Community-Based Recommendations

Machine Learning Complements

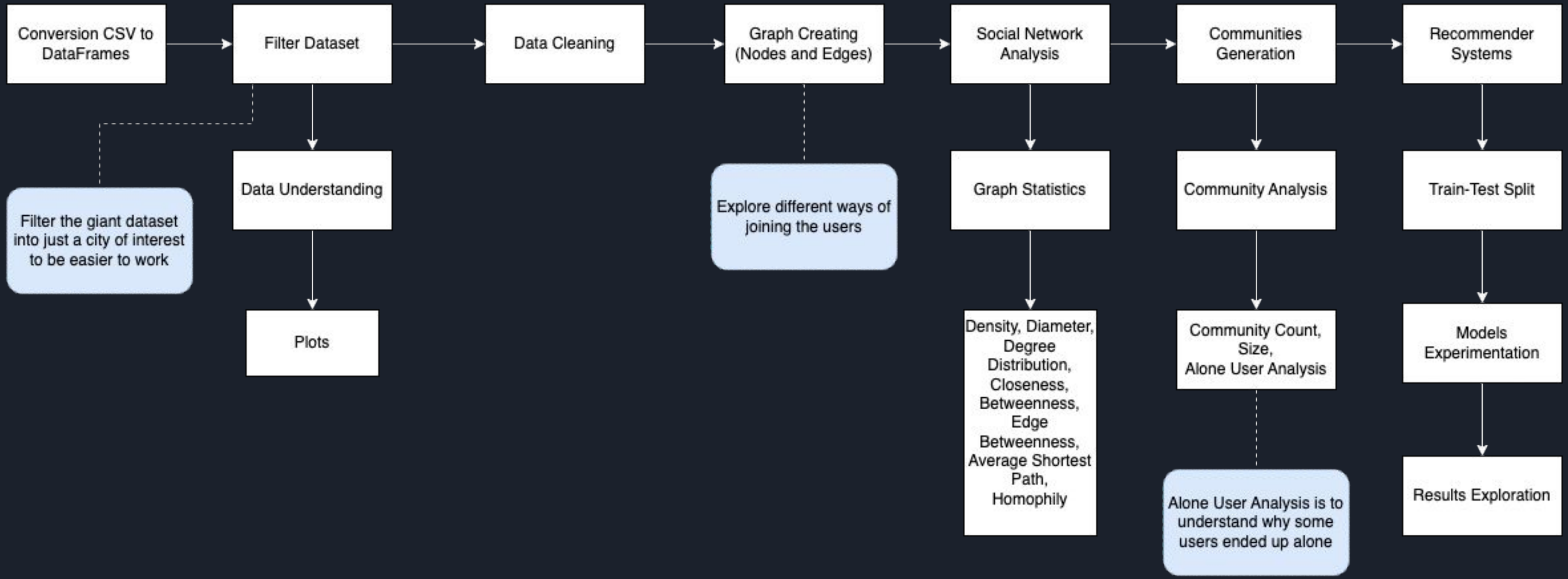
Group C

Bruna Marques, up202007191
Diogo Silva, up202004288
Lia Vieira, up202005042
Pedro Fonseca, up202008307

Domain Description

- **Yelp dataset** - businesses, check-ins, reviews, tips, users
- Extrapolate factors that play a role in **user connections**:
    - friendships, business categories reviewed, etc.
- Develop valuable **business recommendations** for users (in a city)
- Main objectives:
    - generate user communities based on social network analysis
    - apply a recommender system to each community
    - evaluate and understand the results

Project Pipeline



City of Choice

- Cities in the dataset were analyzed
- Picked **Springfield** as the city of choice:
    - Not too big - so that it doesn't take too long to run our models
    - Not too small - so that we have enough data
    - Good ratio between reviews and businesses (**29.02** reviews/business)
    - 384 businesses, 1683 categories, 11145 reviews, 1455 tips, 7707 users

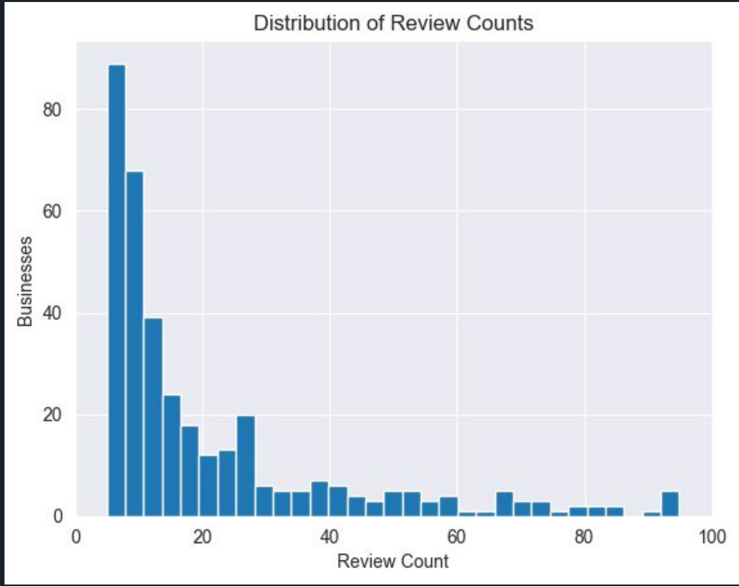
1	City	Businesses * Reviews	review_count	business_count
2	Philadelphia	0.115973	936240	14569
3	Tampa	0.062626	439506	9050
4	New Orleans	0.061678	621361	6209
5	Tucson	0.059431	387254	9250
6	Nashville	0.055060	441053	6971

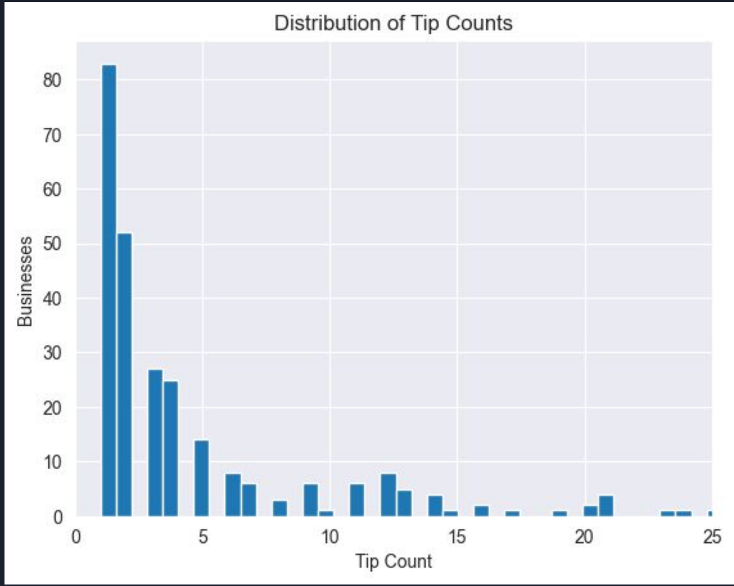
• • •

[illegible]	Bensalem	[illegible]	[illegible]	[illegible]
47	Ardmore	0.002393	15451	376
48	Wayne	0.002329	14669	375
49	Exton	0.002296	12764	419
50	Media	0.002268	14023	372
51	Carpinteria	0.002228	16895	298
52	Norristown	0.002221	11163	448
53	Newark	0.002153	13096	359
54	Conshohocken	0.002084	14631	301
55	Phoenixville	0.002080	12025	365
56	Mount Laurel	0.002075	12691	344
57	Springfield	0.002054	11145	384
58	Lansdale	0.002026	11013	378
59	Oro Valley	0.001879	12520	286
60	Willow Grove	0.001867	11368	311
61	Clearwater Beach	0.001858	21471	163
62	Seminole	0.001850	9665	359
63	Smyrna	0.001848	9468	366
64	Newtown	0.001848	10754	322
65	Langhorne	0.001848	10589	327

