

# CAC Project

## Social Network Analysis Recommender Systems

### Group H

André Barbosa – 202007398

José Araújo – 202007921

José Ribeiro – 202007231





# Table of contents

**01**

## **Introduction**

Project description and  
goals to be achieved

**02**

## **Dataset**

Analysis of the content  
and data processing

**03**

## **SNA**

Social Network analysis  
and implementation

**04**

## **RS**

Recommender System  
implementation and  
analysis

**05**


## **Metrics**

Metrics to evaluate our  
results

**06**

## **Conclusion**

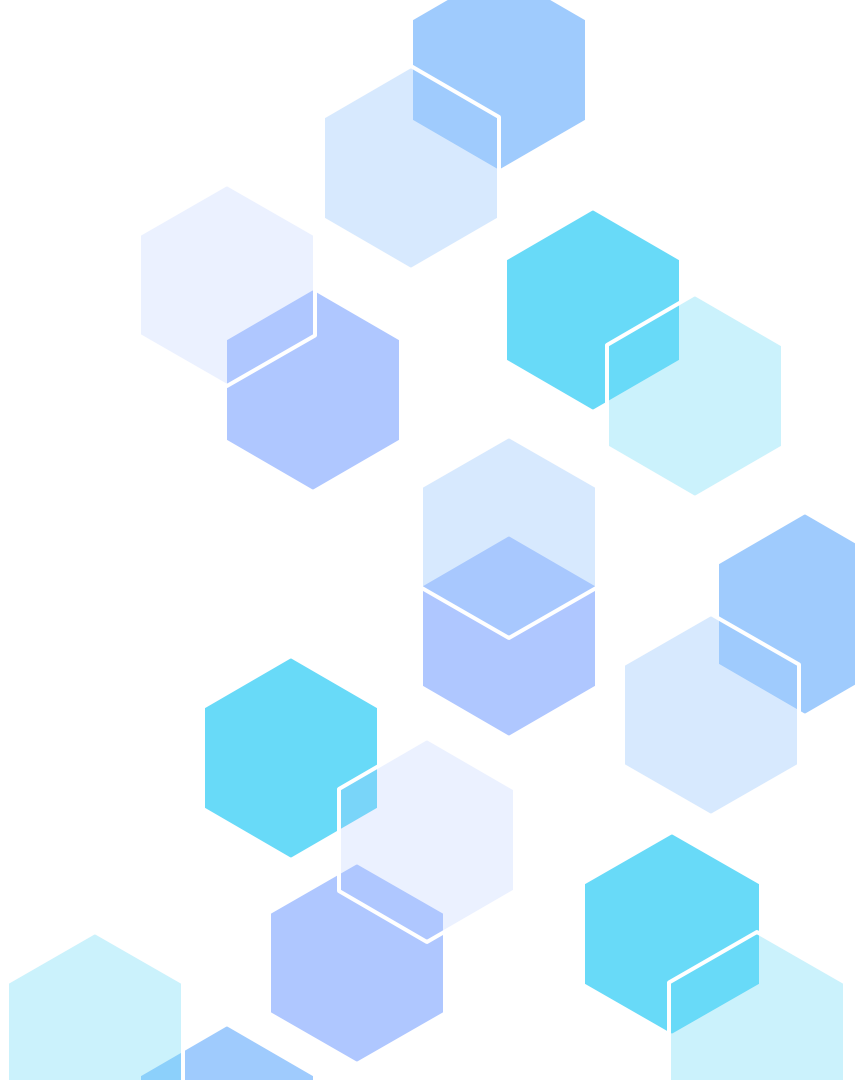
Main takeaways from  
work developed



