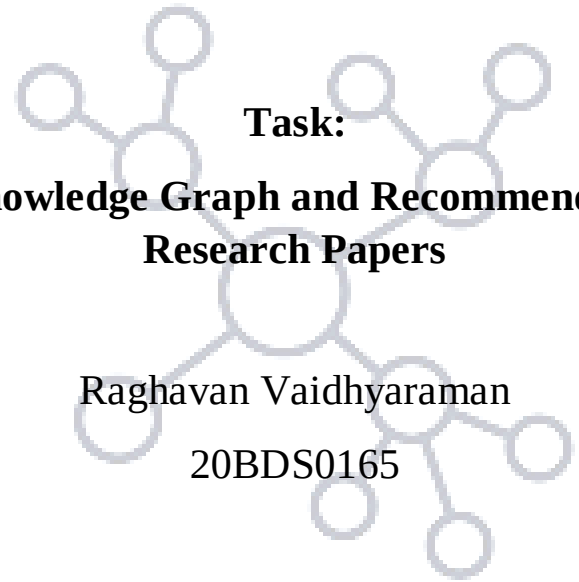
BalkanID

FTE Hiring Task

Data Science Task



Task:

Building a Knowledge Graph and Recommender System for Research Papers

Raghavan Vaidhyaraman

20BDS0165

Computer Science Engineering

Specialization in

Data Science

Github Repository: https://github.com/BalkanID-University/balkanid-fte-hiring-task-vit-vellore-2023-raghavan93513

Problem Statement

Knowledge graphs have emerged as powerful tools for representing and organizing information in a structured and interconnected manner. They enable a deeper understanding of complex domains by capturing relationships and contextual information between entities. By representing knowledge as a graph, with entities as nodes and relationships as edges, knowledge graphs provide a holistic view of the data, facilitating efficient data integration, knowledge discovery, and advanced analytics.

Dataset link: https://bit.ly/balkanid-ds-dataset-download

Part 1: Charting the Research Landscape

You are part of a team of researchers who have embarked on a quest to uncover hidden knowledge and insights within a vast collection of research papers. As you journey through the realm of research, you discover the importance of organizing and modeling metadata effectively. Utilizing the RDF framework and the expressive SPARQL language, build a knowledge graph that represents the interconnectedness of research papers, their authors, citations, and other relevant metadata. Showcase your skills by creating a powerful visualization tool that enables researchers to visualize the results of SPARQL queries as a subgraph.

Note: Draw a diagram to illustrate the ontology of your knowledge graph depicting entities and their relationships using a diagramming software (like lucidcharts, powerpoint, draw.io, etc.). Also, provide at least 5 examples of the visualizations generated by executing SPARQL queries using the visualization tool that you have built.

Part 2: The Search for Knowledge

To aid fellow researchers in their exploration, your task is to design and implement a recommender system that suggests 5 similar and important research papers that a researcher could cite for a new research paper that he/she plans to write. Embrace the power of graph-related algorithms and learning models to generate meaningful recommendations. If implementing graph algorithms for recommendations proves challenging, use any other algorithm, but provide a detailed description of the steps you would take to incorporate graph algorithms in the recommender system.

In new_research_papers.jsonl (https://bit.ly/balkanid-new-data-download) we have provided the title, discipline and abstract of 3 papers that the researcher plans to write. Using your recommender system, for each new research paper, recommend 5 similar and important papers that the researcher could cite (in order).

Note: Importance of a research paper is measured not just by the number of citations that it receives but also by the quality of the citations. The quality of a citation is in turn determined by the number and quality of its own citations.

Part 3: Unveiling the Trailblazers

Problem Statement

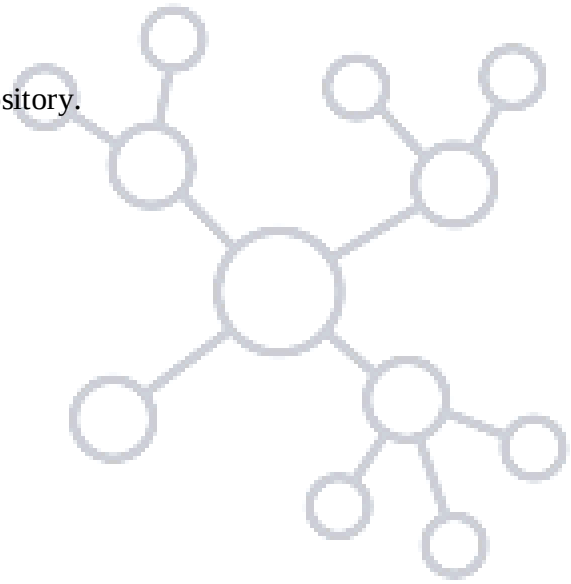
Delve into the heart of impact analysis and embark on a mission to identify the research works that have made the most significant impact on their peers. Explain and implement different methods of evaluating that you have incorporated to calculate how impactful a paper is, in a detailed manner. List the top 5 papers that you think are the most impactful.

Note: Document in a detailed manner, how the impact analysis was executed, along with justifications as to why you chose the methods.

Project is divided into 4 main .ipynb files:-

1. EDA.ipynb
2. Task1.ipynb
3. Task2.ipynb
4. Task3.ipynb

All codes are in the repository.



EDA

This file reads the knowledge graph from a file, analyzes it, and provides various insights such as:

1. **Total Connections**:
   Counts and prints the total number of connections between the nodes in the graph.

```
[2]  ✓ 7.9s
...  The total number of connections in the graph is 1222908
```

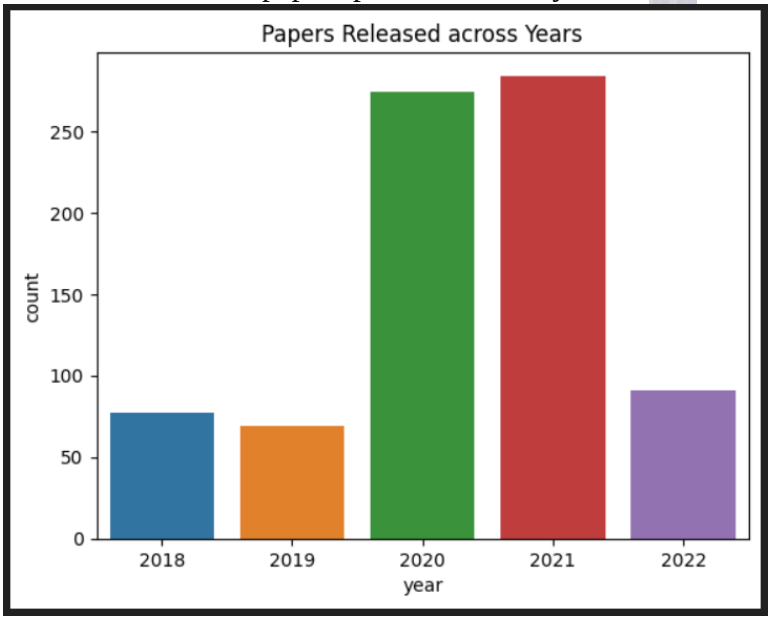
2. **Citation Connections**:
   Checks for specific citation relationships using a defined citation predicate. This gives an insight into how many citation connections exist in the research paper dataset.

```
[3]  ✓ 2.2s
...  The total number of citation connections in the graph is 5745
```

3. **Missing Values**:
   A SPARQL query is used to extract information about papers, such as the title, publication date, and abstract. This data is then used to create a pandas DataFrame. Checks for any missing values in the data.

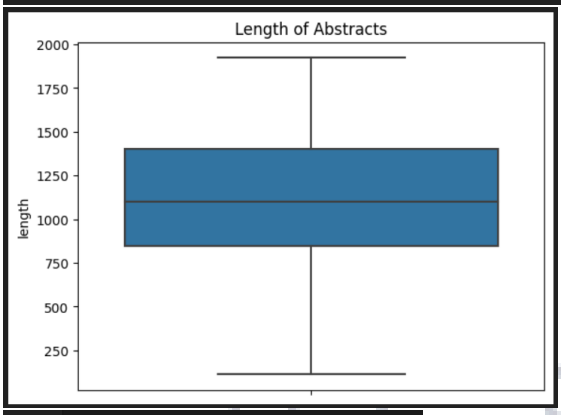
```
any(df.isna().sum())
✓ 0.1s

False
```

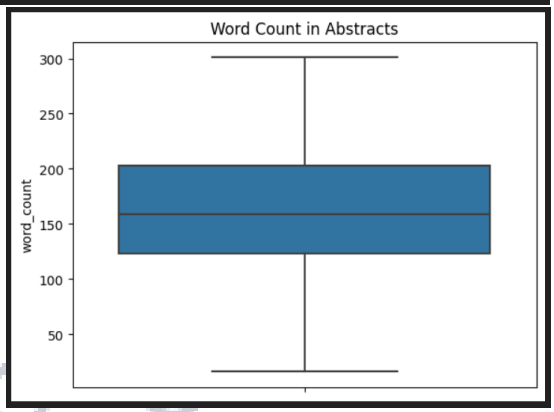
4. **Publication Trend**:
   Plots the number of papers published each year, which can reveal trends over time.



5. **Abstract Analysis**:
   The script also extracts and plots the length and word count of the abstracts. It also calculates the mean and median of the word count.

	title	year	abstract	length	word_count
0	Microfluidic control over topological states i...	2020	Compared to isotropic liquids, orientational...	1064	149
1	Sculpting stable structures in pure liquids	2020	Pure liquids in thermodynamic equilibrium ar...	767	101
2	Locality-Sensitive Hashing for Efficient Web A...	2020	Web application security has become a major ...	1006	151
3	Gravitational wave and electroweak baryogenesi...	2022	We study stochastic gravitational wave produ...	583	78
4	ARA : Aggregated RAPPOR and Analysis for Centr...	2020	Differential privacy(DP) has now become a st...	908	148






```
[11]  ✓ 0.0s
...   Mean of Word Count: 163.76
      Median of Word Count: 159.0
```

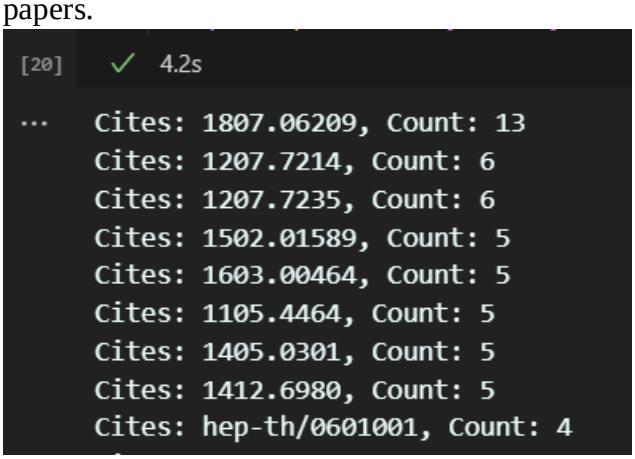
6. **Discipline Analysis**:
   The code finds the number of papers in each discipline and ranks the disciplines by the number of papers. This shows which disciplines have the most research.

```
[12]  ✓ 4.4s
...   Discipline: Physics, Paper Count: 445
      Discipline: Computer Science, Paper Count: 163
      Discipline: Mathematics, Paper Count: 106
      Discipline: Statistics, Paper Count: 43
      Discipline: Quantitative Biology, Paper Count: 18
      Discipline: Electrical Engineering and Systems Science, Paper Count: 16
      Discipline: Quantitative Finance, Paper Count: 4
```

