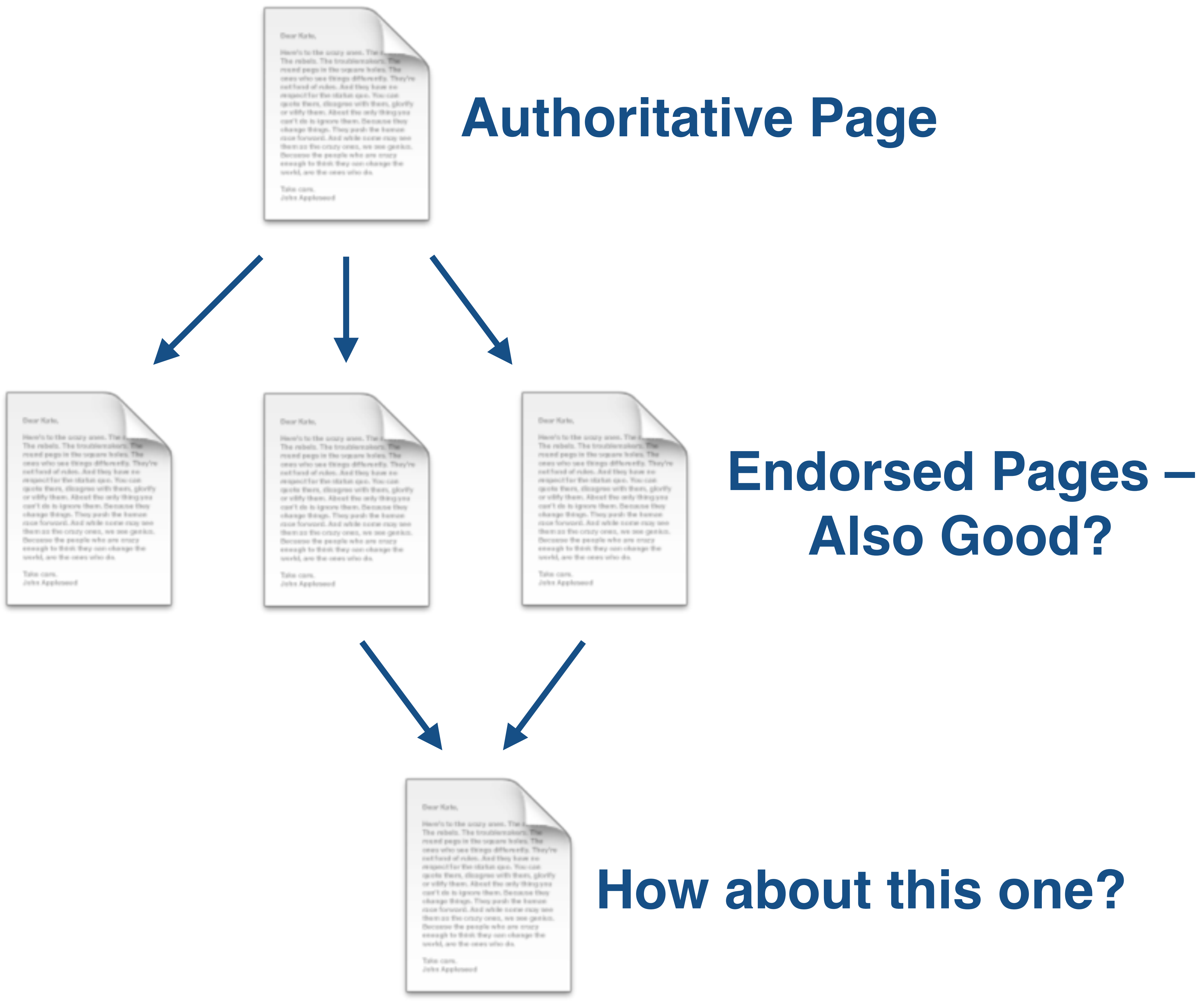
PageRank

Document Understanding, session 3

Link Structure of the Web

The Internet is a graph of web pages that link to each other. In most cases, these links can be seen as endorsements by a page author of the content on some other page.

Building on this assumption, we can create a ranking score for web pages based purely on how many endorsements they receive from high-quality pages. This is PageRank.

