PageRank: The "Flow" Formulation

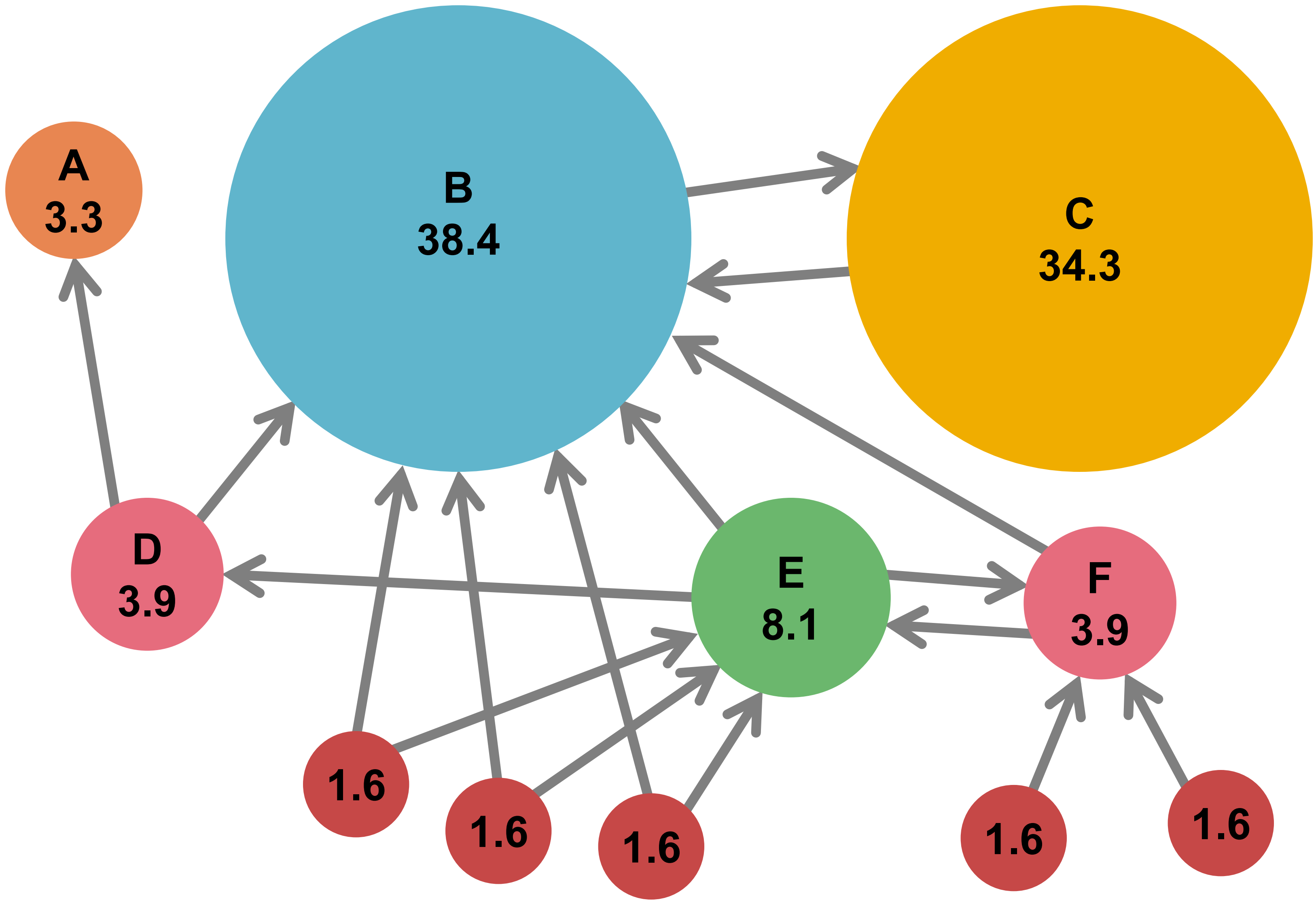
Mining of Massive Datasets
Leskovec, Rajaraman, and Ullman
Stanford University

Links as Votes

- **Idea: Links as votes**
  - **Page is more important if it has more links**
    - In-coming links? Out-going links?
- **Think of in-links as votes:**
  - www.stanford.edu has 23,400 in-links
  - www.joe-schmoe.com has 1 in-link
- **Are all in-links are equal?** 类似于：本文猛不猛，看引用。还是个recursive的
  - Links from important pages count more
  - Recursive question!