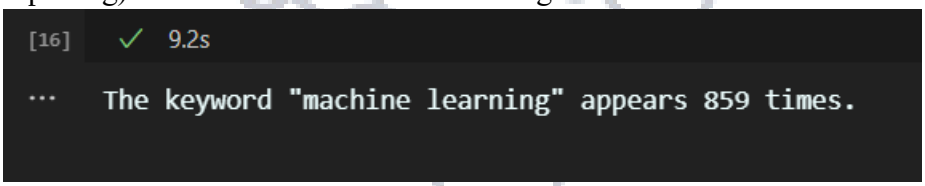
7. **Author Analysis**:
   The code counts the number of papers for each author and shows the authors with the most papers. This can reveal the most prolific authors.

```
[14]  ✓ 4.9s
...   Author: Sansit Patnaik, Paper Count: 7
      Author: Sai Sidhardh, Paper Count: 6
      Author: Kishor D. Kucche, Paper Count: 6
      Author: ATLAS Collaboration, Paper Count: 5
      Author: Fabio Semperlotti, Paper Count: 5
      Author: Siham Aouissi, Paper Count: 5
      Author: Shihao Song, Paper Count: 4
      Author: Anup Das, Paper Count: 4
      Author: Nagarajan Kandasamy, Paper Count: 4
      Author: Markus Sch\"oberl, Paper Count: 4
```

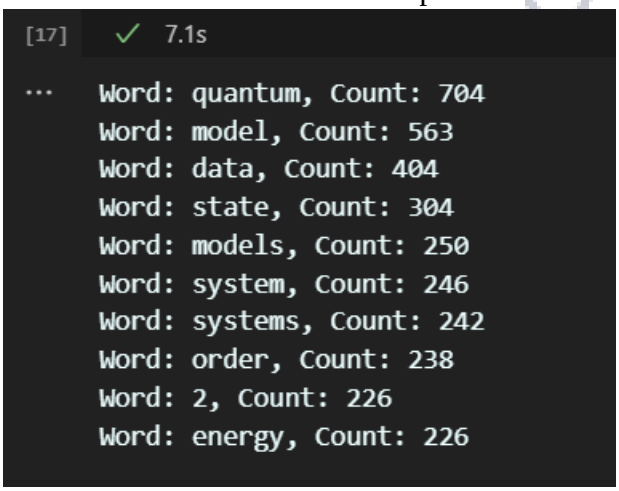
8. **Citation Analysis**:
   It checks the citation counts for each paper and identifies the most frequently cited papers.



9. **Keyword Analysis**:
   The script counts the number of times a given keyword appears in the titles, abstracts, and bodies of the papers. This can help identify the papers most relevant to a specific topic. Eg) Taken here is “machine learning”.



10. **Common Words Analysis**:
    The script counts the number of occurrences of each word in the titles and abstracts of the papers (excluding common stop words), and lists the most common words. This can reveal common themes or topics.



11. **Citation Connection Check**:
    It checks for the existence of citation connections for every node in the graph and prints the nodes that have citation connections.

```
[19]  ✓ 5.2s

...   Node 2011.13660 has citation connection
      Node 2003.05150 has citation connection
      Node 2012.02092 has citation connection
      Node 2011.05889 has citation connection
      Node 2003.13703 has citation connection
      Node 2007.10686 has citation connection
      Node 2007.10686 has citation connection
      Node 2003.05150 has citation connection
      Node 2011.12331 has citation connection
      Node 1805.05212 has citation connection
```

PART - I

RDF and Triples

RDF, or Resource Description Framework, is a standard model for data interchange on the Web. RDF has features that facilitate data merging even if the underlying schemas differ, and it specifically supports the evolution of schemas over time without requiring all the data consumers to be changed.

RDF data is structured in what we call "triples", which is the core concept in RDF. A triple is a basic data record in RDF. Each triple consists of three parts:

1. Subject: This is the entity that the fact or assertion is about. It could be a web resource like a webpage, an XML document, a person, a physical object, etc. Basically, anything that can be identified uniquely on the web can be a subject.
2. Predicate: This is the characteristic or attribute of the subject that is being asserted. It is the property or relationship that is being stated about the subject. For example, if the subject is a person, predicates could be properties like "hasName", "hasAge", etc.
3. Object: This is the value of the property. If the predicate is "hasName", the object could be the name of the person. Objects can either be literal values (like strings, numbers, dates, etc.), or they can be other resources, allowing for connections between different entities.

A Literal consists of two components: the lexical form and the optional datatype or language tag. The lexical form is the actual value of the Literal, while the datatype or language tag provides additional information about the type or language of the value.

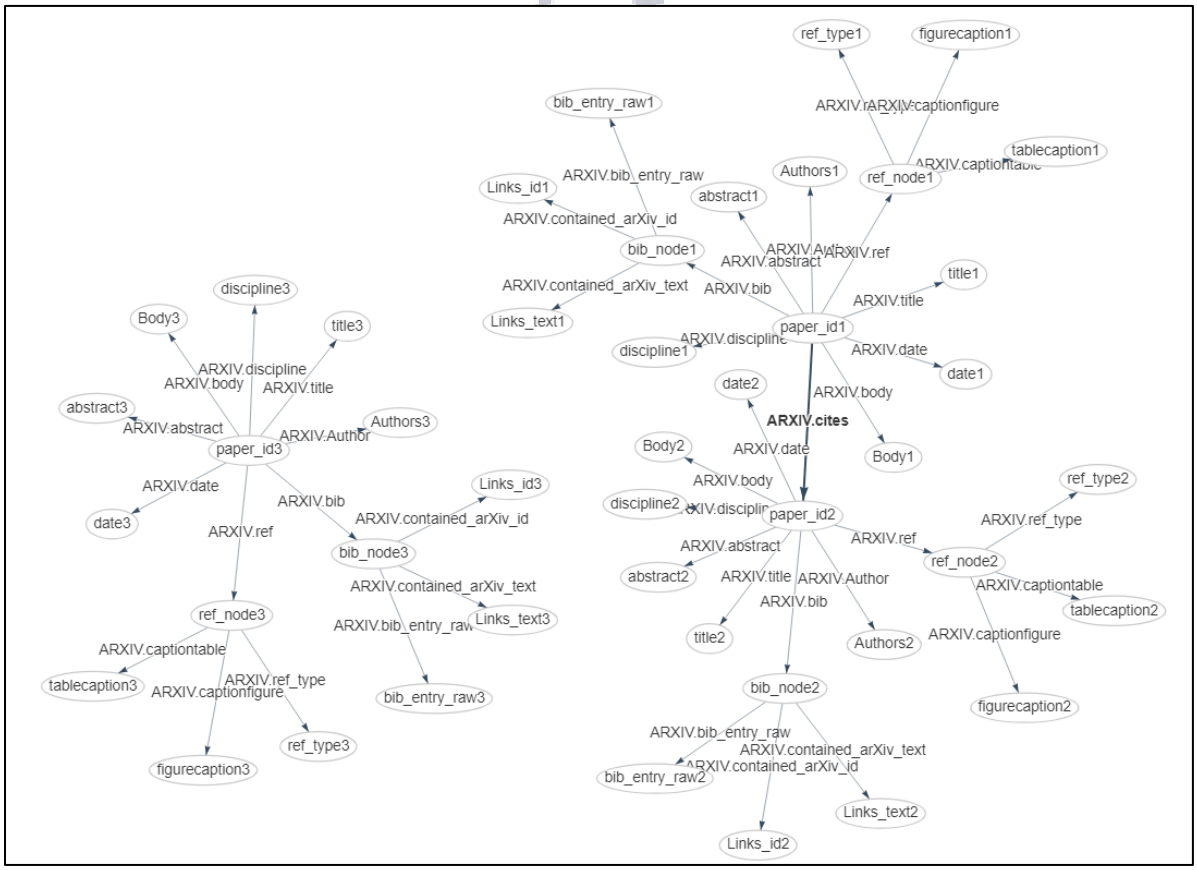
Let's break down the components of a Literal:

- Lexical Form: The lexical form is the actual value of the Literal. It can be a string, number, date, or any other valid value.
- Datatype: The datatype component of a Literal specifies the data type of the value. Common datatypes in RDF include xsd:string for strings, xsd:integer for integers, xsd:date for dates, and so on.
- Language Tag: Instead of a datatype, a Literal can also have a language tag to indicate the natural language of the value. Language tags are represented using the BCP 47 language tag format. For example, a Literal representing the string "Hello" in French would be written as "Bonjour"@fr.

By using Literals, RDF allows for the representation of data values with their associated types or languages within the graph. This enables more precise interpretation and querying of the data in a knowledge graph.

Structure of my Research Paper Knowledge Graph

- paper_id
- metadata
  - authors
  - title
  - update_date
- discipline
- abstract
  - text
- body_text
  - text
- bib_entries
  - bib_entry_raw
  - contained_arXiv_ids
    - id - if equal to any paper_id, it will draw a link between the two papers
    - text
- ref_entries
  - type
  - caption
    - table
    - figure



This graph has 3 papers paper1, paper2 and paper3. In this case paper1 cites paper2 and paper3 is separate.

The above graph was plotted using **GraphGPT**, which is open source. GraphGPT converts unstructured natural language into a knowledge graph. You can even pass in the synopsis of your favorite movie, a passage from a confusing Wikipedia page, or transcript from a video to generate a graph visualization of entities and their relationships.

