HDDT analysis - All bigraph popular entity memberships (non CEDA)

My PhD project

Subject:

My own research, in collaboration with others, has revealed the extensive social connectivity between the roughly 600 members of a 'Quaker Led Network (QLN)' and their involvement within a community of roughly 3000, spread across four organisations in Britain active between 1830 and 1870, which the QLN network helped to set up and staff. I call these, the 'Centres for the Emergence of Discipline of Anthropology in Britain' (CEDA).

Question 1:

What can be revealed if a historian uses data science to study a large historical community over a long period of time by bringing together and integrating metadata from catalogues, indexes, and genealogical data?

Methodology:

I have designed, built and I am now using a suite of open-source and reproducible relational database technologies and digital analytic tools to visualise and scrutinise the entire community of some 3000 activists over 40 years (1830-1870), picking out the Quakers amongst them so that the community can be explored at both group and individual level. I am able to model the 'connected' relationships between the individual members of the CEDA through time, including kinship, education, occupations, locations and organisations.

Question 2:

What is the extent of Quaker involvement in the CEDA, over the 40 year time span researched, and was Quaker kinship as socially cohesive as (say) education or occupation amongst the wider community?

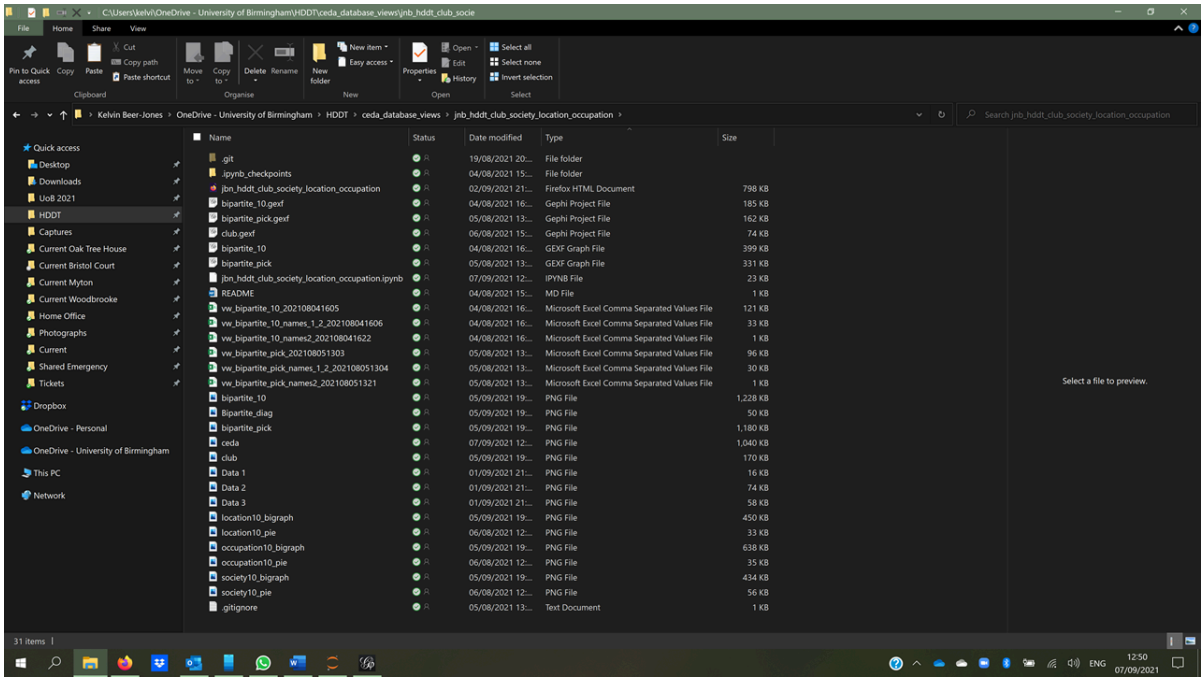
This is a code cell

In [ ]:

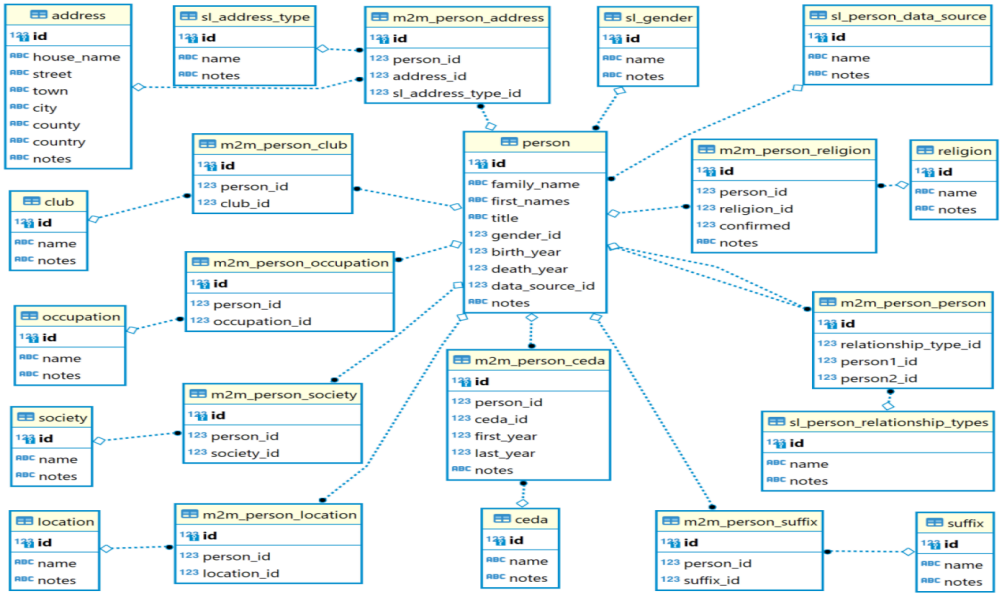
```
# First we call up the python packages we need to perform the analysis:

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
plt.rc('figure', figsize=(20, 10))
from IPython.display import set_matplotlib_formats
set_matplotlib_formats('png', 'pdf')
from operator import itemgetter
import networkx as nx
from networkx.algorithms import community #This part of networkx, for community dete
import nbconvert
import csv

# <img src="xxxx.png">
```

