HITS ALGORITHM

HUBS AND AUTHORITIES ON THE INTERNET

META-SEARCH ENGINES

- Search engine that passes query to several other search engines and integrate results.
  - Submit queries to host sites.
  - Parse resulting HTML pages to extract search results.
  - Integrate multiple rankings into a "consensus" ranking.
  - Present integrated results to user.
- Examples:
  - Metacrawler
  - SavvySearch
  - Dogpile

HTML STRUCTURE & FEATURE WEIGHTING

- Weight tokens under particular HTML tags more heavily:
  - <TITLE> tokens (Google seems to like title matches)
  - <H1>,<H2>… tokens
  - <META> keyword tokens
- Parse page into conceptual sections (e.g. navigation links vs. page content) and weight tokens differently based on section.

BIBLIOMETRICS: CITATION ANALYSIS

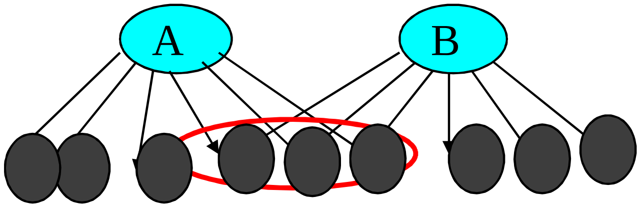
- Many standard documents include *bibliographies* (or *references*), explicit *citations* to other previously published documents.
- Using citations as links, standard corpora can be viewed as a graph.
- The structure of this graph, independent of content, can provide interesting information about the similarity of documents and the structure of information.
- CF corpus includes citation information.

IMPACT FACTOR

- Developed by Garfield in 1972 to measure the importance (quality, influence) of scientific journals.
- Measure of how often papers in the journal are cited by other scientists.
- Computed and published annually by the Institute for Scientific Information (ISI).
- The *impact factor* of a journal *J* in year *Y* is the average number of citations (from indexed documents published in year *Y*) to a paper published in *J* in year *Y*–1 or *Y*–2.
- Does not account for the quality of the citing article.

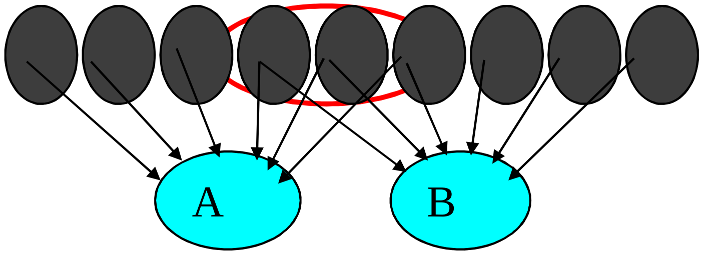
BIBLIOGRAPHIC COUPLING

- Measure of similarity of documents introduced by Kessler in 1963.
- The bibliographic coupling of two documents *A* and *B* is the number of documents cited by *both A* and *B.*
- Size of the intersection of their bibliographies.
- Maybe want to normalize by size of bibliographies?



CO-CITATION

- An alternate citation-based measure of similarity introduced by Small in 1973.
- Number of documents that cite both *A* and *B.*
- Maybe want to normalize by total number of documents citing either *A* or *B* ?



CITATIONS VS. LINKS

- Web links are a bit different than citations:
  - Many links are navigational.
  - Many pages with high in-degree are portals not content providers.
  - Not all links are endorsements.
  - Company websites don’t point to their competitors.
  - Citations to relevant literature is enforced by peer-review.

HITS

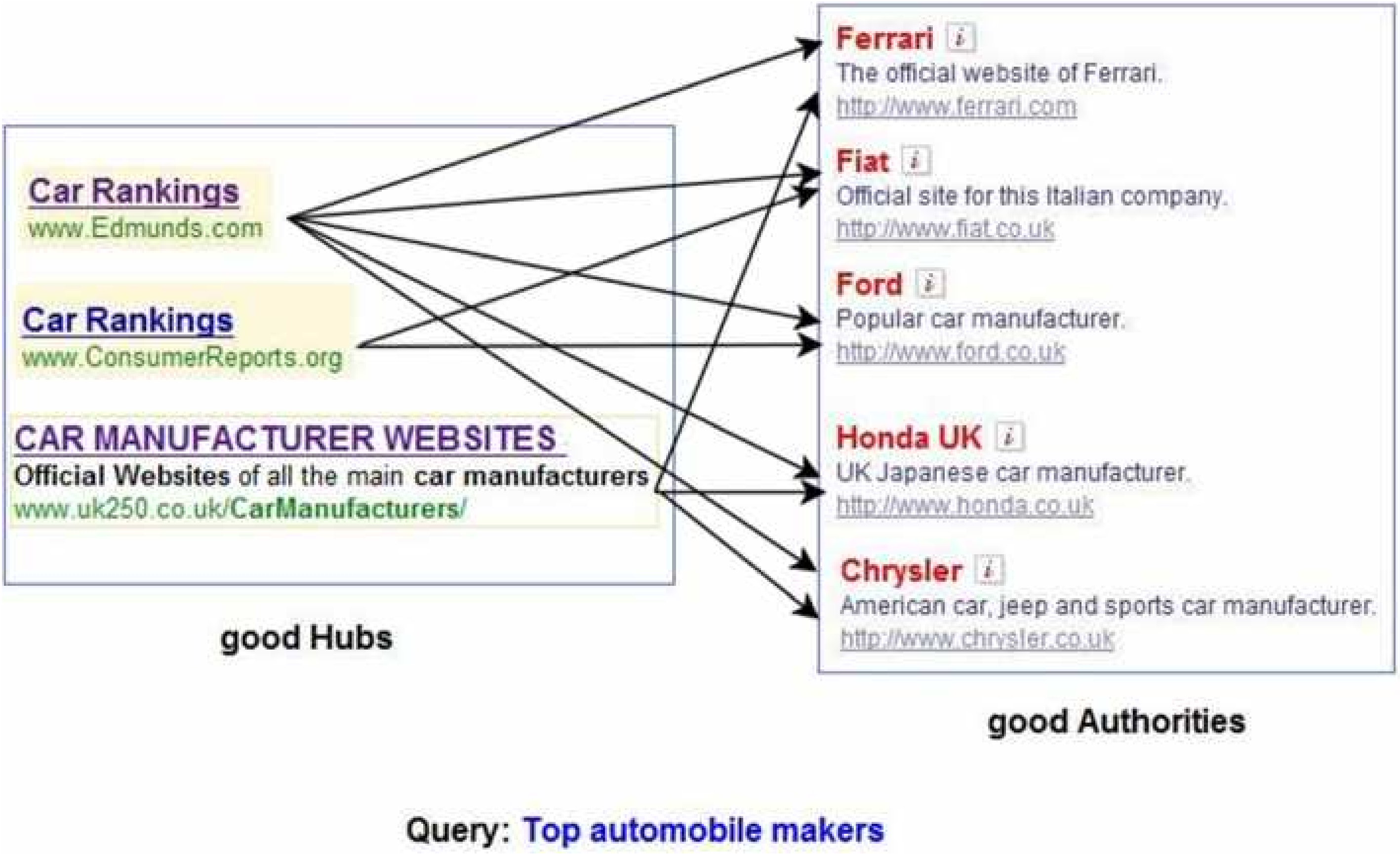
- In the same time that PageRank was being developed, Jon Kleinberg a professor in the Department of Computer Science at Cornell came up with his own solution to the Web Search problem.
- He developed an algorithm that made use of the link structure of the web in order to discover and rank pages relevant for a particular topic. **HITS *(hyperlink-induced topic search)*** is now part of the **Ask** search engine (www.Ask.com).

AUTHORITIES

- *Authorities* are pages that are recognized as providing significant, trustworthy, and useful information on a topic.
- *In-degree* (number of pointers to a page) is one simple measure of authority.
- However in-degree treats all links as equal.
- Should links from pages that are themselves authoritative count more?
- 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. Commercial web sites selling cars might be authorities on the subject as well. These are the ones truly relevant to the given query. These are the ones that the user expects back from the query engine.

HUBS

- *Hubs* are index pages that provide lots of useful links to relevant content pages (topic authorities).
- 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".
- 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.



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**/

good Hubs

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 Authorities

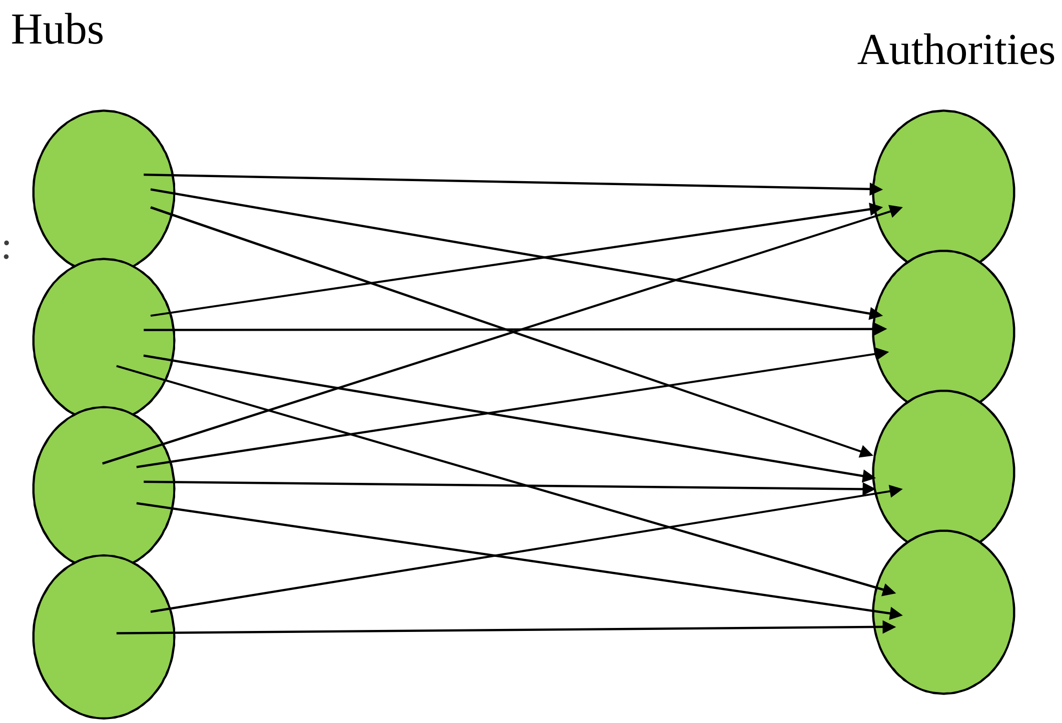
Query: Top automobile makers

HITS

- Algorithm developed by Kleinberg in 1998.
- Attempts to computationally determine hubs and authorities on a particular topic through analysis of a relevant subgraph of the web.
- Based on mutually recursive facts:
  - Hubs point to lots of authorities.
  - Authorities are pointed to by lots of hubs.

HUBS AND AUTHORITIES

- Together they tend to form a bipartite graph:

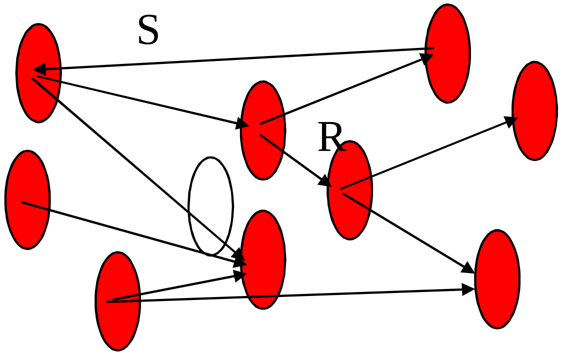


HITS ALGORITHM

- Computes hubs and authorities for a particular topic specified by a normal query.
- First determines a set of relevant pages for the query called the *base* set *S*.
- Analyze the link structure of the web subgraph defined by *S* to find authority and hub pages in this set.

CONSTRUCTING A BASE SUBGRAPH

- For a specific query *Q*, let the set of documents returned by a standard search engine (e.g. VSR) be called the *root* set *R*.
- Initialize *S* to *R*.
- Add to *S* all pages pointed to by any page in *R*.
- Add to *S* all pages that point to any page in *R*.



