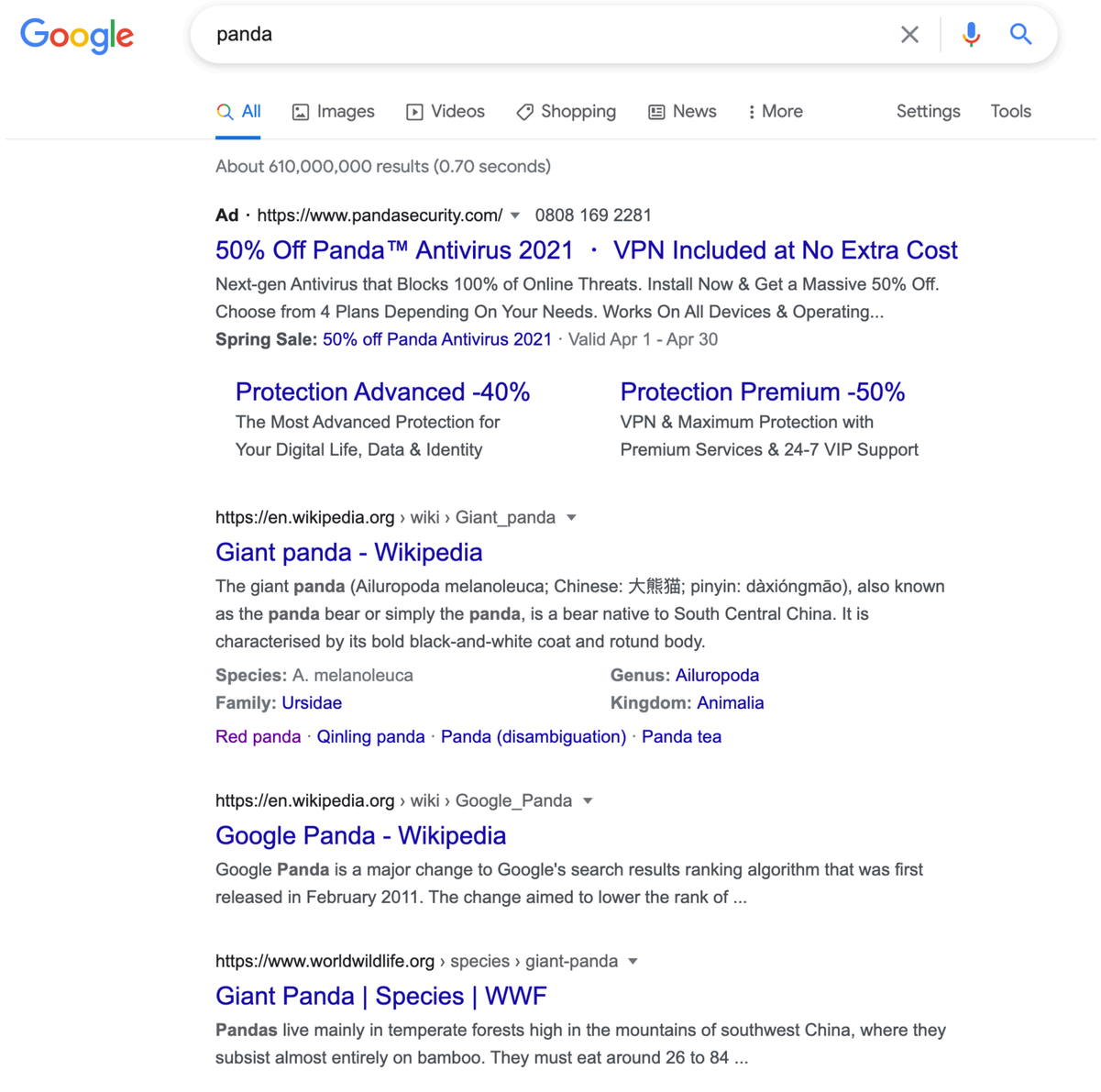
PageRank algorithm: main idea

Procheta Sen

UNIVERSITY OF LIVERPOOL

How does Google decide which pages to show first?

