PageRank

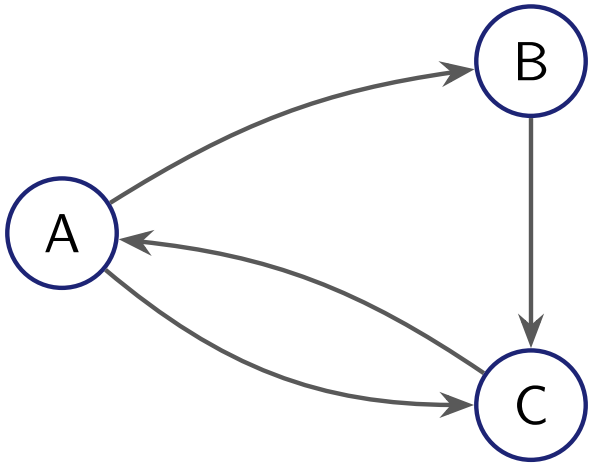
Sehyun Kwon

Mathematical Data Science
Hanyang University ERICA

Finding a page is not enough

A search query may match many pages. Which should appear first?

The words on a page help us decide whether it is relevant. The links between pages provide a different kind of evidence.



We study **PageRank**: a score derived from the link network. It is a ranking signal, not a complete search engine.

Reading: Easley and Kleinberg, *Networks, Crowds, and Markets*, Sections 14.1 and 14.3.

Links as endorsements

Suppose $A \to B$. A link from A directs attention to B. We can model this as an endorsement of B by A.

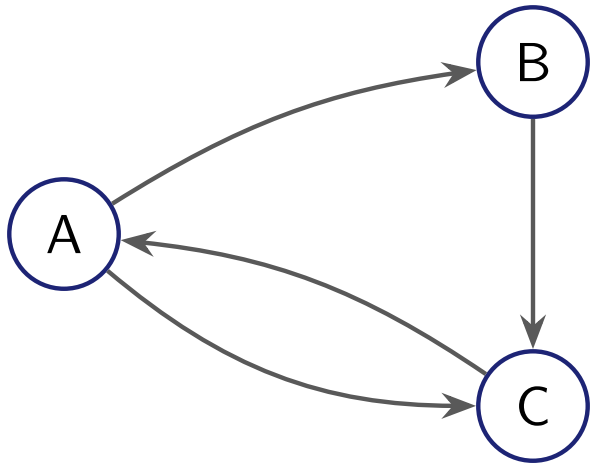


A first ranking rule is simply the **in-degree**. But ten links from influential pages may carry more weight than ten links from obscure pages.

Our assumption: we treat links as endorsements. In reality, a link may express criticism or serve as an advertisement rather than a recommendation.

A first ranking: in-degree

Problem. Rank these pages by the number of links pointing to them.



$$d^-(A) = 1, \qquad d^-(B) = 1, \qquad d^-(C) = 2.$$

In-degree ranks C first and ties A with B.

This rule treats every incoming link equally. It ignores how much attention the linking page receives and how many links it offers.

We will return to this comparison after computing PageRank.

Whose endorsement matters?

We would like a link from an important page to carry more weight.

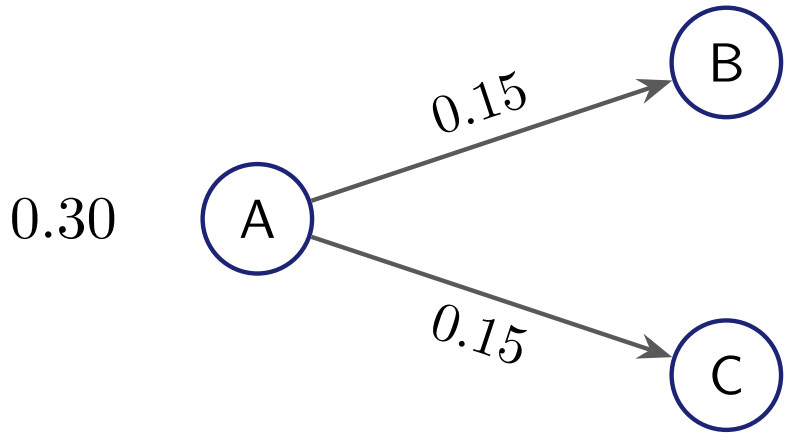
But the importance of that page also depends on links pointing to it.