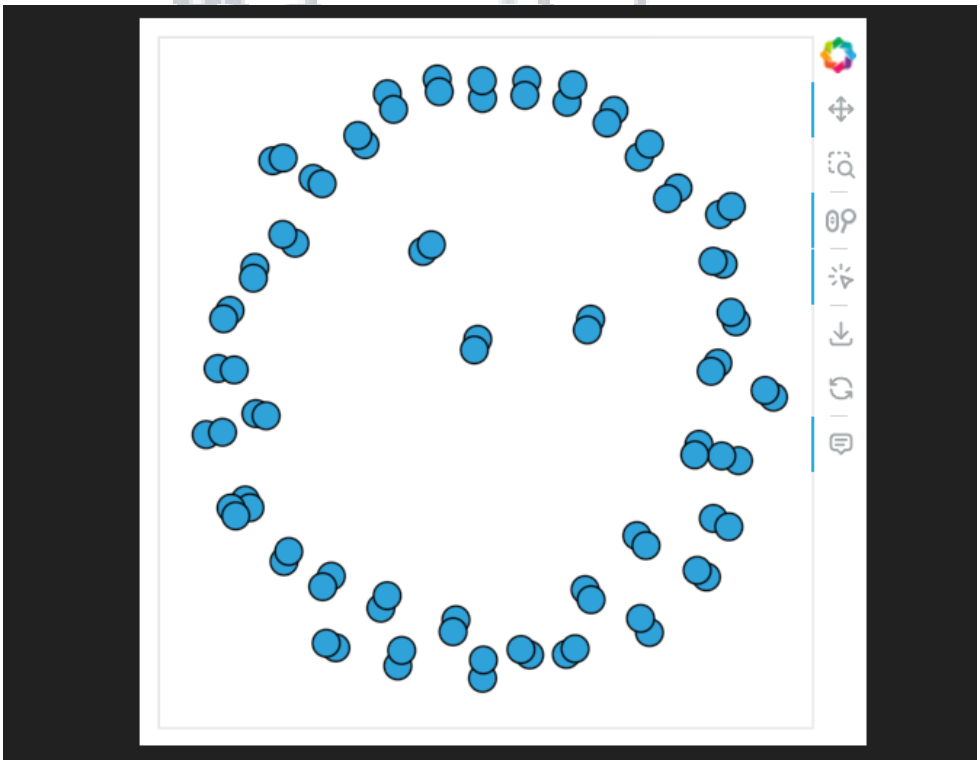
Website: https://graphgpt.vercel.app/

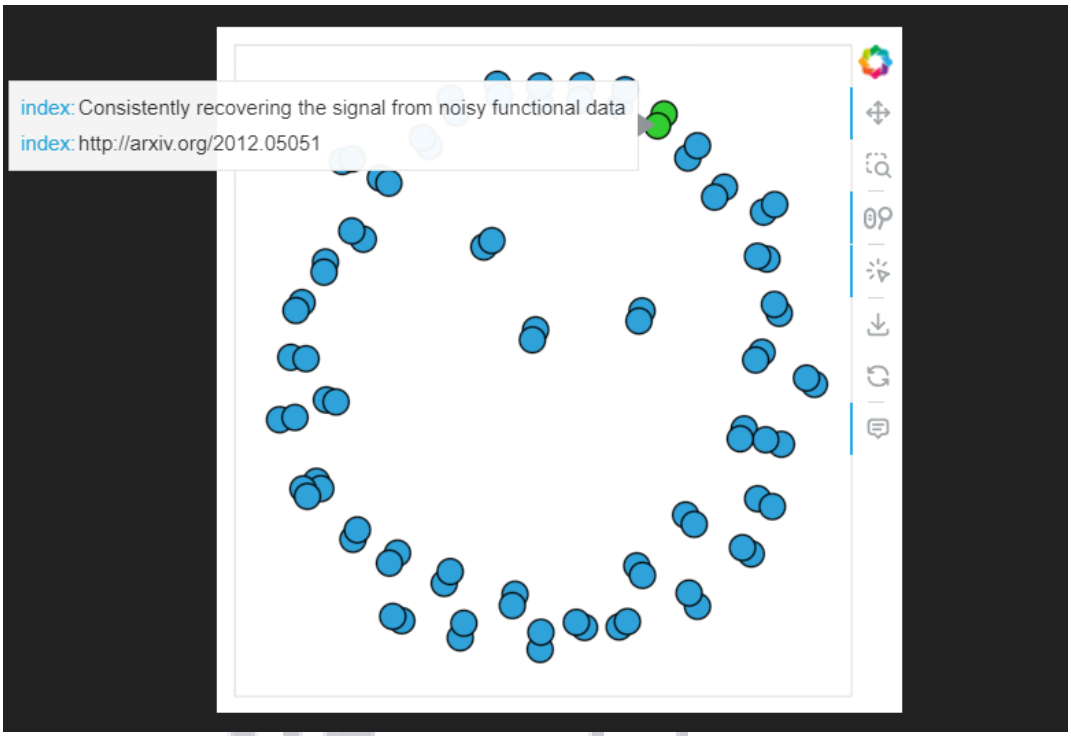
Visualization of the subgraphs obtained using SPARQL querying

HoloViews is an open-source Python library designed to make data analysis and visualization seamless and simple. Holoviews is a powerful Python library for creating interactive visualizations. It provides a high-level interface to efficiently visualize complex datasets.

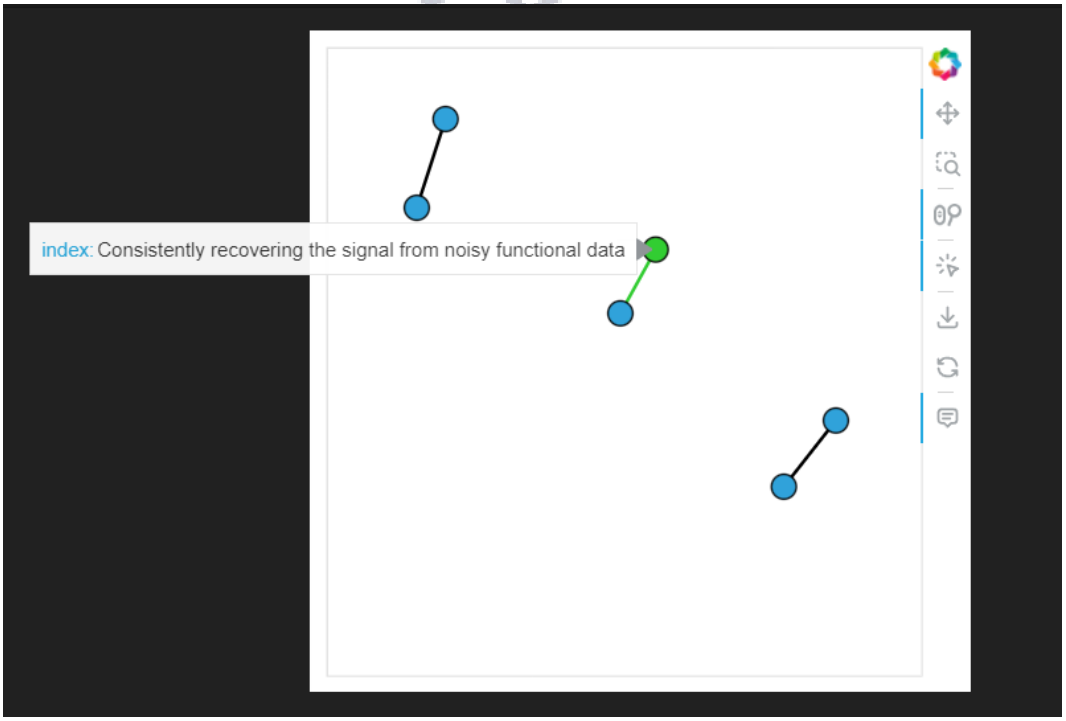
I personally like this Python library for visualization as it displays text only when we hover over a node. So this does not over complicate the Visualization and you see what you want the visualization to display.

1. Query to provide valuable insights about a particular discipline in the research paper knowledge graph. Eg) “Statistics”.

