CS343: Graph Data Science
Homework 02

Ali Muhammad Asad

1 Data Model

Based on the given data, and the provided sample questions, the data model designed is as follows:

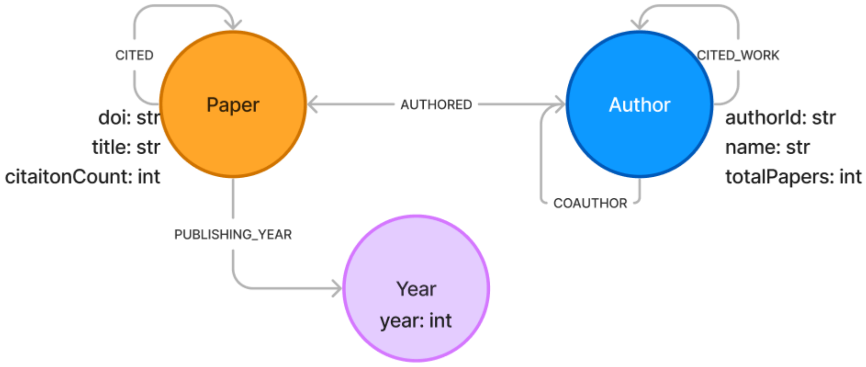


Figure 1: Data Model

The rationale behind the data model is such that each author will also have cited some author's work, which can potentially be used to answer a query that which author had been cited the most which means they were influential. The `totalPapers` attribute can be used to answer which author has published the most or fewest number of papers. Further, the `citationCount` attribute with each paper can also be used to answer the query about which paper had the most or fewest number of citations, or which paper was the most influential based on the number of citations. An additional node and relationship for the year has been added to the data model, where a paper is published in a given year which can potentially answer questions related to the years in which the most papers were published, or the years in which the most citations were made, or the average number of citations per year, etc. The co-author relationship can be used to answer questions related to which authors have collaborated the most, or which authors have collaborated the least, etc. The cited work relationship can be used to answer questions related to which authors have cited each other's work the most, or which authors have cited each other's work the least, etc.

2 Graph Loading Queries

We used the following query to load the data into the graph database, and build the necessary relationships between the nodes (loading the query takes a lot of time though and although alot of the data was loaded, not all of it was able to load due to errors in Neo4J due to which certain batches were crashed. Due to time limitations, a smaller batch size wasn’t feasible). We also use the APOC library to iterate over the CSV files and load the data in batches to speed up the process. The query is as follows:

```
CALL apoc.periodic.iterate(
    "LOAD CSV WITH HEADERS FROM ’file:///papers.csv’ AS row RETURN row",
    "MERGE (p:Paper {doi: row.doi, title: row.title, citationCount: 0})
        MERGE (y:Year {year: coalesce(row.year, ’Unknown’)})
        MERGE (p)-[:PUBLISHED_IN_YEAR]->(y)",
    {batchSize: 10000, iterateList: true, parallel: false})
```

Listing 1: Loading Papers

```
CALL apoc.periodic.iterate(
    "LOAD CSV WITH HEADERS FROM ’file:///references.csv’ AS row RETURN row",
    "MATCH (p1:Paper {doi: row.paper_doi}), (p2:Paper {doi: row.reference_doi})
        MERGE (p1)-[:CITED]->(p2)
        SET p2.citationCount = p2.citationCount + 1",
    {batchSize: 10000, iterateList: true, parallel: false})
```

Listing 2: Citation Relationships

```
CALL apoc.periodic.iterate(
  "LOAD CSV WITH HEADERS FROM ’file:///authors.csv’ AS row RETURN row",
  "MERGE (a:Author {authorId: row.authorId, name: row.name, totalPapers: 0})
   WITH a, row
   MATCH (p:Paper {doi: row.doi})
   MERGE (a)-[:AUTHORED]->(p)
   SET a.totalPapers = a.totalPapers + 1",
  {batchSize: 10000, iterateList: true, parallel: false})
```

