Group Project

Group 2

Group members

Jiang Jane: S5356970
Lei Pan: S5179270
Li Yang: S5315573
Zhou Marina: S5340985

1. Motivation

Sometimes there are stylistic differences between pieces of literature that are written by authors from a certain period, a certain region, etc. Ted Underwood, Kevin Kiley, Wenyi Shang, and Stephen Vaisey (2022) have found that the birth year of authors explains better than the published year of literature in their corpus of 10,830 works of fiction from 1880 to 1999. One of the assignments we have done this semester shows that the published year is the main factor influencing the stylistic relations between 150 English novels published between 1771 and 1930. We are interested in the situations in other corpus and want to find out the factors there if there are stylistic differences between those pieces of literature.

2. Goal of the project

We aim to analyse the factors that affect the stylistic differences of Colonial South Asian English Literature.

3. Research question

Are there stylistic differences between authors of different ethnicities writing in English during the time of the British Colonial Empire?

4. Dataset

We decide to use the Colonial South Asian Literature corpus from Amardeep Singh's dataset (2020) which contains books published between 1853 and 1923.

We have made some changes to the dataset. First, the dataset includes 110 pieces of literature in total while the metadata of this corpus only has information on 101 pieces, so we delete those which are not in the metadata. Then we add other possible attributes that may affect the stylistic differences like gender, birth year, and the original language. Finally, South Asia is one of the ethnicities of the authors in the metadata. We change it to Indian and Bangladesh so that it will be more clear when comparing the ethnicities of the authors.

5. Tools

(1) Stylo

First, we will use Stylo to make cluster analyses in the corpus. We may not use the visualizations in the output since the corpus contains 101 pieces of literature. But Stylo will help us get a rough understanding of the stylistic relationship between the literature. One of the outputs of Stylo is the edges file that shows all the relationships - mainly the weight - between the literature.

(2) Gephi

We will use Gephi to create network visualizations. The edges file from Stylo will be used as input for Gephi. Besides, we will also use a node file with the metadata from Amardeep Singh's dataset (2020).

Different networks will be plotted when using different attributes for nodes. Some networks might not show stylistic differences if there are no clusters or trends at all. Some may have the main factors while some only show part of the reasons. From these networks, we could find out whether there are stylistic differences between 101 pieces of Colonial South Asian literature, and what affects them most if there are any.

6. Result

We have compared six factors in total, including the published year, the original language, gender, birth year, ethnicity, and genre.

