Exploring Patent Citation Networks Using Python

Jeremy McCormick

DS745, Fall 2022

Introduction

Many areas of intellectual research contain referential citation networks. The Stanford SNAP Project maintains several of these large datasets with directed edges from the identifier of a citing work to its citations. The Patent citation network from the SNAP data archive contains almost 4 million patent records from the years 1963 to 1999 with 16.5 million citation links from 1975 to 1999. For about 1.8 million of these records, there are only citations rather than information on the full patent.

Purpose

This analysis will explore a large dataset of citations to discover some interesting patterns and insights. This will include analysis over time of statistics such as the number of citations per patent and breakdowns of references by country. Several smaller sub-graphs will be explored. Since the dataset is so large, subsets and aggregates will be used to visualize parts of the network.

Datasets

The primary dataset is formatted as two columns, the first with the citing paper and the second with the citation. Each of these is a valid USPTO utility patent number. The NBER U.S. Patent Citations Data supplements these links by providing information on these records such as the country of origin and year the patent was granted.

The NBER data is split into a number of files. The following were used in this analysis:

File	Descr	Link
Cite75_99.txt	Pairwise citations data	https://data.nber.org/patents/Cite75_99.zip
pat63_99.txt	Patent information	https://data.nber.org/patents/pat63_99.zip
subcategories.txt	Subcategory classifications	https://data.nber.org/patents/subcategories.txt
countries.txt	Country codes	https://data.nber.org/patents/list_of_countries.txt

Data was loaded from CSV files into a data frame within the Python Pandas library. After processing, the data frame was saved to a binary Parquet format file so that it could be easily reloaded later without needing to re-perform cleaning and transformations.

After reading in the data files and linking them via the patent numbers, the working data set had a large number of rows:

```
Record count: 2923922
```

Each row contains a list of the patents it references as well as those which reference it. This will be used in analysis for calculating aggregate statistics and constructing a graph of the data.

Creating the Citations Graph

The NetworkX packge provides a comprehensive library for studying complex networks. The patent data was loaded into this toolkit with several node attributes including the country of the patent and the year it was granted. In the base data, the edges had no extra attributes but were directed from the citing to cited patent. Each record in the dataset was added to the graph.

```
Nodes added to graph: 2923922
```

Then the edges were created by linking nodes from citing to cited patents. The total edge count corresponds to the total number of citations across all patents. Interestingly, the graph has significantly more nodes after this operation, due to the fact that the citation data contains patents for which there is no additional information in the dataset. These nodes lacking attributes were either included or excluded depending on the type of analysis.

```
Total node count after adding edges: 3942429
```

```
Total edge count: 16512783
```

Data Exploration

Summary Statistics

Summary statistics were computed for the number of citations within a patent and the count of those patents citing it across all records (1963 - 1999). These are shown below.

Name	Mean	Median	Min	Max	Std
Citations	5.65	4	0	770	8.42
Cited by	4.78	3	0	779	7.34

These look quite similar, which is probably expected since they are highly related to each other. The minimum value is zero for each, for patents that have no citations in them or with none citing them, respectively. The maximum value for both is very large compared with the mean, median and standard deviation, indicating that both distributions are highly skewed. This indicates that most patents have both a low number of internal citations and are not cited by very many other patents. Some very few patents are heavily cited, and some small percentage have a large number of internal citations.

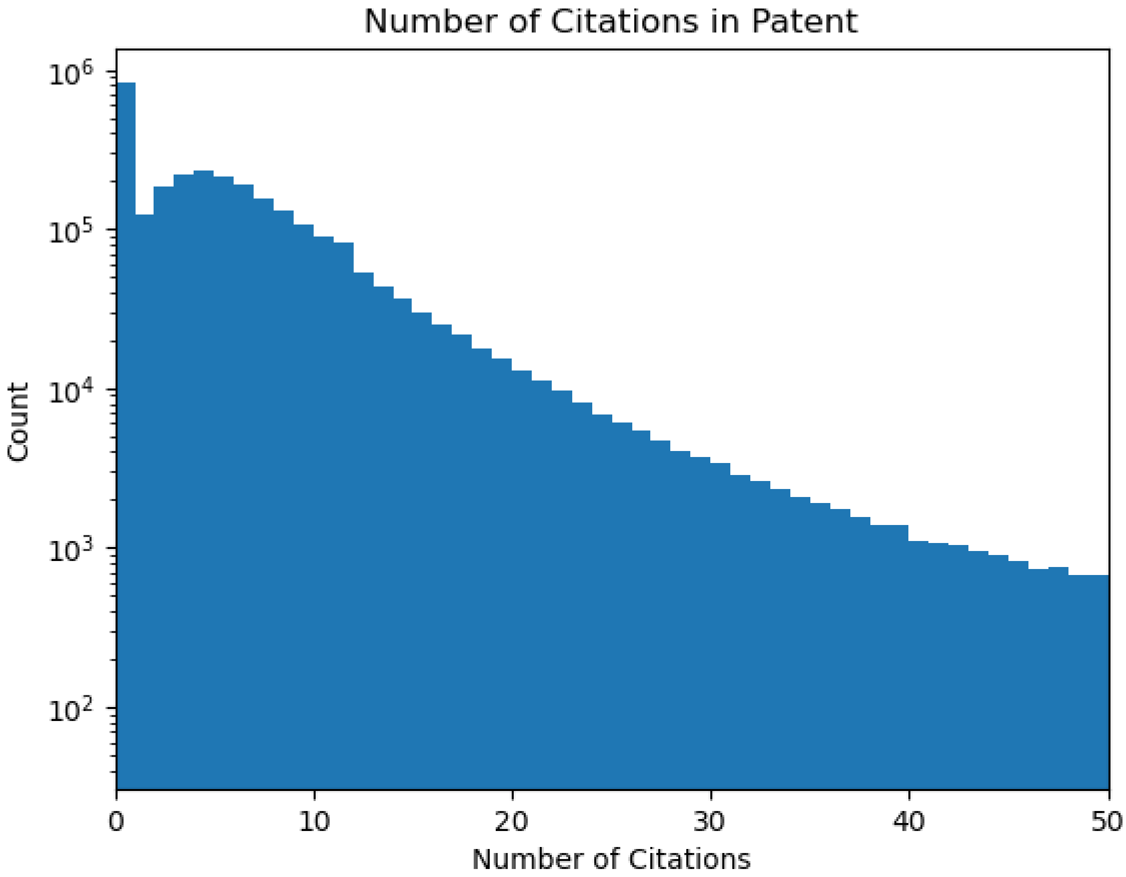
Quantiles are computed below for both incoming and outgoing citations below.

