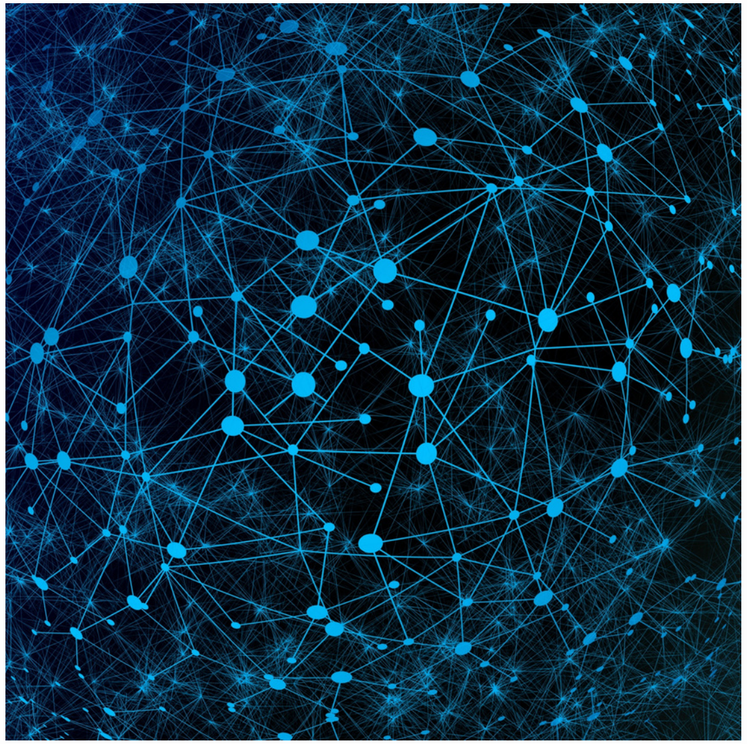
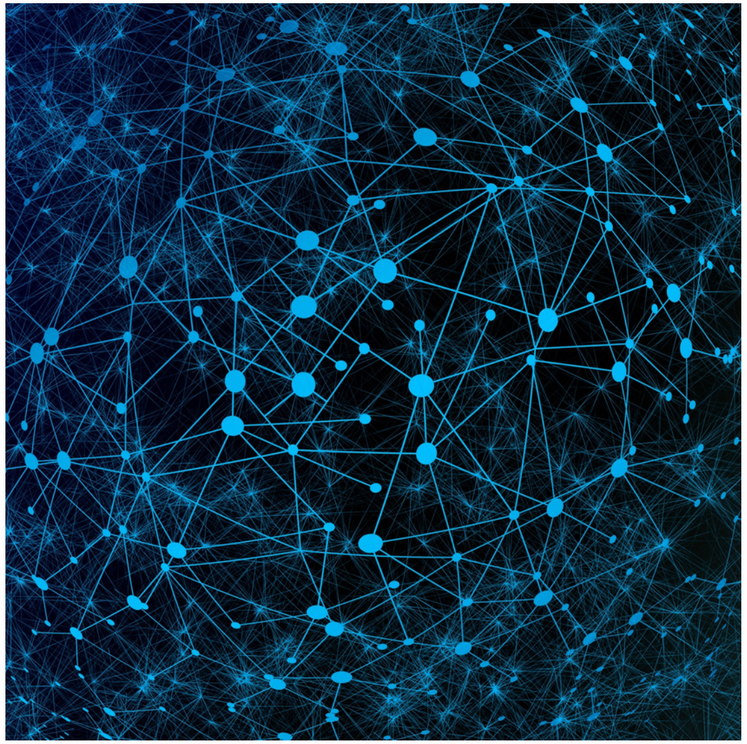
WEB MINING: PAGE RANK