These are the resources in my container for this exercise

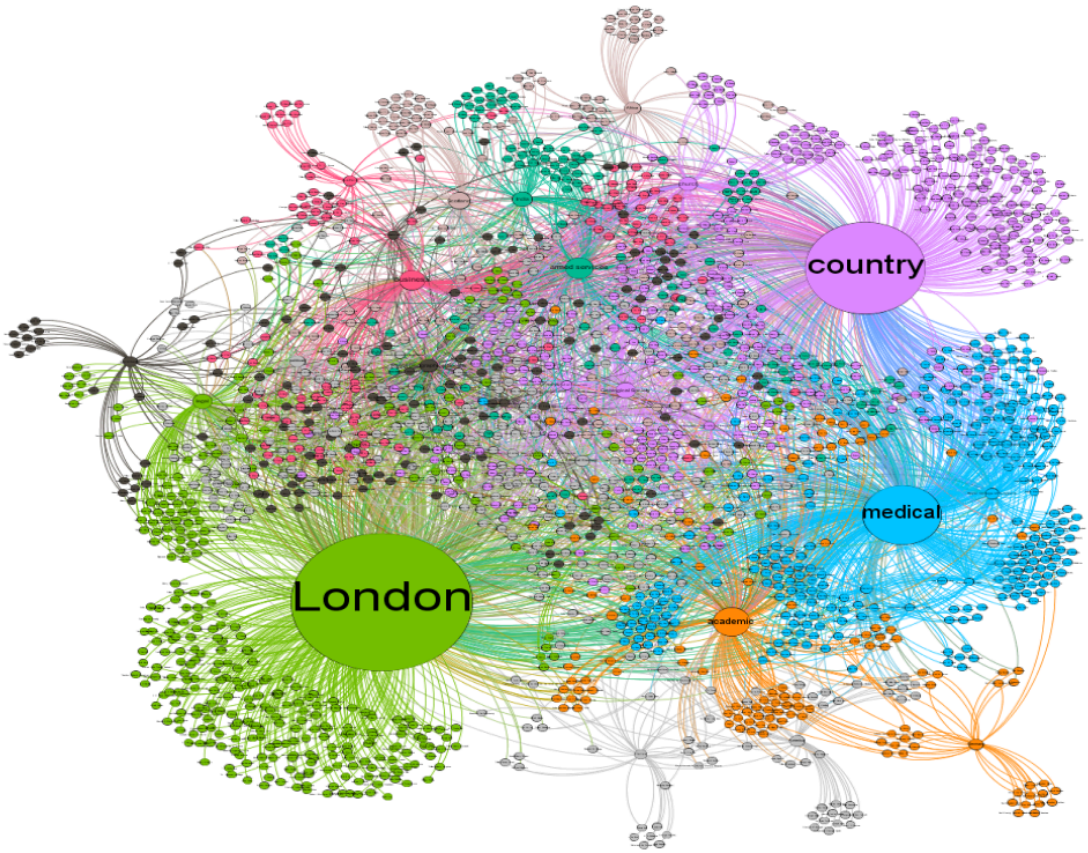
Resource containers are also GitHub repo's facilitating granular version control of all changes made to any resources



This is the structure of the SQLite database (ERD)

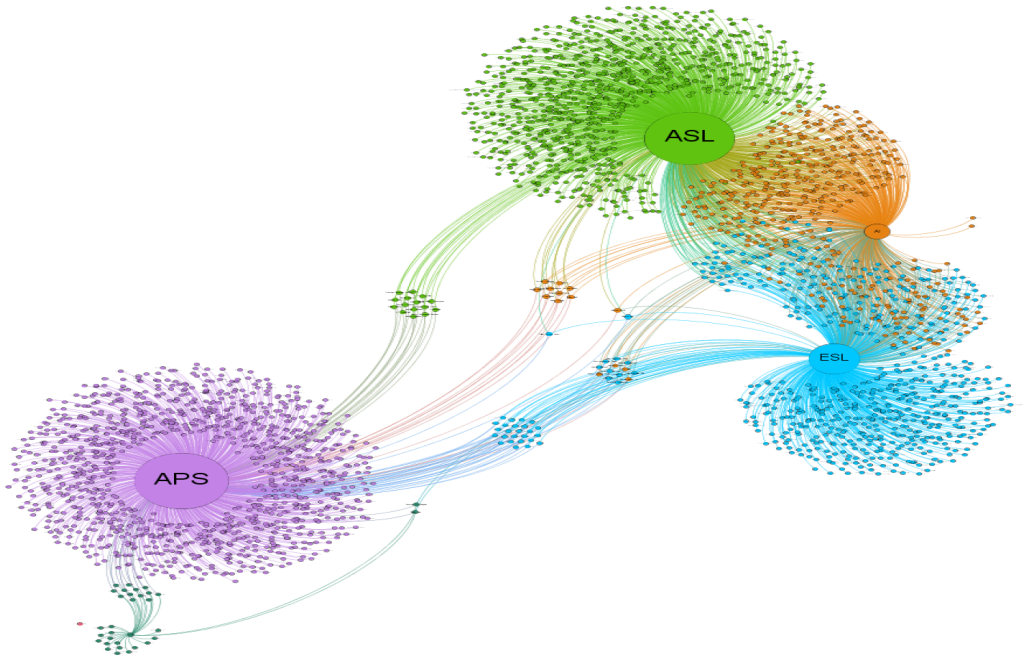


An analysis of the popular memberships of bigraph entities (non CEDA) associated with 3095 members of the CEDA 1830 -1870



This graph shows all of the popular bigraph data in the database. Including all data would result in a 'hairball' where data would be too dense to be capable of analysis at this (the highest) level. (See Most popular entities section below)

