American Physics Society data
Cross Citations Network Map

Rodrigo Da Costa
costa@gatech.edu
Georgia Institute of Technology

Anthony Sena
asena8@gatech.edu
Georgia Institute of Technology

Kamal Kishore Kashyap
kkashyap9@gatech.edu
Georgia Institute of Technology

Ryan J Miller
rmiller307@gatech.edu
Georgia Institute of Technology

1 INTRODUCTION - MOTIVATION

One the of fundamental pillars in academic research is conducting a careful and complete resource of the chosen field and subject. There are many forms and tools that can help in this like Google Scholar, APPsearch, Web of Science, arXiv between others.

Conducting research can be simplified and interpreted by running **network graph analysis**, more precisely, link analysis where one should aim to explore associations between a subset of nodes and their relationships with one another. Given how much the number of academic publications has grown over time, conducting proper academic research started to become harder and harder.

As a benefit, academic data for physics is quite easy to grab and arXiv [1] even publishes the data on Kaggle to facilitate and help on running analysis. But navigating the body of knowledge provided by arXiv, particularly the relationships between publications and area of impact, remains a challenge as cross citation is not yet properly identified and available to users.

2 PROBLEM DEFINITION

In this project, we aim to show the benefit on having a clear cross citation map and summarize the benefits on visualizing physics academic knowledge. We also will propose a novel solution for mapping scientific impact that instead of using the traditional **h-index** method we will measure the degree of the node comparing citations in the same group of knowledge (publication original journal) vs the citations in other groups on knowledge (other journals).

3 SURVEY

There is a large body of work focused on constructing maps of scientific publications using citation data. Areas of focus include: mapping relationships between scientific disciplines, author relationships and methods for clustering and classification of these relationships. [9, 12] Software solutions exist for general purpose visualization of citation networks [7, 13, 14].

Currently, visualizations of physics publications exist but do not always present a clear co-citation map making it difficult to properly represent the network and forcing us to map [3, 4]. Here we propose a purpose-built software solution for the physics community that builds on these ideas.

4 PROPOSED METHOD

4.1 Intuition

The current approach for navigating physics publications is through string-literal searches and traversing citations. In general, publications lack a standardized approach for capturing information such as author name and affiliations making it challenging when attempting to find resources using search strings. For example, authors may be listed in a number of ways: full name, first initial and last name, first and middle initial and last name. Institutional affiliations may differ if inclusive of address and other attributes. This lack of standardization makes knowledge discovery via string searches a challenge for researchers.

Our project aims to use a standardized representation of these publications with a clear cross-citation map to allow for querying based on author name and institution to provide a complete set of references. We also plan to provide a novel visualization atop this data to illustrate relationships between authors.

4.2 Materials

The American Physics Society curates a cross-citation map in a machine readable format (JavaScript Object Notation - JSON) that we will utilize to build a visualization tool. This curated list will allow for a cleaner mapping of author and institution in contrast to applying a heuristic to gather the relevant publication details.

4.3 Description

The following list provides a set of goals for the project. We divided our goals in three phases.

(1) Visualize the relationships between physics academic papers based on digital object identifier (DOI) citations and area of impact (journal id).
(2) Filter and visualize data by author and by institution.
(3) Map out and present Georgia Tech network contributions by area (journal id).

In order to achieve item (2) and (3) we would require to build our own API endpoint and structure the data as the raw inputs are giving as individual json files in a S3 bucket. This structure by itself does not enable us to filter easily. Apart from building our own API endpoint, we also aim to merge the cross-citation data frame together with each document by including citing and cited DOIs per JSON/object.

4.4 Architecture

We aim to create an *Express* server to handle API calls for for the APS raw data (build our own application programming interface (API)) and handle all web interface with the *React framework*. Backend will be done in *Node.js* with our data stored on *MongoDB*. We will use the **MERN stack**. We will use leverage *d3.js* [11] and *argo-graph-lite* [5, 6] for data visualization.