OPREA OLIVIA MARIA-MAGDALENA
IA1B



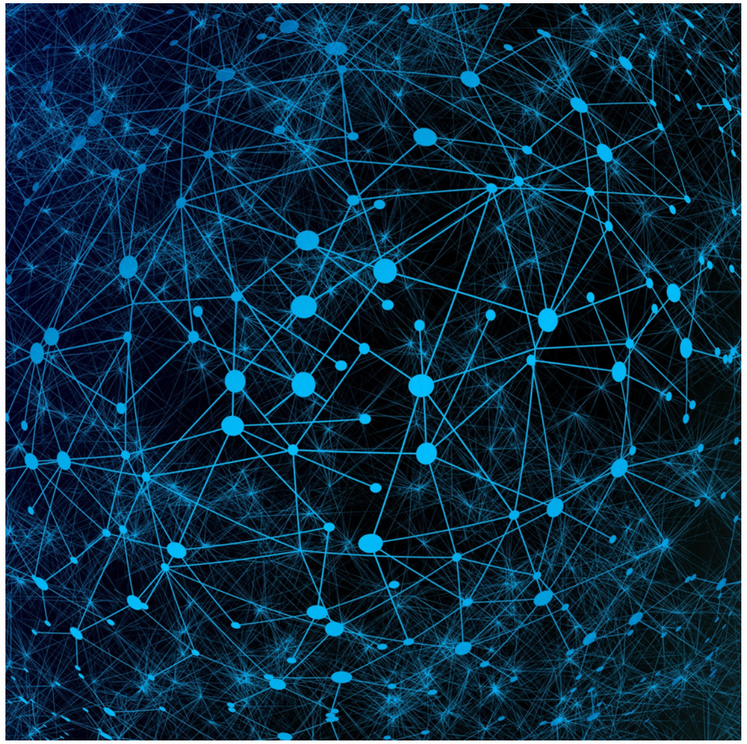
INTRODUCTION

- Web mining is the use of data mining techniques to automatically discover and extract information from web documents and services.
- This can include various types of data, such as web content, web structure, web usage, and other forms of web-related data. Web mining is typically used to identify patterns and trends in data, which can be used to improve the performance of web-based applications and services. Some examples of web mining applications include search engines, web-based recommendation systems, and fraud detection systems.
- One of the most well-known applications of web mining is the calculation of page rank, which is used by the popular search engine Google to help determine the importance or relevance of a particular web page.



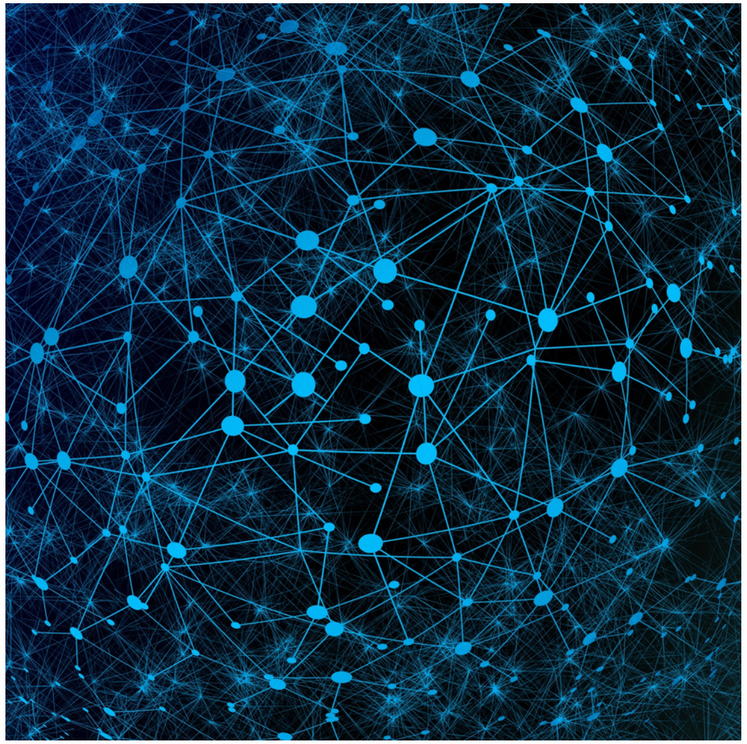
WHAT IS PAGE RANK?

- Page rank is a mathematical algorithm that assigns a numerical weighting to each element of a hyperlinked set of documents, such as the World Wide Web.
- The purpose of page rank is to "measure" the relative importance or relevance of a particular web page within the set.