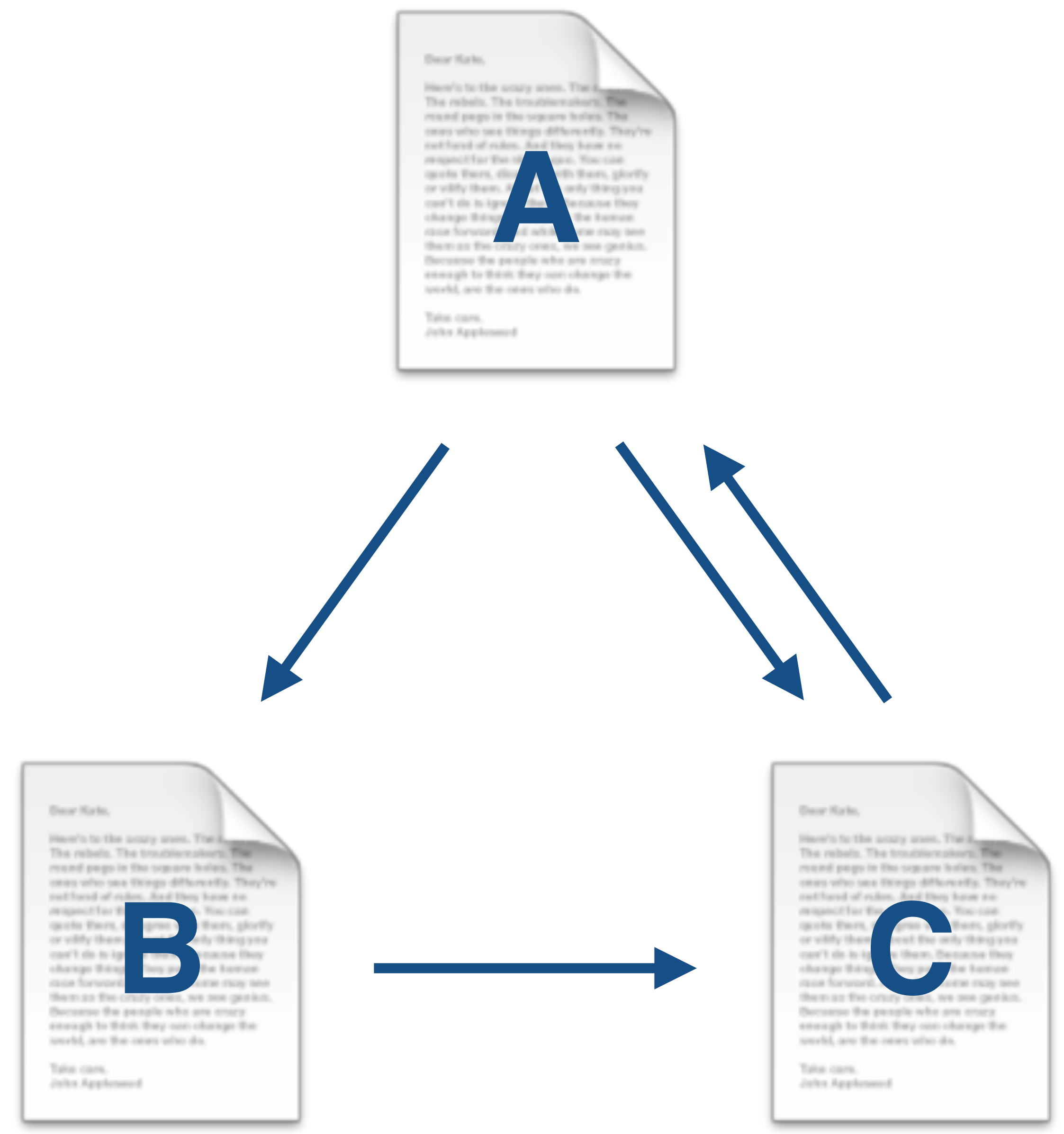


The Random Surfer

Consider the following random experiment:

Start at a web page chosen uniformly at random. At each time t, flip a biased coin (e.g. probability of heads is λ). If the coin comes up heads, follow a link chosen at random from the current page. Otherwise, choose a new page uniformly at random.

The PageRank of a particular page is the expected fraction of visits the surfer would make to it.