Exploratory Data Analysis

Businesses



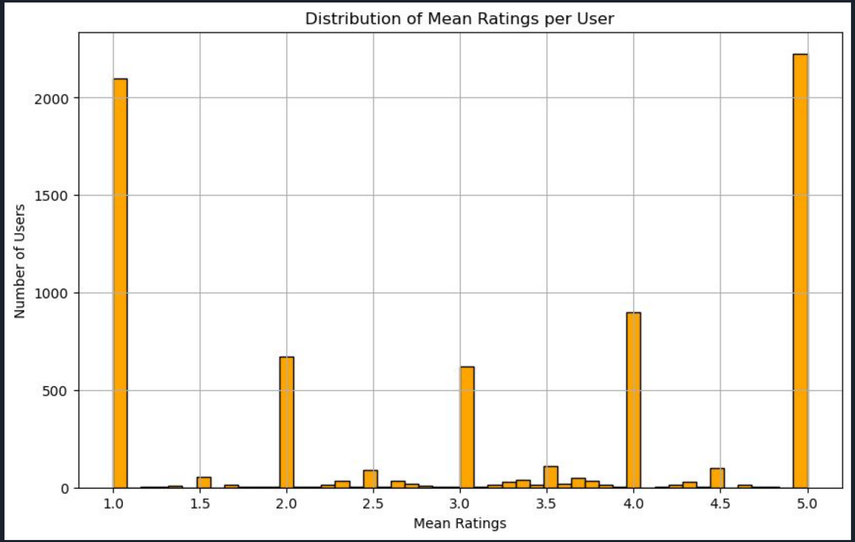


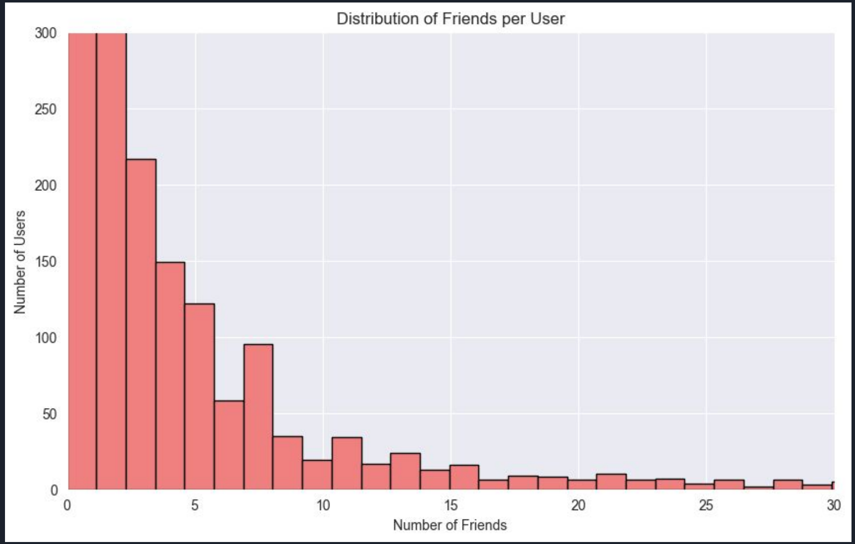
Reviews vs Tips - similar distributions, but review data visibly more dense

Note: all users have either reviewed (7350) or tipped (907) at least once

Exploratory Data Analysis

Users





Bimodal distribution in ratings
(users tend to be extreme)

5521 users with no friends in the city
868 users with just one friend in the city

Data Preparation

- **Data cleaning:**
  - fill null values (user friends, etc.)
  - drop several columns (city, state, etc.), leaving the relevant ones
- **Feature engineering** - new data frames with users main interactions:
  - businesses reviewed
  - businesses tipped
  - categories reviewed/tipped
  - friendships

Graphs - Nodes & Edges 1/2

- **Libraries used** - networkx for the graphs
- **Nodes** - always users (we didn't find it useful to join other types of nodes)
- **Edges**
  - **Friendships** - joins users on their friendships (weightless)
  - **Compliments** - vectorized the compliments and joins 5 Nearest Neighbors (weight = distance between compliments)
  - **Reviews** - users with a review on the same business (weight = count)
  - **Tips** - users with a tip on the same business (weight = count)
  - **Categories** - users with a tip or review on a business of the same category (weight = count)

Graphs - Nodes & Edges 2/2

- **Edges**
  - **Combined** - joins users based on multiple variables - categories, reviews, tips, friendships (weights sorted by ascending rarity of interaction)
  - **Categories and Reviews Combined** - same as combined but just categories and reviews
  - **Priority Combined** - joins users based on a priority list
    - <u>Example</u>: joins users by friendships, then picks users left alone and joins them by using another variable
    - <u>Priority list</u>: friendships, reviews, tips, categories
    - <u>Weight</u>: weight of the connection from each variable logic

Social Network Analysis

Community Detection - Models Used

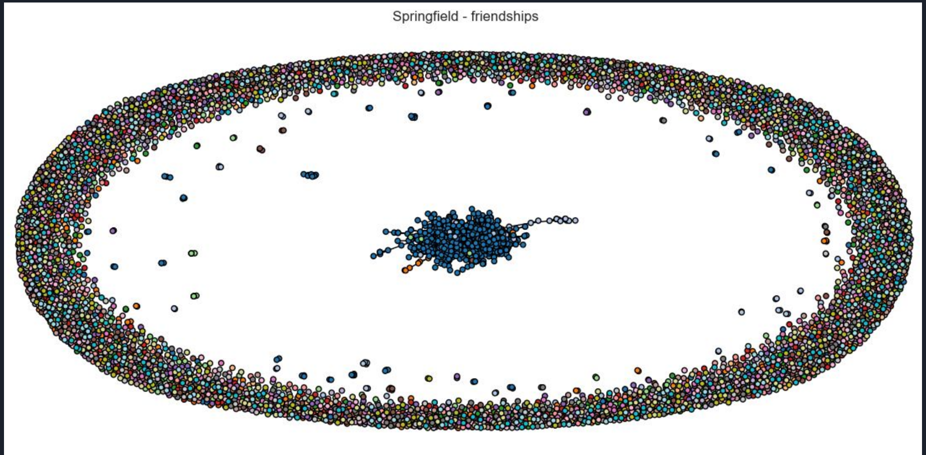
- **Girvan-Newman:**
  - relies on edge-betweenness centrality
  - biased community formation - favoring few large communities initially
  - scalability and community cohesion are limited (large-scale systems)
- **Louvain:**
  - modularity optimization
  - some sensitivity to noise
  - balanced community formation
  - scalability for large datasets
  - selected as the algorithm of choice (outweighs Girvan-Newman's drawbacks)

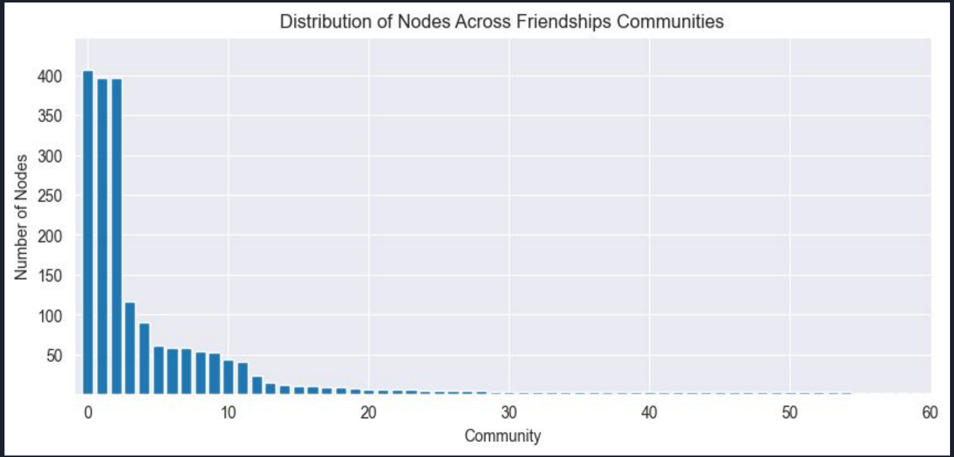
Social Network Analysis

Friendships

- High amount of alone users
- Good results on the ones actually connected
- 3 major communities with roughly 400 users
- 10 communities with roughly 50 users

Alone Users
5521/7707 (71.6%)
Nº of Communities
5662 (**1.36** users/community)



