U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

CAC PROJECT

Community-Based Recommendations

Group B
António Ferreira - up202004735
João Maldonado - up202004244
Tomás Gomes - up202004393

Part #1

CAC 23/24

DATASET FILTERING AND CLEANING

New Orleans

The original Yelp dataset had information regarding **multiple cities in the USA**.
With this in mind, we filtered the data to only include the **business**, **users** and **reviews** regarding **New Orleans**.

Removal of Data

Although it was filtered, the dataset still had **too much unnecessary information**. Therefore, in order to clean the data, we did the following steps:

Removed Non-Existing Friends + **Removed Users with No Friends** **18945313** users → **14555** users

Removal of Users with < 2 reviews **14555** users → **13769** users

Removed Reviews of Non-Existing Users + **Removed Reviews of Non-Existing Businesses** **635364** reviews → **89515** reviews

Removed Business with No Reviews + **Removed Closed Business** **6209** business → **4649** business

Removed Reviews of Closed Business **89515** reviews → **71944** reviews

EXPLORATORY ANALYSIS

To better understand the dataset, it’s essential to explore its features. With this in mind we decided to draw the following plots:

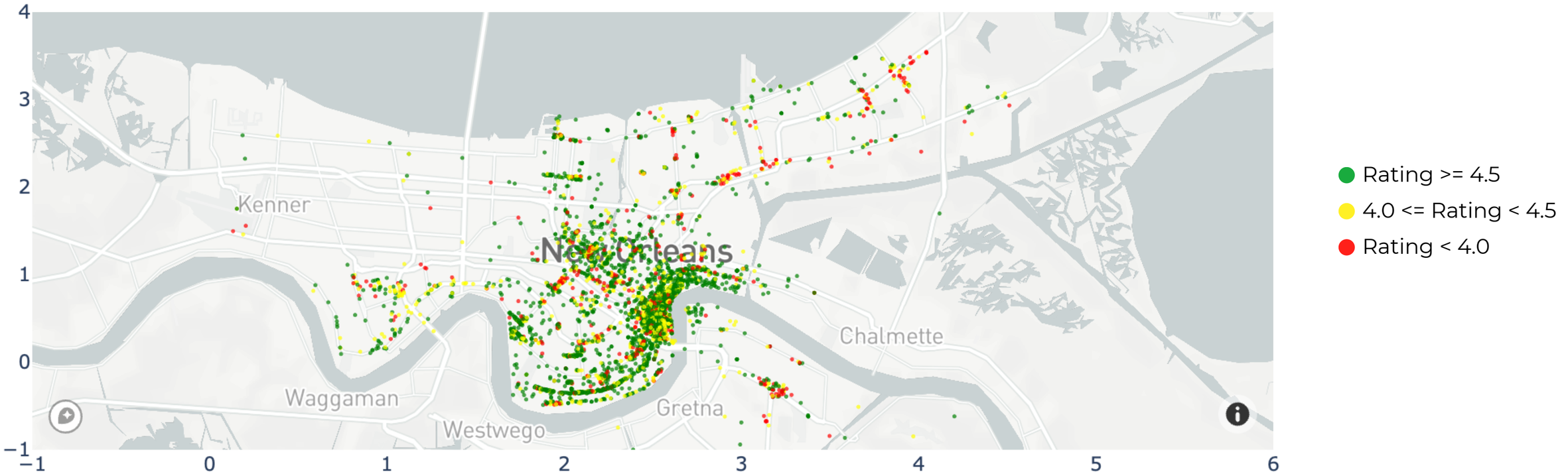


fig 1 - Geographical Distribution of the Business in New Orleans

fig 2 - Most common business categories

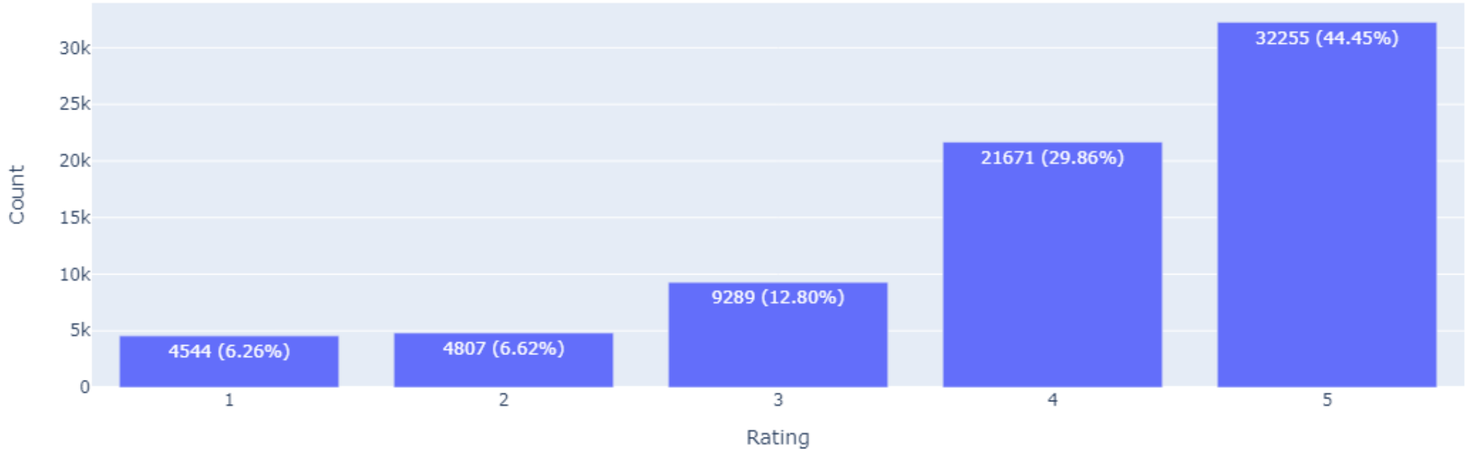
