Carnegie Mellon University

95-885 Data Science and Big Data

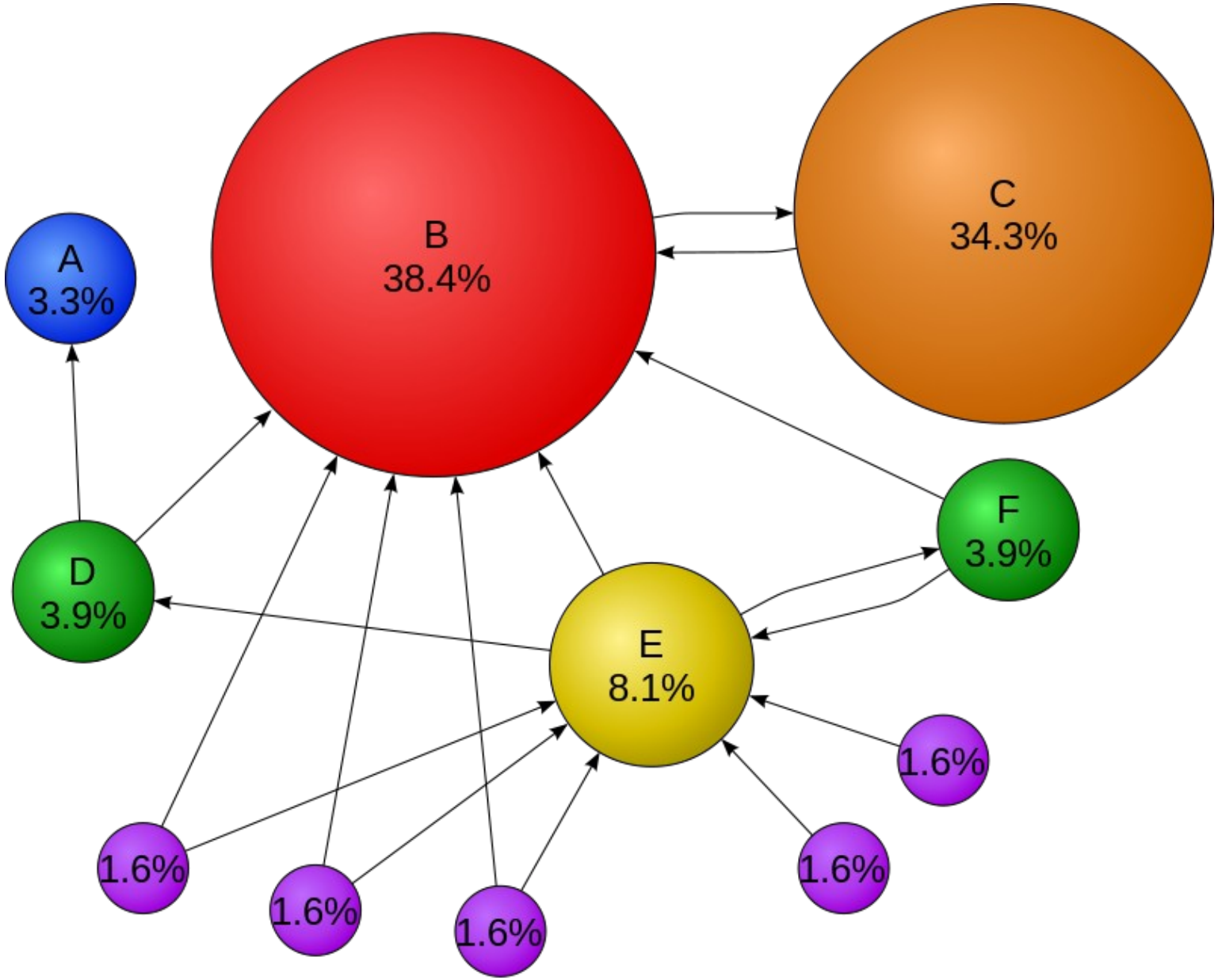
PageRank with MapReduce

Agenda

- The essence of the algorithm
- Simulating with a spreadsheet
- Expressing with a mapper and reducer

THE INTUITION BEHIND THE PAGERANK ALGORITHM

A Sample Set of Interconnected Web Pages



Pagerank (page)
==
Probability of navigating to that page

If you sum all the probabilities on the adjacent figure you will get 100
3.3 + 38.4 + … + 1.6

The Intuition behind the PageRank algorithm

- A -> B means A thinks B is worth something
- A page is important if it is pointed to by other important pages
  - e.g., a link to a page from Wikipedia etc
  - the “slashdot effect”
- Note that this measure of importance is calculated *independent* of the *content* of the page
- “wisdom of the crowds”

Quantifying the Intuition

- We get to a page in one of two ways

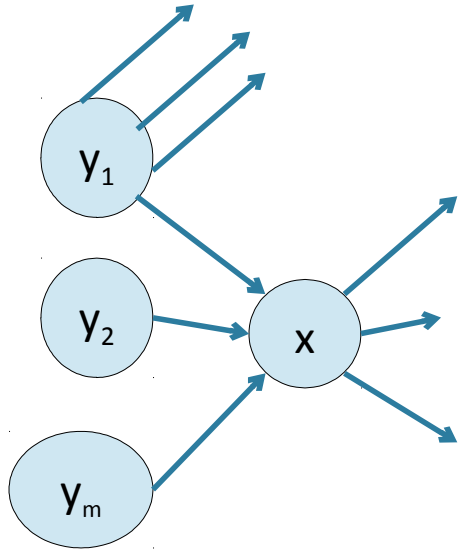
1. By randomly jumping to it
2. By navigating (clicking) to it from the page we currently are on

- Hence the probability of getting to a page x is:
  - P(x) = P(randomly jumping to the page) + P(navigating to it by clicking a link)
- Suppose the probability of deciding to navigate from a page i.e., click a link on the page is β, then the probability of not clicking a link i.e., randomly jumping to a page is 1-β
- If there are a total of N pages then,
  - P(randomly jumping to a page) = $\frac{1-\beta}{N}$

Probability of Navigating to a Page

- Suppose page y_1 links to page x
- Then the probability of navigating to x from y_1 is:
  - $PR(y_1) * 1/out(y_1)$
  - where $PR(y_1)$ is the probability of getting to y_1

and $out(y_1)$ is the number of outlinks from y_1



- We may have several 'y's linking to x so

PR(x) is

  - $PR(y_1)/out(y_1) + \ldots PR(y_m)/out(y_m) +$

- The probability of deciding to navigate in the first place is β
- So the probability of navigating to x from any of the ys (pages that link to x) is: $\beta * \sum_{y->x} \frac{PR(y)}{out(y)}$