(1) The published year

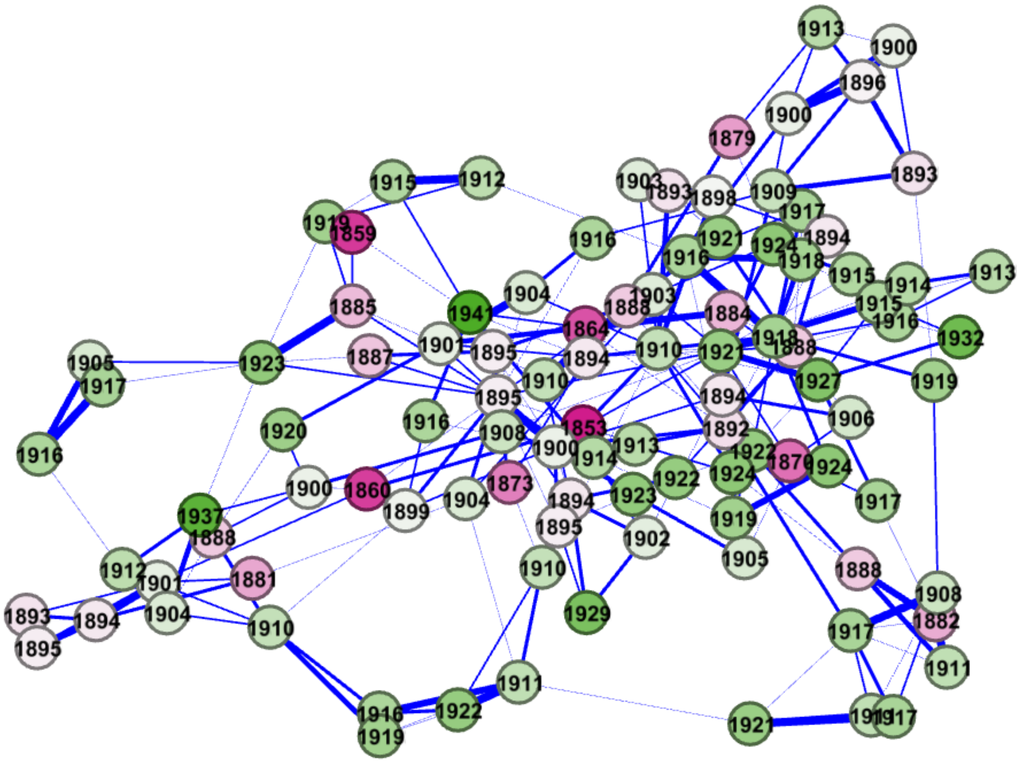

Figure 1. Colonial South Asian English Literature network by the published year

Figure 1 is a network of the 101 pieces of literature from the Colonial South Asian English Literature Corpus categorized by the published year. The redder the node is, the earlier the literature has been published. The greener, the latter. For the blue edges, a broader edge means a stronger connection between the two pieces of literature and vice versa. The same goes for the following figures.

At first glance, we can easily tell that the published year as an attribute does not show a trend. There is no clear tendency from red nodes to green nodes, so the published year is not the main factor that affects the stylistic differences between the literature.

(2) The original language

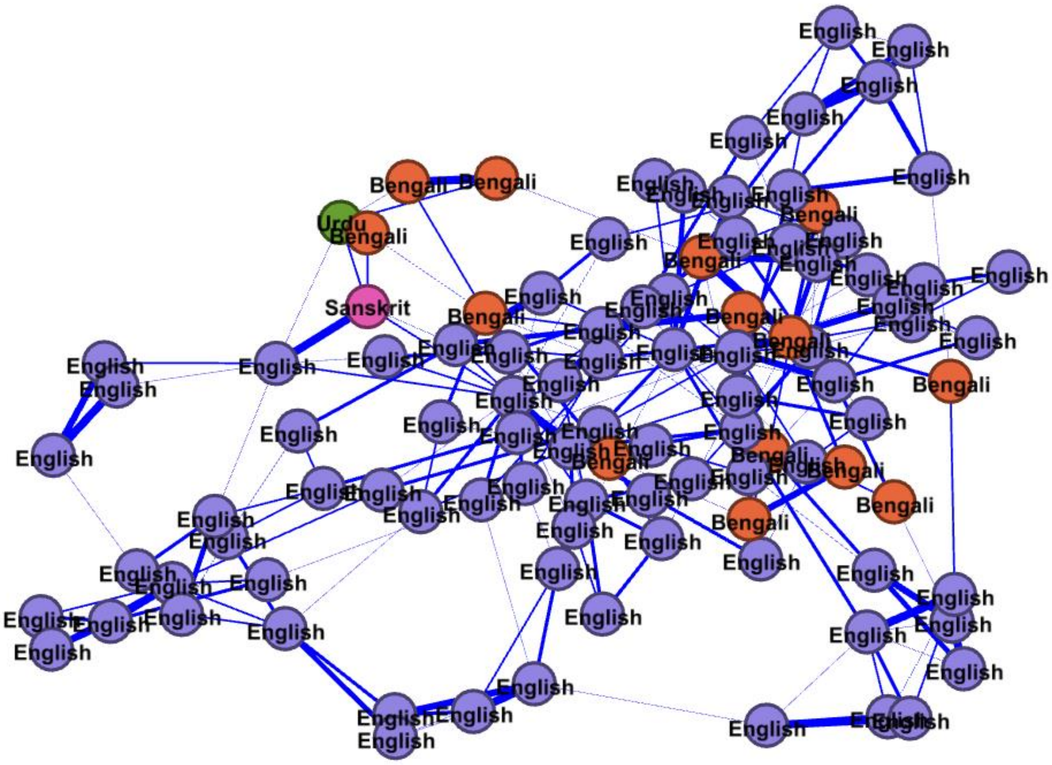

Figure 2. Colonial South Asian English Literature network by the original language

Even though all the literature in the corpus is English, some of them are written in other languages and translated later. When we check whether the original language has any influence on the stylistic difference, we will get Figure 2. Here we can see that the majority of the literature is written in English - the purple nodes, and others do not stay together as a group, for instance, Bengali pieces of literature - the orange nodes - scatter in the picture. Thus the original language is not the factor that we are looking for.