---

01

# Introduction



# Project Description



**SNA**

Creation of a Social Network to  
connect recipes and find  
communities



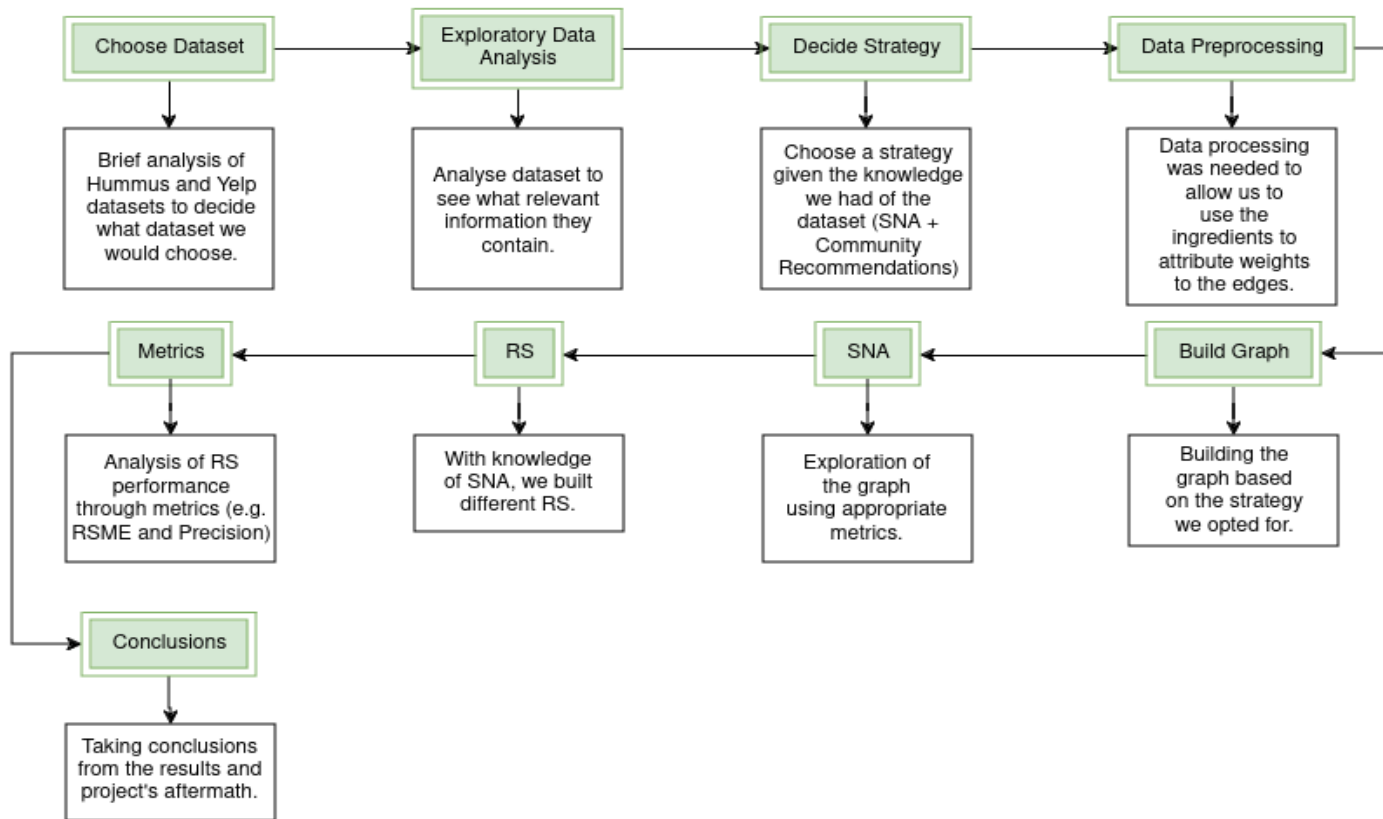
**RS**

Recommend to members other  
recipes from a same cluster on the  
Social Network

# Project Goals

- Community-Based Recommendation approach.
- Social Network created with different amount of edges.
- Communities are created with clusters in the Social Network.
- Multiple Algorithms application to full Social Network and respective clusters.
- Performance and metrics evaluation.
- Recommendation of top-N recipes for a specific member.