Inlinks = Links on other websites that send traffic to your site (进来的).
Inlinks are harder to fake than outlinks 所以通常采用比较多，类似于他引
Outlinks = Links on your site that send people to other sites （出去的）

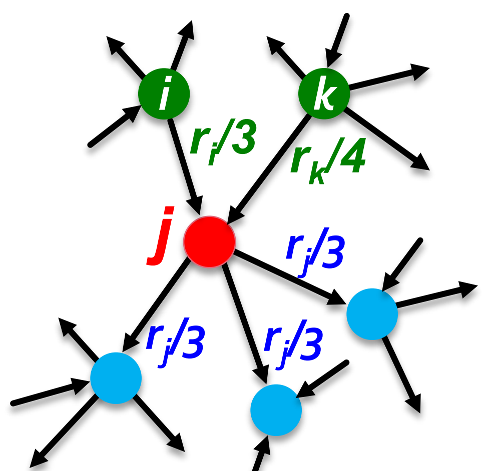
Example: PageRank Scores



Simple Recursive Formulation

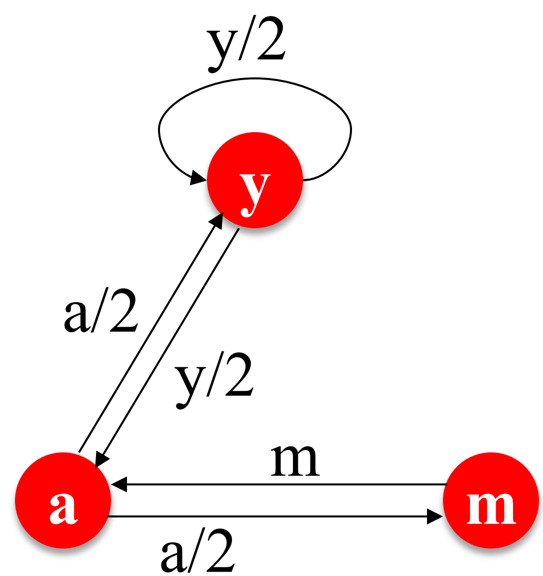
- Each link's vote is proportional to the **importance** of its source page
- If page j with importance r_j has n out-links, each link gets r_j / n votes
- Page j's own importance is the sum of the votes on its in-links

$$r_j = r_i/3 + r_k/4$$



PageRank: The "Flow" Model

The web in 1839



- A "vote" from an important page is worth more
- A page is important if it is pointed to by other important pages
- Define a "rank" r_j for page j

$$r_j = \sum_{i \to j} \frac{r_i}{\mathrm{d}_i}$$

此处的i是所有指向j的i网页的集合

d_i out-degree of node i

"Flow" equations:

$$\mathbf{r}_y = \mathbf{r}_y/2 + \mathbf{r}_a/2$$
$$\mathbf{r}_a = \mathbf{r}_y/2 + \mathbf{r}_m$$
$$\mathbf{r}_m = \mathbf{r}_a/2$$

Solving the Flow Equations

Flow equations:

$$r_y = r_y/2 + r_a/2$$
$$r_a = r_y/2 + r_m$$
$$r_m = r_a/2$$

- **3 equations, 3 unknowns, no constants**
  - No unique solution
  - All solutions equivalent modulo the scale factor
- **Additional constraint forces uniqueness:**
  - $\boldsymbol{r_y + r_a + r_m = 1}$
  - **Solution:** $\boldsymbol{r_y = \frac{2}{5}, r_a = \frac{2}{5}, r_m = \frac{1}{5}}$
- **Gaussian elimination method works for small examples, but we need a better method for large web-size graphs**
- **We need a new formulation!**