1. Data collection:

the dataset is taken from a website called Aminer.  The citation data is extracted from DBLP, ACM, MAG (Microsoft Academic Graph), and other sources. And v is using the version 14 which is about 18 Gb and has around 5 million entries and 36 million citations.

1. Data preparation:

Removing the invalid data point like art works which is irreverent to our study. And removing the fields which are not required or useless fields to our study like useless\_attributes=['lang','volume','v12\_id','v12\_authors','indexed\_abstract','page\_start','page\_end','isbn','issn','doc\_type','url','issue','doi','keywords','abstract','title','authors','year']

1. Data input:

the data input is in this schema

Author; source; destination; contents of source; contents of destinations

1. Data pre-processing:
   1. We need to first convert our datapoints into a graph structure this is done by using the paper id and the references field of the cleaned dataset. With this v construct the citation graph but all the edges are undirected the v take random nodes and that K hope neighborhood of these nodes. the reason for the undirected edges is because it allows to backtrack the citation which increases the number and diversity of the sample.
   2. To find self-citations v take the sample and construct a heterogenous graph which have 2 different types of node and edges which are authors, papers which are nodes and the edges are cite and published.

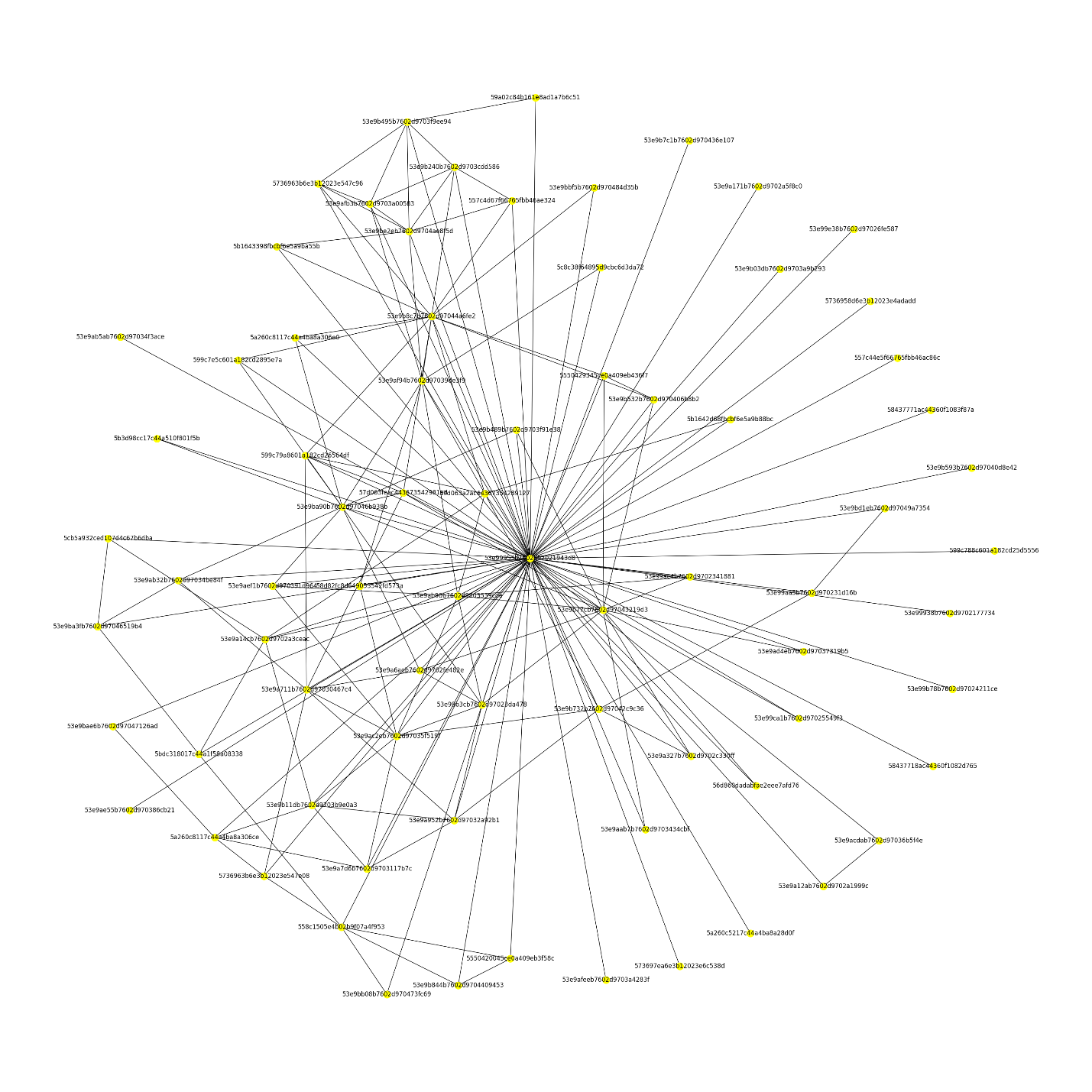
Now v convert this graph into a homogenous graph were all the nodes are the same.

this is done by converting the edges which are cite into a directed edge similar to the ones found in the wild and converting the published edge into a undirected or a bidirectional edge.