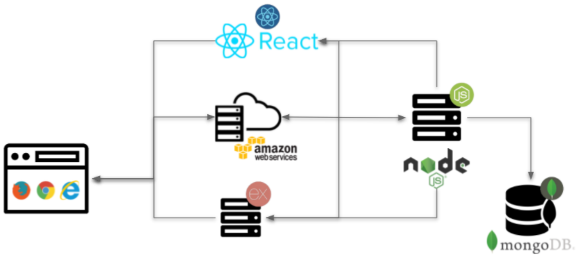


Figure 1: Proposed MERN stack architecture.

4.5 Costs

We plan to utilize the following technology and have provided an estimated cost associate for each component.

Tech	*Estimated Cost (month)*
AWS S3 server	$3
MongoDB Atlas	$25
Heroku	free tier
domain	$10
AWS Route 53	$1
Total	**$39**

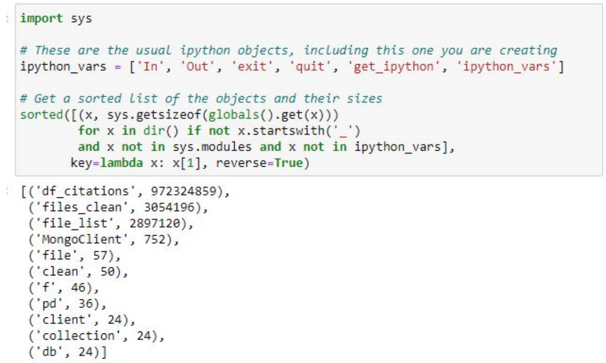
Table 1: Technology estimated cost.

4.6 Data Ingestion

The team has collected the required data in json and csv files that cover the corpus of physics articles. The team merged the cross citation csv to the json files with the purpose of more easily accessing and manipulating the data. Originally we implemented a multi threading python script to re-parse the data and save to our MongoDB database hosted in ATLAS but since memory management was critical we had to revert back to single threading and required a 60+ hour data insertion via MongoDB.

Giving that contributions to the APS data comes all over the world, we had to be carefully make sure that we encode all our data to UTF-8 as special characters are heavily used in titles, names and institutions.

As a team decision, we moved on to merge the raw APS article corpus to the cross citation data so we don't need to manipulate large data frames each interaction.



```
import sys

# These are the usual ipython objects, including this one you are creating
ipython_vars = ['In', 'Out', 'exit', 'quit', 'get_ipython', 'ipython_vars']

# Get a sorted list of the objects and their sizes
sorted([(x, sys.getsizeof(globals().get(x)))
        for x in dir() if not x.startswith('_')
        and x not in sys.modules and x not in ipython_vars],
       key=lambda x: x[1], reverse=True)
```

```
[('df_citations', 972324859),
 ('files_clean', 3054196),
 ('file_list', 2897120),
 ('MongoClient', 752),
 ('file', 57),
 ('clean', 50),
 ('f', 46),
 ('pd', 36),
 ('client', 24),
 ('collection', 24),
 ('db', 24)]
```

Figure 2: Example of memory allocation in our python notebook

That increased our database processing build time from 4h to +60h but at the end we believe that this decision will significantly improve our data analysis capabilities by support MongoDB **find()** query at an author and institution level as well as simplified node navigation as citing and cited nodes got included in the data.

4.7 API Development

API development in NodeJS is complete based on the proposed data structure defined in MongoDB and stubbed methods are in place where we are still working through thoughts on experiments.

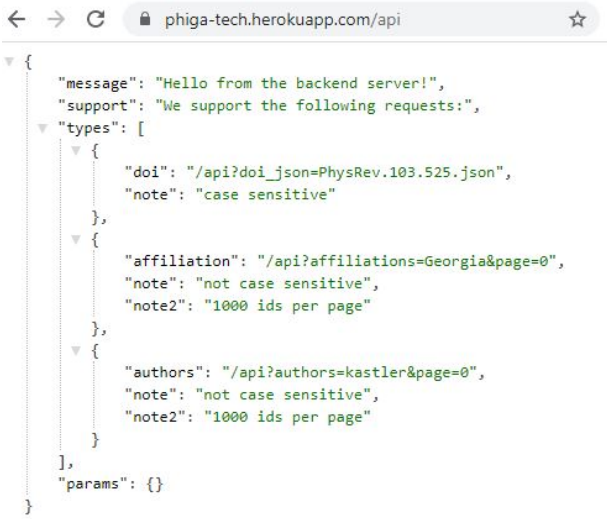
The API is published and fully functional in our Heroku box at: https://phiga-tech.herokuapp.com/api. Where we support three types of data extraction of:

by DOI: /api?doi_json
by affiliations: /api?affiliations
by authors: /api?authors

Examples of the Schema are represented in Figure 3.