Rather than deciding all scores at once, start with equal scores and repeatedly redistribute them along links.

Each round uses only the previous round's scores, which are already known.

A fixed amount of score

We distribute a total score of 1 across the pages. Each page divides its current score equally among its outgoing links.



A sends $0.30/2 = 0.15$ along each link. It does not send 0.30 to each destination.

Adding outgoing links divides the same score among more recipients; it does not create additional score.

Why divide by the sender's out-degree?

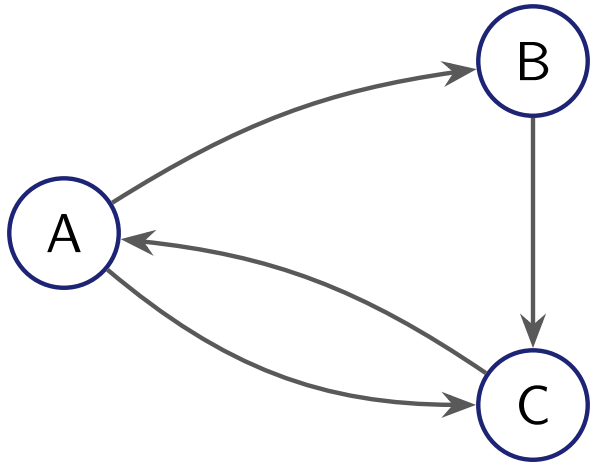
Suppose page u currently has score 0.30.

Outgoing links	Score per link	Total sent
1	0.30	0.30
2	0.15	0.30
3	0.10	0.30

The page splits its score of 0.30 equally among its outgoing links. The total score it sends remains 0.30.

A receiving page adds the contributions from its incoming links. It does not divide that total by its own in-degree.

Our example network



$$N^+(A) = \{B, C\}, \qquad N^+(B) = \{C\}, \qquad N^+(C) = \{A\}.$$

So A splits its score in half, B sends all its score to C, and C sends all its score to A.

Every page has an outgoing link in this example. We will handle pages without one in the next lecture.

The basic update rule

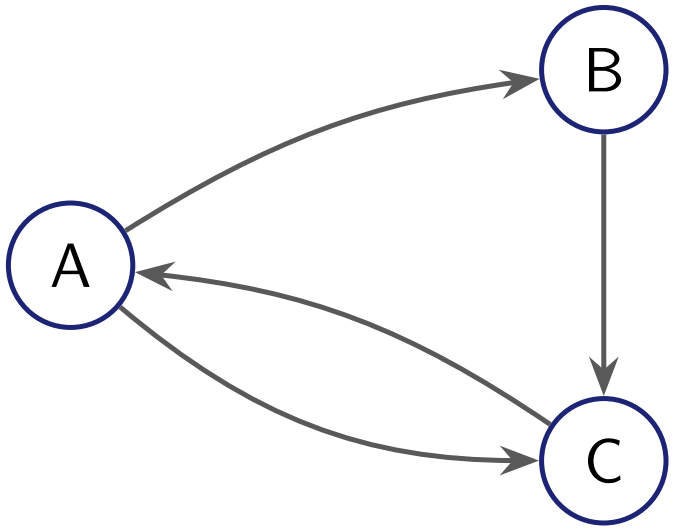
Let $r_t(v)$ be the score of page v after t rounds. For n pages, initialize $r_0(v) = 1/n$.

$$\boxed{r_{t+1}(v) = \sum_{u:\, u \to v} \frac{r_t(u)}{d^+(u)}.}$$

- Sum over pages u that link *to* v.
- Each u contributes its score divided by *its own* out-degree.
- Use the same old-round scores for every page's update.