Listing 3: Loading Authors and Authorship Relationships

```
CALL apoc.periodic.iterate(
    "MATCH (a1:Author)-[:AUTHORED]->(:Paper)<-[:AUTHORED]-(a2:Author) WHERE id(a1) < id(a2) RETURN a1, a2",
    "MERGE (a1)-[:COAUTHOR]->(a2)",
    {batchSize: 10000, iterateList: true, parallel: false})

CALL apoc.periodic.iterate(
    "MATCH (a1:Author)-[:AUTHORED]->(:Paper)<-[:CITED]-(:Paper)<-[:AUTHORED]-(a2:Author) WHERE id(a1) < id(a2) RETURN a1, a2",
    "MERGE (a1)-[:CITED_WORK]->(a2)",
    {batchSize: 10000, iterateList: true, parallel: true})
```

Listing 4: Coauthorship and Cited Work Relationships

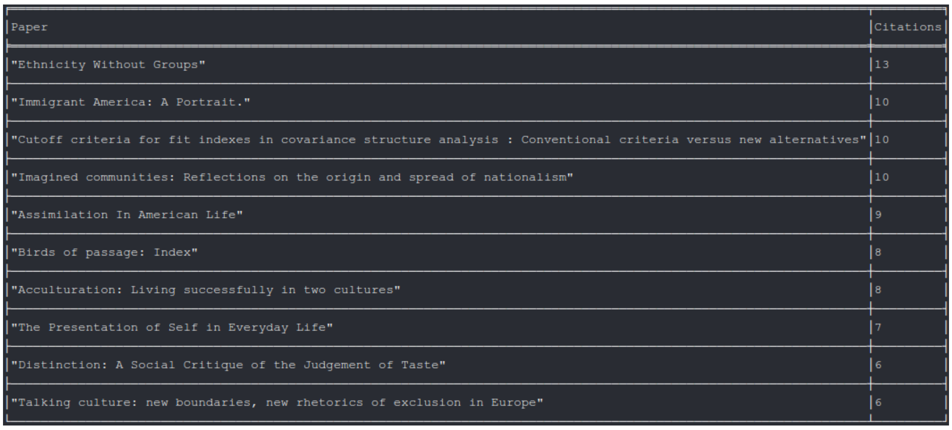
3 Graph Analytic Queries, Analysis, and Interpretation of Results

Based on the above data model, we can answer the following questions using the graph database:

What are the most influential papers in the dataset?

```
MATCH (p:Paper)
RETURN p.title AS Paper, p.citationCount AS Citations
ORDER BY Citations DESC LIMIT 10
```

Listing 5: Most Influential Papers



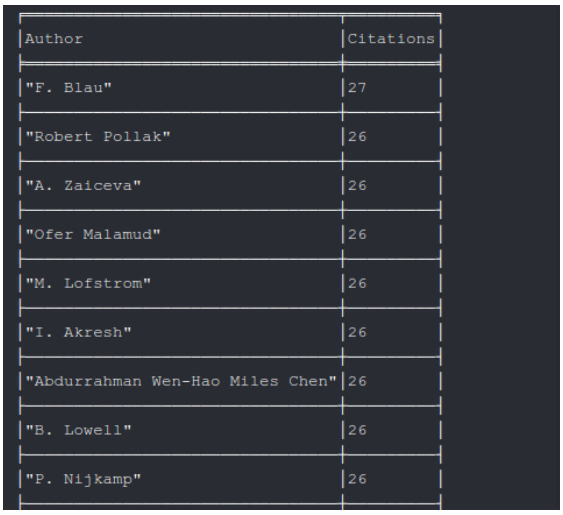
Paper	Citations
"Ethnicity Without Groups"	13
"Immigrant America: A Portrait."	10
"Cutoff criteria for fit indexes in covariance structure analysis : Conventional criteria versus new alternatives"	10
"Imagined communities: Reflections on the origin and spread of nationalism"	10
"Assimilation In American Life"	9
"Birds of passage: Index"	8
"Acculturation: Living successfully in two cultures"	8
"The Presentation of Self in Everyday Life"	7
"Distinction: A Social Critique of the Judgement of Taste"	6
"Talking culture: new boundaries, new rhetorics of exclusion in Europe"	6

The above query returns the top 10 most influential papers in the dataset based on the number of citations they have received.

