Big Data Analysis Project

Under the guidance of Prof. Pushparaj Shetty D.

By:

Kushagra Singh	:: 232CD014
Sameer Prajapati	:: 232CD024
Sohham Seal	:: 232CD030

Agenda

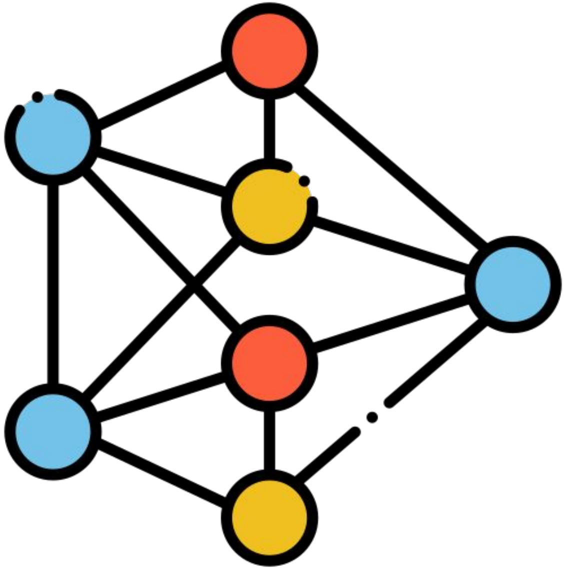
Data Acquisition

Preprocessing

Queries & Results

Neo4J

Data Visualization



Data Acquisition

Scopus:

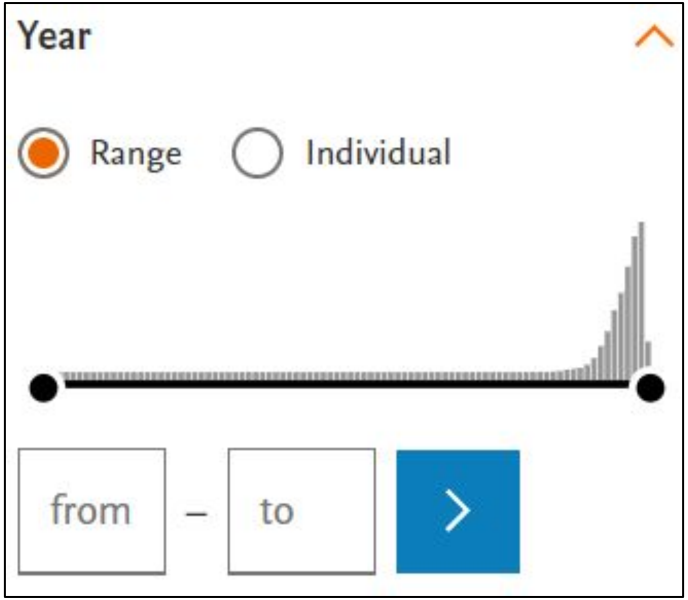
Scopus is a database of peer-reviewed literature, including books, scientific journals, and conference proceedings. It covers research from the fields of science, technology, medicine, social science, and arts and humanities. Scopus is owned and maintained by Elsevier, a publisher of scientific, technical, and medical content.

Data Acquisition

Features of Scopus:

- Structured data: Very easy to adapt in relational databases
- Clean data: No erroneous special characters in strings
- Dense data: Not much missing data
- Numerous filters readily available: A wide variety of filters to choose from

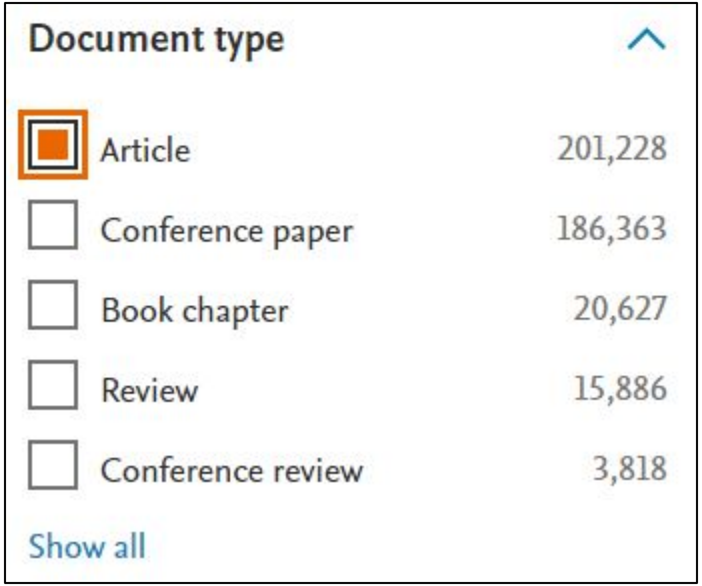
Filters



Year

● Range ○ Individual

from – to

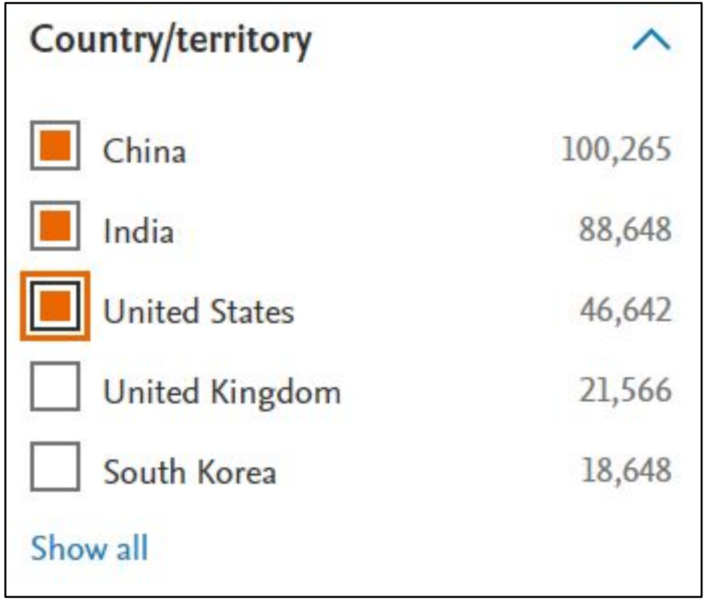


Document type

■	Article	201,228
□	Conference paper	186,363
□	Book chapter	20,627
□	Review	15,886
□	Conference review	3,818

Show all

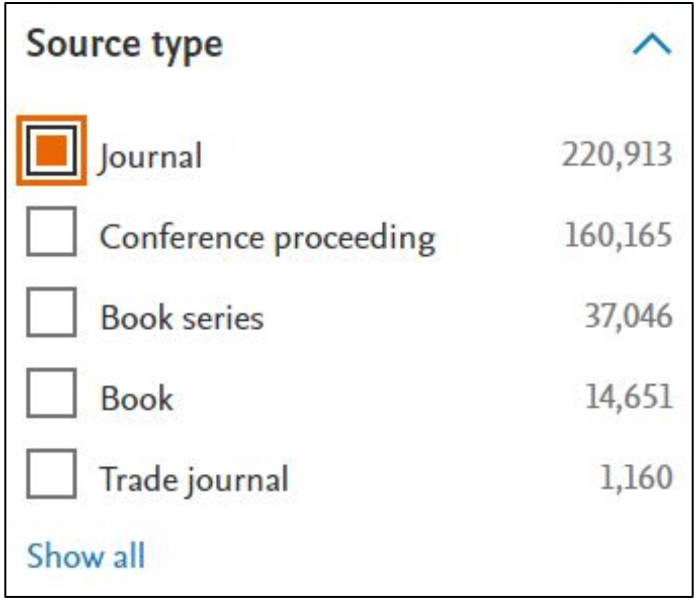
Filters



Country/territory

☑	China	100,265
☑	India	88,648
☑	United States	46,642
□	United Kingdom	21,566
□	South Korea	18,648