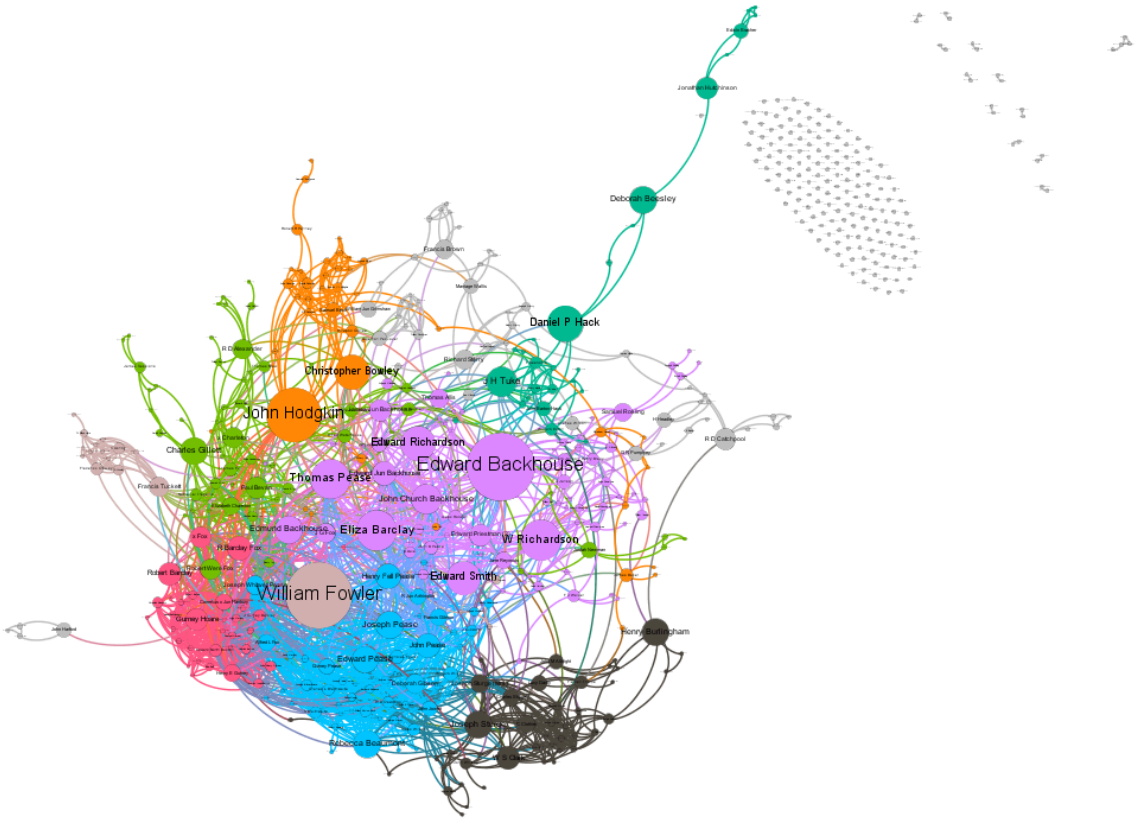
All persons are members of at least one CEDA



Society	abv.	Dates
Quaker Committee on the Aborigines*	QCA	1832/37 - 1846
Aborigines Protection Society	APS	1837 - 1919
Ethnological Society of London	ESL	1843 - 1871
Anthropological Society of London	ASL	1863 - 1871
Anthropological Institute	AI	1843 - 1871
London Anthropological Society**	LAS	1873 - 1874

- Origin Society included in this project but not recognised by RAI. ** not included in this project (beyond 1871 cut off date).

Here are the Quakers we looked at in our Gephi session



Working with a variety of datatables

A 'complete' dataset

Would be one like this, where all of the data can be contained witin a perfect rectangular block of cells ('containers') and every container contains only one data item and every data item can be located by the coordinates 'Row n, Column n'

Table	Col1	Col2	Col3	Col4	Col5
Row1	A	B	C	D	E
Row2	F	G	H	I	J
Row3	K	L	M	N	O
Row4	P	Q	R	S	T
Row5	U	V	W	X	Y

An 'incomplete' dataset

When historical data is used often some data is missing (permanently lost) and the HDDT is able to accept 'Incomplete' datasets. The HDDT does not lose functionality because of the incomplete nature of much historical data.

Table	Col1	Col2	Col3	Col4	Col5
Row1		B	C	D	E
Row2	F		H	I	J
Row3	K	L		N	O
Row4	P	Q	R		T
Row5	U	V	W	X	

An 'irregular' dataset

The HDDT has been designed to accept Irregular datasets. The surviving evidence of the past is not only often Incomplete, it is frequently Irregular, where multiple datasets have different dimensions. (Either because the data in itself is intrinsically different or because different data collectors use different cataloguing methods)

1	A		1.0			
2	B		2.1	Cat	Q	
3	C		3.2	Dog	W	
4	D	fff	4.3	Fish	E	
5	E	ggg			R	
6	F	hhh			T	!
7	G	iii			Y	“

For the HDDT a qualifying dataset is a data set of any dimensions, complete, incomplete or irregular. The only requirement is that all datasets must contain a single common containing one universally shared data item. The HDDT requires all data sets to contain datatables that can be referenced to a PERSON (Name) in one of its rows.

Conflicts between dataset Person (Name)'s are resolved by adopting in this project by nominating the 'RAI dataset' as the 'Authority Index'. With careful matching of Person (Name)'s found in other datasets, the RAI naming rule applies throughout.