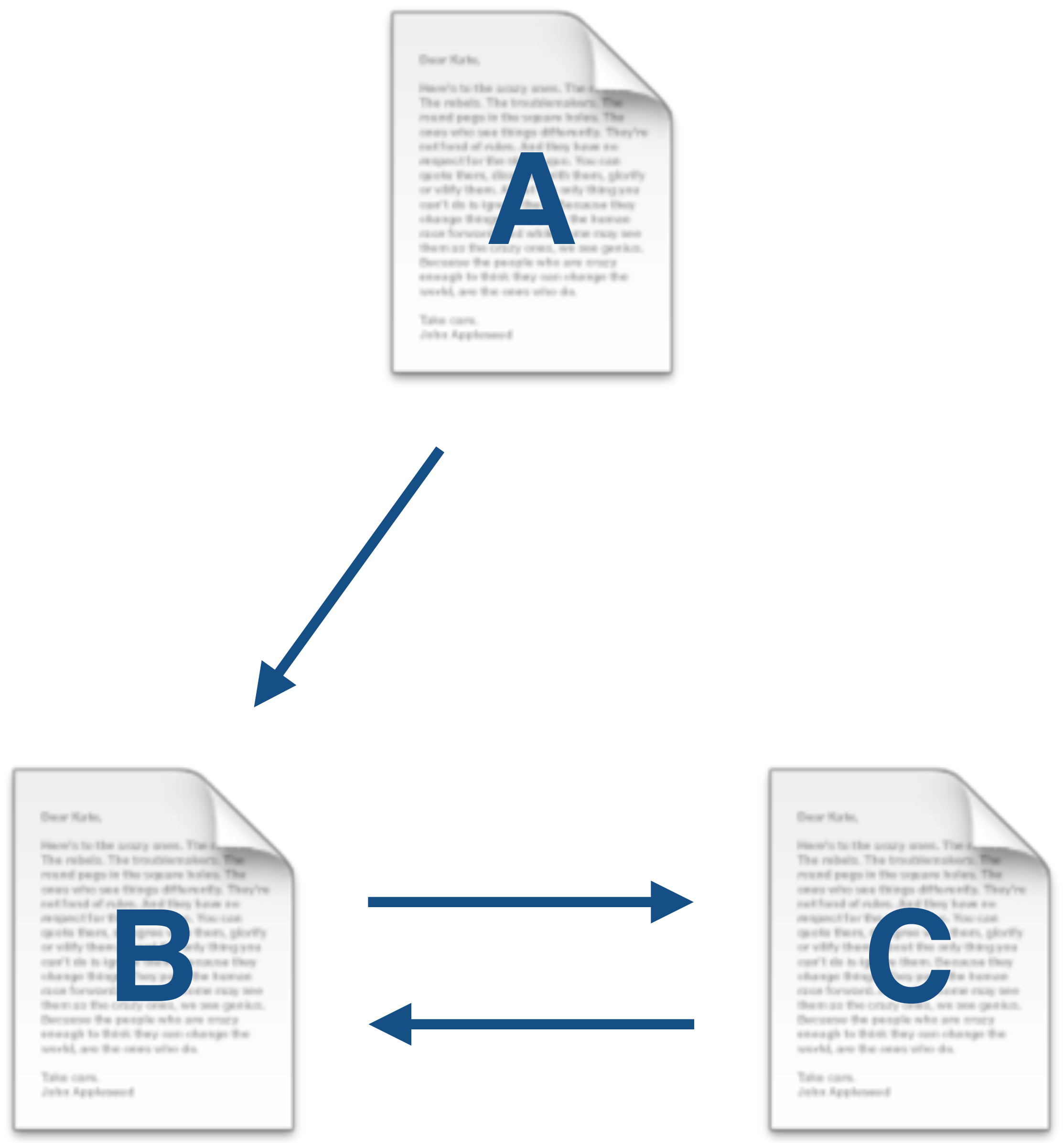


$$PR(C) \approx \frac{1}{2} PR(A) + \frac{1}{1} PR(B)$$

Teleportation in PageRank

The surfer's ability to choose a random page instead of following a link is called *teleportation*.

The surfer needs to teleport in order to escape from dead-end link cycles, and from pages with no out-links.



A trap for naive surfers

Calculating PageRank

More precisely, the PageRank of a page is:

$$PR(u) = \frac{\lambda}{N} + (1-\lambda) \sum_{v \in inlinks(u)} \frac{PR(v)}{|outlinks(v)|}$$

One way to calculate it is to initialize all PageRanks to $1/N$, then iteratively update each page in turn until the process converges.