(3) Gender

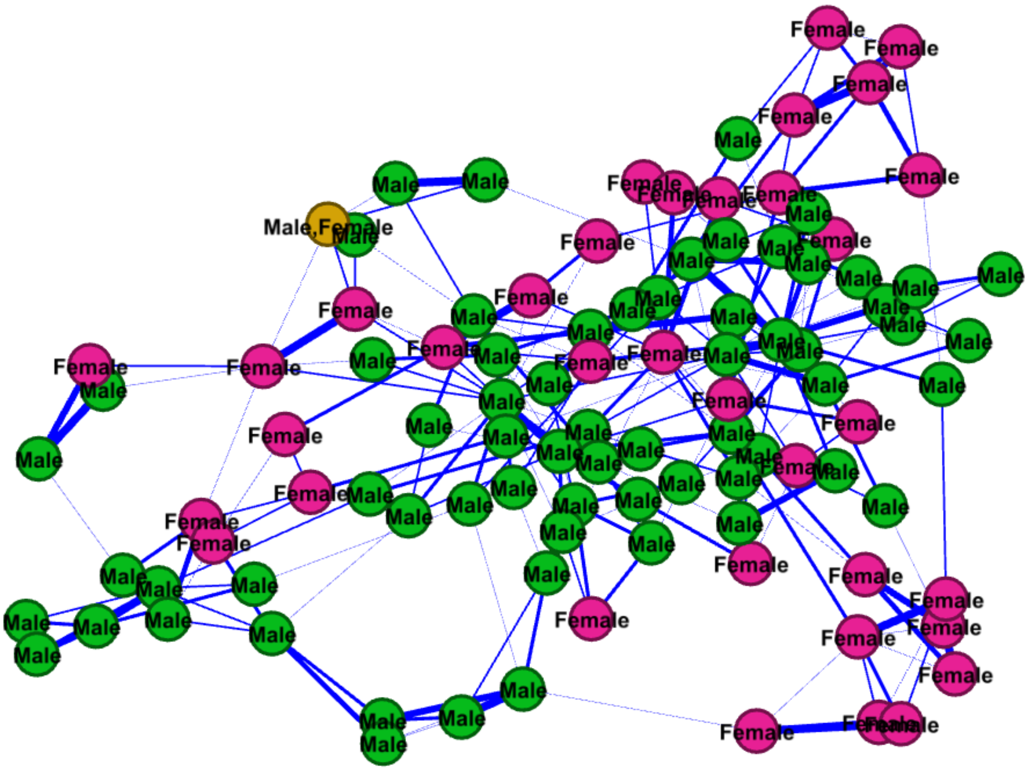

Figure 3. Colonial South Asian English Literature network by gender

The gender of the authors is the third parameter we are going to analyse. Just like the situation in Figure 2, from Figure 3 we can see that most authors are male - the green nodes, with one piece of literature written by a male author and a female author - the yellow node. Since female authors are in everywhere rather than clustering together, gender is not the one that influences the stylistic difference either.