Show all

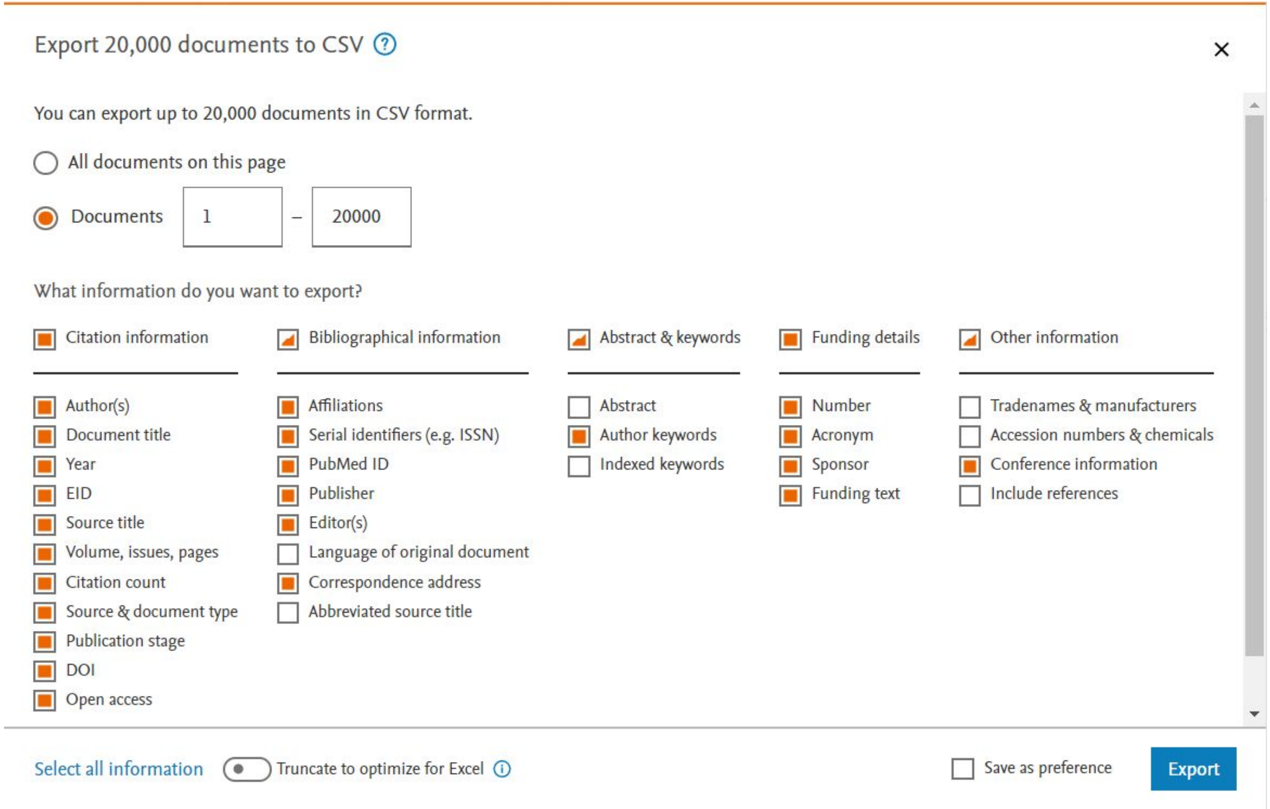


Source type

☑	Journal	220,913
□	Conference proceeding	160,165
□	Book series	37,046
□	Book	14,651
□	Trade journal	1,160

Show all

Download Options



Export 20,000 documents to CSV ?

You can export up to 20,000 documents in CSV format.

○ All documents on this page

● Documents 1 – 20000

What information do you want to export?

☑ Citation information

- ☑ Author(s)
- ☑ Document title
- ☑ Year
- ☑ EID
- ☑ Source title
- ☑ Volume, issues, pages
- ☑ Citation count
- ☑ Source & document type
- ☑ Publication stage
- ☑ DOI
- ☑ Open access

◪ Bibliographical information

- ☑ Affiliations
- ☑ Serial identifiers (e.g. ISSN)
- ☑ PubMed ID
- ☑ Publisher
- ☑ Editor(s)
- □ Language of original document
- ☑ Correspondence address
- □ Abbreviated source title

