Analysis of High-energy physics theory citation network

*Note: Graph Analysis of High-energy physics theory citation network using GraphX Apache Spark

1st Harsh Rawat *20BCS050*
2nd Ashwani Kumar *20BCS023*
3rd Jayant Kumawat *20BCS064*
4th Anil Kumar *20BCS015*
5th Kartik Bhamare *20BCS066*

***Abstract*—In the realm of high-energy physics, understanding the significance and interconnectedness of scholarly papers is of paramount importance. This study utilizes the renowned SNAP dataset, which represents the citation network of scientific papers from arXiv's "High Energy Particle Physics" section. To delve into the intricate web of these citations, we employed Apache Spark's GraphX library, a prominent tool tailored for graph computation. Two fundamental algorithms, PageRank and Connected Components, were implemented using Scala. The former, PageRank, was harnessed to discern the most influential papers within the dataset by evaluating the prominence based on their citation patterns. The latter, Connected Components, was utilized to identify clusters of interrelated papers, shedding light on the major themes and interconnected topics within high-energy physics. The results from this study not only provide a comprehensive view of the citation landscape in the field but also showcase the prowess of scalable graph processing tools in extracting meaningful insights from large-scale scientific datasets.**

***Index Terms*—Apache Spark, GraphX, Connected Components, PageRank**

I. Dataset

In the given study Dataset [1] from snap Stanford is utilized. The High-energy physics theory citation network originates from the arXiv's "High Energy Particle Physics" section. The dataset captures the citation network of scientific papers, where an edge from paper A to paper B signifies that paper A cites paper B. It provides valuable insights into the evolution and dynamics of scientific research in the realm of high-energy physics.

Feature	Value
Nodes	27770
Edges	352807
Nodes in largest WCC	27400 (0.987)
Edges in largest WCC	352542 (0.999)
Nodes in largest SCC	7464 (0.269)
Edges in largest SCC	116268 (0.330)
Average clustering coefficient	0.0.3120
Number of triangles	1478735
Fraction of closed triangles	0.04331
Diameter (longest shortest path)	13
90-percentile effective diameter	5.3

TABLE I
Dataset Statistics

This dataset is of particular interest to researchers aiming to understand the evolution of scientific knowledge in the field of high-energy physics, The influence and importance of particular papers or sets of papers, The patterns of collaboration and citation within the scientific community.

Given Dataset is in the form an edge list. The methodology followed in reading an preprocessing data is as follows:

- Created a local Spark stand-alone cluster.
- Read edge list using GraphLoader API.
- Created a Directed Graph using Edgelist.

The Following Dataset has been extensively studied by the community and acts a good benchmark Dataset and hence is chosen for analysis. The dataset is placed in the Dataset folder in the parent directory.

II. Algorithm

The following Algorithms have been used:

A. PageRank [2]

PageRank is an algorithm developed by Larry Page and Sergey Brin, the founders of Google, originally designed to rank web pages based on their importance and relevance. In the context of the High-energy physics theory citation network, the PageRank algorithm can be applied to rank scientific papers instead of web pages.

Just as the original PageRank algorithm assigns a value of importance to web pages based on the number and quality of links to them, when applied to the SNAP dataset, it can determine the importance of a paper based on the number and quality of its citations. A paper that is frequently cited by other influential papers will have a higher PageRank.

B. Connected Components [3]

The Connected Components algorithm identifies sets of nodes that form a subgraph where any node can reach any other node in the same set, directly or indirectly, and no node in one set can reach a node in another set. When applied to a citation network like the High-energy physics theory dataset, the algorithm can reveal interesting insights.

Using Connected Components on the High-energy Physics Theory Citation Network, we can identify clusters or groups of papers that are interconnected through citations, implying they might be closely related in content or topic.

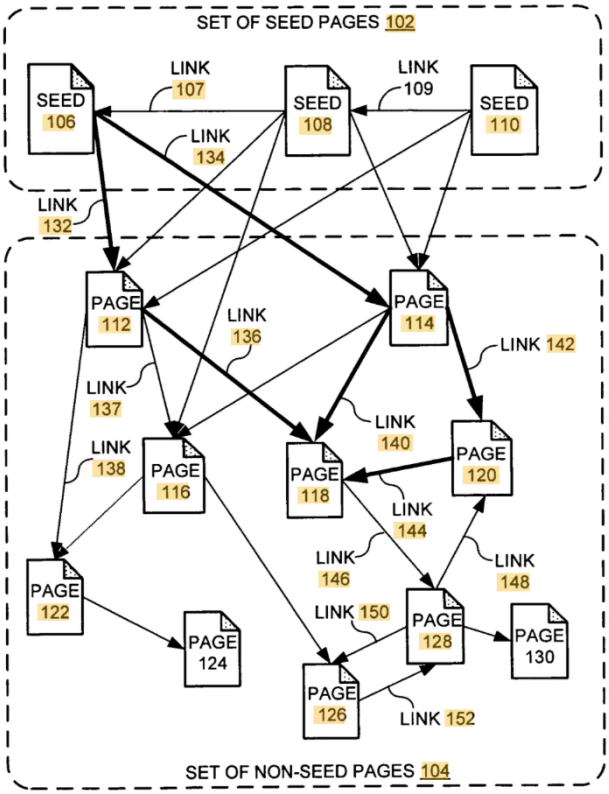

Fig. 1. PageRank

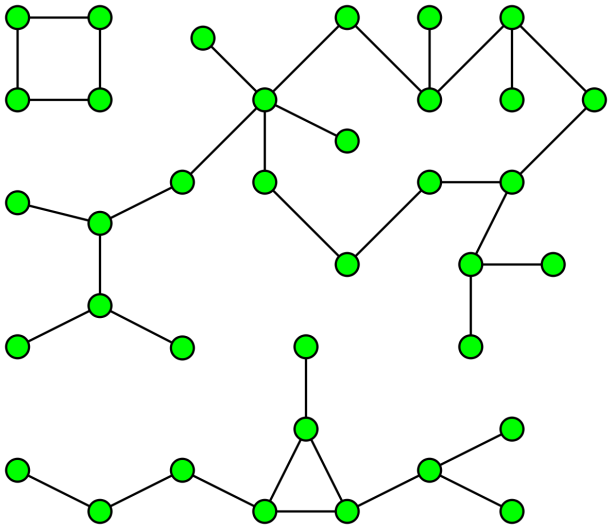

Fig. 2. Connected Components

III. Results

A. PageRank

After running the pagerank algorithm we saved the results in the results directory and in fig-3, we made a table of the top 10 page ranks.

By applying PageRank periodically (e.g., yearly), one can track the rising or waning influence of specific papers or topics in high-energy physics. This can help in identifying emerging trends and fading research area

User	PageRank
9207016	172.9800681646912
9407087	168.96007611985553
9201015	156.57264538289508
9503124	124.11496058644649
9510017	116.90396839658722
9402044	106.09962165392056
9711200	93.51765670513663
9410167	91.36766427200433
9408099	86.76597646942517
9402002	80.40654776355734

Fig. 3. PageRank(10)

B. Connected Components

Seeing fig-4 After running the connected components algorithm, we made a table of top 5 results.

User	Connectecd Component
9910051	9910009
9910009	9910009
9903171	9903171
9905060	9903171
9903179	9903171

Fig. 4. ConnectedComponents(5)

References

[1] Paper: hep-th/0002031 From: Maulik K. Parikh Date: Fri, 4 Feb 2000 17:04:51 GMT (10kb)
Title: Confinement and the AdS/CFT Correspondence Authors: D. S. Berman and Maulik K. Parikh Comments: 12 pages, 1 figure, RevTeX Report-no: SPIN-1999/25, UG-1999/42 Journal-ref: Phys.Lett. B483 (2000) 271-276

[2] https://spark.apache.org/docs/3.0.2/api/scala/org/apache/spark/graphx/lib/PageRank$.html

[3] https://spark.apache.org/docs/latest/api/scala/org/apache/spark/graphx/lib/ConnectedComponents$.html