PageRank Explained

Daniel Wendon-Blixrud

Background

- Early days of the internet
- Only a few websites existed
- A list of every website could be kept in the back of a notebook

The first problem

- Too many websites to keep track of
- Notebooks weren’t big enough
- Websites containing lists of other websites were made to solve this

The second problem

- Too many websites in this list
- Needed a way to search the list to find the site you wanted
- Millions of sites to search through
- How do you search the list, and find the correct website

The solution

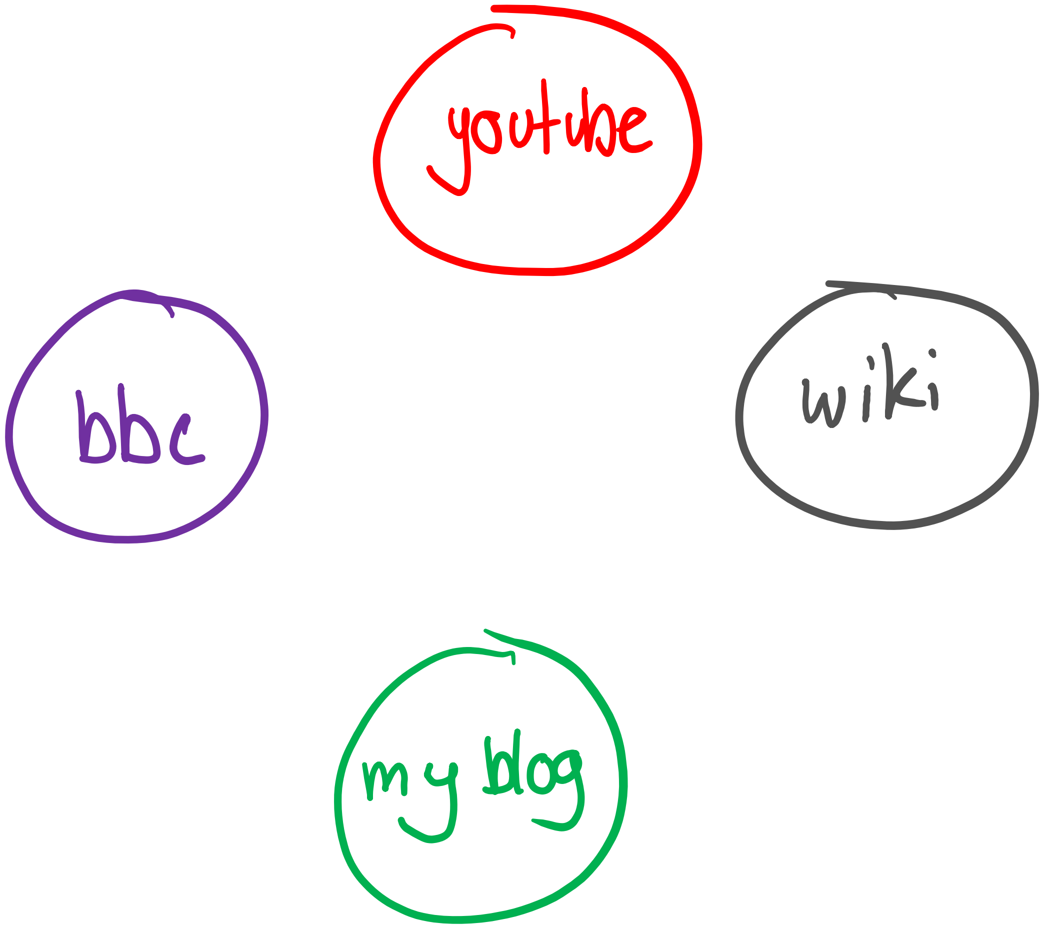
- A way to rank websites was needed
- Larry Page created **Page**Rank, a method of giving each website a score
- As the internet was based off of hyperlinks, could use these inter-site links to an advantage

The process

- Sites which are linked to *more* must be better
- Sites with fewer links to them are worse
- A better site’s link is worth more than a worse site

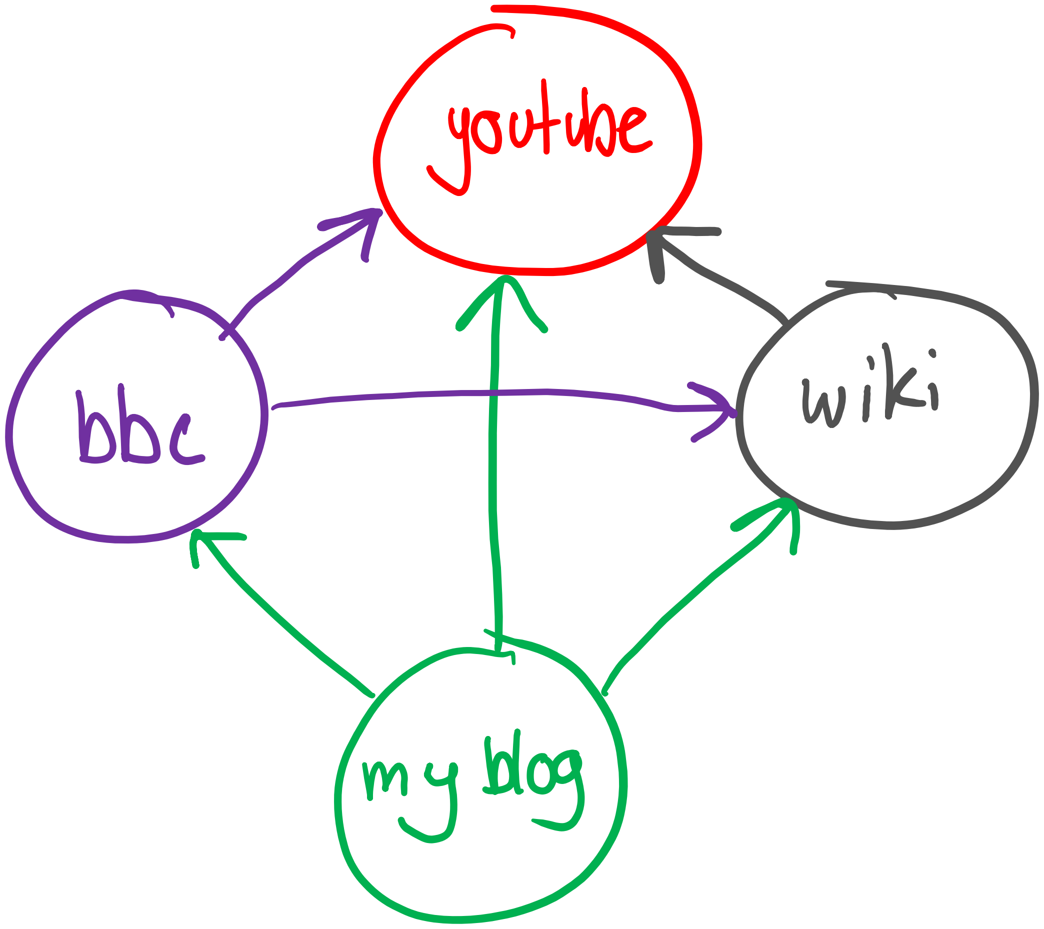
How to score websites

- Imagine only 4 websites exist
- Each website will have different number of users
- Each website will link to different other websites