	10%	25%	50%	75%	90%	99%
Citations	0	0	4	8	13	34
Cited by	0	1	3	6	12	33

These show similar characteristics to the other descriptive statistics, with 25% of the patents having no citations and a quarter being cited only once or not at all. The 99th percentile shows that the maximum value is an extreme number of standard deviations beyond the mean.

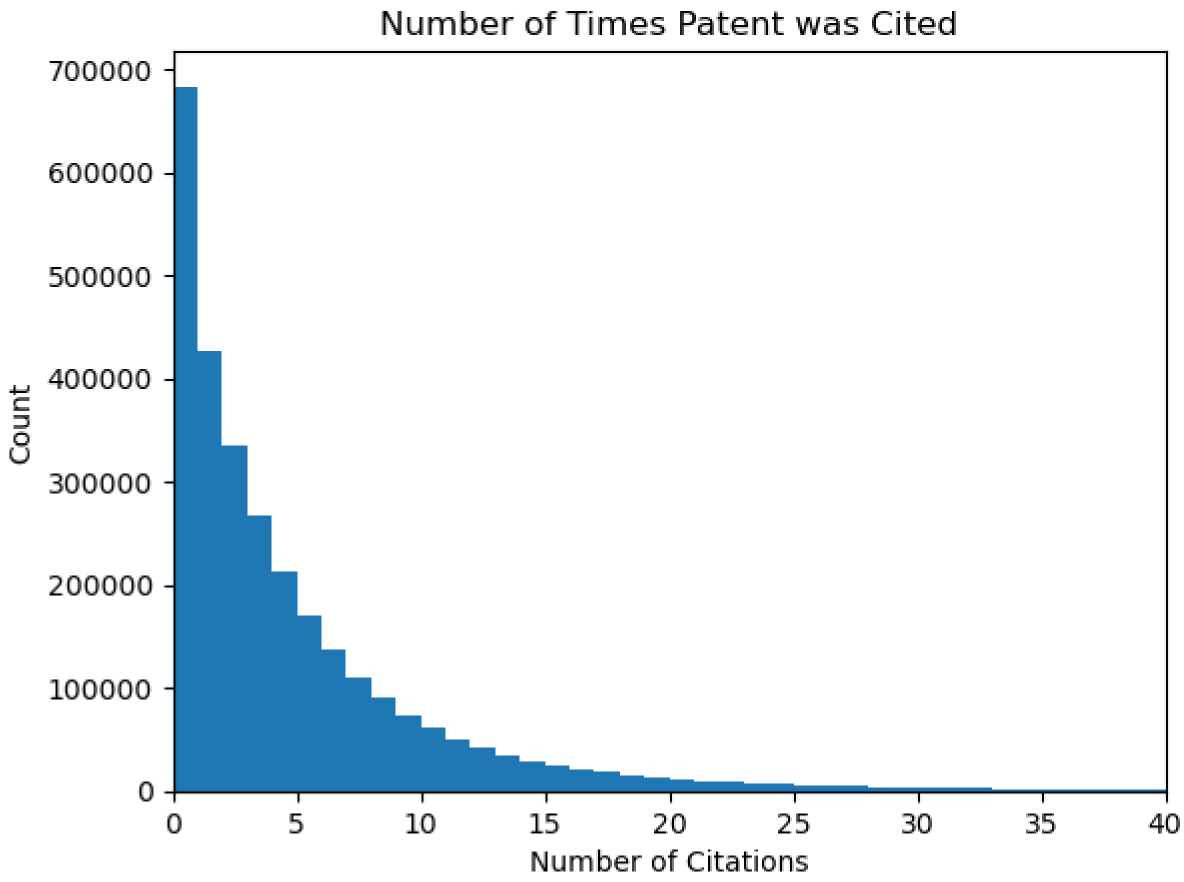
Histograms

Several histograms were used to perform preliminary data exploration on the citation graph. The citation counts and tally of citing patents were cached within each patent record in order faciliate creation of summary statistics across the dataset.



The histogram of number of citations by patent shows a heavily skewed distribution with a large peak at zero for those patents with no citations of other patents at all, which is a common occurrence illustrated by the previously shown summary statistics and quartiles. There is a peak around 4-5 denoting roughly the mean with a very long tail similar to those in a Poisson or Power Law distribution.