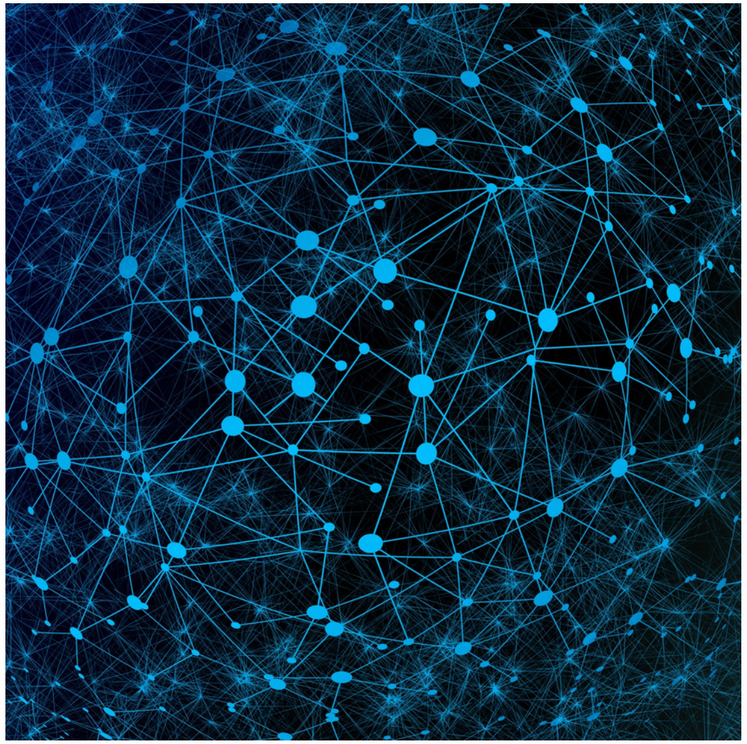
HISTORY OF PAGE RANK

- The page rank algorithm was developed by Google founders Larry Page and Sergey Brin as part of their research project while studying at Stanford University.
- The basic idea behind page rank was to use the link structure of the web to determine the importance of individual web pages.



HOW DOES PAGE RANK WORK?

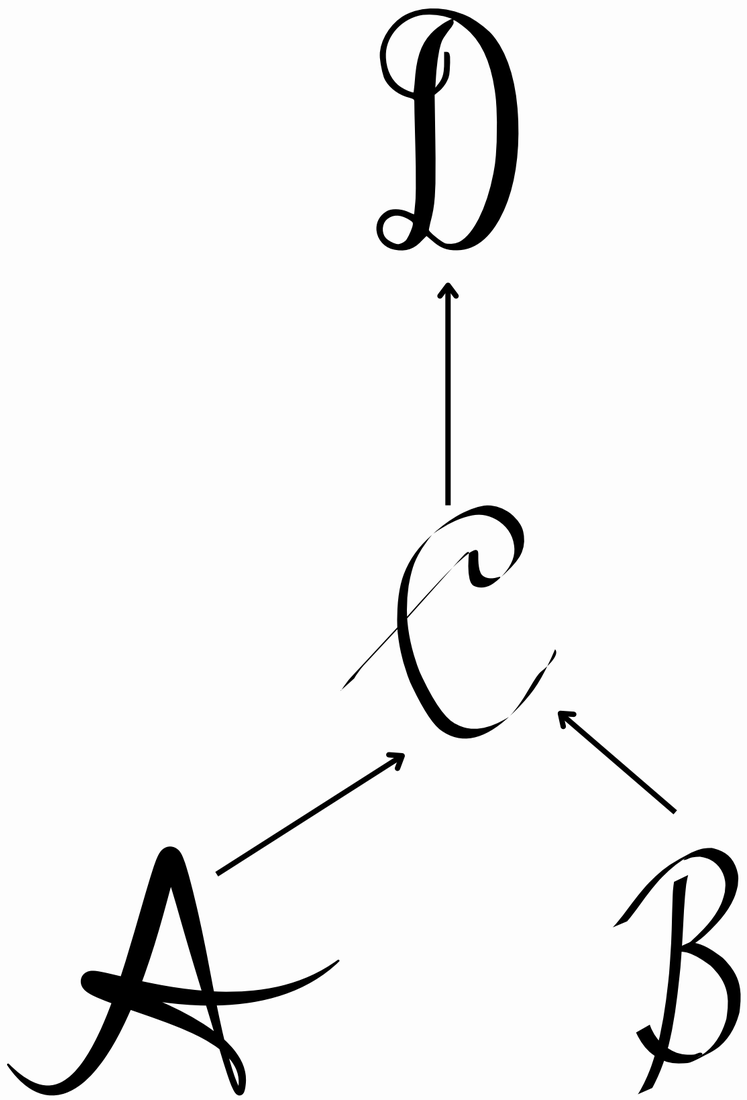
- Page rank works by considering the "votes" or links that a web page receives from other pages on the web.
- More important or relevant pages are likely to receive more links from other websites.
- The page rank algorithm takes into account the quantity and quality of these links, as well as other factors such as the relevance of the linking website and the relevance of the linked-to page to the search query.