Introduction to bigraph analysis

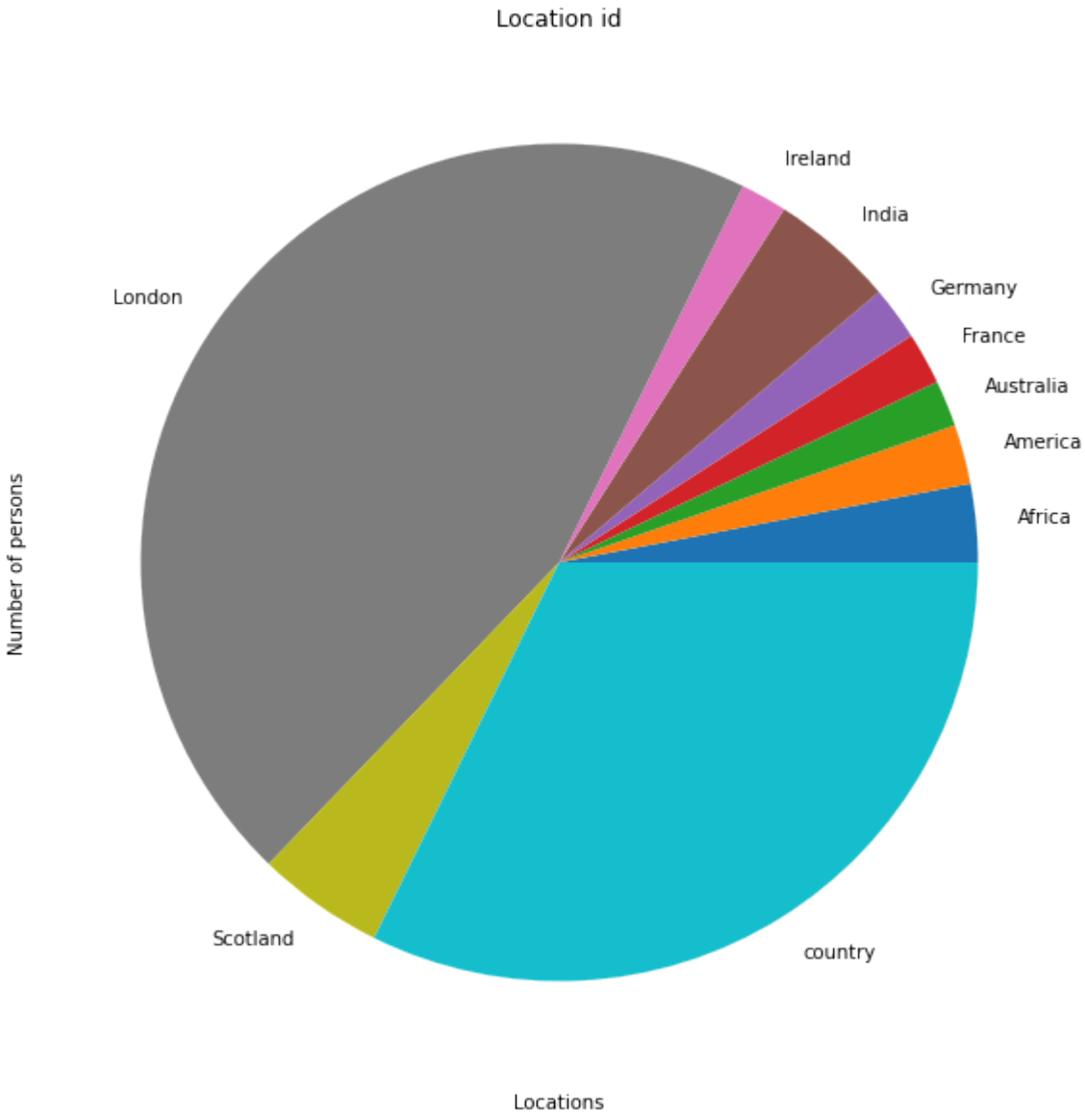
Bigraph data statistics

Table	Rows	Columns	Exc?
ceda	6	1	Yes
person_ceda	3894	4	Yes
club	68	1	*
person_club	323	2	*
location	83	1	
person_location	2061	2	
occupation	93	1	
person_occupation	1883	2	
society	260	1	
person_society	1238	2	

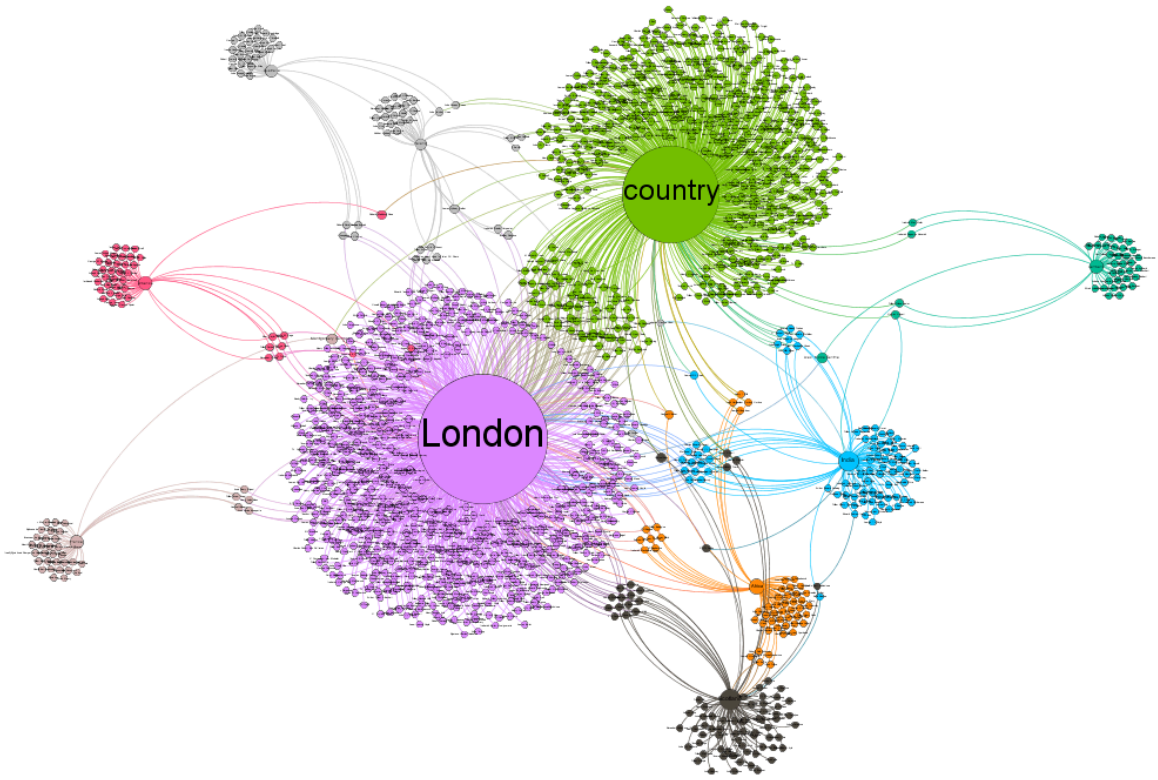
- Due to low levels of poulation of 67 other clubs only the Athenaeum Club is included in analysis

Locations

Top 10 locations pie chart



Top 10 locations bigraph



We can see that London and 'country'(sic) are the most populated locations. Because the 'country' location is an aggregate (and not a specific location) we can think of London and 'country' as a twin centre. Within the twin centre we can see the members of both London and 'country' locations and that the members of each are highly networked. We can also see that the London location contains many members who have no association with any other group (including 'country'). London 1830 - 1870, was densely populated and so it is possible that members of the London location had other modes of association. Because the 'country' location is an aggregate we cannot make the same analysis to the same extent, it is possible that many members in (say) Newcastle had no association with other members in (say) Bristol. We can see the large group of members who were members of both London and 'country' locations. It is highly likely that these members served as conduits of communication and group cohesion. It is interesting to note that only 3 members of this London and 'country' group were members of groups outside of the twin centre.