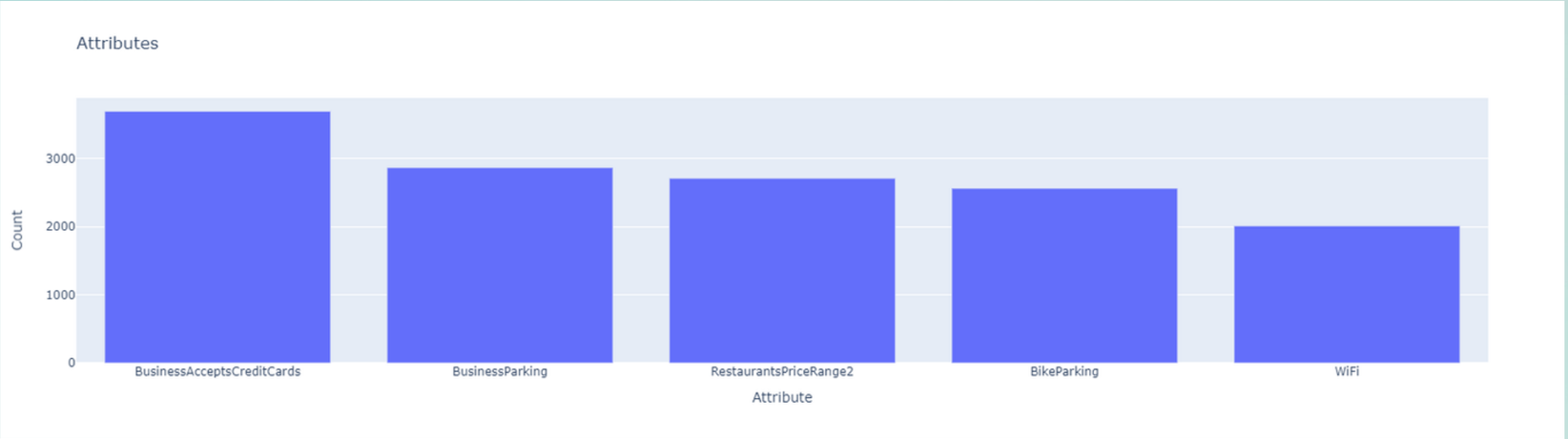


fig 3 - Most common review rating

EXPLORATORY ANALYSIS

EXPLORATORY ANALYSIS



Graph 4 - Top 5 most frequent business attribute

This dataset includes Elite users who are members of the Yelp Elite Squad, a yearly exclusive membership for the most active Yelp users in the community. Regarding the number of Elite users, there are 3125 in total.

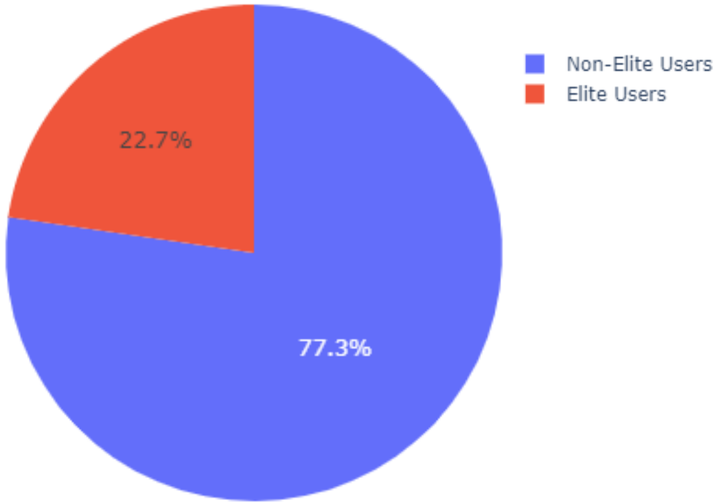


fig 5 - Percentages of Elite and Non-Elite Users

Total # Reviews by Elite Users **26029**

Total # Reviews by Non-Elite Users **45875**

SOCIAL NETWORK ANALYSIS

The first main step in creating user communities, based on SNA, was creating User graphs, **where each node represented a User of the dataset.**

Regarding the **edges**, we used two approaches to connect the Users:

Friendship between Two Users / **Friendship between Two Users** **OR** **>= 5 Reviews in Common Businesses, with Strong Rating Similarity**

NOTE

In order to analyze the **similarity between the rating two users gave to a business**, we used the following method:

- **For each user, there's an array of the ratings they gave to a certain business. This array only contains the common businesses between the two users being analyzed.**
- **We compute the absolute difference between each index of the array.**
- **Finally, we calculate the mean of the differences. If the value is <0.5, then there's a strong similarity between the users' ratings.**