# Project Pipeline

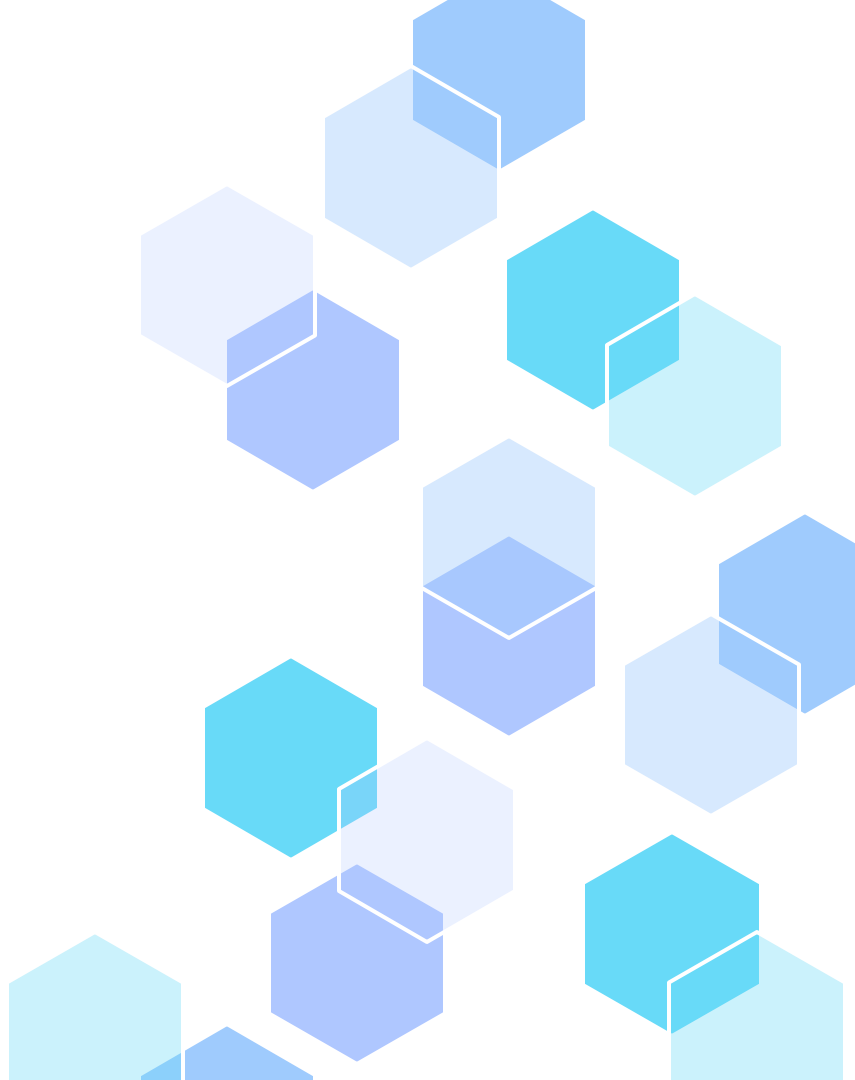


---

# 02

# Dataset

Hummus Dataset



# Data Files

- The Hummus Dataset, used in this project, contains three csv files:
  - members.csv
  - recipes.csv
  - reviews.csv





# Members Dataset

- Includes id, name, description, status, avg\_rating, follows, etc.
- Contains 299583 members.
- Most columns have 75% or more null values.



# Recipes Dataset

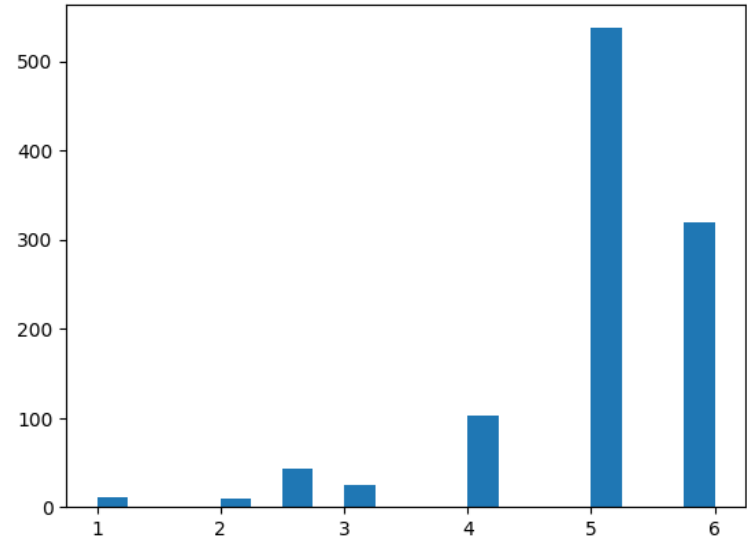
- Includes id, title, description, directions, ingredients, average\_rating, nutritional values and scores.
- Contains 507335 recipes.
- Description and ingredients have a good quantity of textual information for analysis.



# Reviews Dataset

- Main relation between members and recipes, includes both ids, **rating**, text, likes and review\_id.
- Contains 1916424 reviews.
- Rating is the most relevant column to relate member and recipe, despite highly skewed to higher values.

*Ratings distribution (1 – lowest, 6 – highest)*



# Data Preprocessing

As we'll show later, in order to use the ingredients of each recipe in the graphs, we first needed to pre-process the **ingredients** column of the recipes dataset.