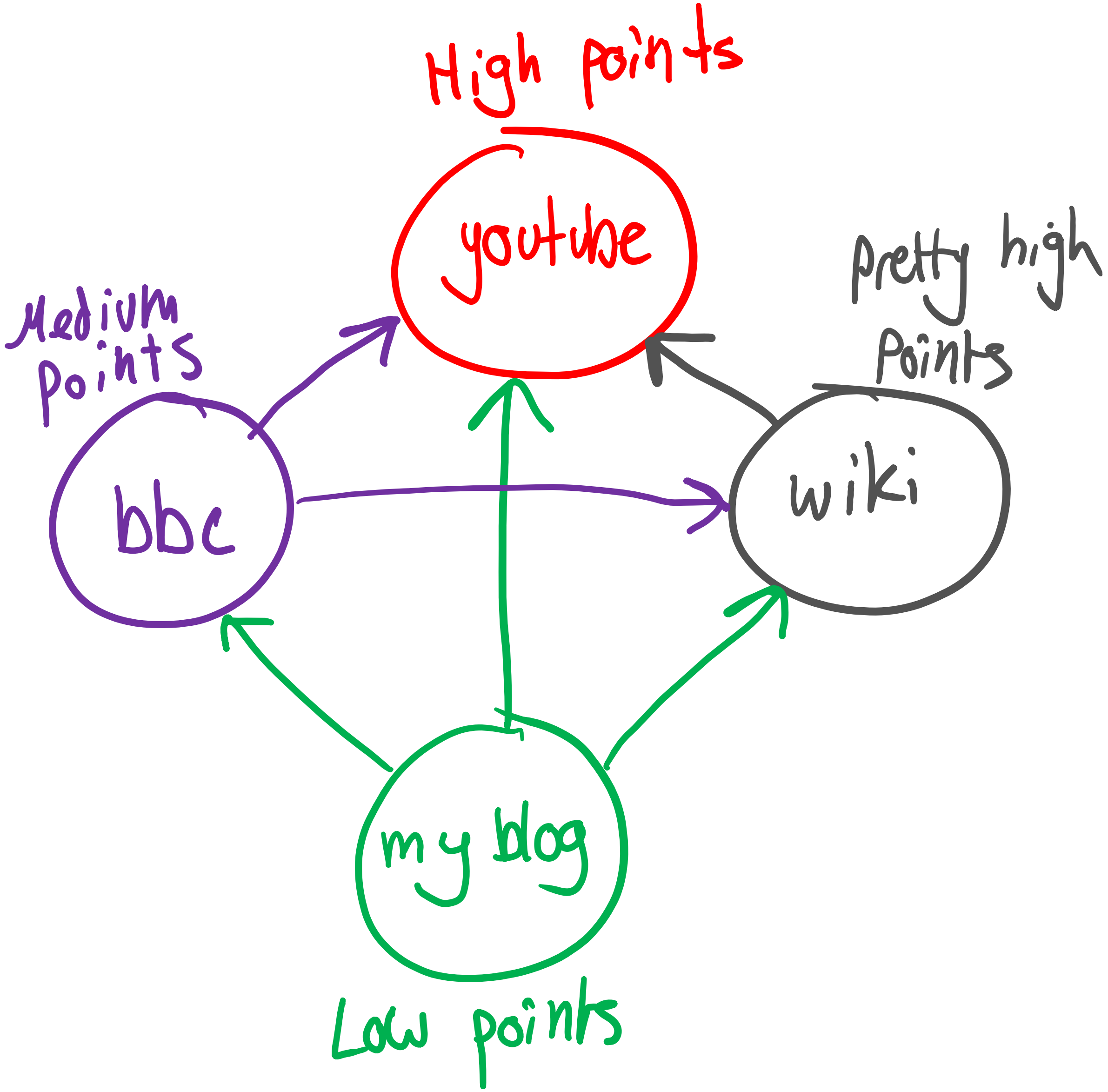
How to score websites

- Show each website's out-links to other sites
- YouTube has lots of in-links, its popular
- My Blog has no in-links, its not very popular
- We want to score YouTube highly, and My Blog low



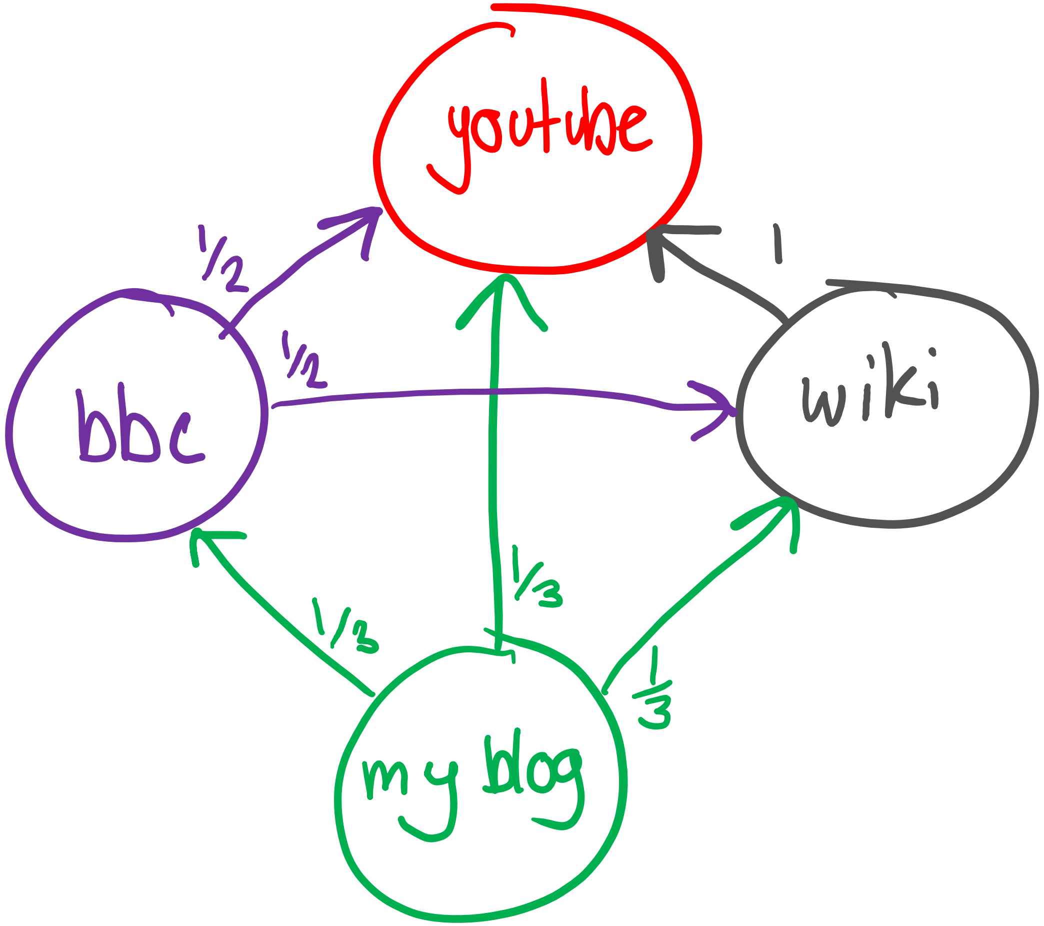
How to score websites

- Each link is worth some 'points'
- The more points a website has, the higher score
- A website with a high score gives more points per link (it is good to be linked to by a popular site)
- A website with loads of out-links, gives fewer points per link (its not that special to be linked by a site who links to everyone)



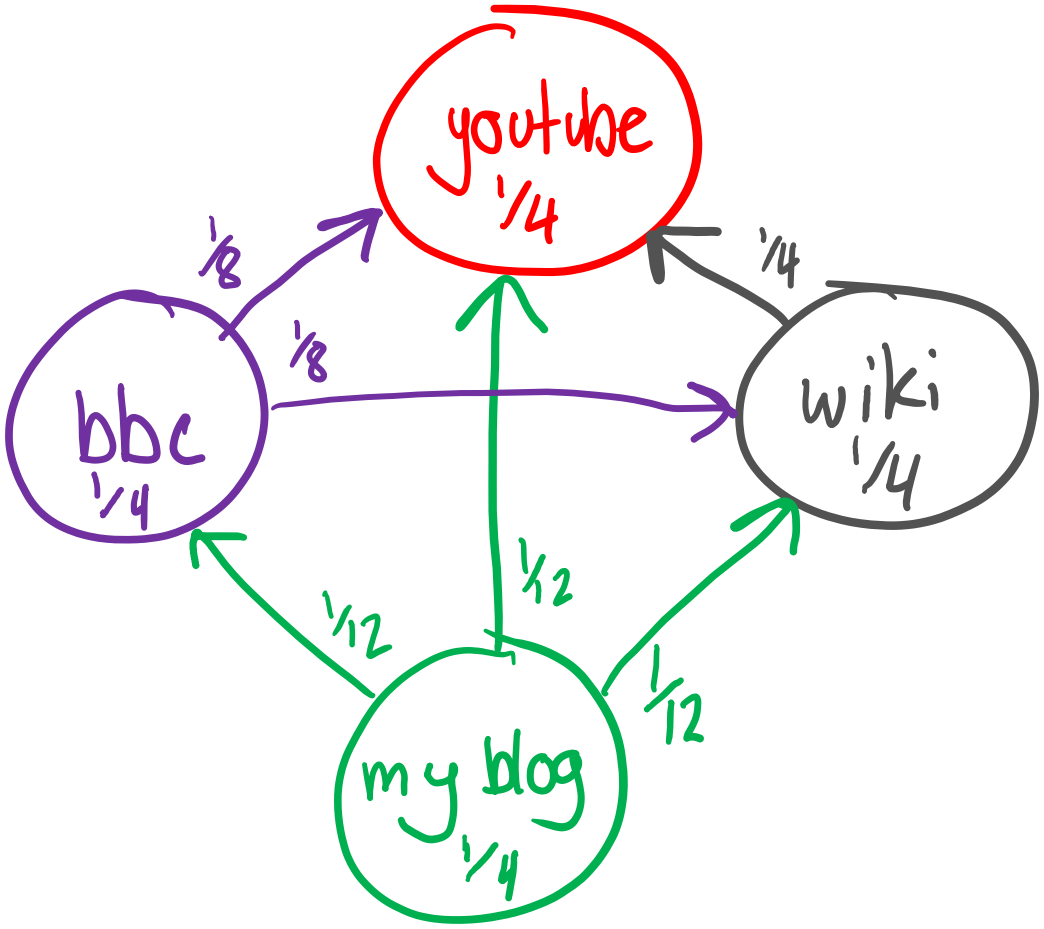
How to score websites

- Give each website a score
- With each out-link, it shares this score to each of the sites it links to
- The more out-links a site has, the less each link is worth
- The more in-links a site has, the higher its score, so the more each link is worth