BASE LIMITATIONS

- To limit computational expense:
  - Limit number of root pages to the top 200 pages retrieved for the query.
  - Limit number of "back-pointer" pages to a random set of at most 50 pages returned by a "reverse link" query.
- To eliminate purely navigational links:
  - Eliminate links between two pages on the same host.
- To eliminate "non-authority-conveying" links:
  - Allow only m (m] 4–8) pages from a given host as pointers to any individual page.

AUTHORITIES AND IN-DEGREE

- Even within the base set *S* for a given query, the nodes with highest in-degree are not necessarily authorities (may just be generally popular pages like Yahoo or Amazon).
- True authority pages are pointed to by a number of hubs (i.e. pages that point to lots of authorities).

ITERATIVE ALGORITHM

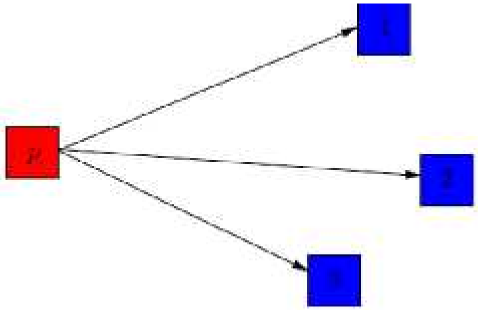
- Use an iterative algorithm to slowly converge on a mutually reinforcing set of hubs and authorities.
- Maintain for each page $p \in S$:
  - Authority score: a_p (vector ***a***)
  - Hub score: h_p (vector ***h***)
- Initialize all $a_p = h_p = 1$
- Maintain normalized scores:

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

HITS UPDATE RULES

- Authorities are pointed to by lots of good hubs:

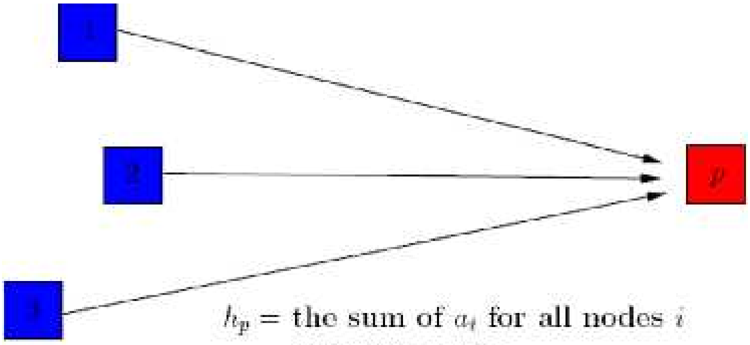
$$a_p = \sum_{q:q \to p} h_q$$



a_p = the sum of h_i for all nodes i pointing to p

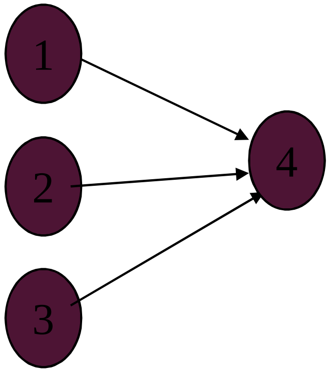
- Hubs point to lots of good authorities:

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



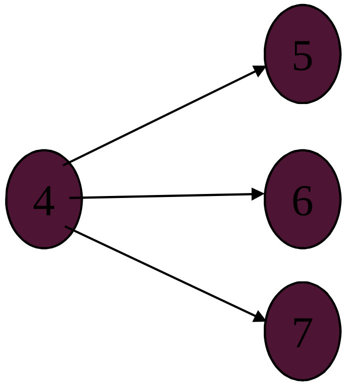
h_p = the sum of a_i for all nodes i pointed to by p

ILLUSTRATED UPDATE RULES



$a_4 = h_1 + h_2 + h_3$

$h_4 = a_5 + a_6 + a_7$



HITS ITERATIVE ALGORITHM

```
Initialize for all $p \in S$: $a_p = h_p = 1$
For i = 1 to k:
    For all $p \in S$:            (update auth. scores)

    For all $p \in S$:            (update hub scores)

    For all $p \in S$: $a_p = a_p/c$    $c$:

    For all $p \in S$: $h_p = h_p/c$    $c$:
```

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

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

$$\sum_{p\in S} (a_p / c)^2 = 1 \quad (\textit{normalize } \mathbf{a})$$

$$\sum_{p\in S} (h_p / c)^2 = 1 \quad (\textit{normalize } \mathbf{h})$$

CONVERGENCE

- Algorithm converges to a *fix-point* if iterated indefinitely.
- Define A to be the adjacency matrix for the subgraph defined by S.
  - $A_{ij} = 1$ for $i \in S, j \in S$ iff $i \rightarrow j$
- Authority vector, ***a***, converges to the principal eigenvector of $A^T A$
- Hub vector, ***h***, converges to the principal eigenvector of AA^T
- In practice, 20 iterations produces fairly stable results.

RESULTS

- Authorities for query: “Java”
  - java.sun.com
  - comp.lang.java FAQ
- Authorities for query “search engine”
  - Yahoo.com
  - Excite.com
  - Lycos.com
  - Altavista.com
- Authorities for query “Gates”
  - Microsoft.com
  - roadahead.com

RESULT COMMENTS

- In most cases, the final authorities were not in the initial root set generated using Altavista.
- Authorities were brought in from linked and reverse-linked pages and then HITS computed their high authority score.

FINDING SIMILAR PAGES USING LINK STRUCTURE

- Given a page, *P*, let *R* (the root set) be *t* (e.g. 200) pages that point to *P*.
- Grow a base set *S* from *R*.
- Run HITS on *S*.
- Return the best authorities in *S* as the best similar-pages for *P*.
- Finds authorities in the "link neighbor-hood" of *P*.

SIMILAR PAGE RESULTS

- Given “honda.com”
  - toyota.com
  - ford.com
  - bmwusa.com
  - saturncars.com
  - nissanmotors.com
  - audi.com
  - volvocars.com

HITS FOR CLUSTERING

- An ambiguous query can result in the principal eigenvector only covering one of the possible meanings.
- Non-principal eigenvectors may contain hubs & authorities for other meanings.
- Example:“jaguar”:
  - Atari video game (principal eigenvector)
  - NFL Football team (2nd non-princ. eigenvector)
  - Automobile (3rd non-princ. eigenvector)

PAGERANK

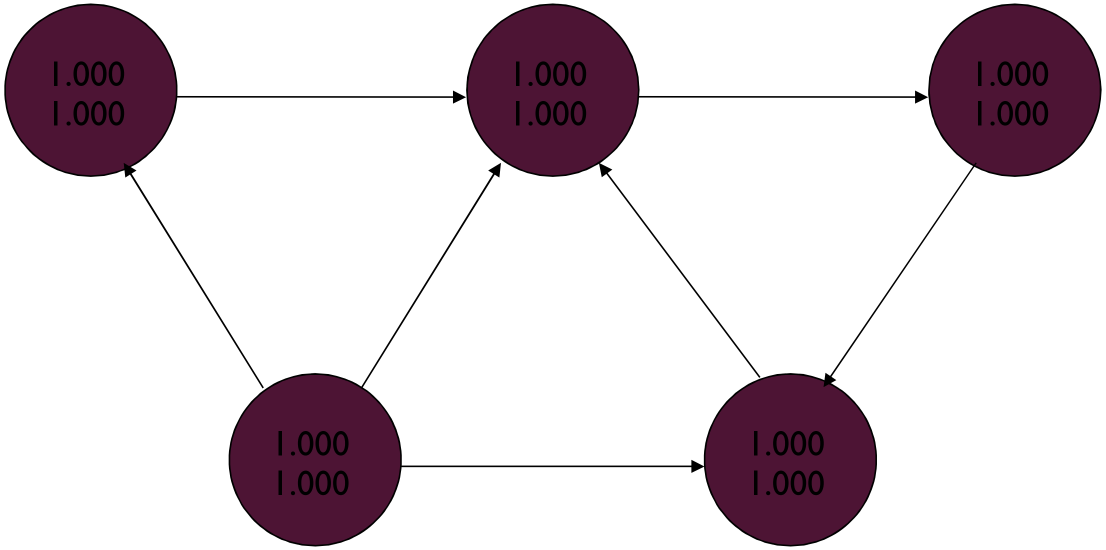
- Alternative link-analysis method used by Google (Brin & Page, 1998).
- Does not attempt to capture the distinction between hubs and authorities.
- Ranks pages just by authority.
- Applied to the entire web rather than a local neighborhood of pages surrounding the results of a query.

HITS EXAMPLE

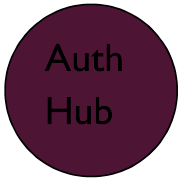
HTTP://WWW.MATH.CORNELL.EDU/~MEC/WINTER2009/RALUCAREMUS/LECTURE4/LECTURE4.HTML

A Simple Example

Update Authority Scores first

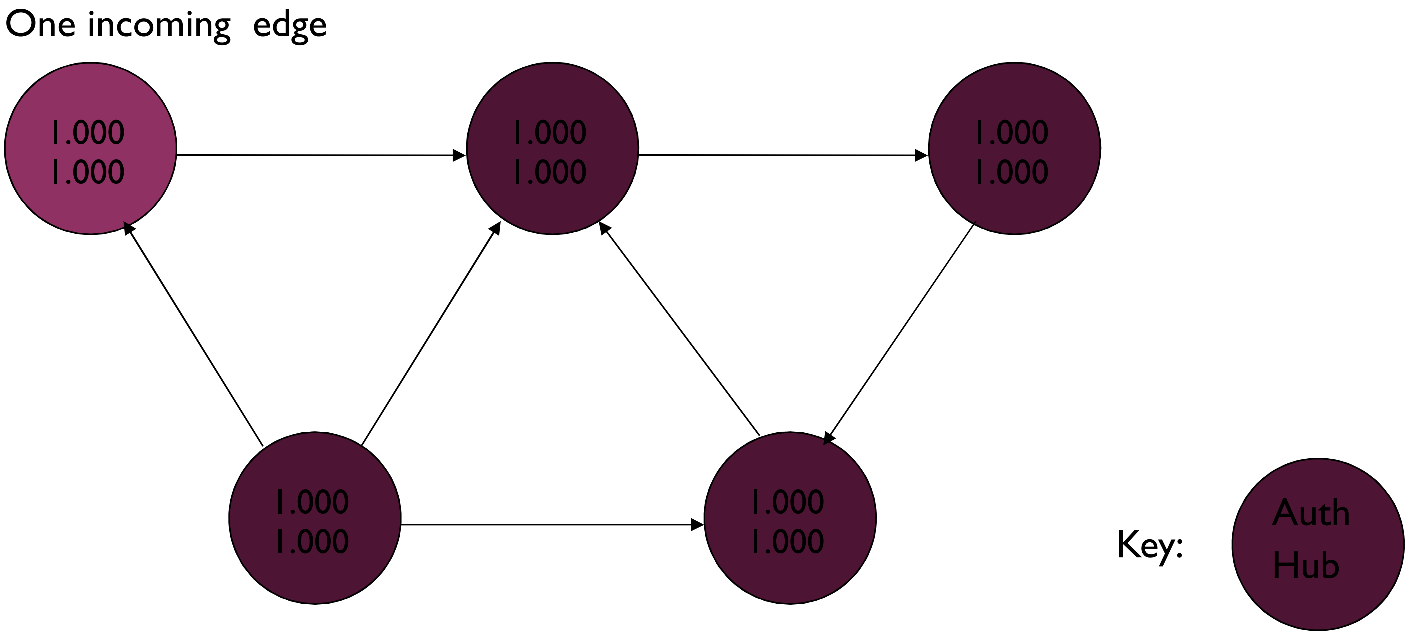


Key:



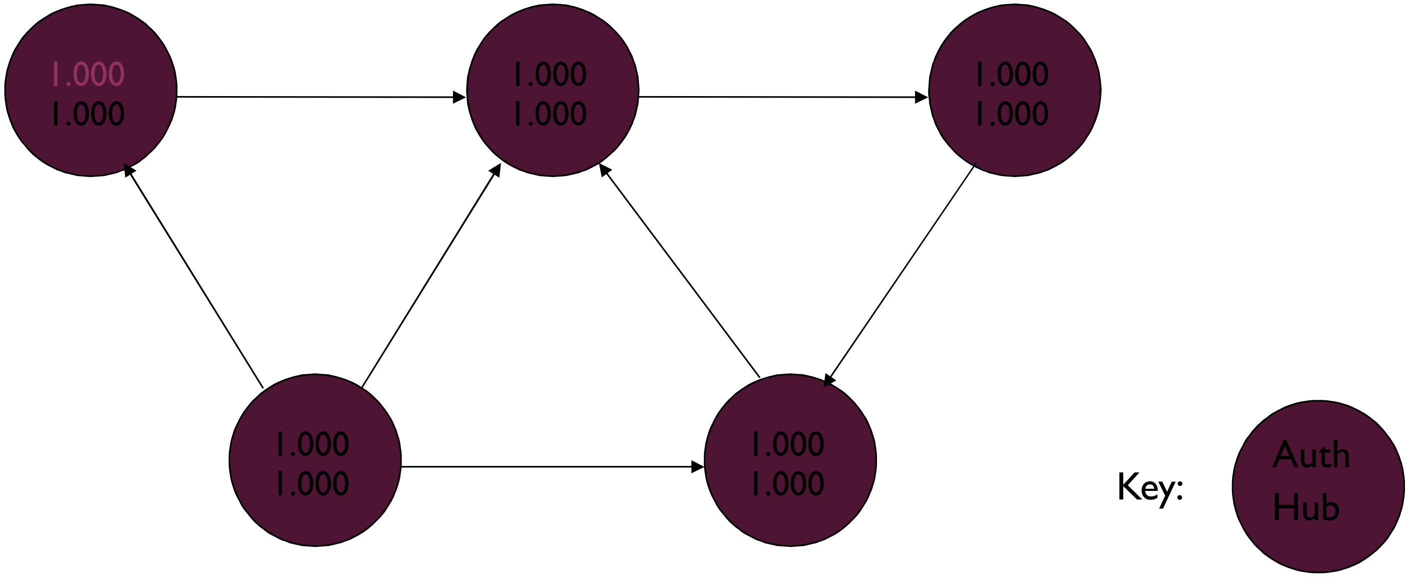
A Simple Example

Update Authority Scores first, using Hub scores



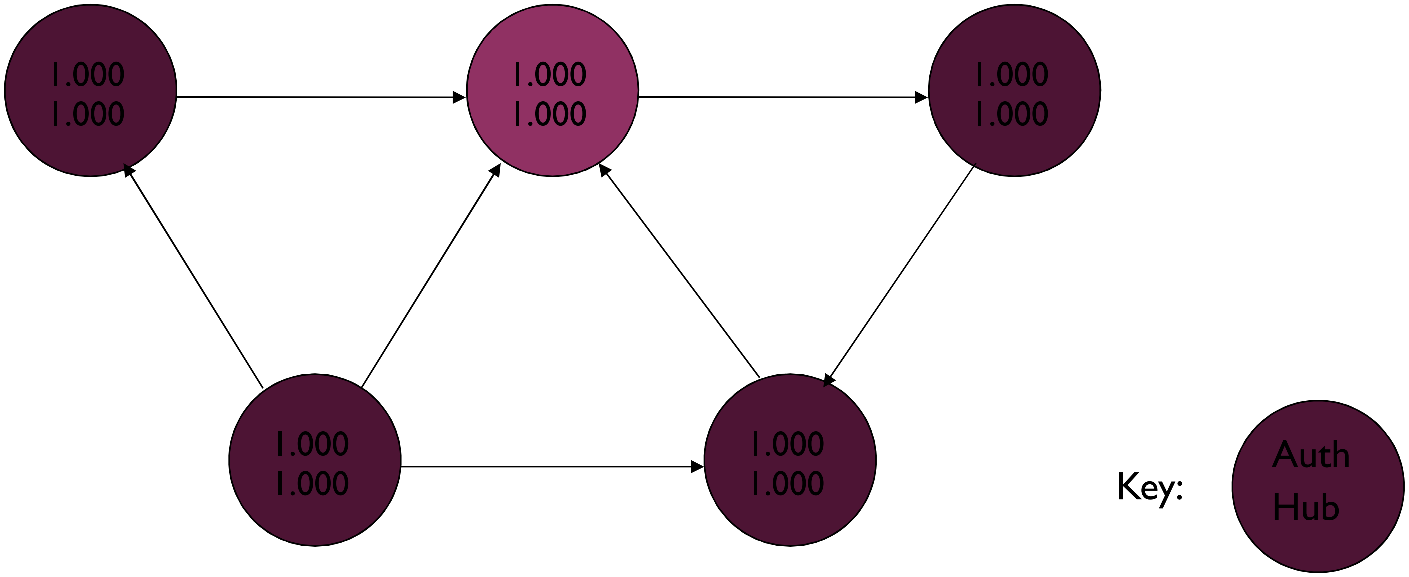
A Simple Example

Update Authority Scores first



A Simple Example

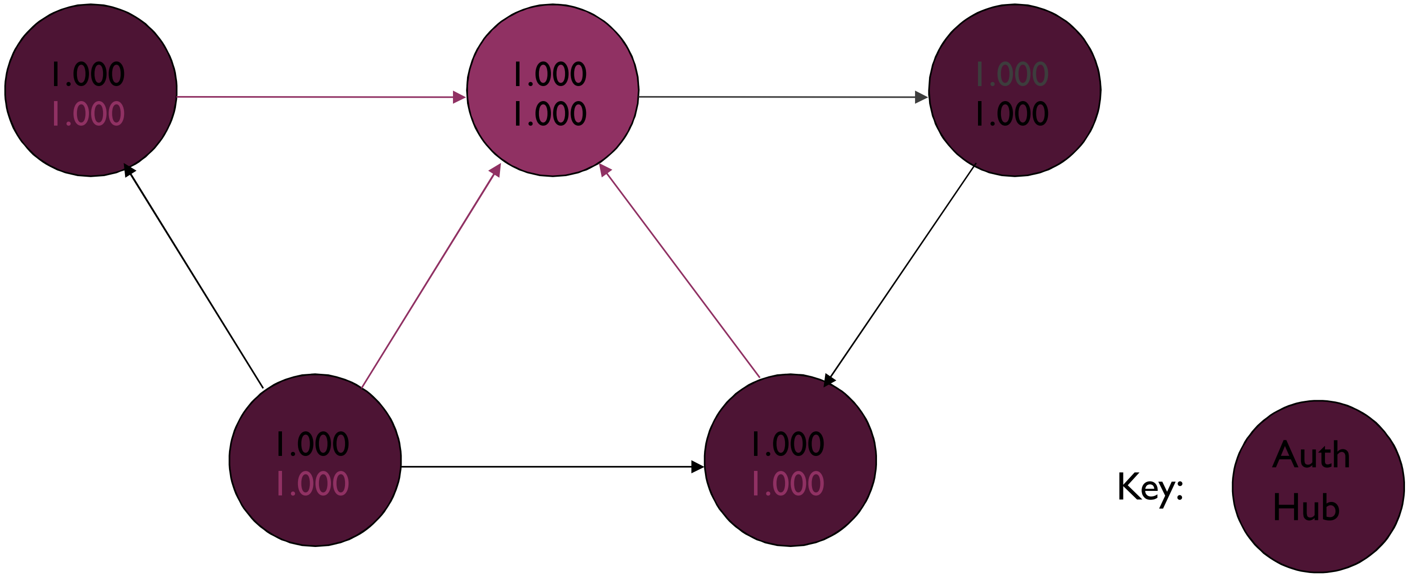
Update Authority Scores first



A Simple Example

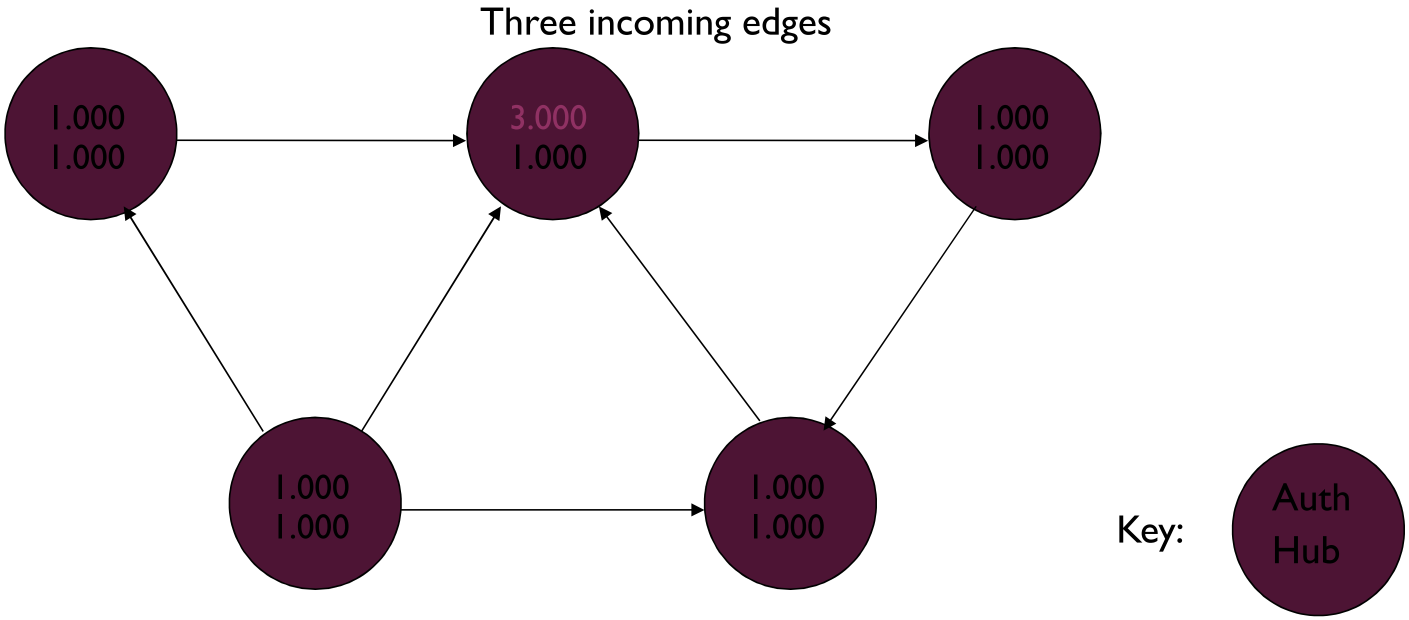