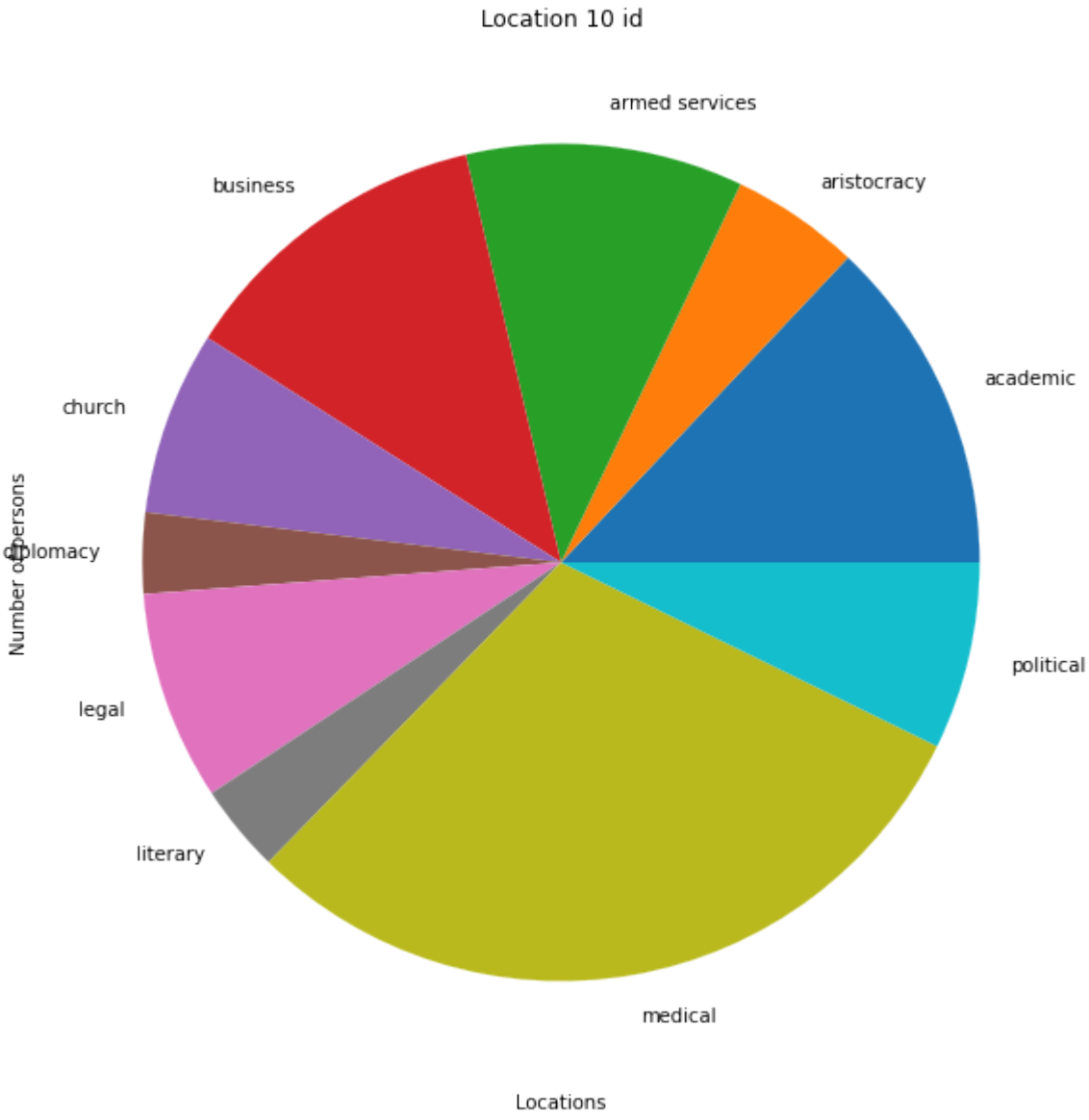
Eight other location each have a membership of around 30 members (we can call these the satellites), all of the satellite groups relate directly to the twin centre with very few members associated with more that one satellite location.

Australia and Ireland have associations with both London and 'country'. The German location is most closely associated with the 'country' group. All of the other locations are strongly associated with the London location.

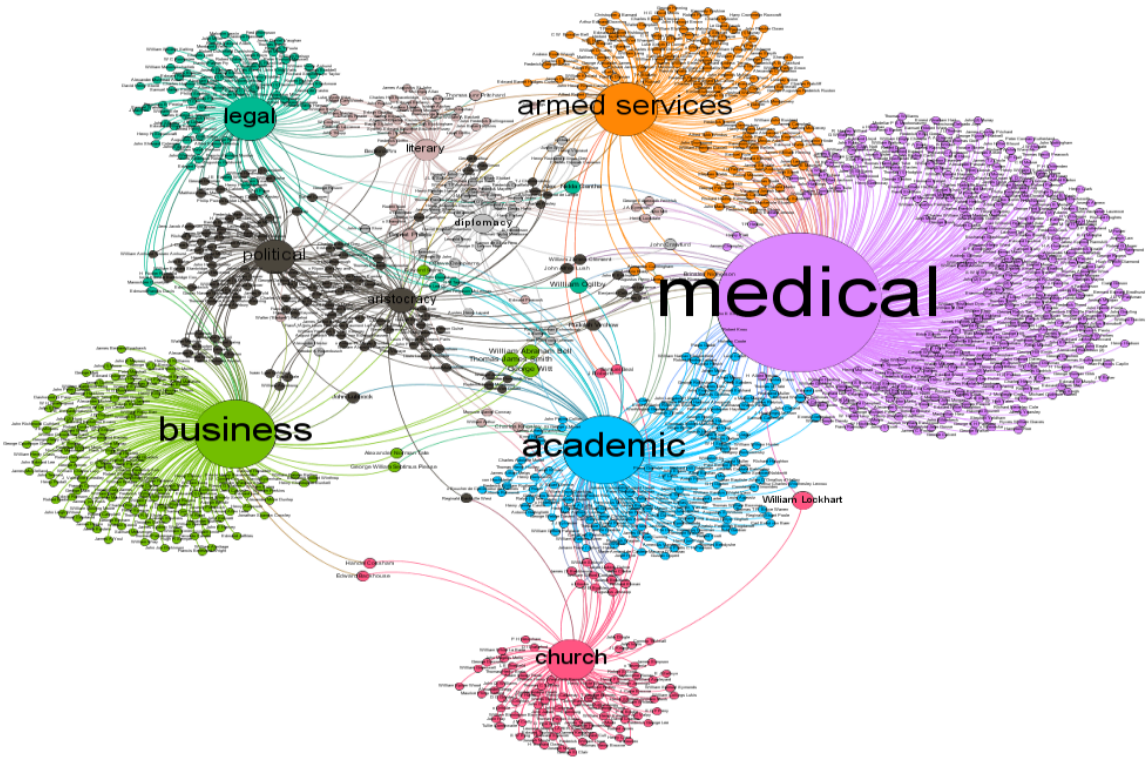
Germany (far right) is the location least associated with London. Alex Nidda Genthe is the only member from Germany who is also a member of the London location. Friedrich Max Muller, Frederick Augustus Haverick and Gustav Oppert each network with 'country' members. William Wilson Hunter is the only 'country' member who also appears in the Germany location. He and Gustav oppert also have a location connection with India.

Occupations

Top 10 occupations pie chart



Top 10 occupations bigraph



We can see that 'medical','academic' and 'armed services' together account for half of the members by occupation. We can also see that the largest three occupational categories each contain many members who have no association with any other occupational group. We can see that the medical categories contain many members who are also members of the other two principal categories ('academic' and 'armed services'). It is highly likely that these members served as conduits of communication and group cohesion amongst the three principal occupational cegories.