Combining the two pieces together we have

$$PR(x) = \frac{1-\beta}{N} + \beta * \sum_{y->x} \frac{PR(y)}{out(y)}$$

Probability of clicking on a link

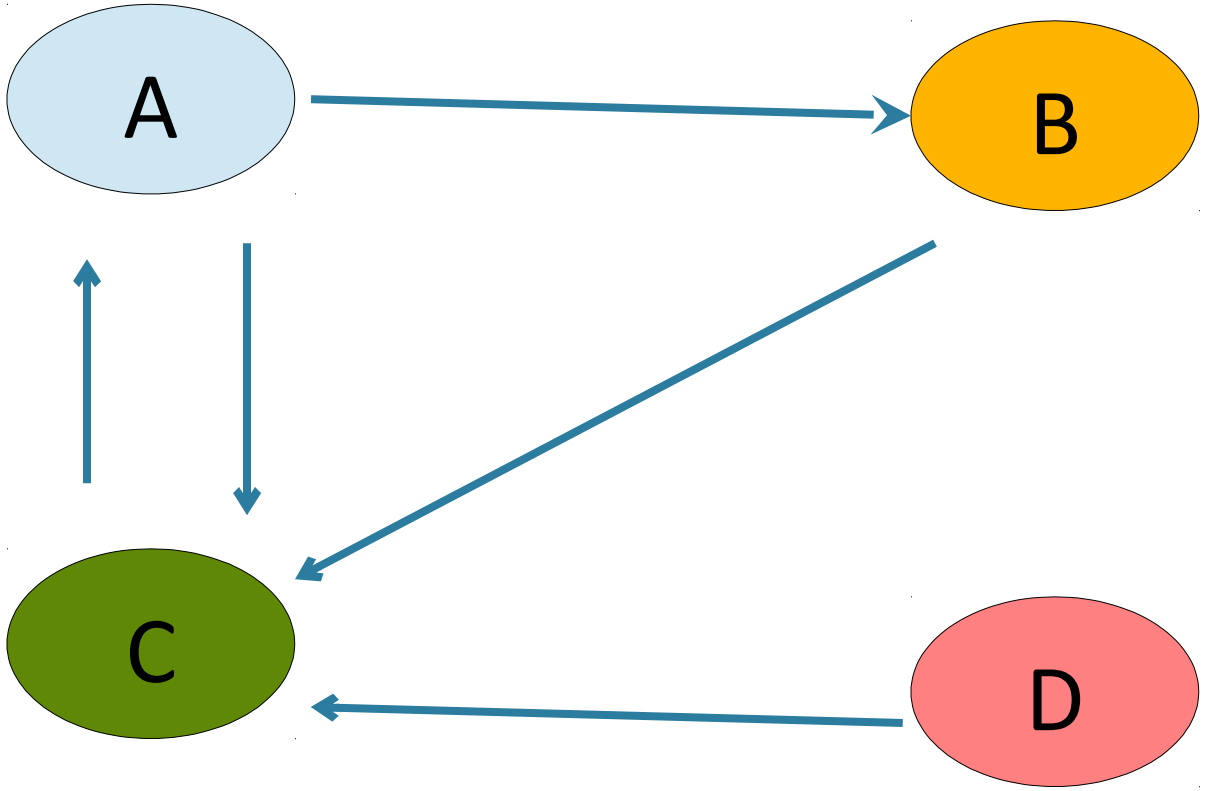
Total number of pages

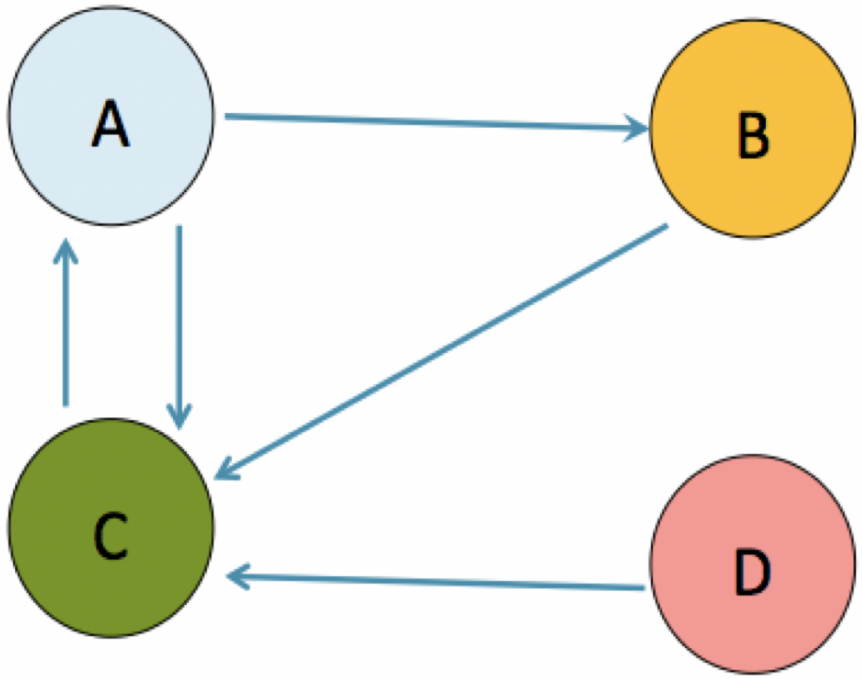
P(x) = P(randomly jumping to the page) + P(navigating to it by clicking)