FRIENDSHIP APPROACH

3179 Edges

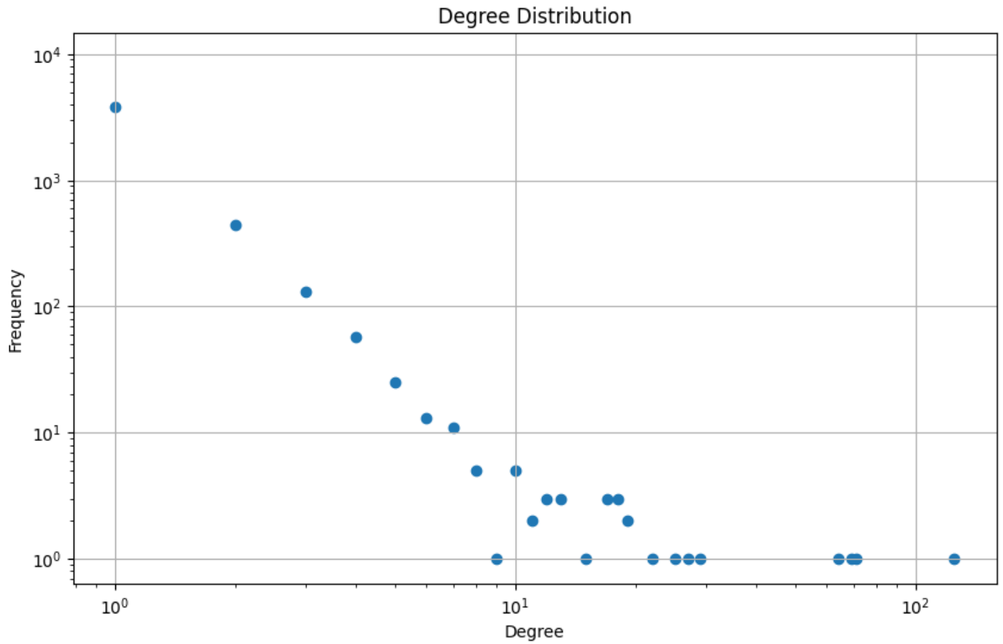


fig 6 - Degree Distribution of the graph where the users are connected by the friendship approach

SIMILARITY APPROACH

4897 Edges

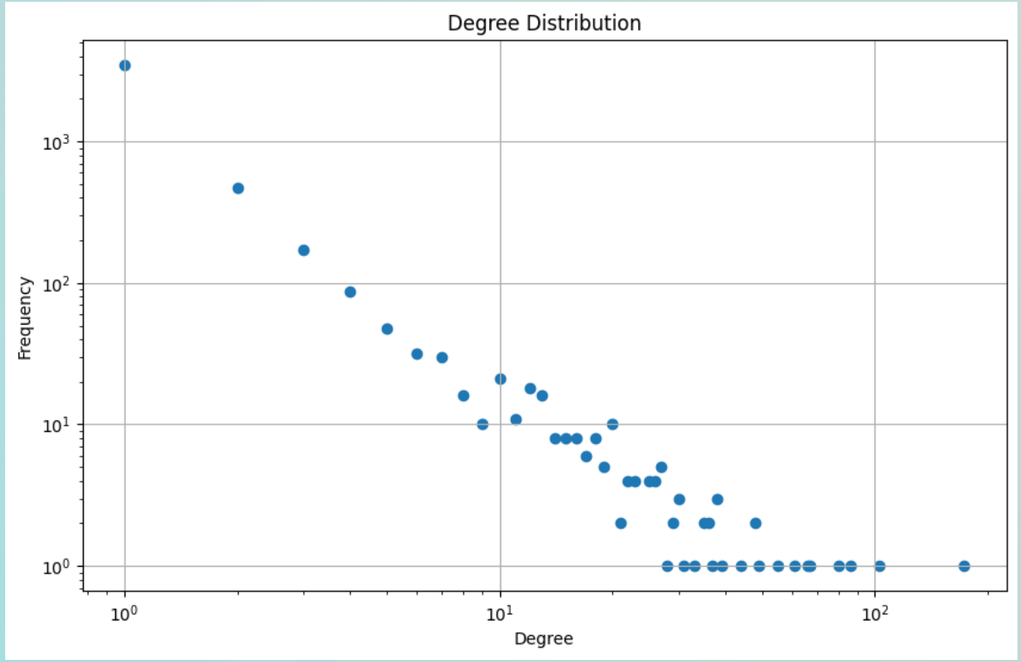


fig 7 - Degree Distribution of the graph where the users are connected by the similarity approach

- **The second graph has more nodes within the range of degree values of 100-200 and has less nodes with only 1 edge (degree = 1). This suggests that the second graph can have an increase of centrality and more density.**
- **There is also less isolated nodes, which suggests a more cohese and less fragmanted graph.**