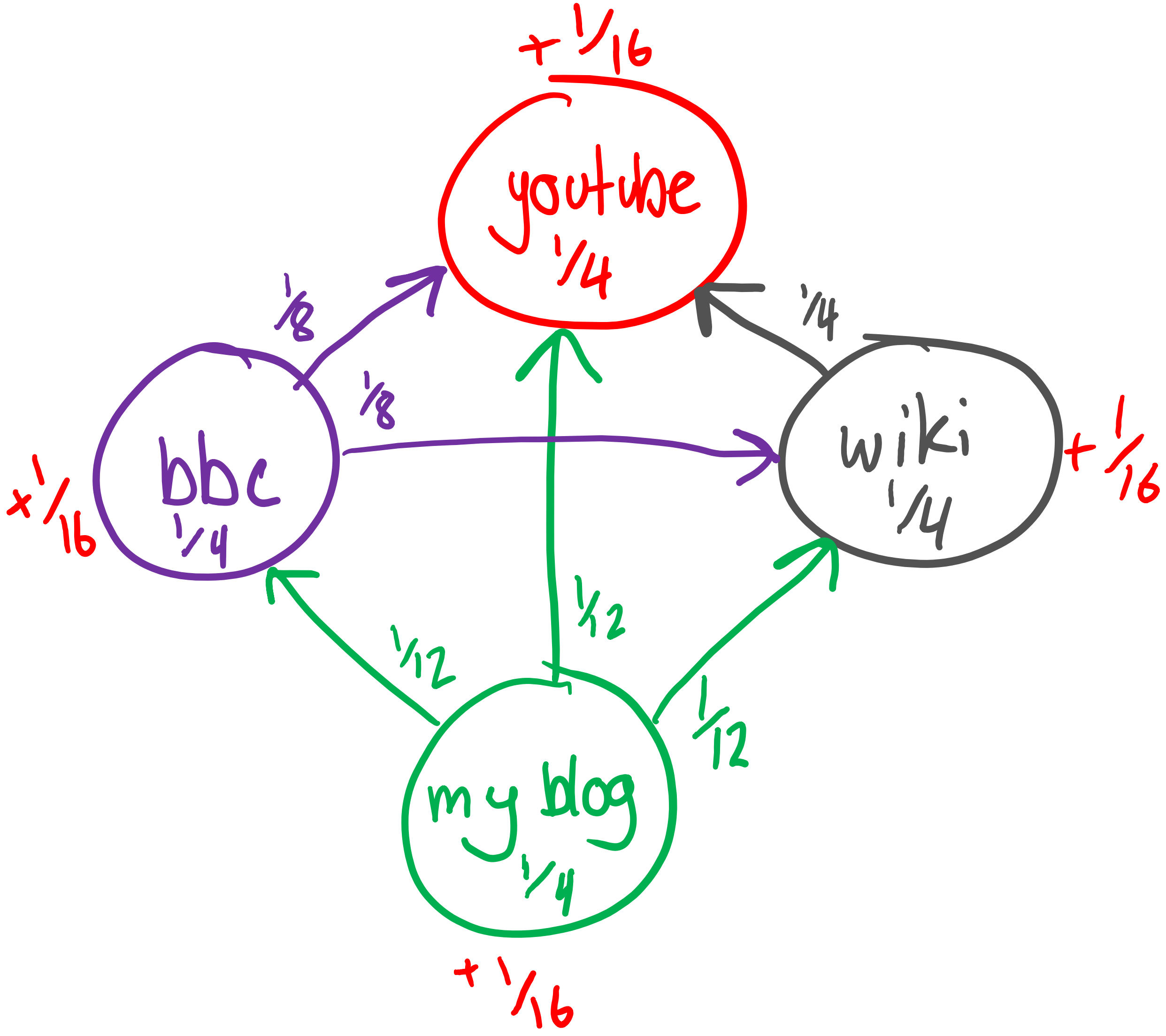
How to score websites

- The total sum of all scores is 1
- Start each site with the same score, 1÷n, there being n sites
- In this case, the starting score for each site is 1÷4, ¼
- Sites share their score evenly with each link
- Eg BBC, with a score of ¼, split between two out-links, each out-link gets ½ of a ¼, an ⅛



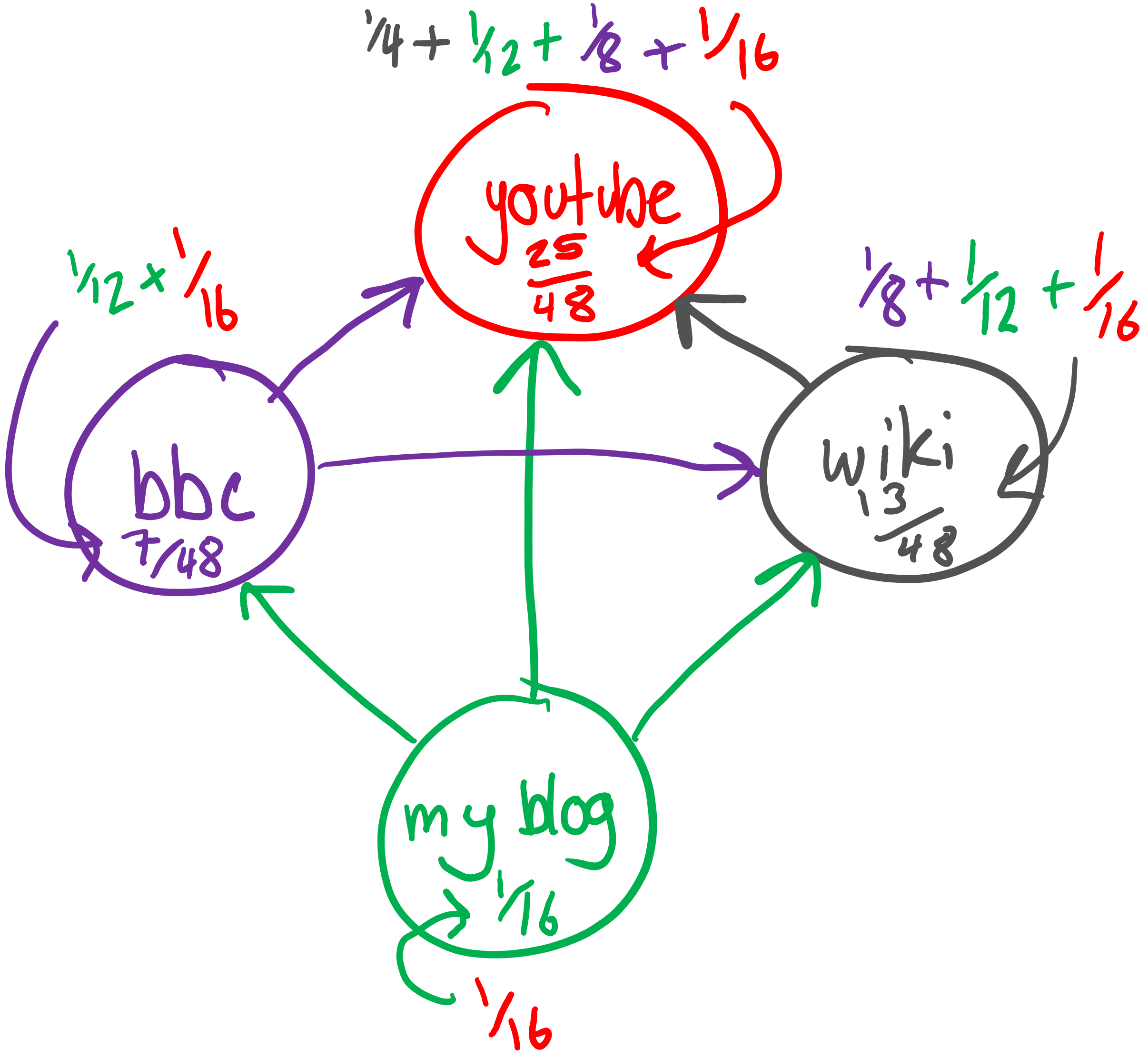
How to score websites

- What about YouTube?
- YouTube has no out-links, so its score is shared equally with all sites (including YouTube)