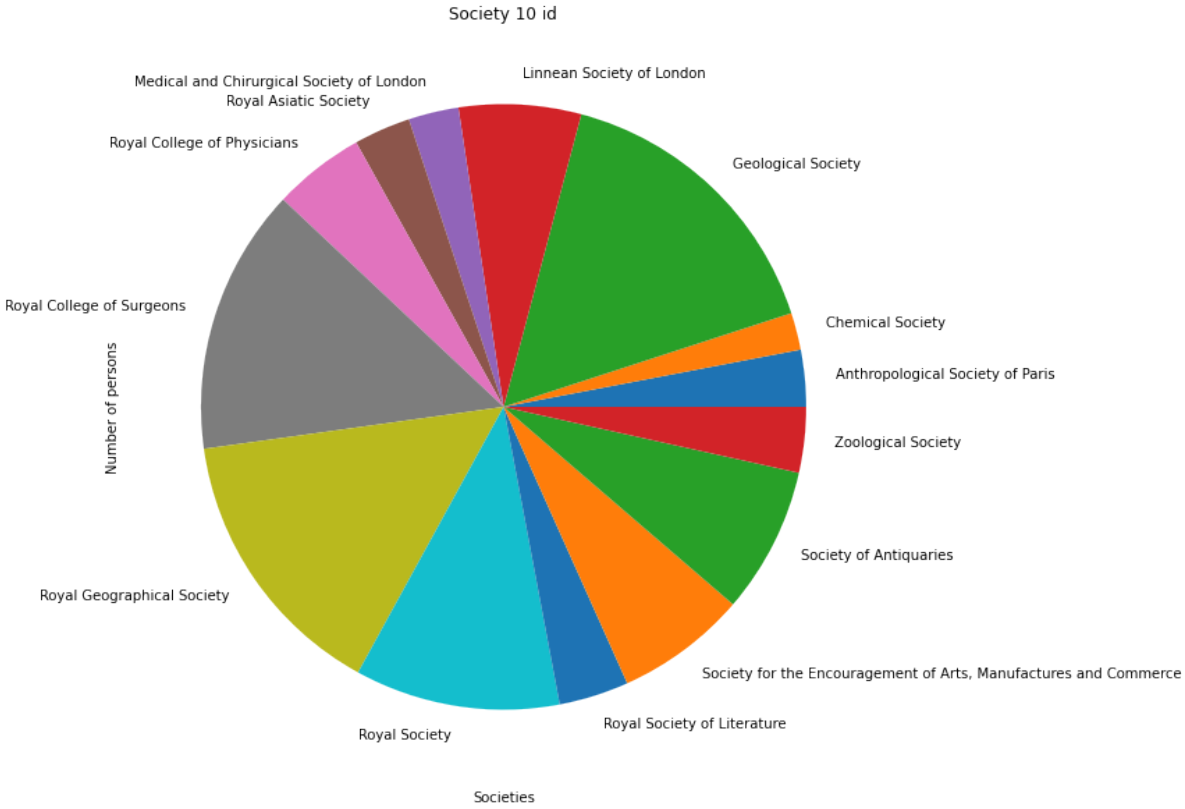
Seven other occupations each have a range of members with literary the lowest and business the highest. All of the satellite groups relate directly to the triple centre with many members also associated with more that one other satellite occupation.

It is surprising the the least networked occupation is 'church' and perhaps less so that 'business' and 'legal'are highly networked.

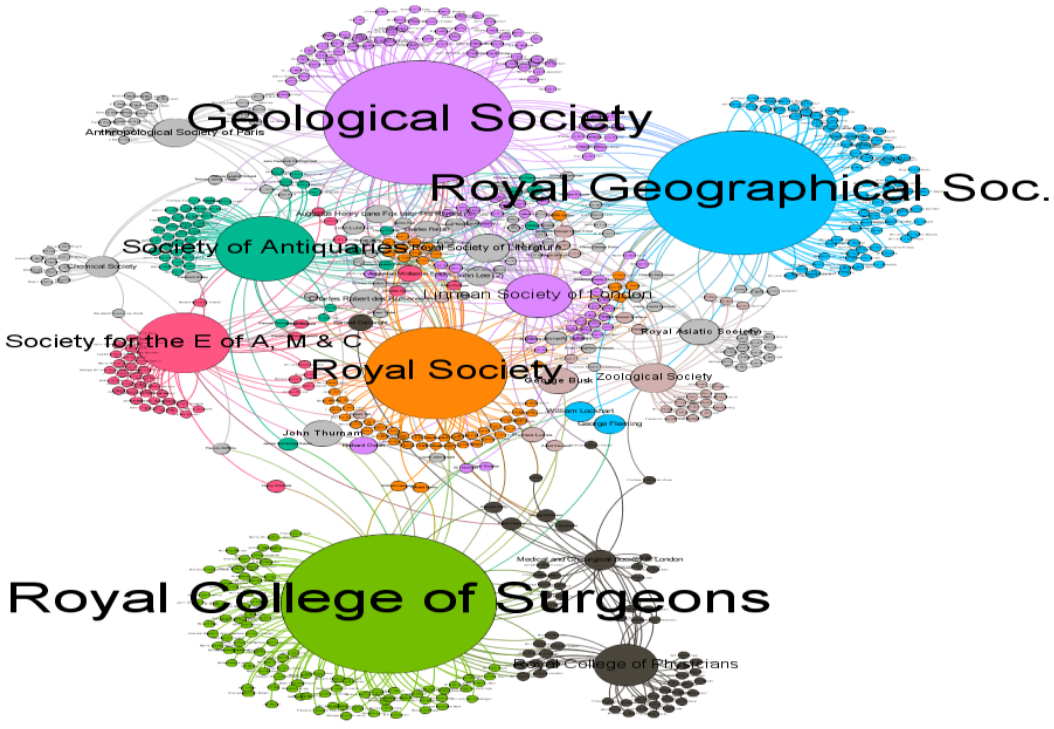
Several individuals form a web of interconnectedness between the members occupations.