◪ Abstract & keywords

- □ Abstract
- ☑ Author keywords
- □ Indexed keywords

☑ Funding details

- ☑ Number
- ☑ Acronym
- ☑ Sponsor
- ☑ Funding text

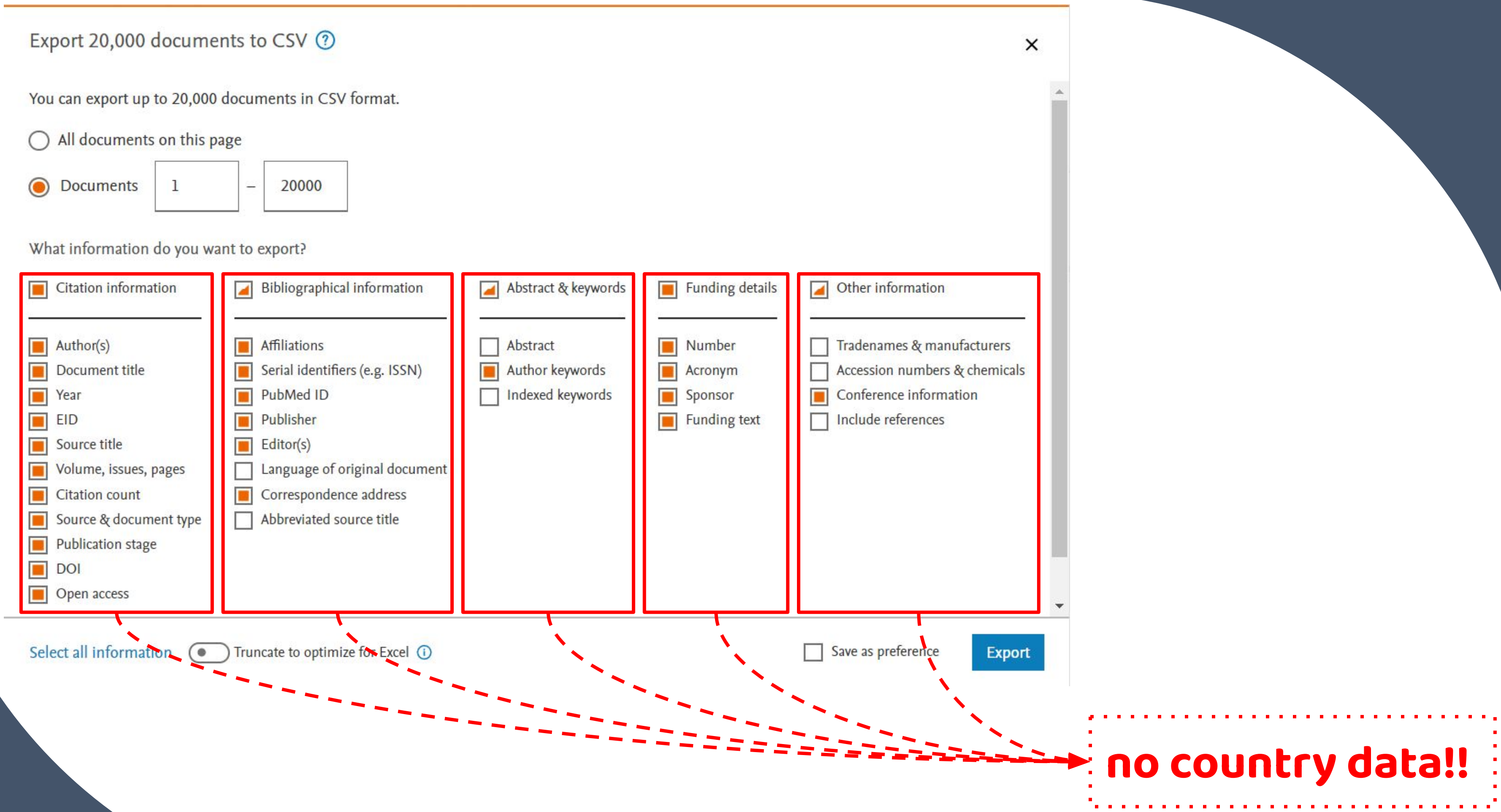
◪ Other information

- □ Tradenames & manufacturers
- □ Accession numbers & chemicals
- ☑ Conference information
- □ Include references

○ Truncate to optimize for Excel ⓘ

□ Save as preference

BUT…



Export 20,000 documents to CSV ?

You can export up to 20,000 documents in CSV format.

○ All documents on this page

● Documents 1 – 20000

What information do you want to export?

☑ Citation information
- ☑ Author(s)
- ☑ Document title
- ☑ Year
- ☑ EID
- ☑ Source title
- ☑ Volume, issues, pages
- ☑ Citation count
- ☑ Source & document type
- ☑ Publication stage
- ☑ DOI
- ☑ Open access

◪ Bibliographical information
- ☑ Affiliations
- ☑ Serial identifiers (e.g. ISSN)
- ☑ PubMed ID
- ☑ Publisher
- ☑ Editor(s)
- □ Language of original document
- ☑ Correspondence address
- □ Abbreviated source title

◪ Abstract & keywords
- □ Abstract
- ☑ Author keywords
- □ Indexed keywords

☑ Funding details
- ☑ Number
- ☑ Acronym
- ☑ Sponsor
- ☑ Funding text

◪ Other information
- □ Tradenames & manufacturers
- □ Accession numbers & chemicals
- ☑ Conference information
- □ Include references

Select all information

○ Truncate to optimize for Excel ⓘ

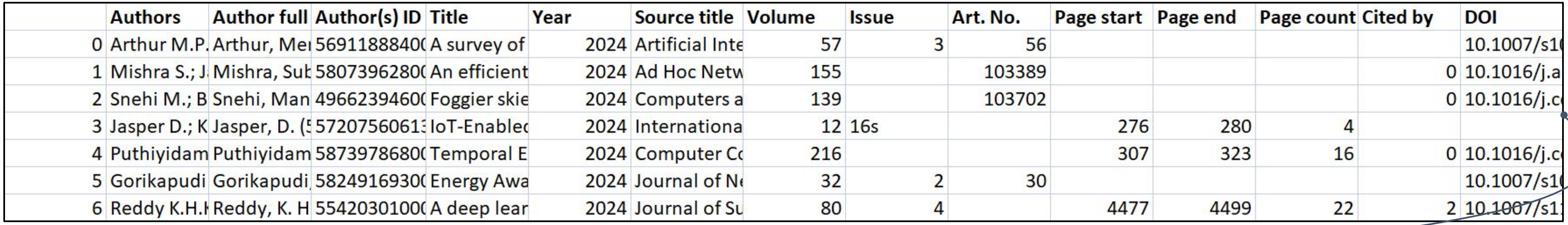
□ Save as preference

no country data!!

Solution:

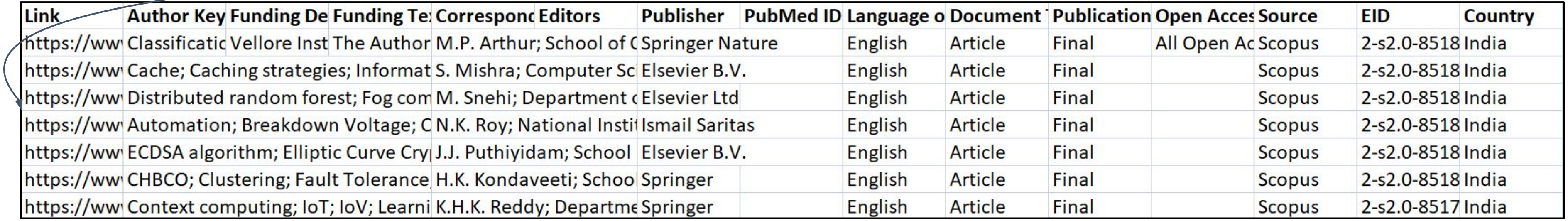
- Select the country in the filter.
- Download using preferred choices.
- Create a ‘country’ column and add the country name for all the records in that particular document.
- Merge all the records downloaded separately together!

Dataset



	Authors	Author full	Author(s) ID	Title	Year	Source title	Volume	Issue	Art. No.	Page start	Page end	Page count	Cited by	DOI
0	Arthur M.P.	Arthur, Me	5691188840(	A survey of	2024	Artificial Inte	57	3	56					10.1007/s1(
1	Mishra S.; J	Mishra, Sub	5807396280(	An efficient	2024	Ad Hoc Netw	155		103389				0	10.1016/j.a
2	Snehi M.; B	Snehi, Man	4966239460(	Foggier skie	2024	Computers a	139		103702				0	10.1016/j.c
3	Jasper D.; K	Jasper, D. (5	5720756061:	IoT-Enablec	2024	Internationa	12	16s		276	280	4		
4	Puthiyidam	Puthiyidam	5873978680(	Temporal E	2024	Computer Cc	216			307	323	16	0	10.1016/j.c
5	Gorikapudi	Gorikapudi,	5824916930(	Energy Awa	2024	Journal of Ne	32	2	30					10.1007/s1(
6	Reddy K.H.I	Reddy, K. H	5542030100(	A deep lear	2024	Journal of Su	80	4		4477	4499	22	2	10.1007/s1