How to score websites

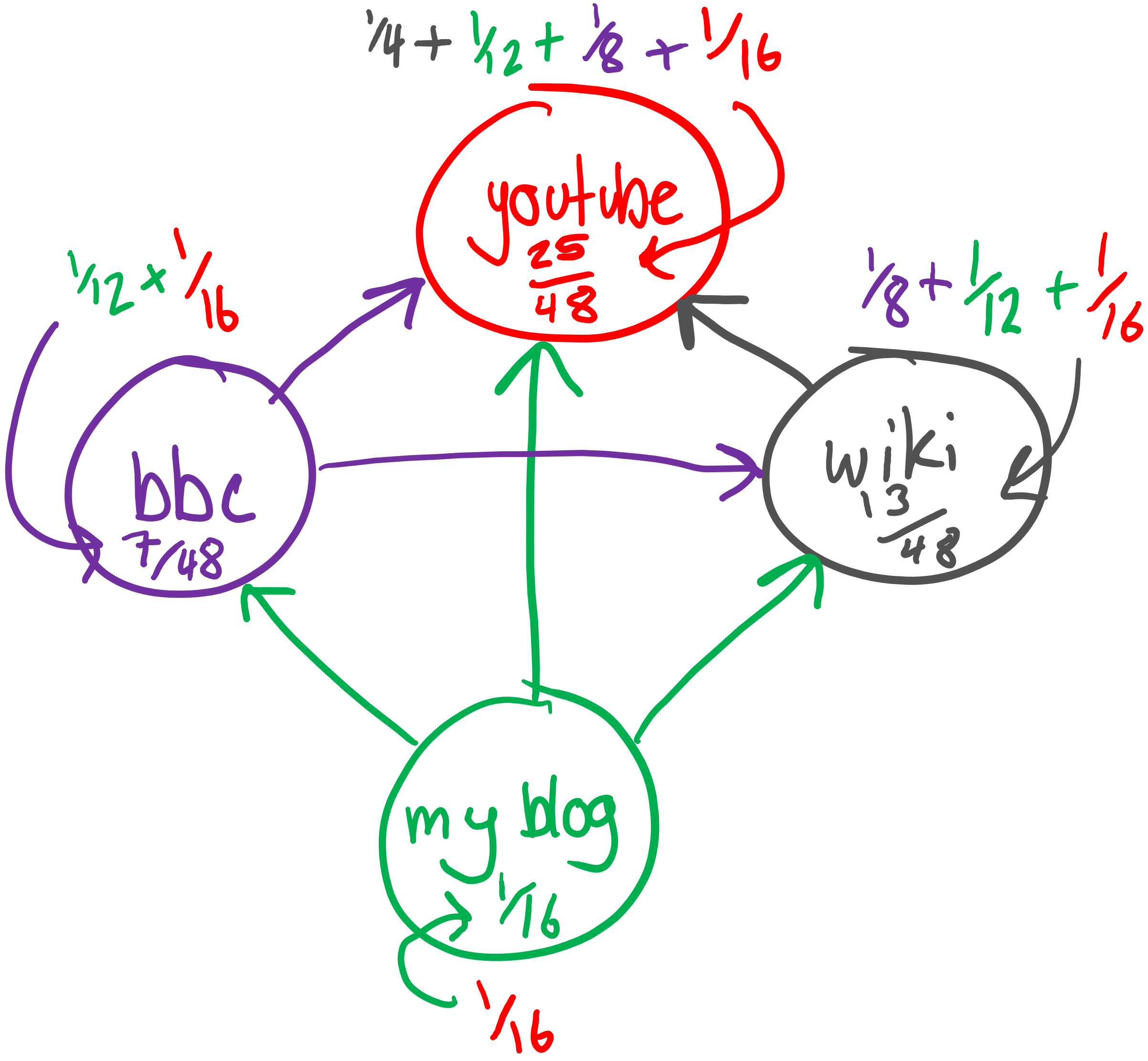
- Each site's current score is replaced by the sum of in-link scores
- This process is repeated, with new scores being shared evenly between out-links, it is iterative
- Each sites score gets more accurate with more iterations
- Google performs 50 iterations for their site indexing



A formula

- A site’s score per out-link, is its own score, divided by the number of out-links
- A site’s new score, is the sum of the scores of it’s in-links
- S( ) is a site’s score
- L( ) is the number of out-links

$$S(Y) = \frac{S(B)}{L(B)} + \frac{S(M)}{L(M)} + \frac{S(W)}{L(W)} + \frac{S(Y)}{L(Y)}$$



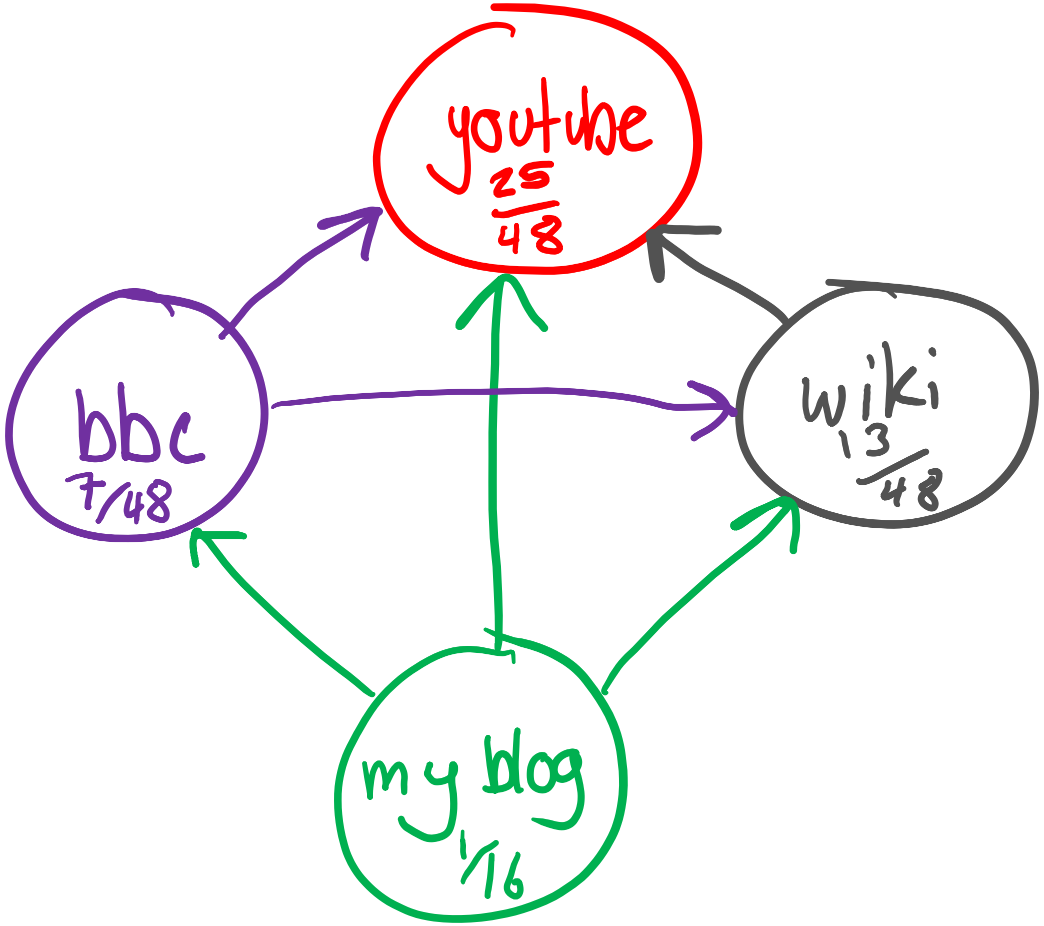
A formula

- Using the formula for the next iteration
- Finding YouTube's new score

$$S(Y) = \frac{S(B)}{L(B)} + \frac{S(M)}{L(M)} + \frac{S(W)}{L(W)} + \frac{S(Y)}{L(Y)}$$

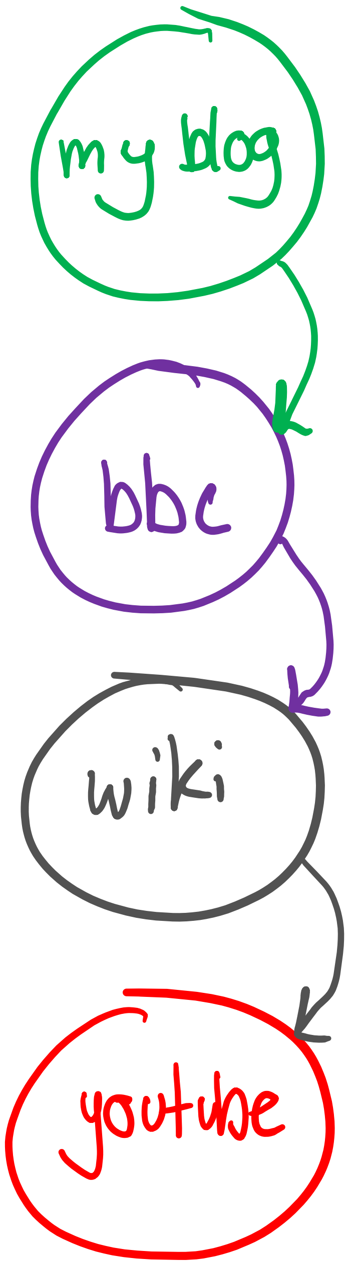
$$S(Y) = \frac{\frac{7}{48}}{2} + \frac{\frac{1}{16}}{3} + \frac{\frac{13}{48}}{1} + \frac{\frac{25}{48}}{4}$$

$$S(Y) = \frac{7}{96} + \frac{1}{48} + \frac{13}{48} + \frac{25}{192} = \frac{95}{192}$$