The final pre-processed version had therefore an additional column named **ingredients\_pp**, which consisted of a list of ingredients (without their quantities) of each recipe.

```
import ast

# x for the row, ing_or_quant for the result column to return, ingredients (0) or
def ing_process(x, ing_or_quant):

    try:
        ing_list = ast.literal_eval(x)
    except:
        print(x)
        return None

    try:
        res = list(ing_list.values())[0]
    except:
        print(ing_list)
        return None

    # for the ingredients return the

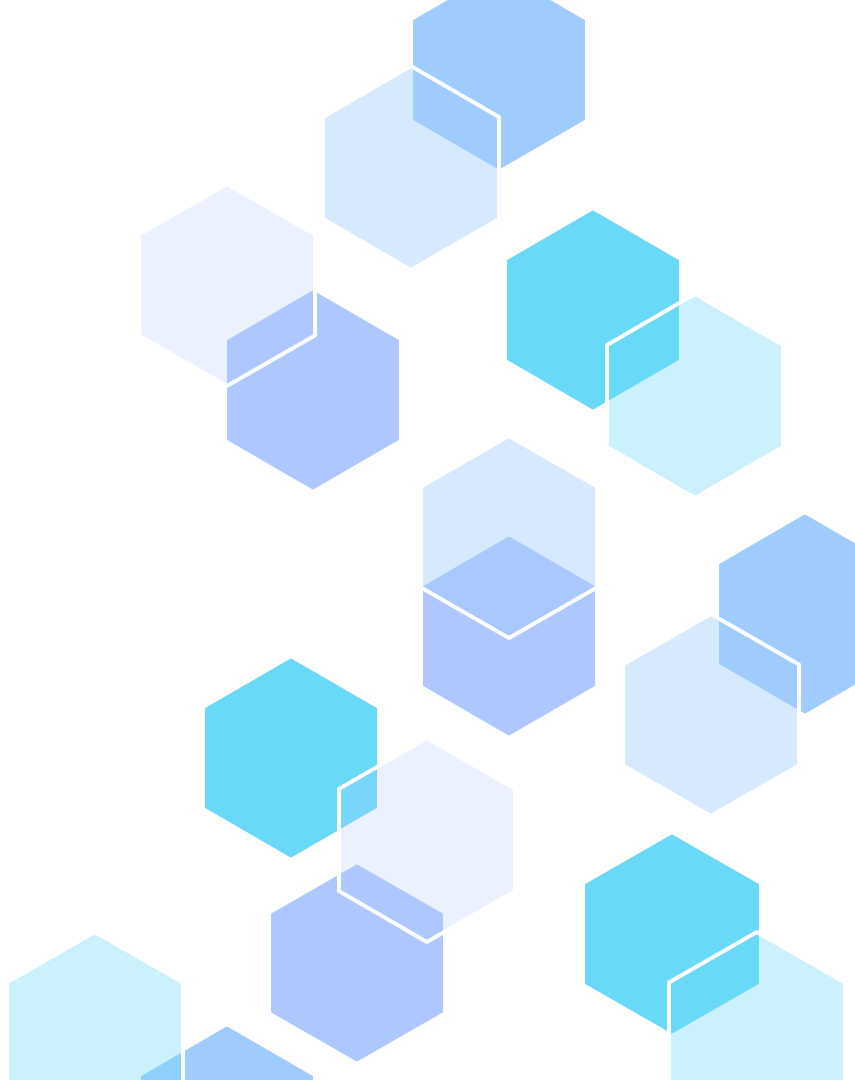
    return [x[ing_or_quant] for x in res]

recipes['ingredients_pp'] = recipes['ingredients'].apply(ing_process, args=(0,))
```

---

# 03

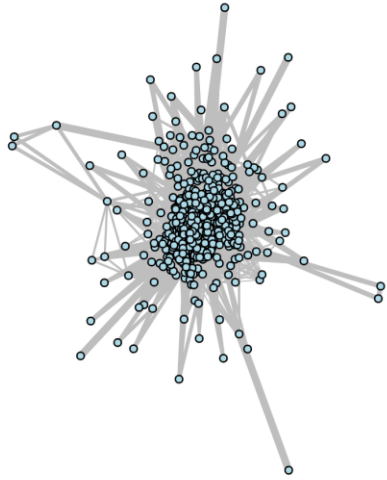
## SNA



# Graph

- **Node:** Recipe
- **Edge:** TF-IDF coefficient
  - Calculate frequency of each ingredient
  - Check if an ingredient is common in both nodes
  - If it is, divide 1 by frequency and add to weight of the edge (inverse of frequency)