FRIENDSHIP APPROACH

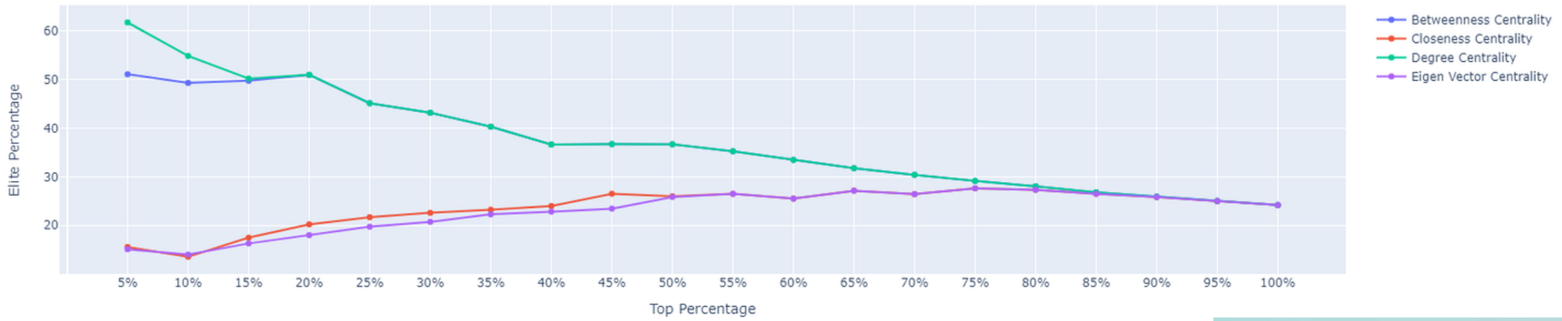


fig 8 - Analysis of the Centrality Behavior of the elite users in the friendship approach

SIMILARITY APPROACH

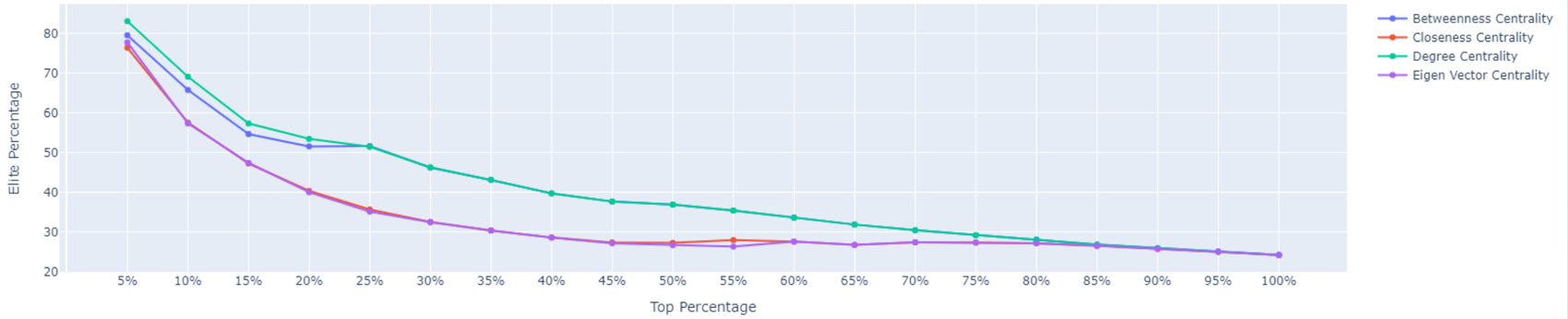


fig 9 - Analysis of the Centrality Behavior of the elite users in the similarity approach

The charts show that the similarity technique **significantly increases** the number of elite users in nodes with the **highest proximity centrality** and **eigen vector centrality**.

- This observation shows that elite users are now **more centrally situated** in the network, allowing for **faster and more efficient connections to other nodes**. It also suggests that elite users are connected to **other critical nodes in the network**.
- This demonstrates the importance of top users in influencing other users on the network. **Elite users can exert significant influence over network dynamics** and **interactions** because they are **strategically positioned and well-connected**.

FRIENDSHIP APPROACH

	Community	Density	Assortativity-Review-Count	Assortativity-Elite	Number of Nodes
0	0	0.0142	-0.2458	-0.6756	141
1	1	0.0180	-0.1278	-0.3108	111
2	2	0.0222	-0.1852	-0.4878	90
3	3	0.0227	-0.1864	-0.5190	88
4	4	0.0541	-0.1846	-0.2125	37

fig 10 - Density and Assortativity (Review Count and Eliteness of users) for the **friendship** approach

SIMILARITY APPROACH

	Community	Density	Assortativity-Review-Count	Assortativity-Elite	Number of Nodes
0	0	0.0026	-0.0089	-0.2774	513
1	1	0.0064	-0.0403	-0.1667	250
2	2	0.0088	-0.1083	-0.6145	195
3	3	0.0093	-0.0531	-0.3384	173
4	4	0.0215	-0.1527	-0.6279	90

fig 11 - Density and Assortativity (Review Count and Eliteness of users) for the **similarity** approach

Community Structure

The community sizes and densities presented suggest varying levels of cohesion and engagement. As it is shown in the tables, smaller communities tend to be more densely connected, which might indicate more intensive interaction among their members.

User Engagement

The **negative assortativity** regarding review counts, in both approaches, across all the communities, suggests that **more active users are connected to less active ones**. This could be essential to spread information about the businesses to all the members in a community. However, in the **similarity approach**, we can see an **increase** in the **assortativity** values, regarding the review count, across all communities. This could be attributed to the social nature of friendships, where individuals might connect with others who have different interests or activity levels. For instance, a user who actively reviews might befriend someone who rarely writes reviews but **shares other common interests.**

This way, in the second approach, we are aggregating users with similar taste or similar activity within communities, which can lead to an increase of recommender systems results.