continued…



Link	Author Key	Funding De	Funding Te:	Correspond	Editors	Publisher	PubMed ID	Language o	Document '	Publication	Open Acces	Source	EID	Country
https://ww	Classificatic	Vellore Inst	The Author	M.P. Arthur; School of (		Springer Nature		English	Article	Final	All Open Ac	Scopus	2-s2.0-8518	India
https://ww	Cache; Caching strategies; Informat			S. Mishra; Computer Sc		Elsevier B.V.		English	Article	Final		Scopus	2-s2.0-8518	India
https://ww	Distributed random forest; Fog com			M. Snehi; Department (		Elsevier Ltd		English	Article	Final		Scopus	2-s2.0-8518	India
https://ww	Automation; Breakdown Voltage; C			N.K. Roy; National Instit		Ismail Saritas		English	Article	Final		Scopus	2-s2.0-8518	India
https://ww	ECDSA algorithm; Elliptic Curve Cry			J.J. Puthiyidam; School		Elsevier B.V.		English	Article	Final		Scopus	2-s2.0-8518	India
https://ww	CHBCO; Clustering; Fault Tolerance,			H.K. Kondaveeti; Schoo		Springer		English	Article	Final		Scopus	2-s2.0-8518	India
https://ww	Context computing; IoT; IoV; Learni			K.H.K. Reddy; Departme		Springer		English	Article	Final		Scopus	2-s2.0-8517	India

Preprocessing

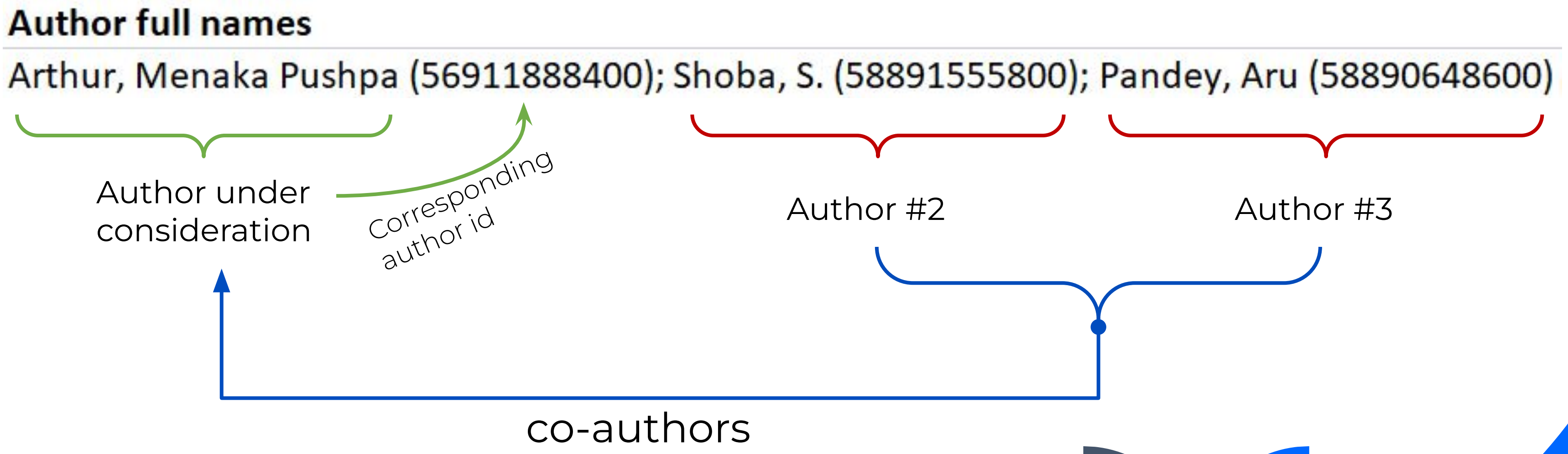
Cleaning, Modifying and Restructuring ...

Column Modifications

Drop excess columns, such as:

- Authors
- Source title
- Volume
- Issue
- Art. No.
- Page start, Page end, Page count,
- DOI
- Link
- Open Access
- EID
- Author Keywords
- Funding Texts
- Correspondence Address
- Editors
- PubMed ID
- Language of Original Document
- Document Type
- Publication Stage
- Source

Co-Author Extraction



Author full names

Arthur, Menaka Pushpa (56911888400); Shoba, S. (58891555800); Pandey, Aru (58890648600)

Author under consideration

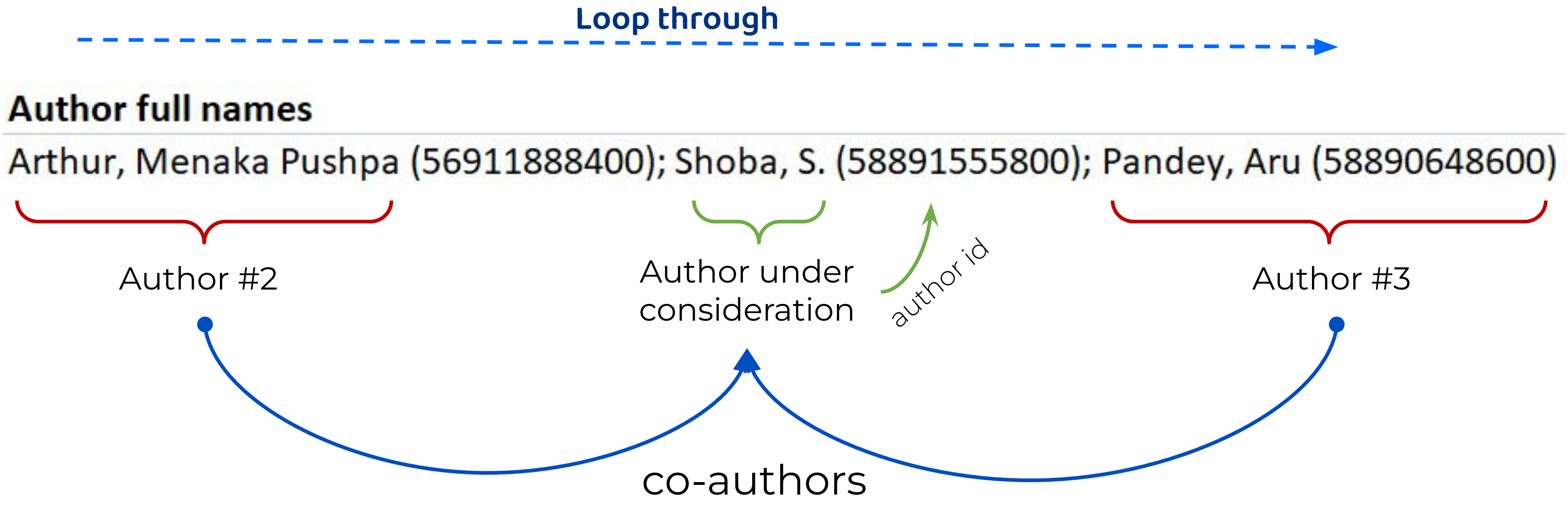
Corresponding author id

Author #2

Author #3

co-authors

Co-Author Extraction



Loop through

Author full names
Arthur, Menaka Pushpa (56911888400); Shoba, S. (58891555800); Pandey, Aru (58890648600)

Author #2

Author under consideration

author id

Author #3

co-authors

and so on ...

