Thuc Tran

Project Overview:

In this project, I was hoping to figure out what domains were accessible by clicking on a link on the main page of the website, and what domains were accessible via those domain's homepage. I had hoped to see if I can mimic something similar to the 6 degrees of separation kind of network effect where you can see an exponentially larger number of websites as you increase the degrees of separation.

Implementation:

The project works by first creating a list of websites, degree of separation, and their origins. This was where the meat of the project was done. This was executed via a breadth first searches, where we get all the links on one level of separation before incrementing to the next level of separation. This would prevent problems where our registration of a site is not the closest path.

As the program goes through and collects links, it also normalizes the URL's so that the Pattern API may be able to actually download the HTML for it. Then there was a function to find the URL's on that page, which were later evaluated for validity, and added to the list as a tuple that also included its separation number and the site that lead to it. These URL's were also added to a to-do list that would be looked at for collecting links on the next level of separation.

After the collection of links, I had used networkX to plot the graph as a directed network graph with the Origin Node being a dark blue, and arrows pointing out to child nodes, getting redder as one gets further from the origin. This was the visualization step.

An interesting choice to handle at this point was whether or not to register websites to the list before they were checked for validity. Due to the requirement of keeping a record of the origin website that directed me to a website, I had chosen to just add them in as the program iterated through the origin. This would necessitate a later-on validity check of all the websites in the list. Although it works, it may be worthwhile in the future to try to review this choice, due to re-downloading the page as a way to check for validity appearing to be a computationally slow operation.

Results:

Unsurprisingly, as one gets further from the origin node, the more and more elements there are. For example, I found that for the first 3 degrees of separation for google.com, there was at most 6 websites in a level. However, for the next three after that, we had 17 sites for 4, 44 sites for 5, and 133 sites for 6 degrees of separation. This was expected, but at the same time, somewhat disappointing, due to a much larger expectation. This is most likely associated with the fact that google.com does not seem to actually have very many external links.

Somewhat surprisingly, we find that some of the big websites that are known for having many links such as google.com or wikipedia.org have very few actual nodes to other pages on their main page. In particular, we found that google.com had just one-two external websites on their homepage to a website not in the form ~.google.com. That one website was youtube.com or schema.org. This may reflect the different websites self-reliance on their own services.