PAGERANK ON A SPREADSHEET

Spreadsheet available on Canvas schedule

Spreadsheet: Working through the PageRank Algorithm

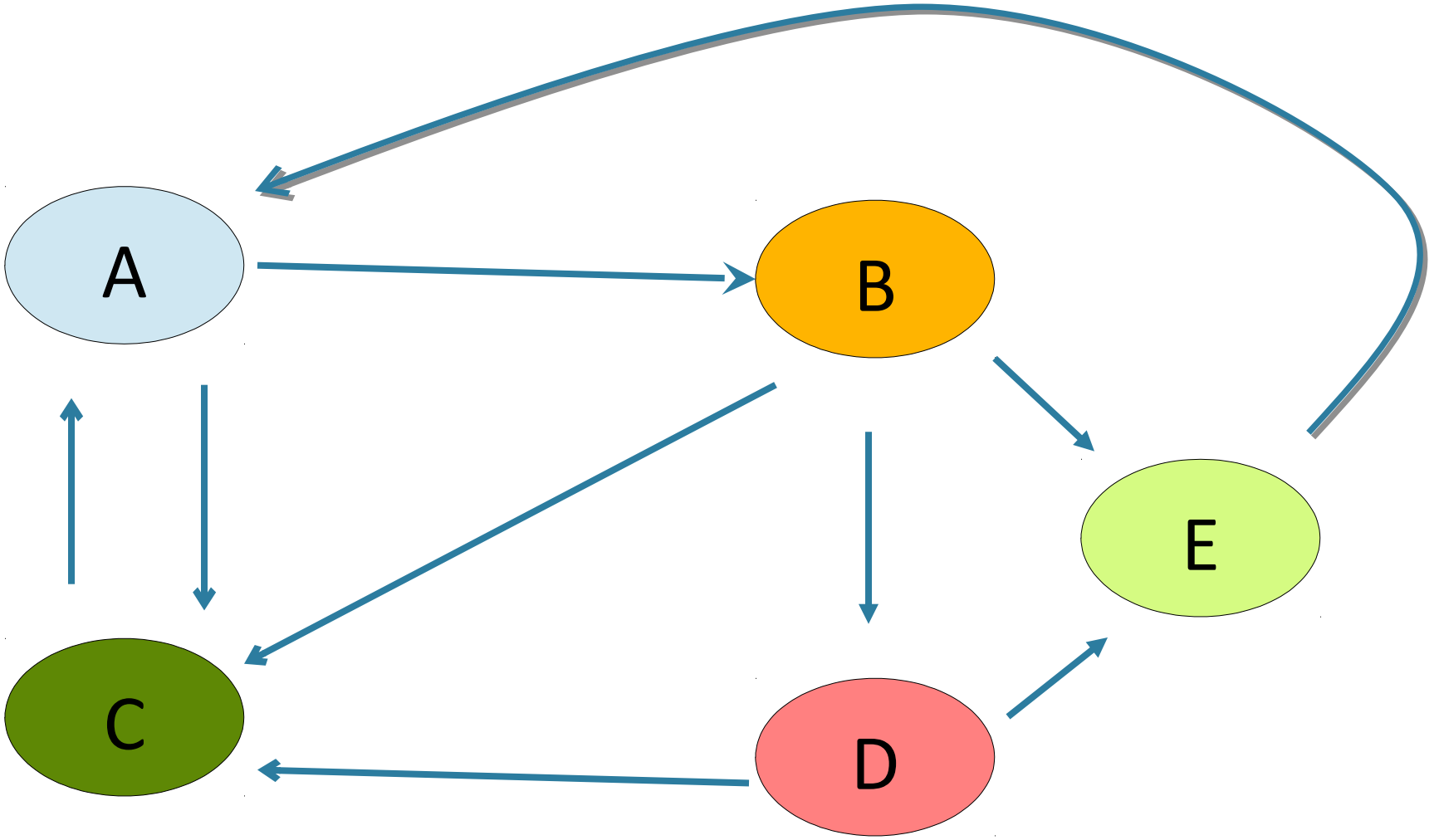


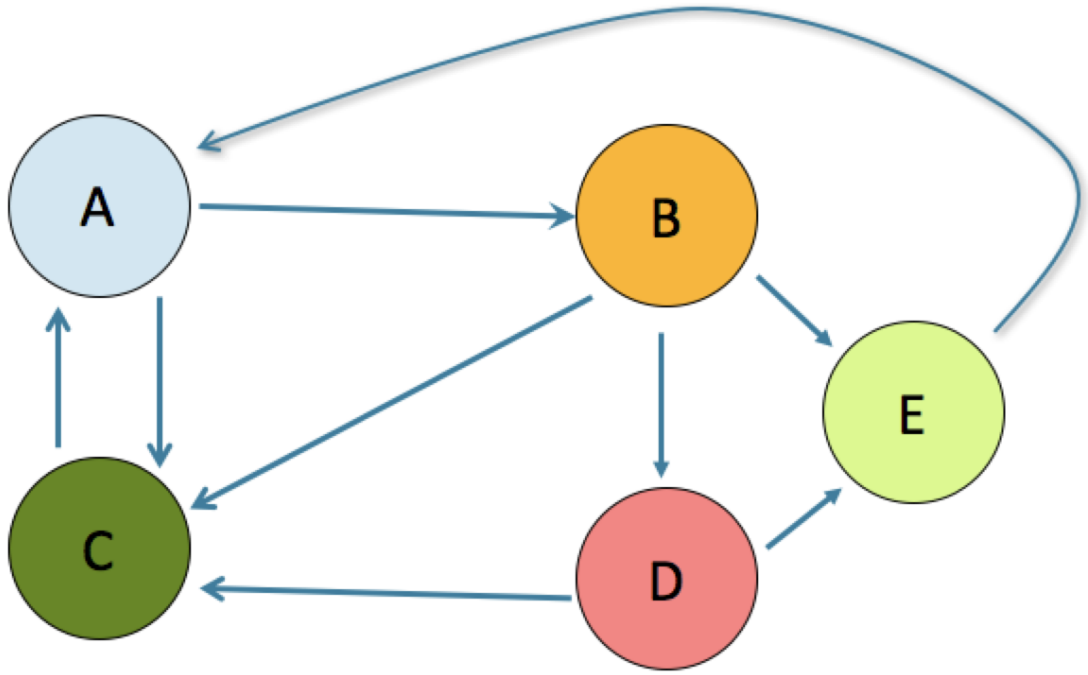


	beta	N
	0.85	4

A	B	C	D		
25%	25%	25%	25%	100%	0
25%	14%	57%	4%	100%	1
52%	14%	30%	4%	100%	2
29%	26%	41%	4%	100%	3
39%	16%	41%	4%	100%	4

Exercise





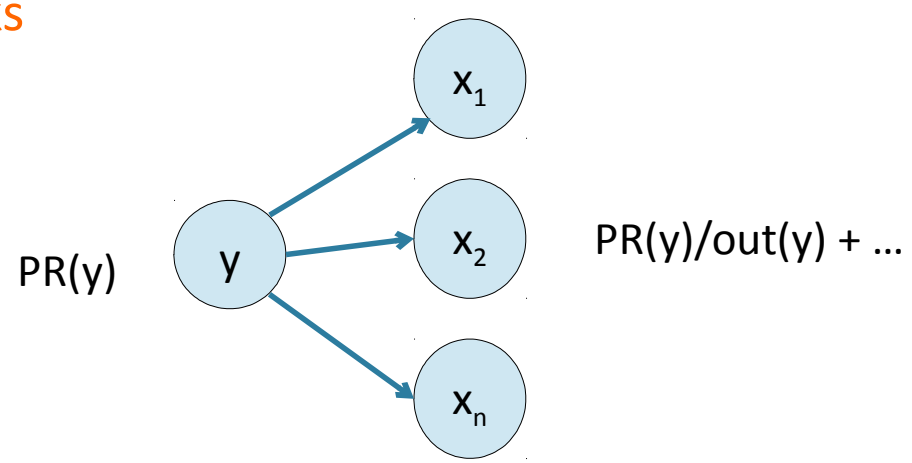
	beta	N
	0.85	5

A	B	C	D	E		
20%	20%	20%	20%	20%	100%	0
37%	12%	26%	9%	17%	100%	1
39%	19%	26%	6%	10%	100%	2
33%	20%	28%	8%	11%	100%	3
36%	17%	26%	9%	12%	100%	4

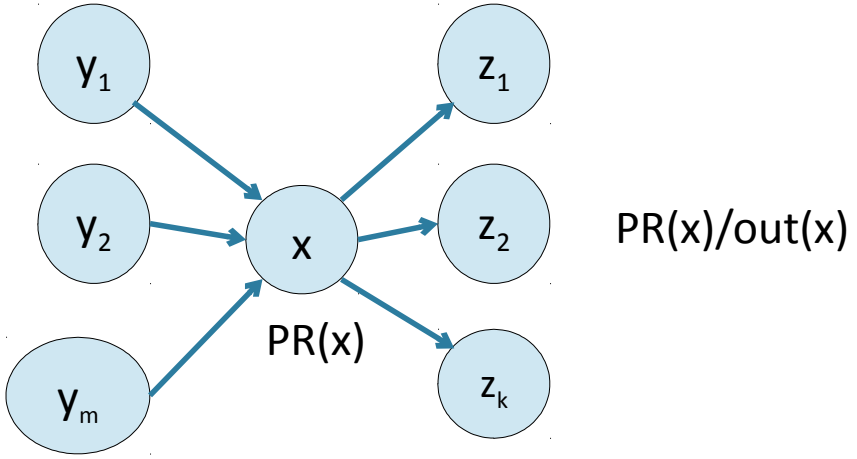
PAGERANK: MAPPER AND REDUCER

Expressing PageRank with MapReduce

- Mapper $\langle y, \{x_1, x_2, \ldots, x_n\} \rangle$ node, out-links
  - for i =1..n, yield $\left\langle x_i, \frac{PR(y)}{out(y)} \right\rangle$
  - yield $\langle y, \{x_1, x_2, \ldots, x_n\} \rangle$



- Reducer $\left\langle x, \left\{ \frac{PR(y_1)}{out(y_1)}, \ldots, \frac{PR(y_m)}{out(y_m)}, \{z_1, \ldots, z_k\} \right\} \right\rangle$ node, ΔPR from in-links
  - compute $PR(x) = \frac{1-\beta}{N} + \beta * \sum_{y->x} \frac{PR(y)}{out(y)}$
  - for j=1..k, yield $\left\langle z_j, \frac{PR(x)}{out(x)} \right\rangle$ $\langle x, \{z_1, z_2, \ldots, z_k\} \rangle$



PR(x) = […] + PR(y1)/out(y1) + ….