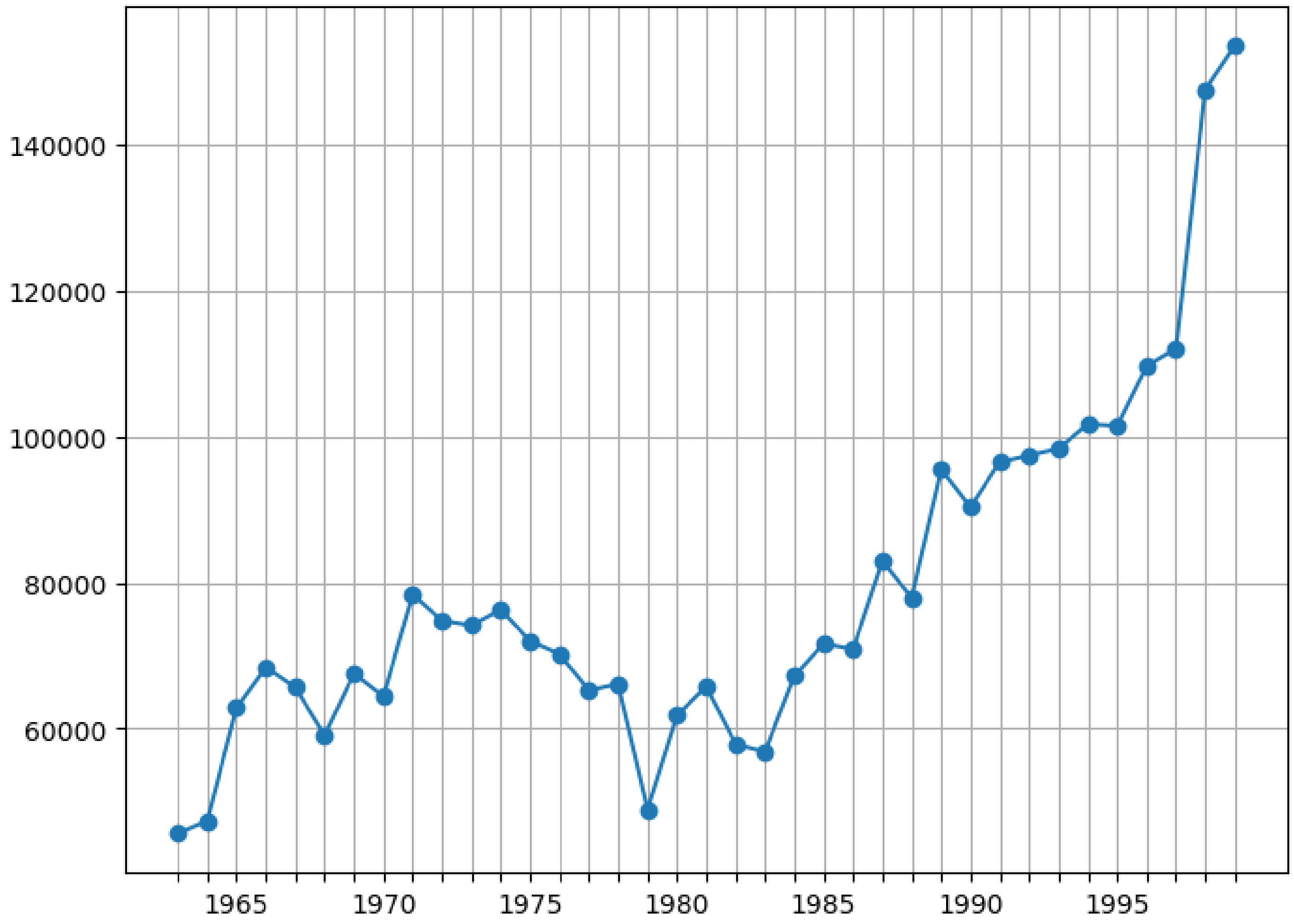
The number of times each patent was cited is shown below in a histogram.The number of times each patent was cited is shown below in a histogram.



This distribution clearly follows a Power Law with an exponential falloff, indicating that most patents are not cited or have only a few external citations, whereas some relative few, technologically significant ones have many. (Incidentally, the most cited patent in the dataset relates to bubble jet printing.)