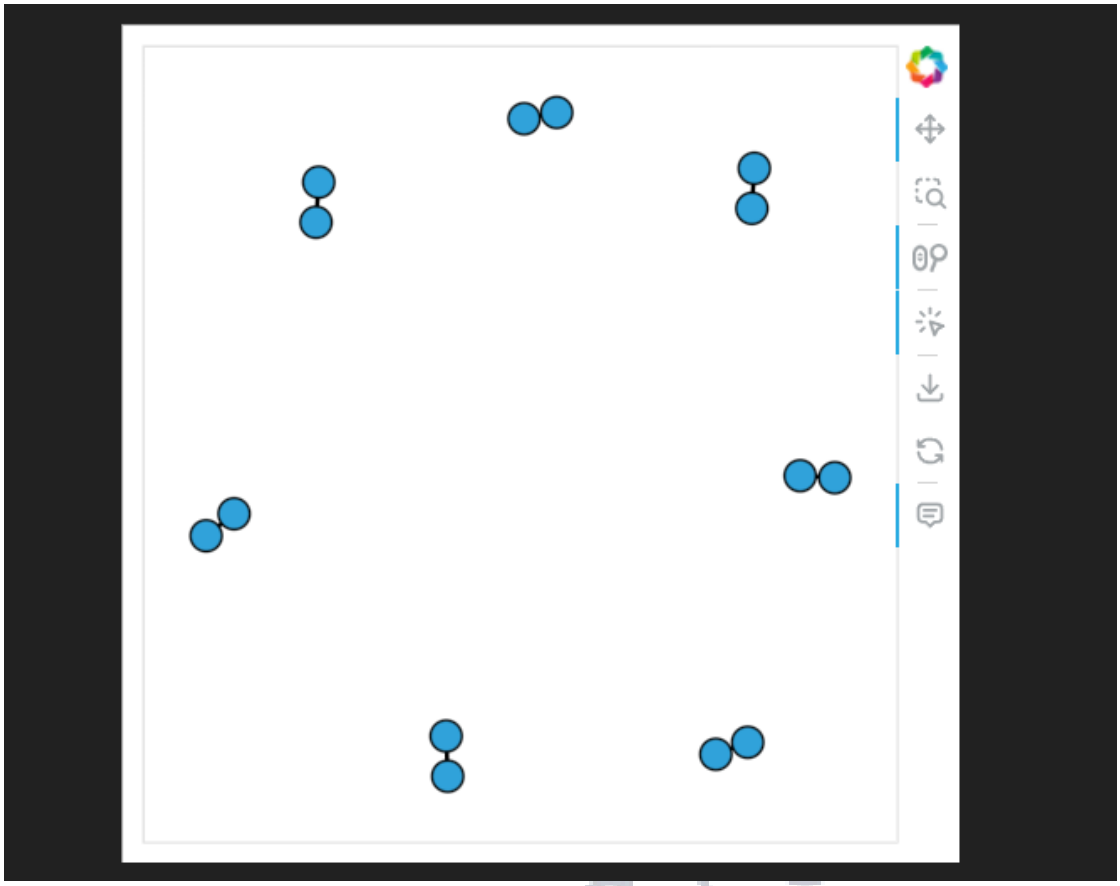




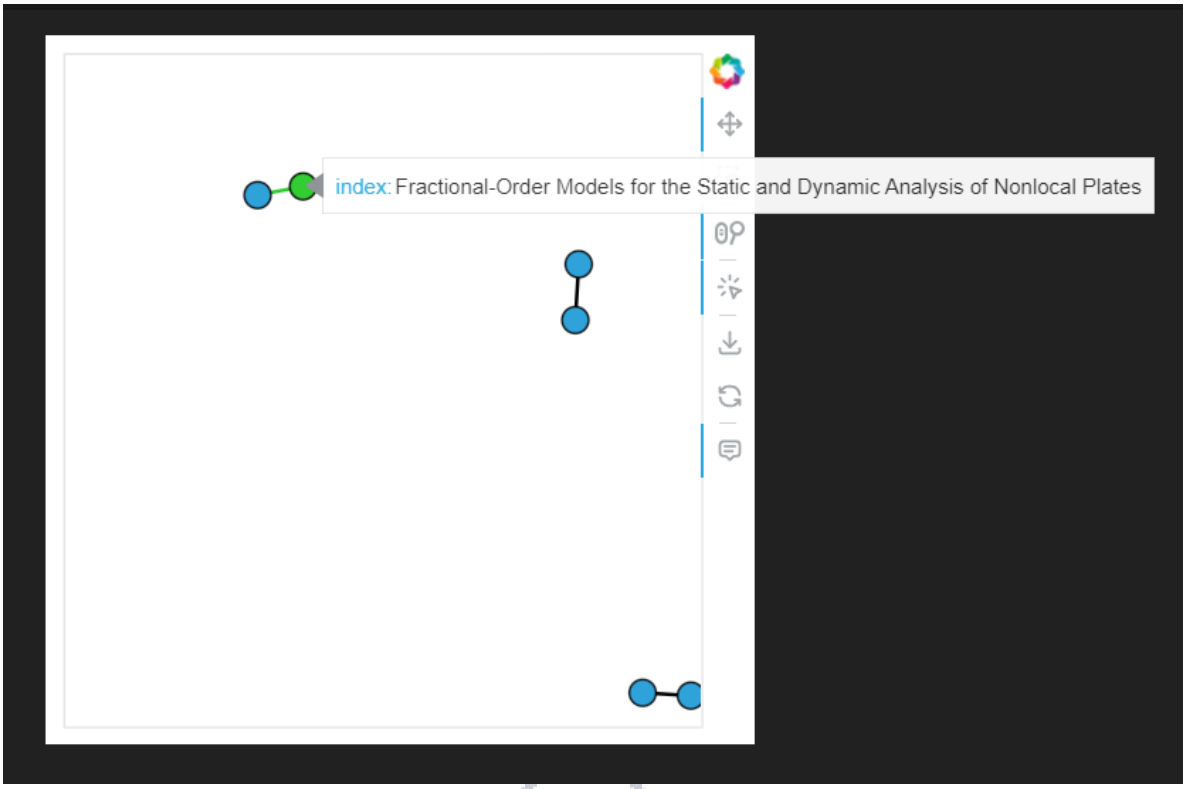
When we zoom in:-



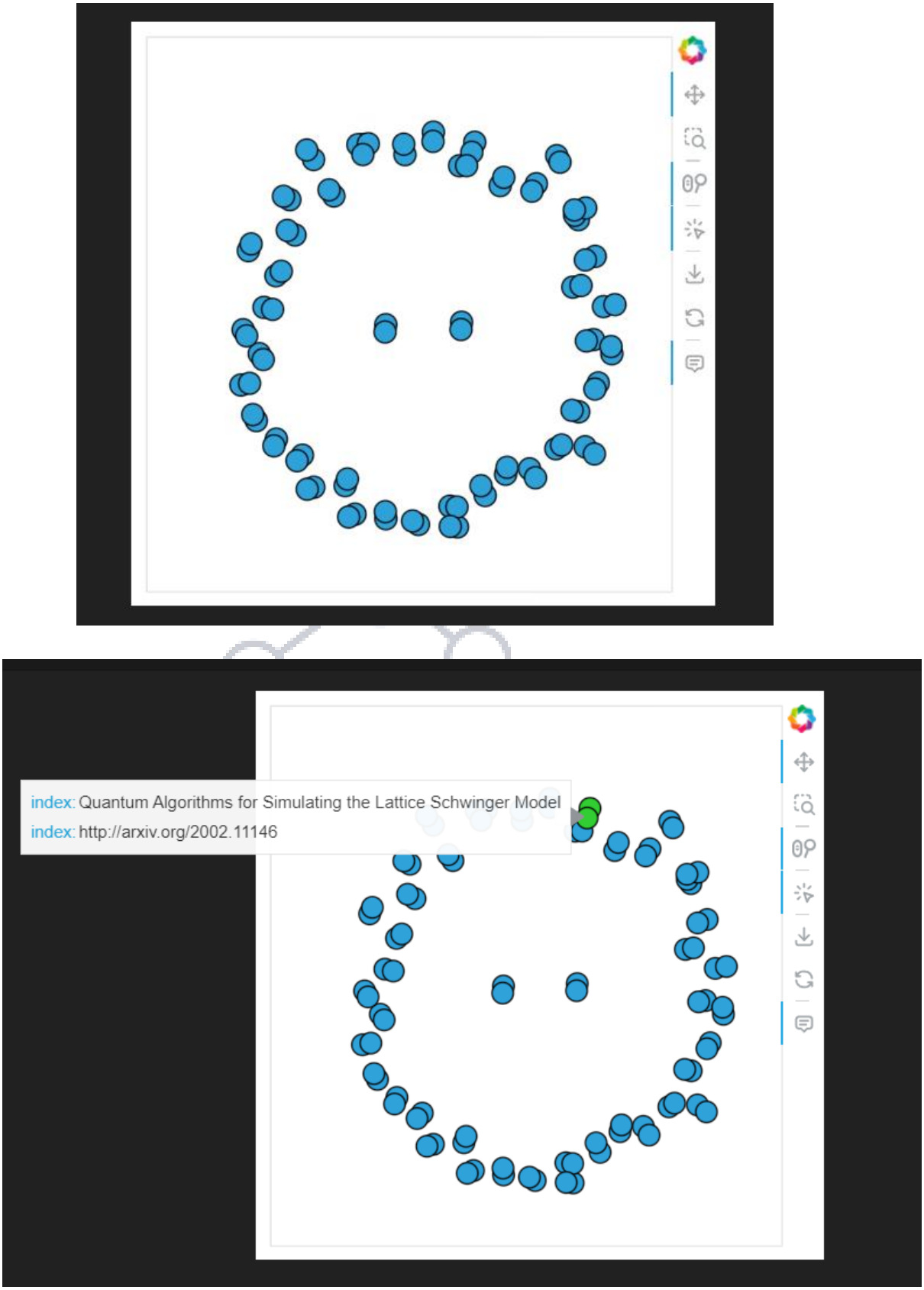
2. Query to retrieve all papers written by a particular author. This can be beneficial for someone who wants to explore all papers published by a certain author. Eg) “Sansit Patnaik”.



When we zoom in:-

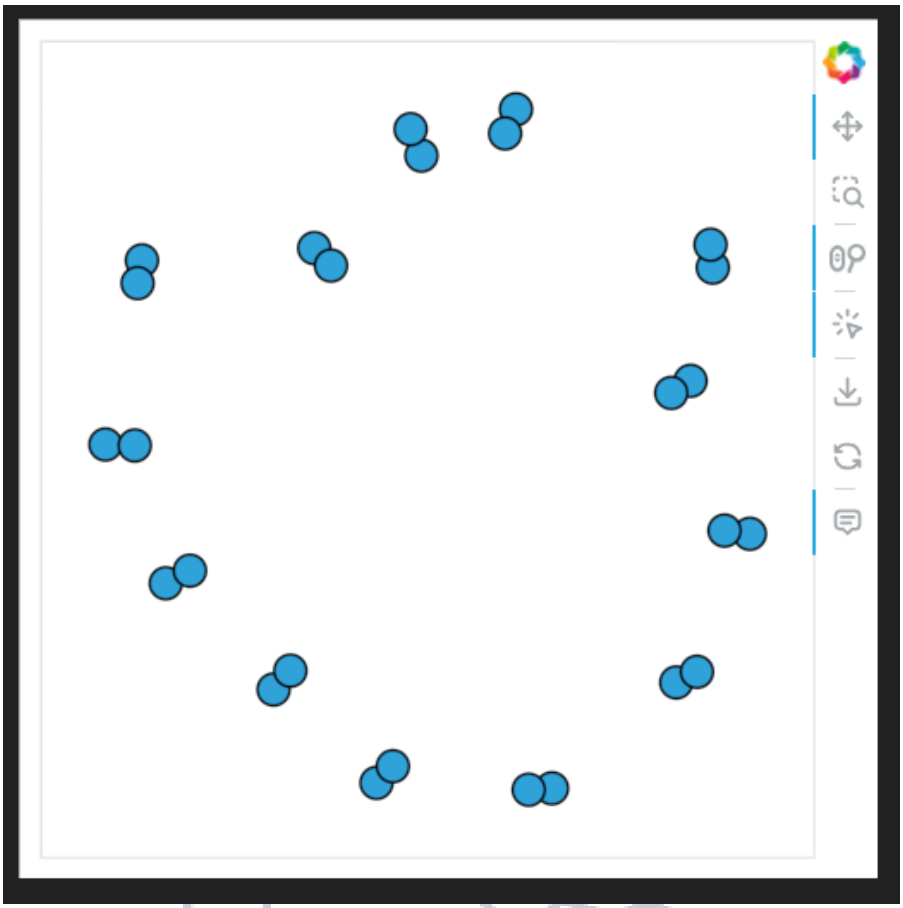


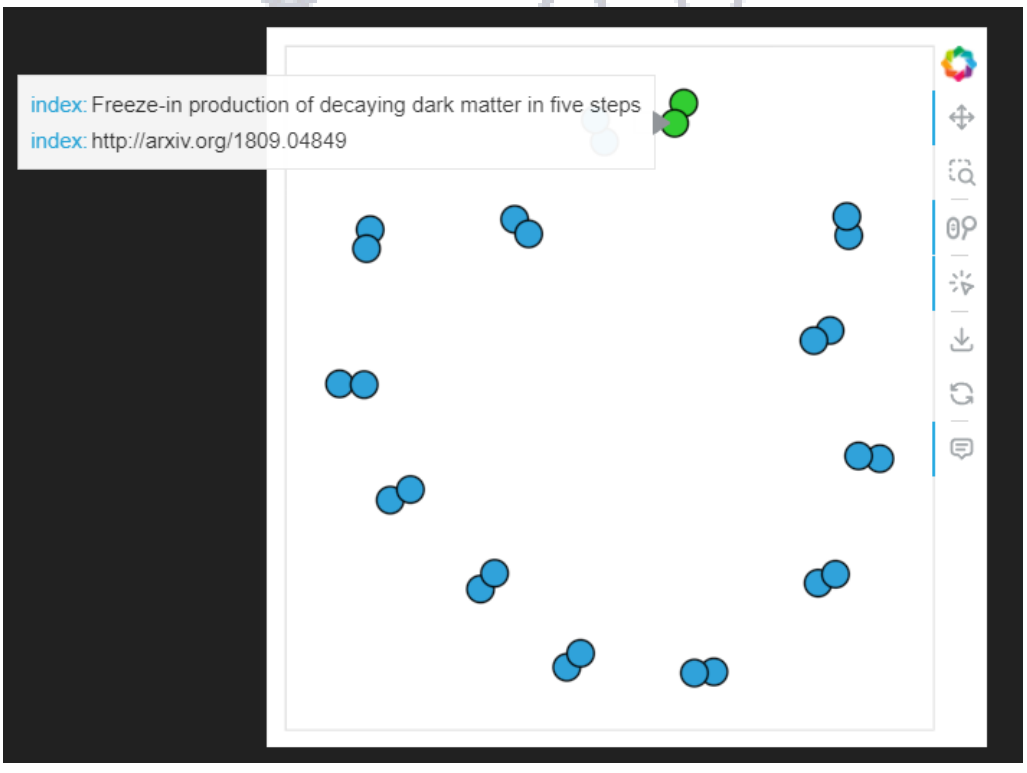
3. Query to find all papers on a specific topic. This could be useful for a researcher looking to find all papers on a specific topic. Eg) “Quantum”.



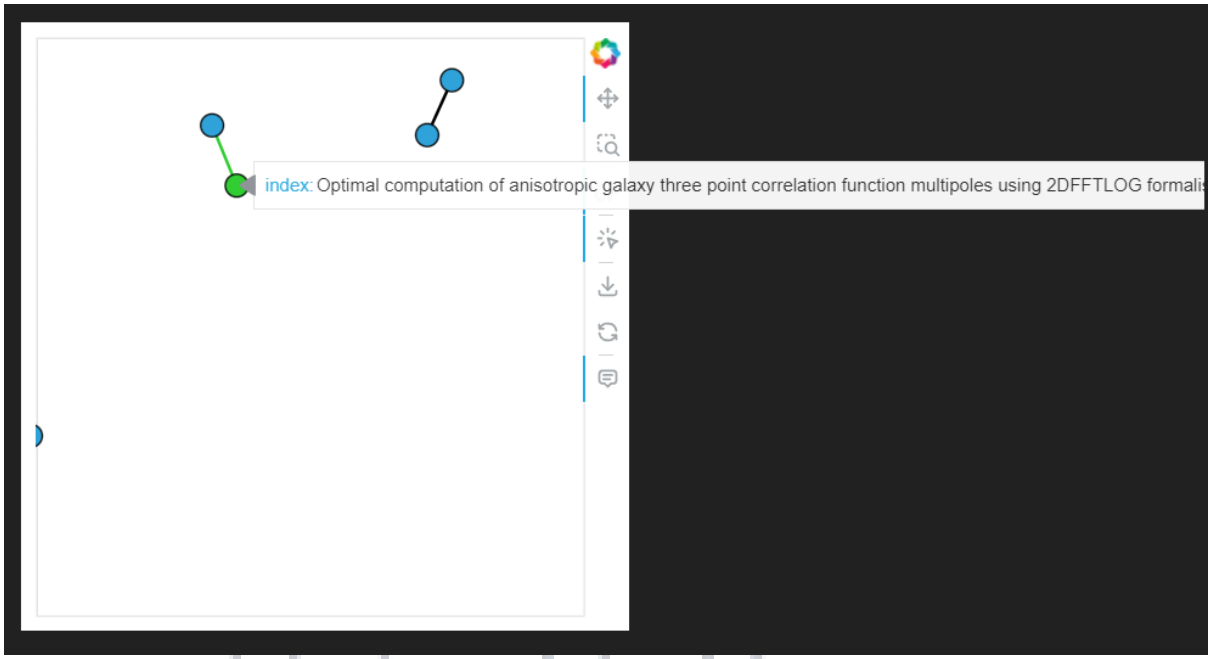
When you zoom in:-

4. Query to find all papers citing a specific paper. This could be helpful for a researcher trying to gauge the impact of a specific work. Eg) “1807.06209”.

