Please **read** this handout "before-or-after" the second lecture (Wednesday of the first week of this module). **It is not assigned for grading**, though it was part of homework last year. It will be previewed and then reviewed in lectures; we encourage further exploration by you.

## Internet Search

The lecture on Wednesday Week 1 (of this unit) asserts that the key innovation of Google's search engine in the 1990s was to "simulate a *random walk* on the graph of the Internet." This is used to compute *ratings* of webpages, which are then used to *rank* the top 10, 20, 30, ..., however-many pages. The name *PageRank* was used for the initial patent—partly from the coincidence of one of its inventors being named Larry Page. However, the details of Google's algorithm have changed significantly over the quarter century since its inception, and the initial patents expired in 2019.

The object of this exercise is to see whether this idea of how Google's search works is easy to verify on the Internet. A major "meta" aspect is that we will specifically use Google's search to try to do so, and the behavior we are trying to verify will influence the results we get. A second part will examine Google's results on what seems to be a simpler current factual question, about which data center is the world's largest. You are not being graded on finding uniquely correct answers to these questions, but on your own attempts to do so and your expressed understanding of what you see.

To set the context, the three main ingredients of an Internet search engine are:

1. A comprehensive index of publicly-accessible webpages. Needs computer space to store it and computer power to process it.
2. A basic initial scoring metric of the relevance of a particular webpage to a particular search query.
3. An algorithm for further processing of webpages and their scores.

What lecture said about *PageRank* pertains to point 3: after the initial scores filter pages by (non-)relavance, the random walk gives the final ratings. There are other algorithms, but Google's is definitely the elephant in the room. Like the 2,500-year-old parable of the blind men and the elephant, different sources will give you different takes on what this essence *is*.

### Part 1

Enter simply `PageRank` into Google. One word, no space, capital `R` as well as capital `P`. Chances are your top hit will be Wikipedia's article of that name. **Do not** read it. The objective is to read the takes on *PageRank* by people trying to be simpler and crisper than Wikipedia. Likewise, skip three or so "Scholarly articles" that Google will offer. But do click under "Things to know" and see if the sources there coincide with some of the regular hits that follow below. Between that and the first two pages of hits, you will likely see some or all of the following, not necessarily in this order:

- A page from `Semrush.com`, a large company that sells tools to businesses to enhance traffic to their webpages.
- A page from `searchengineland.com`.
- A page from `Geeks for Geeks`, which is a coder's version of "How Things Work."
- Two pages from `AHrefs.com` (a company like Semrush): a long new one titled "Evolution of Google PageRank" and an older one saying "PageRank is NOT Dead."
- An article by "Centic" on *Medium*—which might be paywalled. It uses the same graph network diagram as I took from Wikipedia.
- An article by SEO.AI, which like Semrush is concerned with "Search Engine Optimization."
- An article by `positional.com`, another SEO company.
- An article from Amrani Armine of `towarddatascience.com`.
- A three-page handout from Stanford's CS54N class.
- A one-page definition from `WhatIs.com`.

There will also be some PageRank checker apps—we can ignore those—and some more technical sources. You are welcome to consult the latter as well if you wish.[1]

**Your task**: Pick *three* of the above or similar hits that you get (again, not Wikipedia), and for each one, find *three* short phrases or single words that encapsulate how it describes *PageRank*. Your words can be quoted verbatim or paraphrased. If a concept like "Markov chain" seems to be important but you don't know what it is, you can quote it without having to look it up. If you see an equation of the form $PR(A) = \ldots$, pick a phrase on the page that describes what the equation does. For example, here is a more-technical hit I did not incude above: https://neo4j.com/docs/graph-data-science/current/algorithms/page-rank/, for which one might say:

- Neo4j: (a) "A page is only as important as the pages that link to it." (b) "equation is used to iteratively update a candidate solution." (c) Bad stuff occurs when groups of pages have no outgoing links.

Then write a short paragraph as an "executive summary" of what you get about *PageRank* from these sources. If terms that are common to two or more of your sources strike you as important, be sure to include them.