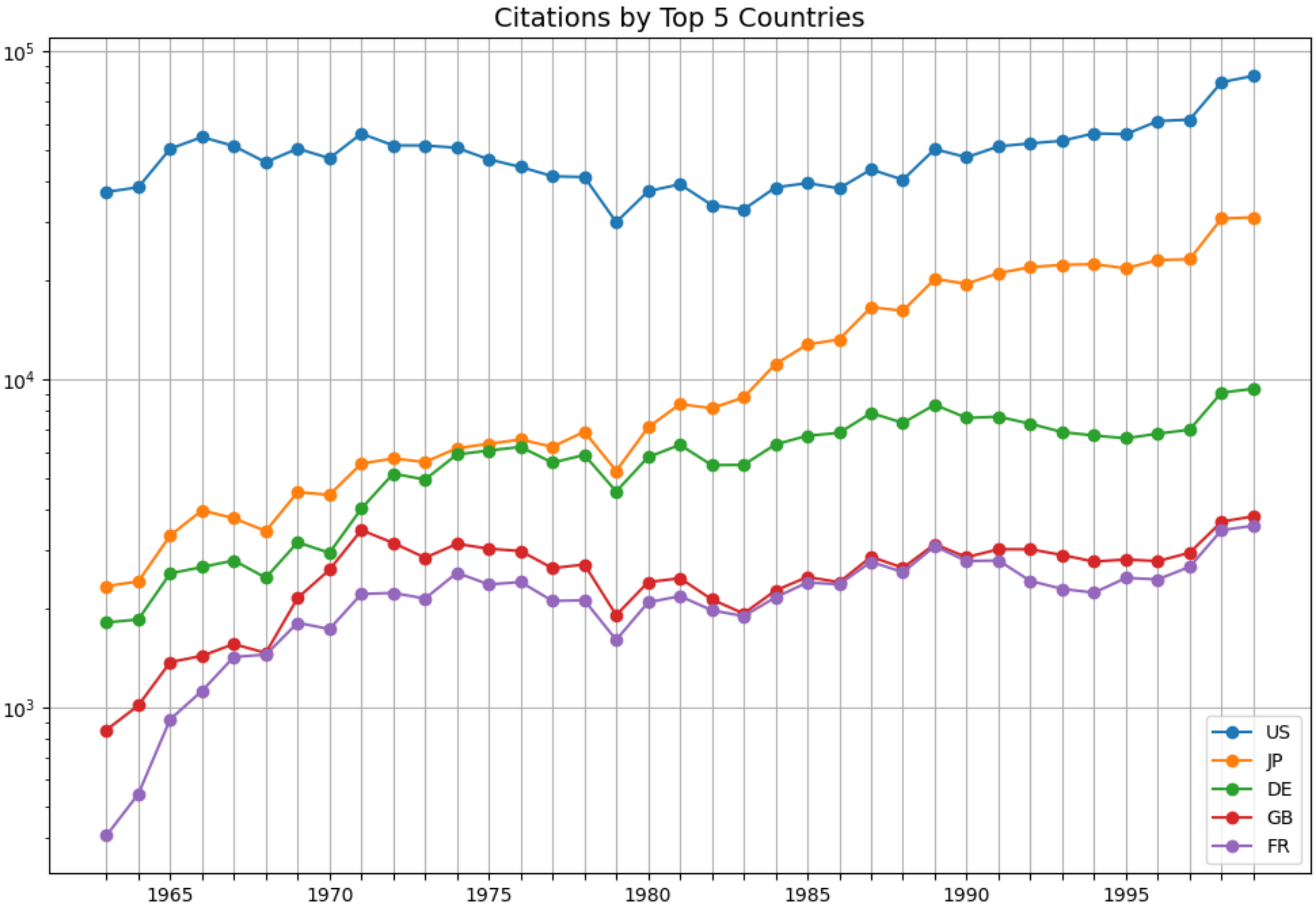
Time Series

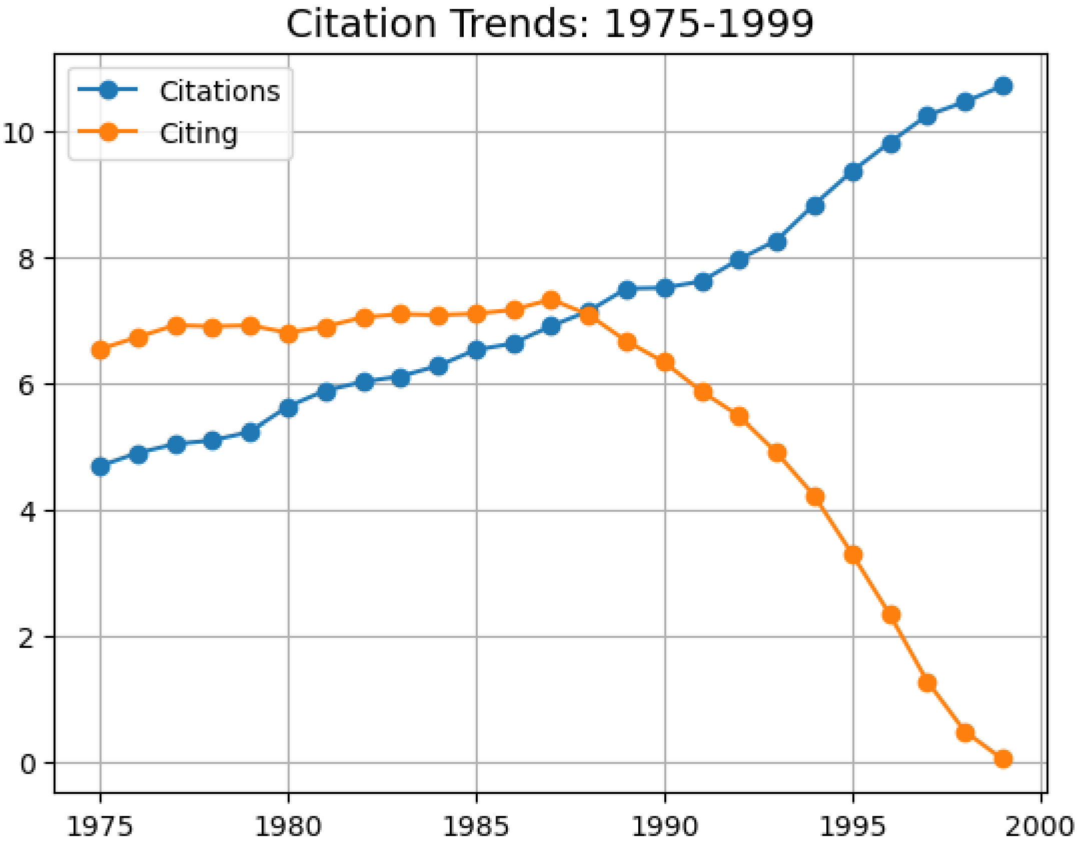
The patents with detailed records have an associated grant year. This can be used to show how the statistics of citations have changed over the time period covered by the entire dataset (1963 - 1999). Below is a time series showing the number of patents by year. This illustrates a strong upward trend in number of patents granted, particularly since the mid 1980's.



```
GYEAR  COUNTRY
1999   US         83906
1998   US         80291
1997   US         61707
1996   US         61104
1994   US         56066
                    ...
1964   FR          1013
1965   JP           919
1963   FR           853
1964   JP           544
1963   JP           407
Length: 185, dtype: int64
```

Most of the patents have country information related to the assignee (the entity requesting the patent). The US patent system is used for intellectual property protection worldwide, and while many requests are made by those in the United States, a large number also originate from abroad. The time series below shows a (log scale) graph of counts from the countries with the top 5 most patents in the database. The age of this dataset is quite telling, as this would likely look much different two decades later (China is not represented at all, for instance).