FRIENDSHIP APPROACH

	Community	Density	Assortativity-Review-Count	Assortativity-Elite	Number of Nodes
0	0	0.0142	-0.2458	-0.6756	141
1	1	0.0180	-0.1278	-0.3108	111
2	2	0.0222	-0.1852	-0.4878	90
3	3	0.0227	-0.1864	-0.5190	88
4	4	0.0541	-0.1846	-0.2125	37

fig 10 - Density and Assortativity (Review Count and Eliteness of users) for the friendship approach

SIMILARITY APPROACH

	Community	Density	Assortativity-Review-Count	Assortativity-Elite	Number of Nodes
0	0	0.0026	-0.0089	-0.2774	513
1	1	0.0064	-0.0403	-0.1667	250
2	2	0.0088	-0.1083	-0.6145	195
3	3	0.0093	-0.0531	-0.3384	173
4	4	0.0215	-0.1527	-0.6279	90

fig 11 - Density and Assortativity (Review Count and Eliteness of users) for the similarity approach

Elite Status

The increase in assortativity from the friendship approach to the similarity approach, regarding the elite status, indicates a potentially beneficial aspect of the network structure. Even though the assortativity remains negative, indicating that nodes with different elite statuses are more likely to connect, this can be advantageous for fostering a community with diverse tastes but varying levels of elite status.

Also, this interconnectedness between elite and non-elite users facilitates the exchange of opinions, recommendations, and influences regarding businesses. Elite users may introduce non-elite users to businesses they haven't reviewed before (and vice-versa).

Overall, this dynamic promotes a more inclusive and diverse community where users of different elite statuses can interact, share experiences, and influence each other's choices. It encourages collaboration and mutual enrichment, contributing to a vibrant and dynamic network ecosystem.

FRIENDSHIP APPROACH

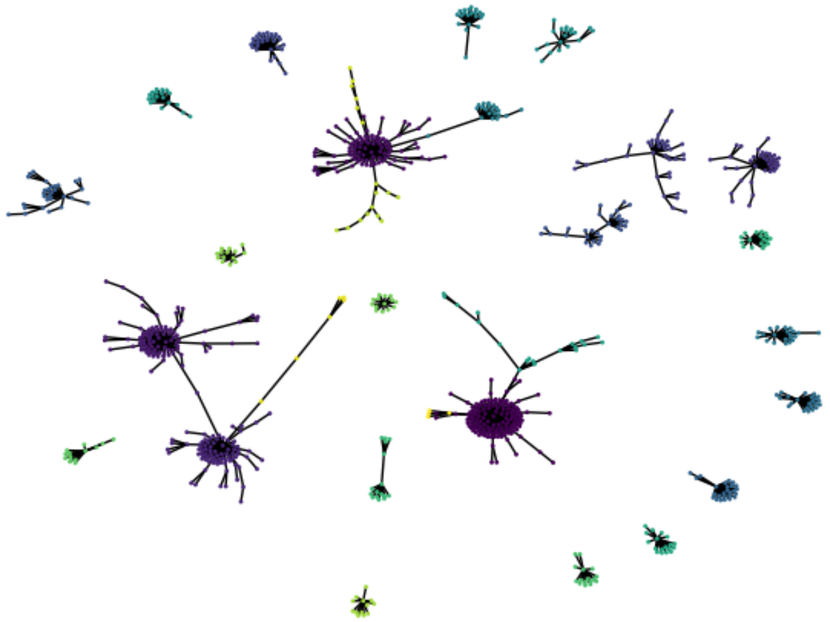


fig 12 - Induced **subgraph** computing communities for top 10 users with most degree, using **friendship approach.**

When more edges were introduced, the **clarity of community split worsened rather than improved.** This effect could be given to the **enhanced connection** caused by **the additional edges**. As more connections are made between nodes across different communities, the boundaries between these communities become less apparent.

SIMILARITY APPROACH

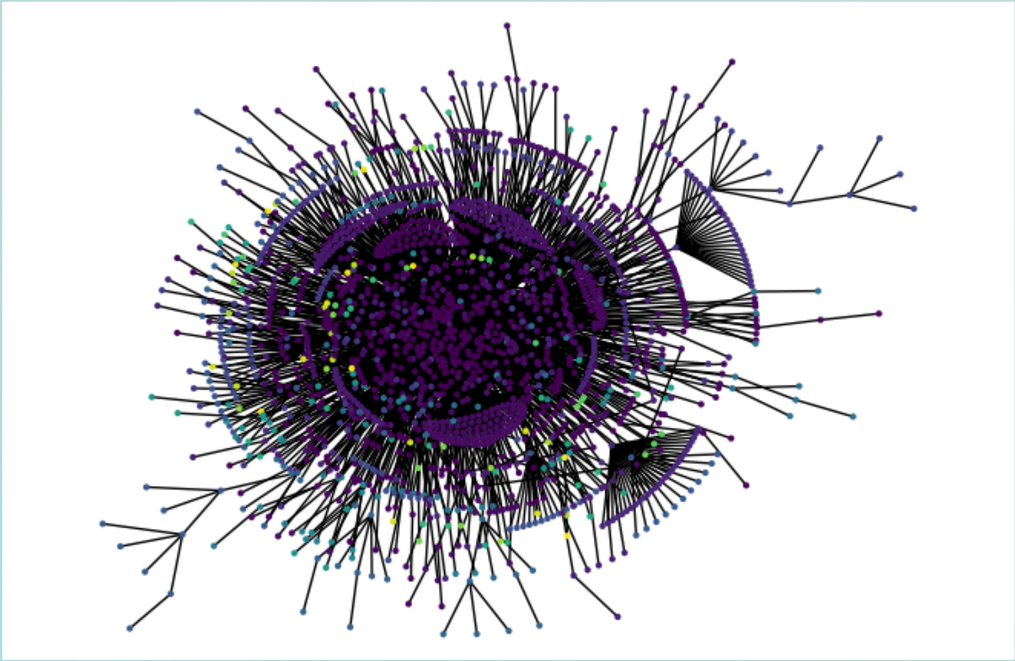


fig 13 -Induced **subgraph** computing communities for top 10 users with most degree, using **similarity approach.**