What are the most influential authors in the dataset?

```
MATCH (a:Author)-[:AUTHORED]->(:Paper)<-[:CITED]-(:Paper)
RETURN a.name AS Author, count(*) AS Citations
ORDER BY Citations DESC LIMIT 10
```

Listing 6: Most Influential Authors



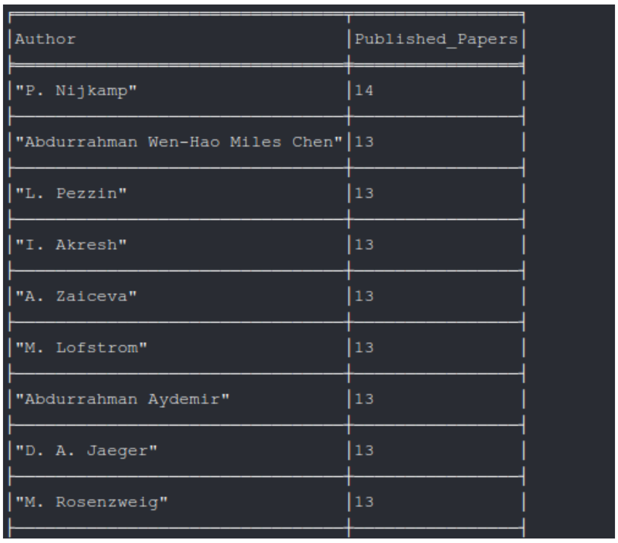
```
╒════════════════════════════════╤═════════╕
│Author                          │Citations│
╞════════════════════════════════╪═════════╡
│"F. Blau"                       │27       │
├────────────────────────────────┼─────────┤
│"Robert Pollak"                 │26       │
├────────────────────────────────┼─────────┤
│"A. Zaiceva"                    │26       │
├────────────────────────────────┼─────────┤
│"Ofer Malamud"                  │26       │
├────────────────────────────────┼─────────┤
│"M. Lofstrom"                   │26       │
├────────────────────────────────┼─────────┤
│"I. Akresh"                     │26       │
├────────────────────────────────┼─────────┤
│"Abdurrahman Wen-Hao Miles Chen"│26       │
├────────────────────────────────┼─────────┤
│"B. Lowell"                     │26       │
├────────────────────────────────┼─────────┤
│"P. Nijkamp"                    │26       │
├────────────────────────────────┼─────────┤
```

The above query works by finding the authors who have authored papers that have been cited the most number of times, and returns the top 10 most influential authors in the dataset. This is done since the number of citations a paper has received is also a good measure of the influence of the author apart from the number of papers they have published, as the number of citations shows how much their work has been referred to by other researchers, while the number of papers shows how much work they have done, and might not be that influential.

Who are the top authors based on the number of papers they have published?

```
MATCH (a:Author)
RETURN a.name AS Author, a.totalPapers AS Published_Papers
ORDER BY Published_Papers DESC LIMIT 10
```