(4) The birth year

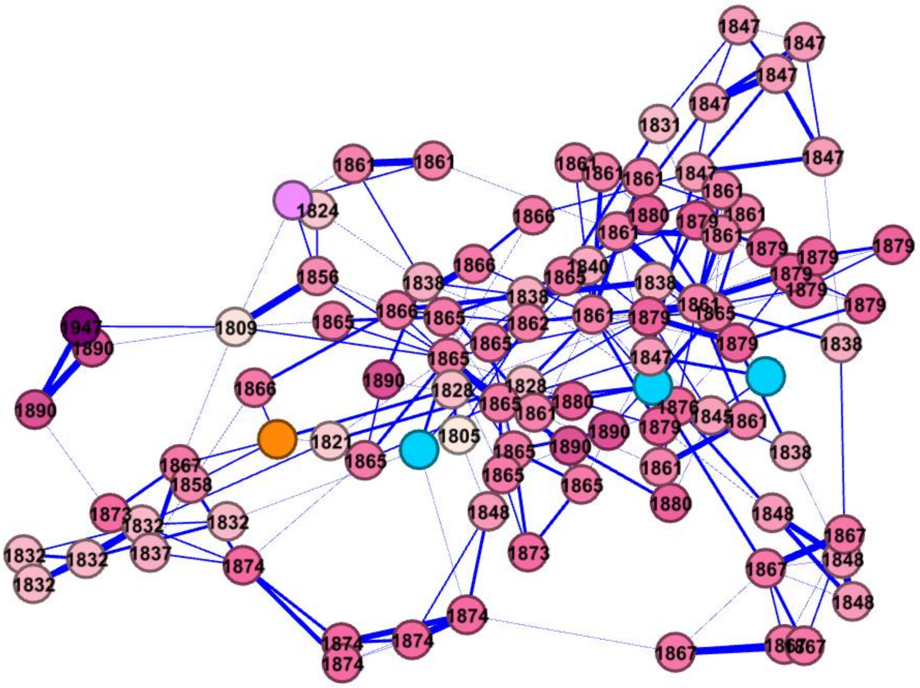

Figure 4. Colonial South Asian English Literature network by the birth year

Then we check the birth year of the authors in the corpus. The redder the node is, the later the author was born. Some nodes are not in other colors, since we could not find the exact birth year of some authors or there are two years due to two authors. From Figure 4, however, we can see that there is no clear trend from light red to dark red. So the birth year does not have much influence there.

(5) Ethnicity

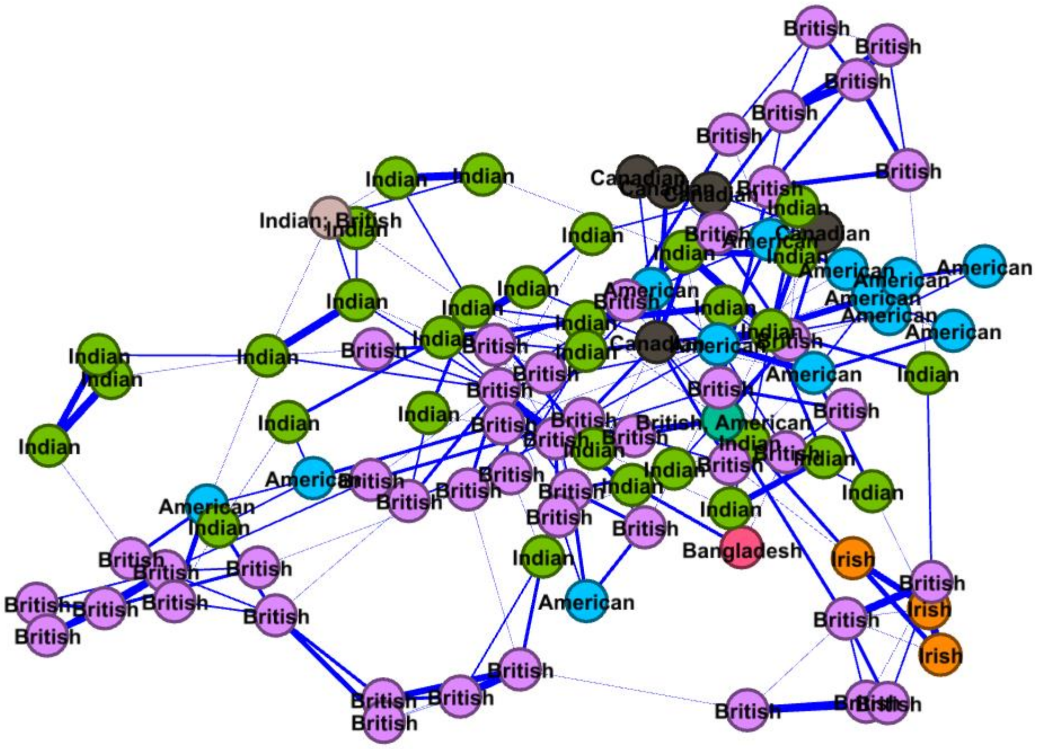

Figure 5. Colonial South Asian English Literature network by ethnicity

Ethnicity could be a reason that leads to stylistic differences between these authors. As we can see from Figure 5, most Indian authors - the green nodes - stay in the northern part and the western part, American authors - the blue nodes - almost cluster together in the eastern part, all the Canadian authors - the brown nodes - are in the northeast part, and all the Irish authors - the orange nodes - are in the southeast part. The only outlier is British authors - the pink nodes, they are the major authors in the dataset and are scattered in every part. This means that British authors have stylistic similarities with authors from different backgrounds. Even though ethnicity could not explain all the stylistic relationships in this network, it is an important factor.

