

PageRank algorithm: main idea

How does Google decide which pages to show first?

The screenshot shows a Google search for the word "panda". The search bar is at the top with the word "panda" entered. Below the search bar, there are tabs for "All", "Images", "Videos", "Shopping", "News", "More", "Settings", and "Tools". The "All" tab is selected. Below the tabs, it says "About 610,000,000 results (0.70 seconds)".

The first result is an advertisement for Panda Antivirus. It says "Ad · https://www.pandasecurity.com/ 0808 169 2281". The headline is "50% Off Panda™ Antivirus 2021 · VPN Included at No Extra Cost". The description says "Next-gen Antivirus that Blocks 100% of Online Threats. Install Now & Get a Massive 50% Off. Choose from 4 Plans Depending On Your Needs. Works On All Devices & Operating...". There is a "Spring Sale: 50% off Panda Antivirus 2021 · Valid Apr 1 - Apr 30". Below the description, there are two columns of text: "Protection Advanced -40% The Most Advanced Protection for Your Digital Life, Data & Identity" and "Protection Premium -50% VPN & Maximum Protection with Premium Services & 24-7 VIP Support".

The second result is from Wikipedia. It says "https://en.wikipedia.org › wiki › Giant_panda". The headline is "Giant panda - Wikipedia". The description says "The giant **panda** (*Ailuropoda melanoleuca*; Chinese: 大熊猫; pinyin: dàxióngmāo), also known as the **panda** bear or simply the **panda**, is a bear native to South Central China. It is characterised by its bold black-and-white coat and rotund body." Below the description, there is a table with two columns: "Species: A. melanoleuca", "Genus: *Ailuropoda*", "Family: *Ursidae*", and "Kingdom: *Animalia*". Below the table, there are links: "Red panda · Qinling panda · Panda (disambiguation) · Panda tea".

The third result is also from Wikipedia. It says "https://en.wikipedia.org › wiki › Google_Panda". The headline is "Google Panda - Wikipedia". The description says "Google **Panda** is a major change to Google's search results ranking algorithm that was first released in February 2011. The change aimed to lower the rank of ...".

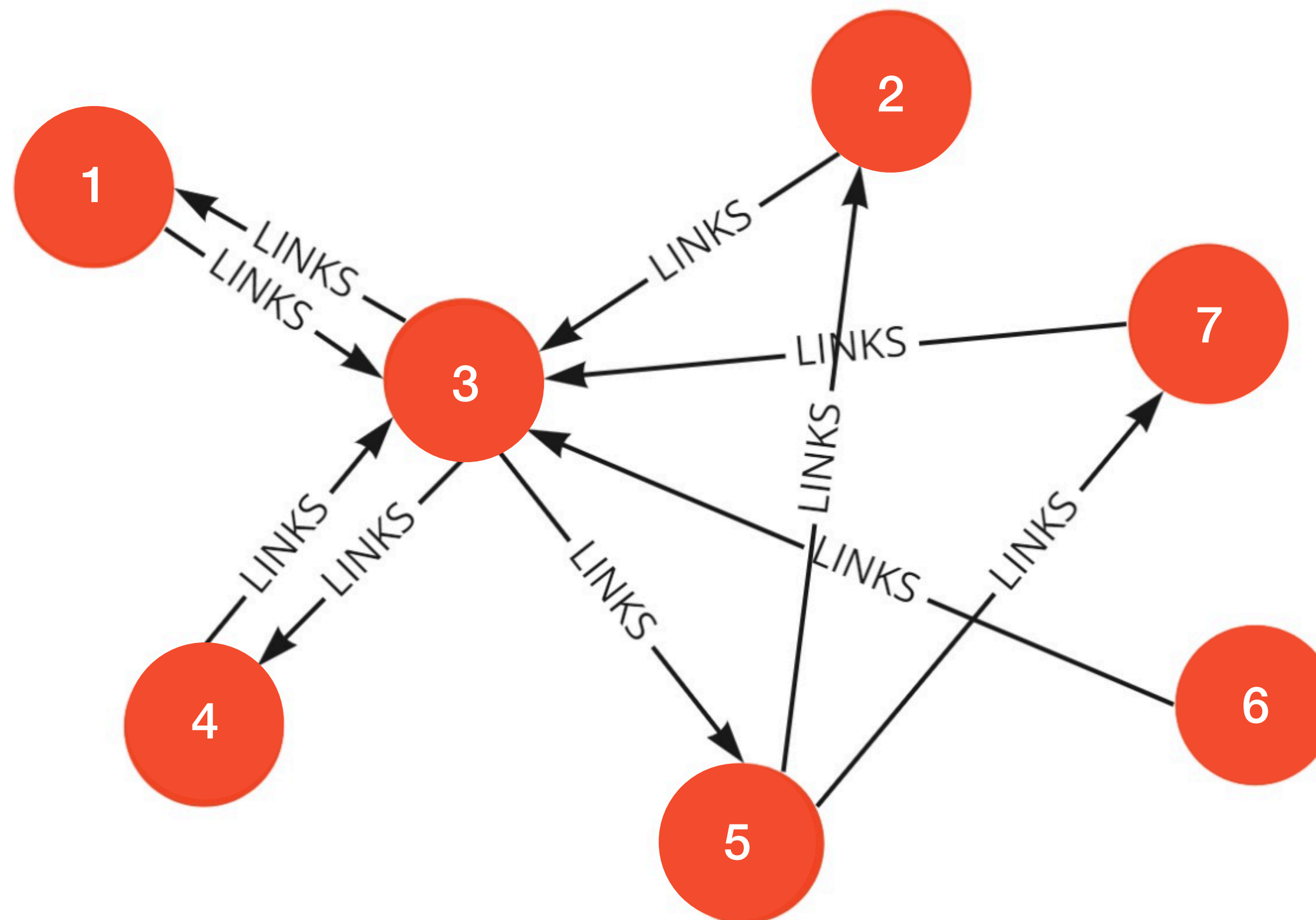
The fourth result is from WWF. It says "https://www.worldwildlife.org › species › giant-panda". The headline is "Giant Panda | Species | WWF". The description says "Pandas live mainly in temperate forests high in the mountains of southwest China, where they subsist almost entirely on bamboo. They must eat around 26 to 84 ...".