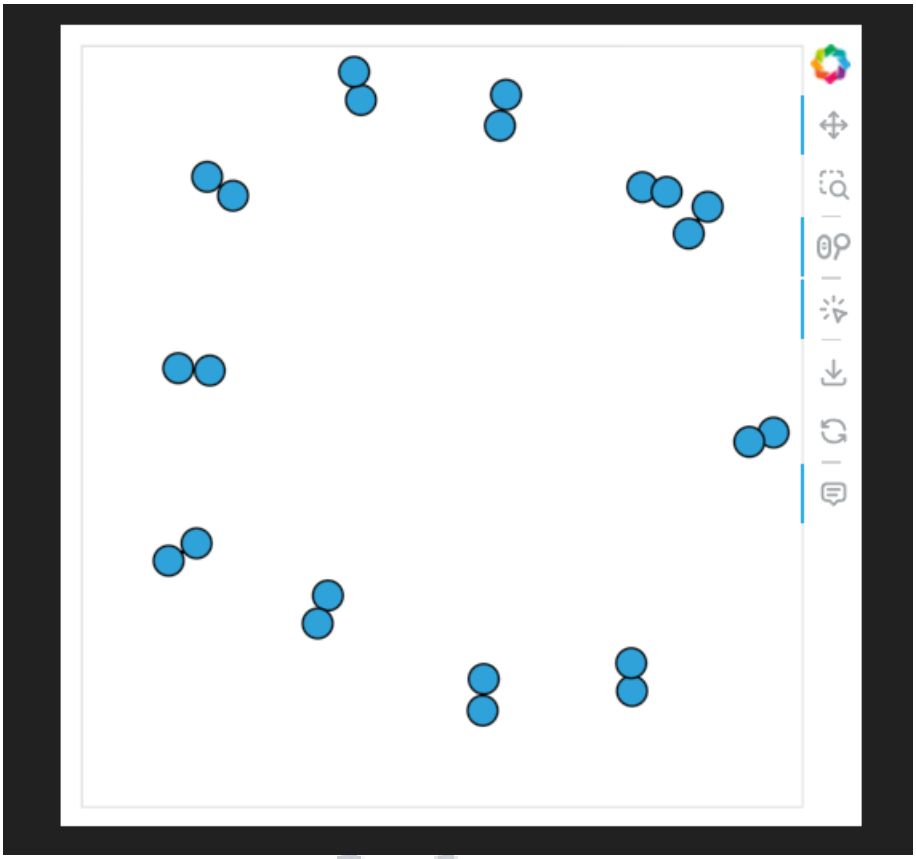


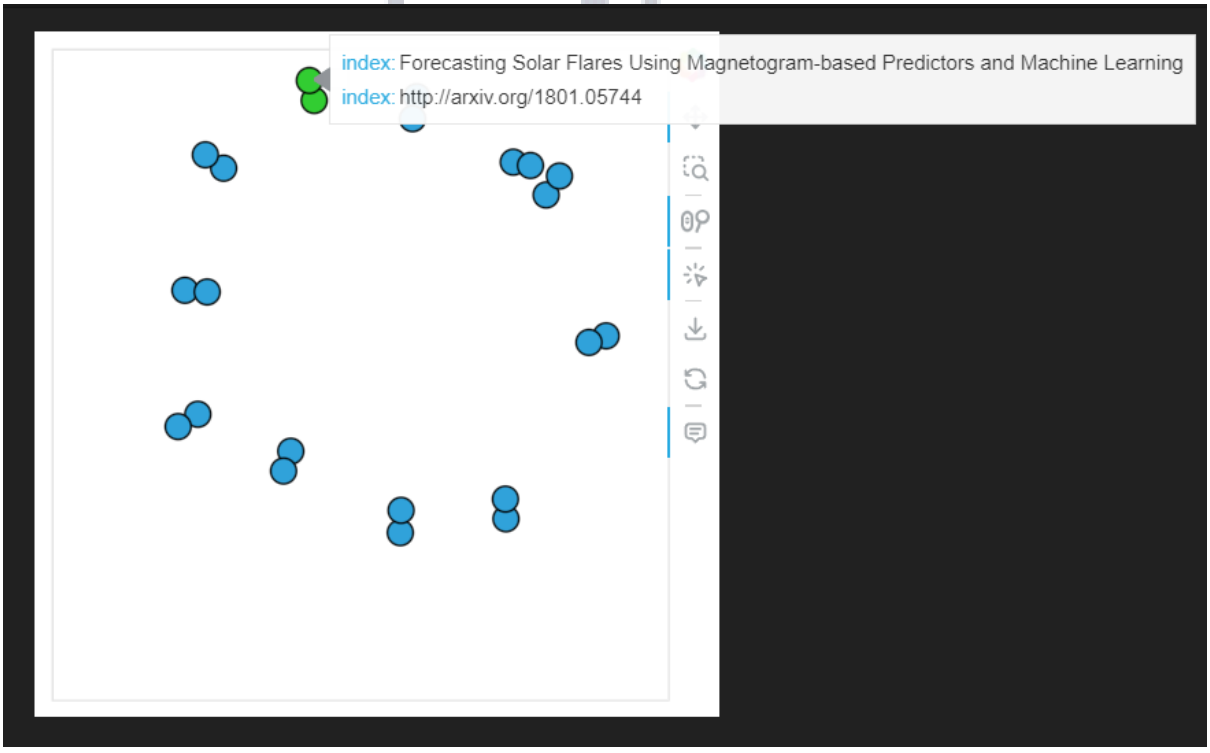


When we zoom in:-



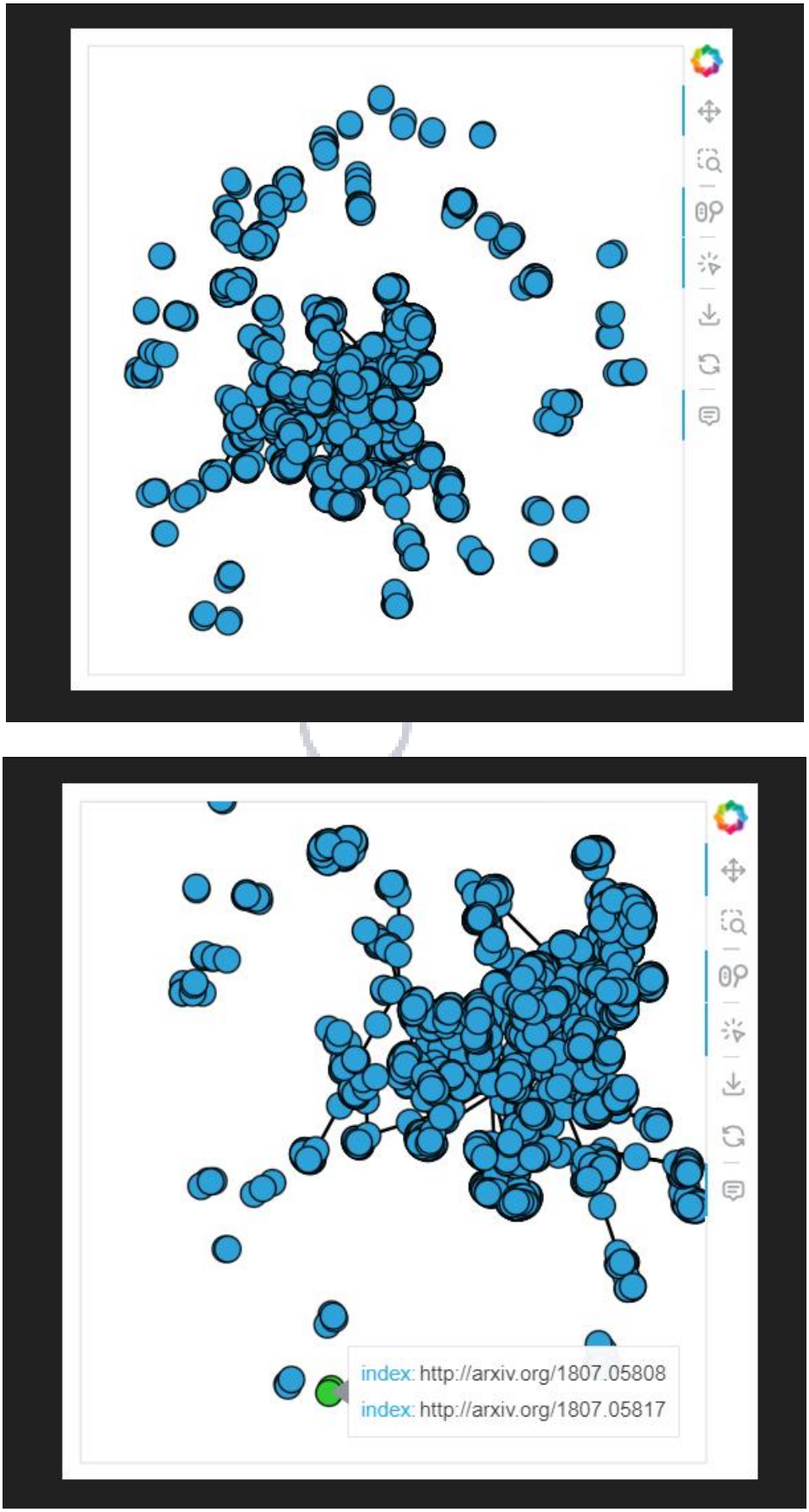
5. Query to give insights into the papers that discuss a specific topic or concept. Eg) “ML”.



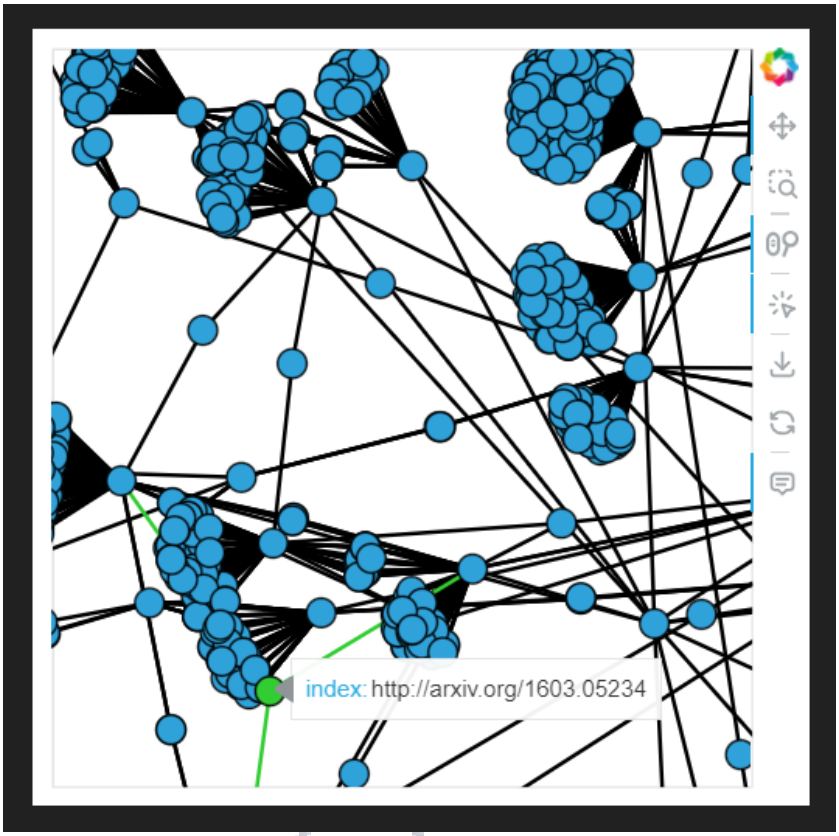


When we zoom in:-

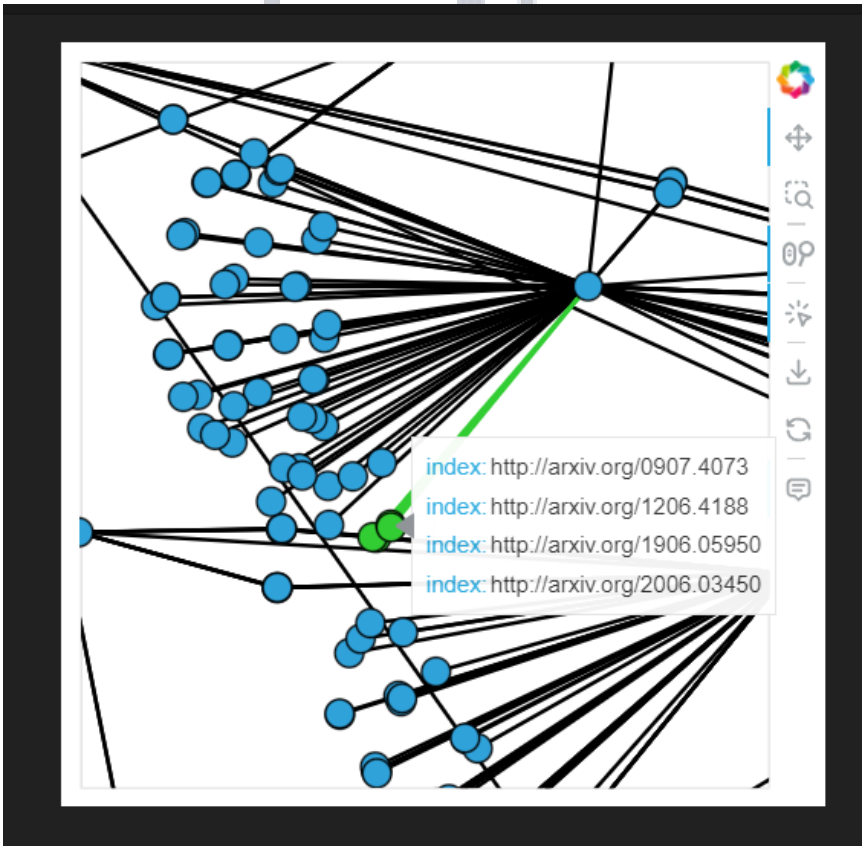
6. Query to get all citations in the knowledge graph