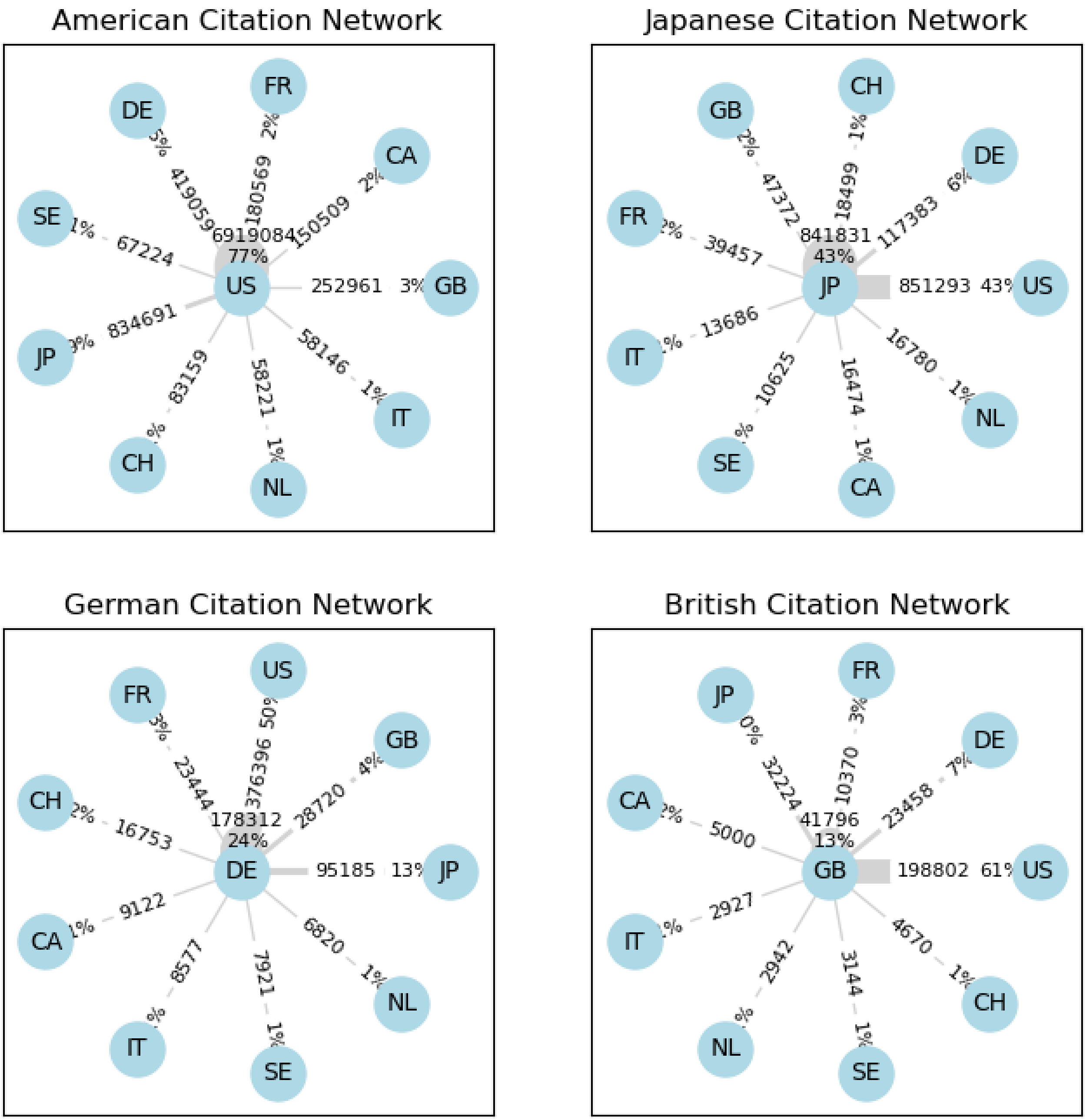

Citations by Top 5 Countries
10^5
10^4
10^3
1965
1970
1975
1980
1985
1990
1995
US
JP
DE
GB
FR


Citation Trends: 1975-1999
Citations
Citing
0
2
4
6
8
10
1975
1980
1985
1990
1995
2000

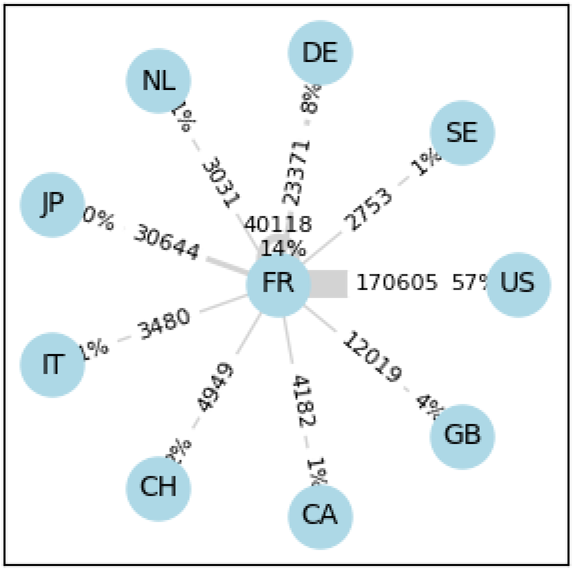
Finally, average citation counts were plotted along with the average of how many times a paper was cited by year of the patent. This shows that important historical patents are more likely to be referenced than newer patents, which show a falloff around the mid-1980's until the last year represented in the dataset. Average internal citations are on an increasing trend, more than doubling between 1975 and 1999. This may correlate to the technical complexity of patents increasing during this period, in terms of how much they relate to and use material from other patents.

Country Relationships Graph

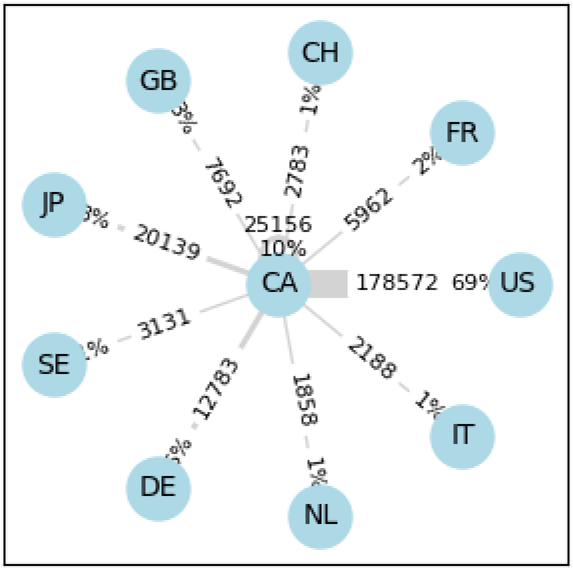
There are many different graph-like relationships that can be explored in this dataset using the associated node attributes. One example is the relationship between different countries. The following example analysis shows the share of citations within papers from one country which reference patents that originated in another one. Only the countries with the top 10 number of total patents in the dataset were used for this analysis.

American Citation Network
US
6919084
77%
DE
5%
419059
FR
2%
180569
CA
2%
150509
SE
1%
67224
GB
252961
3%
JP
9%
834691
CH
%
83159
NL
58221
1%
IT
58146
1%
Japanese Citation Network
JP
841831
43%
GB
2%
47372
CH
1%
18499
DE
6%
117383
FR
2%
39457
US
851293
43%
IT
1%
13686
SE
%
10625
CA
16474
1%
NL
16780
1%
German Citation Network
DE
178312
24%
FR
3%
23444
US
50%
376396
GB
4%
28720
CH
2%
16753
JP
95185
13%
CA
1%
9122
IT
%
8577
SE
7921
1%
NL
6820
1%
British Citation Network
GB
41796
13%
JP
0%
32224
FR
3%
10370
DE
7%
23458
CA
2%
5000
US
198802
61%
IT
1%
2927
NL
%
2942
SE
3144
1%
CH
4670
1%