A standard convergence test is when $\frac{\|new - old\|}{N} < \tau$ for some $\tau \leq 1$. Smaller values of τ are more accurate but take longer to converge.

```
procedure PageRank(G)
      ▷ G is the web graph, consisting of vertices (pages) and edges (links).
   (P, L) ← G                              ▷ Split graph into pages and links
   I ← a vector of length |P|              ▷ The current PageRank estimate
   R ← a vector of length |P|              ▷ The resulting better PageRank estimate
   for all entries I_i ∈ I do
      I_i ← 1/|P|                          ▷ Start with each page being equally likely
   end for
   while R has not converged do
      for all entries R_i ∈ R do
         R_i ← λ/|P|  ▷ Each page has a λ/|P| chance of random selection
      end for
      for all pages p ∈ P do
         Q ← the set of pages such that (p, q) ∈ L and q ∈ P
         if |Q| > 0 then
            for all pages q ∈ Q do
               R_q ← R_q + (1 − λ)I_p/|Q|   ▷ Probability I_p of being at page p
            end for
         else
            for all pages q ∈ P do
               R_q ← R_q + (1 − λ)I_p/|P|
            end for
         end if
         I ← R                             ▷ Update our current PageRank estimate
      end for
   end while
   return R
end procedure
```

PageRank with Linear Algebra

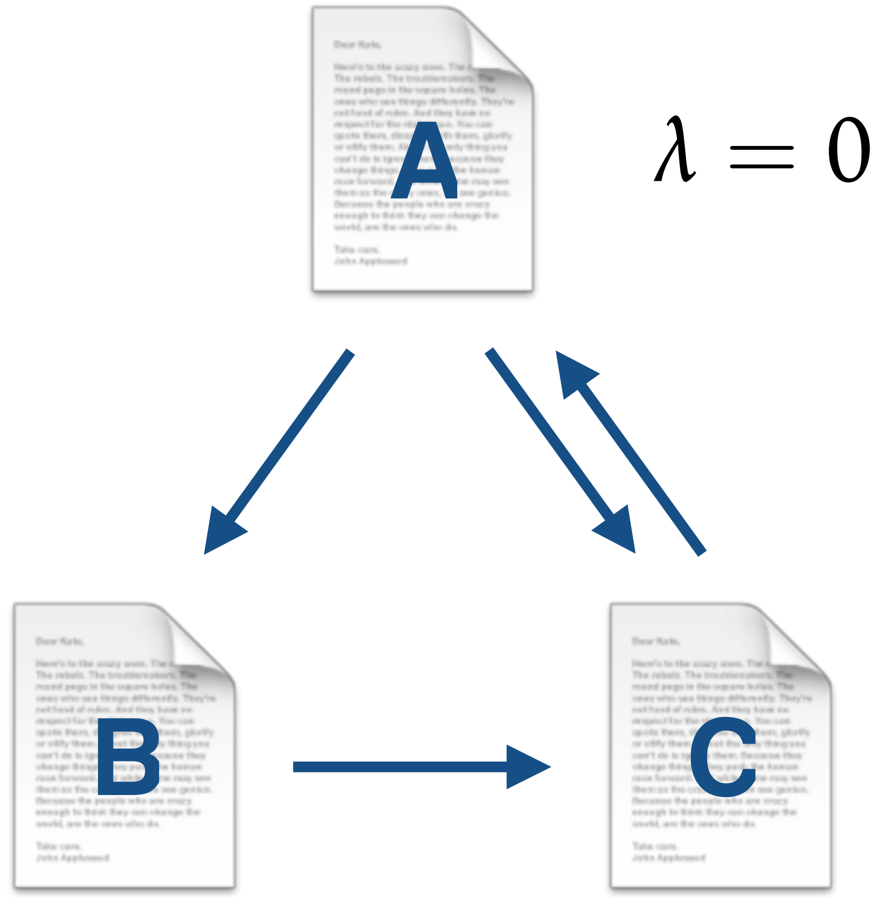
PageRank can also be calculated using the *transition probability matrix* P of the random experiment.

$P_{i,j} \in (0,1)$ is prob. of transition from i to j

$$\forall i, \sum_{j=1}^{N} P_{i,j} = 1$$

The largest eigenvalue of P is 1. The corresponding left eigenvector gives the PageRank of each page.

$$P_{i,j} = \begin{cases} \frac{1}{N} & \text{if } |outlinks(i)| = 0 \\ \frac{\lambda}{N} + \frac{1-\lambda}{|outlinks(i)|} & \text{else if } j \in outlinks(i) \\ \frac{\lambda}{N} & \text{else} \end{cases}$$



$\lambda = 0.3$

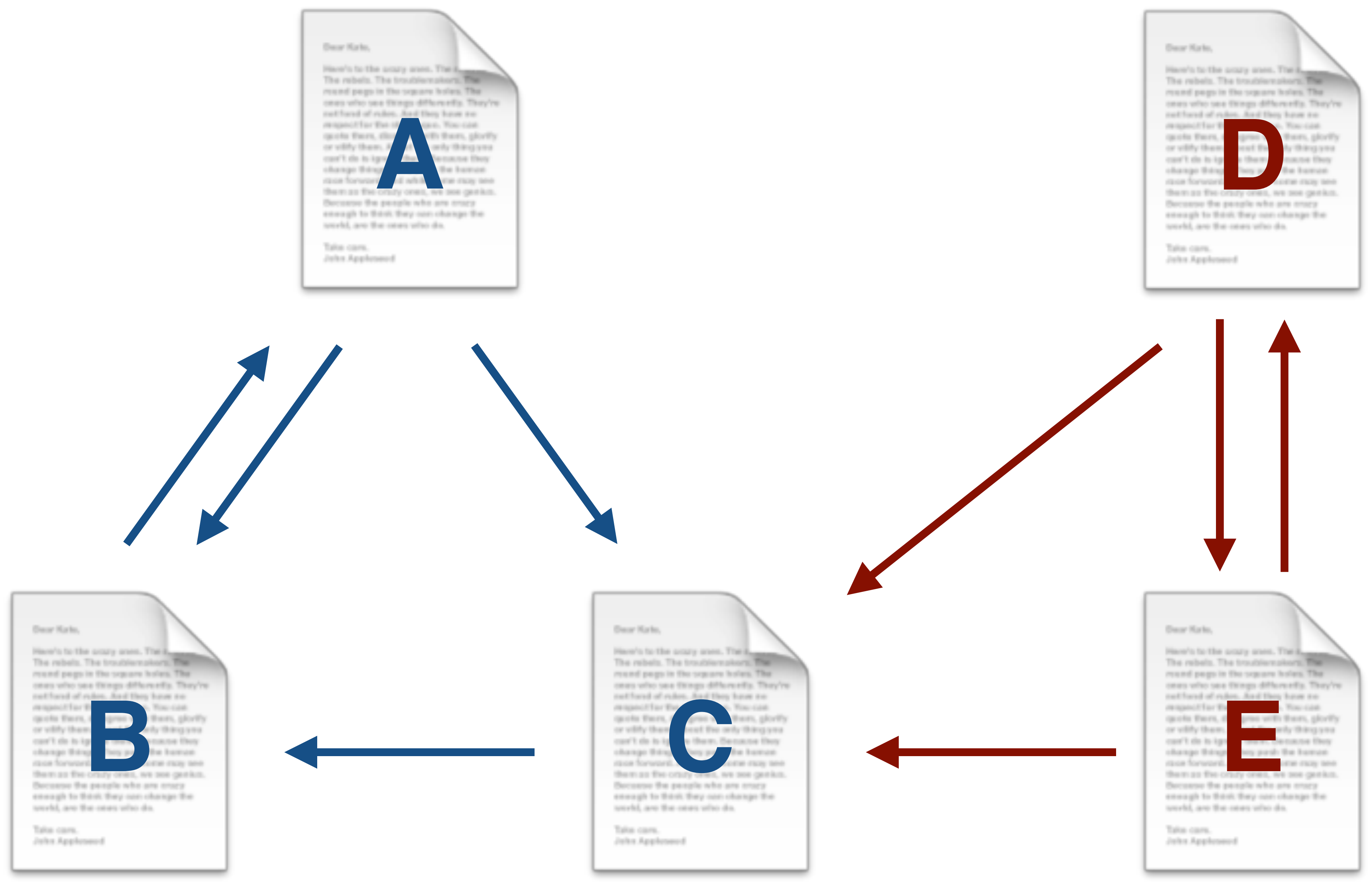
$$\Rightarrow \begin{pmatrix} 2/20 & 9/20 & 9/20 \\ 1/10 & 1/10 & 8/10 \\ 8/10 & 1/10 & 1/10 \end{pmatrix}$$

Problems with PageRank

The original implementation of PageRank has several known flaws. Importantly, it can be easily manipulated.

- Link farms – large collections of inexpensive sites can be created to artificially boost a page's rank by linking to it.

- Link spam – blog comments can link to an unrelated page, causing the blog to artificially "endorse" the page.



A link farm: D and E unfairly boost C's PageRank.

Wrapping Up

PageRank is a query-independent signal of a page’s quality, based on endorsements by other pages online.

It has some issues in its original form, but successive generations have removed some of these issues.

Next, we’ll see an updated form of PageRank which attempts to calculate page quality for a particular user.

Personalized PageRank

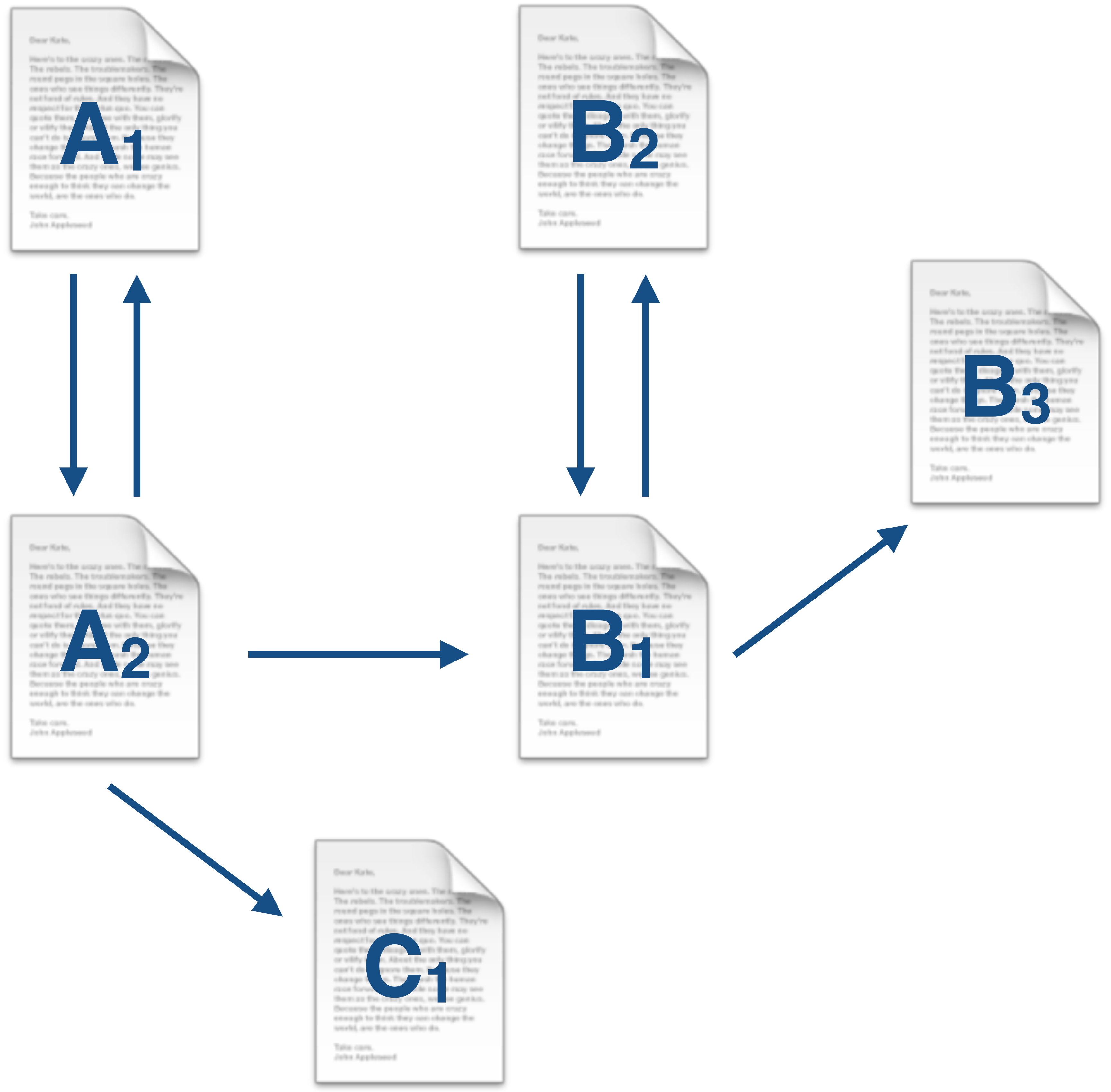
Document Understanding, session 4

Conditional PageRank

The original PageRank score is a distribution over the entire Internet.

We are often interested in quality scores for more restricted subsets of the Internet, e.g. for pages on a particular topic.

The fundamental trick is to modify the teleportation probability and then follow links as usual.



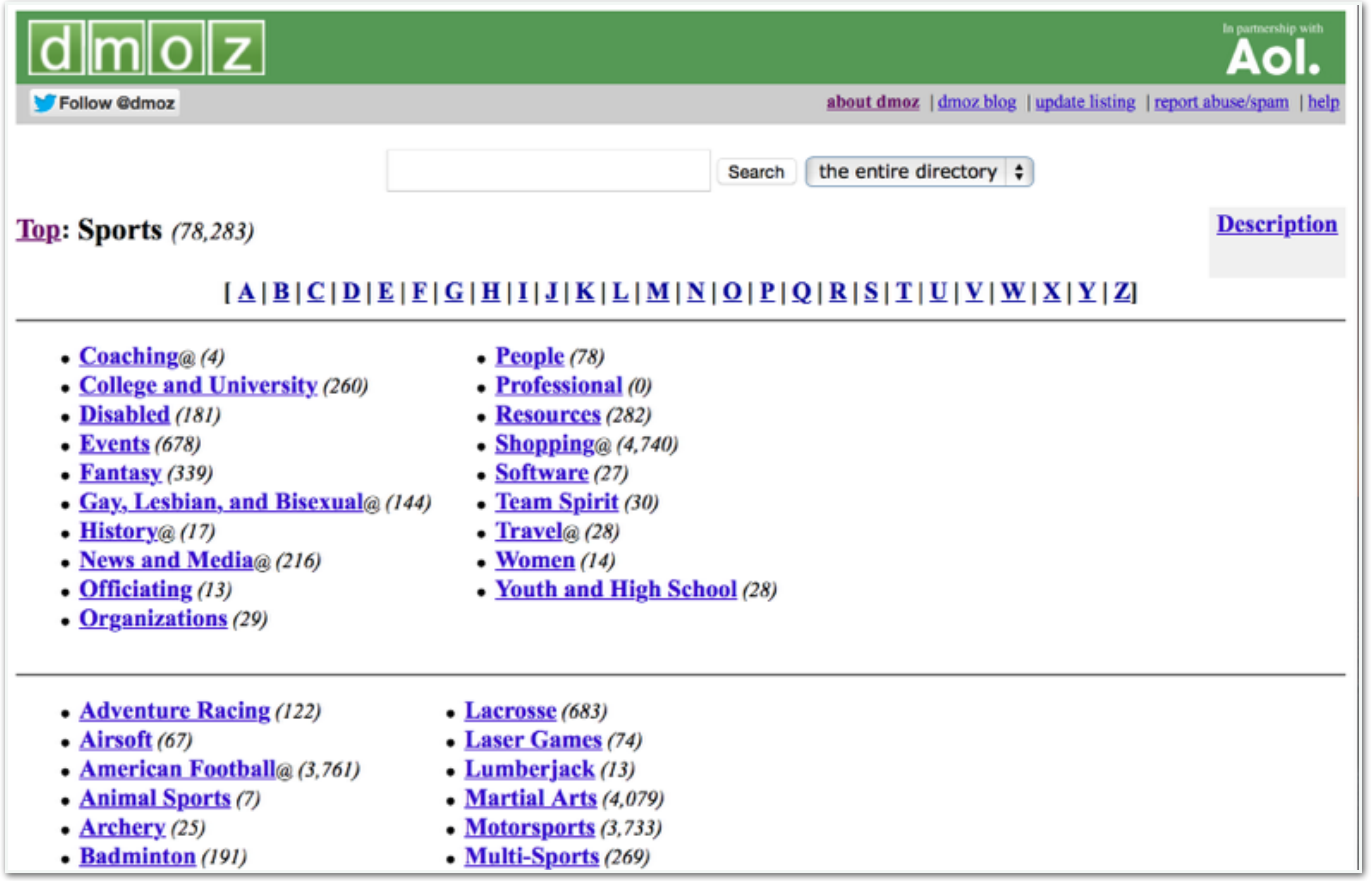
Pages with Topic Labels

Obtaining Page Topic Labels

Topic labels can be obtained from an Internet directory such as dmoz.org or yahoo.com.

Topics can also be inferred using semi-supervised learning: given some labels, we can calculate the most probable topic for unlabeled pages.

We don’t need accurate topic labels for all pages; we will follow links to unlabeled pages.



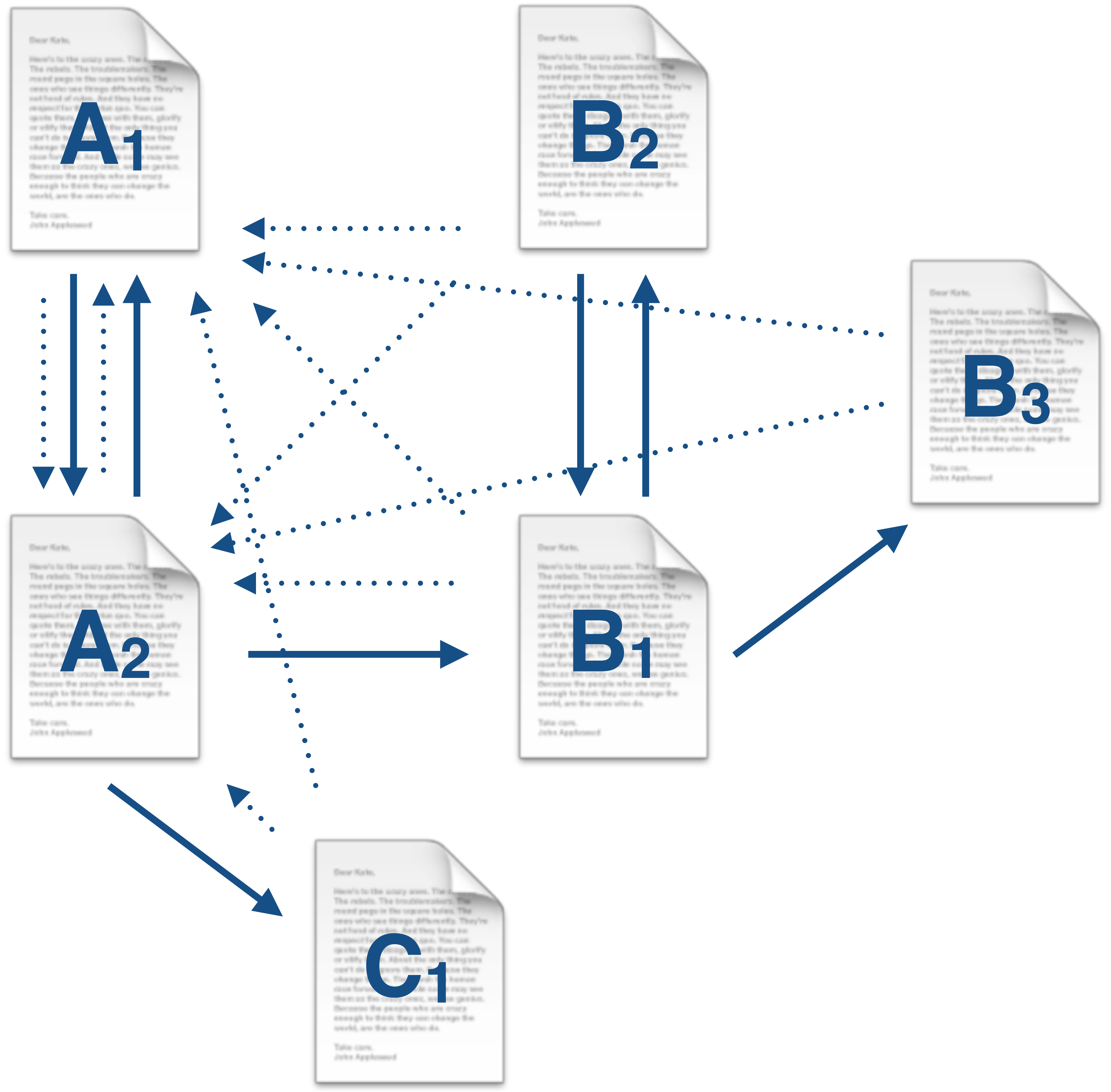
The Open Directory Project

Topic-specific PageRank

Once we have our topic labels, we modify PageRank teleportation to teleport only to the set T of pages with the specified topic t.

Some set $Y \supseteq T$ of pages will have a steady-state PageRank distribution from this process.

The pages in Y have topic-specific PageRank scores for the topic, π_t.



Dotted edges represent teleportation options

Mixing Topics

Suppose a user is interested multiple topics. We can compute a *Personalized PageRank* by teleporting with a distribution according to their interests.

- For instance, 60% of the time we teleport to a sports page and 40% of the time to a politics page.

Recalculating PageRank for each user is prohibitively expensive, but it turns out we don't have to.

The final distribution is just a linear combination of topic-specific PageRank scores: $0.6\pi_s + 0.4\pi_p$.

Does Personalization Help?

Personalized PageRank scores make intuitive sense, but it's not clear that they help much. They tend not to be used in practice due to several concerns.

- Privacy – A detailed log of users' web page preferences can reveal sensitive information about their political opinions, income levels, etc.

- Users change – People gain and lose interests over time, and it isn't clear how to update models. They also run queries related to new topics, and a personalized model might mislead the search engine.

- Clear queries don't need it – If the information need of the query is clear enough, we don't need this kind of topic-based help to perform well.

Wrapping Up

Topic and individual based PageRank scores seem a promising avenue for improving performance of certain queries. However, it’s not clear how to best put them to use in real world situations.

Next, we’ll continue exploring web page topics by learning how to infer topics from the document text alone.