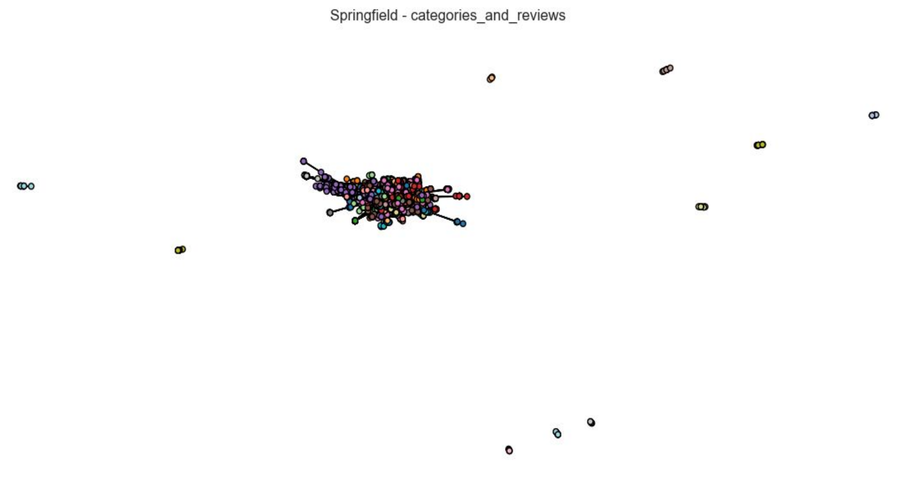


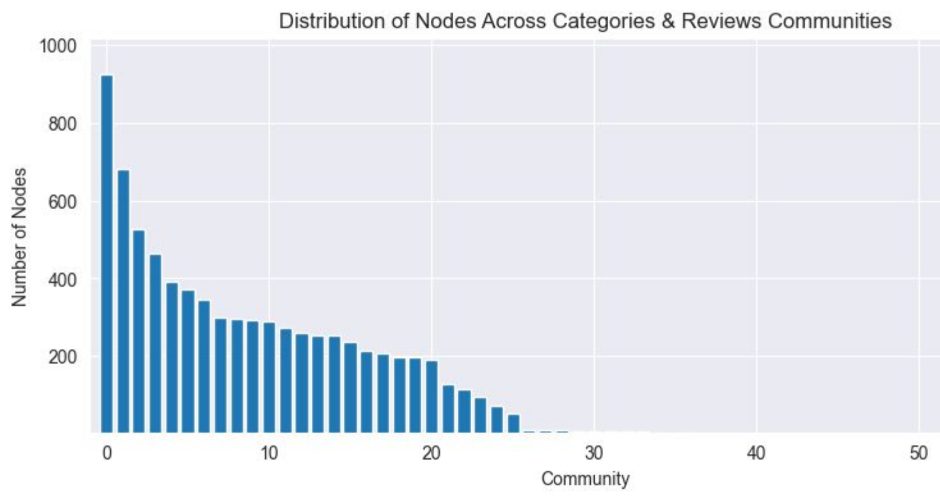
Social Network Analysis

Categories and Reviews

- No alone users - categories also consider tips, which helps cover this
- High number of connections:
    - some hold value but not that much generally

Alone Users
0/7707 (0.0%)
Nº of Communities
37 (**208.30** users/community)



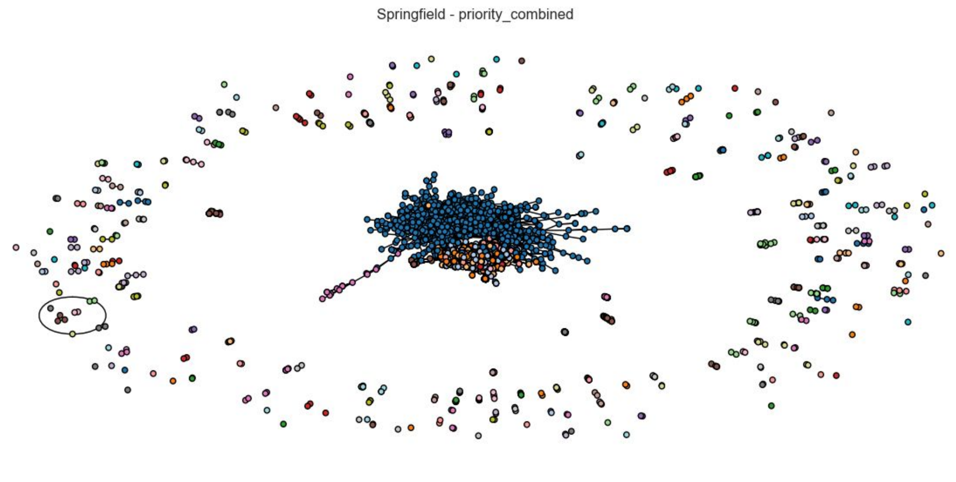


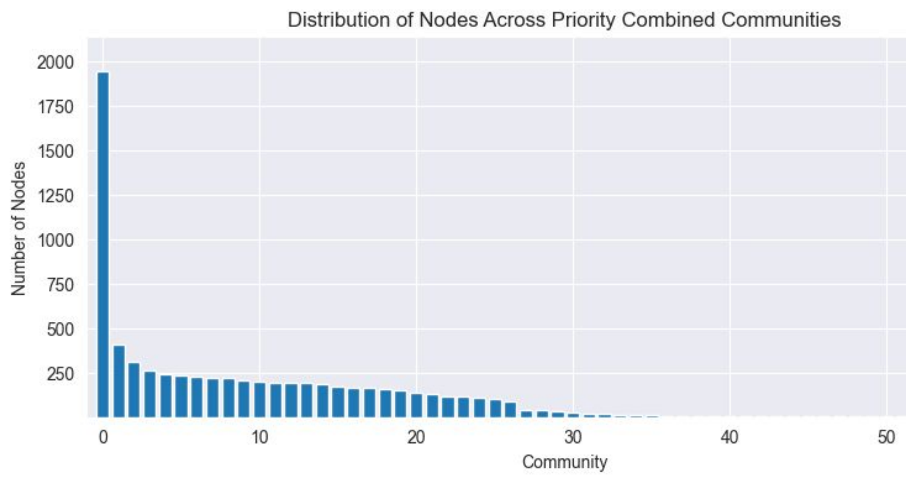
Social Network Analysis

Priority Combined

- Users joined based only on the best value possible
- Low amount of alone users
- Single giant community (almost 2000 users)
- 26 communities with 100-500 users

Alone Users
69/7707 (0.9%)
Nº of Communities
289 (**26.67** users/community)





Social Network Analysis

- Due to the amount of edges in the ‘Categories’ graph it was not possible to run all the statistics.
- Because ‘Combined’ and ‘Categories & Reviews’ also use the ‘Categories’ edges, we were unable to determine the actual statistics but were able to do estimates

	Friendships	Reviews	Tips	Categories	Combined	Categories & Reviews	Priority Combined
Density	0.0002	0.0152	0.0003	0.0187	>0.0187	0.0187	0.0056
Diameter	0.0244	0.0867	0.0083	1.4546	>1.4546	>1.4546	1.0000
Avg. Closeness	0.016	0.338	0.003	0.417	<0.417	<0.417	0.152
Avg. Betweenness (e-05)	2.780	19.136	0.241	17.620	>17.620	>17.620	15.024
Avg. Edge Betweenness (e-05)	0.567	0.069	0.045	0.055	>0.055	>0.055	0.130
Avg. Shortest Path	0.01968	0.0447	0.0066	DNF	DNF	DNF	0.819
Homophily	-0.143	-0.031	0.008	-0.033	-0.035	-0.034	0.162

Social Network Analysis

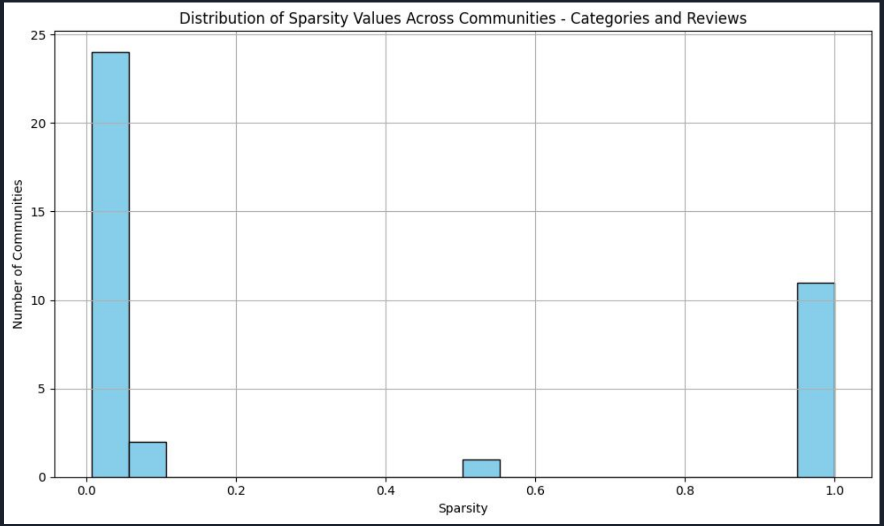
- Selected social network: **Priority Combined**
- 34914 edges of 4 different types (friendships, reviews, tips, categories):
  - **no mixed connections** - separation for more closeness (while keeping alone users to a minimal)
- Decent **closeness centrality** - balanced/trustworthy recommendations:
  - too much would limit serendipity
- Considerable **edge betweenness** - more centralized, with some “bridges”:
  - too much would difficult interpretation of recommendations
- High **homophily**:
  - valuable connections to base recommendations on
- Furthermore, RSs for all SNs were tested:
  - Priority Combined proved to be the best