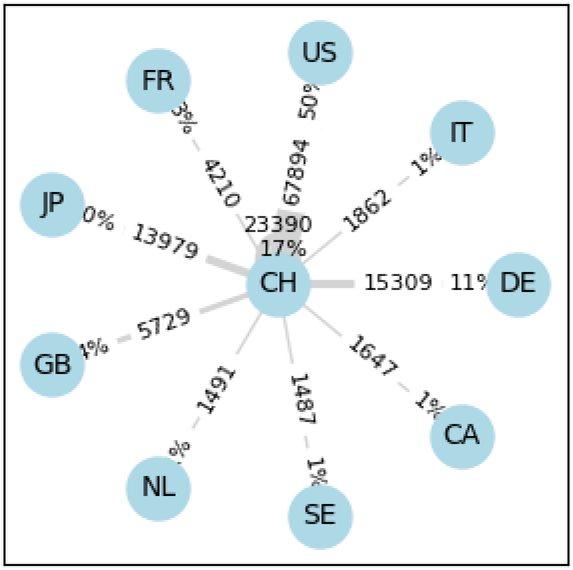
French Citation Network



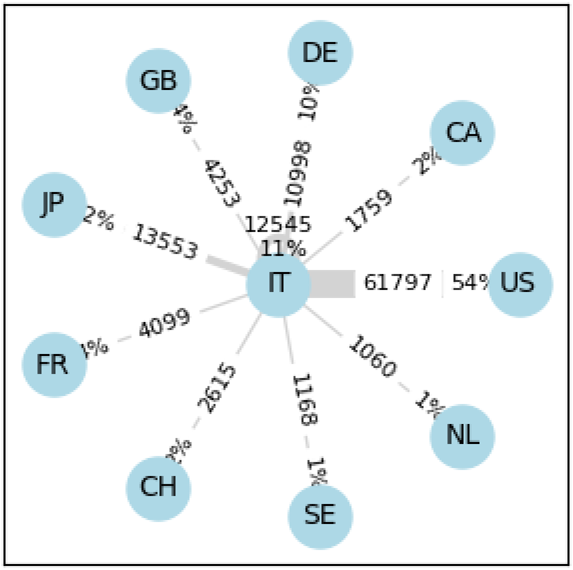
Canadian Citation Network

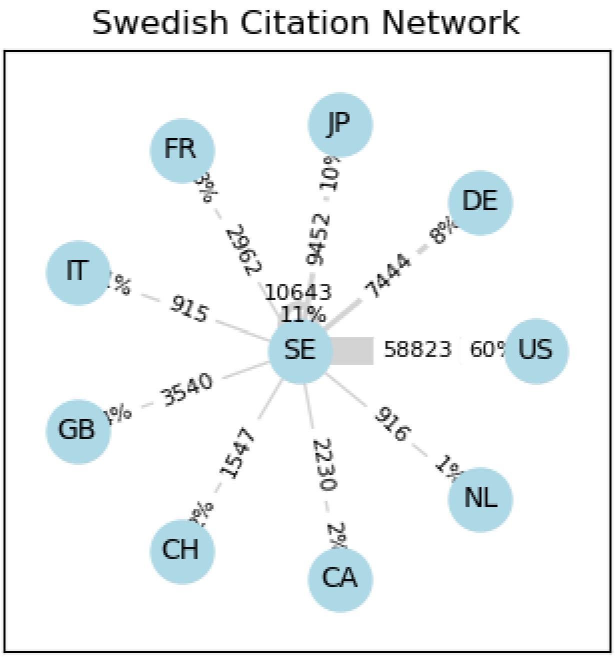


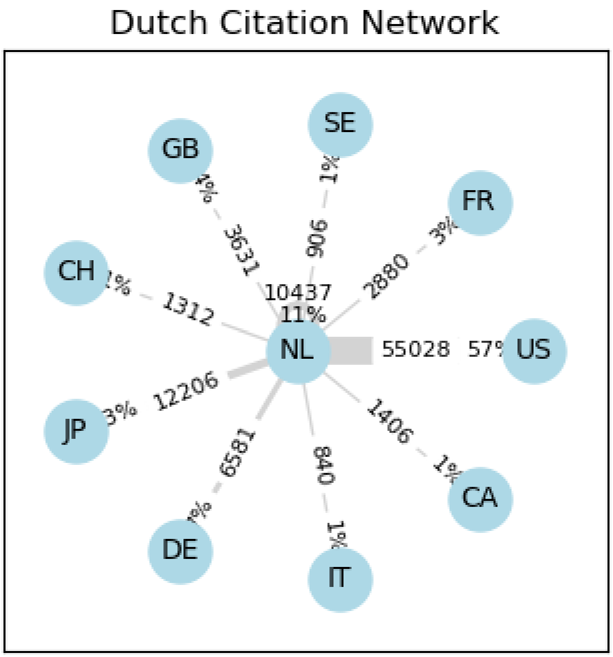
Swiss Citation Network



Italian Citation Network



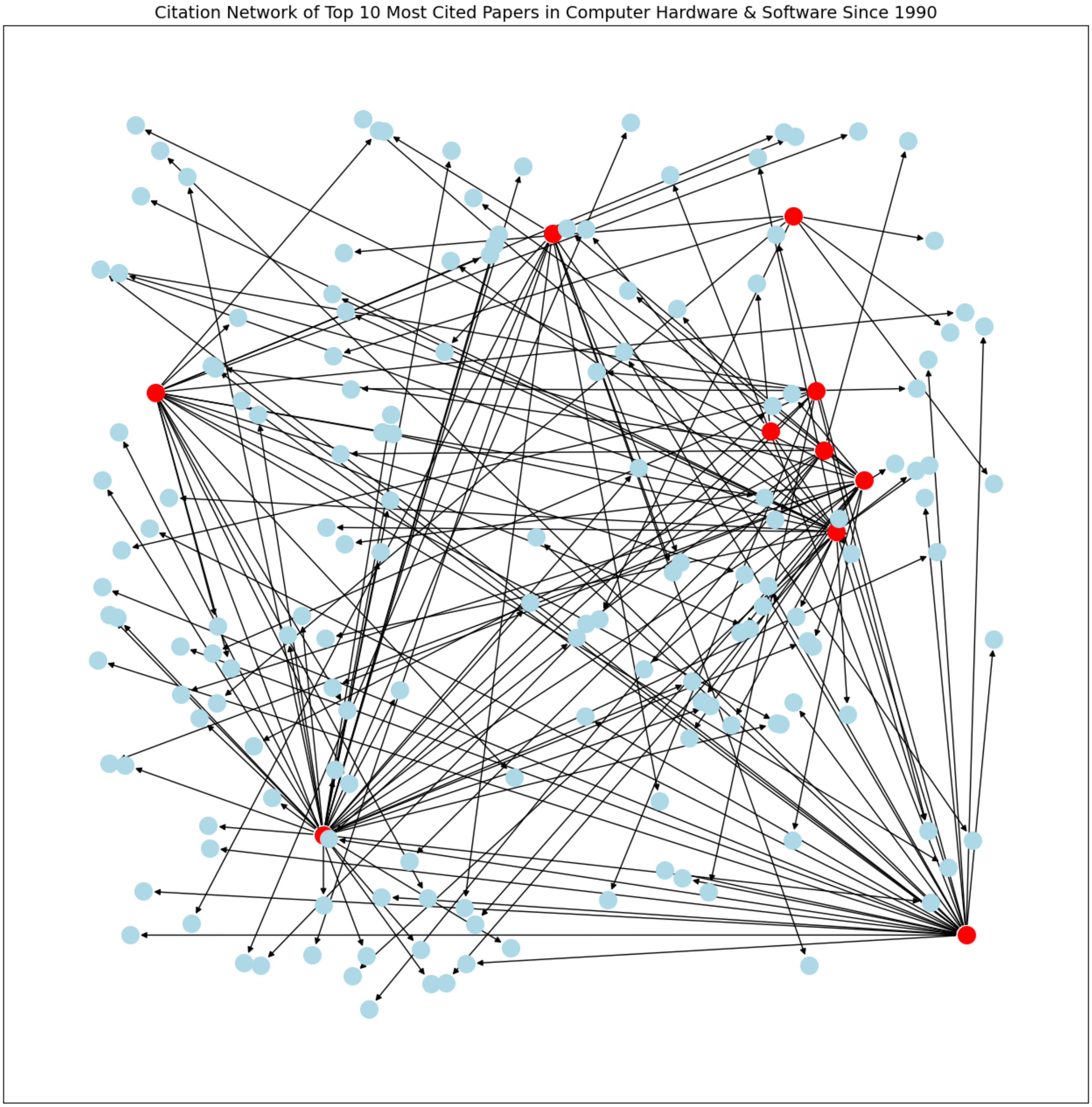




The country which is being explored by each graph is at the center in each of these plots. The number of citations to patents from other countries is shown on the graph edges, which are scaled according to these values. The percentage of the total represented by that reference appears next to the other country. A "self reference" arrow shows how many patents originated from within the same country. These show that patents originating from most other countries primarily reference ones from the United States. But for foreign countries, their own country tends to have a higher than average number compared with all the other (non-US) countries. This may show that while the United States has been intellectually dominant, inventors are also heavily influenced by research within their own countries.

Most Cited Patent Citation Network

Since the graph is so large with millions of nodes, only small subsets can be practically visualized at the node level. Logically, one might examine the most-cited patents within some sub-category. Below is the citation network amongst the top 10 papers in the "Computer Software & Hardware" category (sub-category 22 in the dataset) with these primary nodes shown in red and their references in blue.

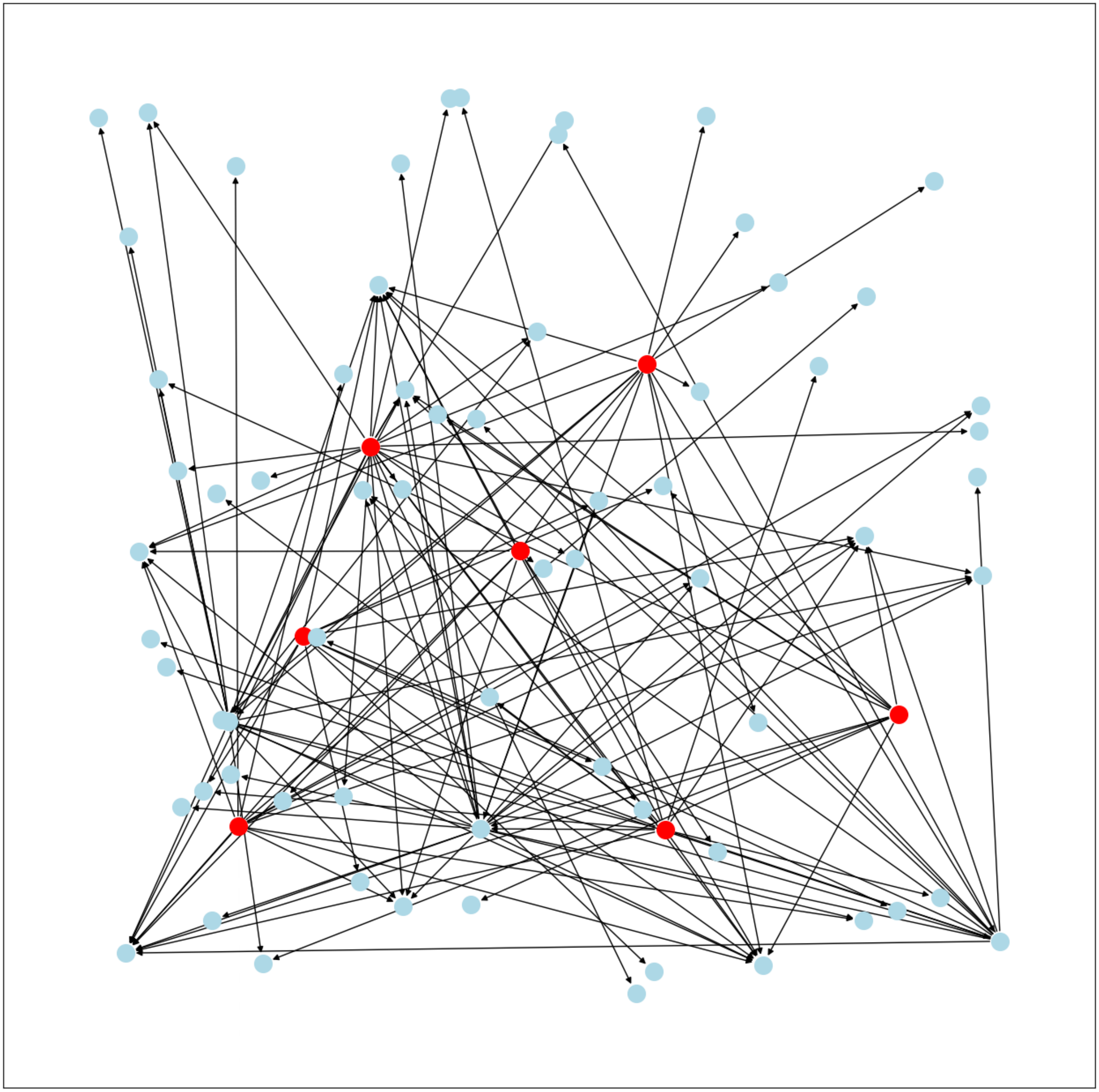

Citation Network of Top 10 Most Cited Papers in Computer Hardware & Software Since 1990

Amongst these papers, there are almost no common citations! This indicates that even within specific sub-categories, it may be uncommon that patents share many or any of the same set of references.