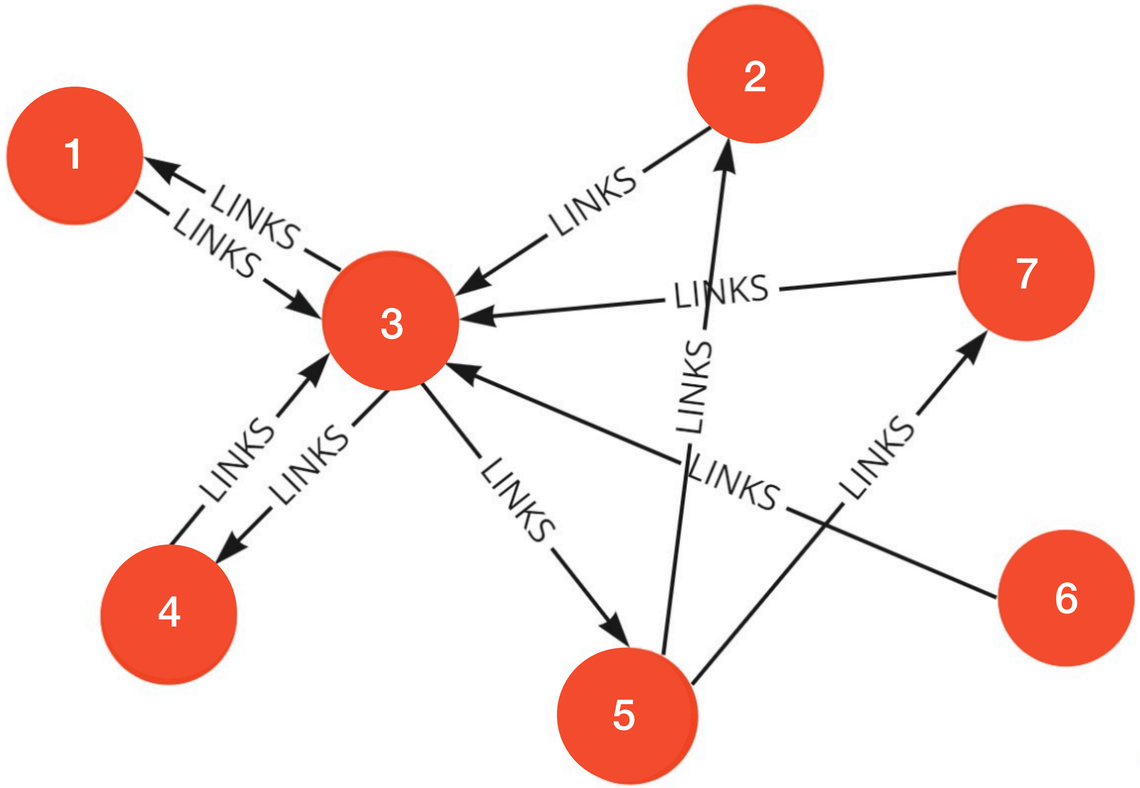


PageRank algorithm

- 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



https://networkx.guide/algorithms/link-analysis/pagerank/

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

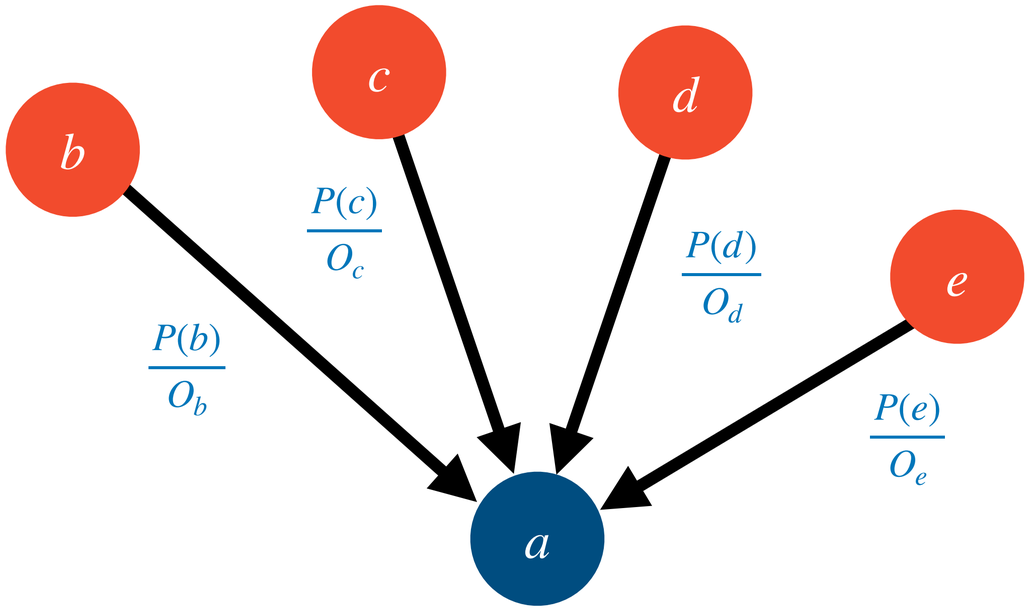
- The importance of page a (i.e. a's PageRank score) is determined by summing up the PageRank scores of all pages that point to a.