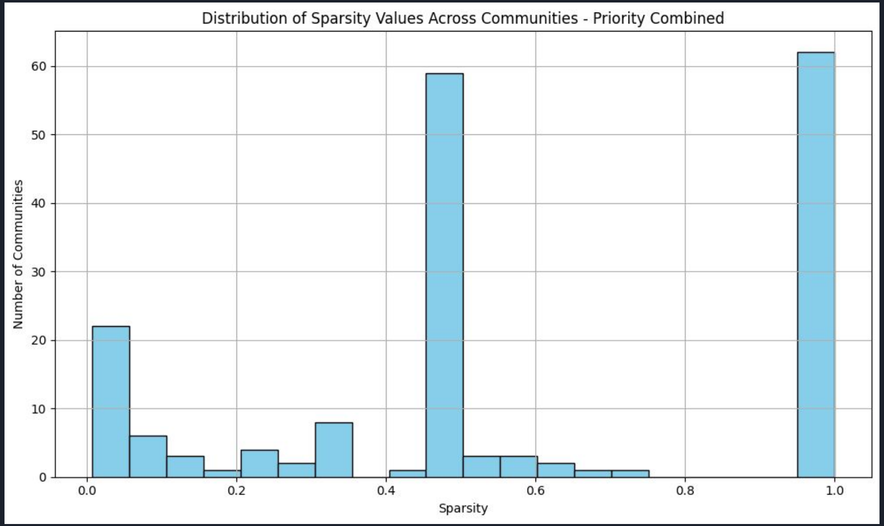
Recommender Systems

- Recommendations were computed **for each community**
- For all recommender systems:
  - users with **fewer than 3 reviews** discarded
- For each user, the reviews were divided :
  - **80%** to training set
  - remaining **20%** to test set
- Criteria for recommending a business:
  - its estimated rating is **superior than 3**

Recommender Systems

Community Matrices Sparsity





Categories and Reviews:

- low sparsity - rich data
- less communities - less diversity

Priority Combined:

- medium sparsity
- more communities - more diversity

Recommender Systems

User Based

- Implements an **user-based** collaborative filtering recommendation system for each community using **KNN** algorithm
- Returns recommendations for each user within each community

Item Based

- Implements an **item-based** collaborative filtering recommendation system for each community using the **KNN** algorithm

	Avg RMSE	MAP	Precision	Recall
Business Tips	1.21	0.56	0.59	0.95
Categories	1.54	0.47	0.58	0.86

	Avg RMSE	MAP	Precision	Recall
Business Tips	1.22	0.53	0.59	0.94
Combined	1.55	0.47	0.59	0.85

Recommender Systems

SVD

- Implements **Singular Value Decomposition**, returning for each user the associated list of tuples containing the recommended business and its estimated rating

	Avg RMSE	MAP	Precision	Recall
Business Tips	1.16	0.54	0.61	0.95
Categories	1.5	0.5	0.61	0.92

Normal Predictor

- Implements **random** recommender algorithm for each community using the *NormalPredictor* class

	Avg RMSE	MAP	Precision	Recall
Business Tips	1.57	0.35	0.57	0.61
Business Reviews	1.96	0.31	0.52	0.56

Recommender System - Content Based

1. The 'categories' list was separated into separate columns, one for each category
2. A **similarity matrix** was then created using cosine similarity
3. For each business positive rated by the user, the function identifies similar businesses based on the cosine similarity matrix
4. Returns the top 3 recommendations based on the accumulated similarity scores
5. The average precision was calculated based on the algorithm's ability to recommend the right businesses

	Precision	**MAP***
Categories and Reviews	0.014	0.03

* mean of average precision across users

Recommender System - Hybrid

- Combines user-based and content-based recommendations
- It prioritizes recommendations common to both approaches and then fills in the remaining slots with recommendations from either approach

	Precision	**MAP***
Categories and Reviews	0.02	0.023

* mean of average precision across users

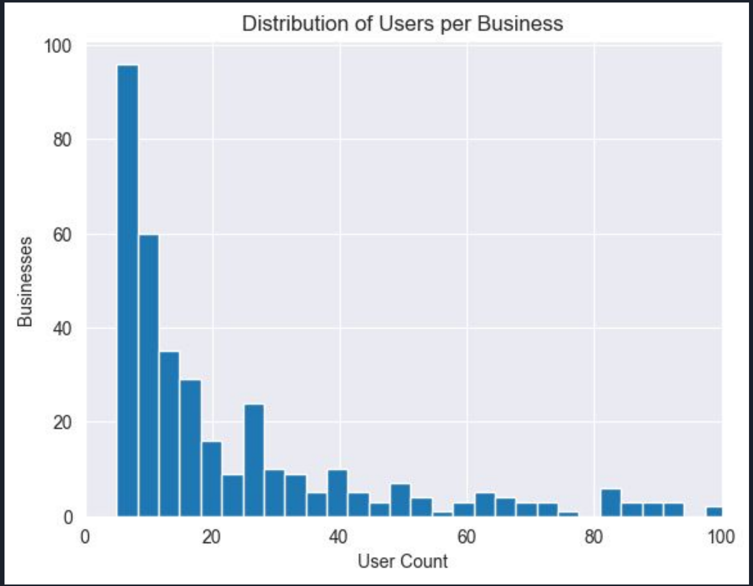
Conclusions

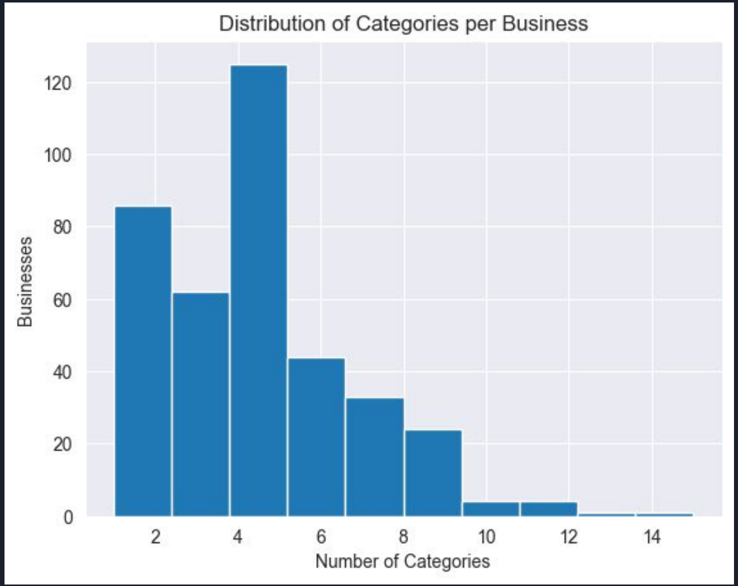
- Overall, the **SVD** yielded the best results compared to other algorithms evaluated
- The **content-based** filtering - which relies on the business categories - had the worst results:
    - indicating that the category information alone may not be sufficient to capture the preferences of users accurately
- Users with a **small number of reviews** and **high sparsity** in the user-business interaction matrix negatively impacted the performance of the recommendation system

Next Steps

- Apply **Natural Language Processing** to the texts of the reviews (and other features)
- Experiment with **Time Series**
- Avoid breaking the dataset so abruptly to allow users to have friends across cities
- Explore new ways of categorizing user nodes and compute different types of assortativity