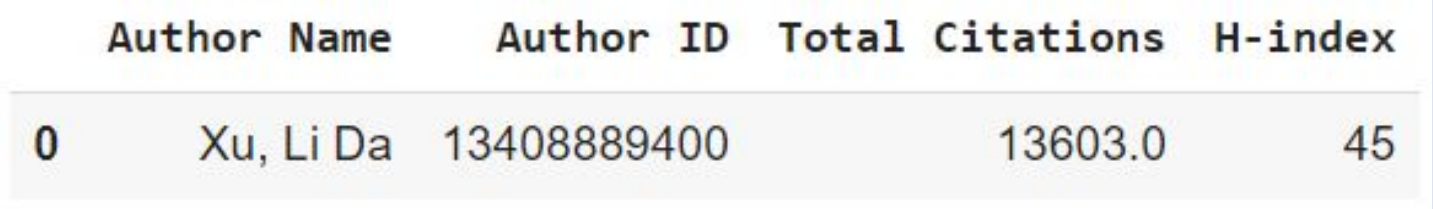
```
{
    Author Name #1: {
                    Author ID: 1234566778,          // Integer
                    Paper Names: [ Paper #1 name, Paper #2 name, … ],          // List
                    Citations: [ Paper #1 citations, Paper #2 citations, … ],          // List
                    Year of Publishing: [ Paper #1 year, Paper #2 year, … ],          // List
                    Funding: [ Paper #1 funding, Paper #2 funding, … ],          // List
                    Country: India          // String
                    Co-Authors: [Co-Author #1, Co-Author #2, … ]          // List
                   },
    Author Name #2: {
                      •
                      •          // Repeat
                      •
                    }, …
}
```

Queries

Solving the queries using Python…

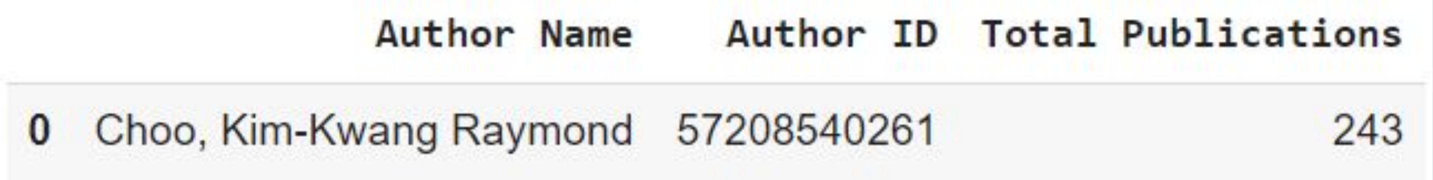
Code available at: link

(a) Highest cited author and his h-index



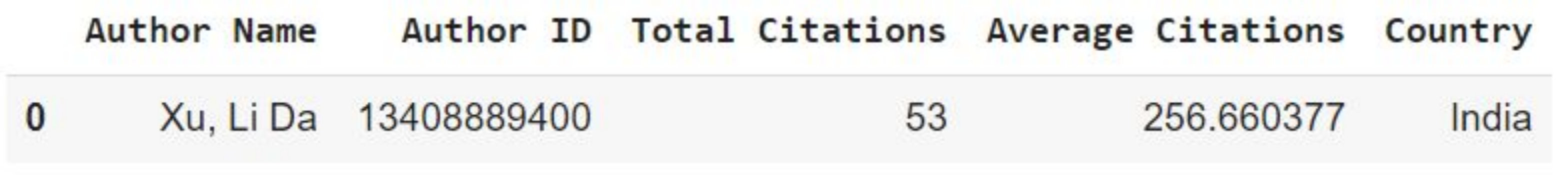
	Author Name	Author ID	Total Citations	H-index
0	Xu, Li Da	13408889400	13603.0	45

(b) Highest publication author



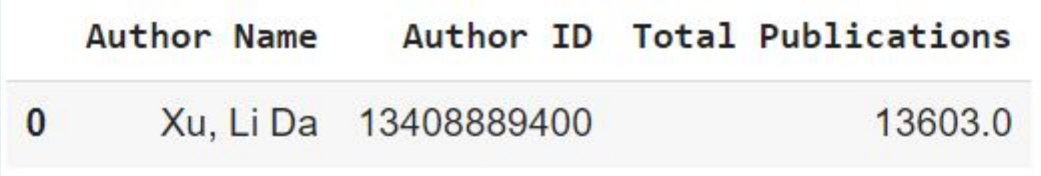
	Author Name	Author ID	Total Publications
0	Choo, Kim-Kwang Raymond	57208540261	243

(c) Highest cited author's avg citations and country name



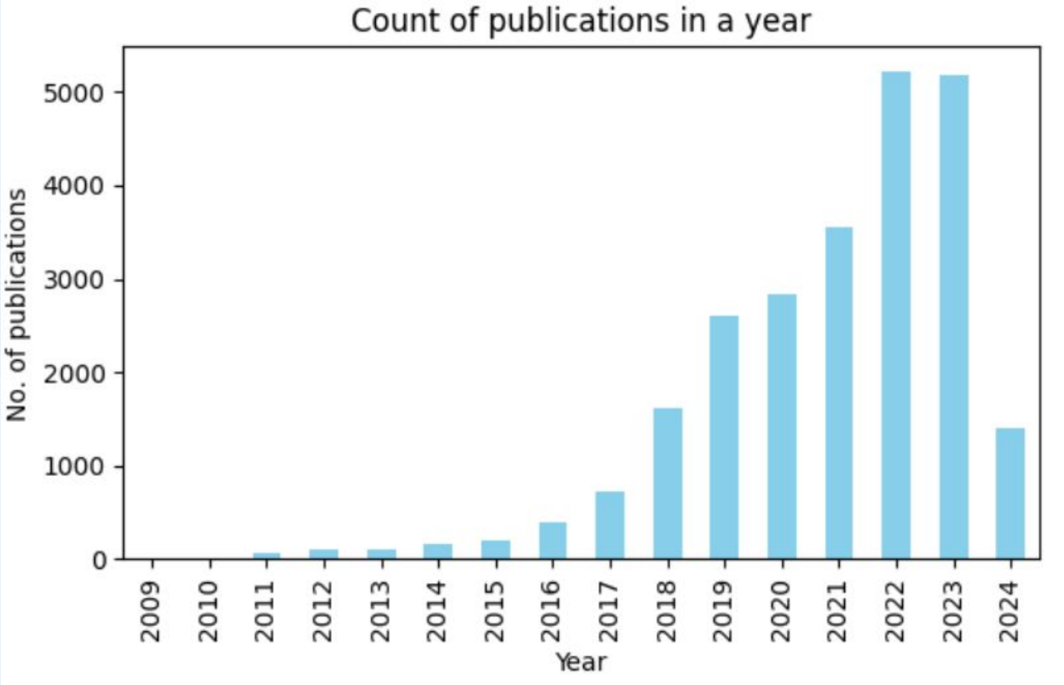
	Author Name	Author ID	Total Citations	Average Citations	Country
0	Xu, Li Da	13408889400	53	256.660377	India