# Graph



- **Node count:** 1000
- **Edge count:** 158128

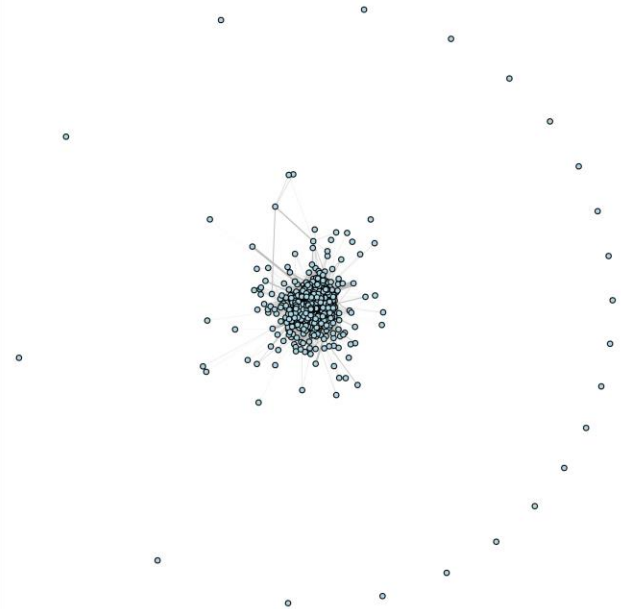
# Graph Simplification

We tried to simplify the graph but noticed that nothing had changed.

- **Node count after simplification:** 1000
- **Edge count after simplification:** 158128

The original graph did not contain any redundancies, unnecessary edges, or nodes that could be combined or removed without changing the structure or properties of the graph.

The graph has 0 loops.





# SNA Statistics Comparison

We applied two different strategies to get the nodes that would compose the Social Network:

- Collect the 1000 nodes that had higher amount of reviews (to collect the most amount of information as possible) – G(reedy) SN – that revealed the best results.
- Collect 1000 reviews that would be balanced in the rating attributed and select the recipes associated with it (to collect a more balanced rating distribution) – B(alanced) SN.

For the first network, we also applied two variations, where we reduced the number of edges, by applying a minimum acceptable weight (strong links remain):

- (weight  $\geq 0.01$ ) – M(edium) G(reedy) SN.
- (weight  $\geq 0.1$ ) – S(trong) G(reedy) SN.

# SNA Statistics Comparison

SNA	Density	Diameter	Avg. Path Len.	Connectivity	Node degree	Closeness	Betweenness	Eccentricity	Bridges	Homophily	Hubs and Authorities
GSN	0.31	5	1.7313	22	665	0.5933	350.1	3.29	4	+0.200	0.5035
BSN	0.2069	5	1.8888	22	608	0.5404	425.48	3.73	13	+0.192	0.3700
MGSN	0.118	5	2.04	35	359	0.496	485.96	3.72	8	+0.148	0.279
SGSN	0.007	10	3.984	163	36	0.265	1031.8	6.03	70	+0.158	0.0822

- GSN - Greedy Social Network
- BSN – Balanced Social Network
- SGSN – Strong Greedy Social Network
- MGSN – Medium Greedy Social Network

# SNA Statistics



## Density

0.3166

The graph is relatively dense, with a significant portion of possible node connections present. Moderate to high connectivity within the graph.



## Diameter

5

The graph is relatively large and interconnected, with paths of moderate length between its nodes.

# SNA Statistics



## Average Path Length

1.7313

The graph is more of a tightly connected graph, where nodes are closer to each other, and information can propagate quickly in the network.



## Connectivity

The graph is not composed of one giant component, being partitioned into distinct subset of nodes, suggesting segmentation within the graph and that some nodes are isolated.

**Connected:** False

**Connected Components:** 22

**Component sizes:**

```
[979, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]
```

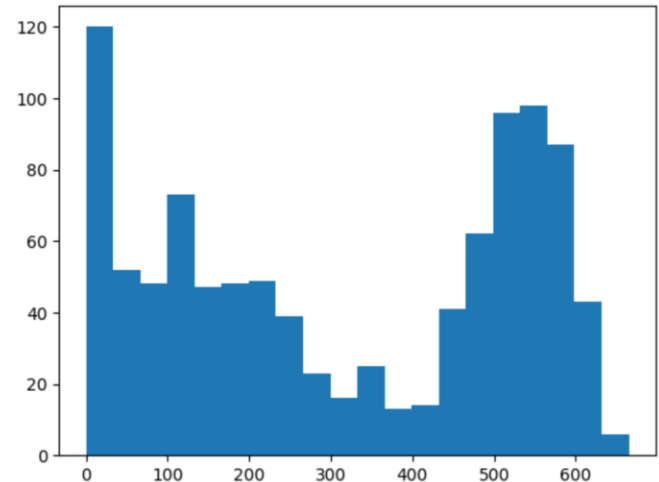
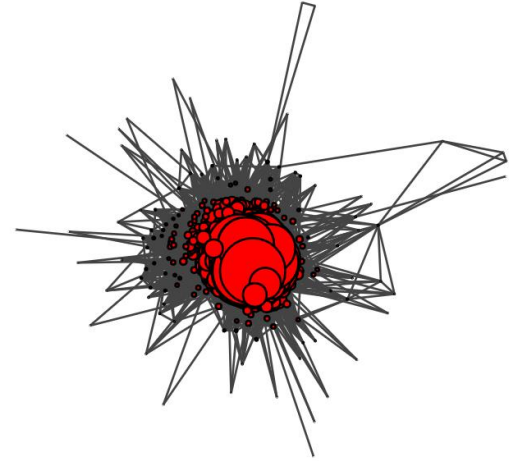
# SNA Statistics