Finally, give your vote as to which of the hits (not Wikipedia's) is the best single source to gain a quick and serviceable understanding of *PageRank*. (The "meta" aspect here is your functioning as a human page-ranker. (There were previously graded points for this.)

[1] I'll add that my own PhD graduate Dr. Arun K. Jagota writes copiously for Toward Data Science, and he has one PageRank example: https://medium.com/towards-data-science/pagerank-illustrated-c056a45a2f60. I've co-written an article on PageRank myself, https://rjlipton.wpcomstaging.com/2014/07/21/shifts-in-algorithm-design/, which has some Pythonic homor. But these do not show high in hits.

**Part 2**

This is a short application, to see if the understanding gained from part 1 is enough for insight. It is intended to have a concrete point *beyond* how rumors mushroom on the Internet. This further point is whether the treating of certain webpages as authoritative sources gets knocked on, and how-and-why Google's algorithm might magnify that.

In my original 2017 edition of the "Internet and Data" slides, I felt on firm enough ground to quote "30%" as the percentage of Web *traffic* consisting of pornography. I linked a secondary source https://ourworldindata.org/internet, from which all mention of porn seems now to have disappeared, but I noted its reference to a 2012 article by Sebastian Anthony of Extreme Tech. Let's see if this figure is still tenable now in 2024.

**Content Note:** The assignment has been structured to avoid getting sexual content in hits. Two racy cartoon images crop up, one of of people dressed for "S&M" in a more-silly-than-sexual way, and the other (used by Huffington Post and credited to a website called PaintBottle that has since disappeared) has fully-dressed figures with mouths agape. No issues were reported in 2022 or 2023. For contrast, if you were to try the J.D. Vance "couch" business, or "chess beads meme" for a story I was actually involved in, you would find lots of "sketchy" comments (in X and Reddit threads for instance) that cannot be controlled for. Search results are variable by person and time, however—so if anything else/worse does come up for you, please let us know. The words `traffic` and `bandwidth` are synonymous with `data` for our purposes, and the former seems *not* to bring up pages with "sex trafficking," but I've avoided it anyway. Look for all three terms but don't care as much about the count of *websites*, because one website could generate a lot of traffic.

1. Enter the search `Internet data percent porn extremetech` (without quotes) into Google. You should see the Extreme Tech link and a BBC link above or below it. Note how the BBC story says that the Extreme Tech article is "regularly quoted for calculating that 30% of all net traffic is generated by porn sites." Count at least 3 other hits that reference Extreme Tech—you can tell from the capsules and need not click on them.

2. Now click the Extreme Tech article, which is by Sebastian Anthony. *Skim* it, but note the hedging in the very last few sentences of the article.

3. Now enter the search `Internet data percent porn`, that is, without `extremetech`. See from the capsules how many hits say "30%" and whether they reference PaintBottle/Huffington Post, Anthony/Extreme Tech (I don't know if the former drew from the latter.) or a 2019 article by the website Statista. Then see if inserting `30` before `percent` in the search changes much. Note the years on the hits, whether any are later than 2017.

Finally, write a paragraph on what you observed about hits referencing each other, either with links or not, and their dates. Give your thoughts on how the nature of Google search may have influenced what you got as hits, and on how reliable you think the "30 percent" figure is now. (Again, there were previously points for this.)

**Part $-\infty$ (Not Graded—but please glance at this as it will be picked up in later lecture discussion)**

And for something else, go to https://wifitalents.com/statistic/pornography-industry/. Note the August 7, 2024 (or later) update stamp. Browse the statistics—looking for '30%' in particular—and then *see if you can get anything out of how they are sourced.*

Then go to https://worldmetrics.org/porn-use-statistics/, which as of this writing has a 7/23/2024 publication date. Here you'll have to click thru the statistics one-at-a-time, like with a slideshow. Are they the same? Same-ish? It is a loop of 40 stats, so if you go backward from 1 you will reach stat #29 with the 30% figure sooner. Again *see if you can get anything out of how they are sourced.*

If you succeed in finding either site's exact sources, let me know. Otherwise, think about implications for truth ever finding a firm root system on the Internet. (This extra new development is why we kept the "porn" example for this HW in Fall 2024. This was not graded for points even last year.)