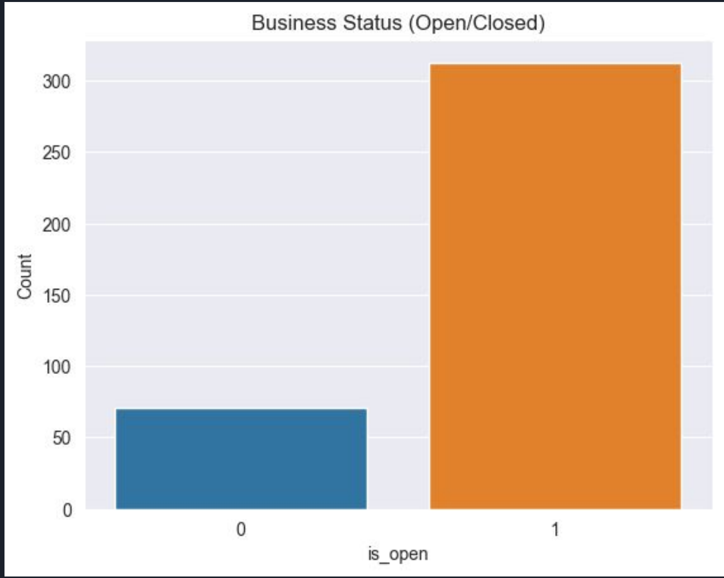
ANNEXES



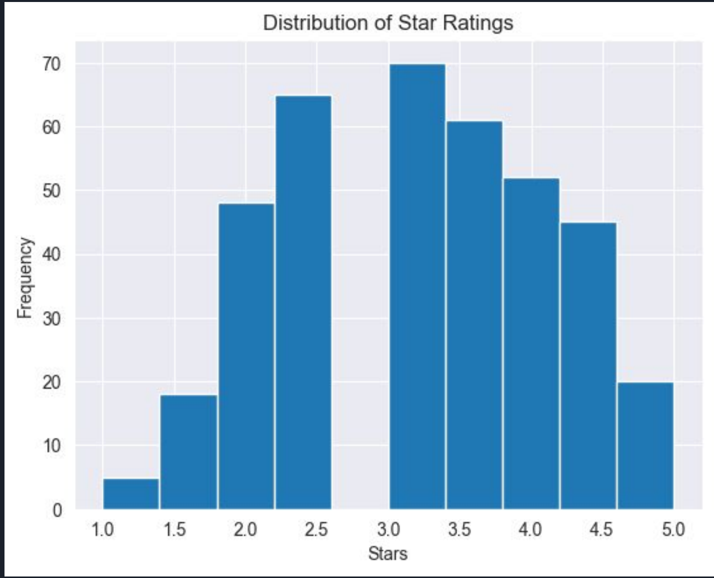


Majority of businesses having few user reviews

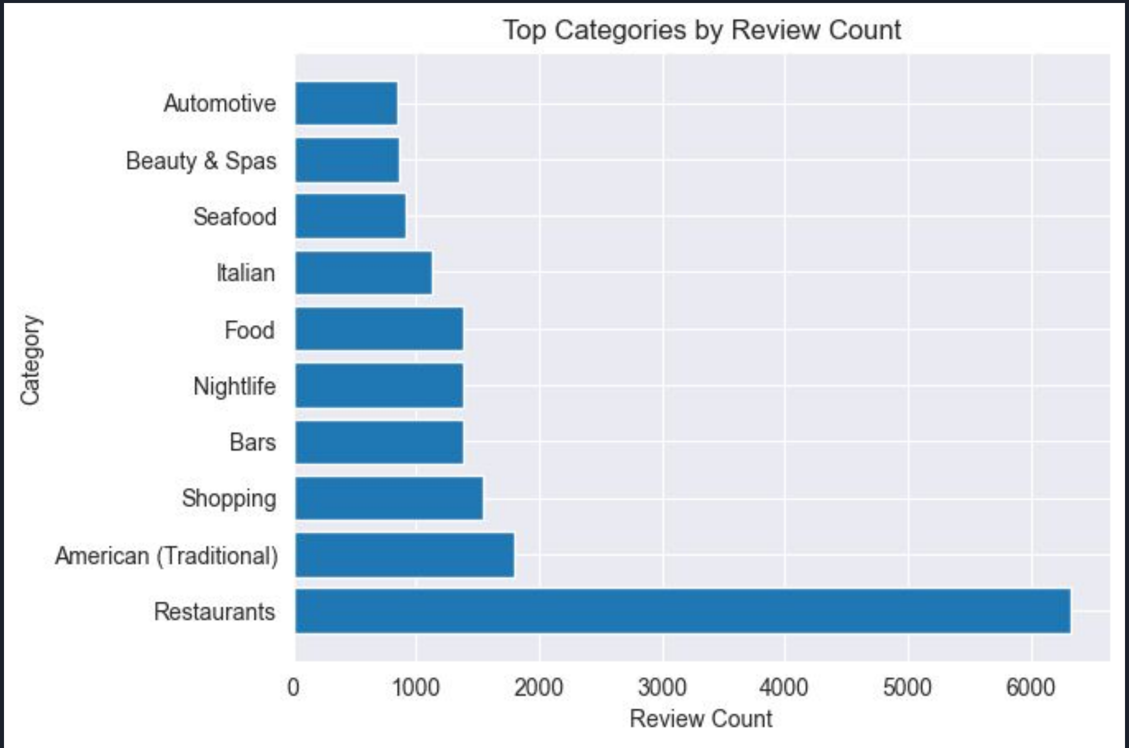
A lot of businesses covering around 5 categories