(d) Total number of publications of the highest cited author

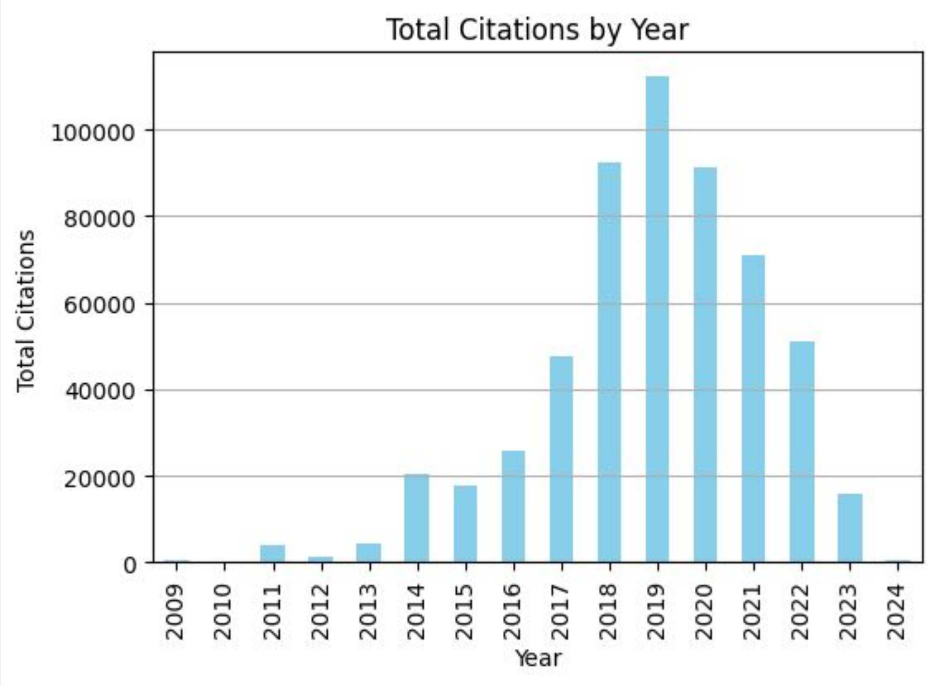


	Author Name	Author ID	Total Publications
0	Xu, Li Da	13408889400	13603.0

(e) Total publications in a year



(f) Total citations in a year



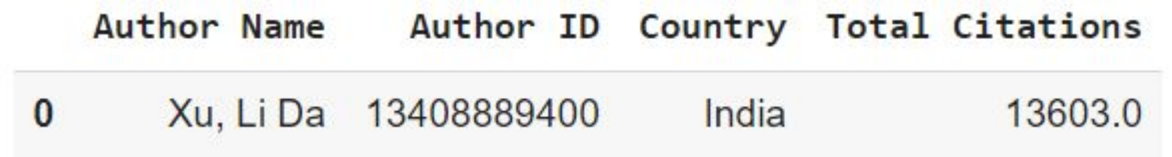
(g) Author(country) having highest co-authorship with indian authors

We will be unable to solve this problem as SCOPUS **does not provide** the data for each of the author separately.

This data need available through the author's profile, which means we need to open each author page and collect the data!!

The data has more than 10,000 authors, therefore, we can do so using **web-scraping** to automatically collect the required data!

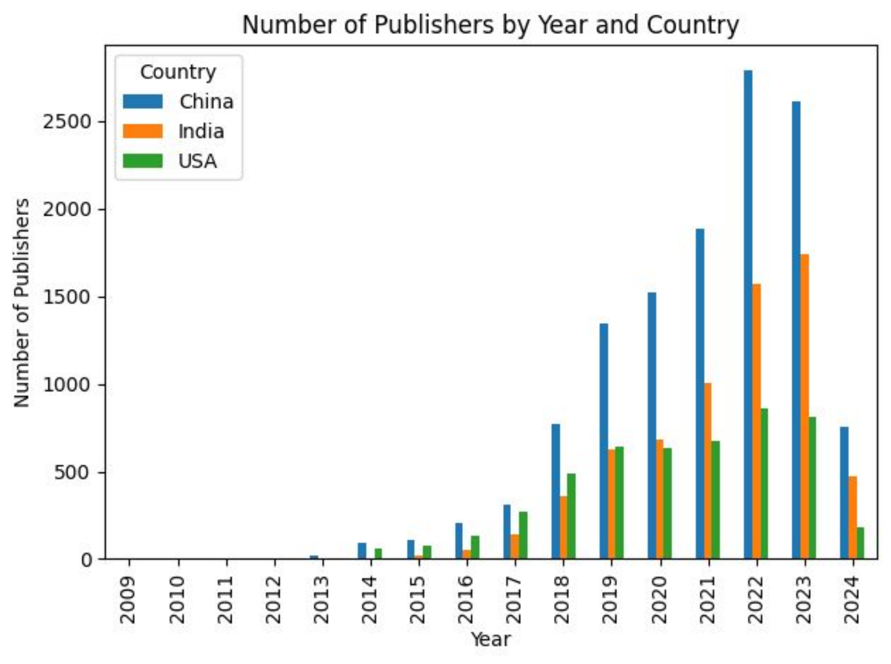
(h) Highest cited author from India and the university



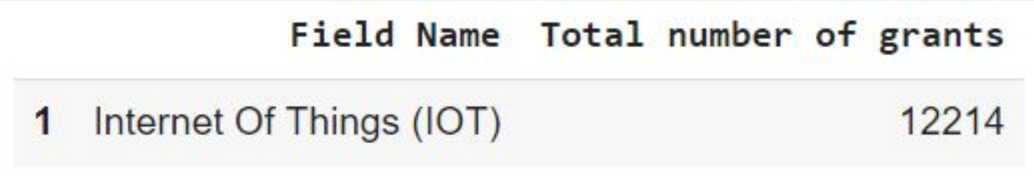
	Author Name	Author ID	Country	Total Citations
0	Xu, Li Da	13408889400	India	13603.0

**** University details had to be dropped as >73% records did not have the university details**

(i) Comparative year wise article publication analysis of India, China and USA

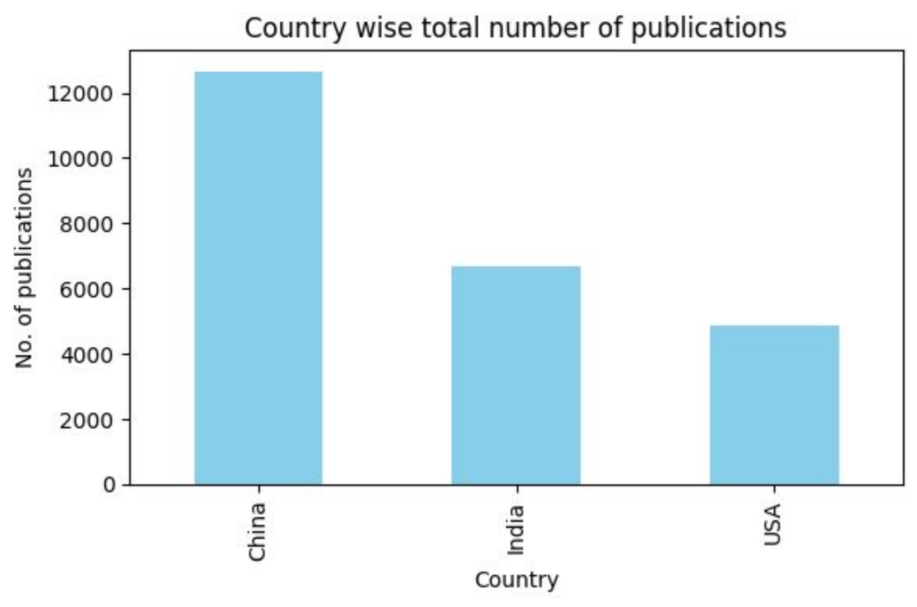


(j) Total number of grants given to the field



	Field Name	Total number of grants
1	Internet Of Things (IOT)	12214

(k) Country wise total number of publication



Neo4j

What do we use it for?