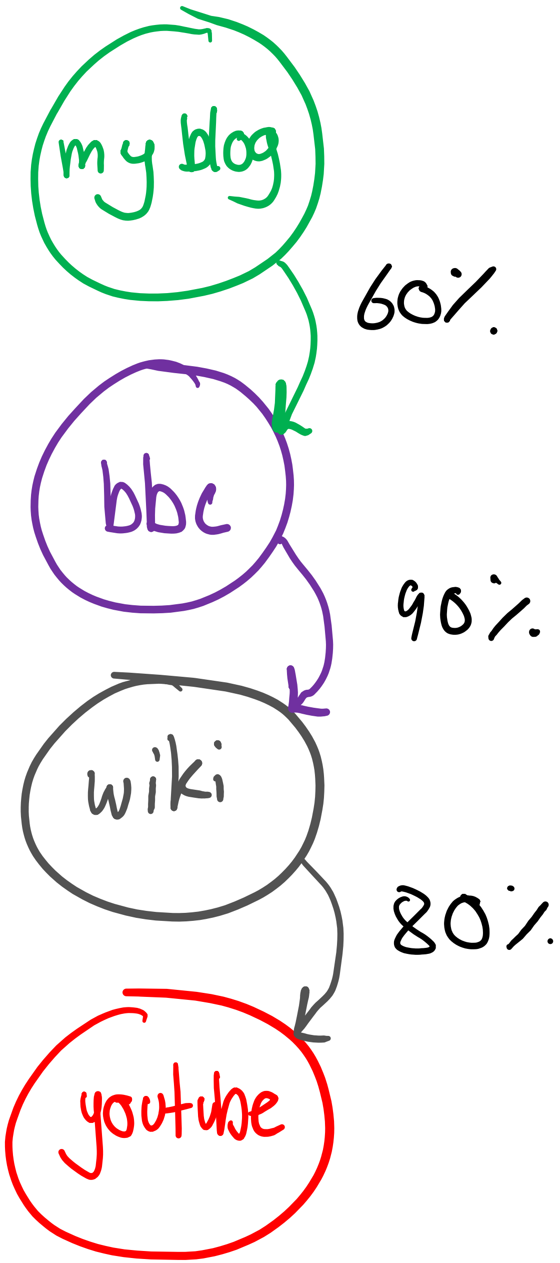
Another problem

- Imagine the same websites, but with this out-link pattern
- A site's score is the probability that a user will visit it
- That's why more in-links means a higher score (a higher chance of being visited from the link)
- But not every user will always click on every single link



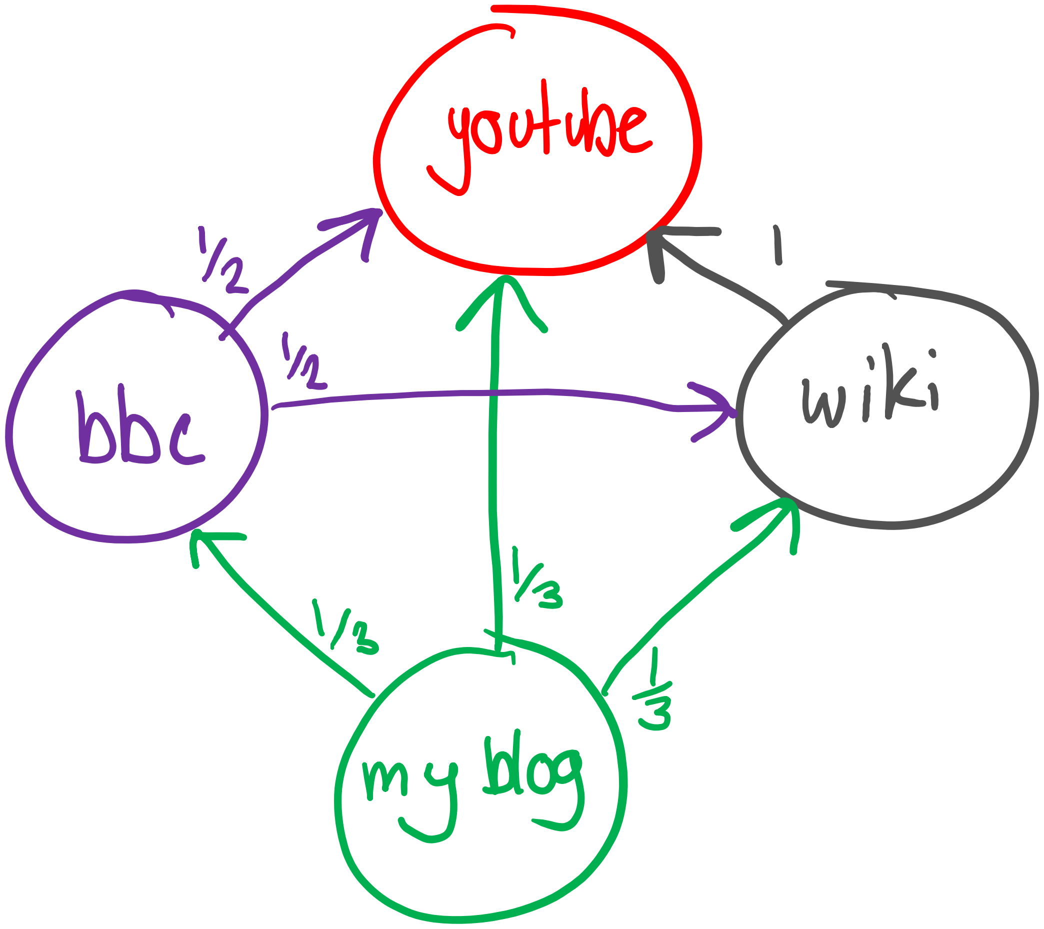
Another problem

- If 100 people visit My Blog, maybe only 60 of them click the link to BBC
- Out of those 60 people, only 54 people click to Wiki
- Out of the 54 Wiki visitors, only 43 click to YouTube link
- Our formula assumes that there is a 100% chance that a user will follow a link on a page



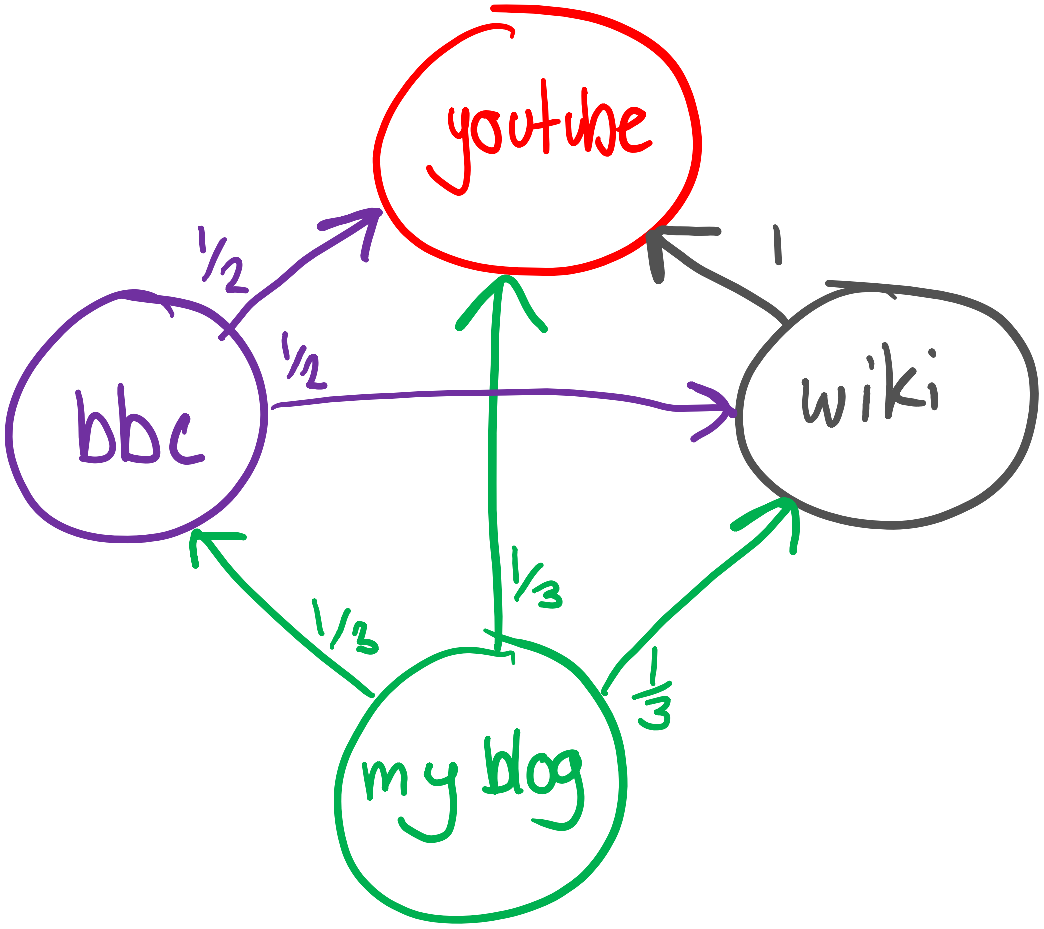
Another problem

- Going back to this diagram, if we look at BBC
- If a user is currently on BBC, there is a ½ chance they visit YouTube, and a ½ chance they visit Wiki
- What if they don't click, and stay at BBC? Or leave their computer and stop clicking links altogether?