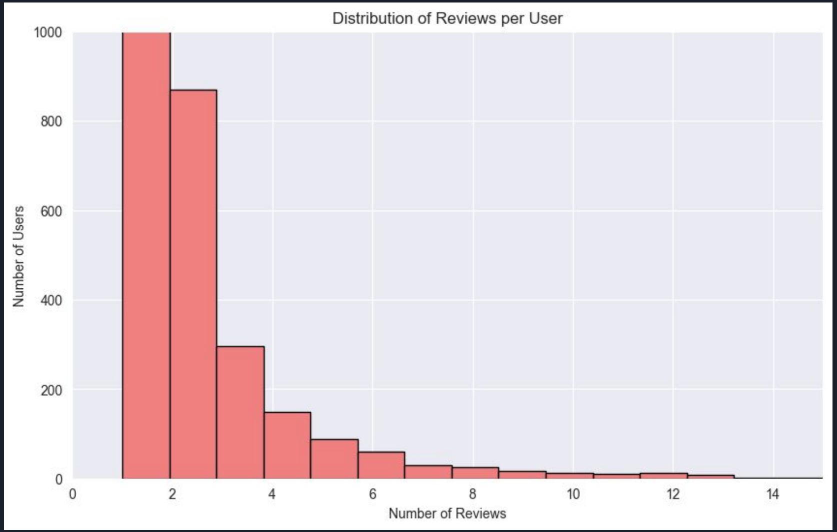
Around 80% of businesses are opened

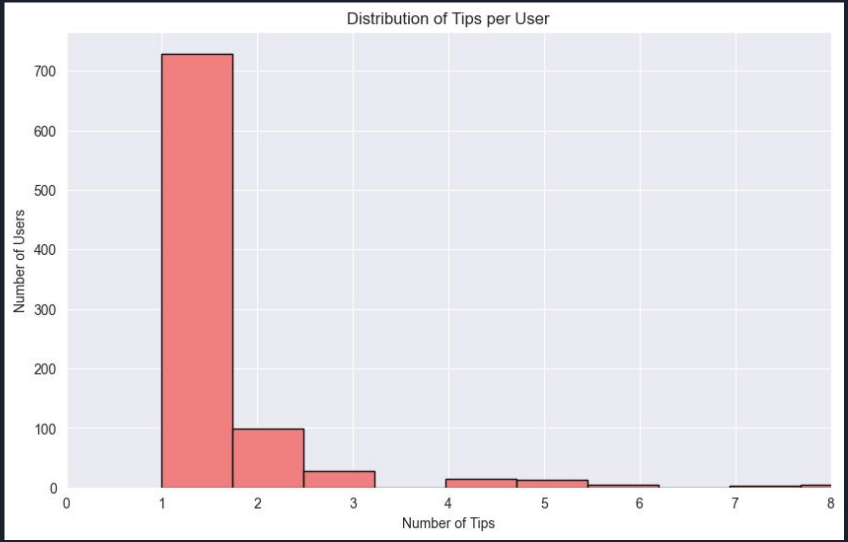


Star ratings distribution across businesses

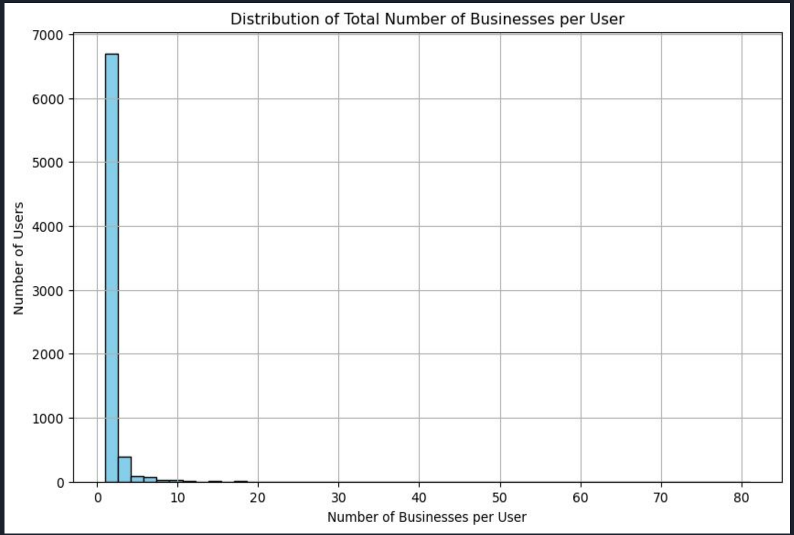


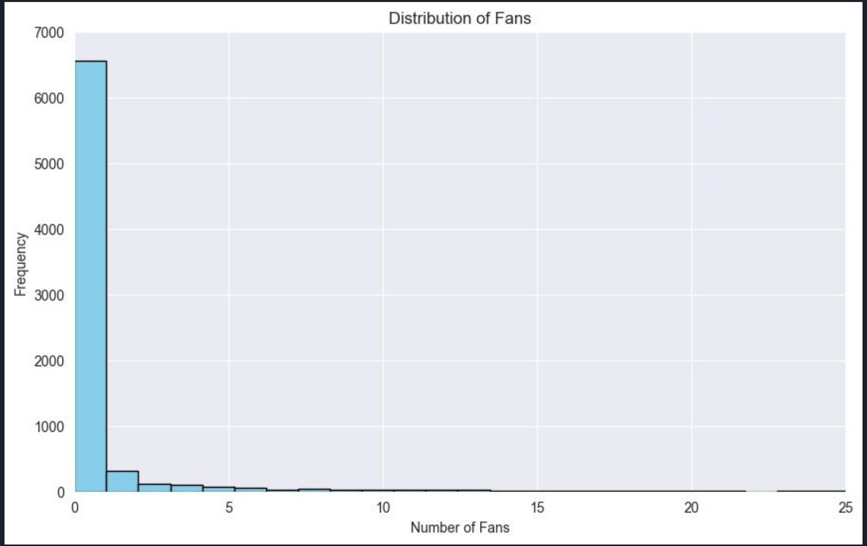
Restaurants clearly dominate in terms of reviews





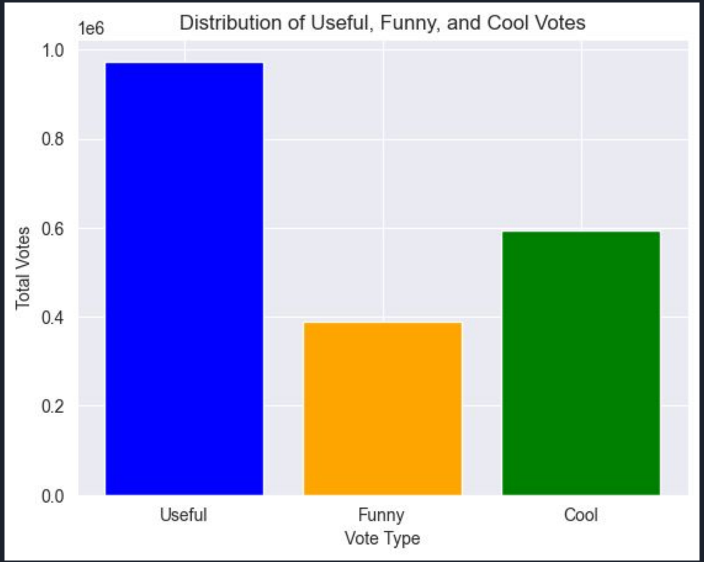
Reviews vs Tips - users mainly opting for reviews rather than tips



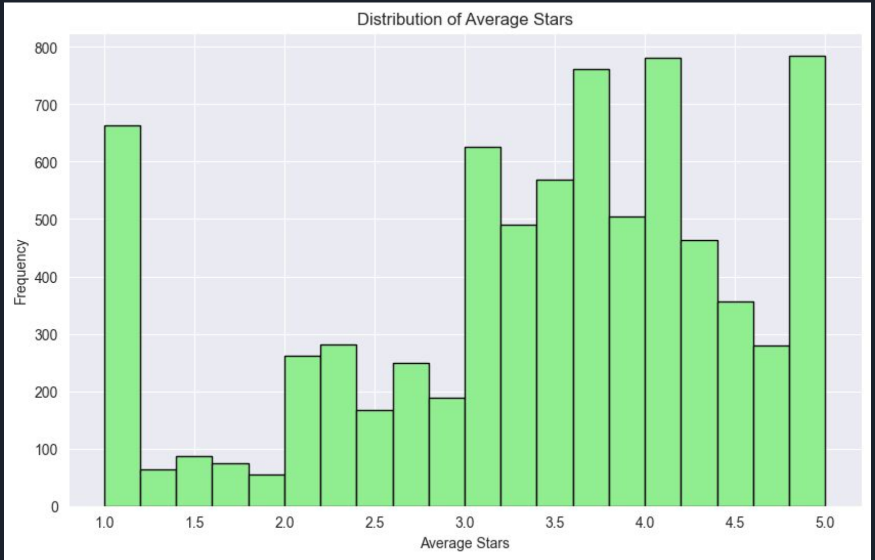


Most users reviewed a small number of businesses

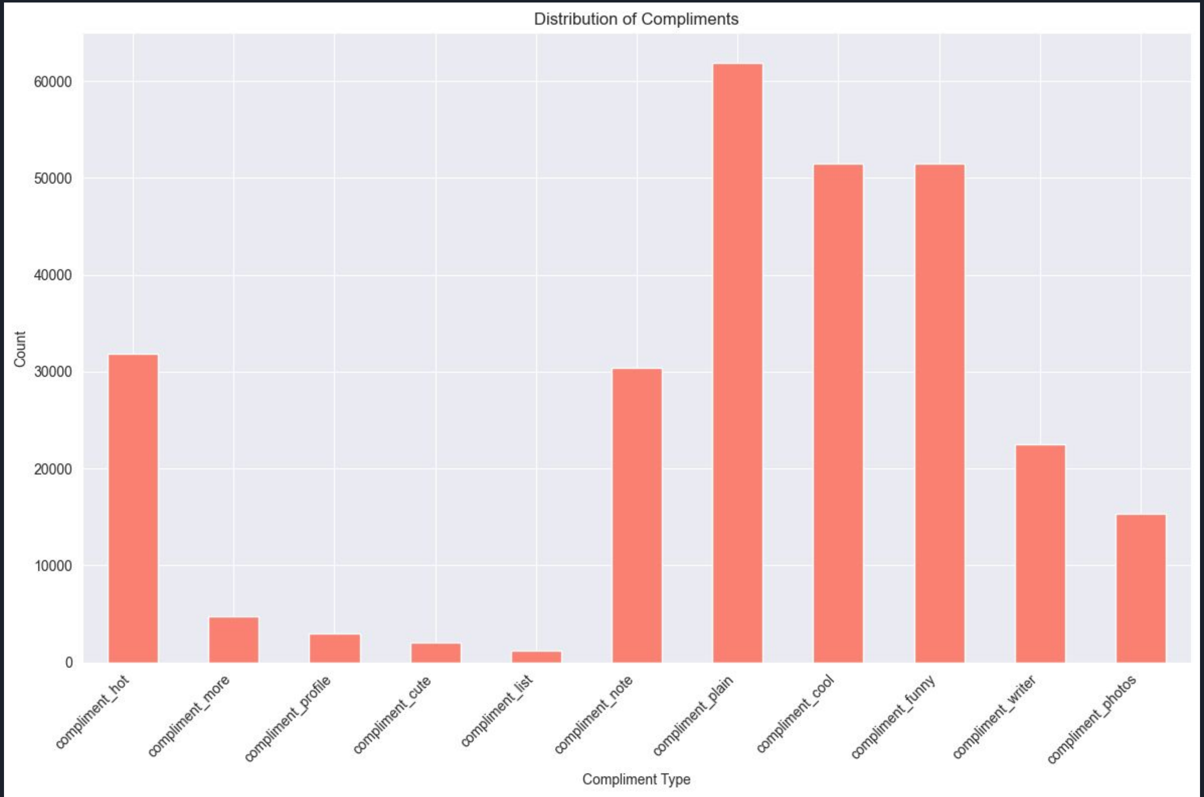
5716 (74.2%) users with 0 fans



Distribution of votes by users

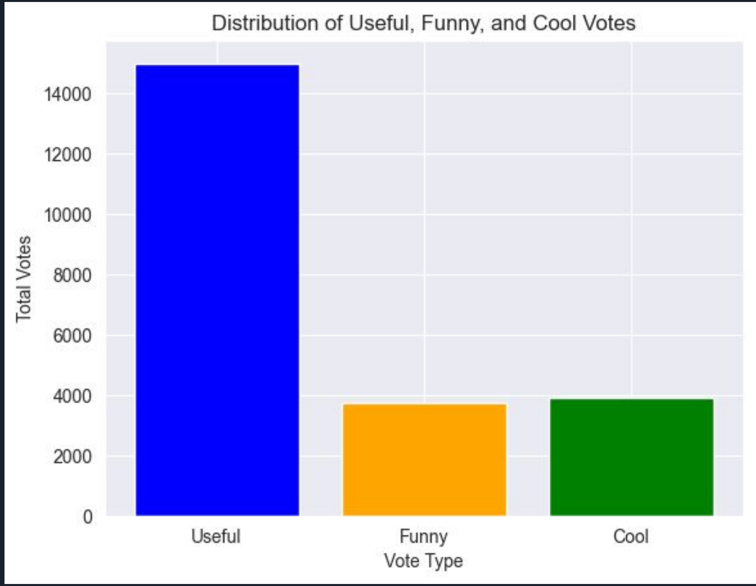


Global average stars per user
(mostly positive)