Expressing PageRank with MapReduce

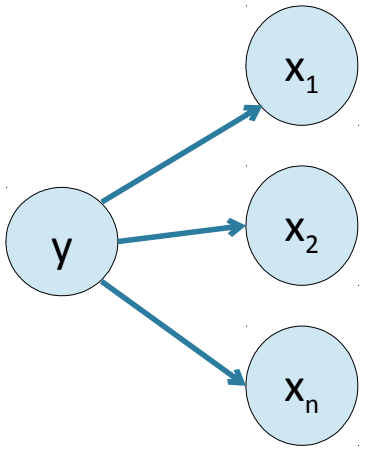
- Mapper

- Reducer

Expressing PageRank with MapReduce

- Mapper

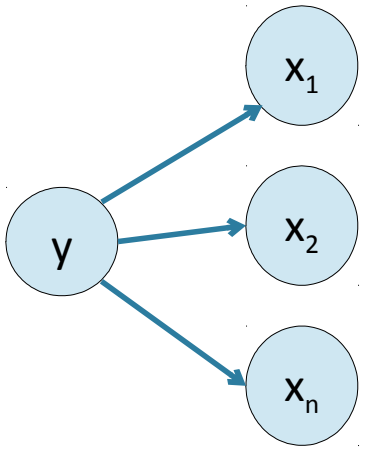
node, out-links



- Reducer

Expressing PageRank with MapReduce

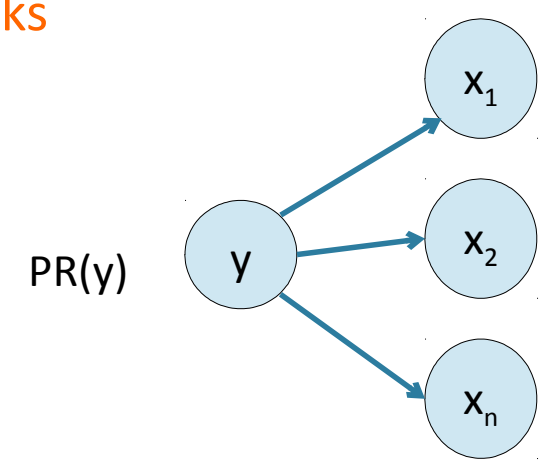
- Mapper $\langle y, \{x_1, x_2, \ldots, x_n\} \rangle$ node, out-links



- Reducer

Expressing PageRank with MapReduce

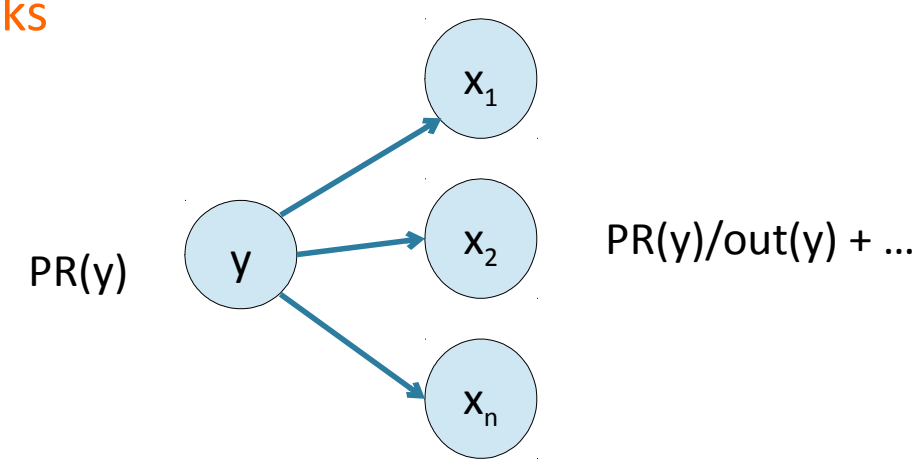
- Mapper $\langle y, \{x_1, x_2, \ldots, x_n\} \rangle$ node, out-links



- Reducer

Expressing PageRank with MapReduce

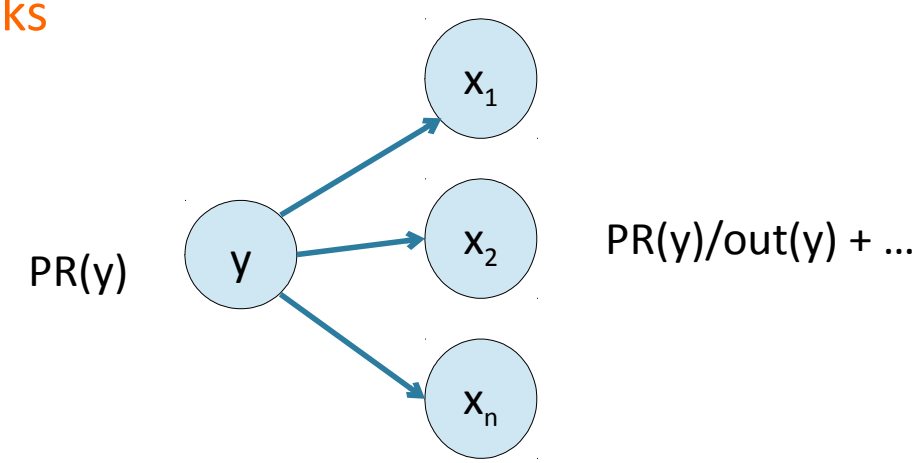
- Mapper $\langle y, \{x_1, x_2, \ldots, x_n\} \rangle$ node, out-links



- Reducer

Expressing PageRank with MapReduce

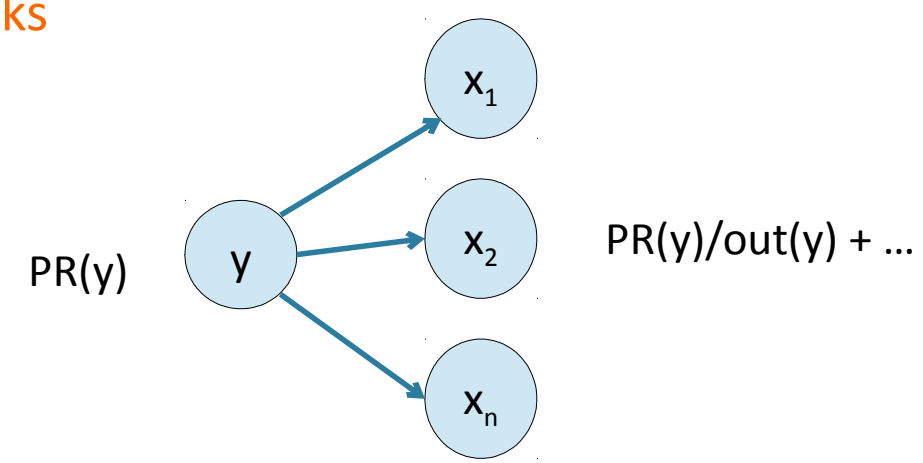
- Mapper $\langle y, \{x_1, x_2, \ldots, x_n\} \rangle$ node, out-links
  - for i =1..n, yield $\left\langle x_i, \frac{PR(y)}{out(y)} \right\rangle$



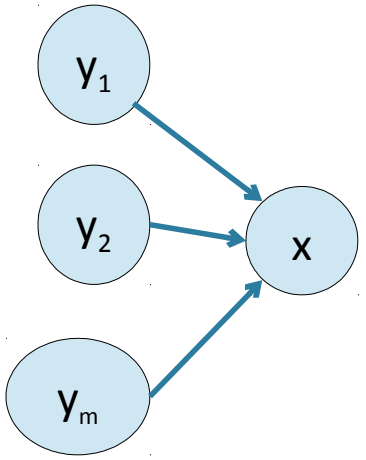
- Reducer

Expressing PageRank with MapReduce

- Mapper $\langle y, \{x_1, x_2, \ldots, x_n\} \rangle$ node, out-links
  - for i =1..n, yield $\left\langle x_i, \frac{PR(y)}{out(y)} \right\rangle$