## Node degrees

- Maximum node degree of 665 and the histogram value distribution suggests that there is a highly connected core of the network
- Minimum node degree of 0 suggests that there are recipes (nodes) in the graph that are not connected to any other nodes.
- There is heterogeneity in the node connectivity.

**Maximum:** 665

**Minimum:** 0

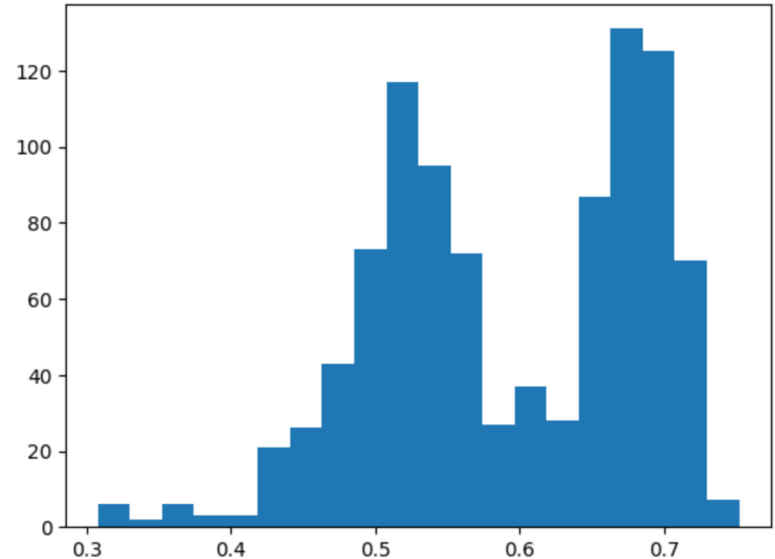


# SNA Statistics

## Closeness

- The average closeness of 0.5933 suggests that the nodes in the graph are relatively efficient in communicating with each other.
- The peaks at 0.45–0.55 and 0.65–0.75 suggest heterogeneity in the efficient communication of nodes.

**Average:** 0.5933

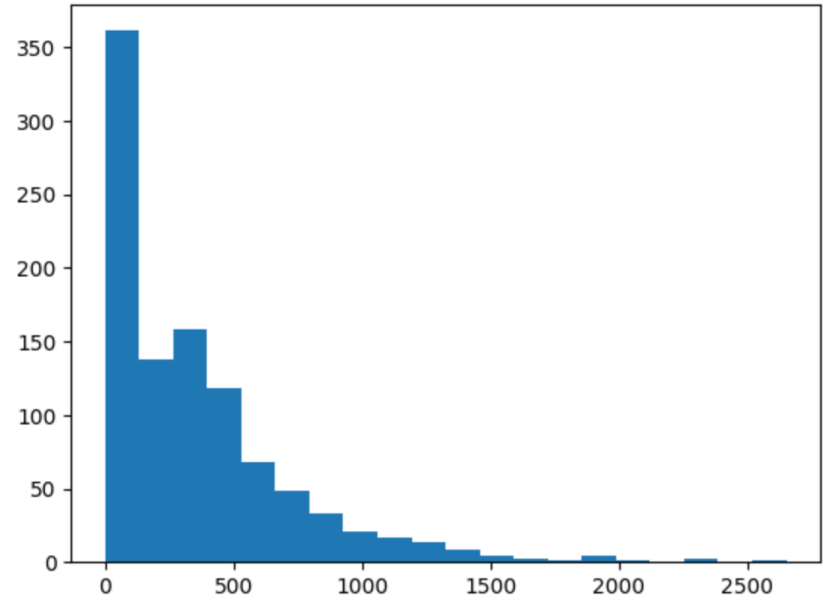


# SNA Statistics

## Betweenness

- Value suggests a node falls on the shortest path between 2 other nodes 350.1 times.
- There are a small number of nodes with high betweenness (hubs) and many nodes with lower values.
- The graph features a community structure, with possible identifiable clusters, and also there are some nodes highly influential with many connections