PageRank algorithm

- Proposed in 1998 by Sergey Brin and Larry Page (the founders of Google).
- PageRank (PR) is an algorithm used by Google Search to rank web pages in their search engine results.
- Algorithm uses the web graph to compute a measure of importance (PageRank) of every webpage.
- PageRank is a static ranking of Web pages, i.e. a PageRank value is computed for each page off-line and it does not depend on search queries.

Web graph

- **Nodes:** webpages
- **Directed edges:** hyperlinks between the pages



Terminology

- **In-links** of page *a*: the hyperlinks that point to page *a* from other pages. Usually, hyperlinks from the same page are ignored.
- **Out-links** of page *a*: the hyperlinks that point out to other pages from page *a*. Usually, hyperlinks to the same site are ignored.

PageRank algorithm idea

- The PageRank score of each page can be regarded as its prestige
- PageRank interprets a hyperlink from page *a* to page *b* as a vote, by page *a*, for page *b*.
- However, unlike Degree Prestige, PageRank looks at more than just the sheer number of votes or links that a page receives. It also analyzes the page that casts the vote.
- Votes casted by pages that are themselves “important” are weighted more heavily and help to make other pages more “important”

PageRank algorithm idea

- The importance of page a (i.e. a 's PageRank score) is determined by summing up the PageRank scores of all pages that point to a .
- Since page may point to many other pages, its PageRank score should be shared among all the pages that it points to.