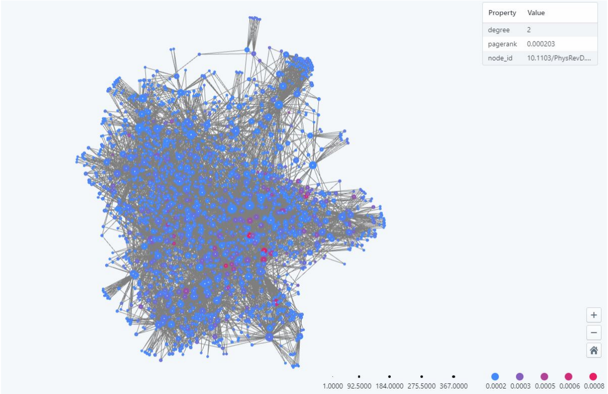
4.8 User Interface

A rough initial network was graphed (Figure 4) using ArgoLite. The network is part of our data discovery efforts to understand our dataset. One of the identified issues is that DOI level representations are not recommended as we have a big dataset. For that reason we decided to focus our initial efforts in building a working API where we could better navigate and interpret the data.



```
{
    "message": "Hello from the backend server!",
    "support": "We support the following requests:",
    "types": [
        {
            "doi": "/api?doi_json=PhysRev.103.525.json",
            "note": "case sensitive"
        },
        {
            "affiliation": "/api?affiliations=Georgia&page=0",
            "note": "not case sensitive",
            "note2": "1000 ids per page"
        },
        {
            "authors": "/api?authors=kastler&page=0",
            "note": "not case sensitive",
            "note2": "1000 ids per page"
        }
    ],
    "params": {}
}
```

Figure 3: API get Schema



Property	Value
degree	2
pagerank	0.000203
node_id	10.1103/PhysRevD....

Figure 4: ArgoLite initial discovery network map

5 EXPERIMENTS/EVALUATION

Our project will provide a user interface that will allow for search and visualization of physics publications. To test the success of this application, we intend to conduct the following experiments.

5.1 Critical authors and their relationship:

We will use a reference set of authors to determine their co-authors and papers that are cited using current search engines. We will then determine how long this process takes to assemble the results. We will then

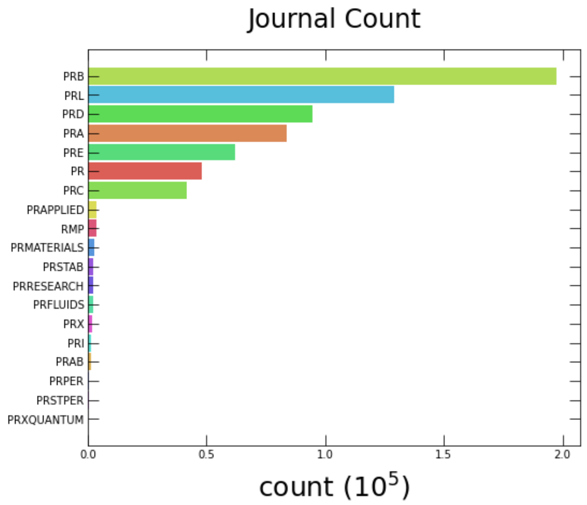


Figure 5: Data Summary with the number of publications for each journal.

