Lecture 2 - *Google's PageRank: Why math helps*

OBJECTIVE: Motivate linear algebra and probability as important and necessary tools for understanding large datasets. We also describe the algorithm at the core of the Google search engine.

◇ PAGERANK

Consider the following mini-web of 3 pages (the data):

The nodes are the webpages and the arrows are links. The numbers are the normalised number of links. We can re-write this directed graph as a **transition matrix**:

⋆

T is a **stochastic matrix**: its columns add up to 1, so that

$$T_{i,j} = P(x_j|x_i)$$

$$\sum_j T_{i,j} = 1$$

In information retrieval, we want to know the "relevance" of each webpage. That is, we want to compute the probability of each webpage: $p(x_i)$ for $i = 1, 2, 3$.

Let's start with a random guess $\pi = (0.5, 0.2, 0.3)$ and "crawl the web" (multiply by T several times). After, say $N = 100$, iterations we get:

$$\pi^T T^N = (0.2, 0.4, 0.4)$$

We soon notice that no matter what initial π we choose, we always converge to $p = (0.2, 0.4, 0.4)$. So

⋆

$$p^T T =$$

⋆

The distribution p is a measure of the relevance of each page. Google uses this. But will this work always? When does it fail?

⋆

The **Perron-Frobenius Theorem** tell us that for any starting point, the chain will converge to the invariant distribution p, as long as T is a stochastic transition matrix that obeys the following properties:

1. **Irreducibility**: For any state of the Markov chain, there is a positive probability of visiting all other states. That is, the matrix T cannot be reduced to separate

smaller matrices, which is also the same as stating that the transition graph is connected.

2. **Aperiodicity**: The chain should not get trapped in cycles.

Google's strategy is to add am matrix of uniform noise E to T:

$$L = T + \epsilon E$$

where ϵ is a small number. L is then normalised. This ensures irreducibility.

How quickly does this algorithm converge? What determines the rate of convergence? Again matrix algebra and spectral theory provide the answers:

⋆