PageRank algorithm idea

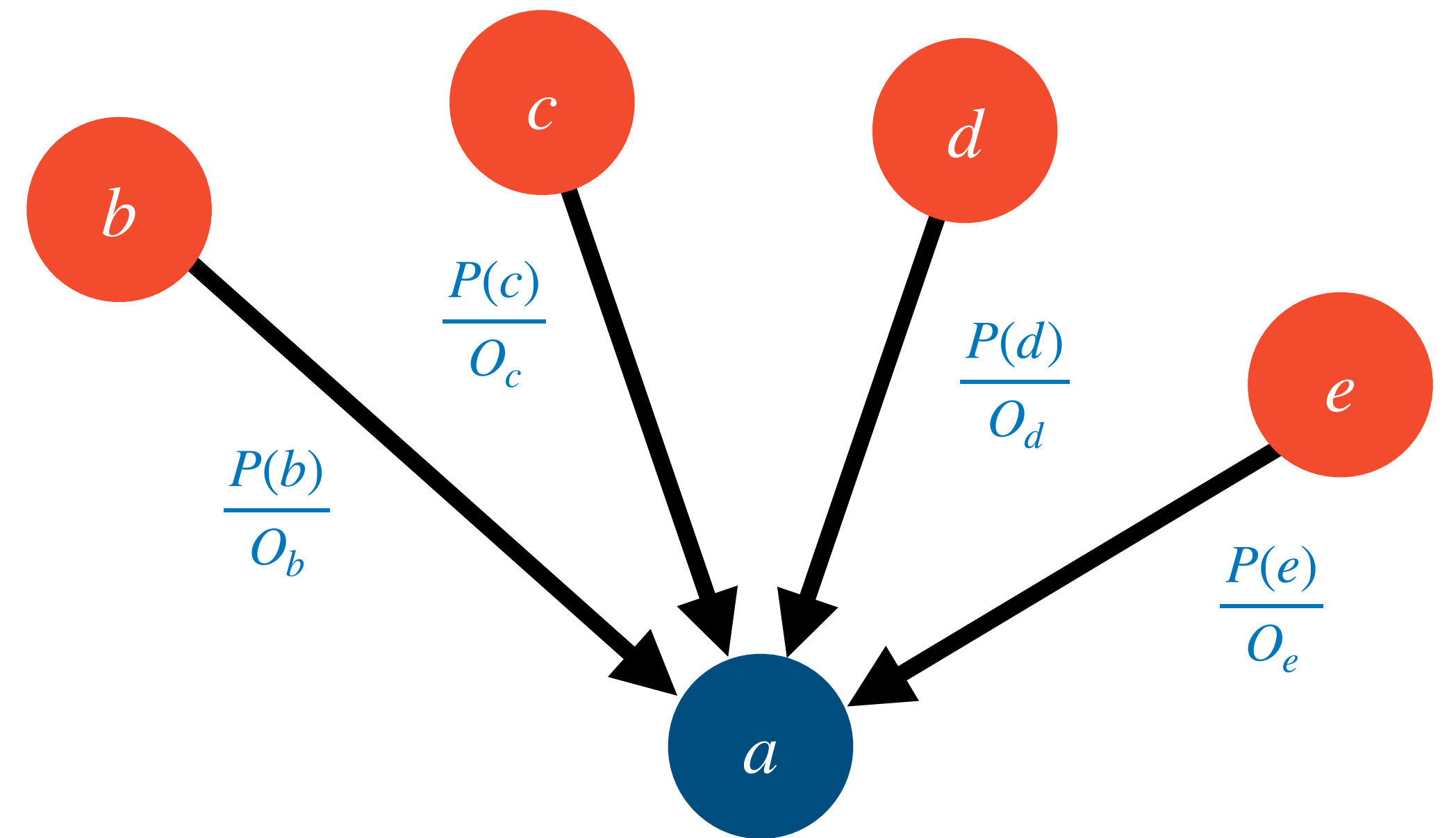
Notation

- $P(a)$ — PageRank score of page a
- O_a — the number of out-links of page a
- E — the set of arcs (directed edges) of the graph

The PageRank score of page a is defined by

$$P(a) = \sum_{(x,a) \in E} \frac{P(x)}{O_x}$$

We have a similar equation for every vertex in the graph, which leads to a system of linear equations.