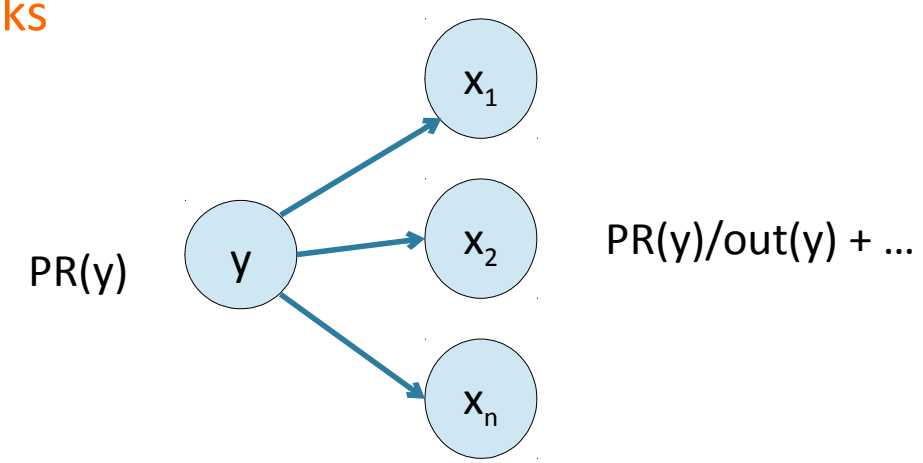


- Reducer

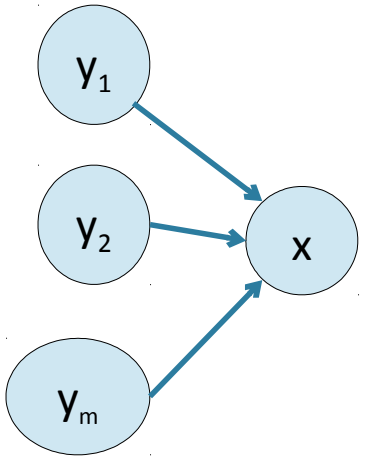


Expressing PageRank with MapReduce

- Mapper $\langle y, \{x_1, x_2, \ldots, x_n\} \rangle$ node, out-links
  - for i =1..n, yield $\left\langle x_i, \frac{PR(y)}{out(y)} \right\rangle$

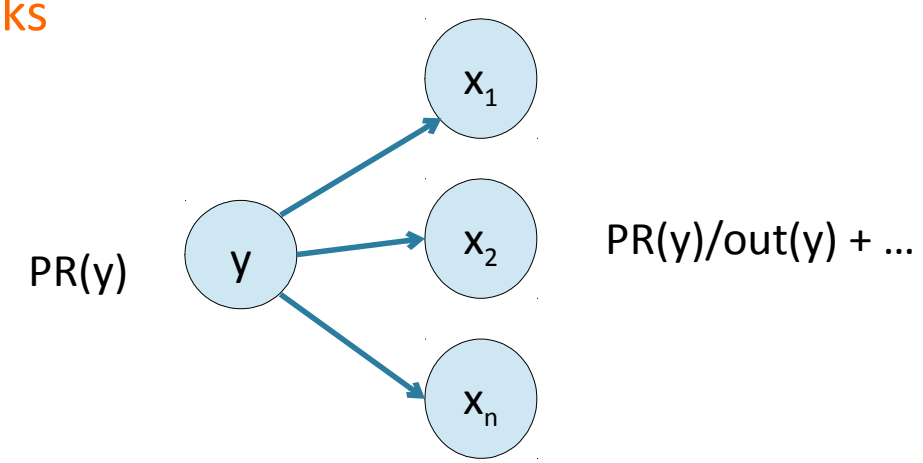


- Reducer node, ΔPR from in-links

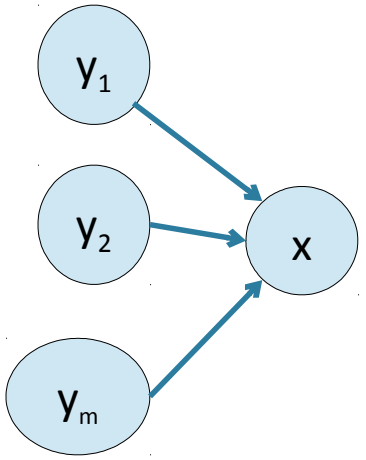


Expressing PageRank with MapReduce

- Mapper $\langle y, \{x_1, x_2, \ldots, x_n\} \rangle$ node, out-links
  - for i =1..n, yield $\left\langle x_i, \frac{PR(y)}{out(y)} \right\rangle$

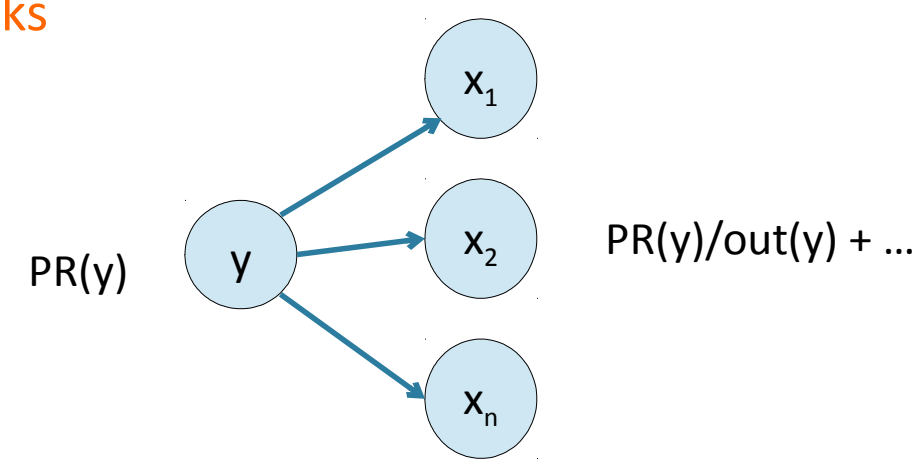


- Reducer $\left\langle x, \left\{ \frac{PR(y_1)}{out(y_1)}, \ldots, \frac{PR(y_m)}{out(y_m)} \right\} \right\rangle$ node, ΔPR from in-links

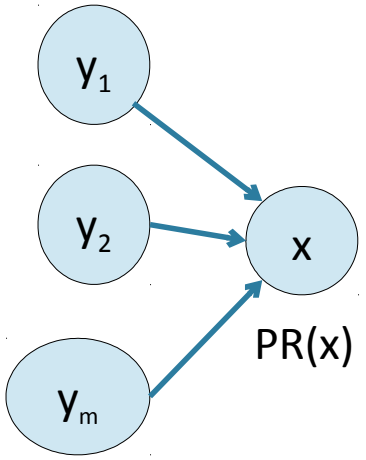


Expressing PageRank with MapReduce

- Mapper $\langle y, \{x_1, x_2, \ldots, x_n\} \rangle$ node, out-links
  - for i =1..n, yield $\left\langle x_i, \frac{PR(y)}{out(y)} \right\rangle$