Societies

Top 10 societies pie chart



Top 10 societies bigraph



We can see that 'Geological Society' and the 'Royal Goegraphical Society' together account for a significant number of members by society. The 'Royal College of Surgeons', the 'Medical and Chirurgical Society' and the 'College of Physicians' form the next largest cluster of memberships of societies. These two clusters each contain many members who have no association with any other society. We can see that the medical group and the geographical group have few members in common. The 'Royal Society' and the 'Linnean Society' in the centre have between them the greatest level of networking amongst all of the societies. It is highly likely that these

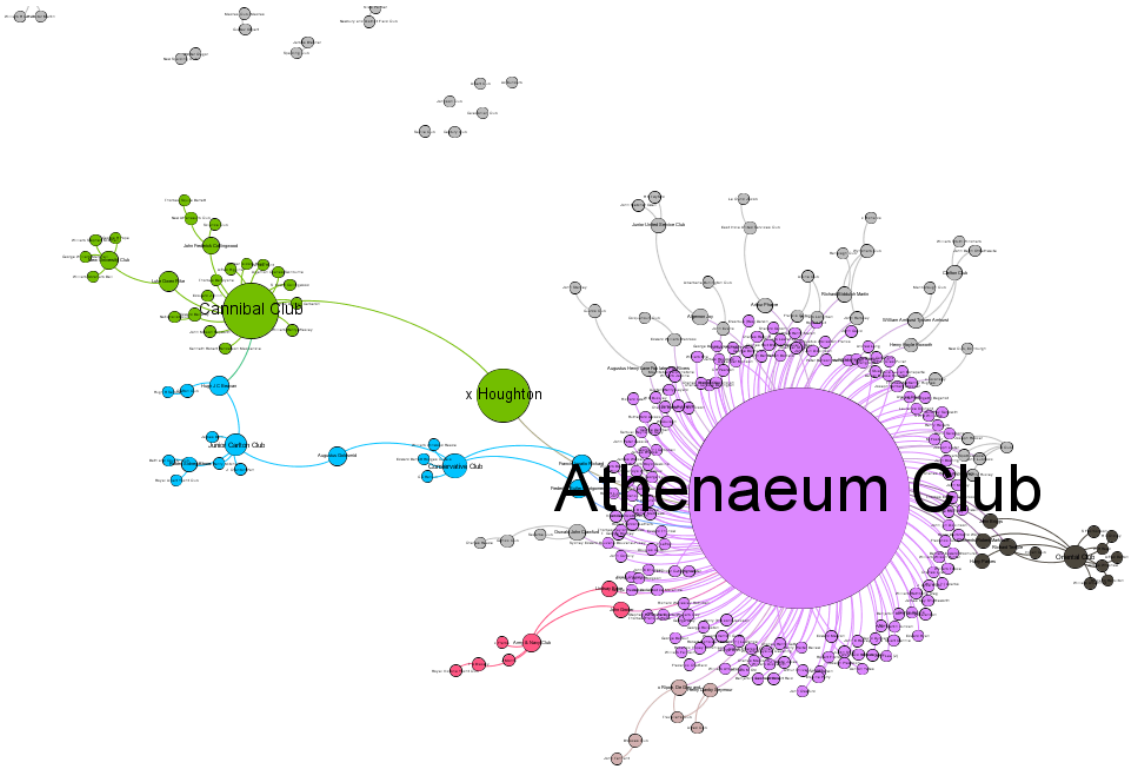
members served as conduits of communication and group cohesion amongst the two principal society groups.

Many other societies have a range of members all of whom are highly interconnected. All of the satellite groups relate most closely to the 'Royal Society' and the 'Linnean Society' rather than to the two larger clusters. Many members of the smaller satellite societies are also associated with more that one other satellite occupation.

It is surprising the the least networked occupation is the 'Royal College of Surgeons' and perhaps less so that the 'Geological Society' and the 'Royal Geographical Society' are highly networked.

Several individuals form a web of interconnectedness between the members of societies.

Clubs



Clubs will not be analysed in this project but the Athenaeum club can be used as an attribute (because it is a singularity).

Most popular bipartite networks combined

1850 members of the community are recorded as members of 35 popular entities (Locations, occupations, societies and the Athenaeum Club). These entities make a sphere of popular interest graph where meetings between members concerning the CEDA may have taken place, equally they may also be places where members might meet up only infrequently or informally. The visual analysis of connectivity between members in single societies and between members of multiple societies indicates the extent that the community is societally connected. The 1850 make up 60% of the entire community.

The graph at the head of this notebook shows the 1850 distributed by popular entity membership with the connectivity between them reflected in those members who are associated with more than one entity.

Iterative Section 1 - (This is an iterative workbook)

As can be seen in the illustrative graph above which has been produced in Gephi by the code cells below to provide an initial overview of the data and its distribution, the graph might be made more meaningful if it did not include societies sparsely populated.

The code cell below and the code cells in the section Iterative Section 2 (below) have therefore been designed so that a second run through the workbook can be made where the second run uses data that excludes low populated occupations identified in the first run through.

In [ ]:

```
# Second we call up the csv files generated from the SQL database that contain infor
# locations and the community members associated with locations. As well as enabling
# we produce a 'node_names' file and a tuples file of edges_attributes to generate t
# produce GefX files for Gephi.

# We can run the code cell twice, first with all data and once all data has been exa
# and a decision made to exclude 'noise' the code block can be run again with newly
# csv files that exclude low populated locations.

#bipartite_10 = pd.read_csv ('vw_bipartite_10_names2_202108041622.csv')
bipartite_pick = pd.read_csv('vw_bipartite_pick_names2_202108051321.csv')

# Use these csv files in the 'with open' statements below to generate bipartite_10.g
#names = pd.read_csv ('vw_bipartite_10_names_1_2_202108041606.csv')# For nodes csv
#tuples = pd.read_csv ('vw_bipartite_10_202108041605.csv')# For edges.csv

# Use these csv files in the 'with open' statements below to generate locations_10.g
bipartite_pick_names = pd.read_csv ('vw_bipartite_pick_names_1_2_202108051304.csv')
bipartite_pick_tuples = pd.read_csv ('vw_bipartite_pick_202108051303.csv') # For edg


with open('vw_bipartite_pick_names_1_2_202108051304.csv', 'r') as nodecsv: # Open th
    nodereader = csv.reader(nodecsv) # Read the csv
    nodes = [n for n in nodereader][1:]# Retrieve the data (using Python list comprh
                                      # to remove the header row
    node_names = [n[0] for n in nodes] # Get a list of only the node names

with open('vw_bipartite_pick_202108051303.csv', 'r') as edgecsv: # Open the file
    edgereader = csv.reader(edgecsv) # Read the csv
    edge_list = list(edgereader) # Convert to list, so can iterate below in for loop

    # Create empty arrays to store edge data and edge attribute data
    edges = []
    edges_attributes = []

    # Fill the arrays with data from CSV
    for e in edge_list[1:]:
        edges.append(tuple(e[0:2])) # Get the first 2 columns (source, target) and a
        # not used this time. edges_attributes.append(tuple(e[2:4]))
        # Get the 3rd and 4th columns (first_year, last_year) and add to array

edge_names = [e[0] for e in edges] # Get a list of only the edge names
```

We begin by listing out and validating all of the popular bipartite (non CEDA) data in the database

In [ ]:
```
# List out the societies to be analysed

# bipartite_10
```

In [ ]:
```
# List out the community members who have been associated with at least one selected

# names
```

In [ ]:
```
# Finally list out the tuples of members and societies
# (Note - some members are associated with more that one society)

# tuples
```

Use pyplot to make an inital visualisation of the data

We can see that many occupations are thinly populated.

