FINAL REPORT

ORGANISATION OF INFORMATION

南京大學

Name: 1. ALCIDES BERNARDO TELLO

2. TULKINOV SHAVKATJON 李 克 涵

Student ID: 1) MG21145001

2) 502022145004

Professor: Yan Jiaqi

Summary of the task

This task aimed to assess students' understanding of information organization methods and techniques through practical operations in a specific literature field.

Here, we presented all the application learnt during the semester and asked for doing a task, plus a contribution using database on cloud computing.

Contents

1. Organizing literature data in Excel or a Databases

1.1 Background

The dataset provided by the teacher consisted of SCI journal papers on blockchain along with their citation data. To work with the dataset effectively, it was essential to ensure its proper organization. Upon examining the dataset, we discovered that the Authors_Full_Names column was not well-structured. Each cell in this column contained multiple authors separated by ";". Hence, the initial task assigned by the teacher was to separate the authors into distinct rows.

To accomplish this task, we leveraged our knowledge from the Information Organization Course and opted to use Excel, a widely used computerized spreadsheet program. Excel provided a convenient platform to perform the required author separation. In order to streamline the process, we installed a tool called "kutools" (refer to Figure 01) that enhanced the functionality of Excel and facilitated the efficient splitting of authors within each cell of the Authors_Full_Names column in the dataset.

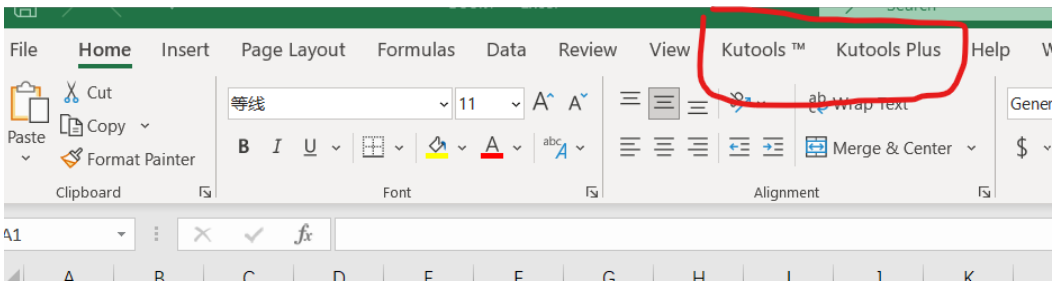


Figure 01. Installed Kutools in Excel

1.2 Splitting the data

Method 1: Using R

To split the data into separate elements, there are two methods you can use: one using R and another using Excel.

In R, you can use the 'str_split' function from the 'stringr' package to split the data. As we want to split a sentence using the delimiter " ; ", we can use the following code:

```
library(stringr)
str_split(Sentence, " ; ")
```

This will split the **Sentence** string wherever it encounters " ; " and return a list of the resulting substrings.

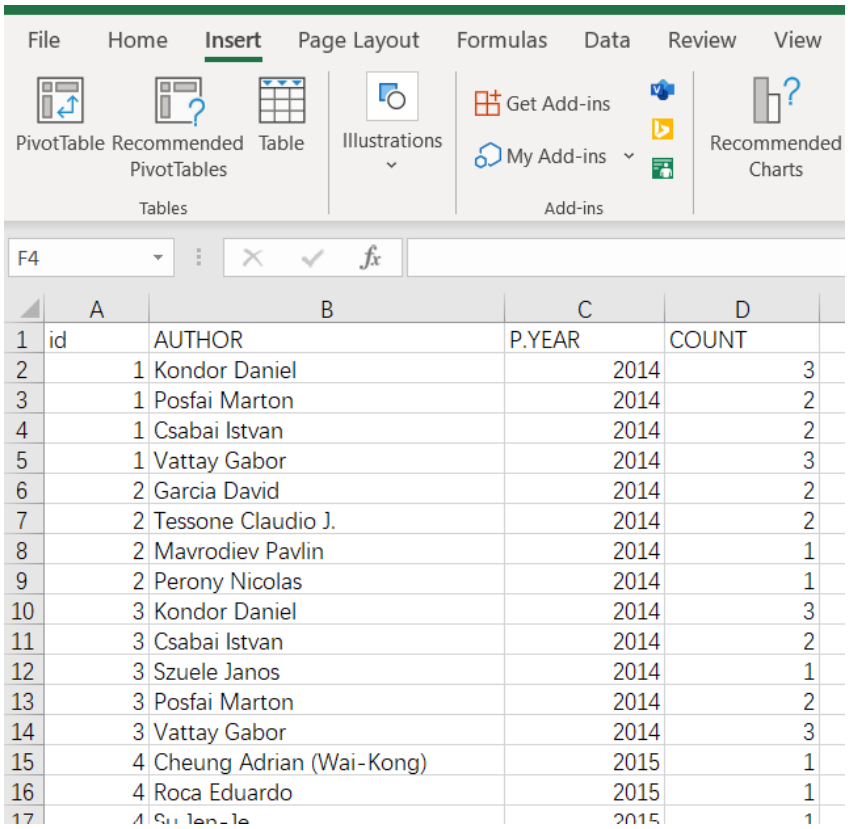
Method 2: Using excel

In Excel, we can use built-in functions to split the data into separate cells. Here's a step-by-step how we did:

1. We selected the column that contains the data we want to split (Authors_Full_Names).
2. Then, go to the "Data" tab in the Excel ribbon.
3. Click on the "Text to Columns" button.
4. In the "Convert Text to Columns Wizard," we choose the "Delimited" option and click "Next."
5. Selected the delimiter that separates our data (in our case, " ; ") and click "Next."
6. Then we choose the destination where we wanted to place the split data (a new column or range).
7. Click "Finish" to complete the splitting process.

Excel splatted the data based on the chosen delimiter and placed the separated elements into the specified destination cells.

After successfully using Kutools, we were able to separate the authors' full names in the dataset. The resulting output, shown in Figure 02, provides information about the authors, their respective IDs, and the number of articles they have written.



	A	B	C	D
1	id	AUTHOR	P.YEAR	COUNT
2	1	Kondor Daniel	2014	3
3	1	Posfai Marton	2014	2
4	1	Csabai Istvan	2014	2
5	1	Vattay Gabor	2014	3
6	2	Garcia David	2014	2
7	2	Tessone Claudio J.	2014	2
8	2	Mavrodiev Pavlin	2014	1
9	2	Perony Nicolas	2014	1
10	3	Kondor Daniel	2014	3
11	3	Csabai Istvan	2014	2
12	3	Szuele Janos	2014	1
13	3	Posfai Marton	2014	2
14	3	Vattay Gabor	2014	3
15	4	Cheung Adrian (Wai-Kong)	2015	1
16	4	Roca Eduardo	2015	1
17	4	Su Jen-Le	2015	1

Figure 02. The count indicates the number of papers for each author

In the figure 02, the "id" column represents the unique ID for each article, which helps to verify the consistency between authors and their respective article IDs. The "AUTHOR" column contains the separated results of each author's name in different rows. The "COUNT" column indicates the count of articles written by each author.

This organization allows for easier analysis and tracking of authors and their contributions to the articles in the dataset.

2. Clustering literature data using database

2.1 Background

Throughout our course, we have learned MySQL for Information Organization. So, in this step, we have applied our knowledge into practice. MySQL is an open-source relational database management system (RDBMS) that follows the relational model. In a relational database, data is organized into tables, where each table consists of rows and columns. Each row represents a record or an entity, and each column represents a specific attribute or field.

In MySQL, tables have a primary key, which uniquely identifies each row or record in the table. This primary key acts as a unique identifier for each record and helps establish relationships between tables.

The relational model in MySQL allows for the creation of relationships between tables using keys. For example, a foreign key can be used to establish a relationship between two tables by referencing the primary key of one table in another table.

2.2 Visualising the results

In order to visualize the results, we focused on a specific author, namely "Kondor Daniel." Using an INNER JOIN operation, we combined the relevant data from multiple tables. The INNER JOIN operation matches each row in one table with every row in the other table, allowing us to query rows that contain columns from both tables. Results shown in Figure 03 below:

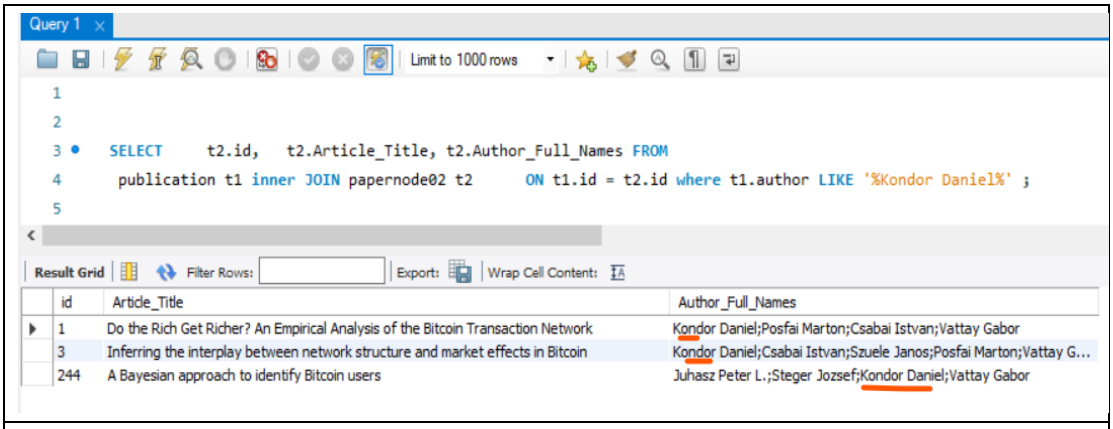


Figure 03: SQL query for selecting the specific author "Kondor Daniel"

3. A webpage using with Node Js with access to MySQL (In addition to cloud computing)

3.1 Background

Node.js is a runtime environment for executing JavaScript code outside of a web browser. It is a lightweight and cross-platform framework that is commonly used for server-side programming. Node.js allows developers to build scalable and efficient web applications by leveraging JavaScript on the server-side. Additionally, Node.js can be used as a server-side proxy to collect data from different third-party resources. It is also capable of building client-side applications, providing a versatile platform for JavaScript development.

3.2 Visualizing the relationship

3.2.1 Local server (using Node Js as server and client)

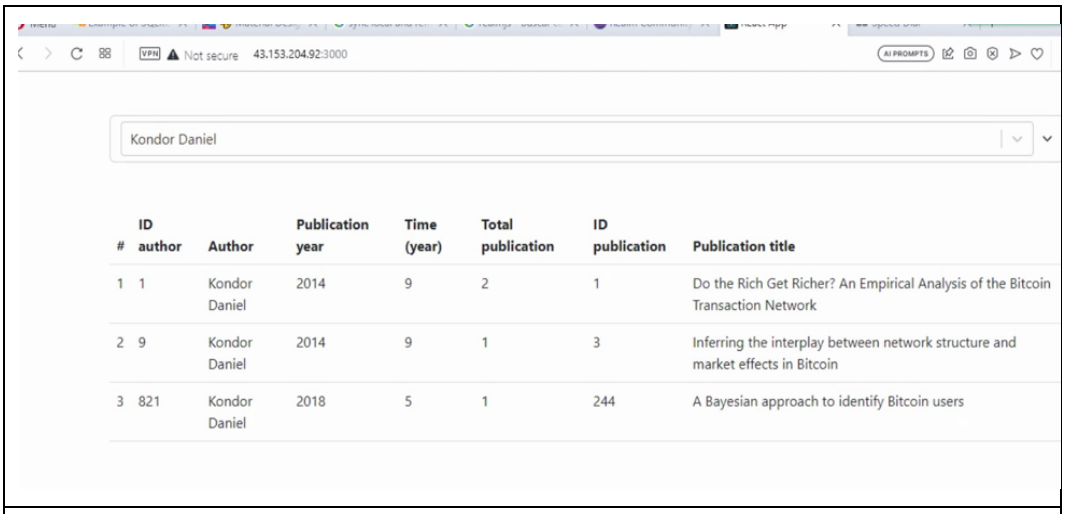
The server-side code, written in Node.js, handles incoming requests and responds with the requested data.

```
  res.header('Access-Control-Allow-Methods', 'GET, POST, OPTIONS, PUT, DELETE');
  res.header('Allow', 'GET, POST, OPTIONS, PUT, DELETE');
  next();
});
// Define a GET request handler

app.get('/author', async (req, res) => {
  const parameterValue = req.query.name;
  try {
    await database.connect();
    const results = await database.executeQuery(`select pu.id id, pu.author author, pu.publication_year publication_year, pu.year year, pu.count c
    inner join article ar on pu.id= ar.id where pu.author = '${parameterValue}'`);
    console.log(results);
    res.json(results); // Return the data as JSON response
  } catch (error) {
    console.error('Error fetching data:', error);
    res.status(500).json({ error: 'An error occurred while fetching data from the database.' });
  } finally {
    database.close();
  }
});

app.get('/authors', async (req, res) => {
  const parameterValue = req.query.name;
  try {
    await database.connect();
    const results = await database.executeQuery(`SELECT id, author FROM publication where author like '${parameterValue}%' limit 10`);
```

Figure 04: Coding on NodeJs

It utilizes an Express.js framework to define routes and handle the specific author's data. After coding and running as shown in figure 04, we created a web page that fetches data from a server using Node.js and displays it dynamically.

The achieved the output shown in Figure 05 below:



#	ID author	Author	Publication year	Time (year)	Total publication	ID publication	Publication title
1	1	Kondor Daniel	2014	9	2	1	Do the Rich Get Richer? An Empirical Analysis of the Bitcoin Transaction Network
2	9	Kondor Daniel	2014	9	1	3	Inferring the interplay between network structure and market effects in Bitcoin
3	821	Kondor Daniel	2018	5	1	244	A Bayesian approach to identify Bitcoin users

Figure 05: Displaying on webpage for a specific author

In Figure 05, we demonstrated how to display information on a webpage for a specific author. We can dynamically display data for a specific author on a webpage, providing an interactive and personalized browsing experience.

3.2.2 Using cloud computing:

In the context of cloud computing, MariaDB can be leveraged as a drop-in replacement for MySQL on a Debian Linux system. By using MariaDB, you can ensure backward compatibility with MySQL while taking advantage of its enhanced features and performance. To facilitate seamless access to your entire infrastructure, Termitus can be employed. Termitus is a cloud

computing platform that provides instant access to your infrastructure, allowing you to efficiently manage and utilize resources. By combining MariaDB as a reliable database management system and Termitus for streamlined infrastructure access, you can optimize your cloud computing environment for improved efficiency, scalability, and performance.

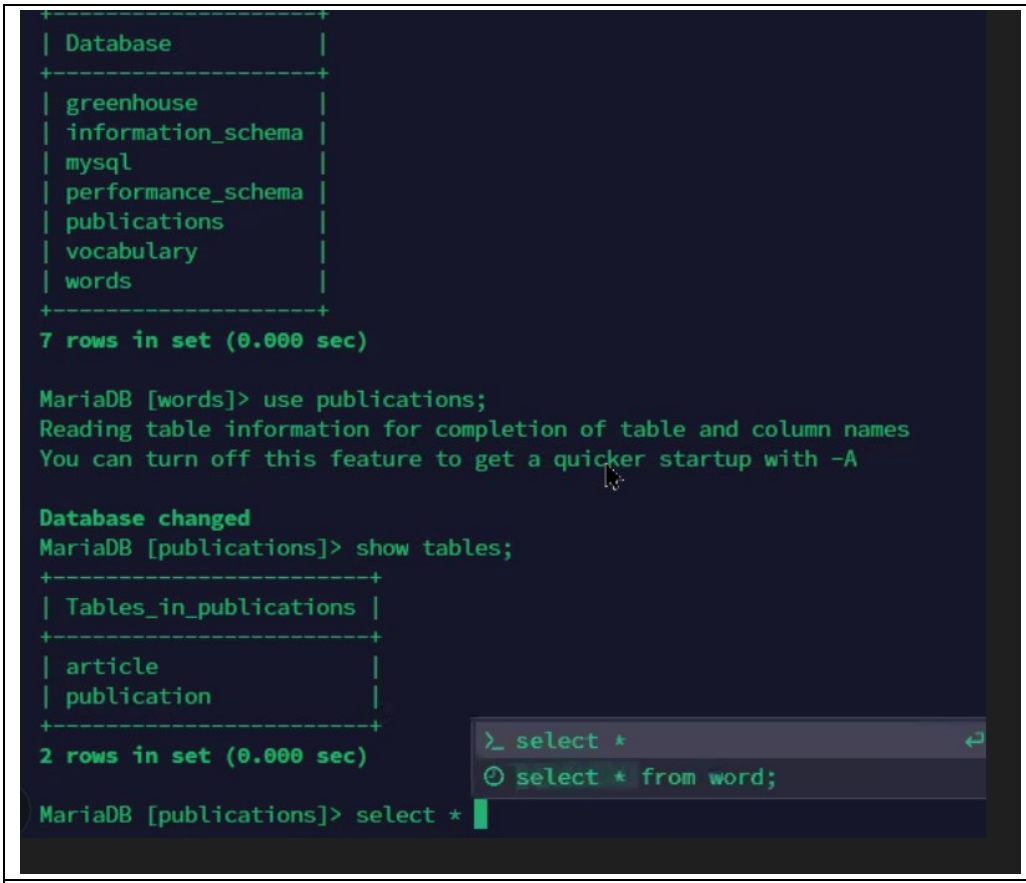


Figure 06: Access to cloud computing using Termitus

In this figure 06, we illustrate the process of accessing cloud computing resources using Termitus. Termitus is a cloud computing platform that provides a seamless and efficient way to manage and utilize your infrastructure.

```
MariaDB [(none)]> select * from publication WHERE author LIKE '%Kondor%';
ERROR 1046 (3D000): No database selected
MariaDB [(none)]> use publications;
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A

Database changed
MariaDB [publications]> select * from publication WHERE author LIKE '%Kondor%';
+-----+---------------+------------------+------+-------+---------------+
| id  | author        | publication_year | year | count | idpublication |
+-----+---------------+------------------+------+-------+---------------+
|   1 | Kondor Daniel | 2014             | 9    | 2     |             1 |
|   9 | Kondor Daniel | 2014             | 9    | 1     |             3 |
| 821 | Kondor Daniel | 2018             | 5    | 1     |           244 |
+-----+---------------+------------------+------+-------+---------------+
3 rows in set (0.006 sec)

MariaDB [publications]> select * from article limit 10;
+----+-----------------------------------------------------------------------------------------------------------+
| id | paper                                                                                                     |
+----+-----------------------------------------------------------------------------------------------------------+
|  1 | Do the Rich Get Richer? An Empirical Analysis of the Bitcoin Transaction Network                          |
|  2 | The digital traces of bubbles: feedback cycles between socio-economic signals in the Bitcoin economy     |
|  3 | Inferring the interplay between network structure and market effects in Bitcoin                           |
|  4 | Crypto-currency bubbles: an application of the Phillips-Shi-Yu (2013) methodology on Mt. Gox bitcoin prices |
|  5 | Bitcoins as an investment or speculative vehicle? A first look                                            |
|  6 | Bitcoin: Economics, Technology, and Governance                                                            |
|  7 | The economics of Bitcoin and similar private digital currencies                                           |
|  8 | What Are the Main Drivers of the Bitcoin Price? Evidence from Wavelet Coherence Analysis                  |
|  9 | The Predecessors of Bitcoin and Their Implications for the Prospect of Virtual Currencies                |
| 10 | Speculative bubbles in Bitcoin markets? An empirical investigation into the fundamental value of Bitcoin  |
+----+-----------------------------------------------------------------------------------------------------------+
10 rows in set (0.000 sec)
```

Figure 07: Visualizing Data on MariaDB for a specific author

In this figure 07, we demonstrate the process of visualizing data on MariaDB for a specific author. MariaDB is a robust and feature-rich open-source relational database management system, which provides powerful tools for data storage and retrieval.

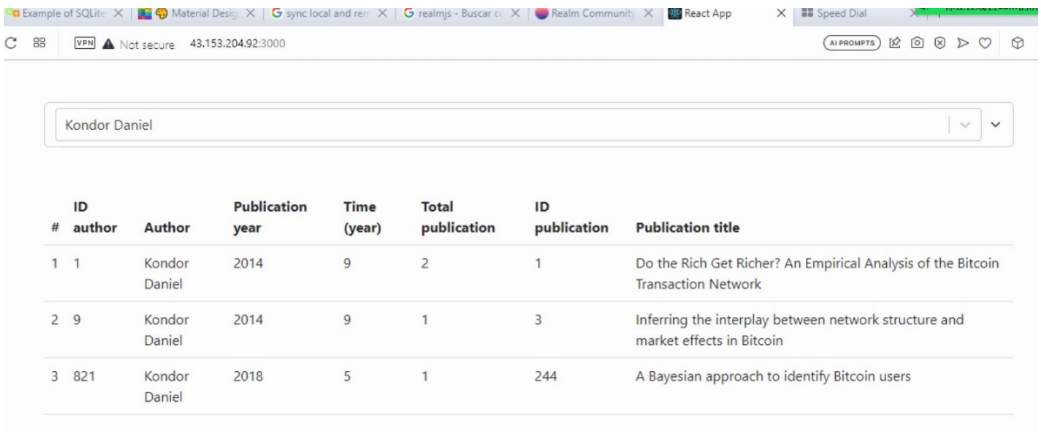


Figure 08: Display from a cloud computing server

In this figure 08, we illustrated the results of displaying data from a cloud computing server. Cloud computing offers the ability to store, process, and access data remotely, providing flexibility and scalability for various applications.

By leveraging cloud computing, organizations and individuals can benefit from the scalability, reliability, and accessibility of remote data storage and processing. The ability to display data from a cloud computing server opens up opportunities for real-time data analysis, collaborative decision-making, and seamless integration with other applications and systems.

4. Clustering literature data using LDA

4.1 Background

In this section, we use did clustering by using LDA. LDA is widely used in various natural language processing tasks, such as document clustering, text classification, and information retrieval. It provides a valuable framework for understanding and analyzing the latent structure of textual data. Latent Dirichlet Allocation (LDA) is a probabilistic generative model used for topic modeling. It estimates the conditional probability of observing a set of words (X) given a specific topic (Y). Symbolically, this can be represented as P(X | Y=y).

In LDA, topics are assumed to be latent variables, meaning they are not directly observed but inferred from the observed data. The model assumes a Dirichlet prior distribution over the topic-word probabilities and the document-topic probabilities. The conditional probability P(X | Y=y) in LDA represents the likelihood of observing a particular set of words (X) given a specific topic assignment (Y=y). It captures the probability distribution of words within topics, indicating the likelihood of certain words appearing in a particular topic.

4.2 Visualising the results

In order to visualize the results, the following steps has been done:

```
from gensim import corpora
import pyLDAvis.gensim_models as gensimvis
import pyLDAvis
```

```
[5]: nltk.download()
```

```
showing info https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/index.xml
```

```
Out[5]: True
```

```
[2]: # Read the data
df = pd.read_excel('papernode.xlsx')
documents = df['Abstract'].tolist()

# download stopwords and punktion
nltk.download('stopwords')
nltk.download('punkt')

stopwords_set = set(stopwords.words('english'))
punctuation_set = set(string.punctuation)

# Preprocessing and tokenization
preprocessed_documents = []
for doc in documents:
    # change all characters to lower
    doc = doc.lower()
    # tokenization
    tokens = word_tokenize(doc)
    # remove stopwords and punktion
    filtered_tokens = [token for token in tokens if token not in stopwords_set and token not in punctuation_set]
    # store the proprocessed data in preprocessed_documents
    preprocessed_documents.append(filtered_tokens)
```

```
[nltk_data] Error loading stopwords: <urlopen error [Errno 11004]
[nltk_data]     getaddrinfo failed>
[nltk_data] Error loading punkt: <urlopen error [Errno 11004]
[nltk_data]     getaddrinfo failed>
```

VPN

Figure 09. Coding for LDA

After executing these codes, here we have output shown in figure 10 below:

```
    # Print the subject heading and corresponding document number for each topic
for topic_id in range(num_topics):
        print(f"Topic {topic_id}:")
        topic_words = lda_model.show_topic(topic_id)
        for word, prob in topic_words:
            print(f"\t{word} ({prob})")
```

```
Topic 0:
        blockchain (0.019345667213201523)
        data (0.017462128773331642)
        proposed (0.015912124887108803)
        consensus (0.00983403343707323)
        system (0.008932783268392086)
        mechanism (0.007932063192129135)
        computing (0.007712614722549915)
        edge (0.0076234606094658375)
        based (0.007593065034598112)
        results (0.0071608396247029305)
Topic 1:
        news (0.010730053298175335)
        cryptocurrencies (0.009706526063382626)
        cryptocurrency (0.009571018628776073)
        blockchain (0.00879131630063057)
        fake (0.00843499694019556)
        content (0.007462773937731981)
        price (0.006533390376716852)
        bitcoin (0.006380567327141762)
        paper (0.006203243974596262)
        vicious (0.006009057629853487)
Topic 2:
        blockchain (0.013742570765316486)
        cryptocurrency (0.00967349112033844)
        contracts (0.008999839425086975)
        cryptocurrencies (0.00885239988565445)
        smart (0.007899200543761253)
        based (0.007104570511728525)
        btc (0.006242180708795786)
        digital (0.006103953812271357)
        system (0.006103404331952335)
        financial (0.005841681733727455)
```

Figure 10. Topics after passing all the abstract trough LDA

This figure 10 illustrates the results of applying Latent Dirichlet Allocation (LDA) to abstract column. Each abstract is represented as a mixture of topics, and the figure 10 shows the identified topics and their corresponding word distributions.

The figure 10 displays the top words associated with each topic. Each topic is represented by a cluster of related words, indicating the underlying theme or subject it represents. The purpose of Figure 10 is to provide an overview of the discovered topics in the abstracts, allowing researchers or readers to gain insights into the main themes present in the dataset.

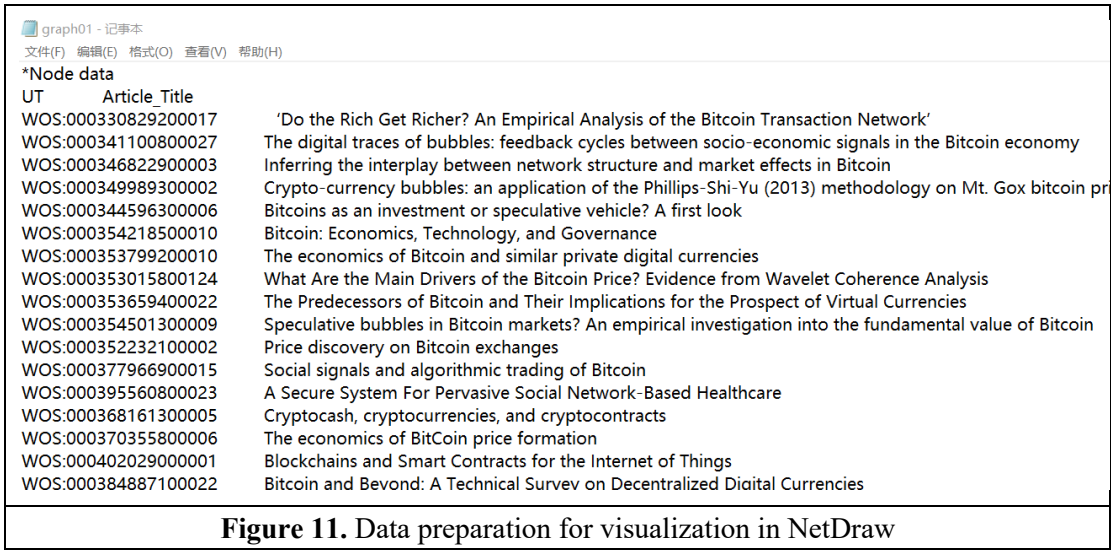
5. Citation networks for different topics using Netdraw

5.1 Background

Based on our knowledge, gained throughout the course, we employed NetDraw to create visual representations of social network data. NetDraw offers the ability to visualize multiple relations associated with the same nodes, allowing users to switch between and combine them. Furthermore, it supports the interpretation of various node attributes by providing options to set colors, sizes, rims, and labels based on these attributes.

5.2 Visualising the results

NetDraw requires a properly formatted dataset to create visualizations. In this figure, we can see the steps involved in preparing the node dataset. This may include extracting relevant information, organizing the data in a suitable format, and ensuring that the necessary attributes for nodes are included. By following these steps, the dataset is prepared and ready to be used in NetDraw to generate visual representations of the social network data. In Figure 11, the process of preparing the node dataset for NetDraw is illustrated:


graph01 - 记事本
文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)

*Node data

UT	Article_Title
WOS:000330829200017	'Do the Rich Get Richer? An Empirical Analysis of the Bitcoin Transaction Network'
WOS:000341100800027	The digital traces of bubbles: feedback cycles between socio-economic signals in the Bitcoin economy
WOS:000346822900003	Inferring the interplay between network structure and market effects in Bitcoin
WOS:000349989300002	Crypto-currency bubbles: an application of the Phillips-Shi-Yu (2013) methodology on Mt. Gox bitcoin pri
WOS:000344596300006	Bitcoins as an investment or speculative vehicle? A first look
WOS:000354218500010	Bitcoin: Economics, Technology, and Governance
WOS:000353799200010	The economics of Bitcoin and similar private digital currencies
WOS:000353015800124	What Are the Main Drivers of the Bitcoin Price? Evidence from Wavelet Coherence Analysis
WOS:000353659400022	The Predecessors of Bitcoin and Their Implications for the Prospect of Virtual Currencies
WOS:000354501300009	Speculative bubbles in Bitcoin markets? An empirical investigation into the fundamental value of Bitcoin
WOS:000352232100002	Price discovery on Bitcoin exchanges
WOS:000377966900015	Social signals and algorithmic trading of Bitcoin
WOS:000395560800023	A Secure System For Pervasive Social Network-Based Healthcare
WOS:000368161300005	Cryptocash, cryptocurrencies, and cryptocontracts
WOS:000370355800006	The economics of BitCoin price formation
WOS:000402029000001	Blockchains and Smart Contracts for the Internet of Things
WOS:000384887100022	Bitcoin and Beyond: A Technical Survey on Decentralized Digital Currencies

Figure 11. Data preparation for visualization in NetDraw

In Figure 12, the process of preparing the link dataset for NetDraw is depicted. NetDraw requires a well-structured dataset to visualize the connections between nodes.

```
WOS:000542571700036     Tushar Wayes;Saha Tapan Kumar;Yuen Chau;Smit
WOS:000542678200031     Aslan Aylin;Sensoy Ahmet    Intraday efficiency-fi
*Tie data
from      to
WOS:000752849700035     WOS:000498404700027
WOS:000752849700035     WOS:000353015800124
WOS:000752849700035     WOS:000548999400001
WOS:000752849700035     WOS:000545288600005
WOS:000758152200001     WOS:000504552000001
WOS:000745515800001     WOS:000703327100001
WOS:000745515800001     WOS:000541127800052
WOS:000745515800001     WOS:000566296200001
WOS:000745515800001     WOS:000582586400022
WOS:000745515800001     WOS:000464140500001
WOS:000745515800001     WOS:000618081600003
WOS:000745515800001     WOS:000612229800006
WOS:000745515800001     WOS:000497160500098
WOS:000745515800001     WOS:000562037900010
WOS:000745515800001     WOS:000640952900001
WOS:000745515800001     WOS:000582585800009
```

Figure 12. Preparing the link dataset for Netdraw

In Figure 13, the visualization on NetDraw is presented. NetDraw is a software tool that allows for the visualization of social network data. It provides various features and options to customize the visualization and enhance the clarity of the network representation.

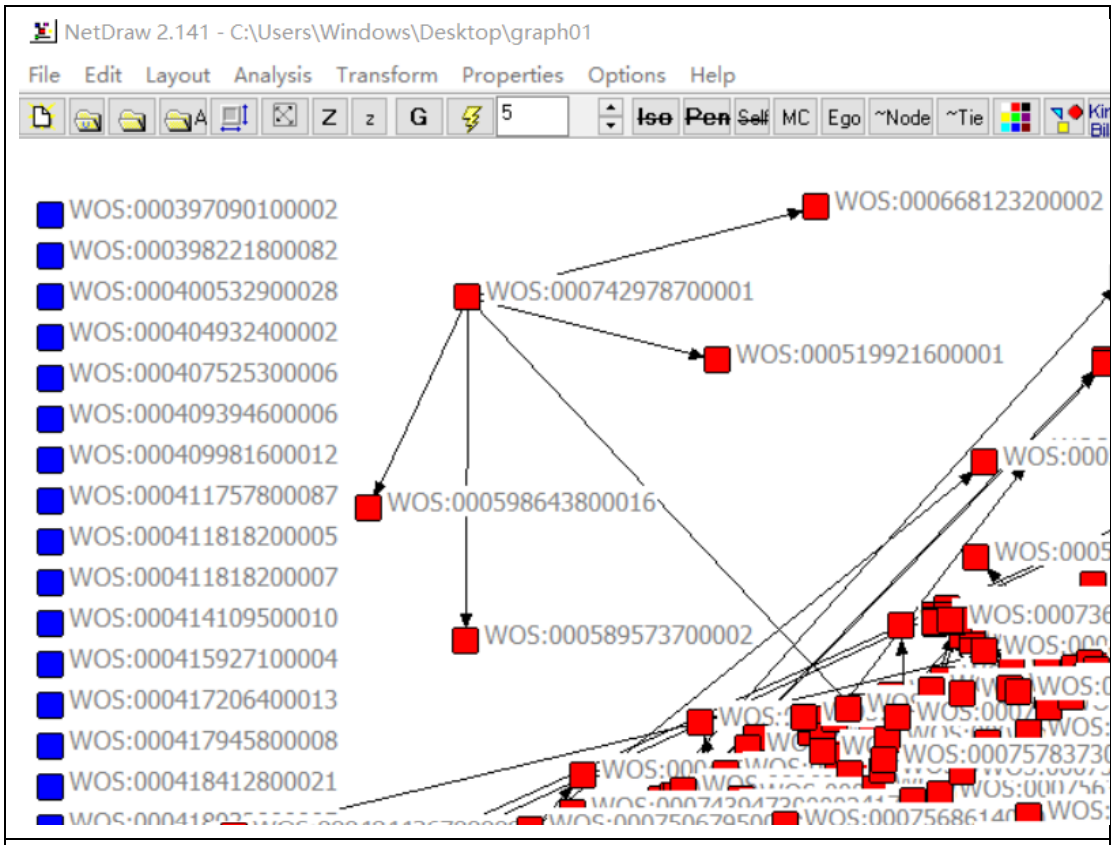


Figure 13. Visualisation on Netdraw

In this figure 13, we can see a visual representation of the network created using NetDraw. The nodes represent entities or individuals, and the links depict the connections or relationships

between them. The visualization helps to understand the structure and patterns within the network, enabling further analysis and insights into the data.

6. Network Analysis and Visualization by Using Neo4j

6.1 Background

In our current section, we have utilized Neo4j, a graph database management system (DBMS) that we recently studied in our course. Neo4j is specifically designed for creating, managing, and analyzing graphs. A graph database stores data in the form of nodes, edges, and properties, which accurately represent and maintain the relationships between various entities. With Neo4j, we can effectively store and navigate complex relationships, allowing for efficient querying and analysis of interconnected data. By leveraging the power of graph databases, we gain valuable insights into the underlying relationships and patterns within our dataset.

6.2 Visualizing the relationship

In order to visualize the relationship between authors and articles in our dataset using Neo4j, we followed these steps:

Step 1: First, we loaded the CSV files containing the node data for "papernode" and "authorPaper". We created nodes of type "papernode" and "authorPaper" and set their properties using the data from the CSV files.

```
LOAD CSV WITH HEADERS FROM "file:///papernode.csv" AS row
CREATE (n:papernode)
SET n = row

LOAD CSV WITH HEADERS FROM "file:///authorPaper.csv" AS row
CREATE (n:authorPaper)
SET n = row
```

Step 2: We loade the CSV file for "authorPaper" one more time and match the "AUTHOR" and "Article" nodes based on their names. We then created a "WROTE" relationship between the matched nodes using the MERGE clause.

```
LOAD CSV WITH HEADERS FROM "file:///authorPaper.csv" AS row
MATCH (a:AUTHOR {name: row.AUTHOR})
MATCH (p:Article {name: row.Article})
MERGE (a)-[:WROTE]->(p)
```

By executing this modified query, we created a relationship between the "Author" nodes and the "Article" nodes based on the "Author_Full_Names" and "Article_Title" columns. The

MATCH clauses match the corresponding nodes, and the **MERGE** clause creates a "WROTE" relationship between them.

Step 3: To visualize the relationship network between authors and articles, we executed the following query:

```
MATCH (a:AUTHOR)-[:WROTE]->(p:Article)
RETURN a, p
```

This query matches the "AUTHOR" nodes that have a "WROTE" relationship with the "Article" nodes. It then returns the matched nodes, which include the authors and the articles they wrote.

Here in Figure 14, we can see the output:

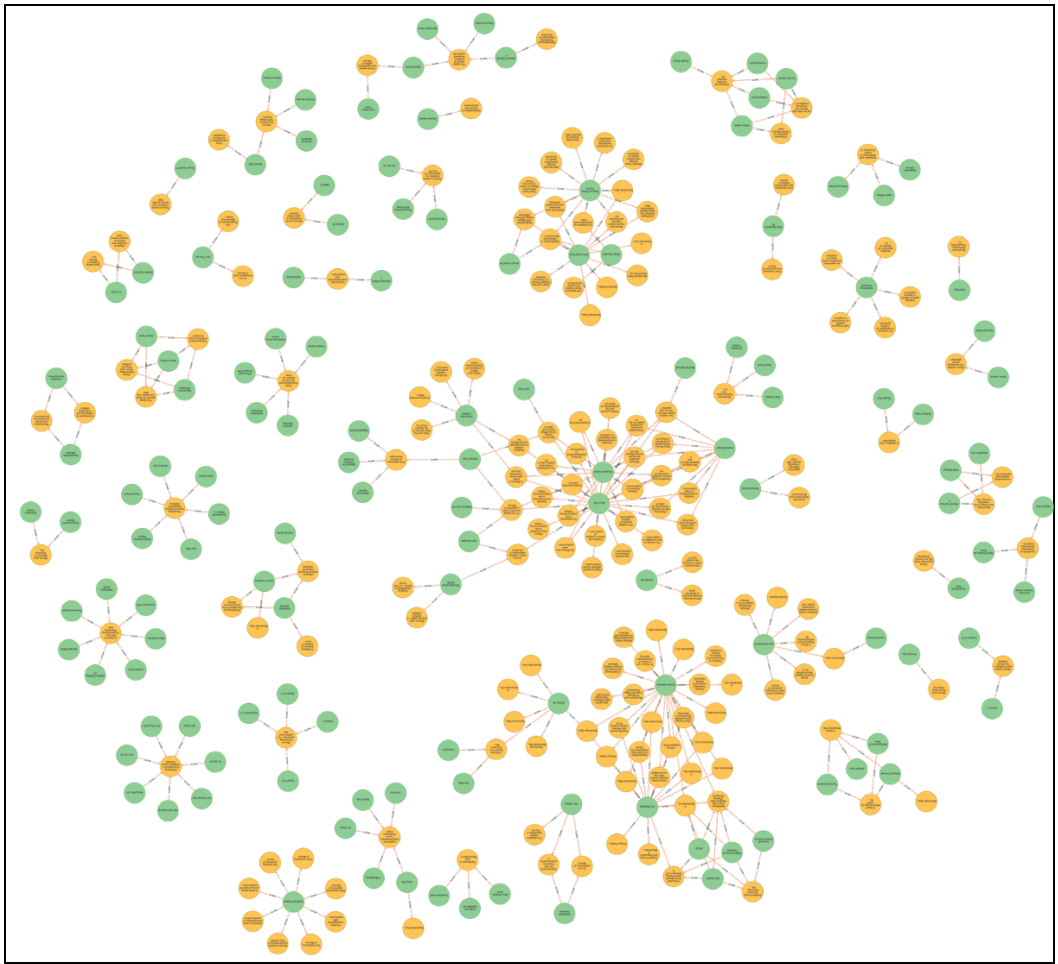


Figure 14. Visualisation on Articles and Authors relationship

Figure 14 represents the visualization of the relationship between articles and authors in the Neo4j graph database. The graph visualization displays nodes representing authors and articles, connected by directed edges. The nodes labeled "AUTHOR" represent authors, and the nodes labeled "Article" represent articles.

The relationships between authors and articles are depicted by the directed edges labeled "WROTE." Each edge connects an author node to an article node, indicating that the author wrote the corresponding article.

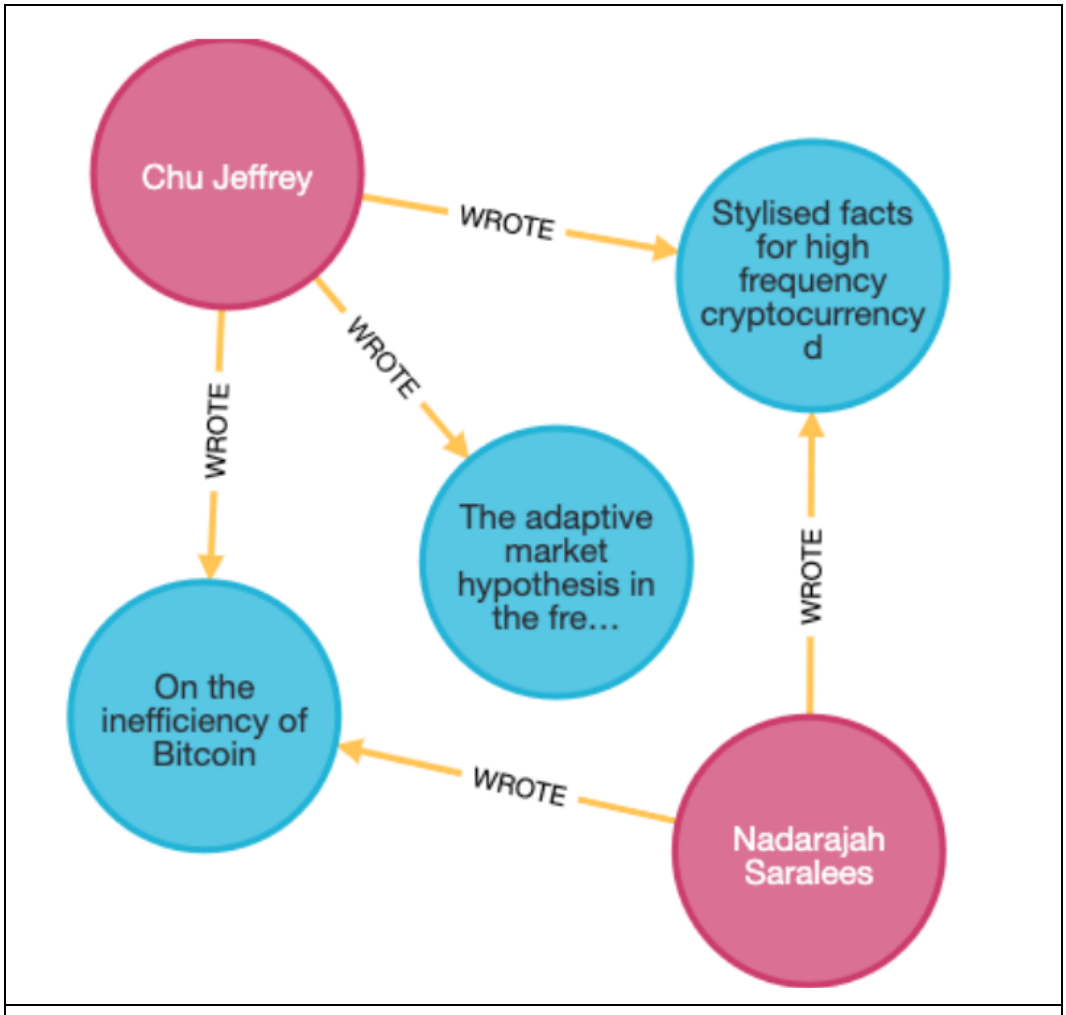


Figure 15. Zooming the relationship

Here in figures 15, 16, 17 u can see closer by zooming it, and we confirmed that the relationship network has been created accurately:

30	WOS:000398221800082	Ouaddah Aafaf;Abou Elkalam Anas;Ouahman Abdellah	FairAccess: a new Blockchain-base
31	WOS:000392568300003	Nadarajah Saralees;Chu Jeffrey	On the inefficiency of Bitcoin
32	WOS:000411757800087	Xia Qi;Sifah Emmanuel Boateng;Asamoah Kwame Omo	MeDShare: Trust-Less Medical Dat

Figure 16. Contrasting the relationship

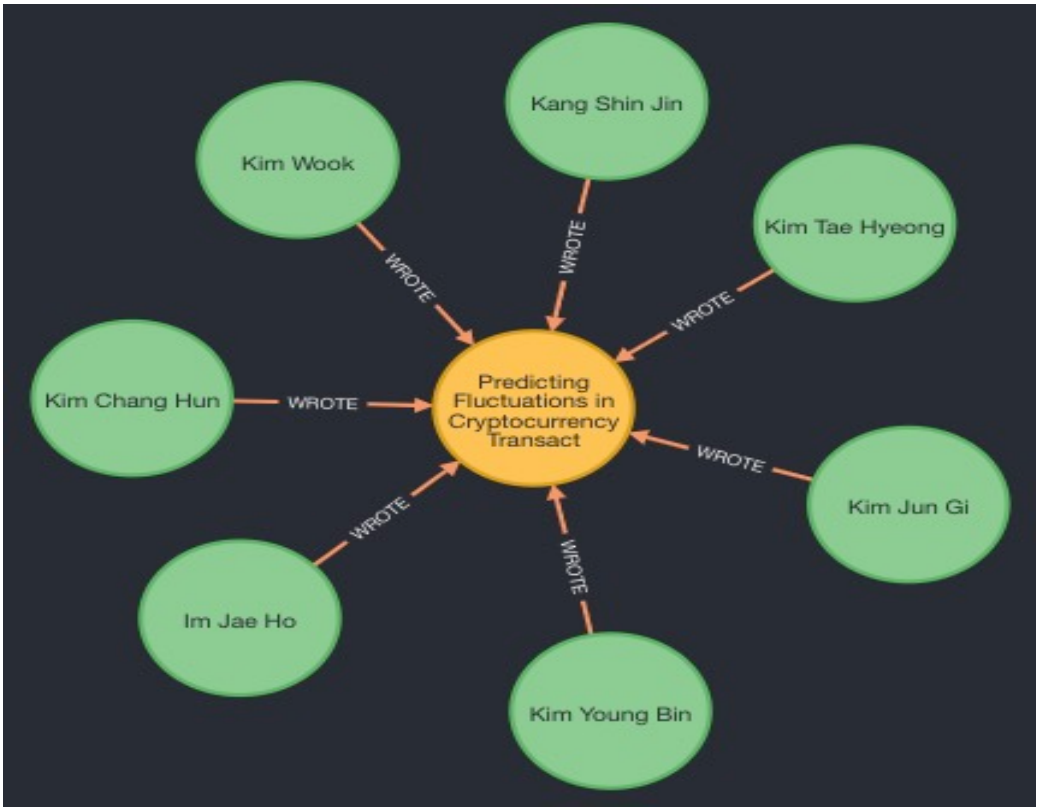


Figure 17. Zoomed visualization

Here, we have done some more visualization on the relationship between Citing and cited columns of citations dataset. The results of the visualization are shown in the following Figures below:

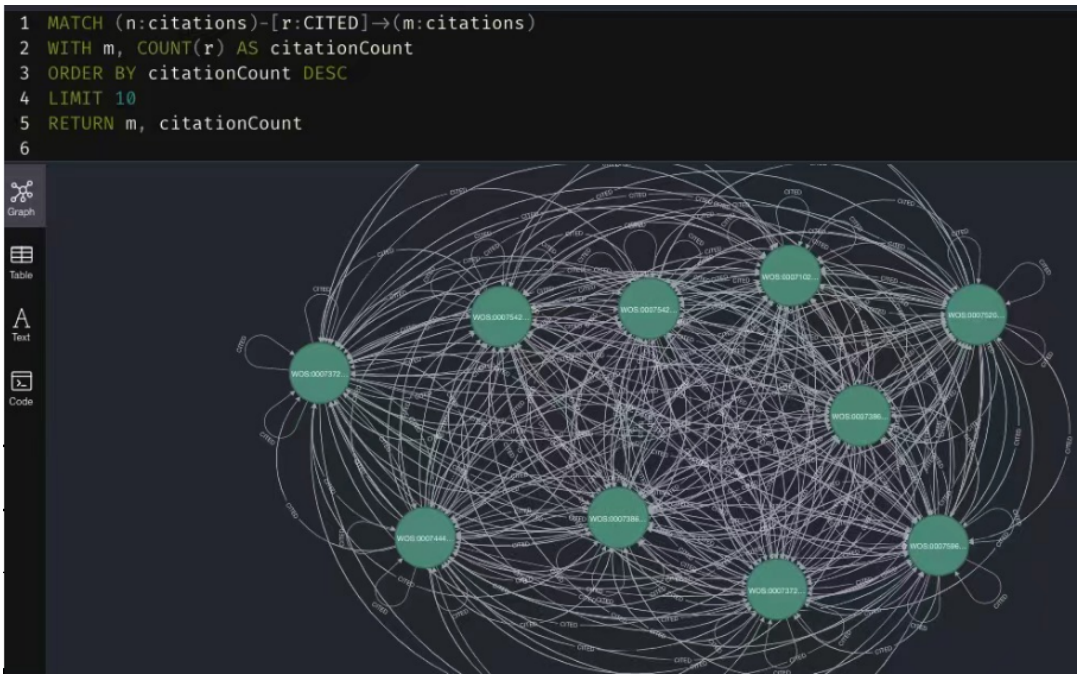


Figure 18. The most cited top ten articles

Then, we have created visualization for the least cited nodes along their citation counts by running these codes:

```
MATCH (n:citations)-[r:CITED]->(m:citations)
WITH m, COUNT(r) AS citationCount
ORDER BY citationCount ASC
LIMIT 20
RETURN m, citationCount
```

Output:

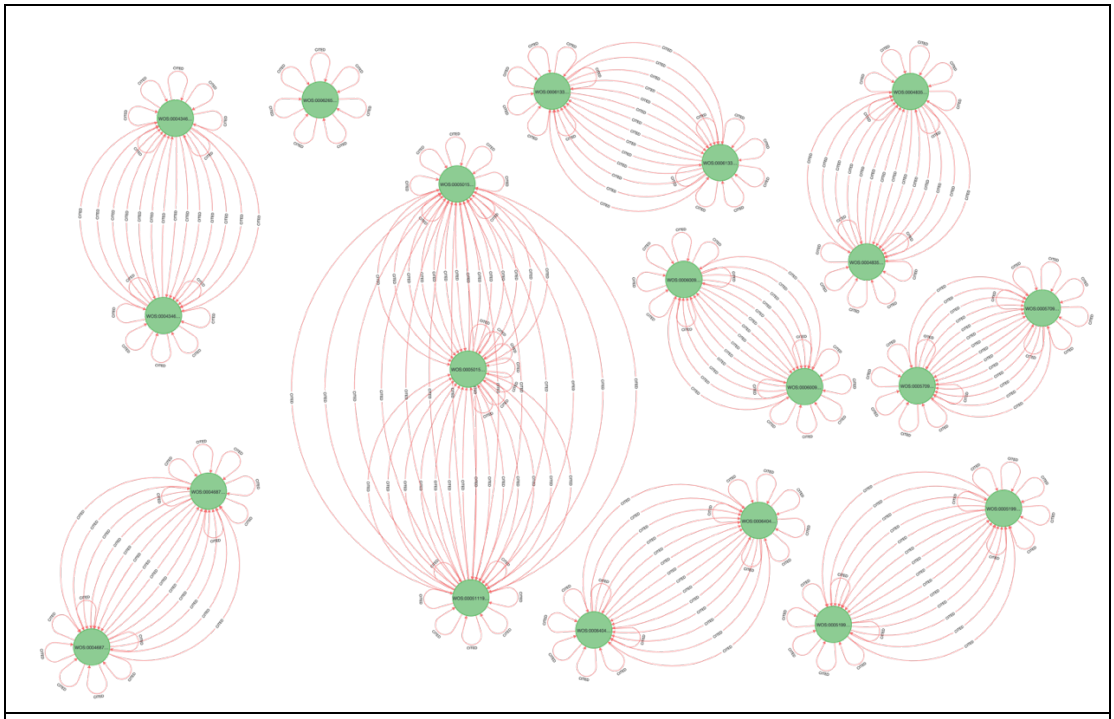


Figure 19. The least cited top 20 nodes.

In the Figure 19, We can see to the 20 nodes (Citation ID) in a graph database that have the lowest number of citations or references associated with them. Citations or references are typically used to indicate the influence or importance of a particular node within the network. Nodes with a high number of citations are often considered more significant or influential within the network, while nodes with a low number of citations may have relatively less impact or recognition. By analyzing the least cited nodes, researchers or analysts can gain insights into the less prominent aspects of the network and potentially uncover hidden patterns or connections that were previously overlooked.

7. Further Analysis

In the next step of the analysis, we used the R programming language to further analyze the data and create visualizations. We started by performing data manipulation and aggregation using the dplyr package in R.

First, we used the ungroup() function on the publications_df data frame to remove any existing grouping. Then, we grouped the data by the publication year (variable: Publication_Year) using the group_by() function.

Next, we used the summarize() function to calculate the total number of publications for each year. We used the sum(publications) syntax within the summarize() function to calculate the sum of the publications for each year. The resulting data frame, named plotted_pub, contains the publication year and the corresponding total number of publications, shown below in table 1.

	Publication_Year	total_publications
1	2014	27
2	2015	32
3	2016	96
4	2017	517
5	2018	1875
6	2019	4978
7	2020	8092
8	2021	6222
9	2022	2800
10	NA	343

Table 1. Publication per year.

Based on the Table 1, we can observe that the number of publications has been steadily increasing from 2014 to 2020, with 2020 having the highest total publications. However, after 2020, there is a decline in the number of publications until 2022. These insights provide an overview of the publication trends over the analyzed period.

To visualize the relationship between the publication year and the total publications, we utilized the ggplot2 library. We created a bar graph shown in Figure 20 below:

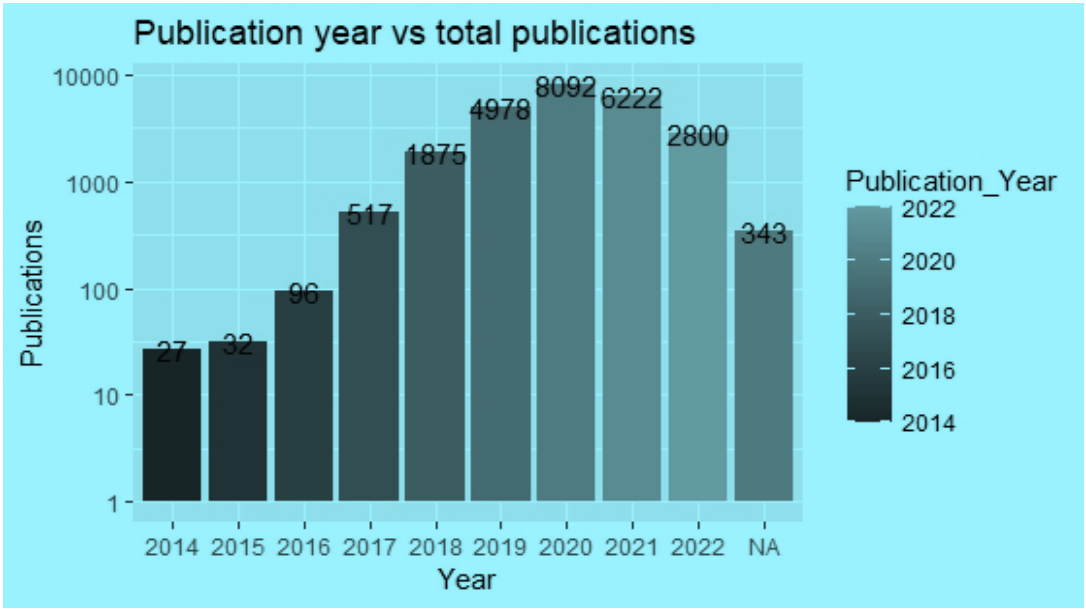


Figure 20. Visualization of the total publications by year.

The figure 20 represents the publication year on the x-axis and the total publications on the y-axis. Each bar represents a specific year, and its height corresponds to the total number of publications for that year.

According to the Figure 20, 2020 had the highest number of publications, with exactly 8092 publications. On the other hand, 2014 had the lowest number of publications, with only 27 publications. It's worth noting that there is a category labeled "NA" in the graph, which indicates that the year for some publications is unknown or missing in the dataset. These publications could not be assigned to a specific year.

8. Classification

We classified the articles into approximately 35 categories by reading their titles and pasting them into another sheet:

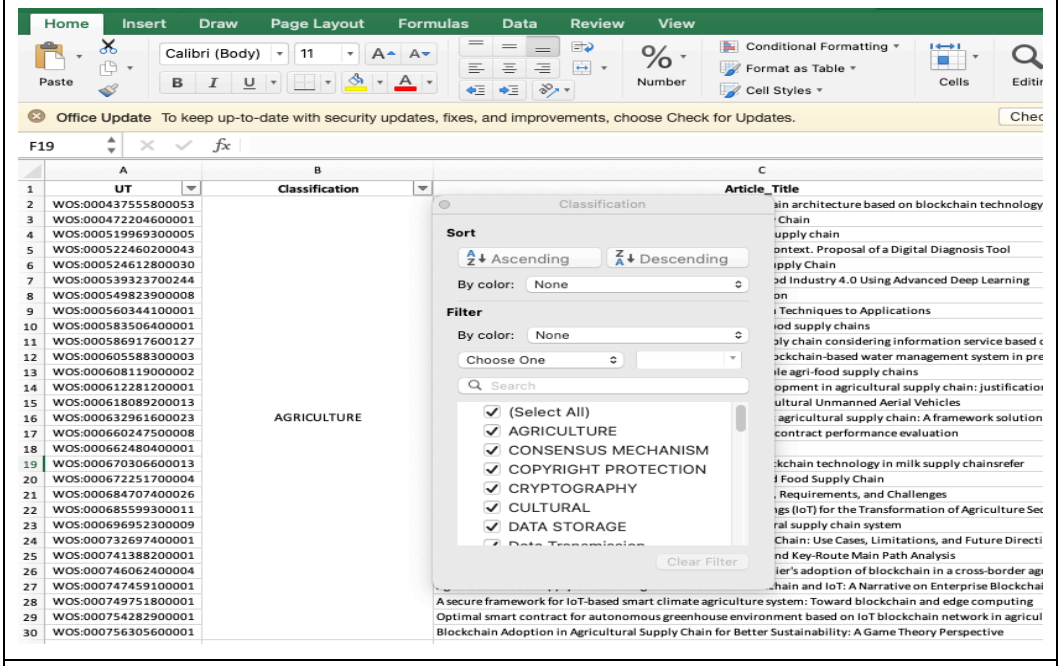


Figure 21. Classification results

9. Access to the cloud

This screenshot shows the implementation on LINUX operating system of our learning. We purchased cloud computing infrastructure as a service to store both the website and the database. This experience has prepared us for any project, whether it be academic or industry-related

```
/* // VISUAL CODE, CREATE FOLDER, COMMAND NPM INIT,  INSTAll
libraries:  NPM INSTALL EXPRESS  (node package manager)
//go to the folder, and type cmd then type code .
// install    libraries:   express (create server port 8089), axios
(post, get data), boostrap (style), sweet-alert ( jod job), react
（framework front end）, react-select (drop down box)
//////   type in the URL  http://localhost:8081/api/authors?name=lu
//  in clien terminal  npm   start
// http://localhost:8081/api/authors?name=Kondor
//http://localhost:8081/api/author?name=Kondor%20Daniel
// on the client folder, do cmd, then wriite code .   to start the
client
T3nc3ntcl@ud
43.153.204.92
T3nc3ntcl@ud
*/
```

10. Conclusion

This project focused on analyzing and visualizing a dataset of scholarly publications on blockchain technology. It involved data preprocessing, network analysis, and visualization using multiple tools and programming languages. The dataset was processed to extract information such as author names, publication years, and citation data. Network analysis techniques revealed collaboration patterns and citation networks among authors.

Visualizations were key in presenting the results. Graphs and charts showcased publication trends over time, influential authors, and relationships between authors and their works. Tools like R, SQL, Excel, Termitus, and Neo4j were used for data manipulation, analysis, and visualization. We have applied all the knowledge we have learned during the course as required.

Reference: Required Task for the Project.

The teacher provided a collection of SCI journal papers on blockchain and their citation data. Based on this data, students organized information about research titles, abstracts, keywords, authors, citation relationships, publication dates, journals, and categories to analyze the current distribution of blockchain research topics, future development trends, collaboration among scholars, highly cited papers and authors, and corresponding tags.

The main tasks may include:

- organizing literature data in **Excel or a databas**e,
- classifying literature data based on existing categorizations and designing tags, clustering literature **data using LDA**,
- presenting literature da**ta through a webpage** using Flask+MySQL,
- conducting network analysis and visualization of cooperation and citation networks for different topics **using Netdraw or G**ephi.
- The data can be further processed, such as separating authors listed in one column, extracting author addresses, and incorporating external information such as author profiles and journal impact factors. Keyword extraction, segmentation, clustering, and statistical analysis can reveal trends in the evolution of topics over time.
- ***This paper aims to assess students' understanding of information organization methods and techniques through practical operations in a specific literature field.***

- ***There is no strict requirement to follow the above suggestions and tips, and students can use their own information organization methods and analytical perspectives.***