Tools we have used:

Neo4j Desktop

A desktop version of the open source software Neo4j.

Available for Windows, Linux & MacOS.

Readily contains Neo4j Browser, Neo4j Bloom and Neo4j ETL Tool.

Neo4j Browser

A developer-focused tool that allows you to execute Cypher queries and visualize the results.

Neo4j Bloom

A beautiful and expressive data visualization tool to quickly explore and freely interact with Neo4j’s graph data platform with no coding required.

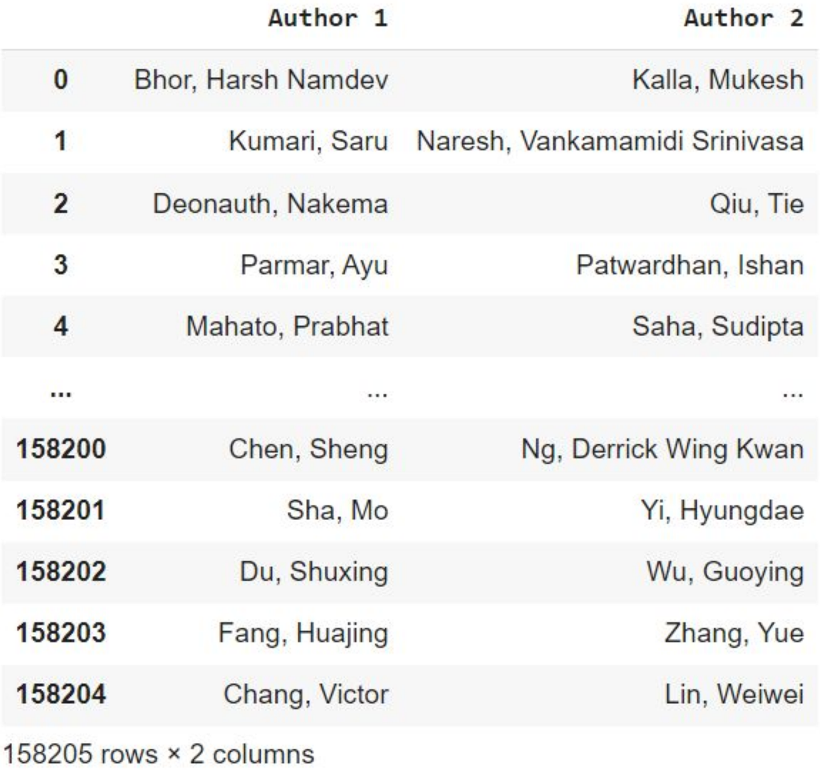
Process Undertaken…

Create a **.CSV** file of only unique relationship edges in python

Import the .CSV file in **Neo4j desktop** and use the browser to run queries for graph creation

Use **Bloom** to visualize the data in a beautifully generated manner.

.CSV file

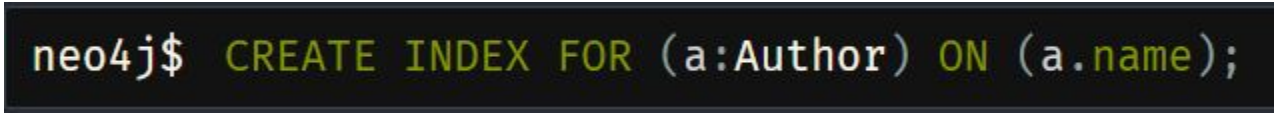


	Author 1	Author 2
0	Bhor, Harsh Namdev	Kalla, Mukesh
1	Kumari, Saru	Naresh, Vankamamidi Srinivasa
2	Deonauth, Nakema	Qiu, Tie
3	Parmar, Ayu	Patwardhan, Ishan
4	Mahato, Prabhat	Saha, Sudipta
...	...	...
158200	Chen, Sheng	Ng, Derrick Wing Kwan
158201	Sha, Mo	Yi, Hyungdae
158202	Du, Shuxing	Wu, Guoying
158203	Fang, Huajing	Zhang, Yue
158204	Chang, Victor	Lin, Weiwei

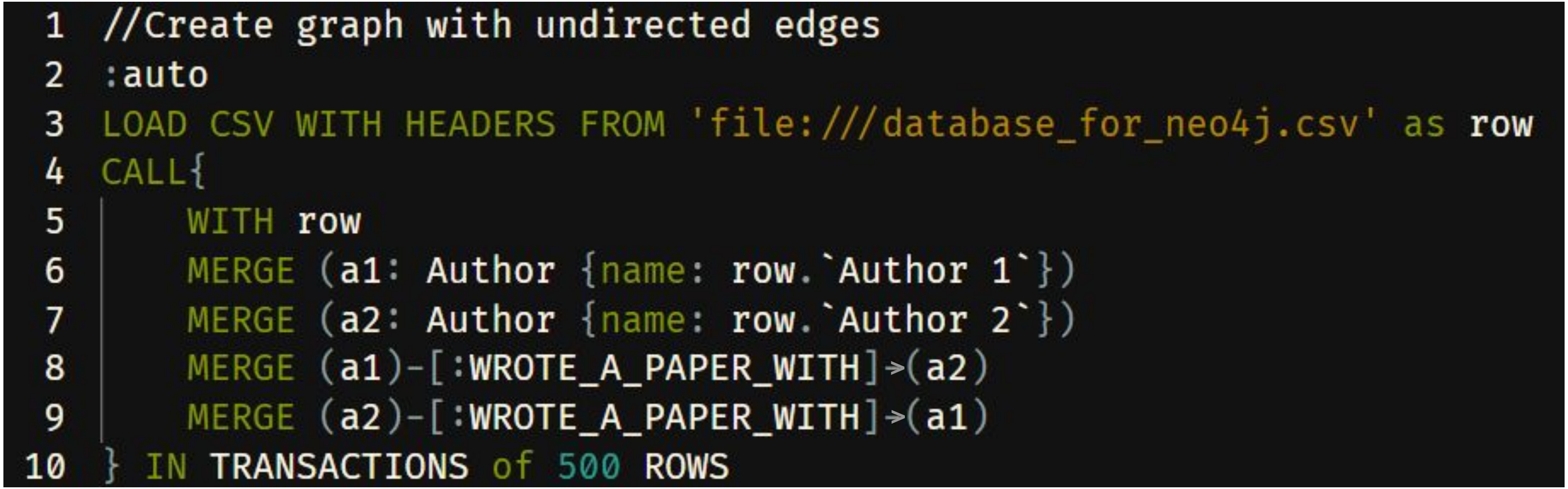
158205 rows × 2 columns

each record is a unique edge for the corresponding graph!