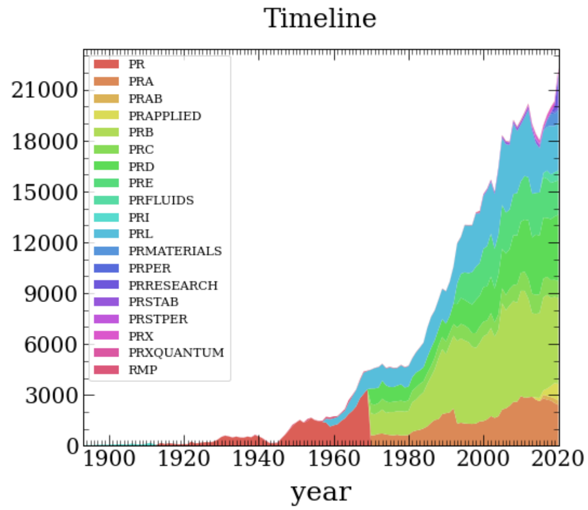


Figure 6: Data Summary with a timeline evolution.

compare this process and time with the implementation provided by our project for comparison.

Moving on to a more deep dive data analysis, we can evaluate ritical authors are identified by searching in our dataset for Nobel[1] prize laureates and evaluate their network and how they connect.We can then analyze their impact on the published papers and to what degree; wrote/co-wrote by or referenced by another paper.

We also have access to each author institution and with that data, we can network Georgia Tech contributions and map each areas of physics Georgia Tech most contribute. We could also use that information to compare institutions on their contributions. We plan to identify critical nodes as papers that have a high page rank and high degree in our dataset. From there we can observe the ratio of how many of these papers are associated with Georgia Tech and to what degree; published by Georgia Tech, co-wrote by a Georgia Tech associate, or references a Georgia Tech paper.
The impact of Georgia Tech in our dataset can be measured by these ratios and put into several tables to be displayed in our poster presentation. Corresponding graphs created by the team in ArgoLite will be presented as well with their respective tables.

6 CONCLUSIONS AND DISCUSSION

6.1 Innovations

Our analysis could have a great impact on how APS (American Physics Society) data is shared today and create a new standard moving away from the file base on S3 buckets to a REST API architecture or even a graphQL architecture in the future. What we built as the infrastructure for this project offers better flexibility, scalability and faster access to the data.

As our innovation points:

- We built an API for the APS dataset that contains not only the article corpus that also includes a cross citation (citing and cited nodes). This is a complete innovation compared to the current APS raw dataset that lives only at an AWS S3 bucket.
- We plan on mapping out how the 19 journals from APS (are of physics) correlates to each other. If we are able to achieve this goal, this is a new analysis and can be leveraged to identify potential clusters.
- As we are able to navigate from node to node (citing and cited) as well as filtering data by institution giving our MongoDB database structure. This can potentially be used to evaluate academic

[1] https://www.britannica.com/topic/Winners-of-the-Nobel-Prize-for-Physics-1856942

institutions on their contributions to physics and create a comparative scale of academic impact.

We plan on offering to the APS (American Physics Society) our infrastructure as a proposal for them to improve their data sharing capabilities once the project is finalized.

6.2 Audience

This work can be used as a base for a better index/impact measurement per paper. That information could help authors in finding more relevant papers for their research.

6.3 Success Potential Impact and Measures

Measures of success will focus on the ability of our test team to identify, interpret, and infer the data presented in the co-citation map solution. The data should be legibly, and appropriately represented in the graph to reflect the insight to it's intended audience. The impact of success may further the understanding and utilization of scholarly media going forward for the academic community.

6.4 Risks and Payoffs

Large unstructured dataset in individual json files (+678k files +4GB). If we are able to structure the data, we will be able to move to goal (2) and (3) and provide a deeper understanding of the data and its relationship to users

6.5 Work Distribution

All team members have contributed similar amount of effort. The coding workload of the project so far has been lead by Mr. Costa due to his expertise and knowledge of the proposal and methods used with other team members helping where applicable and appropriate. Document and proposal writing was completed as a team via Overleaf sharing, conference call and Slack communication. Each team member edited, revised, and contributed their ideas to the submitted version of this paper and the progress for the project overall.

6.6 Findings

The team discovered that China, India, France, Germany, Canada and Russia were identified as the largest countries that contribute to the American physics society papers as shown in Figure 7. India is particularly strong in the PR journal (PR = 'PhysRev', "archive published by the American Physical Society.