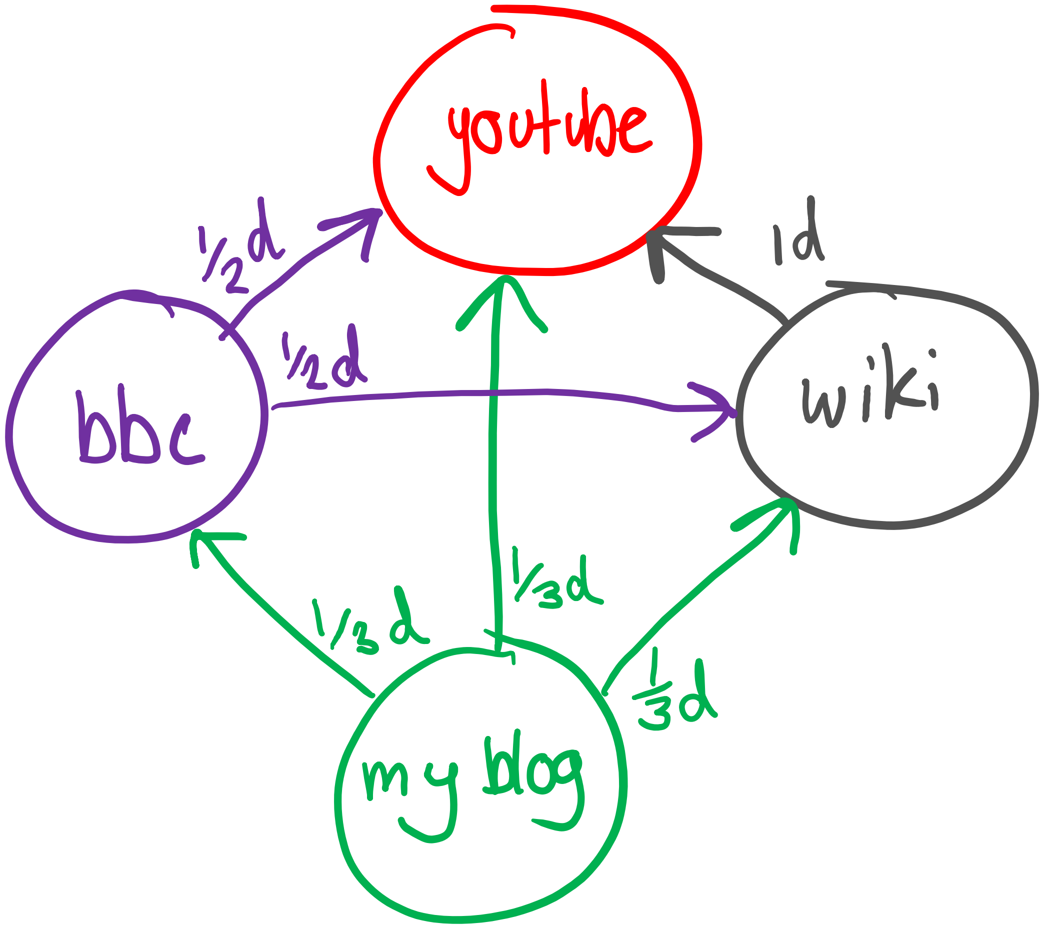
Solution - damping

- We need to account for the probability the user does *nothing*
- We find a probability, that while on a website, the user does not click any links
- This probability has been heavily researched, and is taken as **0.15**
- We take this, and turn it into the probability the user *does* click on at least one link, this is called *d* and is taken as **0.85** (damping)



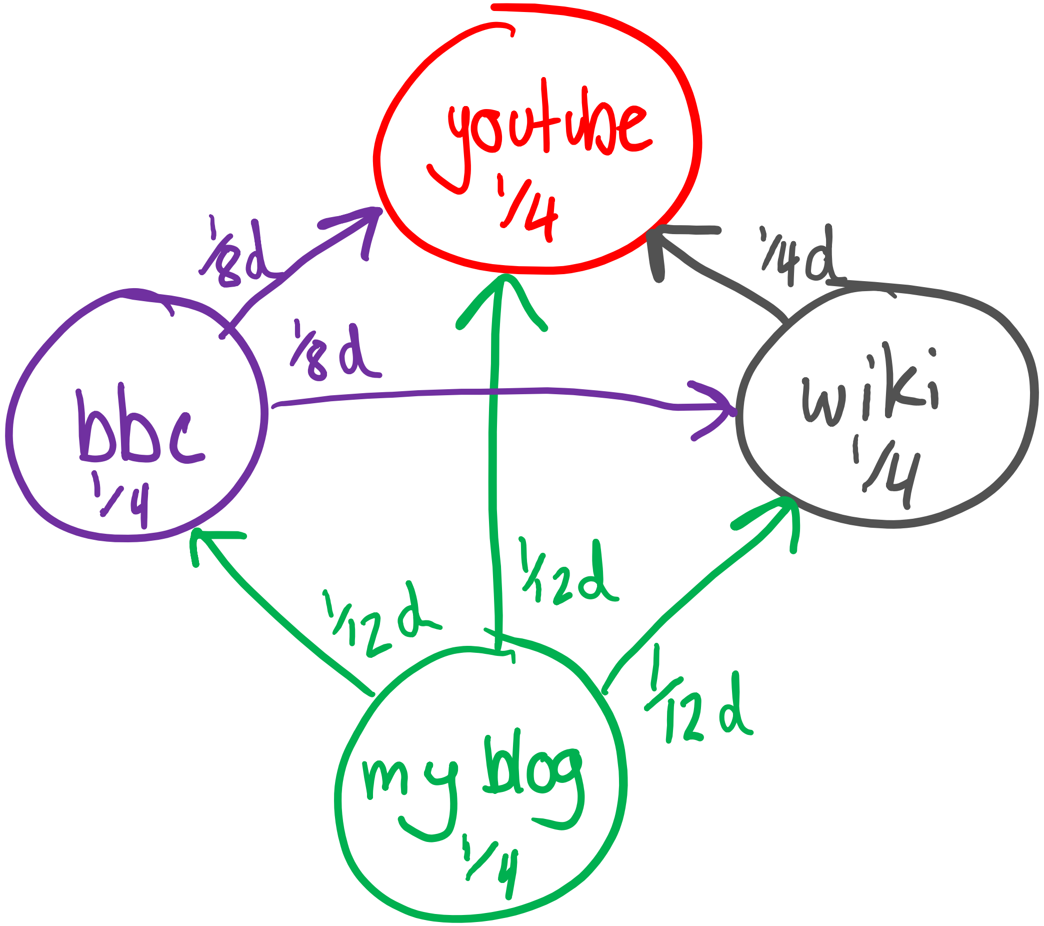
Solution

- How do we use this probability?
- Simply multiply each link's probability of being chosen (aka its score) by the probability that the user actually clicks a link in the first place
- Eg BBC, each link has a ½ probability of being followed, so to account for the user clicking at all, we multiply by 0.85 *(d)* to get $\frac{1}{2} \times d = \frac{17}{40} = 0.425$