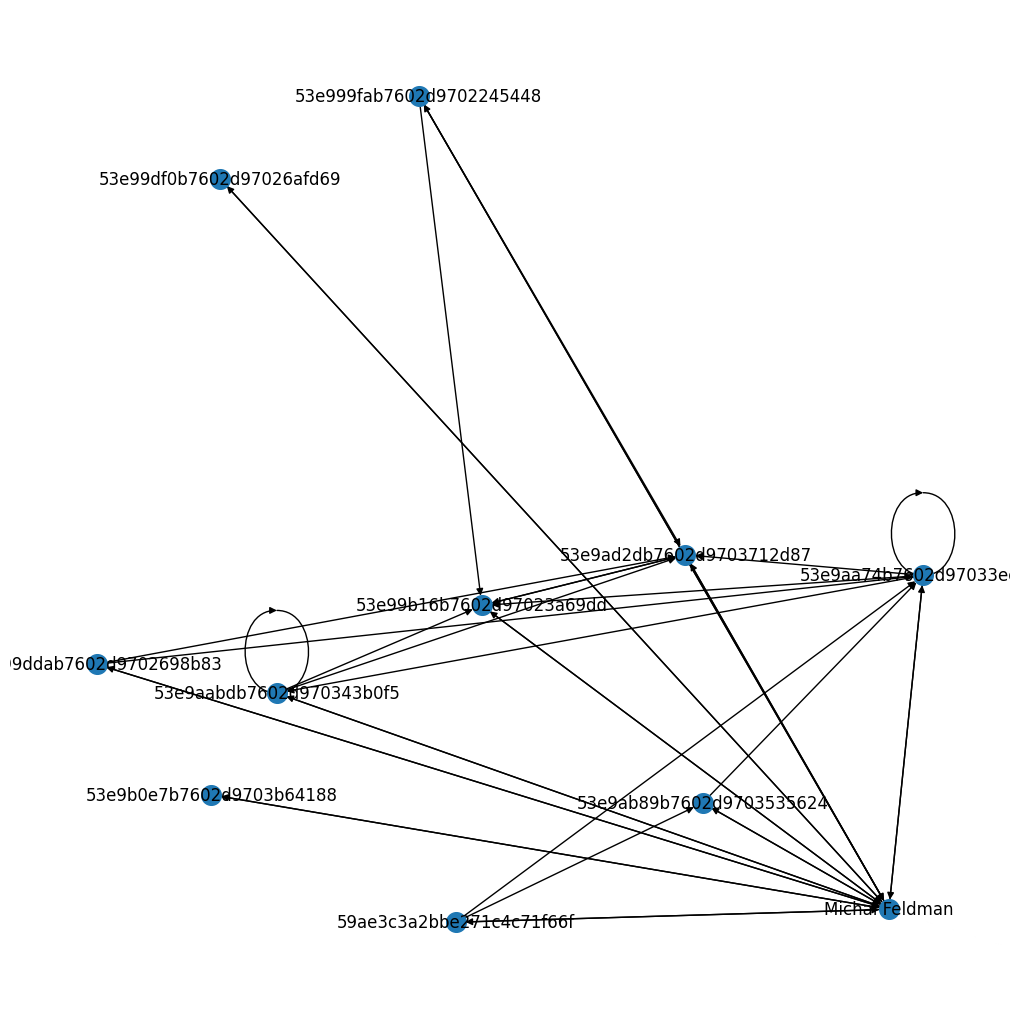
To find the self-citation we take the A^3 of the graph this will tell us which authors are preforming self-citation. Then v take the authors who are do this and take their one hope neighborhood and apply our cycle finding algo .

This algo takes the adjacency which is A and take the first hop of the authors the v take A^2 and take the authors row again. By using this algo v can get the papers which is citing itself which other cycle finding algo could not do .

1. Data Visualization:
   1. This is the visualization for the sample obtained form the dataset where the edges are undirected and it is a 1 hop neighborhood of a paper.



* 1. The is a visualization of a author performing a self-citation



1. Data Interpretation:

In this part v make some of our own features and take the fields which and make use of it

The fields which we are making are

* + 1. TSR(total self citation rate) this tells how much self citation the paper has / total number of citation.
    2. Then v make a summarization or the context of the source paper and the summarization of the destination paper.
    3. And v may take other features based on the model performance.

1. Storage : everything other that the final self citation data is stored in json format which is easier to handle compare CSV.

The self citations dataset is stored in CSV format for easy of human use.