Listing 7: Top Authors Based on Number of Papers

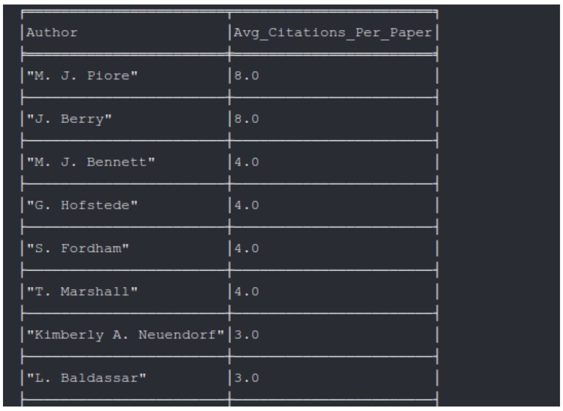


```
╒════════════════════════════════╤════════════════╕
│Author                          │Published_Papers│
╞════════════════════════════════╪════════════════╡
│"P. Nijkamp"                    │14              │
├────────────────────────────────┼────────────────┤
│"Abdurrahman Wen-Hao Miles Chen"│13              │
├────────────────────────────────┼────────────────┤
│"L. Pezzin"                     │13              │
├────────────────────────────────┼────────────────┤
│"I. Akresh"                     │13              │
├────────────────────────────────┼────────────────┤
│"A. Zaiceva"                    │13              │
├────────────────────────────────┼────────────────┤
│"M. Lofstrom"                   │13              │
├────────────────────────────────┼────────────────┤
│"Abdurrahman Aydemir"           │13              │
├────────────────────────────────┼────────────────┤
│"D. A. Jaeger"                  │13              │
├────────────────────────────────┼────────────────┤
│"M. Rosenzweig"                 │13              │
├────────────────────────────────┼────────────────┤
```

Which authors have the highest average number of citations per paper?

```
MATCH (a:Author)-[:AUTHORED]->(p:Paper)
WITH a, avg(p.citationCount) AS avgCitations
RETURN a.name AS Author, avgCitations AS Avg_Citations_Per_Paper
ORDER BY Avg_Citations_Per_Paper DESC LIMIT 10
```

Listing 8: Authors with Highest Average Citations per Paper



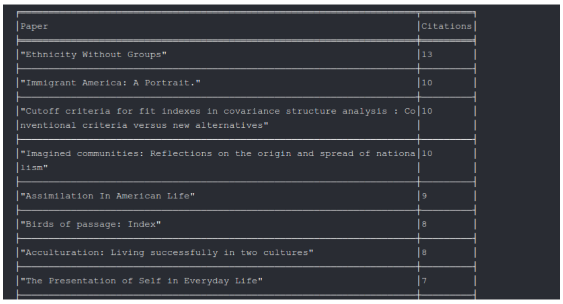
```
╒══════════════════════╤═══════════════════════╕
│Author                │Avg_Citations_Per_Paper│
╞══════════════════════╪═══════════════════════╡
│"M. J. Piore"         │8.0                    │
├──────────────────────┼───────────────────────┤
│"J. Berry"            │8.0                    │
├──────────────────────┼───────────────────────┤
│"M. J. Bennett"       │4.0                    │
├──────────────────────┼───────────────────────┤
│"G. Hofstede"         │4.0                    │
├──────────────────────┼───────────────────────┤
│"S. Fordham"          │4.0                    │
├──────────────────────┼───────────────────────┤
│"T. Marshall"         │4.0                    │
├──────────────────────┼───────────────────────┤
│"Kimberly A. Neuendorf"│3.0                   │
├──────────────────────┼───────────────────────┤
│"L. Baldassar"        │3.0                    │
├──────────────────────┼───────────────────────┤
```

The above query works by finding the average number of citations per paper for each author, and then returns the top 10 authors with the highest average number of citations per paper. This is a good measure of the quality of the work of the author, as it shows how much their work has been cited on average, which is a good measure of the quality of the work.

What are the most cited papers in the dataset?

```
MATCH (p:Paper)
RETURN p.title AS Paper, p.citationCount AS Citations
ORDER BY Citations DESC LIMIT 10
```