Update Authority Scores first

Three incoming edges



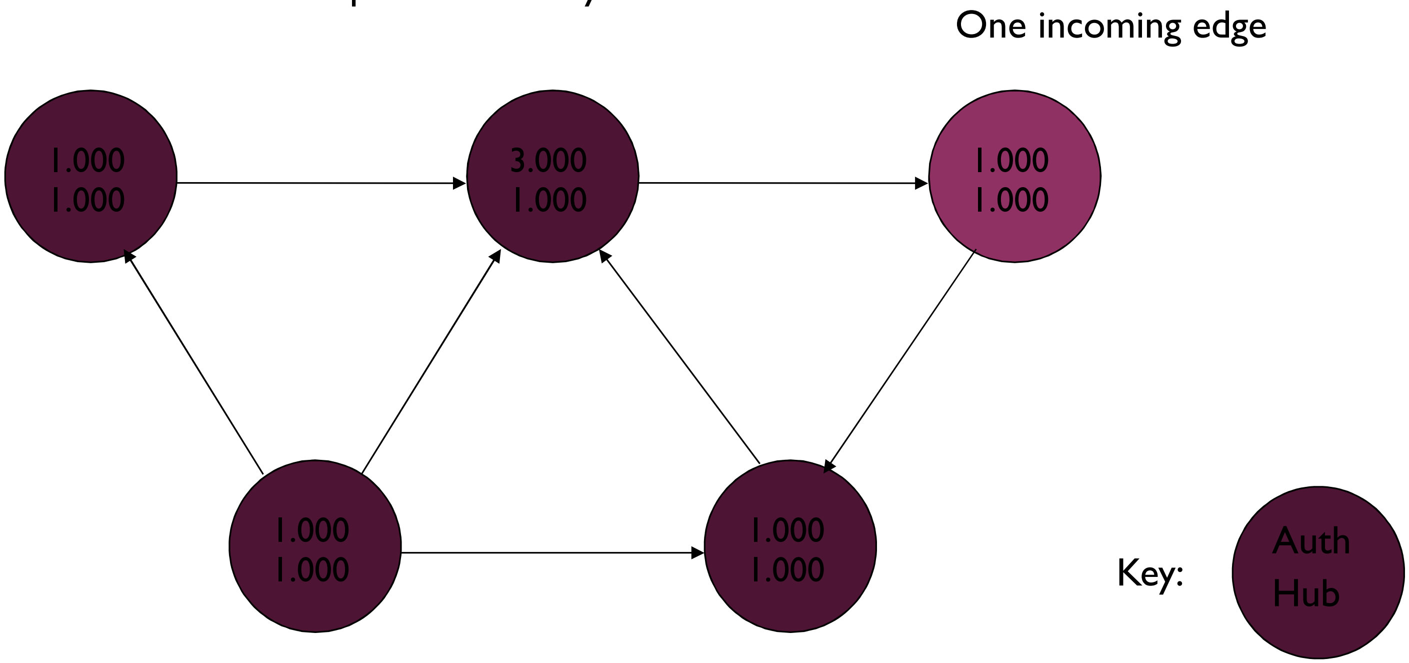
A Simple Example

Update Authority Scores first



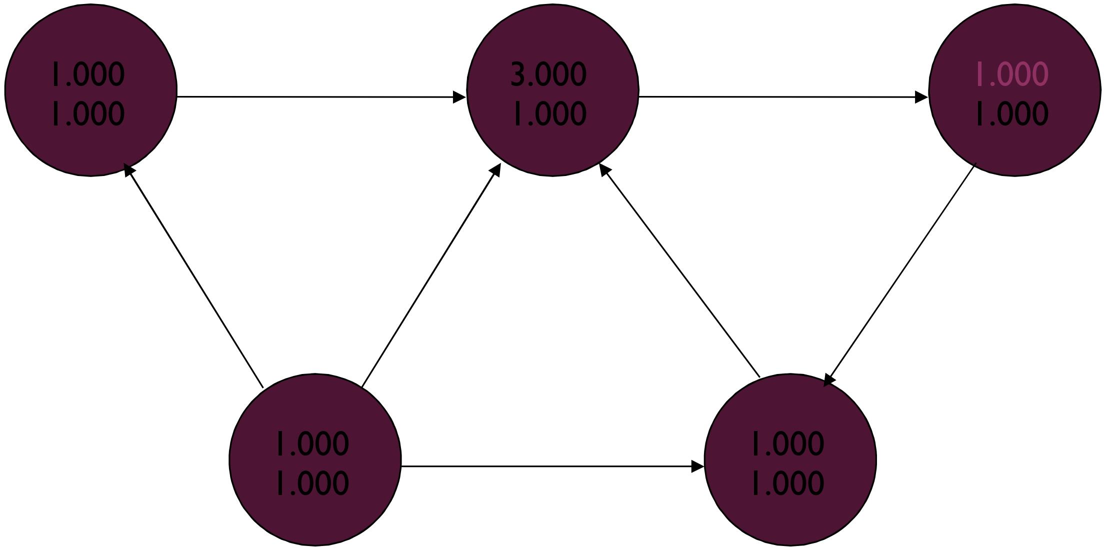
A Simple Example

Update Authority Scores first

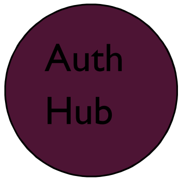


A Simple Example

Update Authority Scores first

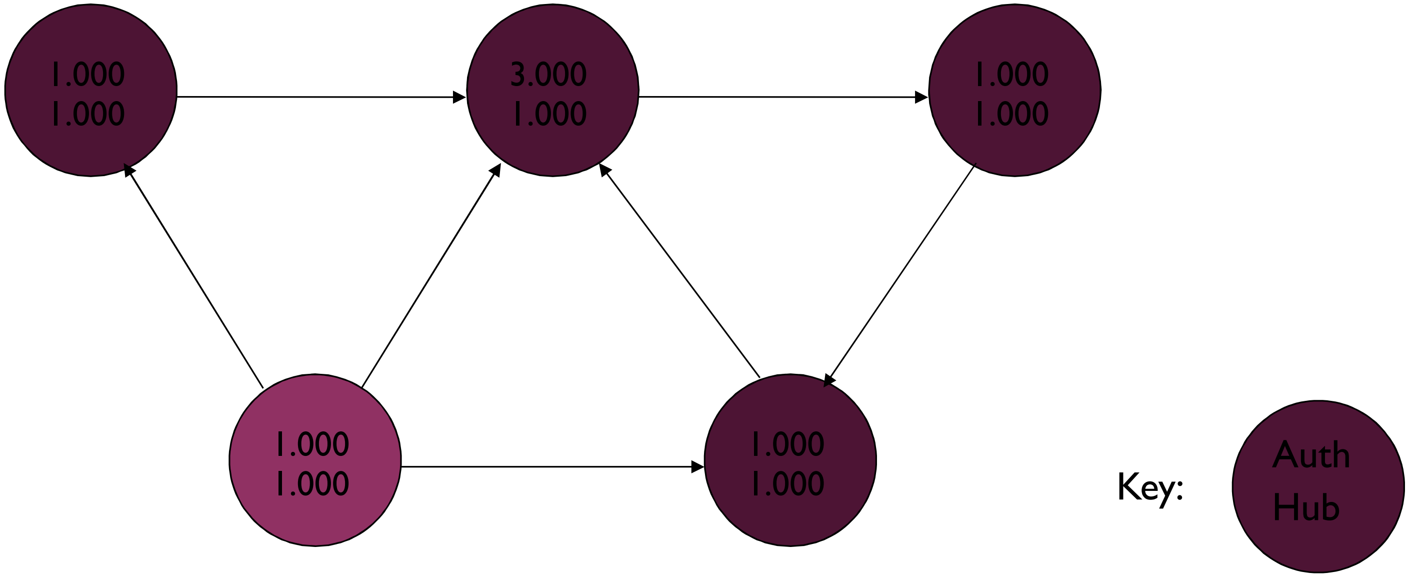


Key:



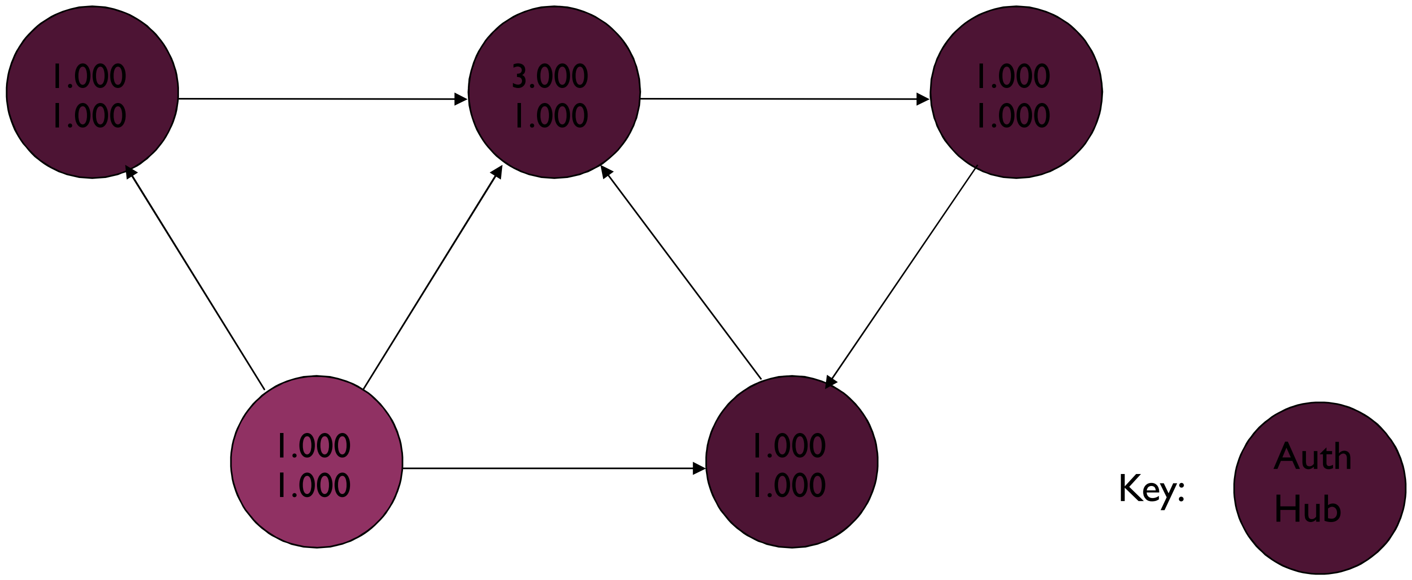
A Simple Example

Update Authority Scores first



A Simple Example

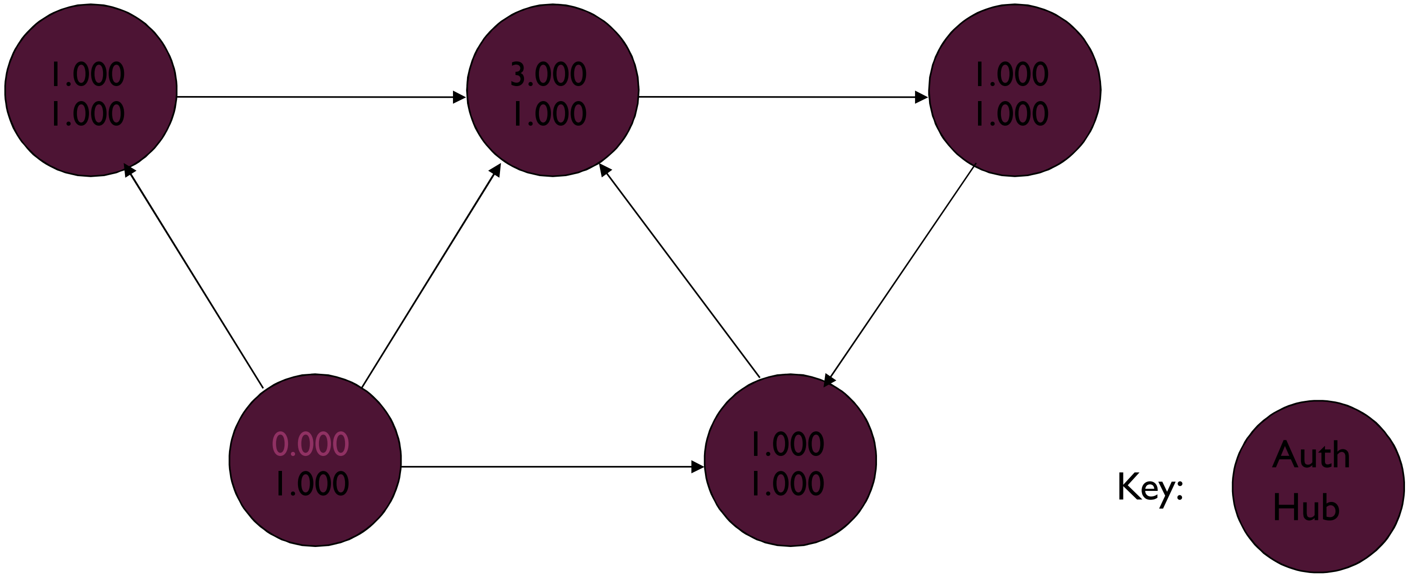
Update Authority Scores first



No Incoming Edges

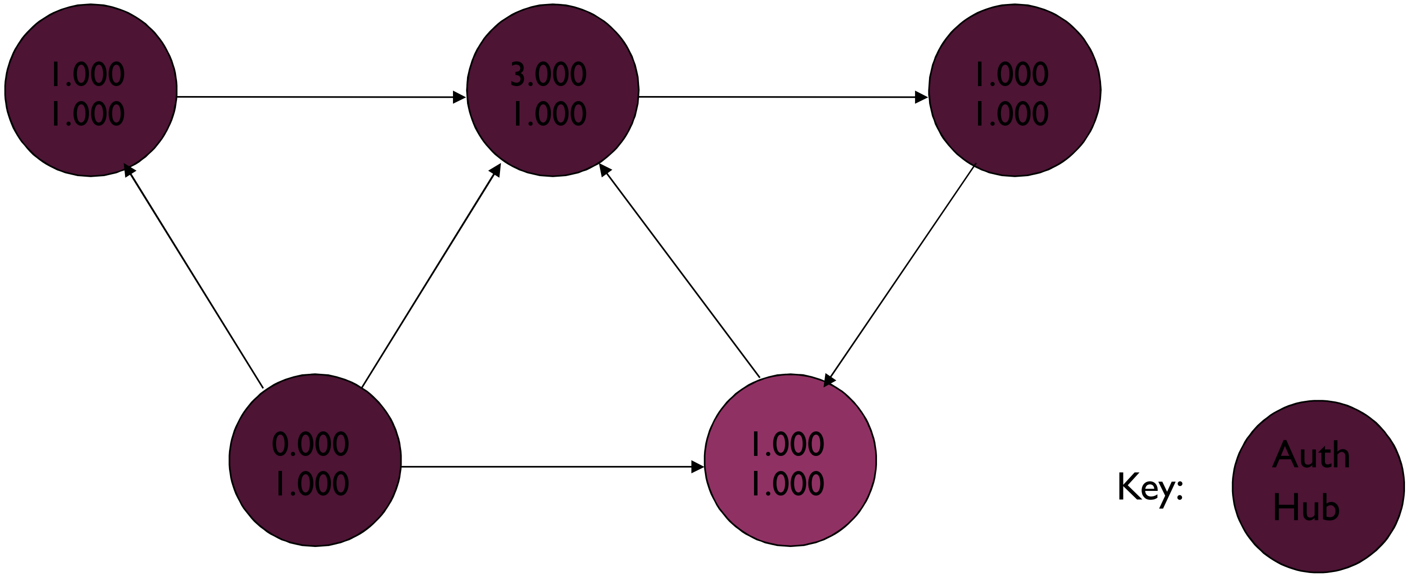
A Simple Example

Update Authority Scores first



A Simple Example

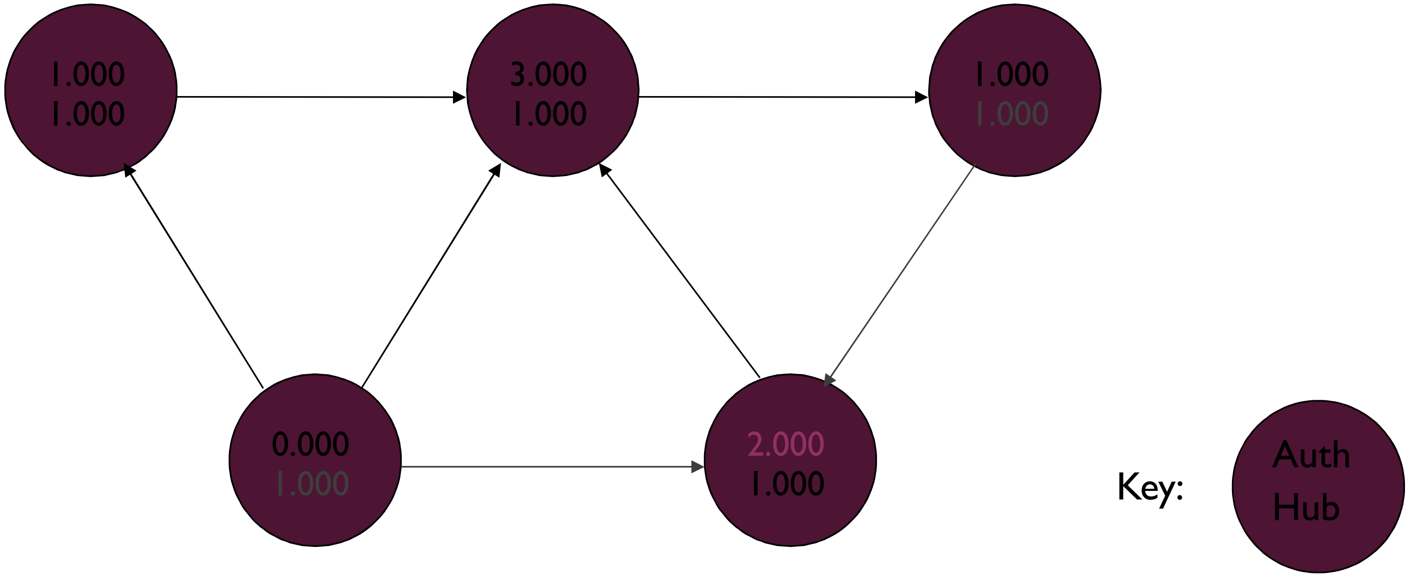
Update Authority Scores first



Two incoming edges

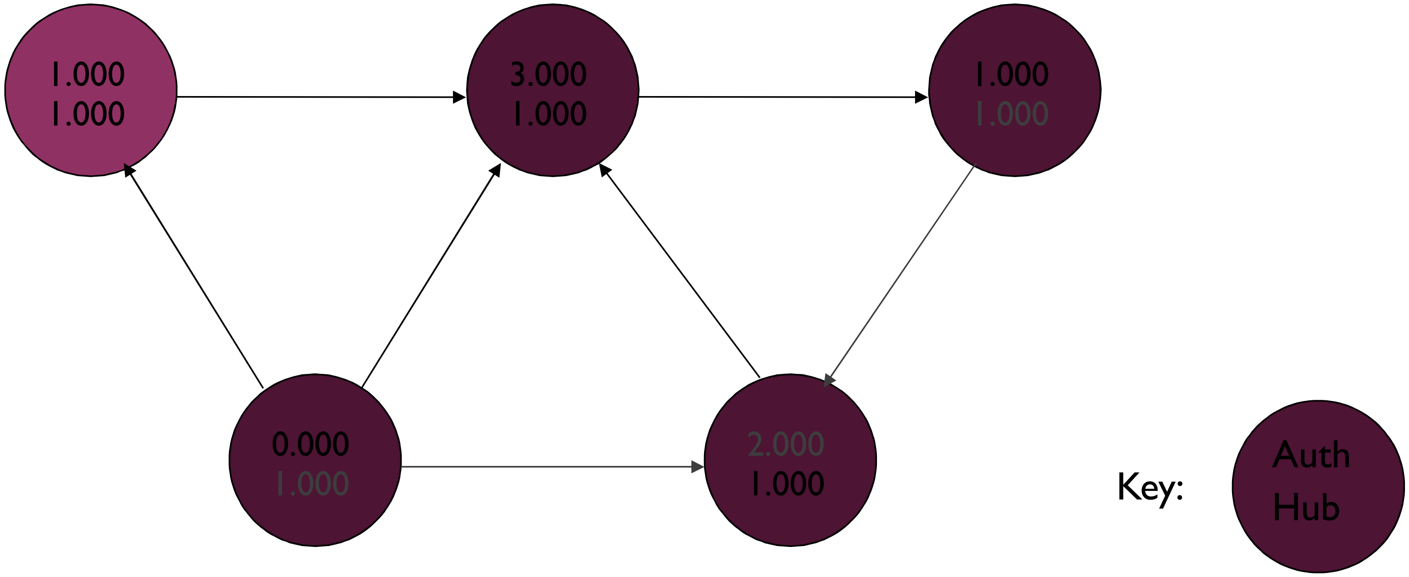
A Simple Example

Update Authority Scores first