- 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

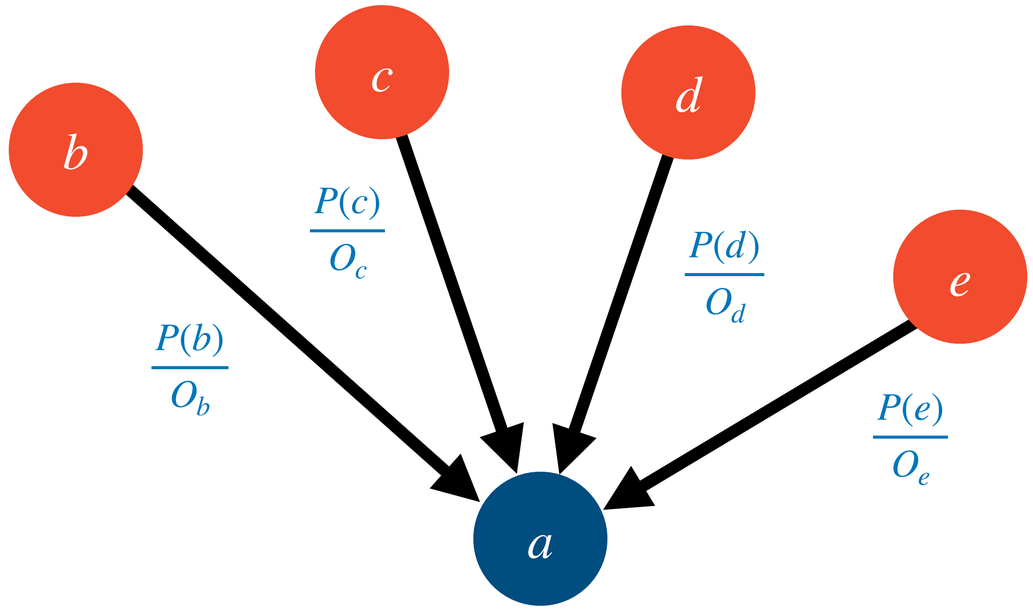
$$P(a) = \sum_{(x,a)\in E} \frac{P(x)}{O_x}$$

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

PageRank algorithm idea

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

$$\overline{P} = (P(1), P(2), \ldots, P(n))^T$$



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

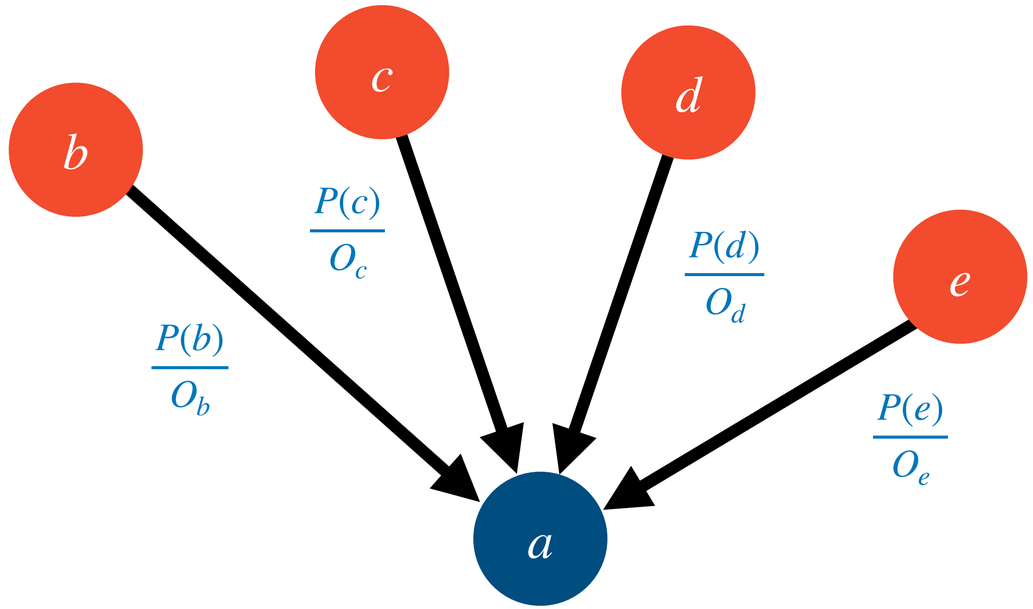
$$\overline{A}_{ij} = \frac{1}{O_i}, \text{ if } (i,j) \in E, \text{ and}$$

$$\overline{A}_{ij} = 0, \text{ otherwise}$$

PageRank algorithm idea

Under this notation we can write the system

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



of n equations in the matrix form as follows

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

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

PageRank algorithm idea

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

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

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

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

PageRank algorithm idea

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