(6) Genre

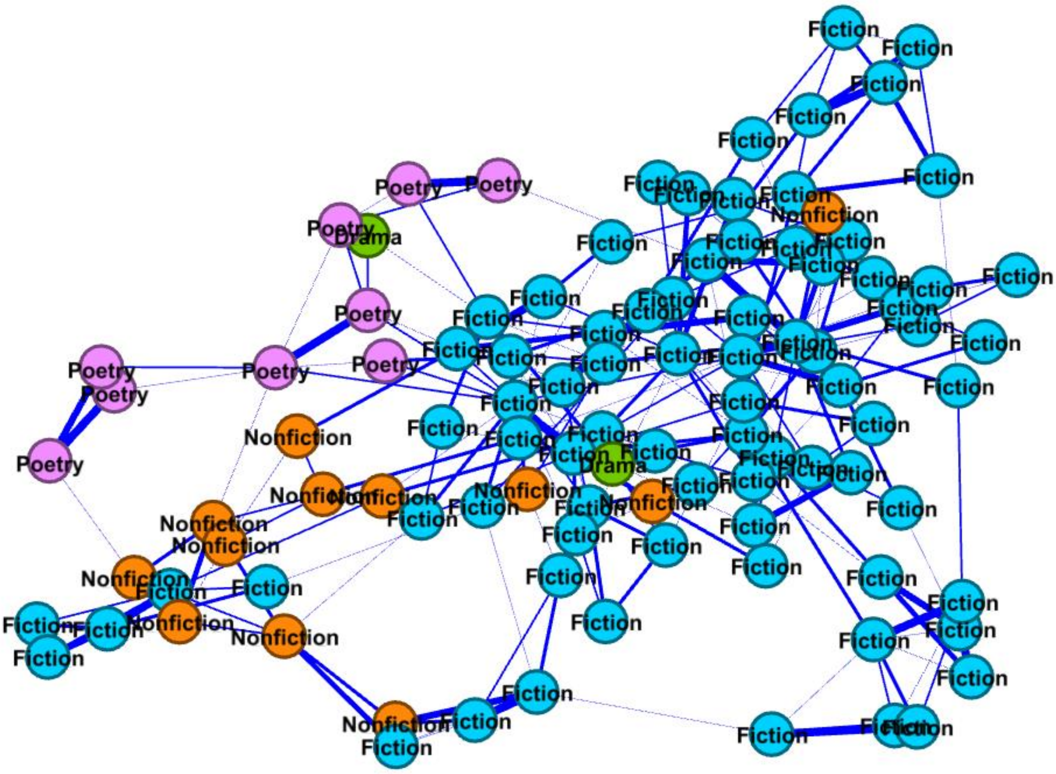

Figure 6. Colonial South Asian English Literature network by genre

We could get Figure 6 when using genre as a variable. Here we can easily tell that there are three clusters - fiction, nonfiction, and poetry. The majority of fiction - the blue nodes - is in the central part and eastern part, most nonfiction - the orange nodes - stays in the southwest part, and all the poetry - the pink nodes - clusters together in the northwest part. The two dramas - the green nodes - are not far from each other, although they do not stay very close. Unlike the factors we have analysed above, there are clear boundaries of these three clusters with only a few exceptions. It is clear now that there are stylistic differences between these pieces of literature and genre is the main factor causing these differences.

1) Ethnicity in genre

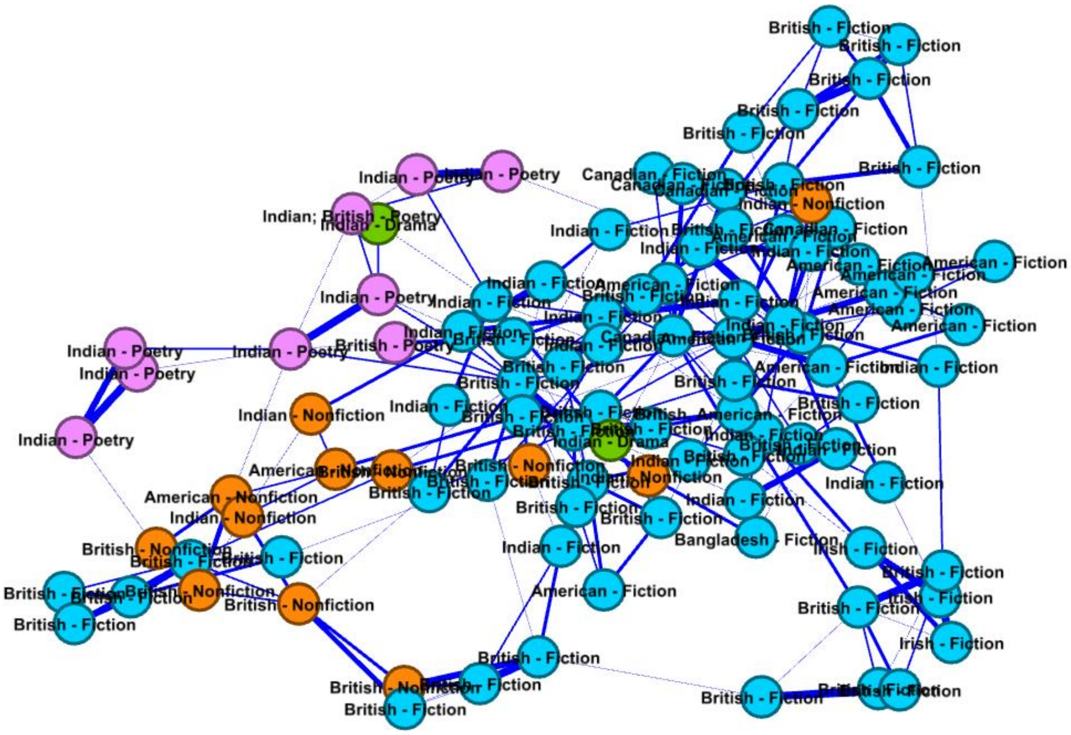

Figure 7. Colonial South Asian English Literature network by genre

We can take a step further after looking at the genre. In Figure 7, the blue, orange, pink, and green nodes still represent different genres: fiction, non-fiction, poetry, and drama, it is just that the attributes on the nodes are not just genres but also ethnicity. We can tell that ethnicity can also tell something here. British authors write most fiction, half of the nonfiction, and only one piece of poetry. Almost all the poetry and half nonfiction are from Indian authors. Apart from some fiction, American authors only write two nonfiction, with no poetry or dramas. Authors from Canadian and Irish only show interest in fiction.

2) Ethnicity in fiction

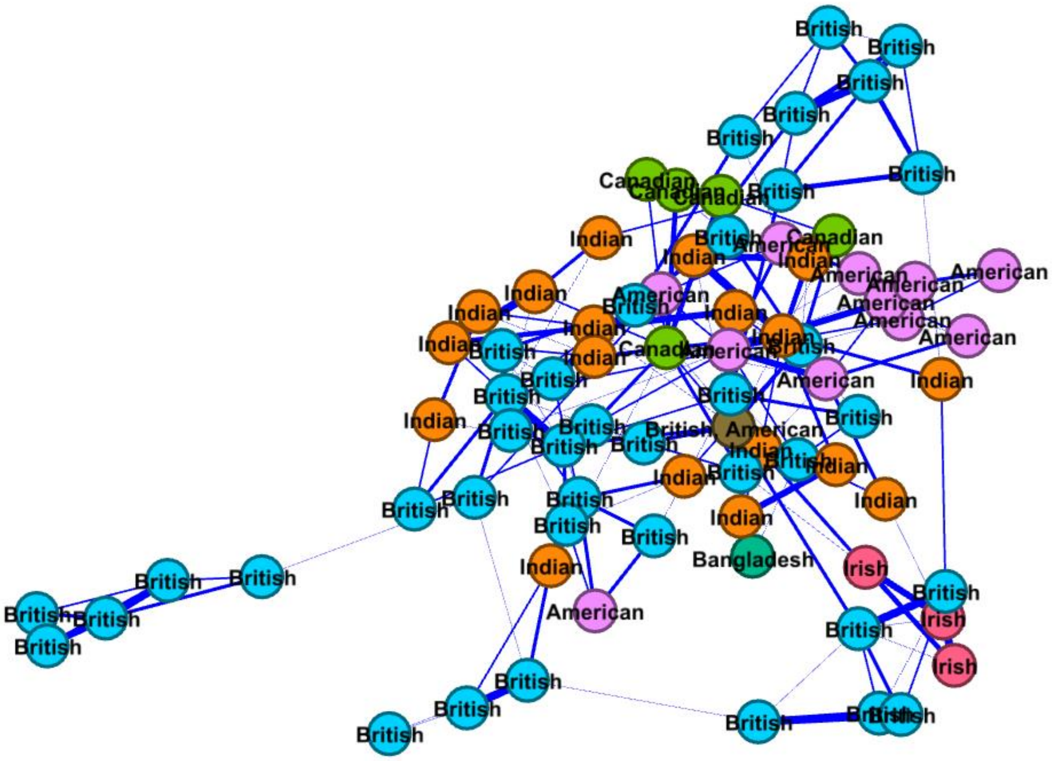

Figure 8. Colonial South Asian English Fiction network by ethnicity

Since the majority of the literature in this dataset is fiction, we can only choose to check ethnicity in fiction. It is shown in Figure 8 that ethnicity influences the stylistic relationship in fiction. The four clusters are Indian authors - the orange nodes, American authors - the pink nodes, Canadian authors - the green nodes, and Irish authors - the red nodes. The boundaries are pretty clear between them. The only outlier here is British authors. They write most of the fiction and they have similar writing styles to authors from all other ethnicities.

3) The birth year in fiction

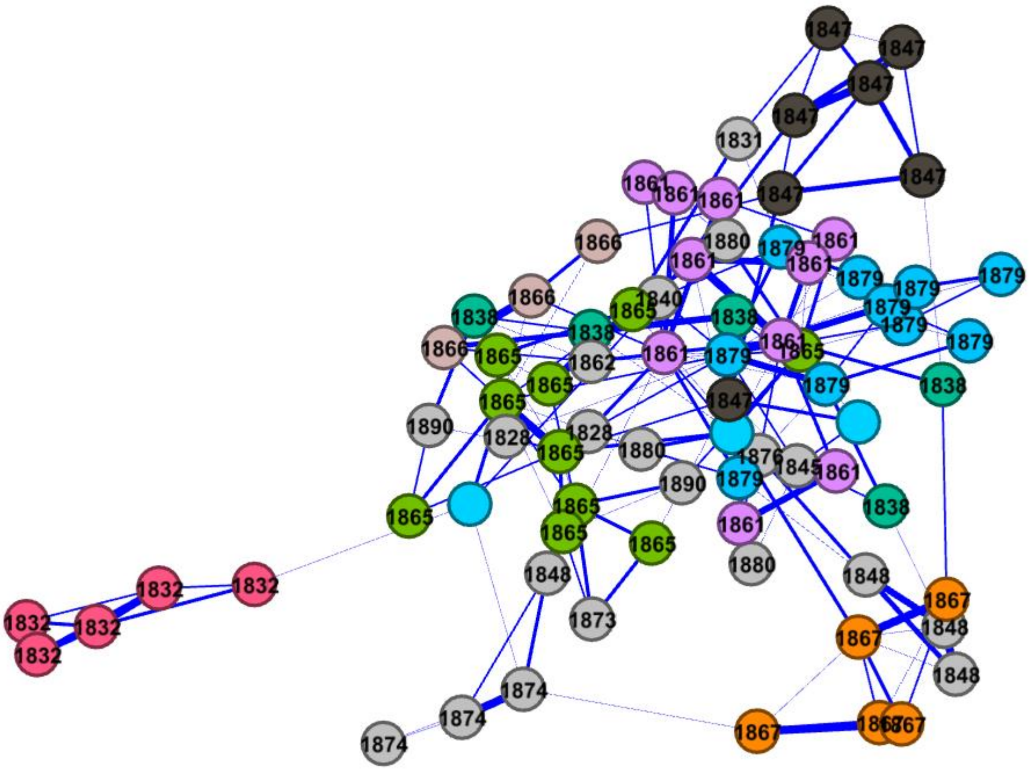

Figure 9. Colonial South Asian English Fiction network by birth year

Similarly, we could also analyse the birth year in fiction. We can tell from Figure 9 that there are many clusters, and authors of the same birth year are always in the same cluster. For example, the red nodes presenting authors born in 1832 are all in the southwest part and far away from others, while most of the authors born in 1847 cluster in the northeast part. This is to say that the birth year of the authors also plays an important role in affecting the stylistic differences in fiction

7. Conclusion

To conclude, there are stylistic differences between authors of different ethnicities writing in English during the time of the British Colonial Empire. Among all the factors, the genre is the most important factor that affects stylistic differences. In the same genre, ethnicity and birth year are the next things causing the differences.

Reference:

Underwood T, Kiley K, Shang W, Vaisey S. (2022, May). Cohort Succession Explains Most Change in Literary Culture. *Sociological Science.*