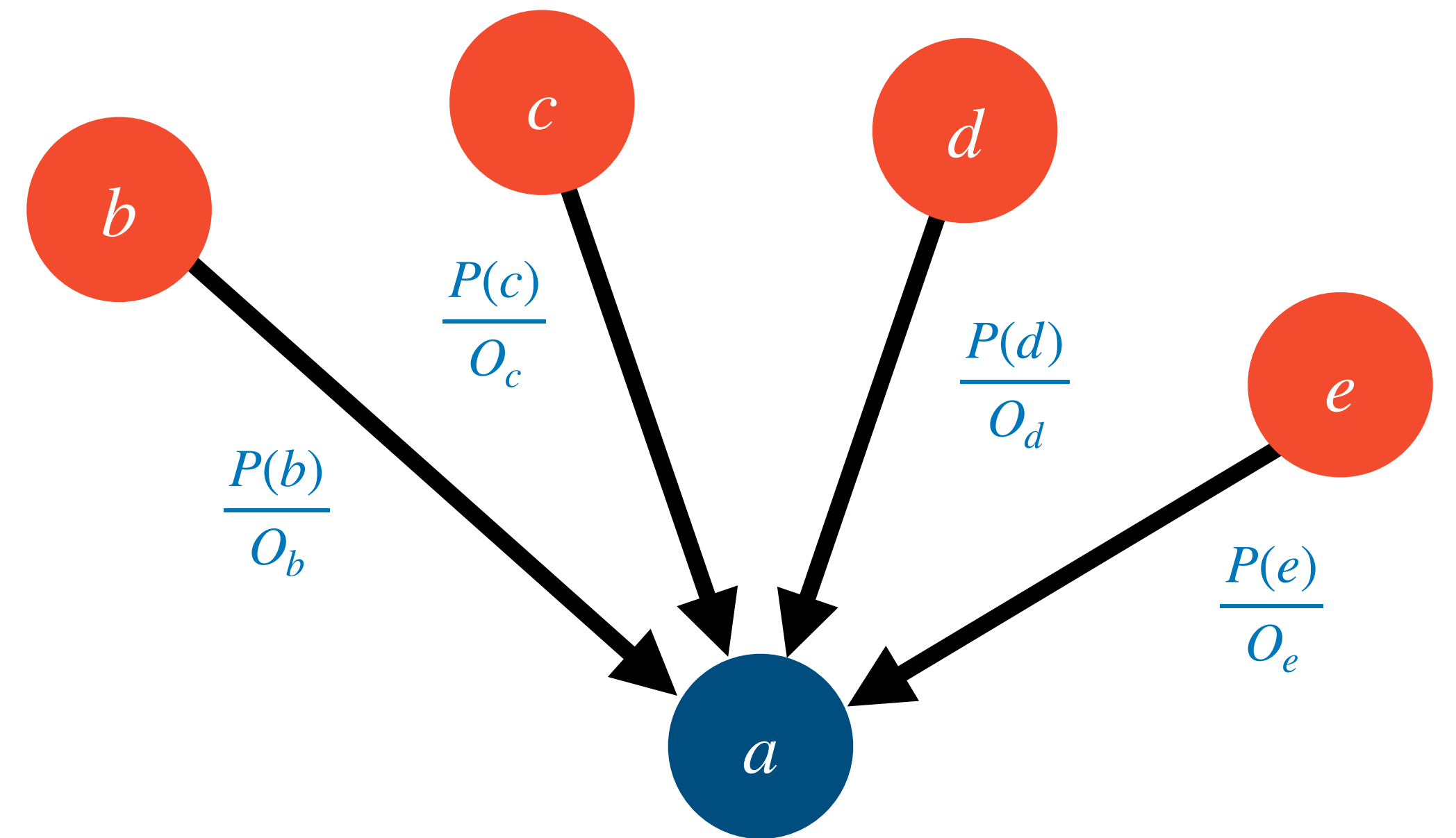
PageRank algorithm idea

- Let $1, 2, \dots, n$ be the vertices of the graph.
- Let \bar{P} be a n -dimensional column vector of PageRank scores of the vertices, i.e.

$$\bar{P} = (P(1), P(2), \dots, P(n))^T$$

- Let \bar{A} be the modified adjacency matrix of our graph with

$$\bar{A}_{ij} = \frac{1}{O_i}, \text{ if } (i, j) \in E, \text{ and}$$
$$\bar{A}_{ij} = 0, \text{ otherwise}$$