We can try to plot a more connected set of nodes by looking at similarities between sets of citations. Since this would be an *$O(N^2)$* operation to compare the citations of each patent against all the others, it makes sense to use a single node instead for this comparison. Starting with the top most-cited patent, Python's difflib can determine a numeric score between 0 (no similarity) to 1 (exactly the same) for two sets of citations based on their patent numbers. Again, to narrow down how many other patents are included, only the specific sub-category can be used (software & hardware). The scores for the most-cited patent versus all these others in the same sub-category are shown below.

Patent	Score
5167024	1.000000
5513361	0.620690
5218704	0.466667
5182810	0.413793
5404546	0.400000
5423045	0.375000
5201059	0.357143
5369771	0.357143
5903746	0.350000
5485623	0.344828

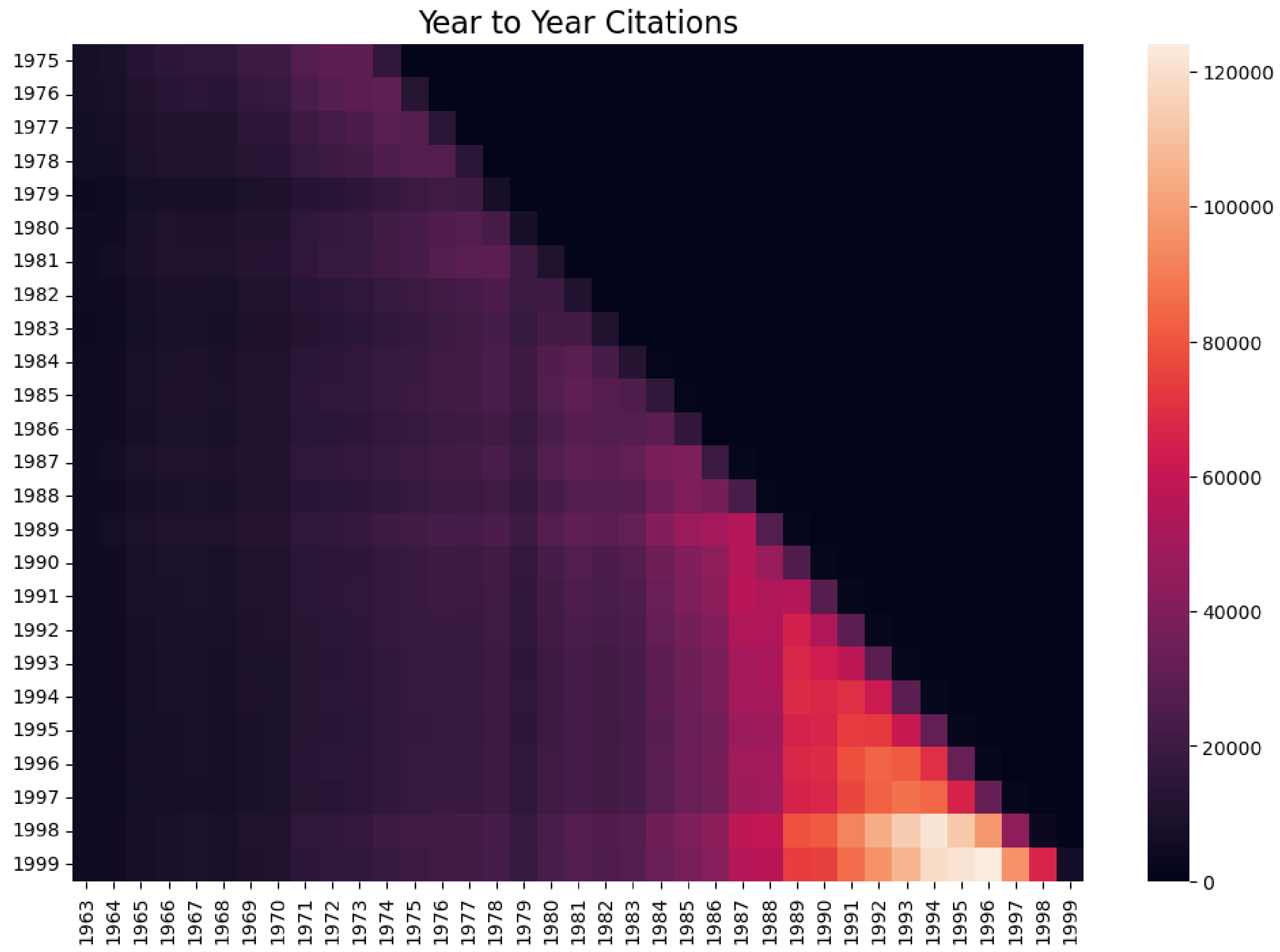
Citation Network of Top Paper in Computer Hardware & Software Since 1990 with Similarity Scores



This looks a lot more interesting! We can see that some of the nodes that were chosen for similarity share quite a few of the same citations. Strategies such as this can be used to find interesting sub-graphs in what is a large and relatively sparsely connected graph.

Year to Year Associations

As a final analysis, the citation relationships between different years will be explored. Initially, I had thought of visualizing this as a graph, but I realized that so many edges were difficult to visualize together. Instead, a colormap can be used to show how many patents from one year cite those from another.



This is perhaps a little difficult to interpret, but one interesting feature is the cluster of citations amongst patents in the 1990's, indicating that this was a very active period, technologically, which could also be seen in prior visualizations such as the time series. Patents in 1999, the last year from this dataset, heavily cite those from the mid-90's. This feature is probably due to inter-related technological advancements in the computing field during this time period.

Reflections

I spent a lot of time data wrangling for this project, and, in retrospect, it might have been more fruitful to pick a sub-network (by sub-category, for instance) on which to focus rather than dealing with the entire dataset. Just loading in the entire dataset from text files and linking by patent numbers took probably 10-15 minutes each time. Big data tools might also have helped, but I was limited to the processing on my laptop. I also noticed that visualizing graphs effectively is challenging, because high density networks with lots of edges tend to become uninterpretable past a certain number of nodes. I am considering exploring this dataset further, as I believe it has a lot of interesting possibilities in terms of visualizing aggregate relationships (such as between countries, sub-categories, etc.) as well as more highly connected sub-graphs.

References

Please see internal hyperlinks for all references. There were no academic journal articles used for this report.