When we zoom in:-



Zoom even more for more clarity:-



PART - II

Our task is to implement a recommender system that suggests 5 similar and important research papers that a researcher could cite for a new research paper that he/she plans to write

The steps taken by me in this project are as follows:-

1. **Data Representation**
   The input data is stored in the form of an RDF graph, where each paper is represented as a node and citations are represented as directed edges between these nodes.

2. **Data Extraction**
   Information about each paper - its title, discipline, abstract, and citation data, is extracted from the RDF graph. The citation data is represented as an adjacency matrix, where each row and column represents a paper and an entry is 1 if the corresponding paper cites the corresponding cited paper.

3. **Feature Generation**
   For each paper, a feature vector is generated by applying a pre-trained BERT model to its abstract. BERT is a transformer-based machine learning model designed to generate meaningful representations of text. These feature vectors, or embeddings, capture semantic information about the abstracts, such as the topics they discuss and the ways in which they discuss them.

   Certain key properties of **BERT model**:-

   - Transformer Architecture: It is a neural network architecture designed to handle sequential data efficiently. Transformers use self-attention mechanisms to capture contextual relationships between words in a sentence, enabling better understanding of the context.

   - Pretraining: BERT is a pretraining model, which means it is first trained on a large corpus of text in an unsupervised manner. During pretraining, BERT learns to predict missing words in sentences by considering both the left and right context, making it bidirectional.

   - Bidirectional Context: The bidirectional nature of BERT is a crucial feature. Unlike GPT (Generative Pretrained Transformer), which are unidirectional, BERT looks at both the left and right contexts of a word during training. This allows it to capture more comprehensive contextual information, leading to better language understanding.

   - Masked Language Model (MLM): BERT's pretraining objective involves masking some words in sentences and training the model to predict the masked words based on their context. This approach enables BERT to learn deep contextual representations of words.

   - Large-Scale Pretraining: BERT is trained on a massive amount of text data, making it capable of learning rich linguistic patterns and semantics. This

large-scale pretraining contributes to BERT's effectiveness in various downstream NLP tasks.

- Contextual Word Embeddings: BERT produces contextual word embeddings, meaning the embedding for a word can vary depending on its context in the sentence. This contextual understanding enhances the model's ability to capture word meanings in different contexts.

- Subword Tokenization: BERT tokenizes words into sub words using WordPiece tokenization, which helps handle out-of-vocabulary words and reduces the vocabulary size.

- Open-Source Implementation: BERT was released as an open-source model by Google, which has led to widespread adoption and usage in the NLP community. The open-source implementation allows researchers and developers to use and improve the model for various applications.

4. **Similarity Calculation**
   The cosine similarity between each pair of papers is calculated based on their BERT embedding. The cosine similarity is a measure of the cosine of the angle between two vectors, which in this case provides a measure of the semantic similarity between two papers. Cosine similarity calculates the angle between these two vectors in the multi-dimensional space. It doesn't consider the magnitude (length) of the vectors; it only looks at the direction. The similarity score is a value between -1 and 1.

   Interpretation of **Cosine Similarity**:-

   - Cosine Similarity = 1: The documents are very similar.
   - Cosine Similarity = -1: The documents are very dissimilar.
   - Cosine Similarity close to 0: The documents are not similar.

5. **Model Training**:
   A logistic regression model is trained to predict whether one paper cites another based on their cosine similarity. Logistic regression is a binary classification model that outputs the probability that a given input belongs to a particular class, which in this case is the class of paper pairs where the first paper cites the second paper.

   Some key properties of **Logistic Regression**:-

   - Binary Classification: Logistic regression is a popular statistical method used for binary classification tasks, where the goal is to predict the probability of an instance belonging to one of two possible classes (e.g., yes/no, true/false, positive/negative).

   - Output Probability: The logistic regression model predicts the probability of the positive class (class 1) using a logistic or sigmoid function. The output probability is a value between 0 and 1.

- Sigmoid Function: The sigmoid function maps the output of the linear regression model (a continuous value) to a probability value between 0 and 1. The formula for the sigmoid function is: S(z) = 1 / (1 + e^(-z)).

How is the logestic regression model trained?

1. **Preprocessing**: For each pair of papers, we calculate the cosine similarity of their BERT-generated abstract embeddings. These similarities serve as our feature set (**X**) for the logistic regression model. The target set (**y**) is a binary value indicating whether one paper cites another.
2. **Training the Model**: The logistic regression model is trained using the **fit** method, which takes the feature set (**X**) and the target set (**y**) as arguments. This method uses the method of Maximum Likelihood Estimation to estimate the model parameters that best fit the training data.

6. **Recommendation Generation**:
   For a new paper, its similarity with each existing paper is calculated, and these similarities are fed into the logistic regression model to predict the probability of citation. The papers are then ranked by their predicted probabilities of citation, and the top papers are recommended.

The input to the recommendation system is taken from the 'new_research_papers.jsonl' file. This recommendation system uses the combination of a knowledge graph to store data, a BERT model to generate meaningful feature vectors, cosine similarity to measure semantic similarity, and a logistic regression model to predict citation based on similarity. These components are integrated into a single pipeline that takes a new paper as input and outputs a list of recommended papers to cite.

The outputs of the system are the IDs and abstracts of the recommended papers, which provide both a reference for locating the recommended papers and a brief summary of their contents. Recommendations are in order. The system is designed to take into account both the semantic content of the papers and the structure of the citation network to provide meaningful and relevant recommendations.

Recommendation 1

'Enhanced Accuracy in Galactic Disc Action Estimates through Perturbed Distribution Functions' (Physics)

```
Recommendations for 'Enhanced Accuracy in Galactic Disc Action Estimates through Perturbed Distribution Functions' (Physics):
ID: 2012.06597
Abstract: {'text': '  In the Gaia era, understanding the effects of the perturbations of the\nGalactic disc is of major importance in the context of dynamical modelling. In\nthis theoretical paper we extend previous work in which, making use of the\nepicyclic approximation, the linearized Boltzmann equation had been used to\nexplicitly compute, away from resonances, the perturbed distribution function\nof a Galactic thin-disc population in the presence of a non-axisymmetric\nperturbation of constant amplitude. Here we improve this theoretical framework\nin two distinct ways in the new code that we present. First, we use better\nestimates for the action-angle variables away from quasi-circular orbits,\ncomputed from the AGAMA software, and we present an efficient routine to\nnumerically re-express any perturbing potential in these coordinates with a\ntypical accuracy at the per cent level. The use of more accurate action\nestimates allows us to identify resonances such as the outer 1:1 bar resonance\nat higher azimuthal velocities than the outer Lindblad resonance (OLR), and to\nextend our previous theoretical results well above the Galactic plane, where we\nexplicitly show how they differ from the epicyclic approximation. In\nparticular, the displacement of resonances in velocity space as a function of\nheight can in principle constrain the 3D structure of the Galactic potential.\nSecond, we allow the perturbation to be time dependent, thereby allowing us to\nmodel the effect of transient spiral arms or a growing bar. The theoretical\nframework and tools presented here will be useful for a thorough analytical\ndynamical modelling of the complex velocity distribution of disc stars as\nmeasured by past and upcoming Gaia data releases.\n'}

ID: 2008.08404
Abstract: {'text': '  The eROSITA X-ray telescope on board the Spectrum-Roentgen-Gamma (SRG)\nmission will measure the position and properties of about 100,000 clusters of\ngalaxies and 3 million active galactic nuclei over the full sky. To study the\nstatistical properties of this ongoing survey, it is key to estimate the\nselection function accurately. We create a set of full sky light-cones using\nthe MultiDark and UNIT dark matter only N-body simulations. We present a novel\nmethod to predict the X-ray emission of galaxy clusters. Given a set of dark\nmatter halo properties (mass, redshift, ellipticity, offset parameter), we\nconstruct an X-ray emissivity profile and image for each halo in the\nlight-cone. We follow the eROSITA scanning strategy to produce a list of X-ray\nphotons on the full sky. We predict scaling relations for the model clusters,\nwhich are in good agreement with the literature. The predicted number density\nof clusters as a function of flux also agrees with previous measurements.\nFinally, we obtain a scatter of 0.21 (0.07, 0.25) for the X-ray luminosity --\nmass (temperature -- mass, luminosity -- temperature) model scaling relations.\nWe provide catalogues with the model photons emitted by clusters and active\ngalactic nuclei. These catalogues will aid the eROSITA end to end simulation\nflow analysis and in particular the source detection process and cataloguing\nmethods.\n'}

ID: 2012.08491
Abstract: {'text': '  We used dedicated SRG/eROSITA X-ray, ASKAP/EMU radio, and DECam optical\nobservations of a 15 sq.deg region around the interacting galaxy cluster system\nA3391/95 to study the warm-hot gas in cluster outskirts and filaments, the\nsurrounding large-scale structure and its formation process. We relate the\nobservations to expectations from cosmological hydrodynamic simulations from\nthe Magneticum suite.\n  We trace the irregular morphology of warm-hot gas of the main clusters from\ntheir centers out to well beyond their characteristic radii, $r_{200}$. Between\nthe two main cluster systems, we observe an emission bridge; thanks to\neROSITA\'s unique soft response and large field of view, we discover tantalizing\nhints for warm gas. Several matter clumps physically surrounding the system are\ndetected. For the "Northern Clump," we provide evidence that it is falling\ntowards A3391 from the hot gas morphology and radio lobe structure of its\ncentral AGN. Many of the extended sources in the field detected by eROSITA are\nknown clusters or new clusters in the background, including a known SZ cluster\nat redshift z=1. We discover an emission filament north of the virial radius,\n$r_{100}$, of A3391 connecting to the Northern Clump and extending south of\nA3395 towards another galaxy cluster. The total projected length of this\ncontinuous warm-hot emission filament is 15 Mpc, running almost 4 degrees\nacross the entire eROSITA observation. The DECam galaxy density map shows\ngalaxy overdensities in the same regions. The new datasets provide impressive\nconfirmation of the theoretically expected structure formation processes on the\nindividual system level, including the surrounding warm-hot intergalactic\nmedium distribution compared to the Magneticum simulation. Our spatially\nresolved findings show that baryons indeed reside in large-scale warm-hot gas\nfilaments with a clumpy structure.\n'}

ID: 2009.00327
Abstract: {'text': '  We present a major update to the 3D coronal rope ejection (3DCORE) technique\nfor modeling coronal mass ejection flux ropes in conjunction with an\nApproximate Bayesian Computation (ABC) algorithm that is used for fitting the\nmodel to in situ magnetic field measurements. The model assumes an empirically\nmotivated torus-like flux rope structure that expands self-similarly within the\nheliosphere, is influenced by a simplified interaction with the solar wind\nenvironment, and carries along an embedded analytical magnetic field. The\nimproved 3DCORE implementation allows us to generate extremely large ensemble\nsimulations which we then use to find global best-fit model parameters using an\nABC sequential Monte Carlo (SMC) algorithm. The usage of this algorithm, under\nsome basic assumptions on the uncertainty of the magnetic field measurements,\nallows us to furthermore generate estimates on the uncertainty of model\nparameters using only a single in situ observation. We apply our model to\nsynthetically generated measurements to prove the validity of our\nimplementation for the fitting procedure. We also present a brief analysis,\nwithin the scope of our model, of an event captured by Parker Solar Probe (PSP)\nshortly after its first fly-by of the Sun on 2018 November 12 at 0.25 AU. The\npresented toolset is also easily extendable to the analysis of events captured\nby multiple spacecraft and will therefore facilitate future multi-point\nstudies.\n'}

ID: 2012.12284
Abstract: {'text': '  We investigate the morphology of the stellar distribution in a sample of\nMilky Way (MW) like galaxies in the TNG50 simulation. Using a local in shell\niterative method (LSIM) as the main approach, we explicitly show evidence of\ntwisting (in about 52% of halos) and stretching (in 48% of them) in the real\nspace. This is matched with the re-orientation observed in the eigenvectors of\nthe inertia tensor and gives us a clear picture of having a re-oriented stellar\ndistribution. We make a comparison between the shape profile of dark matter\n(DM) halo and stellar distribution and quite remarkably see that their radial\nprofiles are fairly close, especially at small galactocentric radii where the\nstellar disk is located. This implies that the DM halo is somewhat aligned with\nstars in response to the baryonic potential. The level of alignment mostly\ndecreases away from the center. We study the impact of substructures in the\norbital circularity parameter. It is demonstrated that in some cases, far away\nsubstructures are counter-rotating compared with the central stars and may flip\nthe sign of total angular momentum and thus the orbital circularity parameter.\nTruncating them above 150 kpc, however, retains the disky structure of the\ngalaxy as per initial selection. Including the impact of substructures in the\nshape of stars, we explicitly show that their contribution is subdominant.\nOverlaying our theoretical results to the observational constraints from\nprevious literature, we establish fair agreement.\n'}
```