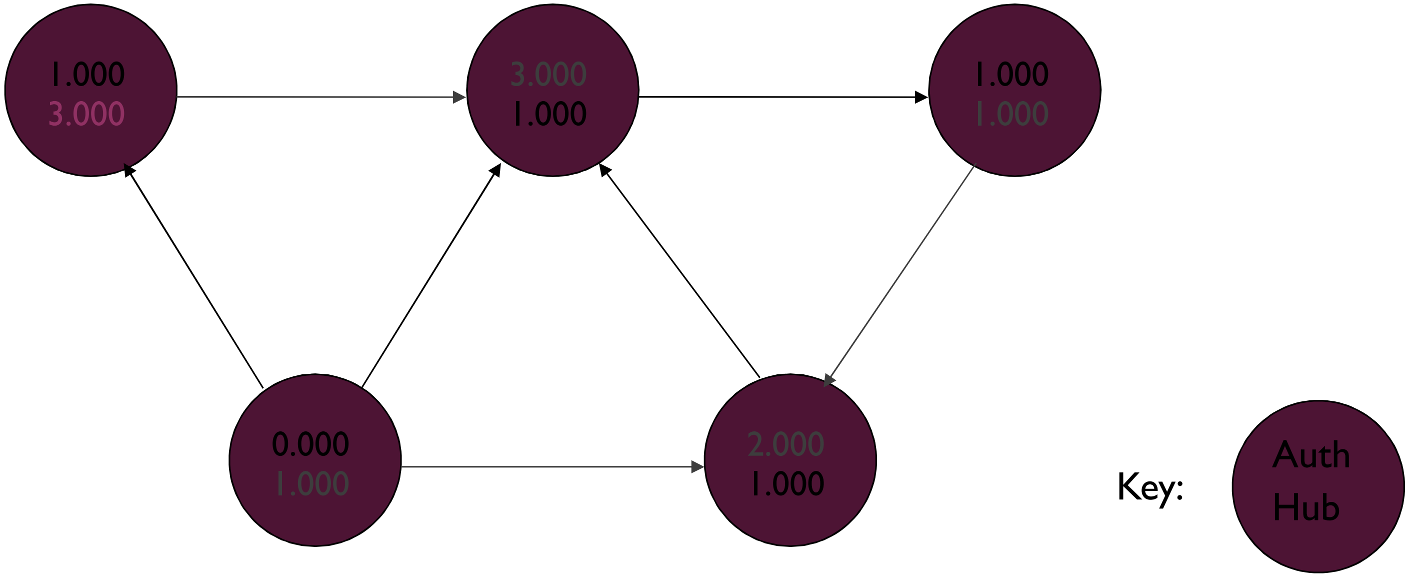
A Simple Example

Update Hub Scores using new Authority Scores



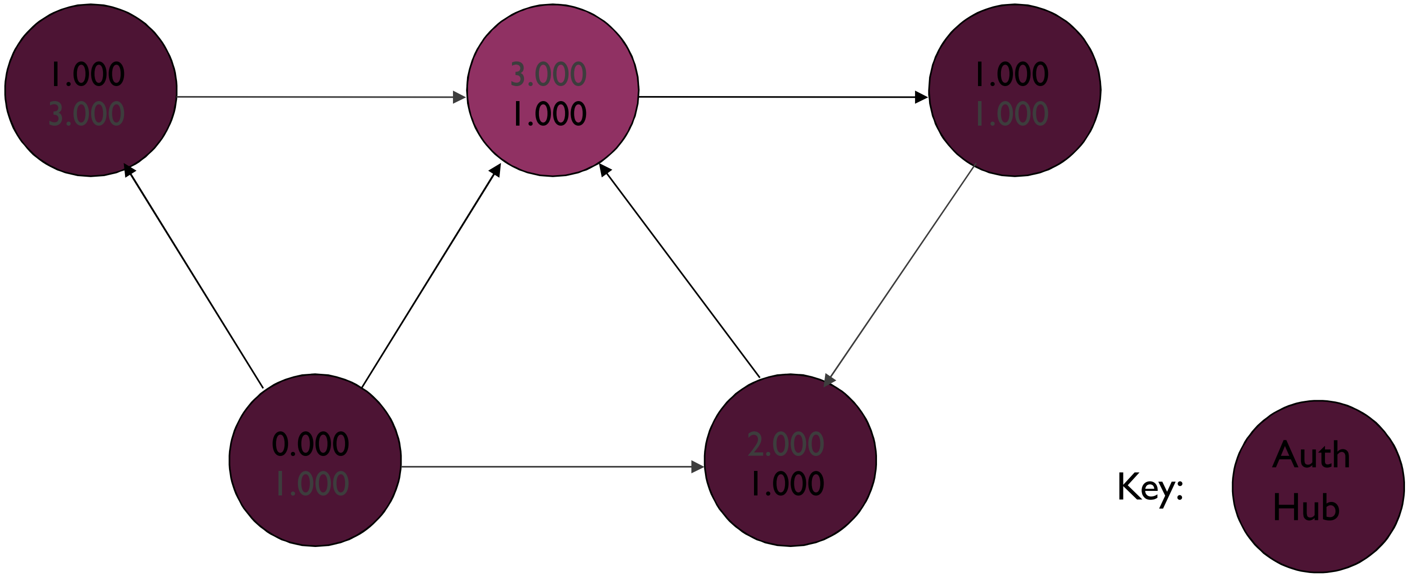
A Simple Example

Update Hub Scores using new Authority Scores



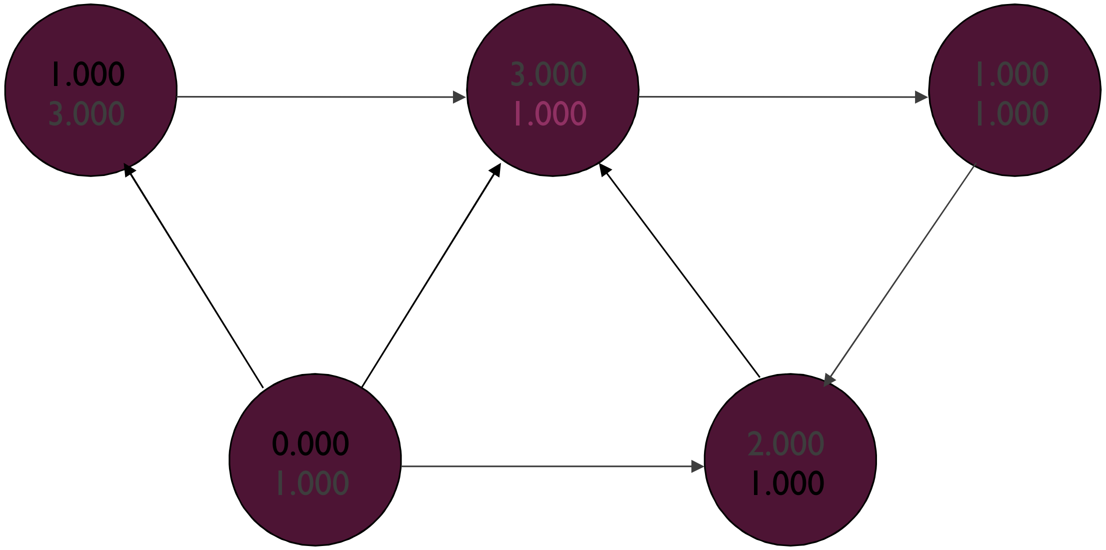
A Simple Example

Update Hub Scores using new Authority Scores

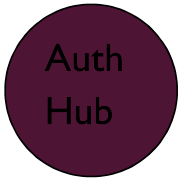


A Simple Example

Update Hub Scores using new Authority Scores

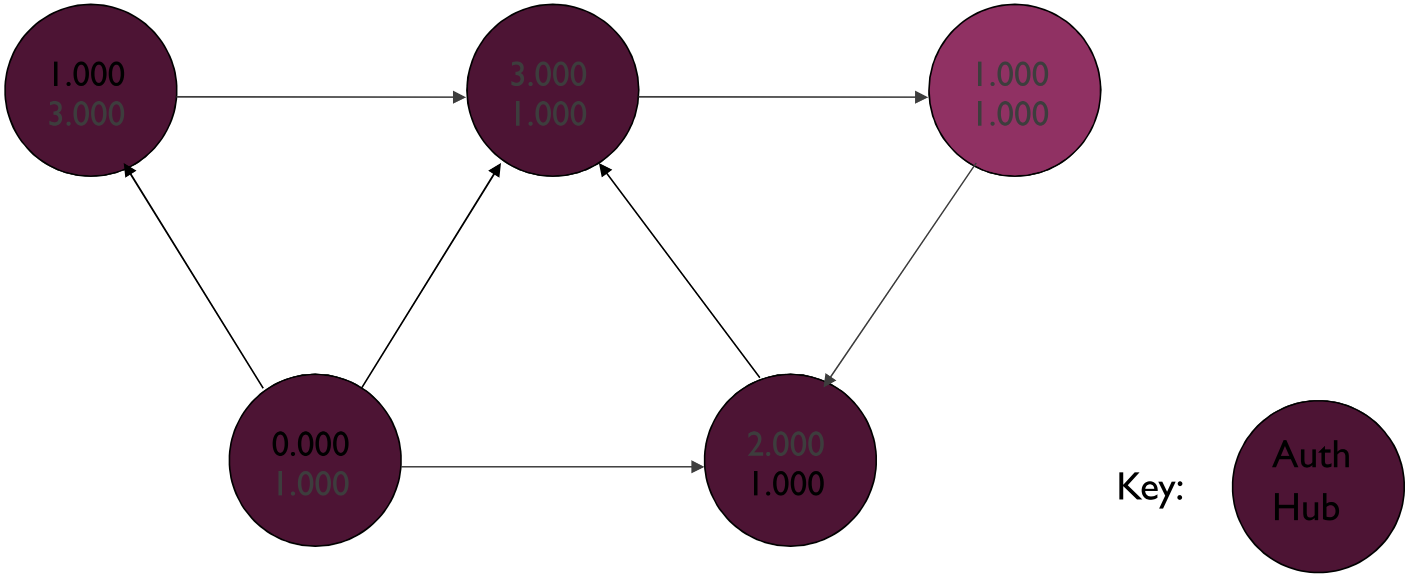


Key:



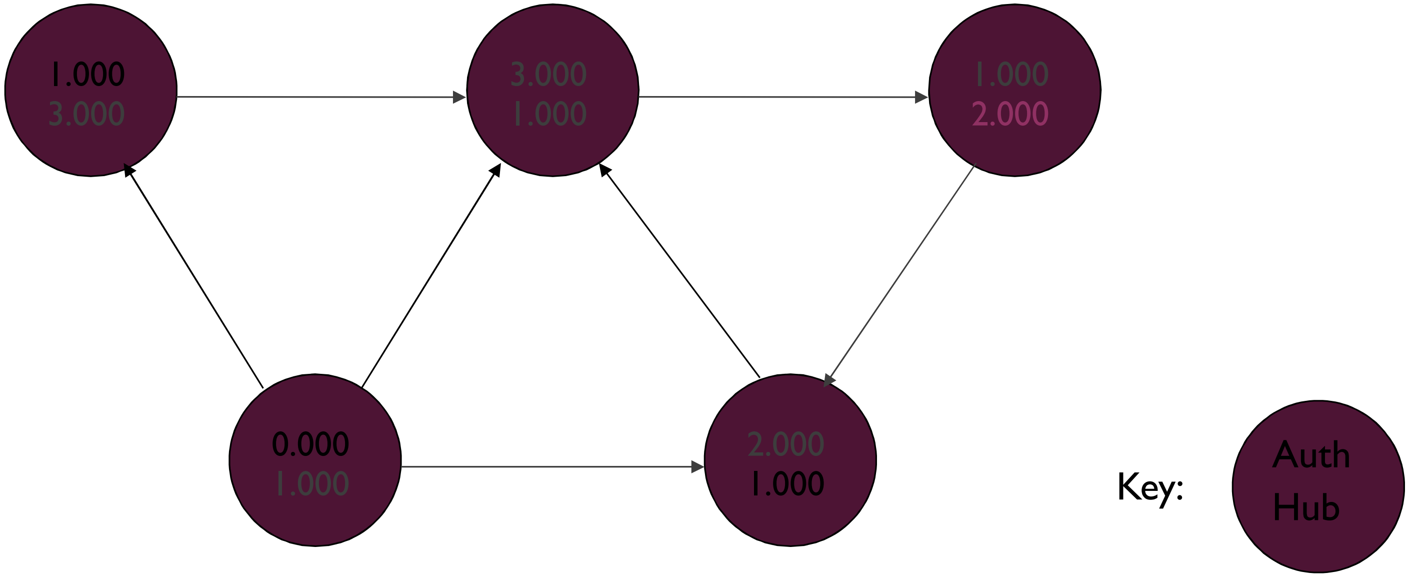
A Simple Example

Update Hub Scores using new Authority Scores



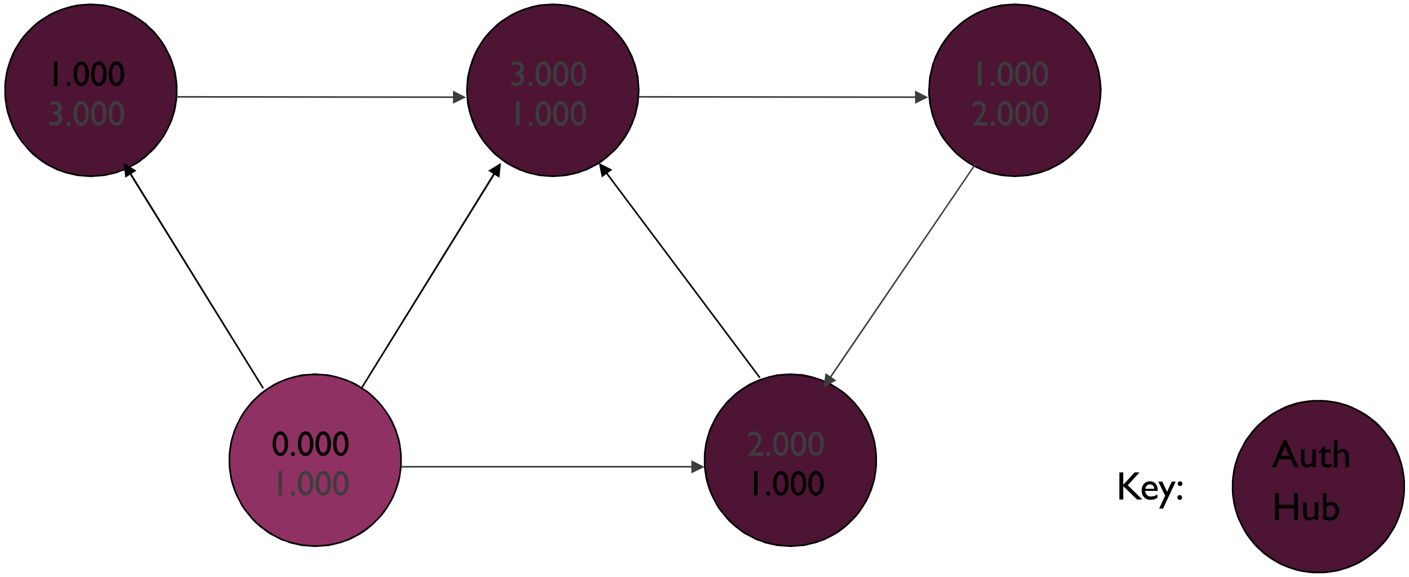
A Simple Example

Update Hub Scores using new Authority Scores



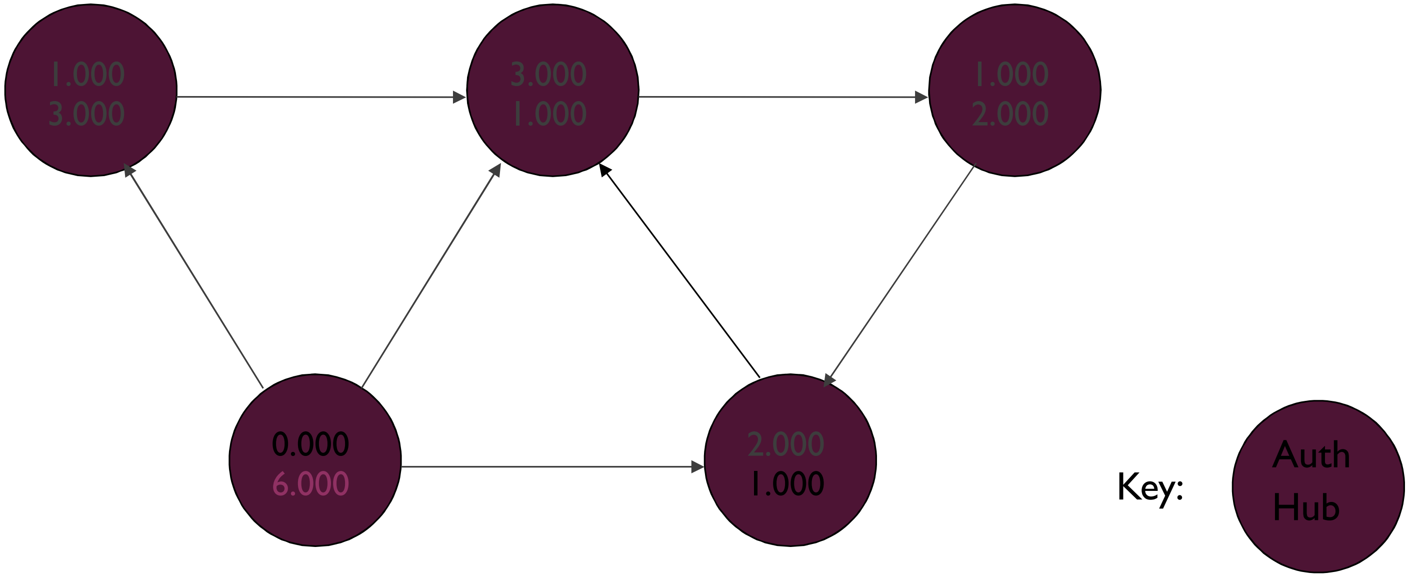
A Simple Example

Update Hub Scores using new Authority Scores



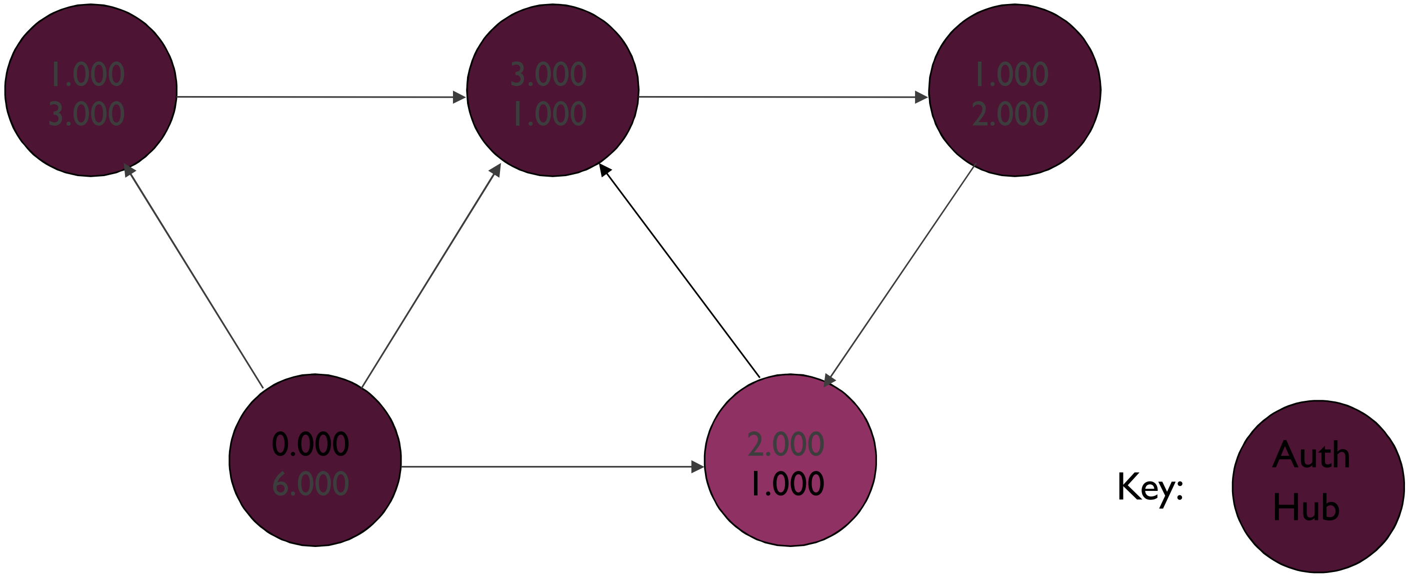
A Simple Example

Update Hub Scores using new Authority Scores

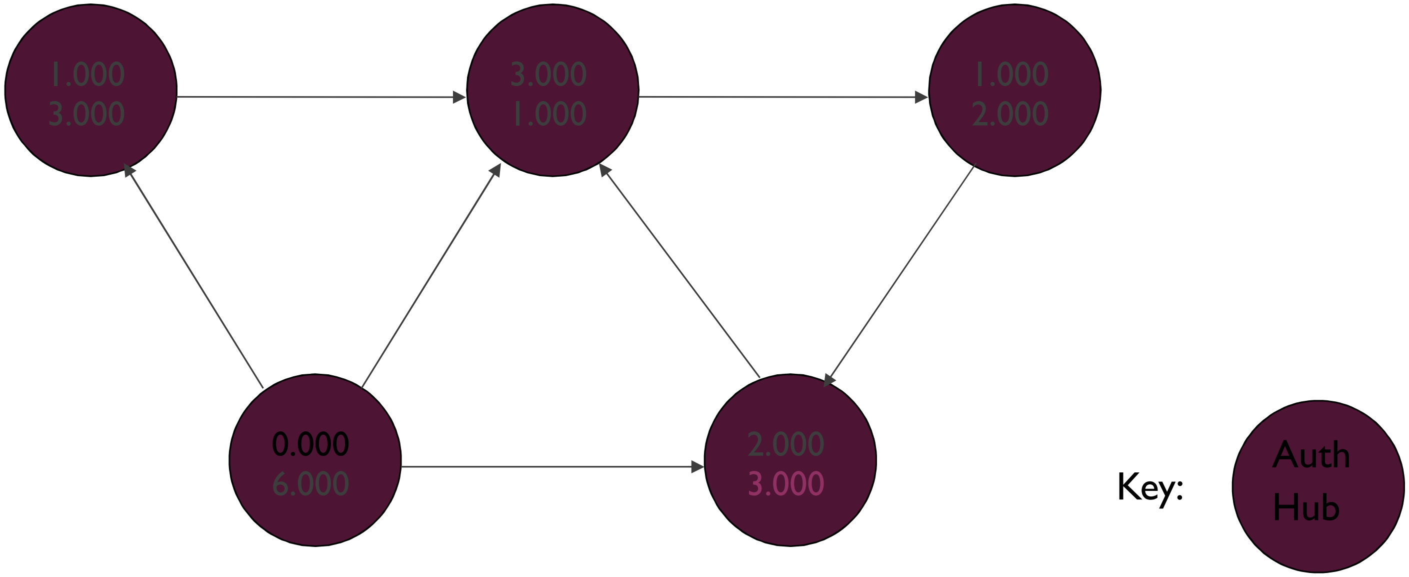


A Simple Example

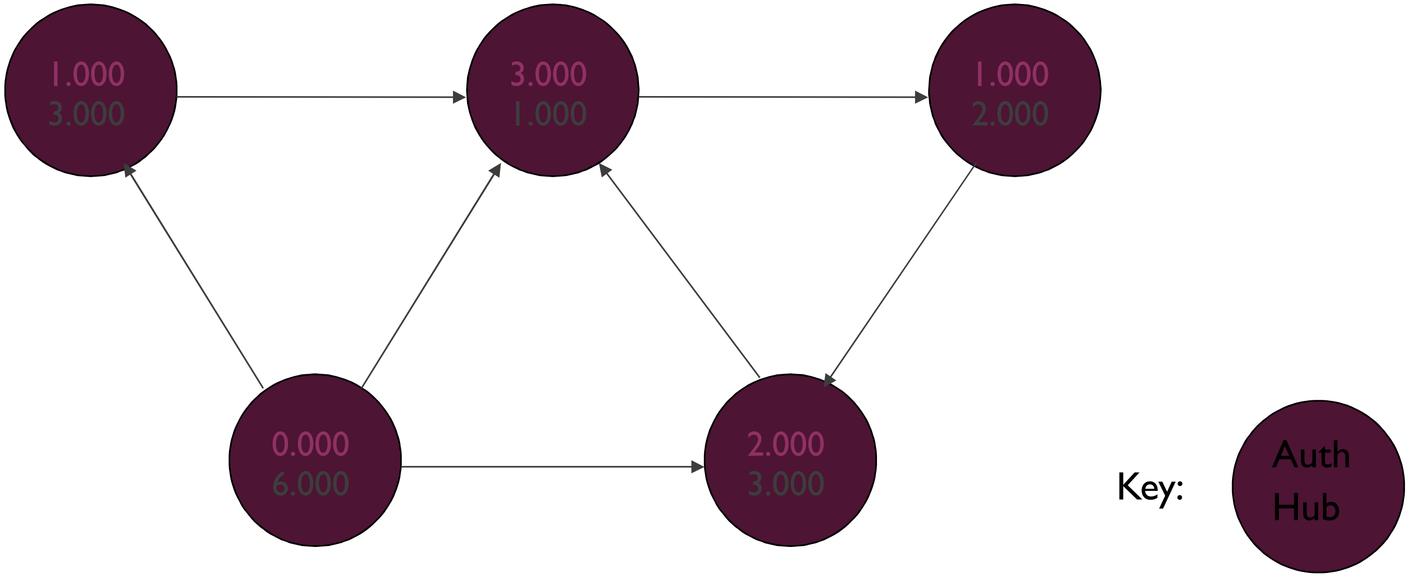
Update Hub Scores using new Authority Scores