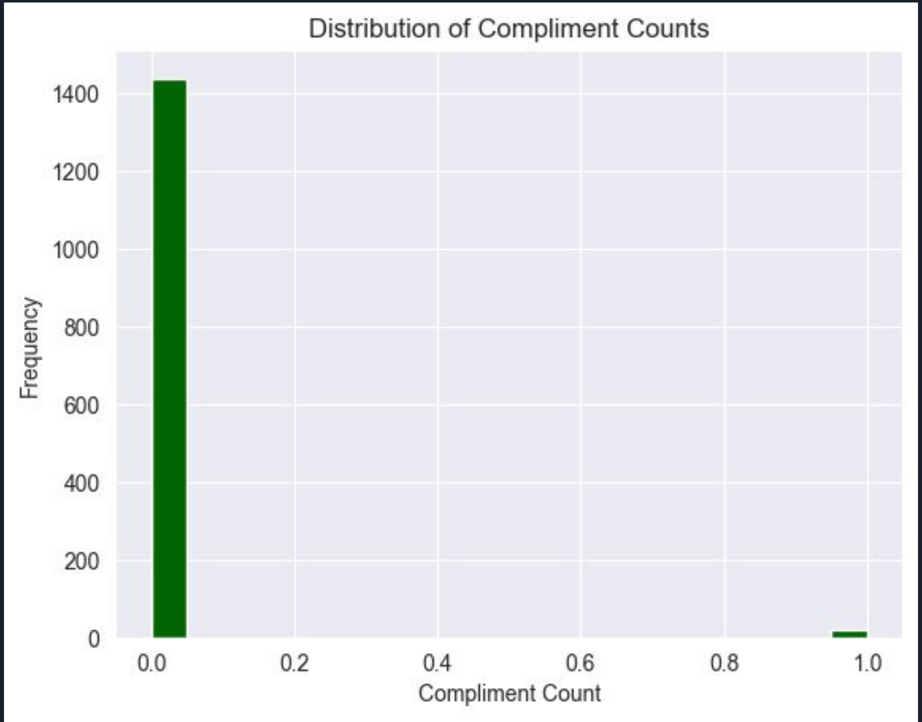


Distribution of compliments received by user

Star ratings distribution across reviews



Distribution of votes in reviews

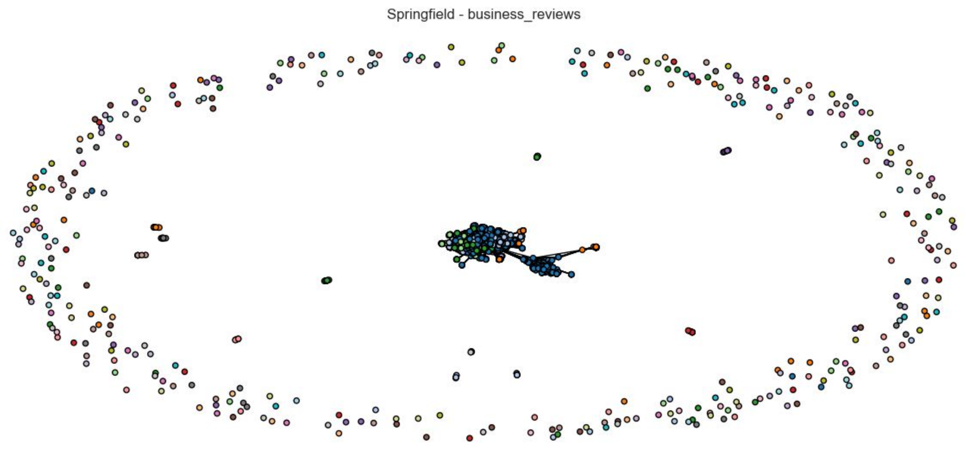


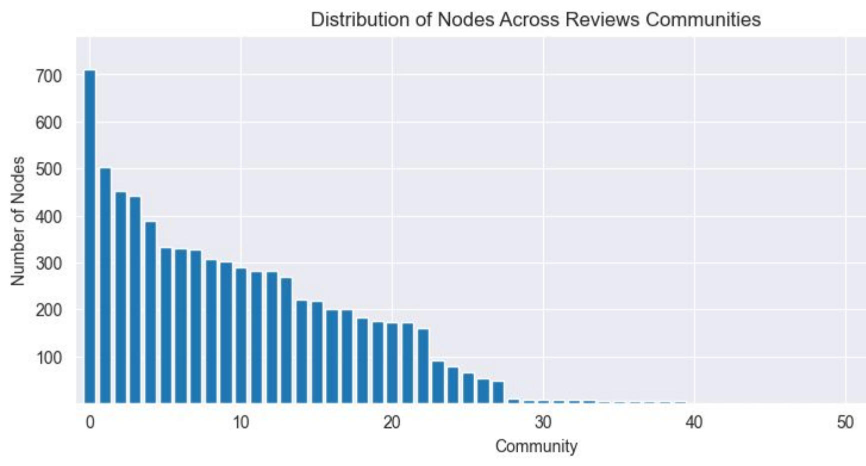
Most tips have 0 compliments

Reviews graph

- Somewhat connected
- Considerable amount of communities (could be even better)

Alone Users
357/7707 (4.6%)
Nº of Communities
397 (**19.4** users/community)

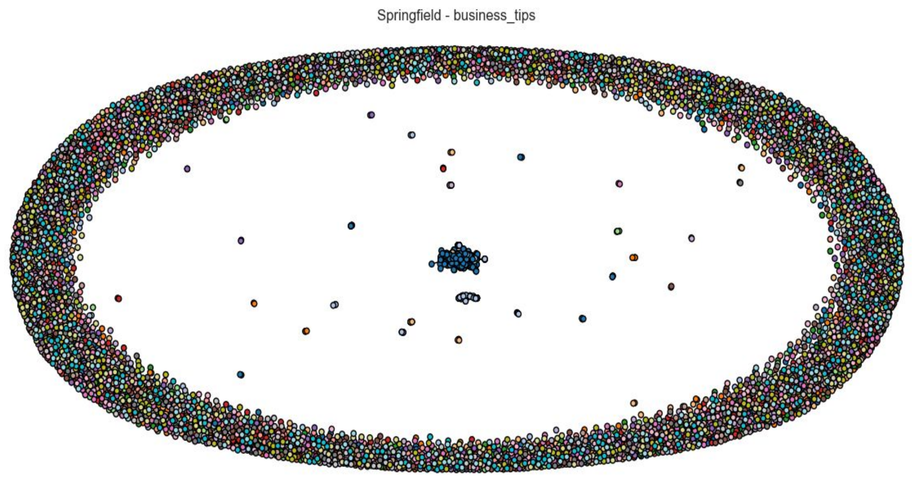


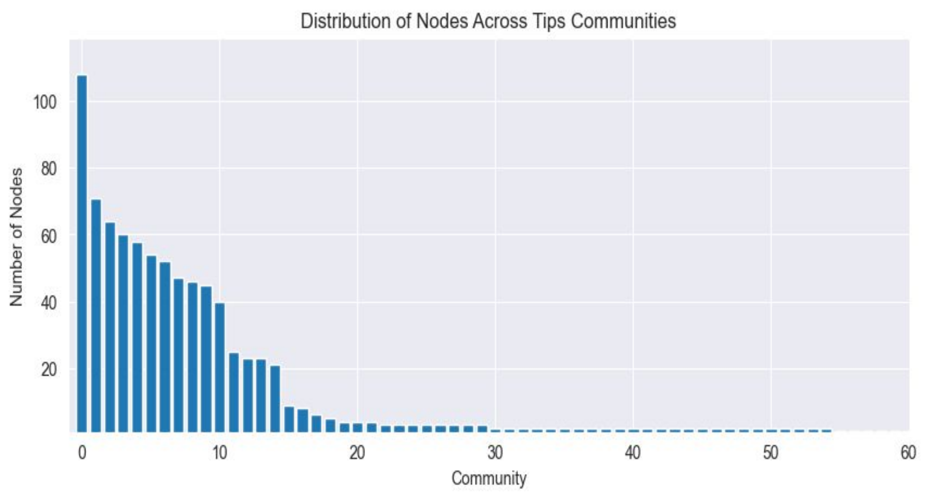


Tips graph

- Not very connected
- Low amount of valuable connections

Alone Users
6856/7707 (89.0%)
Nº of Communities
6911 (**1.12** users/community)