Recommendation 2

'A multimodal analysis of Parkinson's disease patients' (Statistics):


```
Recommendations for 'A multimodal analysis of Parkinson's disease patients' (Statistics):
ID: 2002.05411
Abstract: {'text': "  Background and objectives: Parkinson's disease is a neurological disorder\nthat affects the motor system producing lack of coordination, resting tremor,\nand rigidity. Impairments in handwriting are among the main symptoms of the\ndisease. Handwriting analysis can help in supporting the diagnosis and in\nmonitoring the progress of the disease. This paper aims to evaluate the\nimportance of different groups of features to model handwriting deficits that\nappear due to Parkinson's disease; and how those features are able to\ndiscriminate between Parkinson's disease patients and healthy subjects.\n  Methods: Features based on kinematic, geometrical and non-linear dynamics\nanalyses were evaluated to classify Parkinson's disease and healthy subjects.\nClassifiers based on K-nearest neighbors, support vector machines, and random\nforest were considered.\n  Results: Accuracies of up to $93.1\\%$ were obtained in the classification of\npatients and healthy control subjects. A relevance analysis of the features\nindicated that those related to speed, acceleration, and pressure are the most\ndiscriminant. The automatic classification of patients in different stages of\nthe disease shows $\\kappa$ indexes between $0.36$ and $0.44$. Accuracies of up\nto $83.3\\%$ were obtained in a different dataset used only for validation\npurposes.\n  Conclusions: The results confirmed the negative impact of aging in the\nclassification process when we considered different groups of healthy subjects.\nIn addition, the results reported with the separate validation set comprise a\nstep towards the development of automated tools to support the diagnosis\nprocess in clinical practice.\n"}

ID: 2009.04518
Abstract: {'text': "  In living systems, we often see the emergence of the ingredients necessary\nfor computation -- the capacity for information transmission, storage, and\nmodification -- begging the question of how we may exploit or imitate such\nbiological systems in unconventional computing applications. What can we gain\nfrom artificial life in the advancement of computing technology? Artificial\nlife provides us with powerful tools for understanding the dynamic behavior of\nbiological systems and capturing this behavior in manmade substrates. With this\napproach, we can move towards a new computing paradigm concerned with\nharnessing emergent computation in physical substrates not governed by the\nconstraints of Moore's law and ultimately realize massively parallel and\ndistributed computing technology. In this paper, we argue that the lens of\nartificial life offers valuable perspectives for the advancement of\nhigh-performance computing technology. We first present a brief foundational\nbackground on artificial life and some relevant tools that may be applicable to\nunconventional computing. Two specific substrates are then discussed in detail:\nbiological neurons and ensembles of nanomagnets. These substrates are the focus\nof the authors' ongoing work, and they are illustrative of the two sides of the\napproach outlined here -- the close study of living systems and the\nconstruction of artificial systems to produce life-like behaviors. We conclude\nwith a philosophical discussion on what we can learn from approaching\ncomputation with the curiosity inherent to the study of artificial life. The\nmain contribution of this paper is to present the great potential of using\nartificial life methodologies to uncover and harness the inherent computational\npower of physical substrates toward applications in unconventional\nhigh-performance computing.\n"}

ID: 1812.03503
Abstract: {'text': '  We present an effective post-processing method to reduce the artifacts from\nsparsely reconstructed cone-beam CT (CBCT) images. The proposed method is based\non the state-of-the-art, image-to-image generative models with a perceptual\nloss as regulation. Unlike the traditional CT artifact-reduction approaches,\nour method is trained in an adversarial fashion that yields more perceptually\nrealistic outputs while preserving the anatomical structures. To address the\nstreak artifacts that are inherently local and appear across various scales, we\nfurther propose a novel discriminator architecture based on feature pyramid\nnetworks and a differentially modulated focus map to induce the adversarial\ntraining. Our experimental results show that the proposed method can greatly\ncorrect the cone-beam artifacts from clinical CBCT images reconstructed using\n1/3 projections, and outperforms strong baseline methods both quantitatively\nand qualitatively.\n'}

ID: 2001.08614
Abstract: {'text': "  Wikipedia, the free online encyclopedia that anyone can edit, is one of the\nmost visited sites on the Web and a common source of information for many\nusers. As an encyclopedia, Wikipedia is not a source of original information,\nbut was conceived as a gateway to secondary sources: according to Wikipedia's\nguidelines, facts must be backed up by reliable sources that reflect the full\nspectrum of views on the topic. Although citations lie at the very heart of\nWikipedia, little is known about how users interact with them. To close this\ngap, we built client-side instrumentation for logging all interactions with\nlinks leading from English Wikipedia articles to cited references during one\nmonth, and conducted the first analysis of readers' interaction with citations\non Wikipedia. We find that overall engagement with citations is low: about one\nin 300 page views results in a reference click (0.29% overall; 0.56% on\ndesktop; 0.13% on mobile). Matched observational studies of the factors\nassociated with reference clicking reveal that clicks occur more frequently on\nshorter pages and on pages of lower quality, suggesting that references are\nconsulted more commonly when Wikipedia itself does not contain the information\nsought by the user. Moreover, we observe that recent content, open access\nsources and references about life events (births, deaths, marriages, etc) are\nparticularly popular. Taken together, our findings open the door to a deeper\nunderstanding of Wikipedia's role in a global information economy where\nreliability is ever less certain, and source attribution ever more vital.\n"}

ID: 2002.05412
Abstract: {'text': "  Parkinson's disease is a neurodegenerative disorder characterized by the\npresence of different motor impairments. Information from speech, handwriting,\nand gait signals have been considered to evaluate the neurological state of the\npatients. On the other hand, user models based on Gaussian mixture models -\nuniversal background models (GMM-UBM) and i-vectors are considered the\nstate-of-the-art in biometric applications like speaker verification because\nthey are able to model specific speaker traits. This study introduces the use\nof GMM-UBM and i-vectors to evaluate the neurological state of Parkinson's\npatients using information from speech, handwriting, and gait. The results show\nthe importance of different feature sets from each type of signal in the\nassessment of the neurological state of the patients.\n"}
```