For now, assume every vertex has positive out-degree.

Writing the rule for our graph



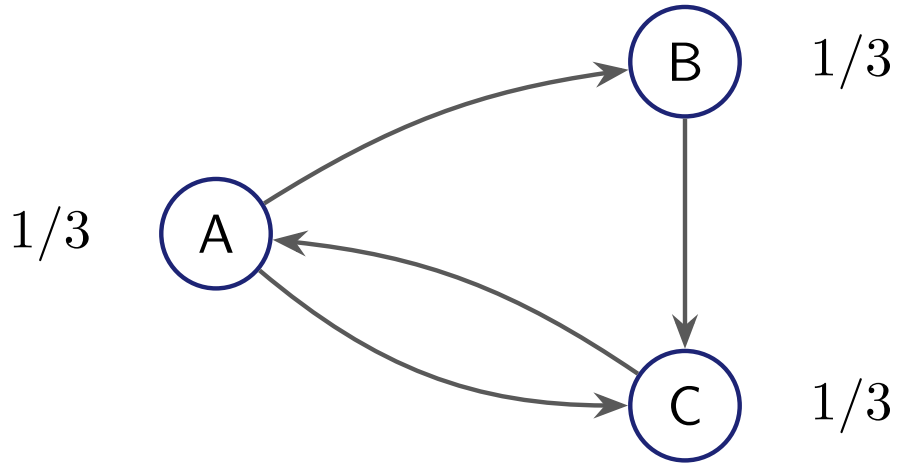
$$r_{t+1}(A) = r_t(C),$$
$$r_{t+1}(B) = \tfrac{1}{2} r_t(A),$$
$$r_{t+1}(C) = \tfrac{1}{2} r_t(A) + r_t(B).$$

Only C links to A. Only A links to B. Both A and B link to C.

The factor $1/2$ appears because A has two outgoing links, not because the receiving page has two incoming links.

Round 1

Problem. Start with $r_0(A) = r_0(B) = r_0(C) = 1/3$. Find each score after one round.



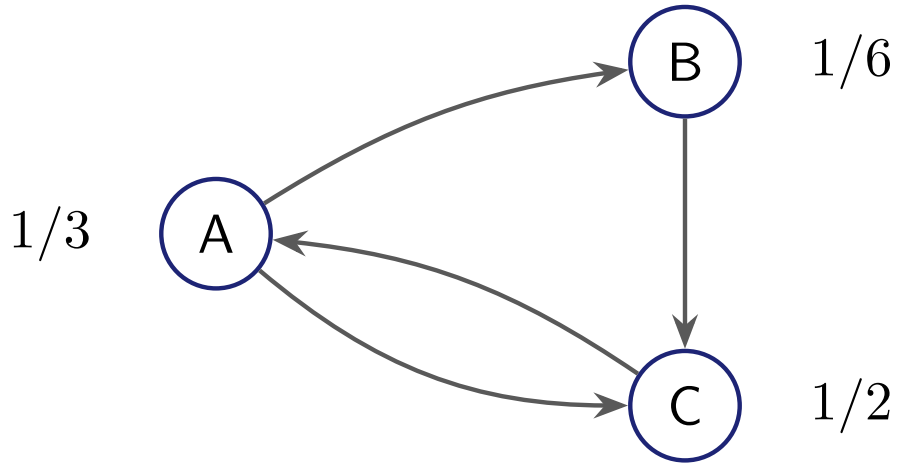
Solution.

$$r_1(A) = \frac{1}{3}, \qquad r_1(B) = \frac{1}{6}, \qquad r_1(C) = \frac{1}{6} + \frac{1}{3} = \frac{1}{2}.$$

The new scores sum to 1. Page C now has the largest score.