PAGE RANK ALGORITHM

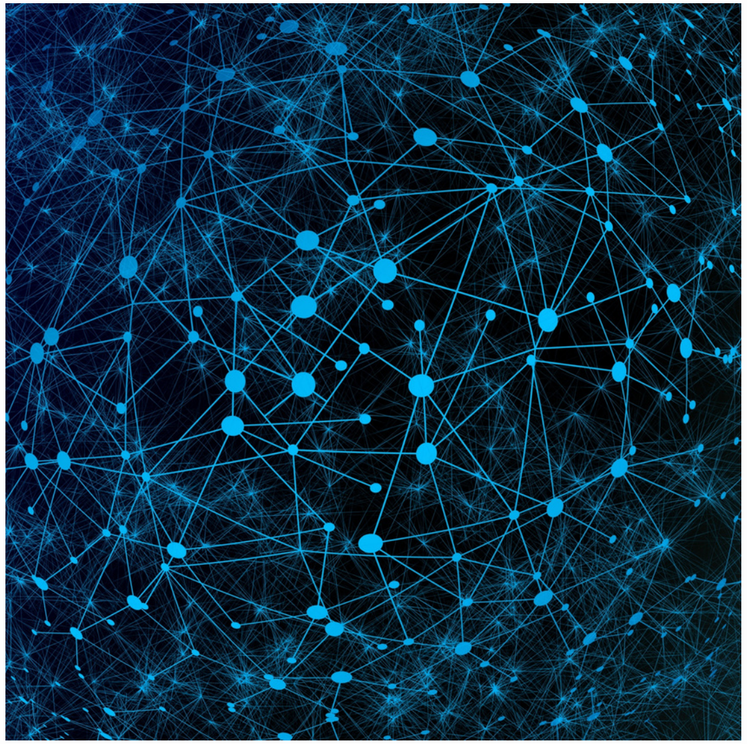
- The page rank algorithm uses a combination of a "damping factor" and the "random surfer model" to calculate the importance of a web page.
- The damping factor is a value between 0 and 1 that is used to simulate the probability that a user will continue clicking on links from one page to another.
- The random surfer model simulates the behavior of a hypothetical user who randomly clicks on links from one page to another, with the probability of clicking on any given link determined by the page rank values of the source and destination pages.

EXAMPLE



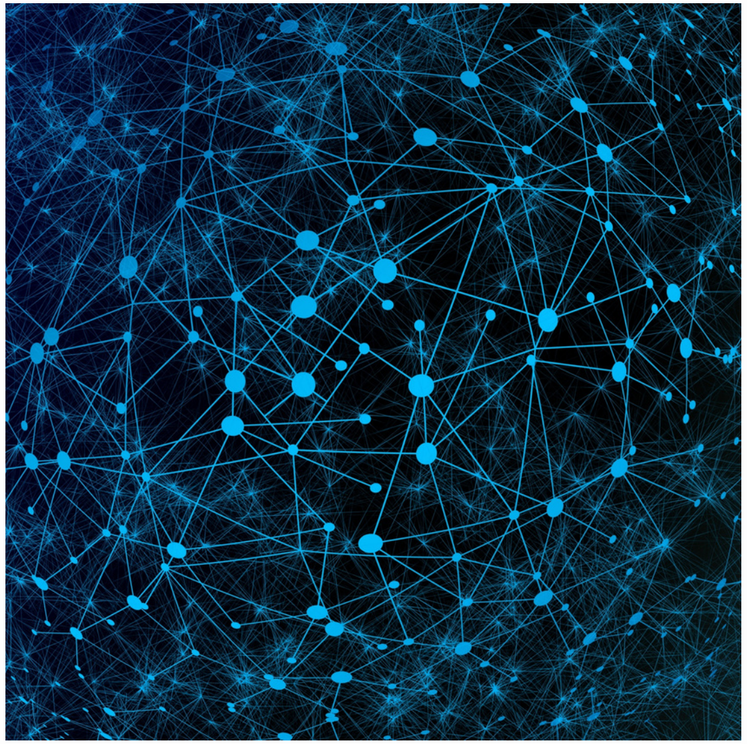
Here is a simple example of how the page rank algorithm might work:

1. Imagine we have a simple network of four web pages, A, B, C, and D, where A and B both link to C, and C links to D.
2. The page rank algorithm would begin by assigning an initial page rank value to each page, based on the number and quality of the links pointing to it. For simplicity, let's say that each page starts with a page rank value of 1.
3. The algorithm would then apply the damping factor to each page's page rank value, to simulate the probability that a user will continue clicking on links from one page to another. Let's say that the damping factor is 0.85.
4. The algorithm would then apply the random surfer model to determine the new page rank values for each page. In this case, the model would calculate the probability that a user would randomly click on a link from A or B to C, and from C to D. These probabilities would be based on the page rank values of the source and destination pages, as well as the damping factor.
5. The algorithm would then use these probabilities to calculate the new page rank values for each page. For example, the page rank value of C would be the sum of the probabilities that a user would click on a link from A or B to C, multiplied by the page rank values of A and B, and multiplied by the damping factor.
6. This process would be repeated iteratively until the page rank values of all pages converge to a stable value.



LIMITATIONS OF PAGE RANK

- Page rank is just one of many factors that search engines use to determine the relevance and importance of web pages.
- Page rank can be manipulated through the use of link schemes and other tactics.
- Page rank is not always a reliable indicator of the true importance or relevance of a web page.



CONCLUSION

- Web mining is the use of data mining techniques to automatically discover and extract information from the web.
- Page rank is a well-known application of web mining that is used by search engines to help determine the importance or relevance of a web page.
- While page rank can be a useful tool, it is important to remember its limitations and the potential for manipulation.

THANK YOU