PageRank algorithm idea

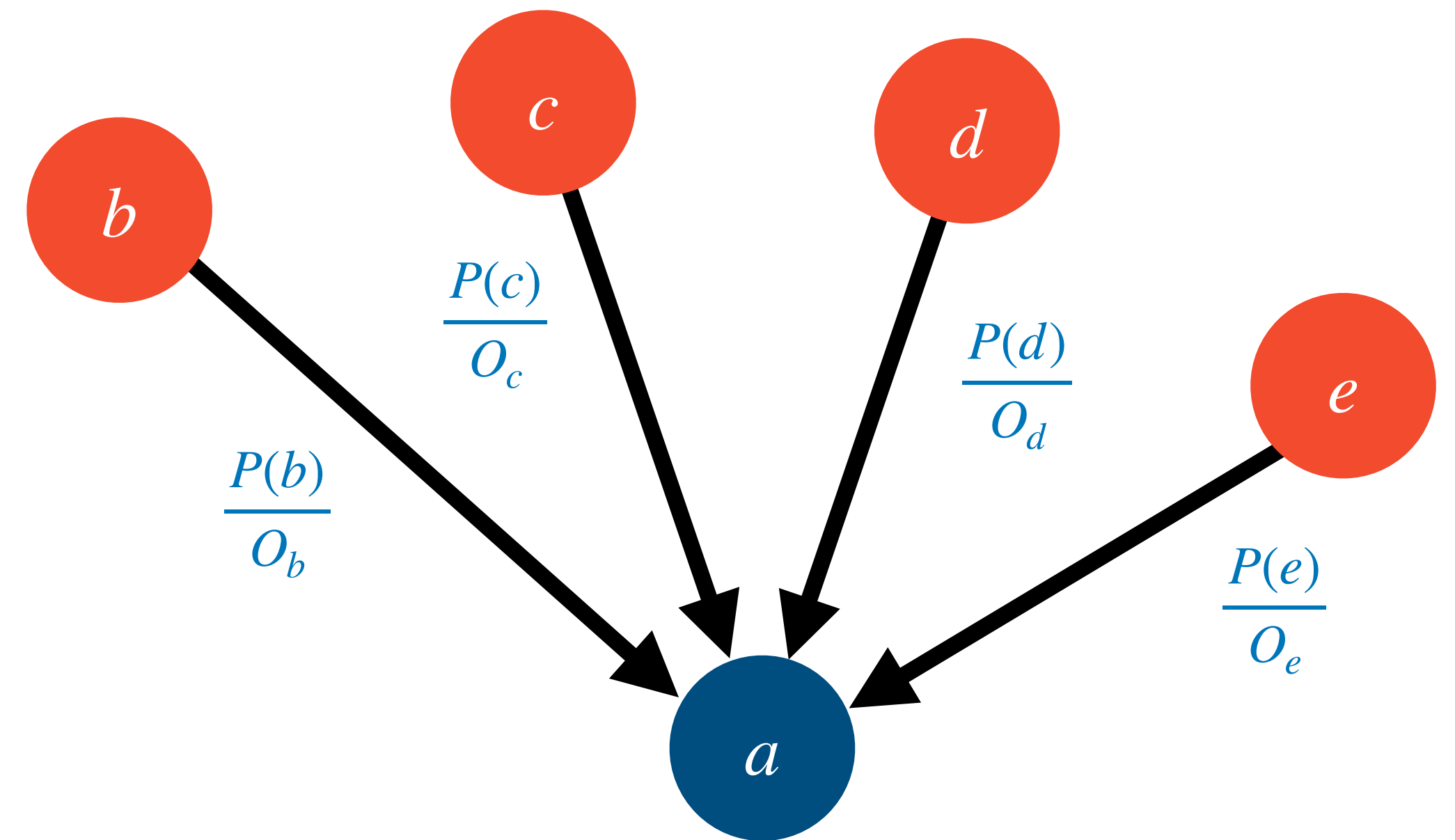
Under this notation we can write the system

$$P(i) = \sum_{(j,i) \in E} \frac{P(j)}{O_j}, \quad i = 1, \dots, n$$

of n equations in the matrix form as follows

$$\bar{P} = \bar{A}^T \bar{P}$$

The solution \bar{P} is an **eigenvector** corresponding to **eigenvalue** of 1.



PageRank algorithm idea

$$\bar{P} = \bar{A}^T \bar{P}$$

The solution \bar{P} is an **eigenvector** corresponding to **eigenvalue** of 1.

If \bar{A} satisfies certain conditions, then 1 is the largest eigenvalue of \bar{A} and the solution \bar{P} can be found by the **power iteration algorithm**:

- Starting with \bar{P}_0
- Iteratively compute $\bar{P}_i = \bar{A} \bar{P}_{i-1}$
- Until $||\bar{P}_i - \bar{P}_{i-1}||_1 \leq \varepsilon$, where ε is some pre-specified threshold

PageRank algorithm idea

Unfortunately, the **conditions** are not satisfied for the real Web graph.