A Simple Example

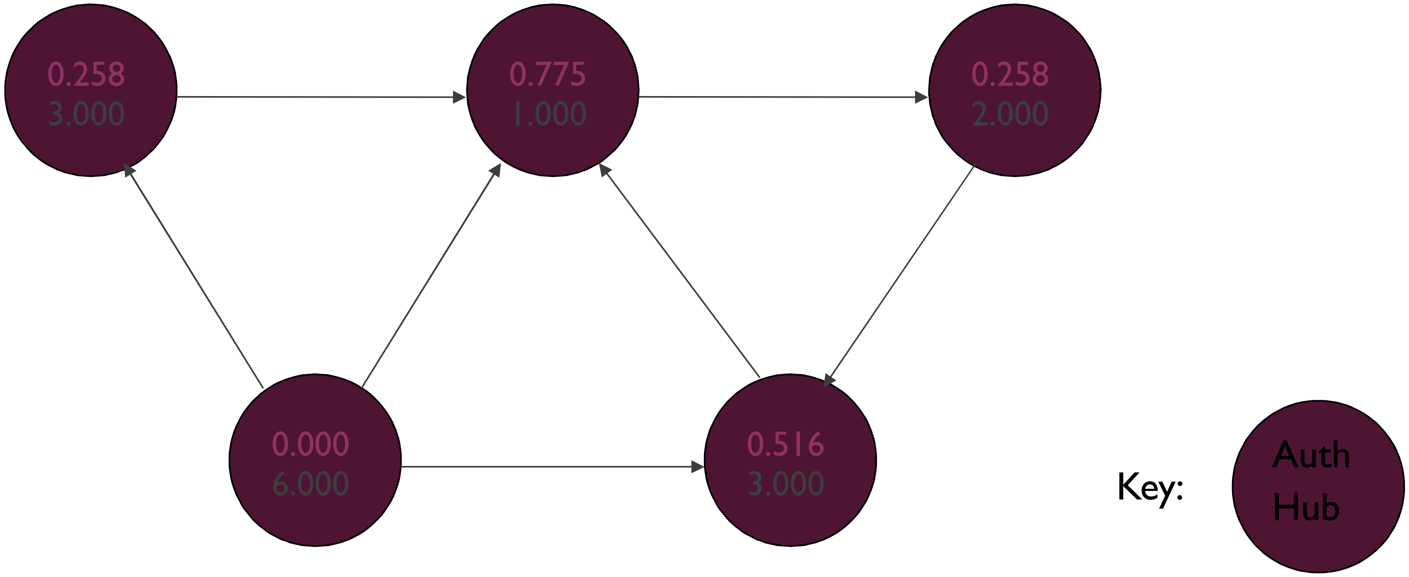


A Simple Example



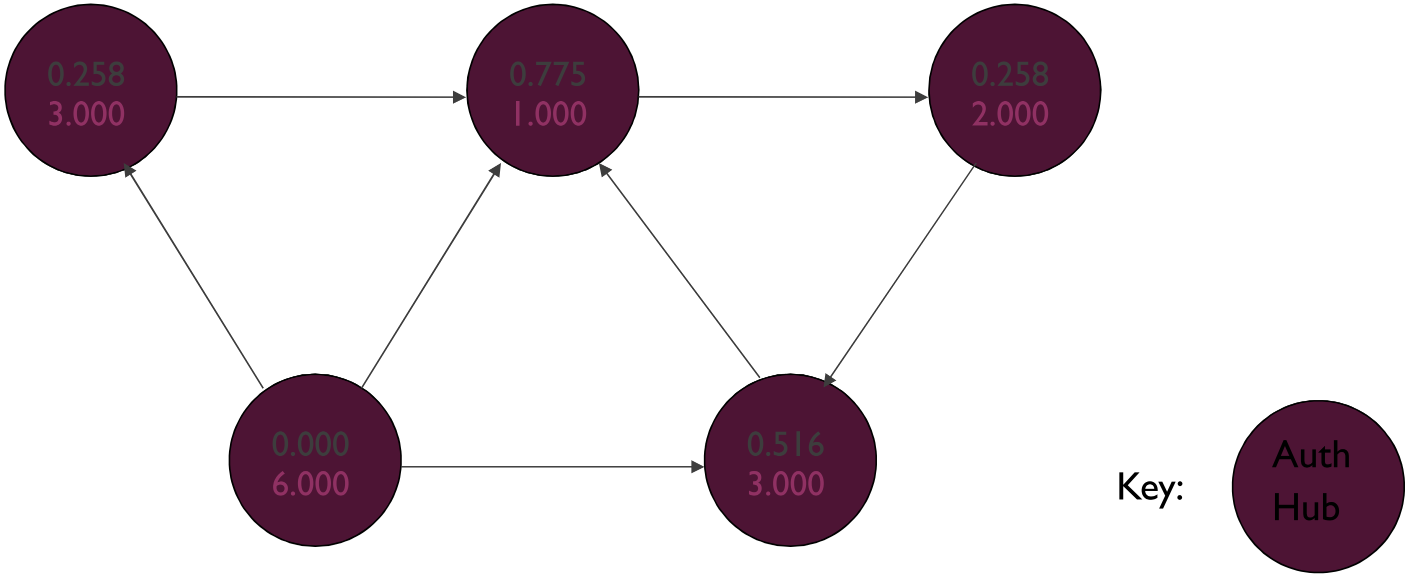
Sum of Squares: 15.000

A Simple Example



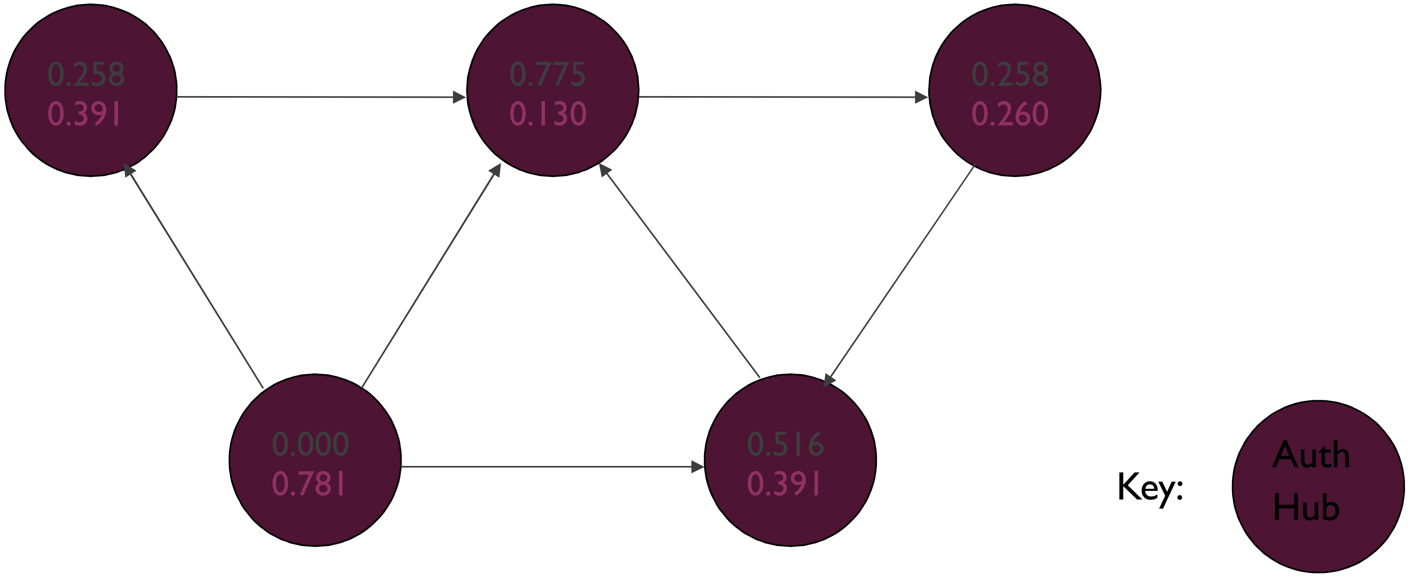
Divide By: 3.873 (sqrt(15))

A Simple Example



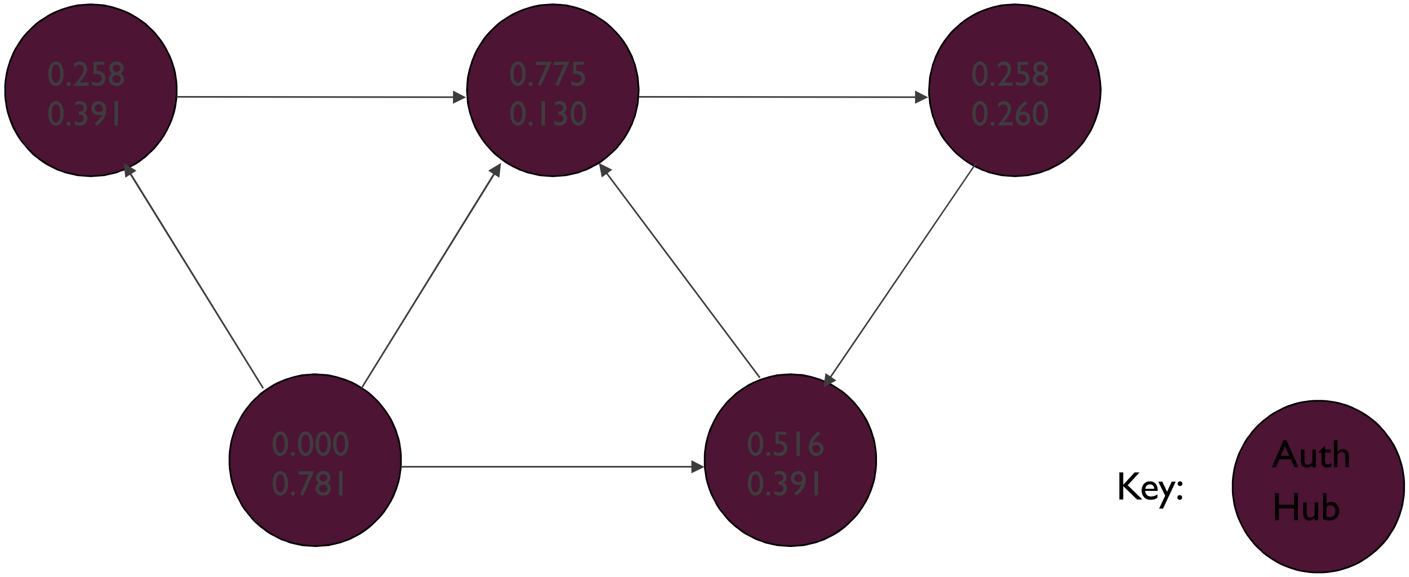
Sum of Squares: 59

A Simple Example



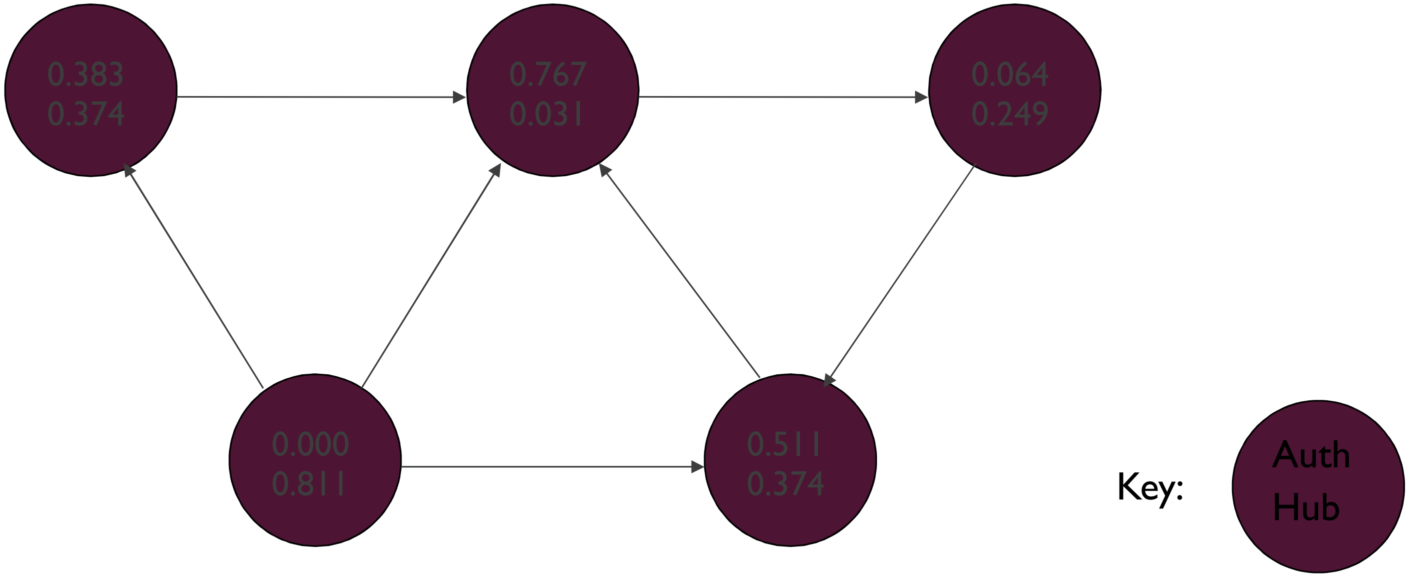
Divide By: 7.681 (sqrt(59))

A Simple Example



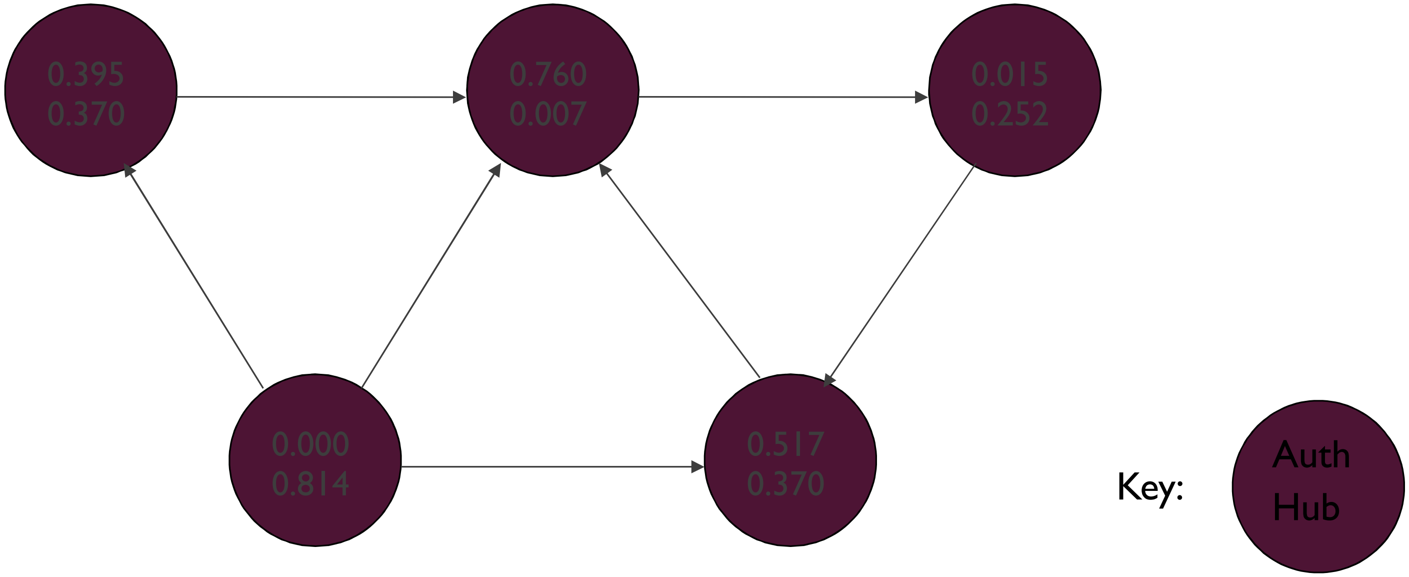
After First Iteration

A Simple Example



After Second Iteration

A Simple Example



After Third Iteration