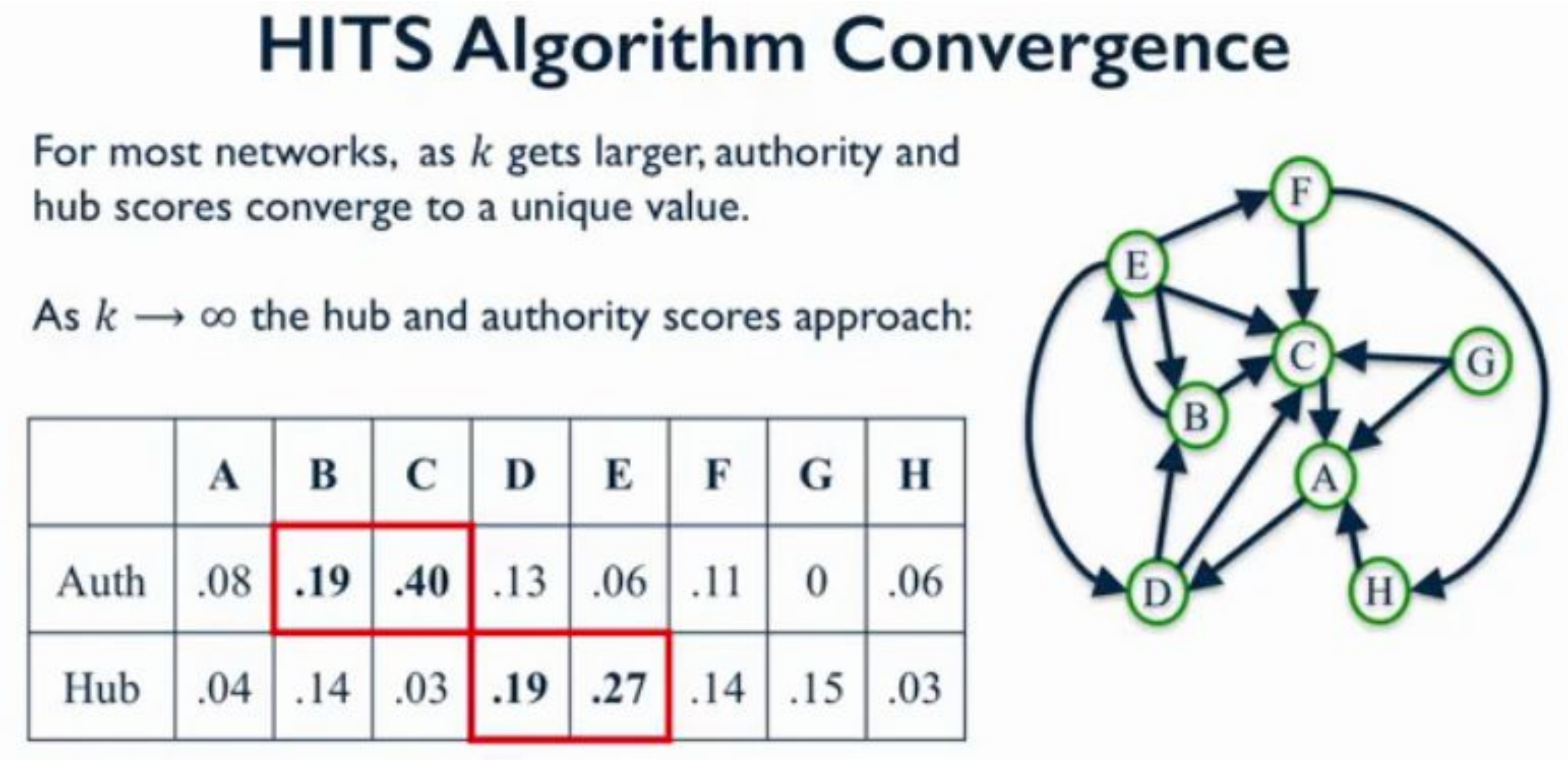
	k	A	B	C	D	E	F	G	H
Auth	2	.11	.17	.34	.14	.06	.11	0	.06
	4	.10	.18	.36	.13	.06	.11	0	.06
	6	.09	.19	.37	.13	.06	.11	0	.06
Hub	2	.04	.13	.07	.16	.22	.13	.18	.07
	4	.04	.14	.05	.18	.25	.14	.17	.04
	6	.04	.14	.04	.18	.26	.14	.16	.04



HITS Algorithm Convergence







HITS Algorithm Convergence

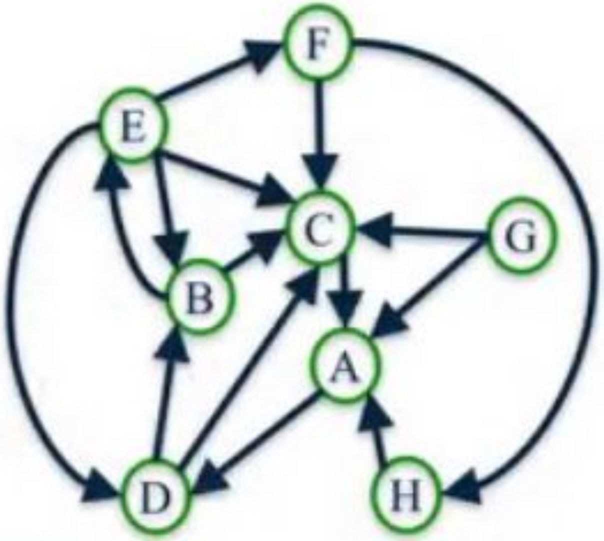
For most networks, as k gets larger, authority and hub scores converge to a unique value.

As $k \longrightarrow \infty$ the hub and authority scores approach:

	A	B	C	D	E	F	G	H
Auth	.08	**.19**	**.40**	.13	.06	.11	0	.06
Hub	.04	.14	.03	**.19**	**.27**	.14	.15	.03

Summary

- The HITS algorithm starts by constructing a *root set* of relevant web pages and expanding it to a *base set.*
- HITS then assigns an authority and hub score to each node in the network.
- Nodes that have incoming edges from *good hubs* are *good authorities*, and nodes that have outgoing edges to *good authorities* are *good hubs*.



- Authority and hub scores converge for most networks.