Increased connectedness can result in a **more cohesive network**, with nodes from diverse groups being more closely linked. As a result, distinguishing between distinct communities becomes **more difficult**, as nodes may have **many connections that straddle different community boundaries**.

FRIENDSHIP APPROACH

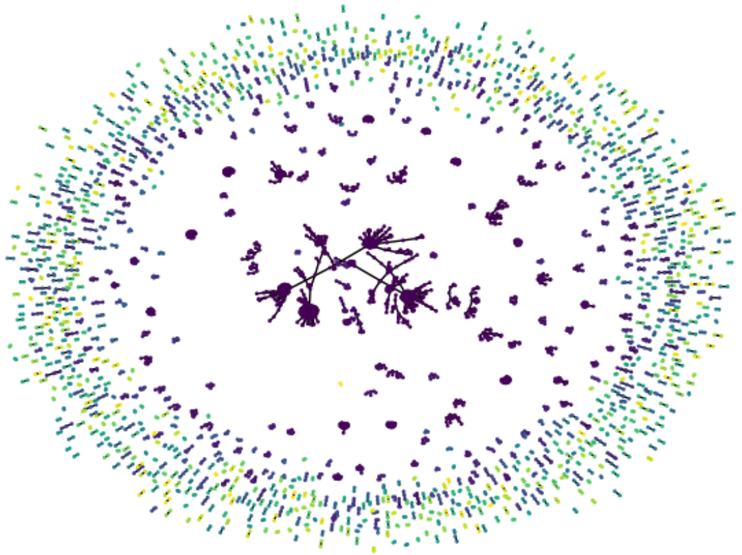


fig 14 - Whole graph where each color represents a community

SIMILARITY APPROACH

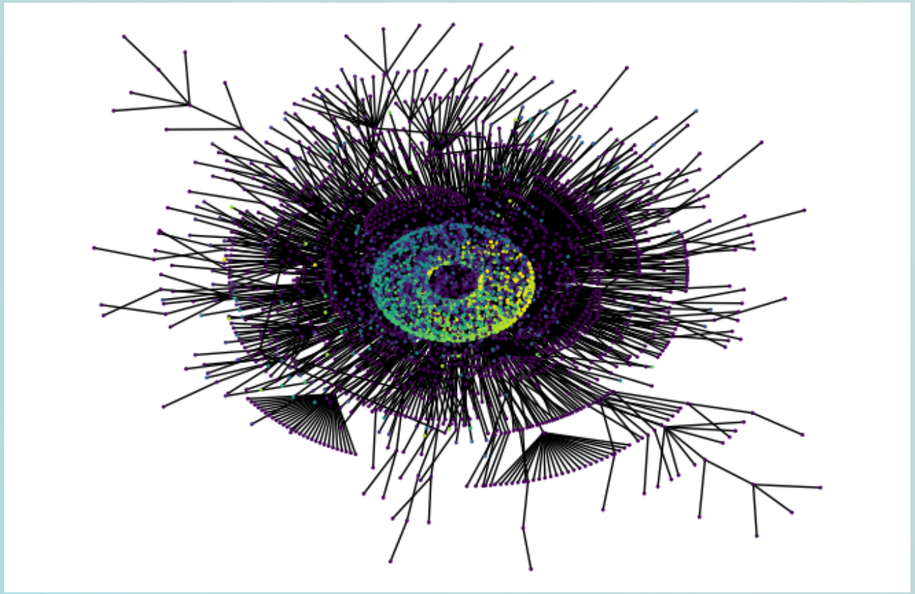


fig 15 -Whole graph where each color represents a community

While the visualization of the subgraph consisting of the top nodes by degree reveals a **clearer separation of communities in the first subgraph**, the overall graph representation, using all nodes, demonstrates that the **second graph exhibits a more cohesive community structure**, characterized by an **increased number of edges, and interconnectivity**. This indicates that the **reduced differentiation of communities in the subgraph does not necessarily imply a negative outcome**.

Moreover, the **cohesive nature of the second graph** presents opportunities to **mitigate sparsity** in the recommender system. By leveraging the main community structure of the second graph, when constructing the **user-item matrix**, it may be possible to **reduce sparsity** and **enhance the effectiveness** of the recommender system. This approach allows for a more **comprehensive representation of user-item interactions** within the dominant community, potentially leading to **more accurate and relevant recommendations**.

RECOMMENDER SYSTEMS

Plan

Given that the main goal of this project is to create **recommender systems for multiple communities** found within the dataset, with the creation of those communities, the objective is now to **recommend businesses to the members of each community**. With this in mind, we tried the following 2 approaches:

User Based **AND** **Content Based**

Also, it's important to note that, firstly, the recommender systems were created **globally** (for all users, without the division in communities) and then **later applied to the communities found**. This way, we can compare the behavior of the two approaches in **different scales**.