Round 2

Problem. Use the scores $r_1(A) = 1/3$, $r_1(B) = 1/6$, and $r_1(C) = 1/2$ to calculate the next round.



Solution.

$$r_2(A) = \frac{1}{2}, \qquad r_2(B) = \frac{1}{6}, \qquad r_2(C) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}.$$

Now A has the largest score. One round is not the final answer.

All updates use the old scores

During round 2, we calculate $r_2(A) = 1/2$ first. Should we immediately use this value to calculate $r_2(B)$?

$$\underbrace{r_2(B) = \tfrac{1}{2} r_1(A) = \tfrac{1}{6}}_{\text{correct}} \qquad \neq \qquad \underbrace{\tfrac{1}{2} r_2(A) = \tfrac{1}{4}}_{\text{mixes rounds}} .$$

Keep two collections of scores: **old** and **new**. Finish the entire new round before replacing the old one.

Otherwise, the result can depend on the order in which the pages are processed.

Why the total stays one

Each page u distributes exactly the score it currently holds:

$$\underbrace{\frac{r_t(u)}{d^+(u)} + \cdots + \frac{r_t(u)}{d^+(u)}}_{d^+(u) \text{ outgoing links}} = r_t(u).$$

Every amount sent is received by some page. Therefore

$$\sum_{v \in V} r_{t+1}(v) = \sum_{u \in V} r_t(u) = 1.$$

No score is created or lost. All scores remain nonnegative.

This check relies on having a destination for every page's score.

Following several rounds

Round t	$r_t(A)$	$r_t(B)$	$r_t(C)$
0	0.3333	0.3333	0.3333
1	0.3333	0.1667	0.5000
2	0.5000	0.1667	0.3333
3	0.3333	0.2500	0.4167
10	0.4062	0.1979	0.3958
20	0.4001	0.1999	0.4001

The scores appear to settle near $(0.4, 0.2, 0.4)$.

This numerical pattern suggests an answer. We can check it directly by applying the update rule once more.

Values are rounded; rounded entries need not sum to exactly one.

An equilibrium score

An **equilibrium** is a score assignment unchanged by one update. For our graph, it must satisfy

$$r(A) = r(C), \quad r(B) = \tfrac{1}{2} r(A), \quad r(C) = \tfrac{1}{2} r(A) + r(B).$$

Let $r(A) = a$. Then $r(B) = a/2$ and $r(C) = a$. Because the scores sum to one,

$$a + \frac{a}{2} + a = 1 \quad \Longrightarrow \quad a = \frac{2}{5}.$$

$$\boxed{r(A) = \frac{2}{5}, \qquad r(B) = \frac{1}{5}, \qquad r(C) = \frac{2}{5}.}$$

Substituting these values into all three update equations leaves them unchanged.

Equilibrium and convergence

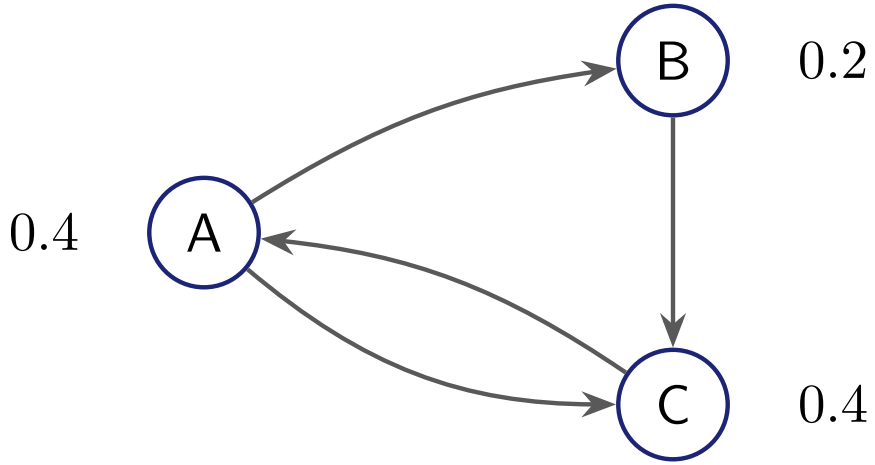
These are two separate questions about the update rule.