Queries



```
neo4j$ CREATE INDEX FOR (a:Author) ON (a.name);
```

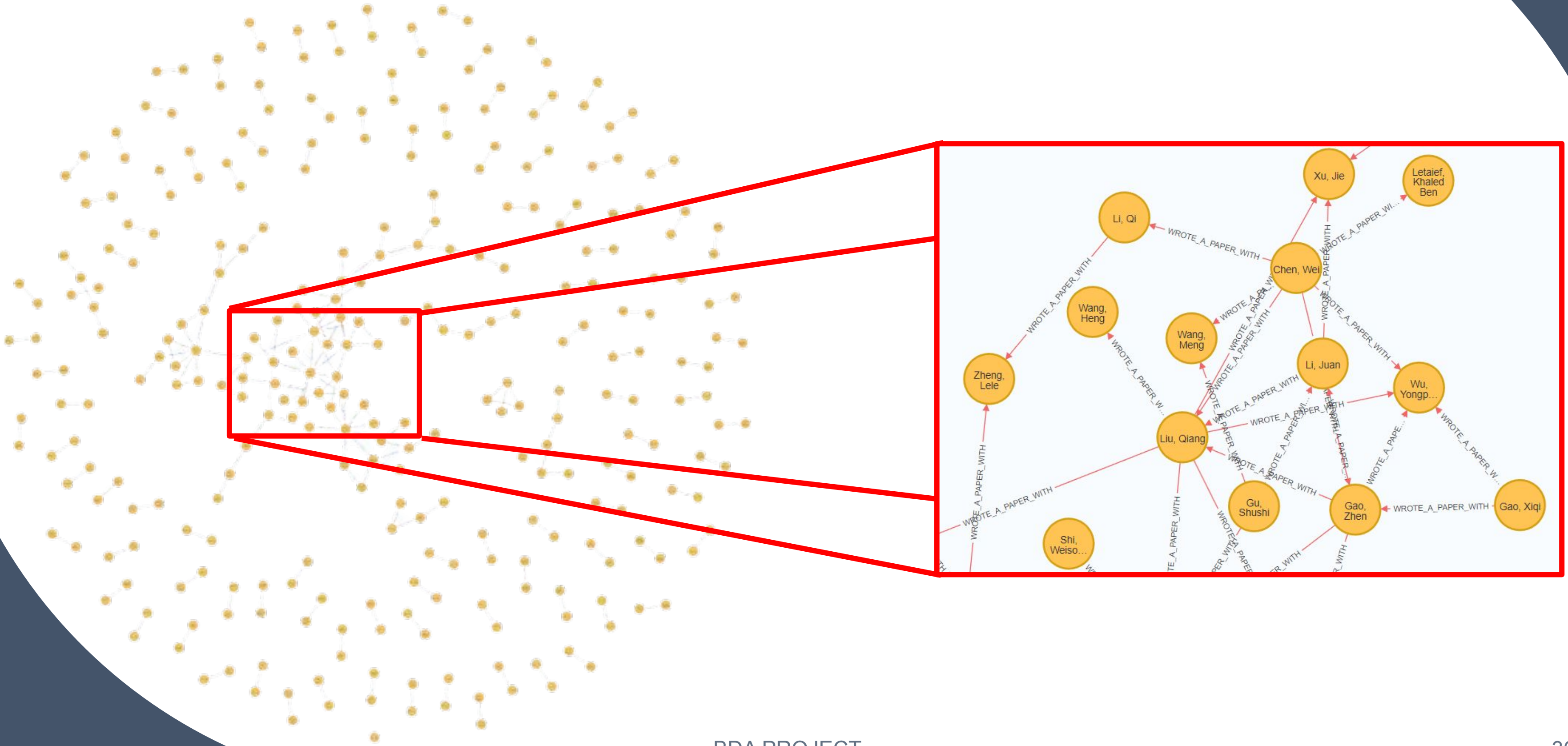


```
//Create graph with undirected edges
:auto
LOAD CSV WITH HEADERS FROM 'file:///database_for_neo4j.csv' as row
CALL{
    WITH row
    MERGE (a1: Author {name: row.`Author 1`})
    MERGE (a2: Author {name: row.`Author 2`})
    MERGE (a1)-[:WROTE_A_PAPER_WITH]->(a2)
    MERGE (a2)-[:WROTE_A_PAPER_WITH]->(a1)
} IN TRANSACTIONS of 500 ROWS
```

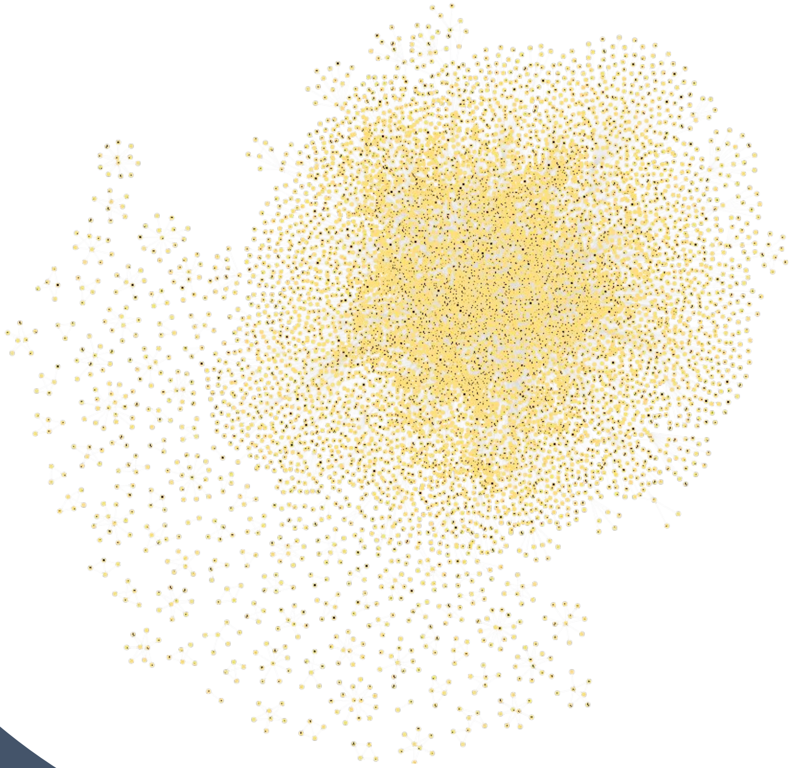
Data Visualization

Co-Author Graph Network, using Neo4J Bloom

Using Neo4j Browser (only 300 at a time)



Using Neo4j Bloom



10,000 Author Nodes

16,640 Relation Edges

DEMONSTRATION!

Thank you

Any Questions??

Link to code