Listing 9: Most Cited Papers



```
╒════════════════════════════════════════════════════════════════════════════╤═════════╕
│Paper                                                                       │Citations│
╞════════════════════════════════════════════════════════════════════════════╪═════════╡
│"Ethnicity Without Groups"                                                  │13       │
├────────────────────────────────────────────────────────────────────────────┼─────────┤
│"Immigrant America: A Portrait."                                            │10       │
├────────────────────────────────────────────────────────────────────────────┼─────────┤
│"Cutoff criteria for fit indexes in covariance structure analysis : Co│10       │
│nventional criteria versus new alternatives"                                │         │
├────────────────────────────────────────────────────────────────────────────┼─────────┤
│"Imagined communities: Reflections on the origin and spread of nationa│10       │
│lism"                                                                       │         │
├────────────────────────────────────────────────────────────────────────────┼─────────┤
│"Assimilation In American Life"                                             │9        │
├────────────────────────────────────────────────────────────────────────────┼─────────┤
│"Birds of passage: Index"                                                   │8        │
├────────────────────────────────────────────────────────────────────────────┼─────────┤
│"Acculturation: Living successfully in two cultures"                        │8        │
├────────────────────────────────────────────────────────────────────────────┼─────────┤
│"The Presentation of Self in Everyday Life"                                 │7        │
├────────────────────────────────────────────────────────────────────────────┼─────────┤
```

The above query returns the top 10 most cited papers in the dataset based on the number of citations they have received. This is a good measure of the quality of the paper, as it shows how much the paper has been cited by other researchers.

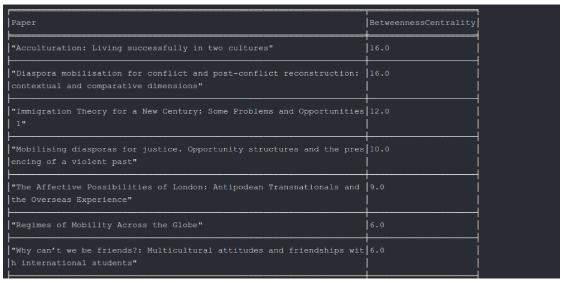
Are there any papers/authors that act as bridges between different clusters in the network (high betweenness centrality)?

```
CALL gds.graph.project( 'papersGraph', 'Paper', 'CITED')

CALL gds.betweenness.stream('papersGraph') YIELD nodeId, score
RETURN gds.util.asNode(nodeId).title AS Paper, score AS BetweennessCentrality
ORDER BY BetweennessCentrality DESC LIMIT 10

CALL gds.graph.drop('papersGraph')
```

Listing 10: Authors with Highest Betweenness Centrality



```
|Paper                                                                |BetweennessCentrality|
|"Acculturation: Living successfully in two cultures"                 |16.0                 |
|"Diaspora mobilisation for conflict and post-conflict reconstruction: |16.0                 |
|contextual and comparative dimensions"                                |                     |
|"Immigration Theory for a New Century: Some Problems and Opportunities|12.0                 |
| 1"                                                                   |                     |
|"Mobilising diasporas for justice. Opportunity structures and the pres|10.0                 |
|encing of a violent past"                                             |                     |
|"The Affective Possibilities of London: Antipodean Transnationals and |9.0                  |
|the Overseas Experience"                                              |                     |
|"Regimes of Mobility Across the Globe"                                |6.0                  |
|"Why can't we be friends?: Multicultural attitudes and friendships wit|6.0                  |
|h international students"                                             |                     |
```

The above query first creates a projection of the graph on the papers and the cited relationships, and then calculates the betweenness centrality of each paper in the graph, and returns the top 10 papers with the highest betweenness centrality. This is a good measure of the importance of the paper in connecting different clusters in the network. The high betweenness centrality score basically tells that a node appears on many shortest paths, and therefore, can be interpreted that it is likely serving as a bridge between different clusters in the network.

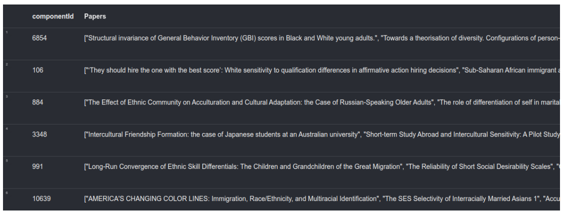
Do we have disconnected communities in the network? If yes, what are they?

```
CALL gds.graph.project( 'papersGraph', 'Paper', 'CITED')

CALL gds.wcc.stream('papersGraph') YIELD nodeId, componentId
RETURN componentId, collect(gds.util.asNode(nodeId).title) AS Papers
ORDER BY size(Papers) DESC

CALL gds.graph.drop('papersGraph')
```