**Average:** 350.1

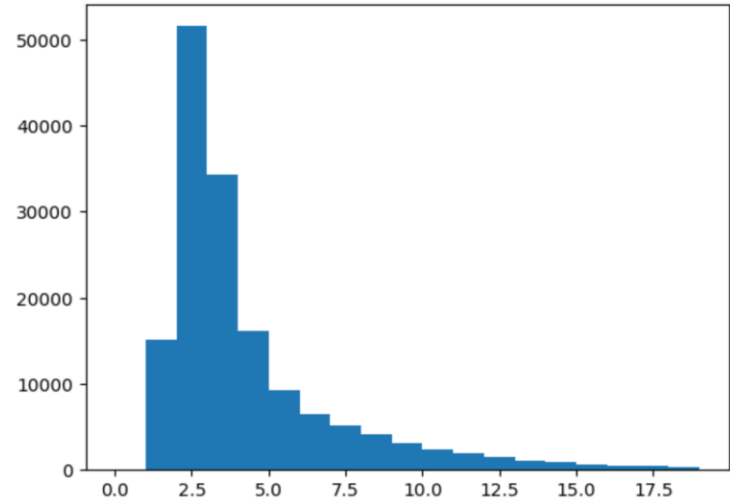
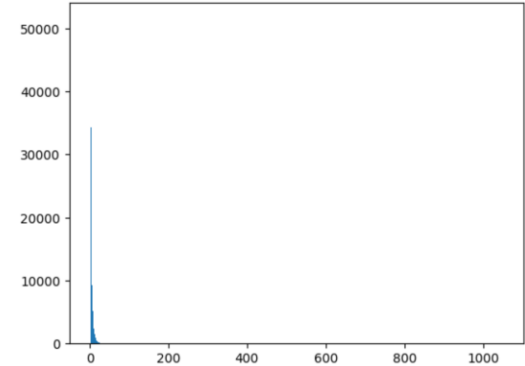


# SNA Statistics

## Edge betweenness

- An edge falls on the shortest path between two other nodes 5.2415 times.
- The graph likely has a dense structure within communities, with many alternative paths between nodes within those communities.
- Removing a high edge betweenness centrality edge could have a more significant impact on network flow compared to removing an edge within a community.

**Average:** 5.2415





# SNA Statistics

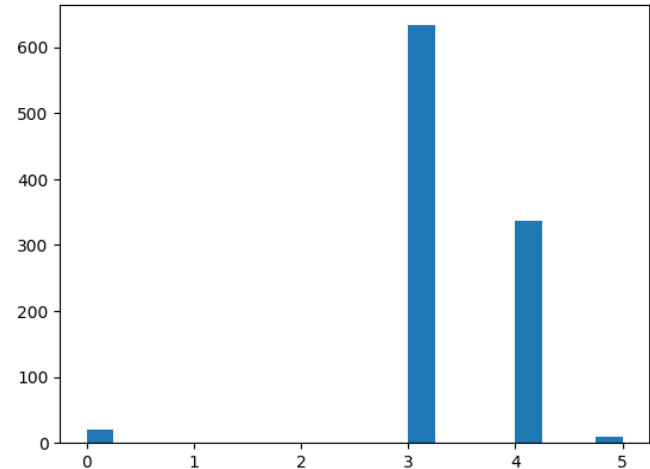
## Eccentricity

- The graph is densely connected
- The maximum value being 5 (diameter) suggests that information disseminates fast through the network, in small few steps

**Average:** 3.292

**Maximum:** 5

**Minimum:** 0



# SNA Statistics



## Bridges

These bridges represent some level of vulnerability due to being weak points in the network.

N° of bridges: **4**



## Homophily

- Weak but considerable tendency for recipes in the graph to be connected to other recipes similar to it.
- Nodes with many connections are more likely to connect to other hubs, and nodes with few connections are more likely to connect to other low-degree nodes.
- There is still some randomness in this values.