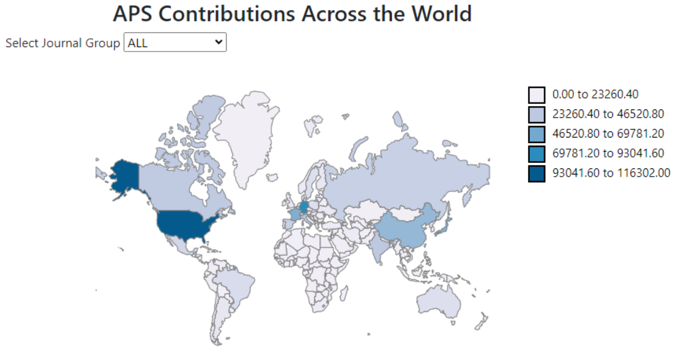


Figure 7: Choropleth map using d3js to visualize countries per contribution (volume of publication).

PROLA provides immediate access to the APS journal collection dating back to the first volume of each journal. A subscription to PROLA gives access to all journal content, except for the current year and the preceding three years"). The PRL and PRB are the two journals with the most publications - and they exercise influence in most of the other journals as seen below in Figure 8.

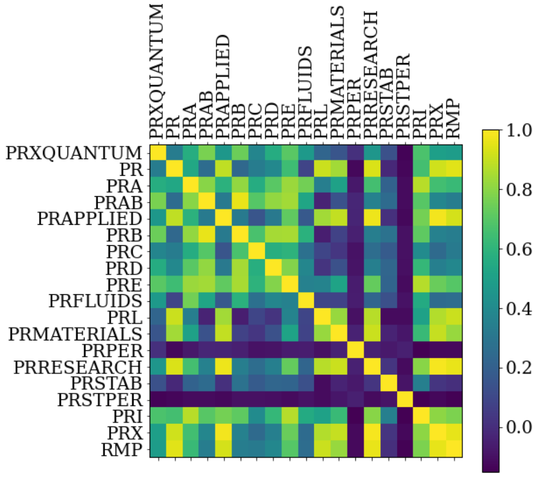


Figure 8: Correlation Matrix of Physics Journals.

The American Physics Society (APS) curates a cross-citation map in a machine-readable format (JavaScript Object Notation - JSON) that we utilized to index the full set of physics publications. 678,916 nodes/publications since 1913 in 19 journals and 8,850,334 edges/cross references as seen below in Figure 9.

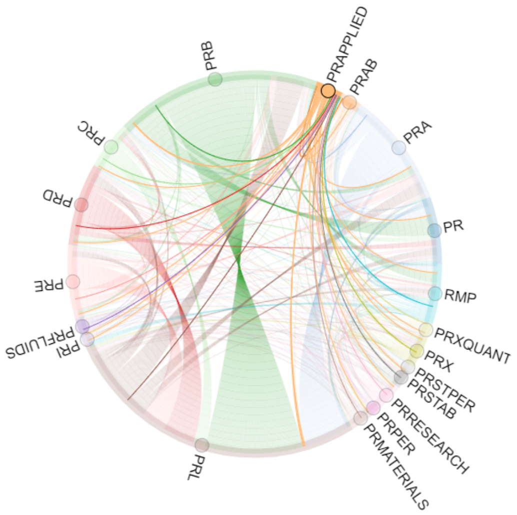


Figure 9: Chord plot that shows the proportion of publications, by journal, and their cross-citations amongst the network.

Our solution provides a web-based explorer that leverages the cross citations network API and the APS data set loaded into MongoDB to provide quick and easy access to publications and their cross citations.

6.7 Future Progression

The team was not able to finish analysis on individual institute's papers and Georgia Tech's influence on physics papers. If given more time, we would have pursued this venture to see not only Georgia Tech's impact but also the impact of all institutes over time. Furthermore, we would want to analyze other fields individually and all academic fields. The APS (American Physics Society) has 19 major publication journals each addressing a area of physics. We plan to determine what are the areas (journals) that have more contributions year by year. Since the data that we have also contains the cross citations (citing and cited nodes), we are also able to map out how each area impacts the other and create an **Hierarchical Edge Bundling** map showing what are the areas of physics that are more correlated.