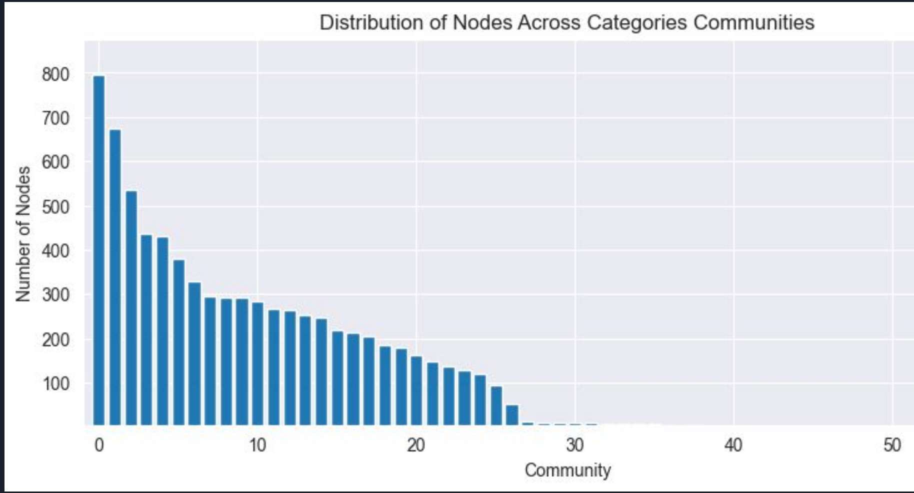




Categories graph

- Too connected
- Edges are not that valuable or strong

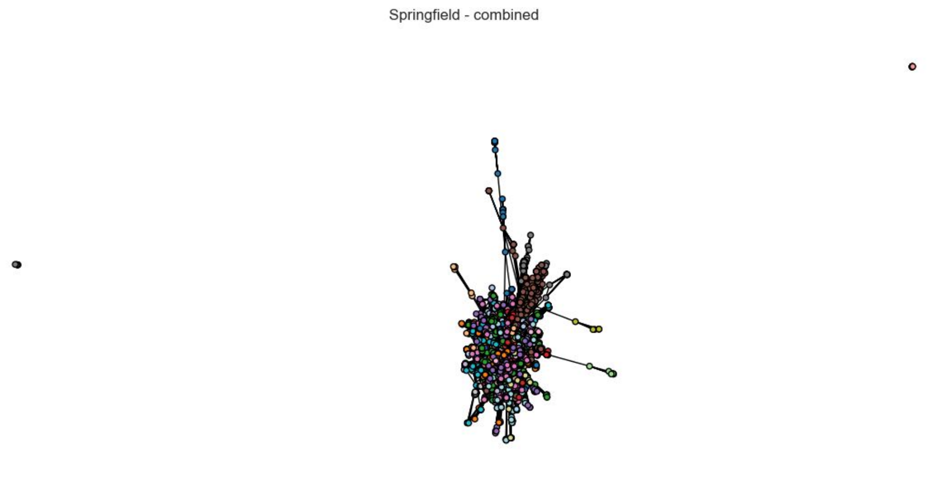
Alone Users
0/7707 (0.0%)
Nº of Communities
39 (**197.62** users/community)

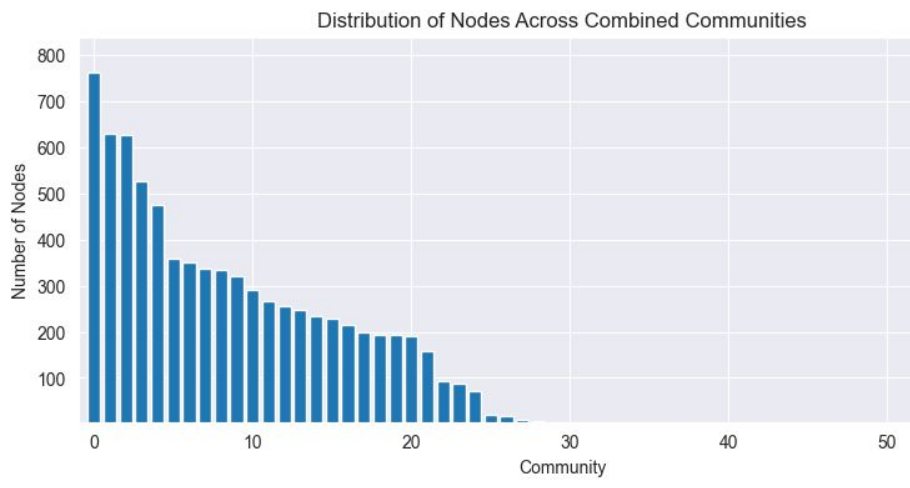


Combined graph

- A lot of connections
- Value is not that strong in several cases

Alone Users
0/7707 (0.0%)
Nº of Communities
31 (**248.61** users/community)





City	Connection	Algo	Precision	Recall	F1	MAP	Avg rmse
Springfield	categories	SVD	0.61	0.92	0.73	0.5	1.5
Springfield	business_reviews	SVD	0.6	0.91	0.72	0.5	1.42
Springfield	business_tips	SVD	0.61	0.95	0.74	0.54	1.16
Springfield	categories_and_reviews	SVD	0.6	0.91	0.72	0.49	1.46
Springfield	combined	SVD	0.6	0.91	0.72	0.5	1.45
Springfield	friendships	SVD	0.6	0.95	0.73	0.55	1.28
Springfield	priority_combined	SVD	0.61	0.9	0.73	0.49	1.37
Springfield	threshold_categories	SVD	0.6	0.91	0.73	0.5	1.41
Springfield	categories	NormalPredictor	0.57	0.61	0.59	0.33	1.91
Springfield	business_reviews	NormalPredictor	0.52	0.56	0.54	0.31	1.96
Springfield	business_tips	NormalPredictor	0.57	0.61	0.59	0.35	1.57
Springfield	categories_and_reviews	NormalPredictor	0.57	0.6	0.58	0.33	1.89
Springfield	combined	NormalPredictor	0.55	0.58	0.57	0.32	1.84
Springfield	friendships	NormalPredictor	0.59	0.68	0.63	0.39	1.67
Springfield	priority_combined	NormalPredictor	0.57	0.61	0.59	0.34	1.64
Springfield	threshold_categories	NormalPredictor	0.55	0.55	0.55	0.3	1.95
Springfield	categories	KNNBasic user_based	0.58	0.86	0.69	0.47	1.54
Springfield	business_reviews	KNNBasic user_based	0.57	0.89	0.7	0.5	1.51
Springfield	business_tips	KNNBasic user_based	0.59	0.95	0.73	0.56	1.21
Springfield	categories_and_reviews	KNNBasic user_based	0.57	0.86	0.69	0.47	1.53
Springfield	combined	KNNBasic user_based	0.58	0.86	0.69	0.48	1.53
Springfield	friendships	KNNBasic user_based	0.59	0.87	0.7	0.51	1.3
Springfield	priority_combined	KNNBasic user_based	0.59	0.82	0.69	0.46	1.42
Springfield	threshold_categories	KNNBasic user_based	0.58	0.86	0.69	0.48	1.48
Springfield	categories	KNNBasic item_based	0.58	0.86	0.69	0.46	1.55
Springfield	business_reviews	KNNBasic item_based	0.58	0.89	0.7	0.49	1.5
Springfield	business_tips	KNNBasic item_based	0.59	0.94	0.73	0.53	1.22
Springfield	categories_and_reviews	KNNBasic item_based	0.58	0.85	0.69	0.46	1.53
Springfield	combined	KNNBasic item_based	0.59	0.85	0.69	0.47	1.55
Springfield	friendships	KNNBasic item_based	0.59	0.89	0.71	0.51	1.3
Springfield	priority_combined	KNNBasic item_based	0.59	0.82	0.69	0.45	1.42
Springfield	threshold_categories	KNNBasic item_based	0.59	0.86	0.7	0.46	1.5

Table of results of metrics for each algorithm for each type of connection