The first approach tackled was the **User Based** recommender system.

RECOMMENDER SYSTEMS

In order to simulate real-life scenarios, where we **base future predictions on past facts**, we decided to split the training set and test set as follows: **80% of the reviews in chronological order were used as the training data**, and the **remaining 20% were used as the test data**. With this setup, we calculated the RMSE, fit time, and test time for all algorithms and found that **SVD() performed the best.**

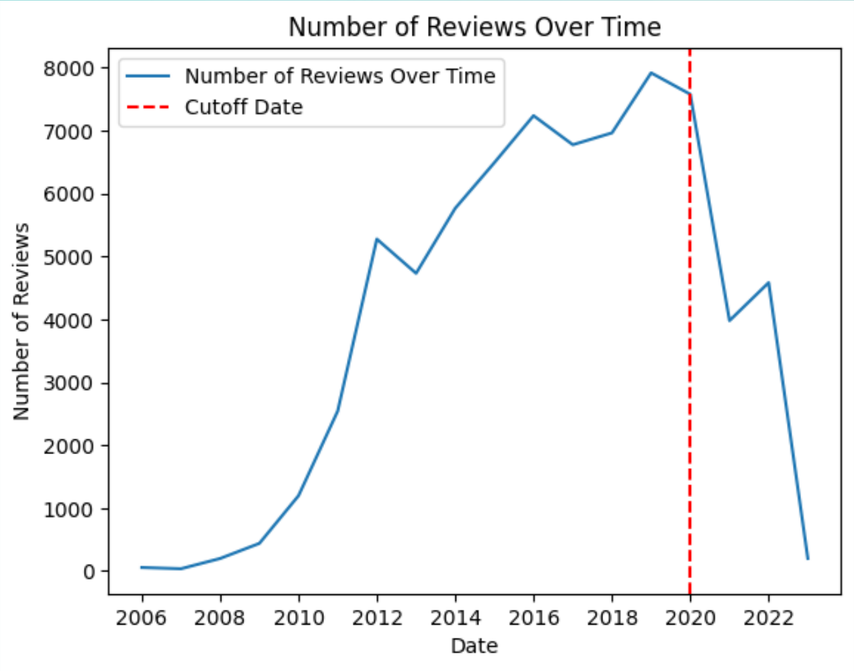


fig 16 - Train-Test Split

RMSE	Train Time	Test Time	Algorithm
1.138000	0.720860	0.104445	SVD
1.213047	2.104502	0.526781	KNNBasic
1.233476	2.871465	1.264984	KNNWithZScore
1.227495	2.333627	0.705222	KNNWithMeans
1.233476	2.718749	0.685497	KNNWithZScore
1.571694	0.049043	0.084323	NormalPredictor

fig 17 - RMSE, Train Time and Test Time for different Algorithms

However, when dividing the data this way, there might be a problem in the **test set where there aren't enough reviews from certain users to compare with the predictions**.

RECOMMENDER SYSTEMS

Collaborative Filtering

Mean RMSE	Mean Fit Time	Mean Test Time	Algorithm
1.079147	0.840373	0.091410	SVD
1.173829	2.451080	1.135614	KNNBasic
1.193781	2.503634	1.177059	KNNWithMeans
1.204212	2.959654	1.215880	KNNWithZScore
1.531086	0.070738	0.064760	NormalPredictor

fig 18 - Mean RMSE, Train Time and Test Time for different **User-Based Algorithms,** with cross-validation technique.

Mean RMSE	Mean Fit Time	Mean Test Time	Algorithm
1.166402	0.493343	0.826982	KNNWithMeans
1.172126	0.646897	0.860676	KNNWithZScore
1.213978	0.411007	0.805247	KNNBasic
1.525703	0.079656	0.121157	NormalPredictor

fig 19 - Mean RMSE, Train Time and Test Time for different **Item-Based Algorithms,** with cross-validation technique.

To address this issue, we also conducted a **cross-validation with 5 folds**, where we **compared all algorithms**, either **User**-**Based** or **Item**-**based**, using the **Normal Predictor** algorithm as the **baseline**, and **calculated RMSE, fit time, and test time again**. Once more, **SVD() emerged as the top performer**.

SVD() excels due to its ability to handle sparse matrices, which is the case in our dataset.
Furthermore, we performed a grid search to determine the best parameters to apply to SVD().

RECOMMENDER SYSTEMS

Content Based

Regarding the **Content Based** approach, we **vectorized the business's categories using the Term Frequency-Inverse Document Frequency Vectorizer library.** Consequently, we obtained a **matrix** where each line represents a **business** and each column symbolizes a **category of the *categories vocabulary***. Each value of the matrix represents the **TF-IDF weight of the category in the respective business.**

After that, the **cosine similarity** between all pairs of businesses is calculated, based on their TF-IDF representations. This measure, **quantifies the similarity between two businesses**, in terms of their categories. With this in mind, to recommend businesses, for a specific user, we:

1. **Compute the cosine similarity scores between each business the user has reviewed and all the other business in the dataset.**
2. **Select the top N most similar business, to each of the user's reviewed businesses.**
3. **Aggregate the recommended businesses and return the best, based on the frequency of occurrence in the recommendations.**

RECOMMENDER SYSTEMS

We evaluated the **mAP** with **precision@10** for both **user-based** and **content-based** recommendation system approaches, considering that a business is **relevant if the user has reviewed it with, at least, a rating of 4 stars**. Also, it's important to note that:

- we chose users that have made many reviews, in order to have a solid ground truth;
- we didn't compute the rank average precision, since we consider that highly ranked hits are not our main concern in this context.

For the **user-based approach**, after computing the top ten users with the most reviews, we split that data into training and testing sets, with the **training set corresponding to 70%**. After that, we considered the test set to be our **ground-truth**. Inside the test set, we considered relevant documents as stated above. Lastly, the obtained mAP for this approach was **0.04**. With the vast amount of businesses available to recommend (4656) , this result is expected, as for example, for the top user with the most reviews, there are 212 reviews with rating >4, which represents 4% of total users.

For the **content-based approach**, the mAP obtained was 0.12, which is a better result than the previous approach. Once again, this continues being low, which can be explained by the fact that there are a lot of businesses available in our dataset that can have similarities in categories with the businesses that the user reviewed.

RECOMMENDER SYSTEMS

Communities - Collaborative Filtering User-Based

As we have a set of communities for both friendship and similarity graph approaches, we computed **RMSE** and **MAE** using **cross validation** and the **SVD** algorithm for the five communities with the most users.

The results suggests that the similarity approach has better results. Despite having radically more nodes than the friendship approach, the similarity aproach gets a lower RMSE and MAE, which indicates that the it can be a good community split to have more efficiency in the recommender system.

Community ID	Mean RMSE	Mean MAE	Number of Nodes	Fit Time	Test Time
0	1.026036	0.796502	131	0.029225	0.003503
1	1.134771	0.885534	104	0.011103	0.001302
2	1.152765	0.911604	85	0.010647	0.001017
3	1.107483	0.869412	85	0.007994	0.001420
4	1.047081	0.811819	35	0.005346	0.000200

fig 20 - Mean RMSE and MAE for the 5 communities with the most users, from the friendship approach

Community ID	Mean RMSE	Mean MAE	Number of Nodes	Fit Time	Test Time
0	0.926827	0.706431	490	0.129968	0.022021
1	1.064783	0.847006	239	0.029190	0.004111
2	1.015482	0.792426	182	0.020832	0.002817
3	0.909743	0.655766	172	0.017745	0.002585
4	1.179811	0.934930	83	0.007360	0.001007

fig 21 - Mean RMSE and MAE for the 5 communities with the most users, from the similarity approach

RECOMMENDER SYSTEMS

Communities - Content-Based

The evaluation of the content-based recommendation approach, where **only businesses reviewed within each community were considered**, reveals noteworthy insights. The **similarity-based approach** demonstrates a **slightly superior** overall mean precision score of **0.092** compared to **0.084** in the **friendship-based approach**. This discrepancy is particularly significant given the observation that the **similarity-based approach** offers more than **double** the **businesses available** for recommendation within each community.

The evaluation of the content-based recommendation approach, where **only businesses reviewed within each community were considered**, reveals noteworthy insights. The **similarity-based approach** demonstrates a **slightly superior** overall mean precision score of **0.092** compared to **0.084** in the **friendship-based approach**. This discrepancy is particularly significant given the observation that the **similarity-based approach** offers more than **double** the **businesses available** for recommendation within each community.

Community	Num of Nodes	Num of Businesses	Mean Average Precision
0	131	1027	0.11
1	104	697	0.05
2	85	666	0.09
3	85	608	0.02
4	35	461	0.15

fig 22 - mAP for each community in **friendship approach**

Community	Num of Nodes	Num of Businesses	Mean Average Precision
0	490	2695	0.08
1	239	1688	0.08
2	182	1201	0.13
3	172	1364	0.08
4	83	575	0.09

fig 23 - mAP for each community in **similarity approach**

RECOMMENDER SYSTEMS

Cold Start

New Users

Whenever a **new user** enters a community, there **won't be any reviews available**, which impacts the recommender algorithm due to it **not finding any similarities and unavailability of data**.

Therefore, we decided on a system where it recommends the top 10 most popular businesses, among the community a new user is in.

New Businesses

Whenever a **new business** enters the network, it **won't have any reviews available**. This will **not affect the content-based recommender** system, considering that the content used is only the business **categories** and this information will be available.

However, since there are no user reviews the **user-based** approach will **fail** to recommend this business to users. One way to fix this would be to use external aid like paid ads or to actively promote new businesses to users.

CONCLUSION

Choosing between content-based and collaborative filtering approaches can be challenging, especially when both demonstrate similar results. However, given the context that we **aim to match similar tastes in businesses** and recognize that **friends may not always share identical preferences** in businesses, the **community-based similarity approach** appears to be more suitable. It offers better understanding of user preferences within localized communities.

While a **hybrid** approach may **seem promising,** the abundance of available businesses makes it unlikely to encounter **overlaps between collaborative filtering** and **content-based recommendations**. This makes the hybrid approach essentially a **random selection** from the lists generated by both methods.

In conclusion, while both content-based and collaborative filtering approaches yield comparable outcomes, the community-based similarity approach stands out in its ability to adapt to diverse tastes within communities. However, the feasibility of a hybrid approach is limited by the lack of overlap between recommendation lists. This underscores the importance of considering contextual factors and the unique characteristics of recommendation datasets when designing and implementing recommendation systems.