- **Equilibrium:** if we start with these scores, does one round leave them unchanged? In other words, do the scores remain unchanged under further updates?
- **Convergence:** do repeated rounds approach those scores from our chosen starting point?

We verified that $(0.4, 0.2, 0.4)$ is an equilibrium. The table also shows the iterates approaching it in this example.

Solving the equilibrium equations alone does not prove convergence. Next lecture, we will see a graph whose iterates keep alternating.

What the scores mean



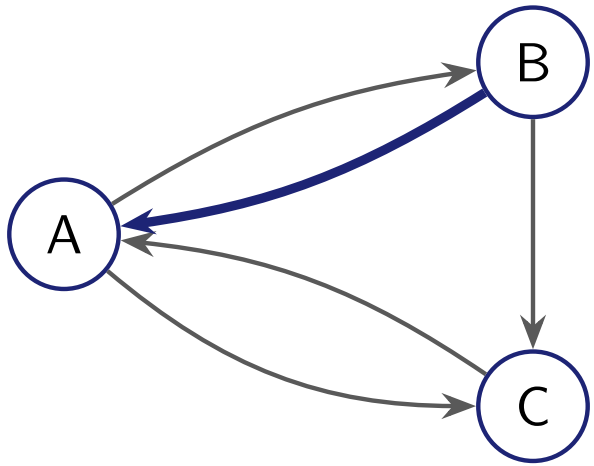
A has only one incoming link, while C has two. Yet they have equal equilibrium scores.

Why? A receives all of C's score, 0.4. Page C receives 0.2 from A and 0.2 from B.

PageRank accounts for **the scores of the linking pages** and **how many outgoing links they share those scores among**.

Changing one link

Problem. Add the link $B \to A$, leaving every other link unchanged. How might this affect the scores of A, B, and C?



Now B divides its score between A and C. Previously, it sent all its score to C.

Only A gains an incoming link. Could the score of B also change, even though no new link points to B?

The changed graph: one round

Solution. Start from the old equilibrium $(0.4, 0.2, 0.4)$. Apply the basic update using the *new* links:

$$r_{\text{new}}(A) = \tfrac{1}{2}(0.2) + 0.4 = 0.5,$$
$$r_{\text{new}}(B) = \tfrac{1}{2}(0.4) = 0.2,$$
$$r_{\text{new}}(C) = \tfrac{1}{2}(0.4) + \tfrac{1}{2}(0.2) = 0.3.$$

Immediately, A gains 0.1 and C loses 0.1. Page B has not changed yet.

In the following round, B receives half of A's new score: $0.5/2 = 0.25$. A link change can affect pages indirectly.

The changed graph: equilibrium

Write $a = r(A)$, $b = r(B)$, and $c = r(C)$. At equilibrium,

$$a = \frac{b}{2} + c, \qquad b = \frac{a}{2}, \qquad c = \frac{a+b}{2}.$$

Hence $b = a/2$ and $c = 3a/4$. Normalization gives

$$a + \frac{a}{2} + \frac{3a}{4} = 1 \quad \Longrightarrow \quad (a, b, c) = \left(\frac{4}{9}, \frac{2}{9}, \frac{1}{3}\right).$$

	A	B	C
Original equilibrium	0.4000	0.2000	0.4000
With $B \to A$	0.4444	0.2222	0.3333

B gains score without gaining an incoming link. Its source A has become more important. The ranking depends on the network beyond direct neighbors.

Next lecture

The basic update rule worked well in our example, but it may not work well on every graph.

Next lecture, we will introduce random walks and use them to improve the PageRank algorithm.