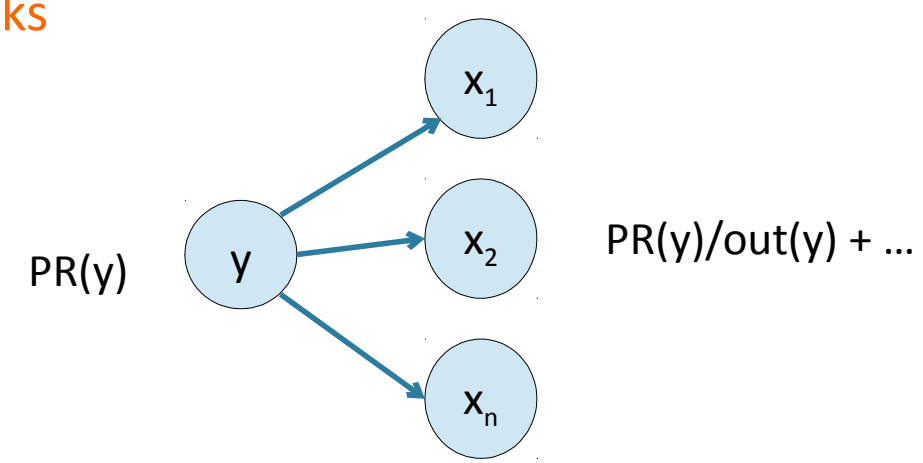


- Reducer $\left\langle x, \left\{ \frac{PR(y_1)}{out(y_1)}, \ldots, \frac{PR(y_m)}{out(y_m)} \right\} \right\rangle$ node, ΔPR from in-links

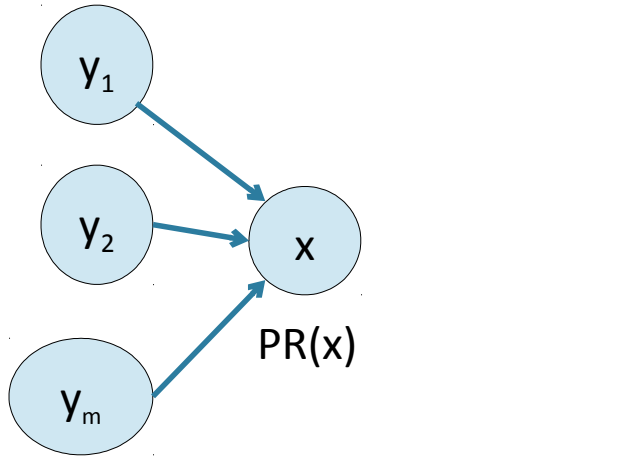


Expressing PageRank with MapReduce

- Mapper $\langle y, \{x_1, x_2, \ldots, x_n\} \rangle$ node, out-links
  - for i =1..n, yield $\left\langle x_i, \frac{PR(y)}{out(y)} \right\rangle$

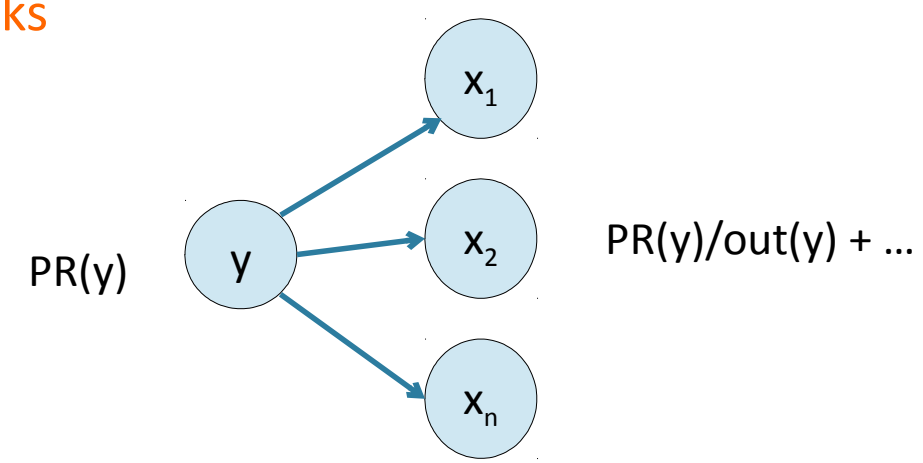


- Reducer $\left\langle x, \left\{ \frac{PR(y_1)}{out(y_1)}, \ldots, \frac{PR(y_m)}{out(y_m)} \right\} \right\rangle$ node, ΔPR from in-links
  - Compute $PR(x) = \frac{1-\beta}{N} + \beta * \sum_{y->x} \frac{PR(y)}{out(y)}$

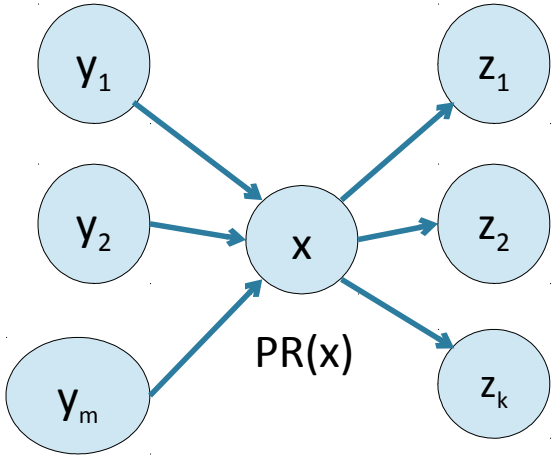


Expressing PageRank with MapReduce

- Mapper $\langle y, \{x_1, x_2, \ldots, x_n\} \rangle$ node, out-links
  - for i =1..n, yield $\left\langle x_i, \frac{PR(y)}{out(y)} \right\rangle$



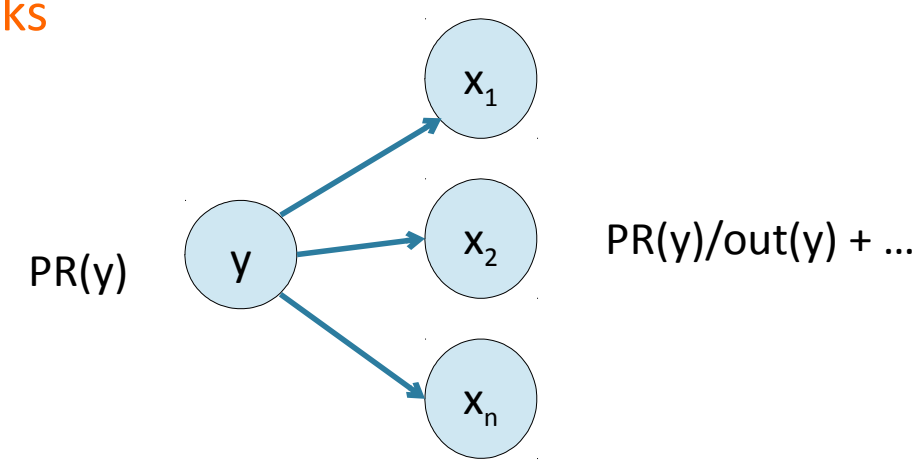
- Reducer $\left\langle x, \left\{ \frac{PR(y_1)}{out(y_1)}, \ldots, \frac{PR(y_m)}{out(y_m)} \right\} \right\rangle$ node, ΔPR from in-links
  - Compute $PR(x) = \frac{1-\beta}{N} + \beta * \sum_{y->x} \frac{PR(y)}{out(y)}$



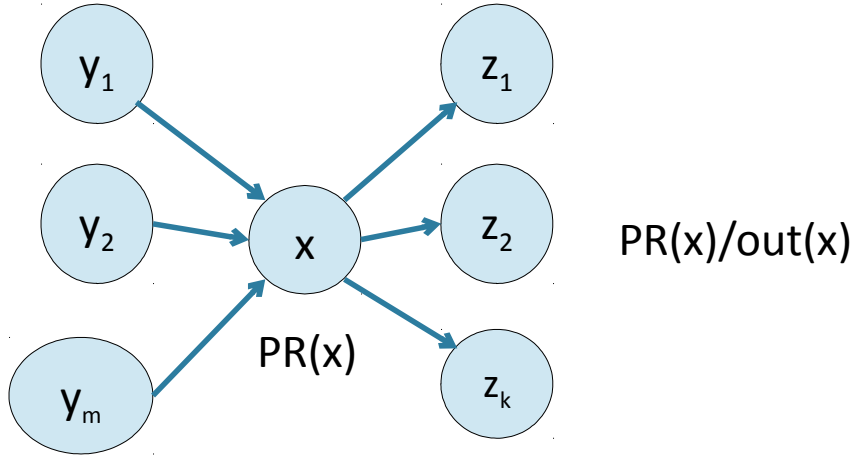
PR(x) = […] + PR(y1)/out(y1) + ….

Expressing PageRank with MapReduce

- Mapper $\langle y, \{x_1, x_2, \ldots, x_n\} \rangle$ node, out-links
  - for i =1..n, yield $\left\langle x_i, \frac{PR(y)}{out(y)} \right\rangle$



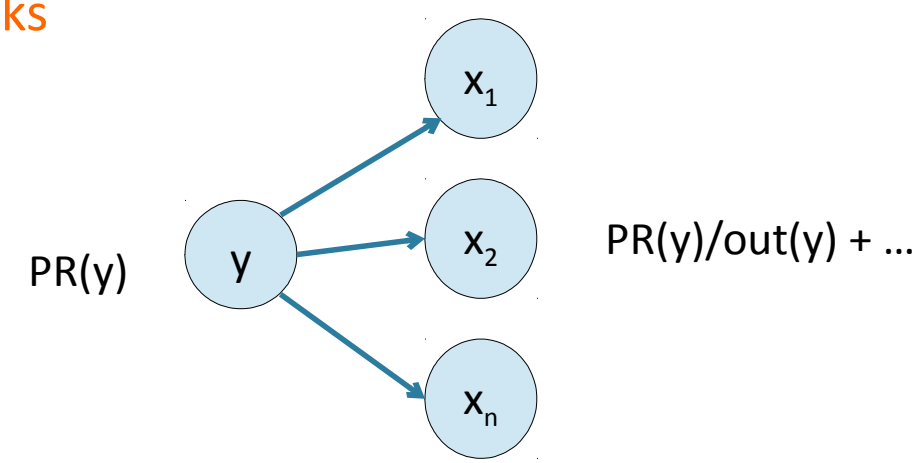
- Reducer $\left\langle x, \left\{ \frac{PR(y_1)}{out(y_1)}, \ldots, \frac{PR(y_m)}{out(y_m)} \right\} \right\rangle$ node, ΔPR from in-links
  - Compute $PR(x) = \frac{1-\beta}{N} + \beta * \sum_{y->x} \frac{PR(y)}{out(y)}$



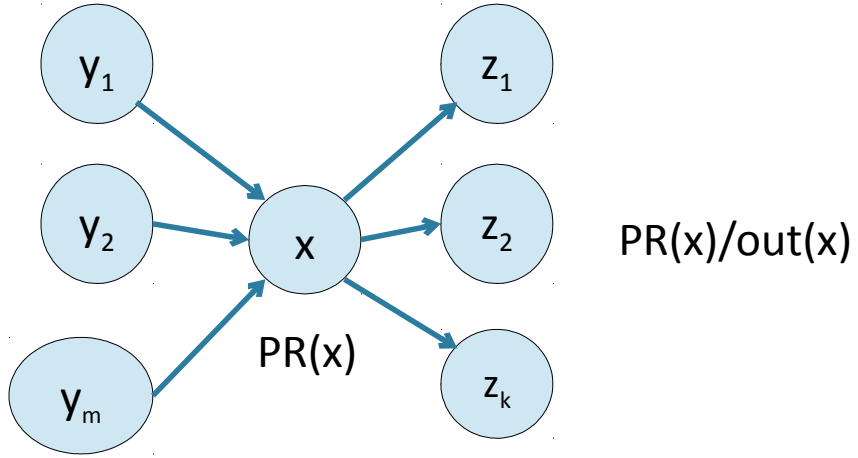
PR(x) = […] + PR(y1)/out(y1) + ….

Expressing PageRank with MapReduce

- Mapper $\langle y, \{x_1, x_2, \ldots, x_n\} \rangle$ node, out-links
  - for i =1..n, yield $\left\langle x_i, \frac{PR(y)}{out(y)} \right\rangle$



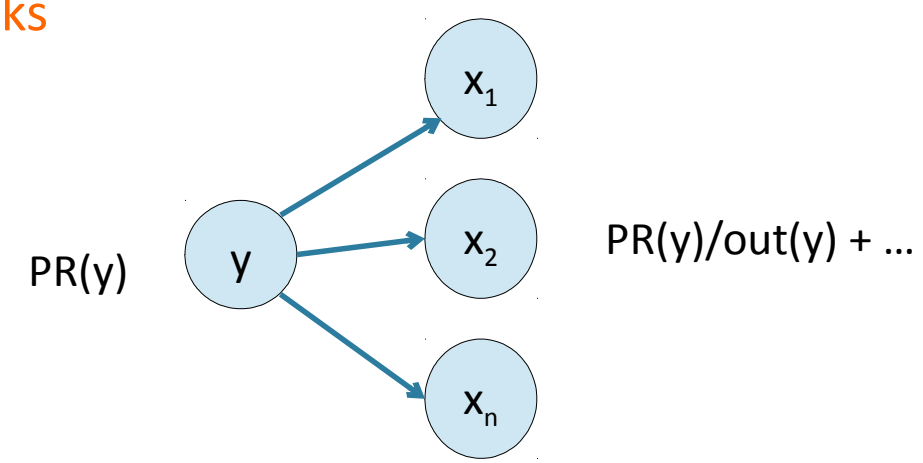
- Reducer $\left\langle x, \left\{ \frac{PR(y_1)}{out(y_1)}, \ldots, \frac{PR(y_m)}{out(y_m)} \right\} \right\rangle$ node, ΔPR from in-links
  - Compute $PR(x) = \frac{1-\beta}{N} + \beta * \sum_{y->x} \frac{PR(y)}{out(y)}$
  - for j=1, k yield $\left\langle z_j, \frac{PR(x)}{out(x)} \right\rangle$



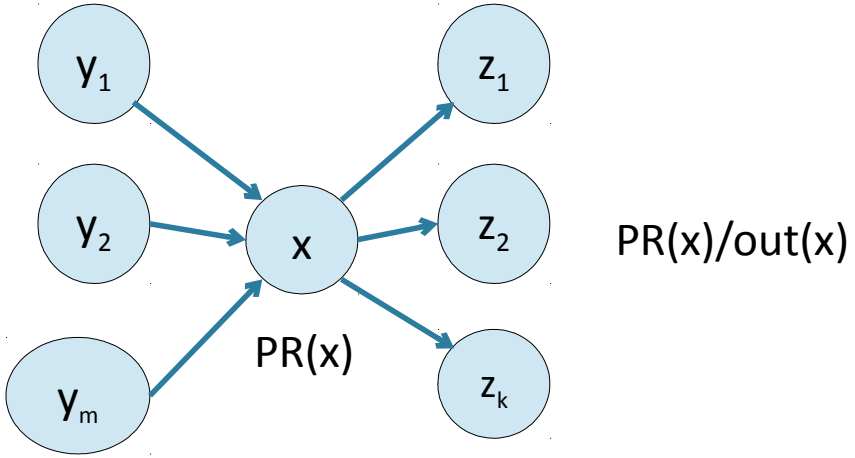
PR(x) = […] + PR(y1)/out(y1) + ….

Expressing PageRank with MapReduce

- Mapper $\langle y, \{x_1, x_2, \ldots, x_n\} \rangle$ node, out-links
  - for i =1..n, yield $\left\langle x_i, \frac{PR(y)}{out(y)} \right\rangle$
  - yield $\langle y, \{x_1, x_2, \ldots, x_n\} \rangle$



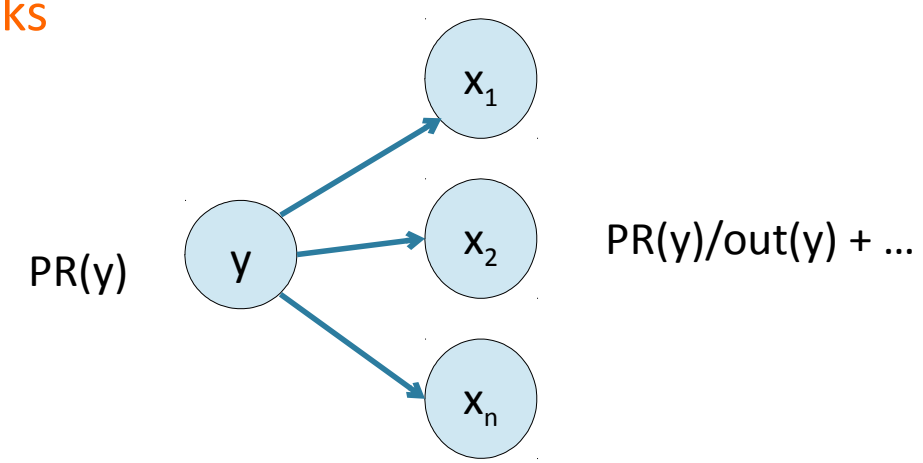
- Reducer $\left\langle x, \left\{ \frac{PR(y_1)}{out(y_1)}, \ldots, \frac{PR(y_m)}{out(y_m)} \right\} \right\rangle$ node, ΔPR from in-links
  - Compute $PR(x) = \frac{1-\beta}{N} + \beta * \sum_{y->x} \frac{PR(y)}{out(y)}$
  - for j=1, k yield $\left\langle z_j, \frac{PR(x)}{out(x)} \right\rangle$



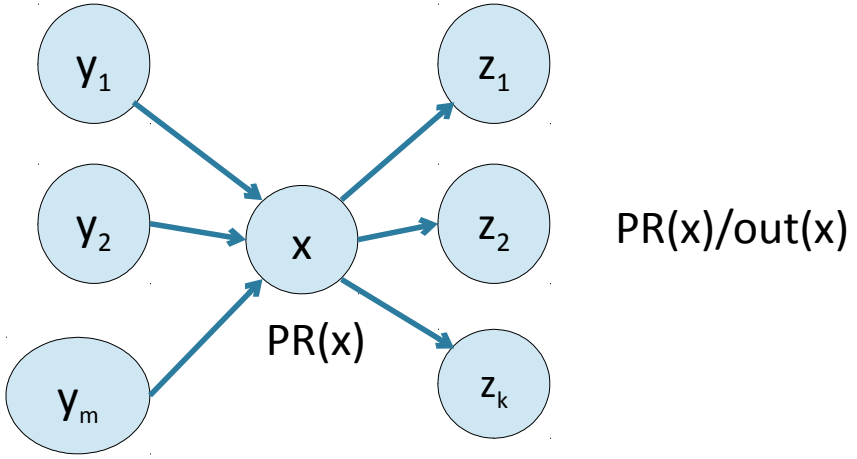
PR(x) = […] + PR(y1)/out(y1) + ….

Expressing PageRank with MapReduce

- Mapper $\langle y, \{x_1, x_2, \ldots, x_n\} \rangle$ node, out-links
  - for i =1..n, yield $\left\langle x_i, \frac{PR(y)}{out(y)} \right\rangle$
  - yield $\langle y, \{x_1, x_2, \ldots, x_n\} \rangle$



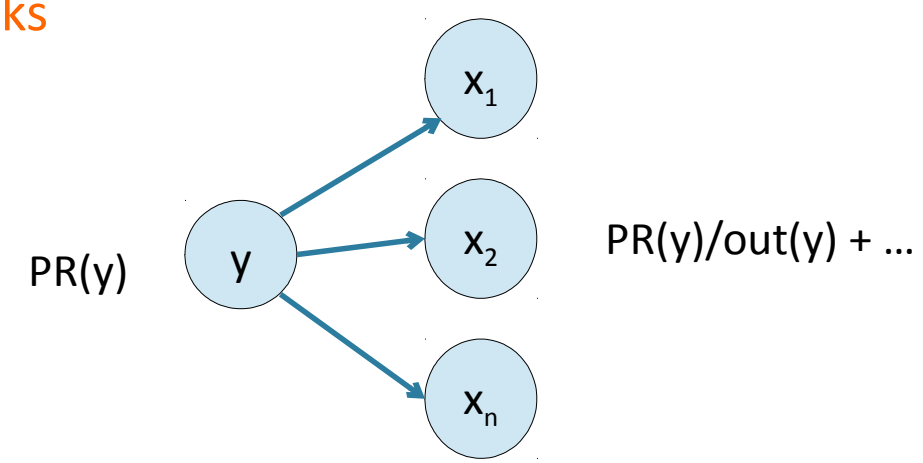
- Reducer $\left\langle x, \left\{ \frac{PR(y_1)}{out(y_1)}, \ldots, \frac{PR(y_m)}{out(y_m)}, \{z_1, \ldots, z_k\} \right\} \right\rangle$ node, ΔPR from in-links
  - compute $PR(x) = \frac{1-\beta}{N} + \beta * \sum_{y->x} \frac{PR(y)}{out(y)}$
  - for j=1..k, yield $\left\langle z_j, \frac{PR(x)}{out(x)} \right\rangle$



PR(x) = […] + PR(y1)/out(y1) + ….

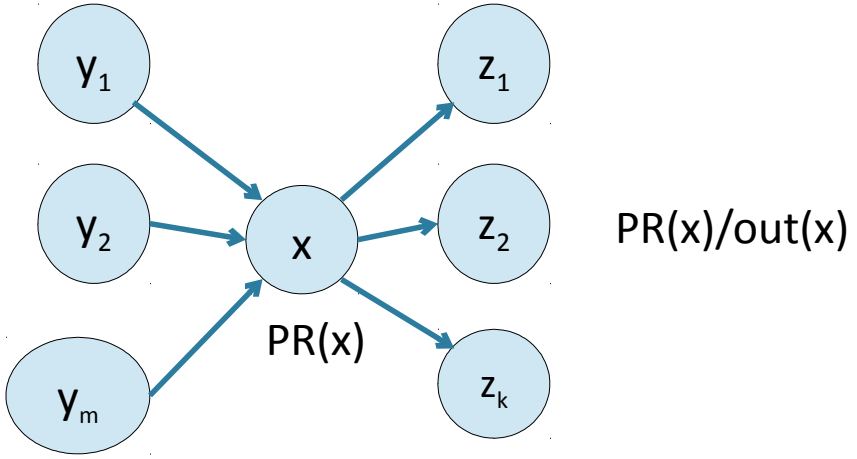
Expressing PageRank with MapReduce

- Mapper $\langle y, \{x_1, x_2, \ldots, x_n\} \rangle$ node, out-links
  - for i =1..n, yield $\left\langle x_i, \frac{PR(y)}{out(y)} \right\rangle$
  - yield $\langle y, \{x_1, x_2, \ldots, x_n\} \rangle$



- Reducer $\left\langle x, \left\{ \frac{PR(y_1)}{out(y_1)}, \ldots, \frac{PR(y_m)}{out(y_m)}, \{z_1, \ldots, z_k\} \right\} \right\rangle$ node, + ΔPR from in-links
  - compute $PR(x) = \frac{1-\beta}{N} + \beta * \sum_{y->x} \frac{PR(y)}{out(y)}$
  - for j=1..k, yield $\left\langle z_j, \frac{PR(x)}{out(x)} \right\rangle$

$\langle x, \{z_1, z_2, \ldots, z_k\} \rangle$



PR(x) = […] + PR(y1)/out(y1) + ….