We can also see that whilst 'Royal college of Surgeons','Geological Society' and 'Royal Geographical Society' are the largest societal segments several other occupations are well represented.

none of the inital segmentation requires qualification.

In [ ]:
```
tuples.groupby('Target')['Source'].nunique().plot(kind='pie')
plt.title ("Popular_bipartite id")
plt.xlabel ("Group")
plt.ylabel ("Number of persons")
plt.show()
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
<ipython-input-6-b54b19db52e7> in <module>
----> 1 tuples.groupby('Target')['Source'].nunique().plot(kind='pie')
      2 plt.title ("Popular_bipartite id")
      3 plt.xlabel ("Group")
      4 plt.ylabel ("Number of persons")
      5 plt.show()

NameError: name 'tuples' is not defined
```

Iterative Section 2 - prepare the data for rendering as a graph in Gephi

Caution - this section depends on the selections made under 'Iterative Section 1' above

If the initial analysis suggests that a more insightful visualisation might be made by refing the data to be analysed, return to the database and make a new Nodes (Names) csv file and a new Tuples csv file conaining only well populated groups.

Then return to Iterative Section 1 codeblock in the workbook and replace the csv files in the 'with open' code lines with the refined datasets.

Finally reset the nx.write_gexf (xxx.gexf) xxx statement to a new file name.

Then run all code blocks again and make a more insightful gexf file. Use that to produce an improved network graph for Stage 2 analysis.

Warning. - Ensure that the statement 'nx.write_gexf' in the last code cell in this section points to a new output file for Gephi. (eg., G, 'xxxx_10.gexf')Failure to set this value correctly will result in the previously generated .gexf file being overwritten instead.

In [ ]:
```
print("Nodes length: ", len(node_names))
print("Edges length: ", len(edges))
# not used this time. print("Edges attributes length: ", len(edges_attributes)) # Th
```

In [ ]:
```
# First check that the data is correctly formatted

print("First 5 nodes:", node_names[0:5])
print("First 5 edges:", edges[0:5])
# not used this time. print("First 5 edges attributes:", edges_attributes[0:5])

# The output will appear below this code cell.
```

In [ ]:
```
# We use NetworkX to build the graph data into a table

G = nx.Graph()
G.add_nodes_from(node_names)
G.add_edges_from(edges)
print(nx.info(G))
```

In [ ]:
```
# Finally we can write a gexf file which will be placed in the root directory.
# We can then open the file in Gephi and visualise the network.

#nx.write_gexf(G, 'bipartite_pick.gexf')
```

Stage 2 - Bipartite analysis with 'noise' removed (low populated groups excluded).

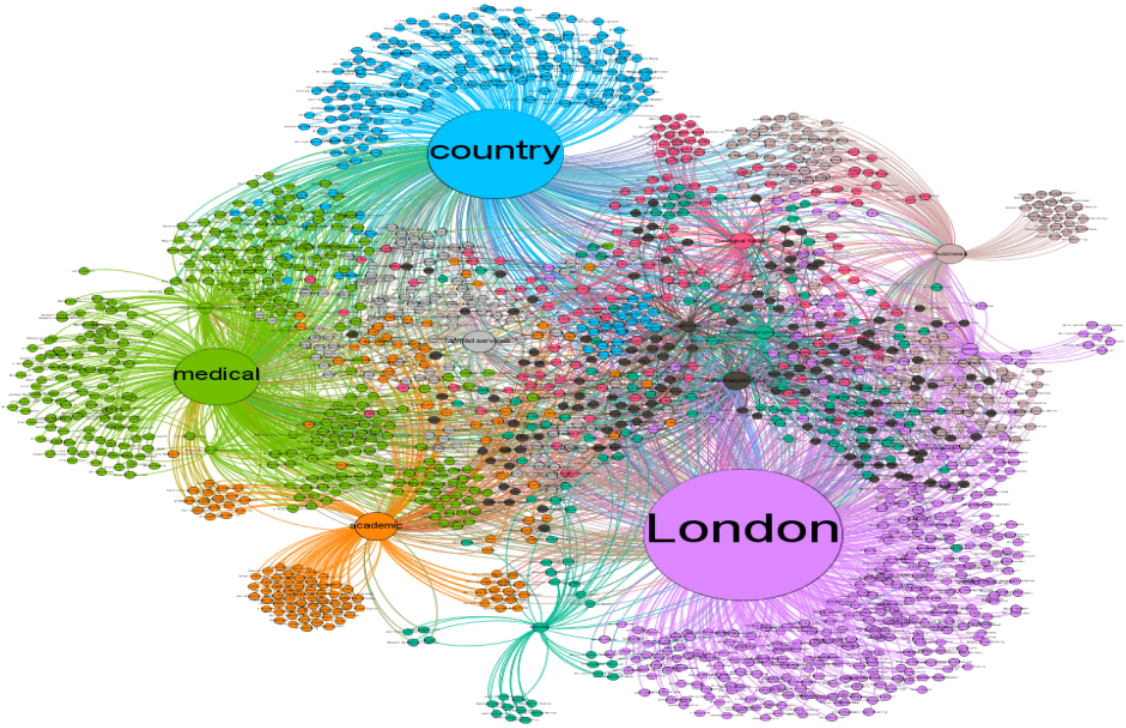
We now re-run the code to generate a new gexf file for gephi. We use the refined pair of nodes (Names) and Tuples files generated in the SQL database that include only the top 17 groups.

In [ ]:
```
bipartite_pick
```

In [ ]:
```
bipartite_pick_tuples
```

We now have a simplified graph of the dataset, but can it be analysed more easily?

In [ ]:
```
bipartite_pick_tuples.groupby('Target')['Source'].nunique().plot(kind='pie')
plt.title ("Bipartite_pick id")
plt.xlabel ("Groups")
plt.ylabel ("Number of persons")
plt.show()
```



The community members most well connected (60%) are densely networked indicating that the CEDA members are able to bring to the task of developing the discipline of anthropology in Britain considerable shared skills, information and knowledge.

END

In [ ]:

In [ ]: