Faculty Hiring Networks

Analysis of Faculty Hiring Networks in Computer Science

Alexandros Frangos

alexandros.frangos.17@ucl.ac.uk

COMP00123 - Complex Networks

2021

Department of Computer Science

University College London

Abstract
The faculty hiring network plays an important role in the prestige of an institution, the educational outcomes, the career paths of individuals and more generally the increase in education standards. Analysing the faculty hiring network, important considerations such as social inequalities including gender discrimination can be detected and universities can take appropriate measures to improve their hiring procedures. This paper uses data of 206 different institutions in the USA and Canada and 4988 people who got their doctorate from institution u and then got a job at an institution v in the field of Computer Science. Using a directed-multigraph network structure, interconnections between different institutions are analysed as well as the importance that the prestige score of an institution plays in both hiring and sending graduates to other institutions. Analysis of the importance of gender is also done and findings concerning the rank position that women tend to have indicates a gender discrimination. [1]

Background and Introduction:
Analyzing the faculty hiring network, important characteristics of institutions, individuals and interactions between those two can be found. The hiring process has an enormous effect on both institutions and individuals. An institution's prestige and educational outcome largely depends on the quality of its professors, therefore hiring the right people is of significant importance for institutions. Equally, individuals who decide to work in academia want to work in a place where they have great career prospects and can show high performance. It is obvious that the hiring procedure plays an important role and therefore fairness measures need to be taken into account to make sure that neither the institutions nor the individuals are disadvantaged.

Research Questions:
The research questions were split among two categories. Below you can see the details.

<u>General Network Properties Research Questions:</u>
1) Which are the top 10 institutions at which most people leave and which are the top 10 universities at which people come? Are there any common universities?
2) What is the relationship between the degree of each node (institution) and the prestige of each institution?
3) Do people end up working at better institutions than the one they got their doctorate from?
4) Are there any universities who collaborate with each other?

<u>Gender Specific Research Questions:</u>
5) Is there any correlation between gender and prestige of the institution?
6) Is there any correlation between gender and rank position?

Literature Survey

“Systematic inequality and hierarchy in faculty hiring networks”

The above paper analysed data of multiple disciplines which are Computer Science, History and Business, located in the USA and Canada. The paper discusses that lower-ranked universities have more difficulty in improving their prestige as they produce a smaller number of faculty candidates who compete for a small number of positions against a large pool of candidates from higher prestige institutions. Therefore, under this definition of prestige, an institution’s prestige is increased by ‘sending’ faculty to higher prestige universities and not by hiring faculty from higher prestige institutions. Moreover an important finding of this paper is that an institution's prestige is highly correlated with the geographical region of the institution. In more detail, institutions in the Northeast are highly over-represented in the high prestige group, whereas institutions in the South are highly under-represented. On average, the high-prestige group includes 23% more institutions than expected in the Northeast, and 22% fewer than expected from the South. The paper relates this finding with two historical facts: (1) the Ivy League institutions are located in the Northeast and (2) the South had fewer resources to invest in higher educational institutions, emphasizing that patterns in current data due to historical events suggests a role for positive feedback mechanisms. [2]

“Gender, Productivity, and Prestige in Computer Science Faculty Hiring Network”

The paper above shows the role that gender plays in faculty hiring networks. The authors of the paper used a dataset on the hiring of assistant professors from 1970 to 2011 and the most important findings are as follows. They found (1) evidence of unexpected gender imbalanced hiring patterns in specific universities, (2) significant differences between men and women that move up ranking in an institution and (3) evidence that higher ranked institutions have more success in hiring females but this may limit lower ranked institutions from having the same success. One limitation of this paper is that the data includes only hiring of assistant professors and not higher ranked professors such as associate or full professors. [3]

“Emergent structures in faculty hiring networks, and the effects of mobility on academic performance”

The above paper uses data concerning the South African Faculty Hiring Network and the purpose of it is to analyse the structure of the network and compare the subsequent scientific performance of scholars with different changes in the prestige hierarchy. An important finding of the paper is that universities that have high in-degree (hiring from many different institutions) tend to place their doctoral graduates to a high number of different institutions. In other words, they found out that the in-degree of an institution is highly correlated with the out-degree of an institution. Moreover they found out that people who work at the same university they got their doctorate exhibit higher performances in the long-term than those who do move up and down the prestige hierarchy. For those who work at a different university than the one they got their doctorate from, their finding suggests that those who got higher prestige doctorates had higher performances in the long term. [4]

Methodology

Software and Programming Languages used

Microsoft Excel - Used to remove all the columns that weren't necessary for the analysis of this network. Excel was also used to convert the .txt file to a .csv one.

Python - The programming language used to analyse the network was Python. Many libraries of python were used which are the following: Pandas for manipulating the dataset, NumPy for executing some mathematical operations, NetworkX for the creation of the graph and the calculation of many properties of the network and webweb which is a library used for interactive network visualisations through the web browser. [5]

Gephi - Gephi was used for the visualisation of the network as it provides some extra functions such as Closeness Centrality which webweb does not.

Data Set Description:

The Dataset was collected between May 2011 and August 2013 for the field of Computer Science. The dataset contains two files. The first one is named edges.csv which has 4989 rows and 4 columns and contains the following columns:

- Source - Institution a person received their doctorate from.
- Target - Institution person end ups working for
- Gender - Gender of the person. It can take the values of male or female
- Rank - The rank of the person that end ups working for the 'Target' institution. It can take the values of Assistant, Associate and Full.

The second file is named nodes.csv which has 206 rows and 4 columns and contains the following information for each node:

- Node Id - Id which is used to identify each node
- NRC95 - Ranking of each institution according to the NRC95 index.

- Pi - Prestige score of each institution which was assigned according to the MVR sampling algorithm

- Region - Geographical Region of each institution.

Network Description:
The two files above were used to create the network. The sources.csv file contains all the directed edges (u,v) denoting a person who received their doctorate from university u and works at university v during the collection period. Each edge also has two weights which are gender and rank. Each node has attributes set to it which are the NRC95, Pi and Region which are described above. Moreover, the graph network is a multigraph, therefore we can have multiple edges for the same pair of nodes. To summarise, a multigraph directed network is created which has 206 nodes and 4989 edges. Below you can find the code snippet (Code Snippet 1) that is used to create the network graph in Python3 using the networkx library.

```
# Reading the files.
edges = pd.read_csv("computer_science_edge.csv")
nodes = pd.read_csv("cs_nodes.csv")
# Creating the network
g = nx.from_pandas_edgelist(edges, source='Source',
target='Target', edge_attr=['gender','rank'],
create_using=nx.MultiDiGraph())
# Setting up the node attributes for each node
nx.set_node_attributes(g, pd.Series(nodes.region,
index=nodes.Node).to_dict(), 'region')
nx.set_node_attributes(g,
pd.Series(nodes.Institution.values,
index=nodes.Node).to_dict(), 'Institution')
nx.set_node_attributes(g, pd.Series(nodes.NRC95.values,
index=nodes.Node).to_dict(), 'NRC95')
nx.set_node_attributes(g, pd.Series(nodes.pi.values,
index=nodes.Node).to_dict(), 'pi')
```

Code Snippet 1 - Creation of the network

Results

<u>1) Which are the institutions at which most people leave (first 10) and which are the universities at which people come (first 10)? Are there any common universities?</u>

To answer this research question I found the 10 institutions which have the highest in degree and the 10 institutions which have the highest out degree. Implementation of the functions is in the Appendix (Code Snippet 2 and Code Snippet 3)

Institution	Prestige	Region
Carnegie Mellon University	9.28	Northeast
Georgia Tech	36.69	Northeast
MIT	3.52	West
University of Waterloo	67.81	West
Indiana University Midwest	116.75	MidWest
Cornell University Northeast	8.29	Northeast
University Of Illinois, Urbana	13.43	Midwest
DePaul University	174.47	Midwest
University of Toronto	24.79	Northeast

Table 1: The 10 Institutions with the highest-in degree

Institution	Prestige	Region
MIT	3.52	West
UC Berkeley	2.31	Northeast
Stanford University	2.23	West
Carnegie Mellon University	9.28	Northeast
University of Illinois	13.43	MidWest
University of Toronto	24.79	Northeast

Cornell University	8.29	Northeast
Purdue University	38.27	Midwest
University of Texas, Austin	19.44	Northeast

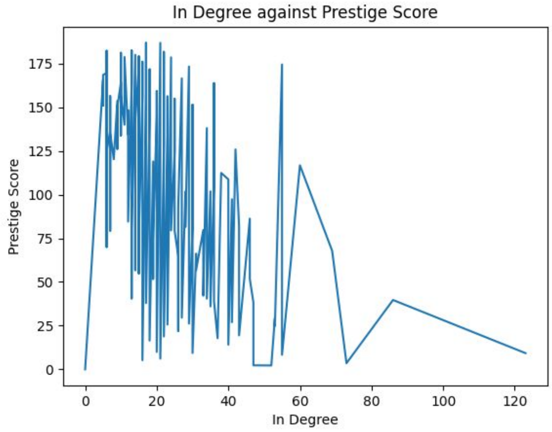
Table 2: The 10 Institutions with the highest-out degree

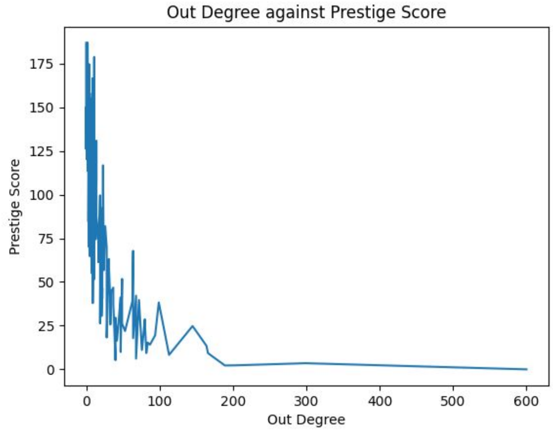
From the results it is seen that 5 out of 10 institutions, more specifically: MIT, University of Illinois, University of Toronto, Carnegie Mellon University, Cornell University appear on both lists. These 5 universities have high prestige, so it is worth finding out if there exists any correlation between the prestige of the university and in/out degree.

<u>2) Is there any relationship between the in/out degree of a university and its prestige?</u>

The outcome from the 2nd research question formed the 3rd research question. To answer this question I created a function that plots the in/out degree of a node against the prestige.The implementation of the function for in-degree/out-degree against prestige is found in the Appendix Code Snippet 4. Below both graphs can be seen (Figure 1, Figure 2) at which it can be seen that there is a moderate negative relationship between out-degree and prestige. On the other hand, the graph of in-degree and prestige is more random, without showing a strong indication of a positive/negative correlation.

The results from this research question were unexpected from my perspective.
I was expecting that institutions with high in-degree institutions would have had higher prestige as more people would like to work at these institutions, and on the other hand high out-degree institutions would have had lower prestige scores as less people would like to work at these institutions. A possible explanation of this, is that people get their doctorate from high-prestige institutions and then choose to work at an institution with less prestige which probably doesn't have as high pressure and demanding environment.





3) Do people end up working at institutions with higher prestige than the one they got their doctorate from?

The purpose of this research question is to see the prestige of the university people end up working for compared to the prestige of the university they got their doctorate from. To do this we calculate the difference in prestige of each edge. If the difference is negative, it means that the prestige of the university a person ends up working for is higher. If the difference is positive, the prestige of the university a person ends up working for is lower. If the difference is zero it means that a person ends up working in the same university that she/he got their doctorate from, as all universities have different prestige scores.

The number of people who end up working for a higher prestige institution is 484 and the number of people who end up working for a lower prestige institution is 4149. The number of people who stayed in the same university is 355. Two graph networks are also constructed, one which shows all the edges with a negative prestige difference, and another one which shows all the edges with positive prestige difference. The table below and the graphs (Figure 3 and Figure 4) show the properties of each network. By comparing the properties of each property we see that the average degree of Graph 2 is much less than Graph 3. This means that the Graph 2 network isn't as connected as the Graph 3 network and there are much less 'exchanges' of people who end up working at a higher prestige university. On the other hand, Graph 3 network has a high average in/out degree and it is a lot more connected. From Figure 3 and Figure 4, it is also notable that the high degree nodes are also the high prestige nodes which confirms the results I got from research question 2. On figures pi, stands for prestige.

	Graph 2 (Negative Prestige Difference)	Graph 3 (Positive Prestige Difference)
# Nodes	144	206
# Edges	484	4149
Avg In-Degree	3.36	20.14
Avg Out-Degree	3.36	20.14

Table 3: Properties of Graph 2 and Graph 3 networks.

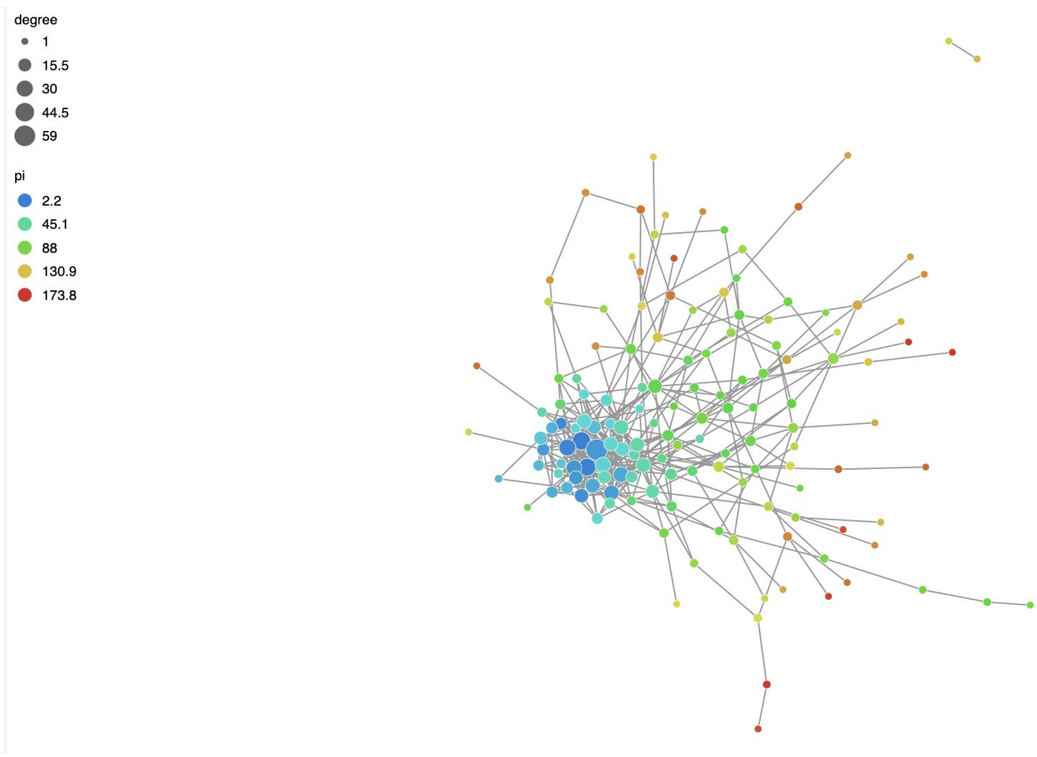


Figure 3: Graph 2 network node sizes scaled by degree coloured by prestige

Figure 4: Graph 3 network node sizes scaled by degree coloured by prestige

4) Do universities collaborate with each other?

This research question tries to find out if some specific institutions collaborate with each other. Collaboration in this question, is defined as the total number of links between two institutions i.e the sum of all edges between two nodes considering the edges as undirected. I first ran a function to get all the institutions which have more than 10 links between them. The table 4 summarises the results along with the prestige of each institution and the absolute value of the prestige difference between the institutions involved. From the table, it can be seen that the total number of institutions is 9. Moreover Stanford University appears 4 times, Carnegie Mellon University 4 times, MIT 3 times, UC Berkeley 3 times, and the University of Georgia, University of Illinois, University of Michigan, UC San Diego and Cornell University just 1 time.

The results give an indication that some institutions tend to send PhD graduates to each other. Moreover it is evident that the institutions who have the highest number of links between them are also high prestige institutions, showing that indeed many high prestige universities tend to collaborate with each other in terms of hiring faculty.. Specifically the average prestige of the 9 universities listed in the table below is 15.1 compared to the total average prestige of all institutions which is 102.5. Implementation of the function used in the Appendix (Code Snippet 5)

Institution A	Institution B	Total Links	Prestige A	Prestige B	Absolute Prestige Difference
Stanford University	MIT	22	2.23	3.52	1.29
UC Berkeley	MIT	20	2.31	3.52	1.21
UC Berkeley	Carnegie Mellon University	20	2.31	9.28	6.97
MIT	Carnegie Mellon University	19	3.52	9.28	5.76
Stanford University	UC Berkeley	17	2.23	2.31	0.080
Stanford University	Carnegie Mellon University	16	2.23	9.28	7.05

Carnegie Mellon University	University of Georgia	12	9.28	39.69	30.41
University of Illinois	University of Michigan	10	13.43	29.37	15.94
UC Berkeley	UC San Diego	10	2.31	27.93	25.62
Stanford University	Cornell	10	2.23	8.29	6.06

Table 4: Institutions which have 10 or more links between them in total.

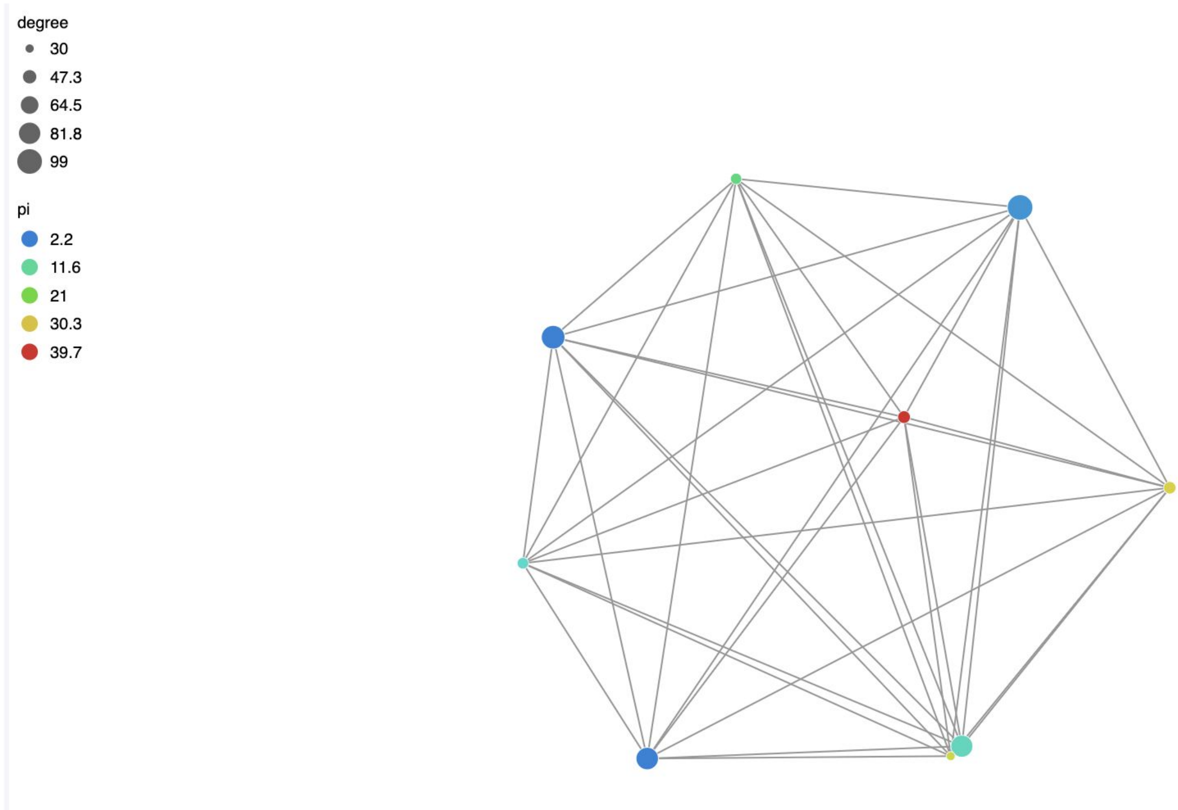


Figure 5: Network Visualisation of the 9 institutions listed in Table 4, scale by degree, coloured by prestige

The absolute prestige difference in 7 of the 10 institutions was less than 10, therefore I wanted to see if the prestige difference and the number of links between two institutions correlate with each other. Figure 9 below shows a scatter plot of the results I got. Figure 9 doesn't show any correlation, as it can be seen that there are institutions that have a low number of links between them and have a low prestige difference, and there are institutions which have a large number of links between them and have a low prestige difference. To conclude, prestige difference doesn't play an important role in the collaboration between institutions on hiring individuals, but it should be noted that in the 10 highest linked institutions the difference in prestige was low.

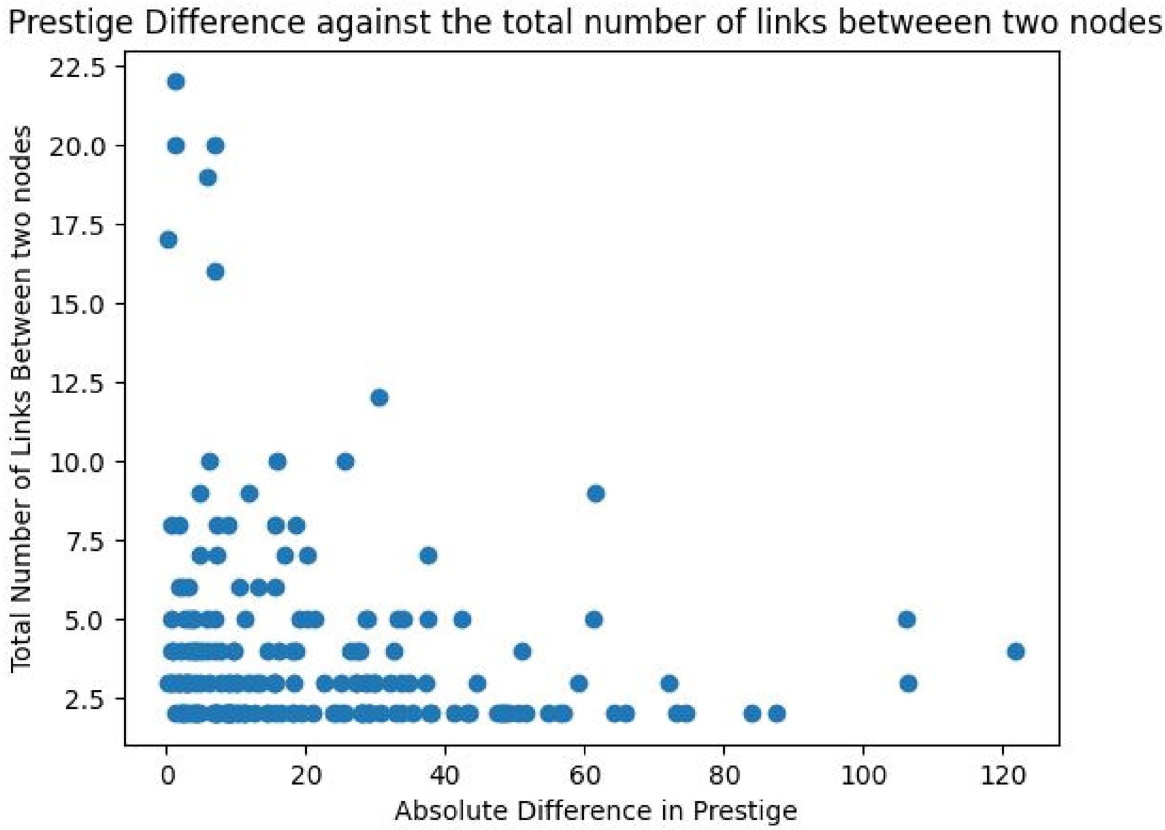


Figure 6: Scatter plot of Prestige Difference against the number of links between two nodes.

<u>5) Are there any notable differences between men and women on the prestige of the institution they end up working for?</u>

To start with, in this network women account for 15% of edges and the other 85% are men, therefore it is obvious that the percentage of men in the computer science faculty hiring network is significantly larger. To answer this question I first found the distribution of the prestige score men and women end up working for. From Figure 7 and figure 8, it is seen that there aren't any statistically significant differences between the two distributions which show that either men or women end up working at higher prestige institutions. By then constructing the visualisation of networks at which first only female edges are shown (Figure 9), and where only male edges are shown (Figure 10) we see that the network of men is a lot more connected with a much larger degree than the women one. Moreover there are 7 institutions at which female edges do not exist. Implementation of function used can be found in the Appendix (Code Snippet 7)

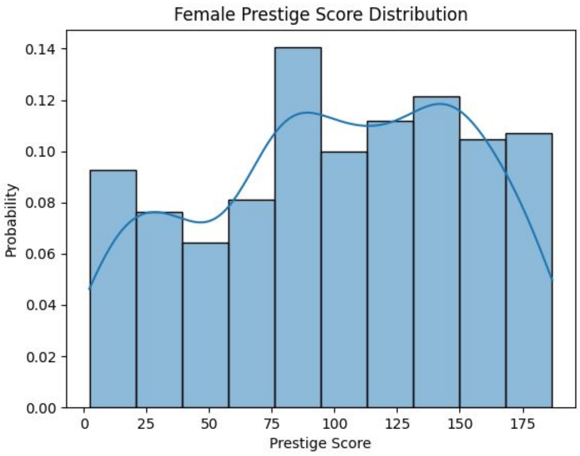


Figure 7: Female Prestige Distribution

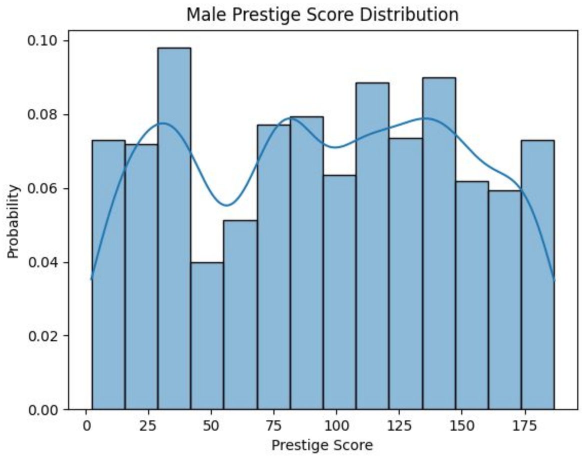


Figure 8: Male Prestige Distribution

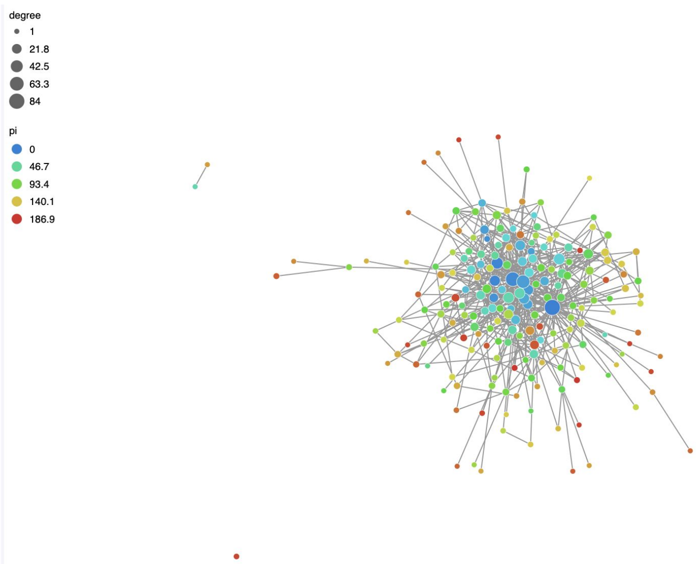


Figure 9: Female Network

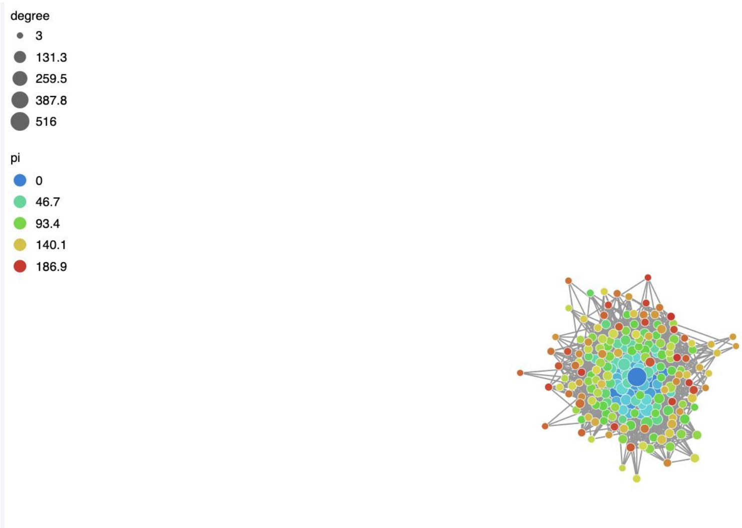


Figure 10: Male Network

I also ran the function used in Research Question 3 on the two graphs (Figure 9, Figure 10) to see if gender plays any role in finding out if people end up working at a higher prestige university than the one they got their doctorate from. The table below summarises the results. It can be seen that data are distributed equally between male and female genders as male in all three cases account on average the 85% and female the rest 15% which is analogous to the distribution of men and women in the original network at which men account for the 85% of edges and females account for the 15% of edges.

	Male	Female	Total
Higher Prestige	426 - 88%	58 - 12%	484
Lower Prestige	3530 - 85%	619 - 15%	4149
Same Prestige	292 - 82%	63 - 18%	355
Total	4248	740	4988

Table 5: Statistics of men and females who work at an institution that has higher,lower and same prestige

6) Is there any correlation between gender and prestige of the institution?

I also analysed if there is any correlation between gender and the rank position a person gets at an institution. The table below summarises the results. One important observation is that most men (47%) are Full Professors (the highest rank) whereas most women (38%) are Assistant Professors (the lowest rank). A large proportion of females end up working as Full Professors but the fact that 38% of women work as Assistant Professors compared to only 16% of men indicates a discrimination against women with the respect to the rank.

	Full	Associate	Assistant	Total
Men	2013 - 47%	1980 - 46%	665 - 16%	4284
Female	273 - 37%	192 - 25%	275 -38%	740
Total	2286	2172	940	4988

Table 6: Statistics of Men and Females who get a job rank of Full, Associate and Assistant Professors.

Discussion

Through this study, a number of interesting results were found. To start with it was shown that high out-degree institutions are also high prestige institutions. In other words, institutions which send their PhD graduates to many institutions are also high prestige institutions. Moreover it was also shown that graduates tend to work at institutions that have a lower prestige than the one they got their doctorate from. By also counting the number of links between each pair of nodes an analysis on institutions collaboration was done. Results showed that the institutions which had the highest number of links between them were also high prestige institutions, although by comparing the data from the whole network I found out that the prestige difference between two institutions doesn't have a correlation with the number of links between two institutions.

The impact that gender has in the faculty hiring network was also analysed. By analyzing the prestige distribution of the universities that both women and men work, it was found that the prestige distributions are similar and the differences found do not have any statistical significance. Moreover, the data showed both men and women tend to go work at a lower prestige institution than the one they got their doctorate from, without major differences. However, by comparing the rank of the job that people get, a significant difference between men and women was found. Specifically, it was found out that 38% of women get the job of Assistant Professors, whereas only 16% of men get the job of Assistant Professors. Adding to this, 47% of men get the job of Full Professors whereas 37% of women get the job of Full Professor. That is evidence that gender does play role when compared with the rank of the position which indicates some degree of gender discrimination.

Overall, I believe that the project has produced a number of interesting insights but it also had some limitations, which could be an incentive for future work. To start with, other network properties could be computed and analyzed. Examples of metrics are betweenness centrality and closeness centrality, although for the purposes of this project I believe that in and out degree were more suitable to use. Moreover, data of only the Computer Science Faculty Network were analysed and not from other disciplines which also have publicly available data such as Business and History. By analyzing those data as well, similarities could be found between different university disciplines as well as differences.

Bibliography:

1) Aaronclauset.github.io. 2015. Faculty Hiring Networks. [online] Available at: <https://aaronclauset.github.io/facultyhiring/>
2) Clauset, A., Arbesman, S. and Larremore, D., 2015. Systematic inequality and hierarchy in faculty hiring networks. Science Advances, 1(1), p.e1400005.
3) F Way, S., Clauset, A. and Larremore, D., 2016. Gender, Productivity, and Prestige in Computer Science Faculty Hiring Networks.
4) Cowan, R. and Rossello, G., 2018. Emergent structures in faculty hiring networks, and the effects of mobility on academic performance. Scientometrics, 117(1), pp.527-562.

References:

5) Webwebpage.github.io. 2021. webweb - webweb. [online] Available at: <https://webwebpage.github.io/>

Appendix

```
def highestIn(g):
      degrees = sorted(g.in_degree, key=lambda x: x[1],
reverse=True)
      degrees = degrees[:10]

      highest_nodes = [i[0] for i in degrees]

      institution = nx.get_node_attributes(g,'Institution')
      region = nx.get_node_attributes(g,'region')
      prestige = nx.get_node_attributes(g,'pi')
      for i in highest_nodes:
            print(institution[i], region[i], prestige[i])
```

Code Snippet 2: Implementation of the highestIn function

```
def highestOut(g):
      degrees = sorted(g.out_degree, key=lambda x: x[1],
reverse=True)
      degrees = degrees[:10]

      highest_nodes = [i[0] for i in degrees]

      institution = nx.get_node_attributes(g,'Institution')
```

```
      region = nx.get_node_attributes(g,'region')
      prestige = nx.get_node_attributes(g,'pi')
      for i in highest_nodes:
            print(institution[i], region[i], prestige[i])
```

Code Snippet 3: Implementation of the highestOut function

```
def InDegreePrestigeComparison(g):
      degrees = sorted(g.out_degree, key=lambda x: x[1],
reverse=True)

      nodes = [i[0] for i in degrees]
      degree = [i[1] for i in degrees]
      prestige_list = []

      for node in nodes:
            prestige = nx.get_node_attributes(g,'pi')
            prestige_list.append(prestige[node])

      fig = plt . figure ()
      ax = fig . add_subplot (111)
      ax.plot(degree,prestige_list)
      ax.set_xscale ('linear')
      ax.set_yscale ('linear')
      plt.title("In Degree against Prestige Score")
      plt.xlabel("In Degree")
      plt.ylabel("Prestige Score")
      plt.savefig("In Degree against Prestige Score")
      plt.show()
```

Code Snippet 4: Implementation of inDegreePrestigeComparison function. Same function was used to calculator the out degree vs prestige.

```
def institutionsCollaboration(g):
      source = []
      target = []
      prestige = nx.get_node_attributes(g,'pi')
      name = nx.get_node_attributes(g,'Institution')
      region = nx.get_node_attributes(g,'region')
```

```
    for u,v,a in g.edges(data=True):
          source.append(u)
          target.append(v)


    edges = list(zip(source,target))
    counter=Counter(edges)
    edges2 = {}

k = list(counter.items())
    for i in k:
          for j in k:
                if i[0] == tuple(reversed(j[0])) and i[0] != j[0]:
                      prestige_difference = abs(prestige[i[0][0]]
- prestige[i[0][1]])
                      #edges2[i[0]] = i[1] + j[1]
                      edges2[prestige_difference] = i[1] + j[1]

                      k.remove(j)

    p_diff = []
    la = list(edges2.items())
    for i in la:
          p_diff.append(i[1])

    items = sorted (edges2.items())


    fig = plt . figure ()
    ax = fig . add_subplot (111)
    ax.scatter([ k for (k , v ) in items ] , [ v for (k ,v ) in
items ])
    plt.xlabel(' Absolute Difference in Prestige')
    plt.ylabel('Total Number of Links Between two nodes')
    plt.title("Prestige Difference against the total number of
links betweeen two nodes ")

    fig.savefig ( "Uni Collaborationsns.png" )
    plt.ylim(0, 24,4)
    plt.show()
```

Code Snippet 5: Implementation of the function institutionsCollaboration. The function was manually changed to test other properties as well outlined in the report.

```
# Plots histogram of gender with prestige of the university women
and men end up working for.
def gender(g):
      count_females = 0
      count_males = 0

      ranking_female = []
      rank_female = []
      ranking_male = []

      ranking=nx.get_node_attributes(g,'pi')

      for u,v,a in g.edges(data=True):
            if a['gender'] == 'M':
                  count_males+=1

                  ranking_male.append(ranking[v])
            elif a['gender'] == 'F':
                  count_females+=1

                  ranking_female.append(ranking[v])


      sns.histplot(ranking_female,stat='probability')
      plt.title("Female Location Distribution")
      plt.xlabel("Prestige Score")
      plt.savefig("Female - Location - Histogram")

      plt.show()


      ranking_male = np.array(ranking_male)
      ranking_female = np.array(ranking_female)

      return count_males, count_females
```

Code Snippet 6: Implementation of gender function, that plots the distribution of the prestige of the institutions that men and women are working. The function was manually changed to test some other properties as well.