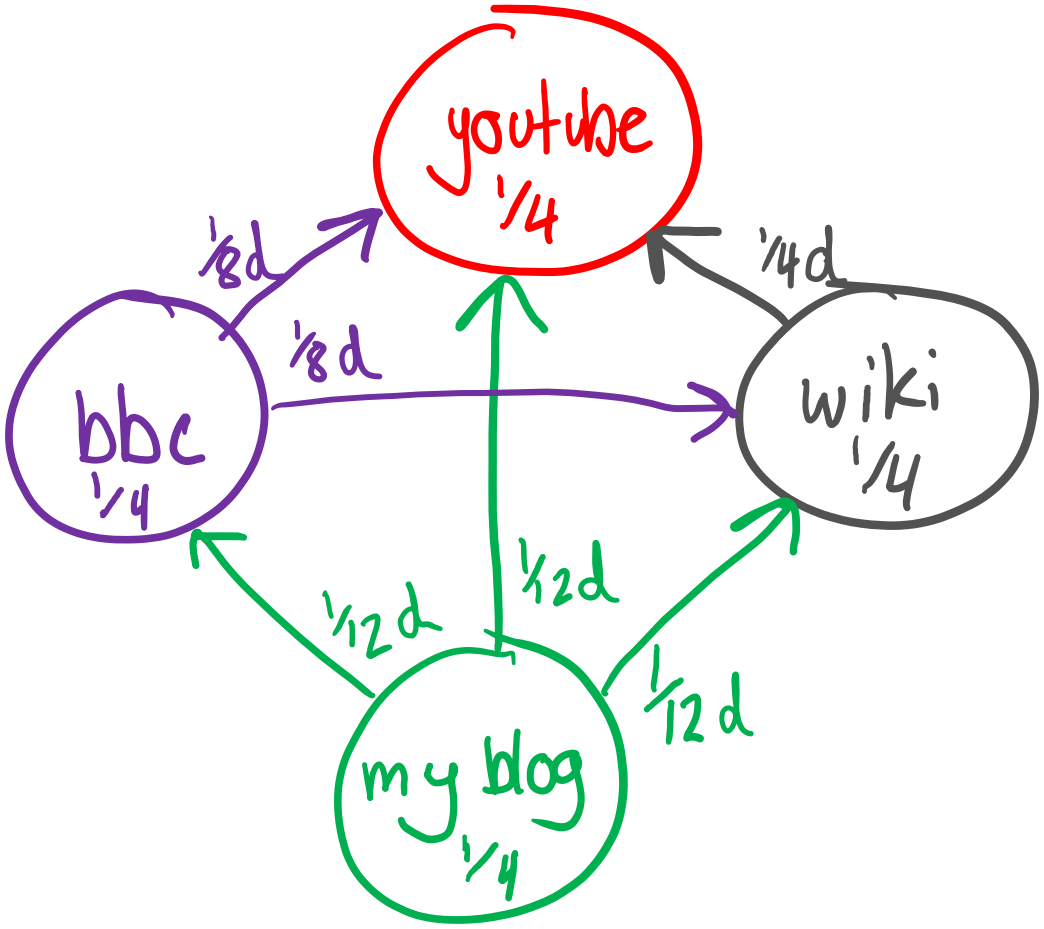
Solution

- Don’t forget we also multiply by the site’s own score
- Before, each BBC link had a ½ chance of being chosen, *if the user was already on BBC*
- A site’s score is the probability that a user is currently on the site, so by multiplying the site’s score by the link’s chance of being clicked, we get the absolute probability of a user clicking on the link



Solution

- This 'absolute probability' is just the chance of the link being pressed if we don't know where the user is
- Eg if 100 people randomly clicked these links, each BBC out-link has a ⅛*d* chance of being clicked
- Remember we also multiply by *d* to decrease each probability, and account for the off-chance the user doesn't do anything



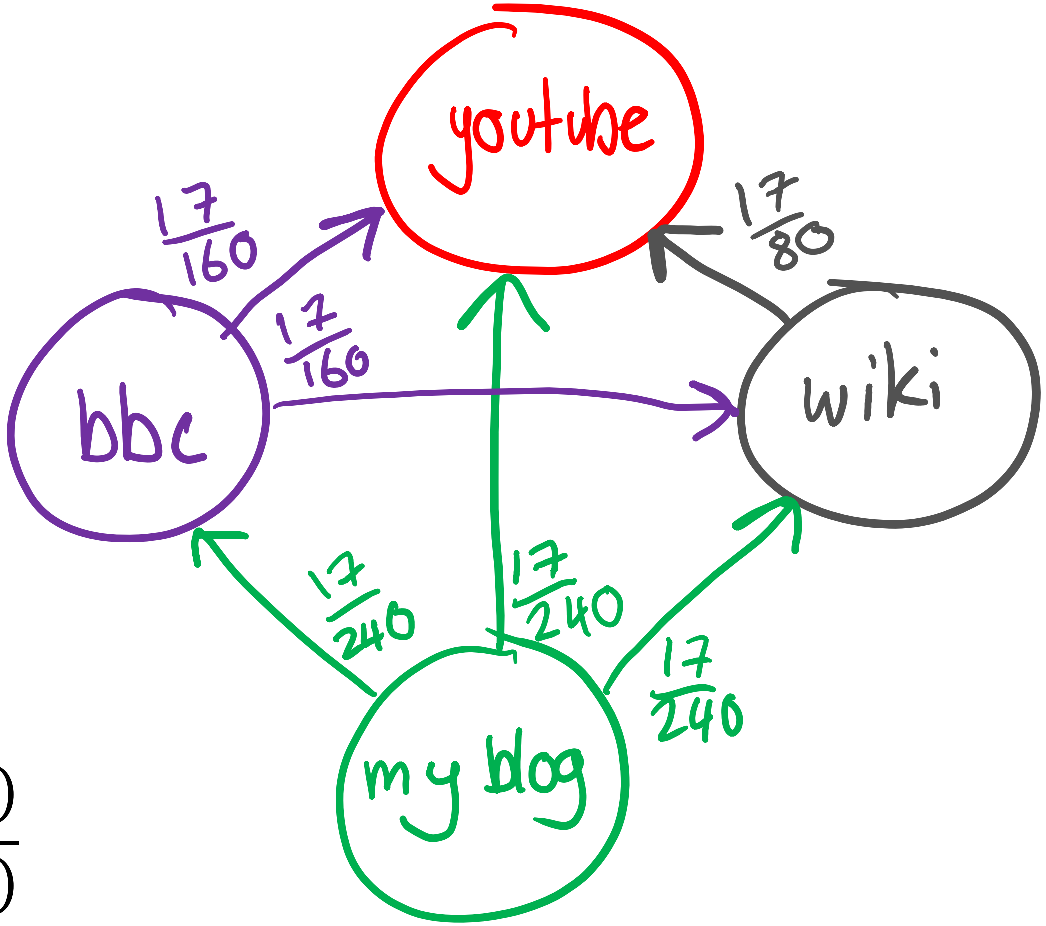
Formula update

- So how do we account for this in our formula?
- Recall our first formula

$$S(Y) = \frac{S(B)}{L(B)} + \frac{S(M)}{L(M)} + \frac{S(W)}{L(W)} + \frac{S(Y)}{L(Y)}$$

- But now, each in-link's score is being multiplied by 0.85 *(d)*

$$S(Y) = d\frac{S(B)}{L(B)} + d\frac{S(M)}{L(M)} + d\frac{S(W)}{L(W)} + d\frac{S(Y)}{L(Y)}$$