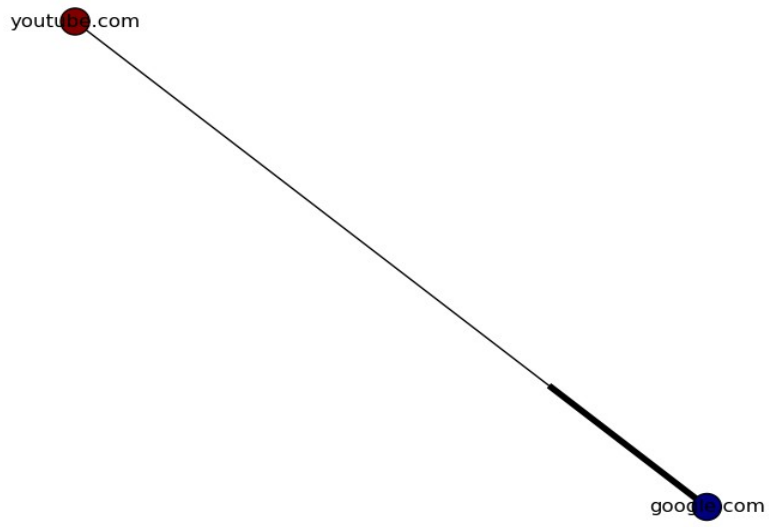

Figure 1: 1 Degree of Separation for Google.com

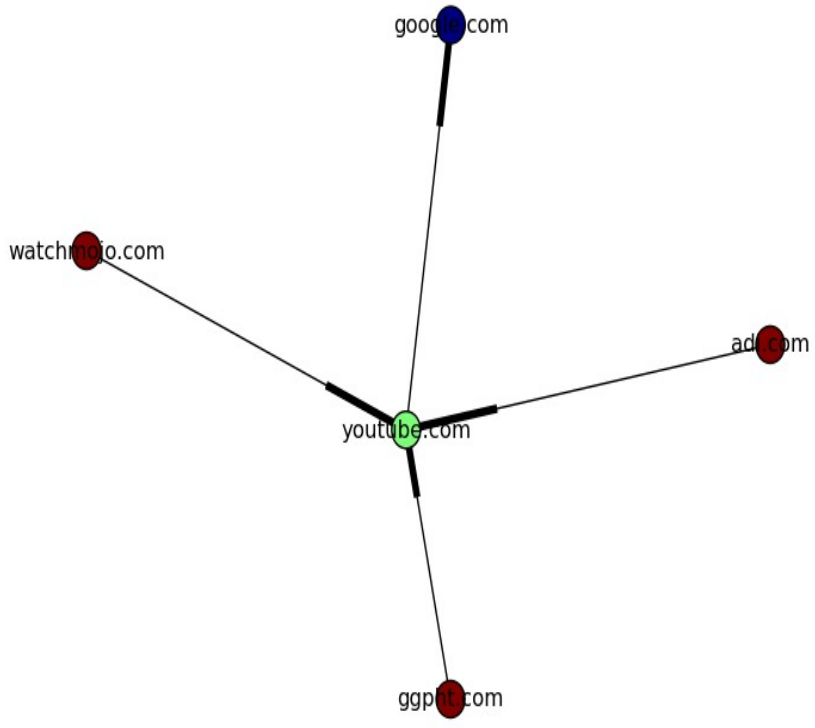

Figure 1: 2 Degree of Separation for Google.com

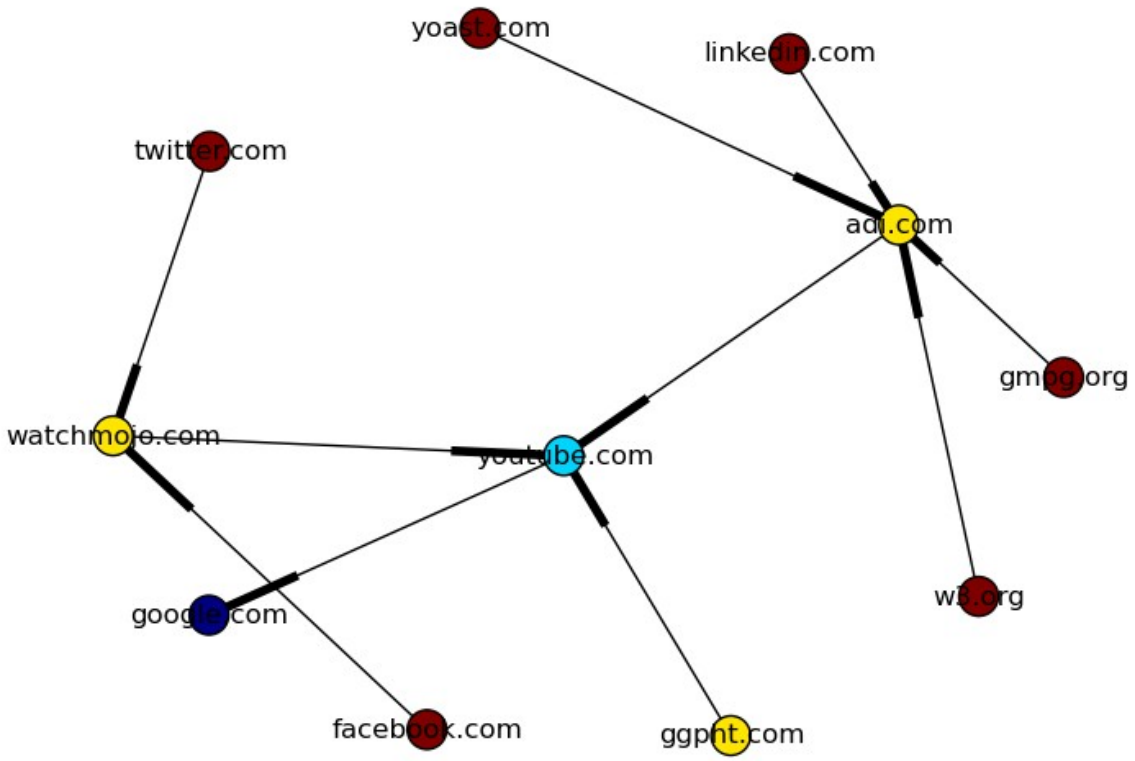

Figure 3: 3 Degree of Separation for Google.com

Reflection:

I think initially, I had mis-scoped my project due to wanting to have a record of every link and vertex. However, the computer hardware made that a bit difficult to do that in a reasonably fast way. Hence, a lot of the code had to be written, and more checks to limit the volume of websites. In addition, for my process, I think that writing down pseudo-code or even just running through an example run of each function manually via python had helped me to at least crystallize what issues that I would be able to expect, and what kind of cases to handle. As it turns out however, after a later-on commenting and refactoring of code, that a lot of these cases became extraneous . In addition, I had spent a large amount of time hunting down particular issues that may have been handled more easily with a better testing plan. i.e. (www..org and www..com were coming up as valid links for a while)

Figure 4: 6 Degrees of Separation for Google.com