Listing 11: Disconnected Communities



	componentId	Papers
1	6854	["Structural invariance of General Behavior Inventory (GBI) scores in Black and White young adults.", "Towards a theorisation of diversity. Configurations of person-
2	106	["'They should hire the one with the best score': White sensitivity to qualification differences in affirmative action hiring decisions", "Sub-Saharan African immigrant a
3	884	["The Effect of Ethnic Community on Acculturation and Cultural Adaptation: the Case of Russian-Speaking Older Adults", "The role of differentiation of self in marital
4	3348	["Intercultural Friendship Formation: the case of Japanese students at an Australian university", "Short-term Study Abroad and Intercultural Sensitivity: A Pilot Study
5	991	["Long-Run Convergence of Ethnic Skill Differentials: The Children and Grandchildren of the Great Migration", "The Reliability of Short Social Desirability Scales", "
6	10639	["AMERICA'S CHANGING COLOR LINES: Immigration, Race/Ethnicity, and Multiracial Identification", "The SES Selectivity of Interracially Married Asians 1", "Accu

The above query uses the `wcc.stream` algorithm which is basically a function that finds sets of connected nodes in the graph, where all nodes form some sort of a connected component and can reach each other. In our context, this can represent a cluster of papers that cite each other but don't cite papers outside of the cluster. This can indicate that there are disconnected communities in the network.

Do we have citation cycles?

```
MATCH cycle=(p:Paper)-[:CITED*]->(p)
RETURN cycle LIMIT 10
```

Listing 12: Citation Cycles

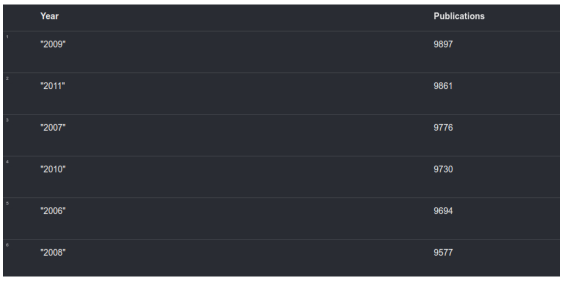
The above query did not return any results, which basically means that for the loaded data, there are no citation cycles in the network. This is a good thing as citation cycles can indicate that there is some sort of a problem with the data, as a paper cannot cite itself, and if there are citation cycles, it can indicate that there is some sort of a problem with the data.

However, it is entirely possible that a citation cycle exists in the data, but it was not loaded into the graph database due to some error in the data loading process.

Which year(s) had the most publications?

```
MATCH (p:Paper)-[:PUBLISHED_IN_YEAR]->(y:Year)
RETURN y.year AS Year, count(p) AS Publications
ORDER BY Publications DESC LIMIT 10
```

Listing 13: Year with Most Publications



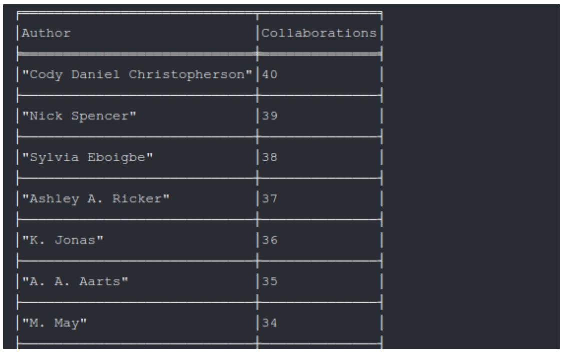
	Year	Publications
1	"2009"	9897
2	"2011"	9861
3	"2007"	9776
4	"2010"	9730
5	"2006"	9694
6	"2008"	9577

The above query returns the top 10 years with the most number of publications in the dataset. This can be a good measure of the activity in the field, as it shows in which years the most number of papers were published.

Which author(s) collaborate the most with others?

```
MATCH (a:Author)-[:COAUTHOR]->(other:Author)
RETURN a.name AS Author, count(distinct other) AS Collaborations
ORDER BY Collaborations DESC LIMIT 10
```

Listing 14: Authors with Most Collaborations



```
╒══════════════════════════════╤══════════════╕
│Author                        │Collaborations│
╞══════════════════════════════╪══════════════╡
│"Cody Daniel Christopherson"  │40            │
├──────────────────────────────┼──────────────┤
│"Nick Spencer"                │39            │
├──────────────────────────────┼──────────────┤
│"Sylvia Eboigbe"              │38            │
├──────────────────────────────┼──────────────┤
│"Ashley A. Ricker"            │37            │
├──────────────────────────────┼──────────────┤
│"K. Jonas"                    │36            │
├──────────────────────────────┼──────────────┤
│"A. A. Aarts"                 │35            │
├──────────────────────────────┼──────────────┤
│"M. May"                      │34            │
├──────────────────────────────┼──────────────┤
```

The above query returns the top 10 authors who have collaborated the most with other authors in the dataset. This can be a good measure of the social activity of the author, as it shows how much the author has collaborated with other authors, and can be a good measure of the social activity of the author.

Are there author(s) who have cited their own work?

```
MATCH (a:Author)-[:AUTHORED]->(:Paper)<-[:CITED]-(:Paper)<-[:AUTHORED]-(a)
RETURN a.name AS Author, count(*) AS SelfCitations
ORDER BY SelfCitations DESC LIMIT 10
```

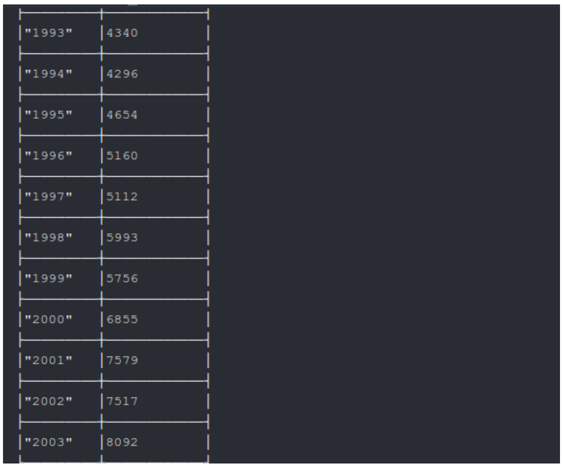
Listing 15: Authors who have Cited their Own Work

We didn't get any results for this query, which means that for the loaded data, there are no authors who have cited their own work. This is a good thing as it is generally considered unethical to cite your own work, and if an author is found to be citing their own work, it can be considered as self-plagiarism.

Publiation Trends over Time

```
MATCH (p:Paper)-[:PUBLISHED_IN_YEAR]->(y:Year)
RETURN y.year AS Year, count(p) AS Publications
ORDER BY Year
```

Listing 16: Publication Trends over Time



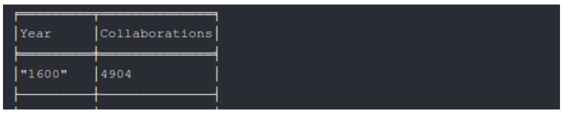
```
│"1993"    │4340        │
│"1994"    │4296        │
│"1995"    │4654        │
│"1996"    │5160        │
│"1997"    │5112        │
│"1998"    │5993        │
│"1999"    │5756        │
│"2000"    │6855        │
│"2001"    │7579        │
│"2002"    │7517        │
│"2003"    │8092        │
```

The above query returns the number of publications over time, which can be used to analyze the publication trends over time, and can be used to see how the number of publications has changed over time.

Collaboration trends over Time

```
MATCH (a1:Author)-[:AUTHORED]->(:Paper)<-[:AUTHORED]-(a2:Author), (p:Paper)-[:
    PUBLISHED_IN_YEAR]->(y:Year)
WHERE id(a1) < id(a2)
RETURN y.year AS Year, count(distinct a1) AS Collaborations
ORDER BY Year
```

Listing 17: Collaboration Trends over Time



```
│Year      │Collaborations│
│"1600"    │4904          │
```