REFERENCES

[1] arXiv. 1996. *arXiv.org - physics*. https://arxiv.org/archive/physics

[2] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. 2009. Gephi: An Open Source Software for Exploring and Manipulating Networks. http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154

[3] Colin B. Clement, Matthew Bierbaum, Kevin P. O'Keeffe, and Alexander A. Alemi. 2019. On the Use of ArXiv as a Dataset. arXiv:1905.00075 [cs.IR]

[4] Zhuoren Jiang, Yue Yin, Liangcai Gao, Yao Lu, and Xiaozhong Liu. 2018. Cross-language Citation Recommendation via Hierarchical Representation Learning on Heterogeneous Graph. arXiv:1812.11709 [cs.IR]

[5] Siwei Li, Zhiyan Zhou, Anish Upadhayay, Omar Shaikh, Scott Freitas, Haekyu Park, Zijie J. Wang, Susanta Routray, Matthew Hull, and Duen Horng Chau. 2020. Argo Lite: Open-Source Interactive Graph Exploration and Visualization in Browsers. In *Proceedings of the 29th ACM International Conference on Information Knowledge Management* (Virtual Event, Ireland) *(CIKM '20)*. Association for Computing Machinery, New York, NY, USA, 3071–3076. https://doi.org/10.1145/3340531.3412877

[6] Siwei Li, Zhiyan Zhou, Anish Upadhayay, Omar Shaikh, Scott Freitas, Haekyu Park, Zijie J. Wang, Susanta Routray, Matthew Hull, and Duen Horng Chau. 2020. Argo Lite: Open-Source Interactive Graph Exploration and Visualization in Browsers. In *Proceedings of the 29th ACM International Conference on Information Knowledge Management* (Virtual Event, Ireland) *(CIKM '20)*. Association for Computing Machinery, New York, NY, USA, 3071–3076. https://doi.org/10.1145/3340531.3412877

[7] Xia Lin, Howard D. White, and Jan Buzydlowski. 2003. Real-time author co-citation mapping for online searching. *Information Processing & Management* 39, 5 (2003), 689–706. https://doi.org/10.1016/S0306-4573(02)00037-7

[8] Ayaka Saka and Masatsura Igami. 2007. Mapping Modern Science Using Co-citation Analysis. In *2007 11th International Conference Information Visualization (IV '07)*. 453–458. https://doi.org/10.1109/IV.2007.77

[9] Henry Small. 1999. Visualizing Science by Citation Mapping. *JASIS* 50 (07 1999), 799–813. https://doi.org/10.1002/(SICI)1097-4571(1999)50:93.3.CO;2-7

[10] Supaporn Tantanasiriwong and Choochart Haruechaiyasak. 2014. Cross-domain citation recommendation based on Co-Citation Selection. *2014 11th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)* 1, 1 (May 2014), 1–4. https://doi.org/10.1109/ECTICon.2014.6839810

[11] Swizec Teller. 2013. *Data Visualization with D3.Js*. Packt Publishing.

[12] Nees Jan van Eck and Ludo Waltman. 2006. VOS: A New Method for Visualizing Similarities between Objects. *ERIM report series research in management Erasmus Research Institute of Management* ERS-2006-020-LIS (April 2006). http://hdl.handle.net/1765/7654

[13] Nees Jan van Eck and Ludo Waltman. 2014. CitNetExplorer: A new software tool for analyzing and visualizing citation networks. *Journal of Informetrics* 8, 4 (2014), 802–823. https://doi.org/10.1016/j.joi.2014.07.006

[14] Ludo Waltman, Nees Jan van Eck, and Ed C.M. Noyons. 2010. A unified approach to mapping and clustering of bibliometric networks. *Journal of Informetrics* 4, 4 (2010), 629–635. https://doi.org/10.1016/j.joi.2010.07.002

[15] Hongjiang Yue. 2010. Core and visualization analysis based on network of co-citation. *2010 2nd IEEE International Conference on Information Management and Engineering* 1, 1 (April 2010), 266–269. https://doi.org/10.1109/ICIME.2010.5478291