# Algorithms used in the project

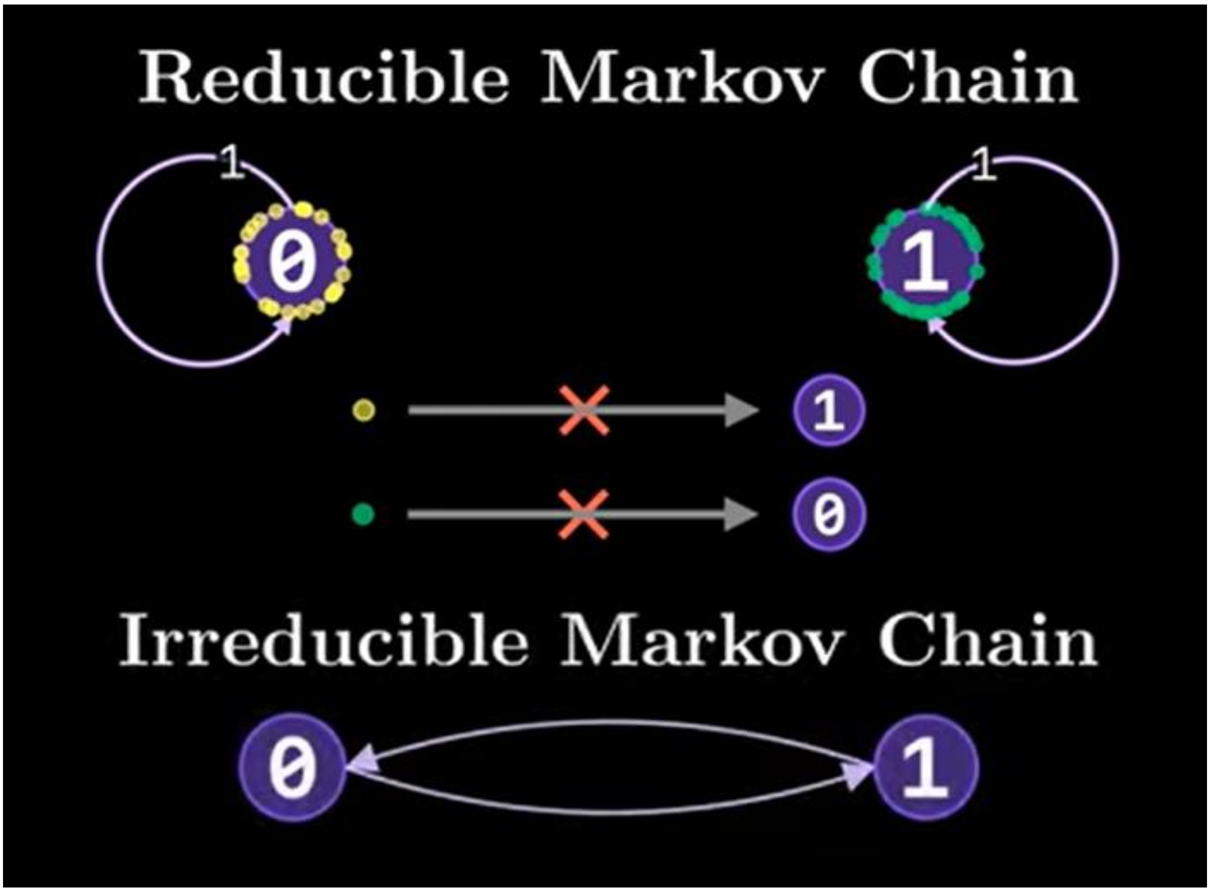
## PageRanker Algorithm:

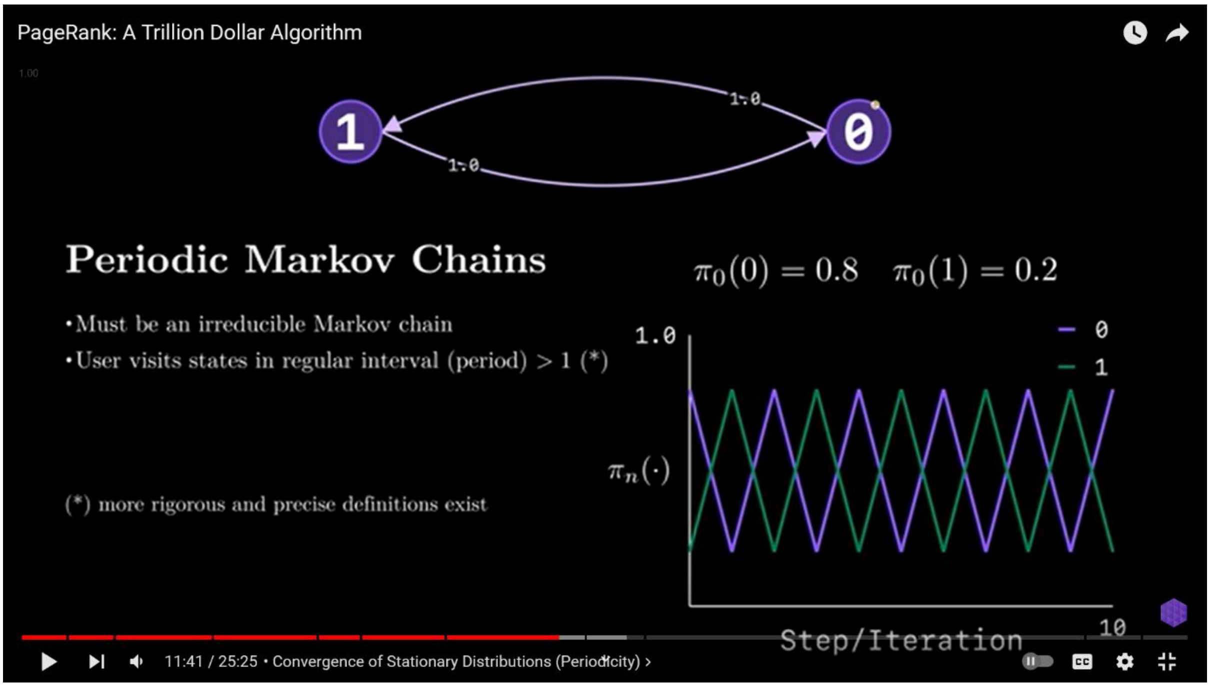
We used a simple pageRanker algorithm, where we calculate a transition matrix – one that represents connections between links through a matrix – and iteratively multiply it by a vector of popularity (initialized as 1/n for each element, where n is the total number of links (6000)).

This algorithm has 2 main assumptions:

1- The Markov chain is irreducible.
2- The Markov chain is aperiodic.

Which means that it must be ergodic, else it bears the risk of never converging.





PageRank: A Trillion Dollar Algorithm

Periodic Markov Chains

- Must be an irreducible Markov chain
- User visits states in regular interval (period) > 1 (*)

(*) more rigorous and precise definitions exist

$\pi_0(0) = 0.8 \quad \pi_0(1) = 0.2$

11:41 / 25:25 • Convergence of Stationary Distributions (Periodicity) >

For a Markov chain to be irreducible, it must avoid having independent graphs, as for the periodicity, it means that the popularity score never converges, but instead keeps oscillating periodically. As we have no means of guaranteeing either of these conditions, we simply place a limit on the number of iterations (1000 in our case), if the popularity vector has not converged by then, we will simply break the loop and set the popularity to the values obtained by then.

We check the divergence by calculating the distance between the old popularity vector and the new one (i.e., the dot product of the difference between the 2 vectors) and comparing it to a threshold.

Normally, multiplying 2 matrices requires a complexity of O(N^3) (or a bit less using an optimized algorithm). However, in case of multiplying a matrix by a vector, the complexity can be reduced to O(N^2)

## Relevance Calculation:

The relevance is obtained by calculating the term frequency of each word (normalized) and the IDF, and then multiply them together.

## isSpam Calculation:

We compare the value of the TF to a certain threshold (0.2 in our case). If it surpasses it, then we consider it spam, otherwise it is not spam.