Recommendation 3

'LOGO2-BongradPlus' (Computer Science):

```
Recommendations for 'LOGO2-BongradPlus' (Computer Science):
ID: 2010.00763
Abstract: {'text': "  Humans have an inherent ability to learn novel concepts from only a few\nsamples and generalize these concepts to different situations. Even though\ntoday's machine learning models excel with a plethora of training data on\nstandard recognition tasks, a considerable gap exists between machine-level\npattern recognition and human-level concept learning. To narrow this gap, the\nBongard problems (BPs) were introduced as an inspirational challenge for visual\ncognition in intelligent systems. Despite new advances in representation\nlearning and learning to learn, BPs remain a daunting challenge for modern AI.\nInspired by the original one hundred BPs, we propose a new benchmark\nBongard-LOGO for human-level concept learning and reasoning. We develop a\nprogram-guided generation technique to produce a large set of\nhuman-interpretable visual cognition problems in action-oriented LOGO language.\nOur benchmark captures three core properties of human cognition: 1)\ncontext-dependent perception, in which the same object may have disparate\ninterpretations given different contexts; 2) analogy-making perception, in\nwhich some meaningful concepts are traded off for other meaningful concepts;\nand 3) perception with a few samples but infinite vocabulary. In experiments,\nwe show that the state-of-the-art deep learning methods perform substantially\nworse than human subjects, implying that they fail to capture core human\ncognition properties. Finally, we discuss research directions towards a general\narchitecture for visual reasoning to tackle this benchmark.\n"}

ID: 2011.02157
Abstract: {'text': '  At the dawn of a new decade, particle physics faces the challenge of\nexplaining the mystery of dark matter, the origin of matter over antimatter in\nthe Universe, the apparent fine-tuning of the electro-weak scale, and many\nother aspects of fundamental physics. Perhaps the most striking frontier to\nemerge in the search for answers involves new physics at mass scales comparable\nto familiar matter, below the GeV scale, but with very feeble interaction\nstrength. New theoretical ideas to address dark matter and other fundamental\nquestions predict such feebly interacting particles (FIPs) at these scales, and\nindeed, existing data may even provide hints of this possibility. Emboldened by\nthe lessons of the LHC, a vibrant experimental program to discover such physics\nis under way, guided by a systematic theoretical approach firmly grounded on\nthe underlying principles of the Standard Model. We give an overview of these\nefforts, their motivations, and the decadal goals that animate the community\ninvolved in the search for FIPs, with special focus on accelerator-based\nexperiments.\n'}

ID: 2009.04518
Abstract: {'text': "  In living systems, we often see the emergence of the ingredients necessary\nfor computation -- the capacity for information transmission, storage, and\nmodification -- begging the question of how we may exploit or imitate such\nbiological systems in unconventional computing applications. What can we gain\nfrom artificial life in the advancement of computing technology? Artificial\nlife provides us with powerful tools for understanding the dynamic behavior of\nbiological systems and capturing this behavior in manmade substrates. With this\napproach, we can move towards a new computing paradigm concerned with\nharnessing emergent computation in physical substrates not governed by the\nconstraints of Moore's law and ultimately realize massively parallel and\ndistributed computing technology. In this paper, we argue that the lens of\nartificial life offers valuable perspectives for the advancement of\nhigh-performance computing technology. We first present a brief foundational\nbackground on artificial life and some relevant tools that may be applicable to\nunconventional computing. Two specific substrates are then discussed in detail:\nbiological neurons and ensembles of nanomagnets. These substrates are the focus\nof the authors' ongoing work, and they are illustrative of the two sides of the\napproach outlined here -- the close study of living systems and the\nconstruction of artificial systems to produce life-like behaviors. We conclude\nwith a philosophical discussion on what we can learn from approaching\ncomputation with the curiosity inherent to the study of artificial life. The\nmain contribution of this paper is to present the great potential of using\nartificial life methodologies to uncover and harness the inherent computational\npower of physical substrates toward applications in unconventional\nhigh-performance computing.\n"}

ID: 2010.10783
Abstract: {'text': '  Representation learning on user-item graph for recommendation has evolved\nfrom using single ID or interaction history to exploiting higher-order\nneighbors. This leads to the success of graph convolution networks (GCNs) for\nrecommendation such as PinSage and LightGCN. Despite effectiveness, we argue\nthat they suffer from two limitations: (1) high-degree nodes exert larger\nimpact on the representation learning, deteriorating the recommendations of\nlow-degree (long-tail) items; and (2) representations are vulnerable to noisy\ninteractions, as the neighborhood aggregation scheme further enlarges the\nimpact of observed edges.\n  In this work, we explore self-supervised learning on user-item graph, so as\nto improve the accuracy and robustness of GCNs for recommendation. The idea is\nto supplement the classical supervised task of recommendation with an auxiliary\nself-supervised task, which reinforces node representation learning via\nself-discrimination. Specifically, we generate multiple views of a node,\nmaximizing the agreement between different views of the same node compared to\nthat of other nodes. We devise three operators to generate the views -- node\ndropout, edge dropout, and random walk -- that change the graph structure in\ndifferent manners. We term this new learning paradigm as\n\\textit{Self-supervised Graph Learning} (SGL), implementing it on the\nstate-of-the-art model LightGCN. Through theoretical analyses, we find that SGL\nhas the ability of automatically mining hard negatives. Empirical studies on\nthree benchmark datasets demonstrate the effectiveness of SGL, which improves\nthe recommendation accuracy, especially on long-tail items, and the robustness\nagainst interaction noises. Our implementations are available at\n\\url{https://github.com/wujcan/SGL}.\n'}

ID: 2010.00403
Abstract: {'text': '  The field of Artificial Intelligence (AI) is going through a period of great\nexpectations, introducing a certain level of anxiety in research, business and\nalso policy. This anxiety is further energised by an AI race narrative that\nmakes people believe they might be missing out. Whether real or not, a belief\nin this narrative may be detrimental as some stake-holders will feel obliged to\ncut corners on safety precautions, or ignore societal consequences just to\n"win". Starting from a baseline model that describes a broad class of\ntechnology races where winners draw a significant benefit compared to others\n(such as AI advances, patent race, pharmaceutical technologies), we investigate\nhere how positive (rewards) and negative (punishments) incentives may\nbeneficially influence the outcomes. We uncover conditions in which punishment\nis either capable of reducing the development speed of unsafe participants or\nhas the capacity to reduce innovation through over-regulation. Alternatively,\nwe show that, in several scenarios, rewarding those that follow safety measures\nmay increase the development speed while ensuring safe choices. Moreover, in\n{the latter} regimes, rewards do not suffer from the issue of over-regulation\nas is the case for punishment. Overall, our findings provide valuable insights\nto the nature and kinds of regulatory actions most suitable to improve safety\ncompliance in the contexts of both smooth and sudden technological shifts.\n'}
```

The above implementation is not a graph-related algorithm or learning model to generate meaningful recommendations.

Detailed step-wise procedure I would take to incorporate graph algorithms in the recommender systems are:-

1. **Data Representation**:
   Similar to before, I would represent my data as a graph. Each paper would be a node in the graph, and each citation would be a directed edge. I can also create edges based on the cosine similarity between the BERT embeddings of the papers.

2. **Node Feature Generation**:
   I can use BERT or SciBERT (it is a BERT model trained on scientific text. SciBERT is trained on papers from the corpus) to generate these feature vectors for each paper. Huggingface or GPT Text embeddings can also be used. Techniques such as PCA or autoencoders could be useful to reduce their dimensionality if required. Apart from the semantic content, I can include other features such as the number of citations the paper has, the impact factor of the journal it's published in, or the h-index of the authors. This would allow my GNN to also consider the importance of a paper when generating recommendations.

3. **Creating the Graph Neural Network**:
   Libraries such as PyTorch Geometric or DeepGraph Library (DGL) can be used to create the GNN. Graph Convolutional Networks (GCNs) are a common choice for this task. Each layer of a GCN uses the graph structure and the node features to learn a new representation for each node. This representation encapsulates both the node's own features and its context within the graph, providing a rich input for the recommendation system.

4. **Training the GNN**:
   Training the GNN to perform link prediction is a good idea. I can treat existing citations as positive examples and randomly selected non-existing links as negative examples. Then I can train my GNN as a binary classifier to predict whether a link should exist between two papers. For node classification, I can classify papers into different fields or topics. The topics can either be predefined (if I have this information available) or they can be learned by running a topic modeling algorithm like Latent Dirichlet Allocation (LDA) on the paper abstracts.

5. **Generating Recommendations**:
   Once the GNN is trained, I can indeed generate recommendations by feeding a new paper and all existing papers into the GNN and computing the likelihood of a link for each pair. This approach assumes that similar papers are more likely to cite each other, which is a reasonable assumption for academic papers.

Incorporating GNNs in the recommender system would allow me to leverage the structure of the citation network in generating my recommendations, in addition to the semantic content of the papers. GNNs can capture complex patterns in the graph structure that could be indicative of citation relationships, which could potentially improve the quality of my recommendations. However, GNNs are also more complex and computationally intensive than traditional machine learning models, so I would need to carefully consider whether their potential benefits outweigh their costs for this specific application.

PART - III

The algorithms used for impact analysis to identify the research works that have made the most significant impact on their peers are:-

1. **PageRank:**

    - PageRank is a method used to measure the importance or influence of a paper in your research knowledge graph.
    - It considers both the number of citations a paper receives and the importance of the papers that cite it.
    - The advantage of using Pagerank is that it accounts for the quality and relevance of the papers citing a particular paper.
    - In other words, if a highly influential paper cites another paper, it will give more weight to that citation than a less influential paper.
    - This way, Pagerank helps identify papers that have a significant impact on the overall network of research papers, considering the connections and influence of each citation.
    - Output:-

```
Paper ID: 1605.02688, PageRank Score: 0.0005306937790931206
Paper ID: 1011.0352, PageRank Score: 0.0004946854805135191
Paper ID: 1412.6980, PageRank Score: 0.0004745626324922857
Paper ID: quant-ph/9705052, PageRank Score: 0.0004000020722544107
Paper ID: 1105.4464, PageRank Score: 0.0003678372466247786
```

2. **HITS - Hub Score:**

    - HITS stands for "Hyperlink-Induced Topic Search," and in this context, we focus on the Hub Score component.
    - The Hub Score measures the importance of a paper based on its ability to cite other relevant and influential papers.
    - The advantage of using the Hub Score is that it highlights papers that act as excellent sources of information and references.
    - A paper with a high Hub Score is like a central hub that connects to many other important papers, making it a valuable resource for researchers and indicating that it has had a considerable impact in the field.
    - I am not using Authority score in HITS score because
        - Authority Score is more relevant when you are interested in finding papers that are highly cited by other papers.

- While this information is undoubtedly valuable, it might not directly indicate the paper's centrality or its role as a hub within the research network.
- The Authority Score helps to identify papers that are considered authoritative or influential due to the number of citations they have received.

- Output:- (will consider only Hub Score)

```
Node: 1207.7214, Authority Score: 0.0010448481862151613
Node: 1207.7235, Authority Score: 0.001044848186215161
Node: 1201.4330, Authority Score: 0.001020359768844416
Node: 1910.06275, Authority Score: 0.000999650972012424
Node: 1712.09737, Authority Score: 0.000999650972012424

Node: 2009.00516, Hub Score: 0.8604648613325913
Node: 2008.06494, Hub Score: 0.03426114842935145
Node: 1805.00736, Hub Score: 0.01974284603624696
Node: 2007.08542, Hub Score: 0.0185946434793888
Node: 1802.09886, Hub Score: 0.018535168449002274
```

3. **Eigenvector Centrality:**

- Eigenvector Centrality is another method to determine the importance of a paper in the research knowledge graph.
- It calculates the centrality of a paper by considering not only the number of its citations but also the importance of the papers that cite it.
- The advantage of Eigenvector Centrality is that it gives more weight to citations from papers that are themselves highly central and influential.
- A citation from a paper that is well-connected and cited by many other important papers will contribute more to the centrality of the cited paper.
- This helps to identify papers that are not just cited frequently but are cited by other crucial papers, indicating their significant impact on the research community.
- Example to understand why Eigenvector Centrality and why not the other centrality types:-

  Imagine you're at a large social event, like a big party or conference. You want to figure out who the most "important" or influential people are at the event. One way you could try to figure this out is by looking at who has the most connections -- who is talking to the most people. This is the basic idea behind "degree centrality": the more connections a node (in this case, a person) has, the more important it is. “Closeness centrality” can be used if we want to find

how close a person is to someone.

But this isn't always the best measure. For example, let's say Person A knows lots of people, but all of those people aren't very well-connected themselves. On the other hand, Person B knows fewer people, but those people are all very well-connected. In this case, we might think of Person B as being more "important" or influential, because they're connected to other influential people.

This is the basic idea behind "eigenvector centrality": a node is considered important if it's connected to other nodes that are important.

- In the question there was a line stating “Note: Importance of a research paper is measured not just by the number of citations that it receives but also by the quality of the citations. The quality of a citation is in turn determined by the number and quality of its own citations.” This is why I chose Eigenvector Centrality Score instead of Degree Centrality Score.

- Output:-

```
Paper ID: 2009.00516, Eigenvector Centrality Score: 0.6946376053239093
Paper ID: Paper, Eigenvector Centrality Score: 0.12768624963298328
Paper ID: 2008.06494, Eigenvector Centrality Score: 0.05400320088322996
Paper ID: 1805.00736, Eigenvector Centrality Score: 0.043124126025963494
Paper ID: 2012.07714, Eigenvector Centrality Score: 0.03892029018082468
```

Normalization of PageRank scores, Hub scores, and Eigenvector Centrality for each paper:-

It is important to ensure:-

1. **Comparable Scale:** Normalization brings all scores to a comparable scale. Without normalization, scores produced by different metrics might have different scales or ranges, making direct comparisons or aggregations (like your weighted average) inaccurate or misleading.
2. **Prevents Bias:** Without normalization, one scoring method might dominate the final combined score simply because its raw scores are naturally higher, not because it's necessarily a better indicator of relevance or importance. By normalizing the scores, each method contributes equally to the final score, as intended by your 1/3 weightage.
3. **Improved Stability:** Normalized metrics are generally more stable to small changes in the network, making your system more robust. For instance, adding a few nodes might significantly change raw PageRank scores but have a much smaller impact on the normalized scores.
4. **Interpretability:** Normalized scores often have a clear interpretation, such as "proportion of total importance". This can make them easier to understand and work with.

5. **Faster Convergence:** When used within iterative algorithms (as in the case of PageRank), normalized scores can help the algorithm converge more quickly.

Considering the importance of normalization, I decided to normalize the PageRank scores, Hub scores, and Eigenvector Centrality with equal weightage in my recommender.

Output:-

```
Paper ID: 2009.00516, Final Score: [0.66838086]
Paper ID: 1605.02688, Final Score: [0.33351089]
Paper ID: 1011.0352, Final Score: [0.29804897]
Paper ID: 1412.6980, Final Score: [0.27853691]
Paper ID: quant-ph/9705052, Final Score: [0.20516886]
```

The top 5 impactful papers are:-

1. 2009.00516
2. 1605.02688
3. 1011.0352
4. 1412.6980
5. quant-ph/9705052