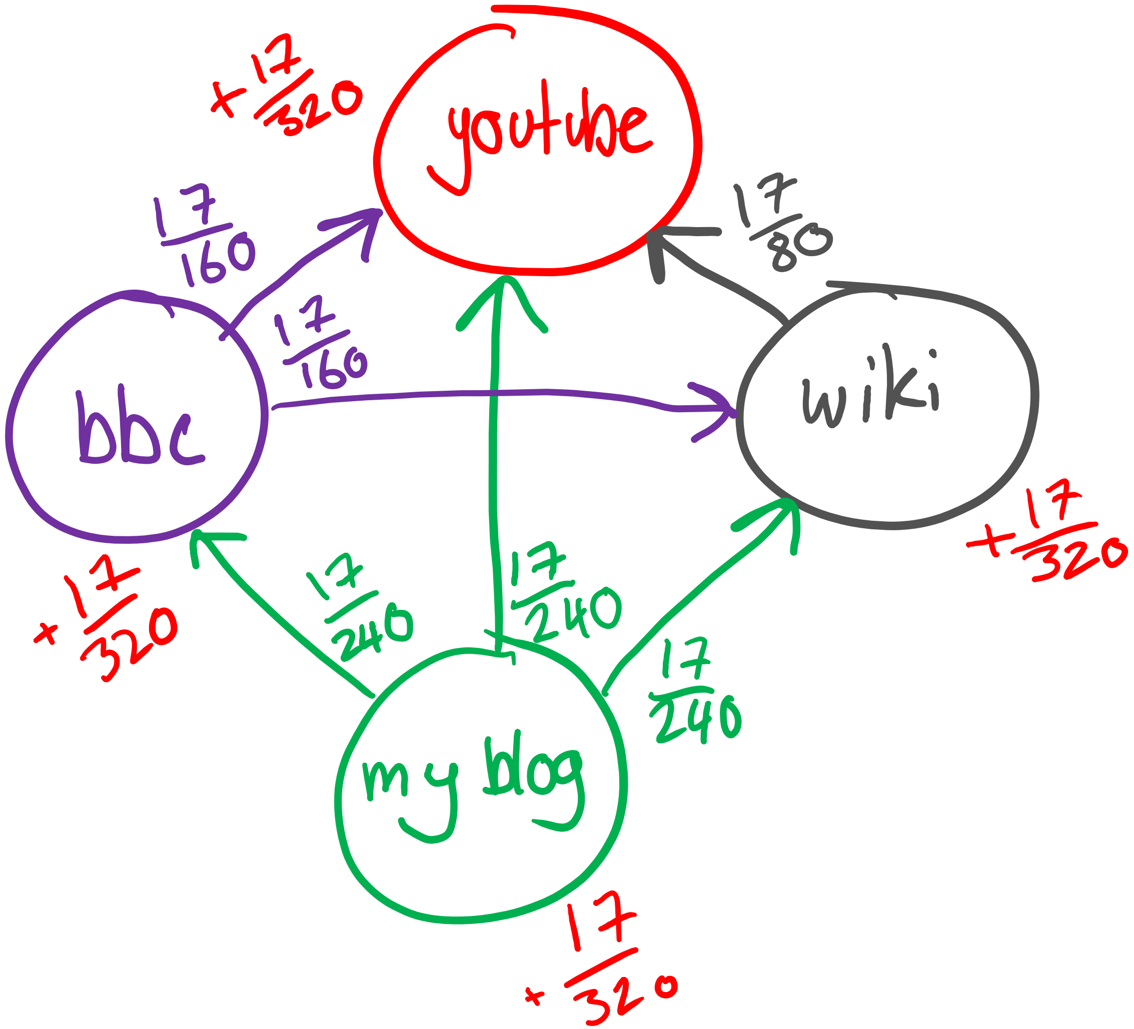


Formula update

- Using some simple maths, we can factorise *d*

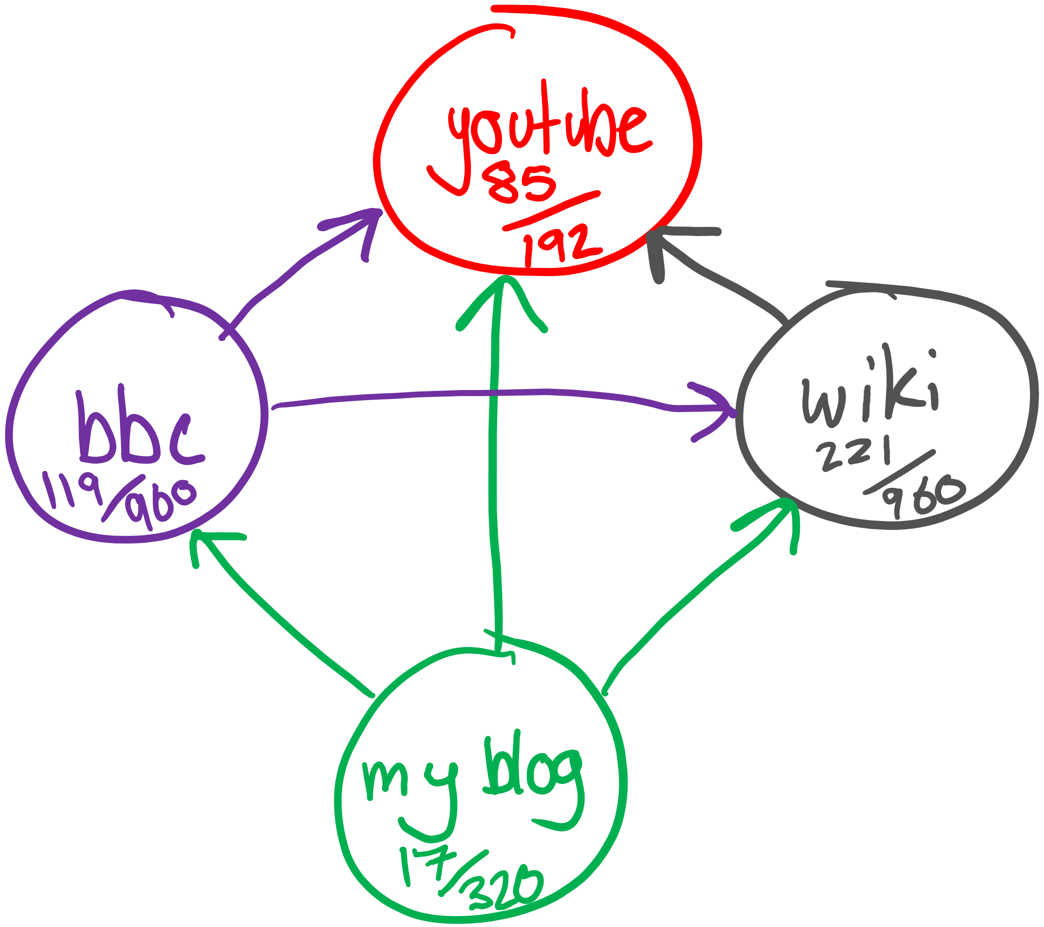
$$S(Y) = d\left(\frac{S(B)}{L(B)} + \frac{S(M)}{L(M)} + \frac{S(W)}{L(W)} + \frac{S(Y)}{L(Y)}\right)$$

- Now remember, all this formula is doing is adding all the in-link scores for a site, and multiplying them by *d*, the probability the user clicks on a link in the first place



Another problem

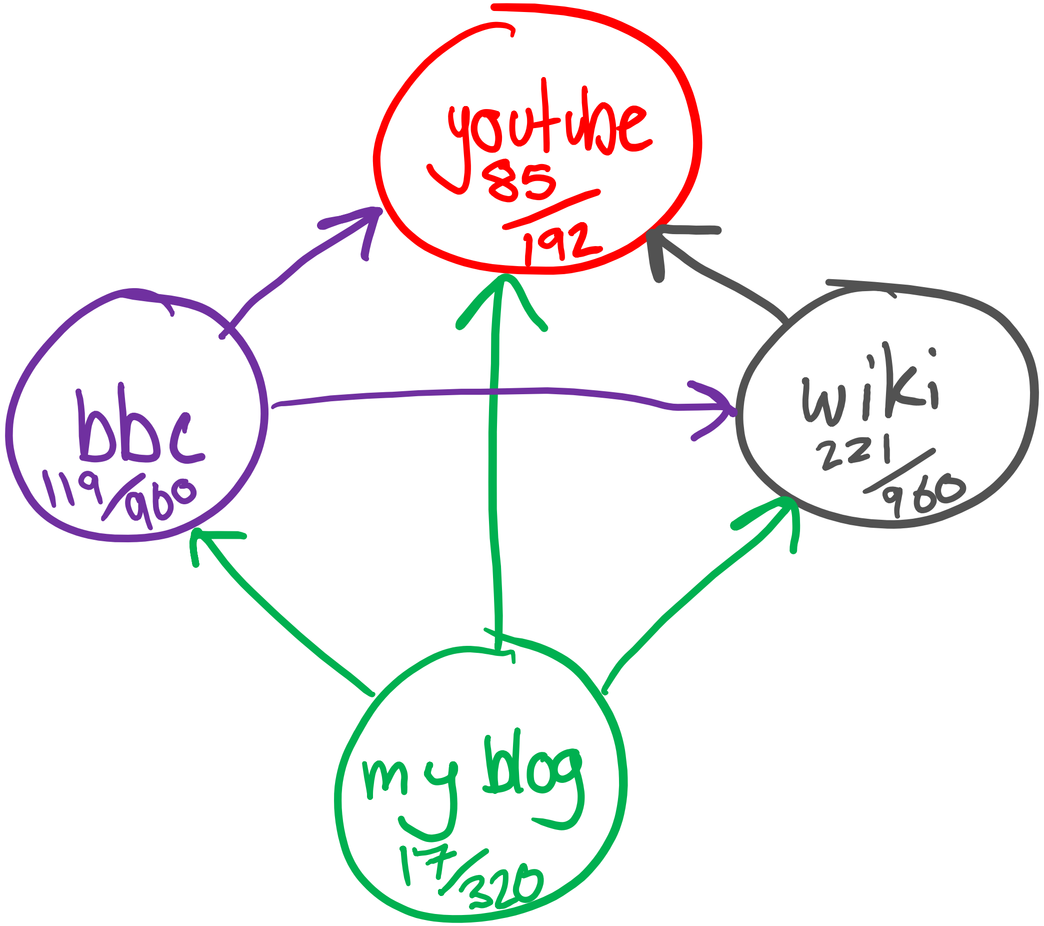
- Now we have a more accurate measure of the probability of a link being followed
- But our total probability doesn’t add up to 1
- Before our total was 1, we multiplied everything by 0.85, so now our total is $1 \times 0.85 = 0.85$
- To fix this, we can simply add on whatever extra we need to get back to 1



Another solution

$\frac{85}{192} + 119/960 + 221/960 + 17/320 = 0.85$

- As you can see our total is indeed 0.85, and so we need $1 - d = 1 - 0.85 = 0.15$ to get back to 1
- An easy way to do this, is simply to share the missing *(1-d)* score between all our websites, so we add on *(1-d)÷n* to each site's score
- We therefore add $\frac{1-d}{n} = \frac{1-0.85}{n} = \frac{0.15}{4} = \frac{3}{80} = 0.0375$



Another formula update

- To make one final change to our formula, we now need to add on our *(1-d)÷n* to each site

$$S(Y) = \frac{1-d}{n} + d\left(\frac{S(B)}{L(B)} + \frac{S(M)}{L(M)} + \frac{S(W)}{L(W)} + \frac{S(Y)}{L(Y)}\right)$$

- Low and behold, we have our final formula to calculate a given site's PageRank score for *n* websites, with a damping value *d*