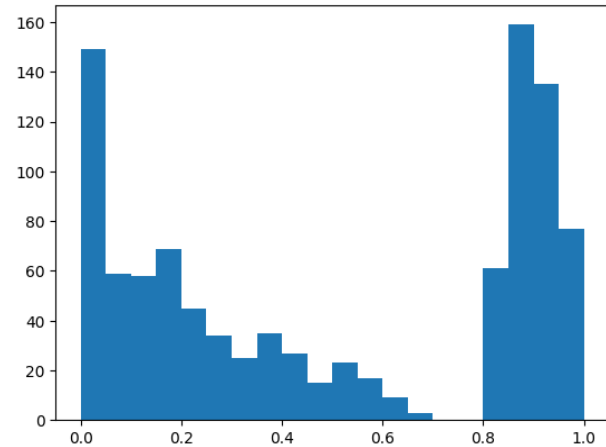
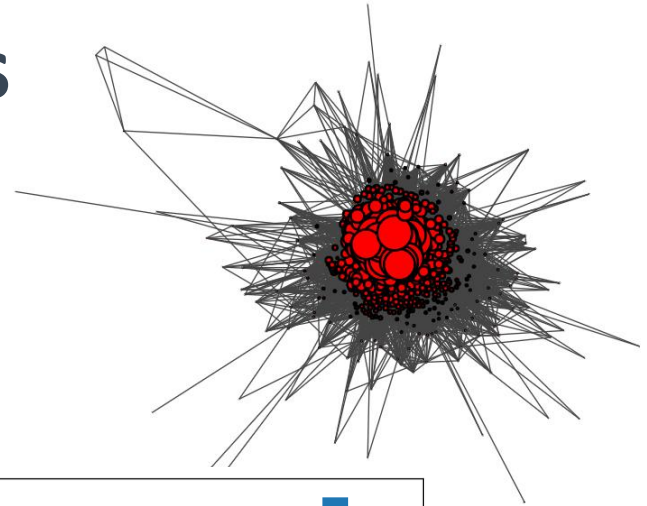
**Assortativity degree: +0.200**

# SNA Statistics

## Hubs and Authorities

- A significant portion of nodes are hubs and authorities
- A significant portion of nodes are peripheral nodes with not much influence in the network.
- There is a well-defined core of highly connected and influential nodes, and a larger periphery of less connected nodes.
- Hubs are the same as authorities given that the graph is undirected.

**Average:** 0.5035



# SNA Statistics

## Small World Theory

Suggests that even in large networks, most individuals are reachable from any other individual through a small number of intermediaries, applying the notion of "six degrees of separation."

### Conclusion:

- Short average path lengths comparable to other graphs
- High level of clustering
- Properties of SW verified

Average path length	1.7313
Average clustering coefficient	0.7505
* Random graph APL	1.6834
* Random graph ACC	0.3166
* Regular graph APL	250.2503
* Regular graph ACC	0.0

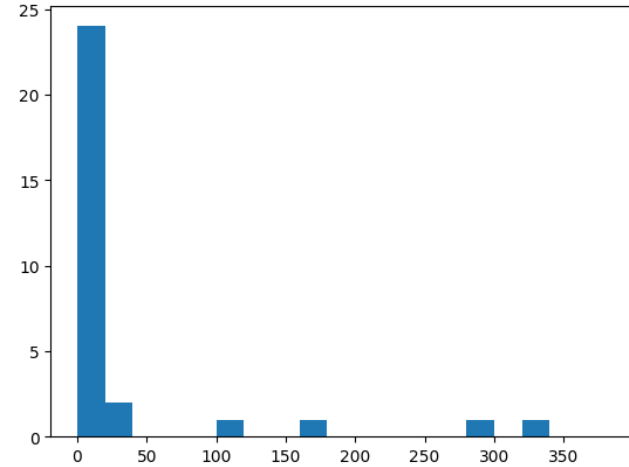
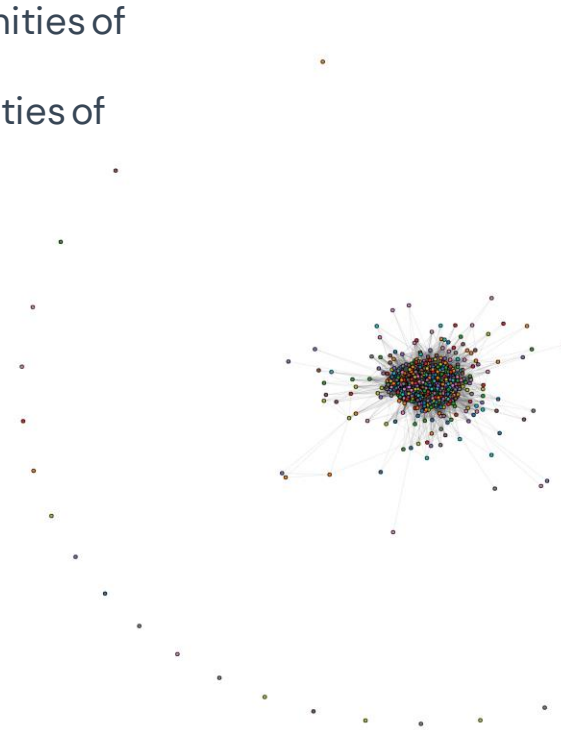
\* Similar graphs to our graph to serve as a comparison.

# SNA Statistics

## Clusters

- High number of small communities of recipes
- There are few larger communities of recipes

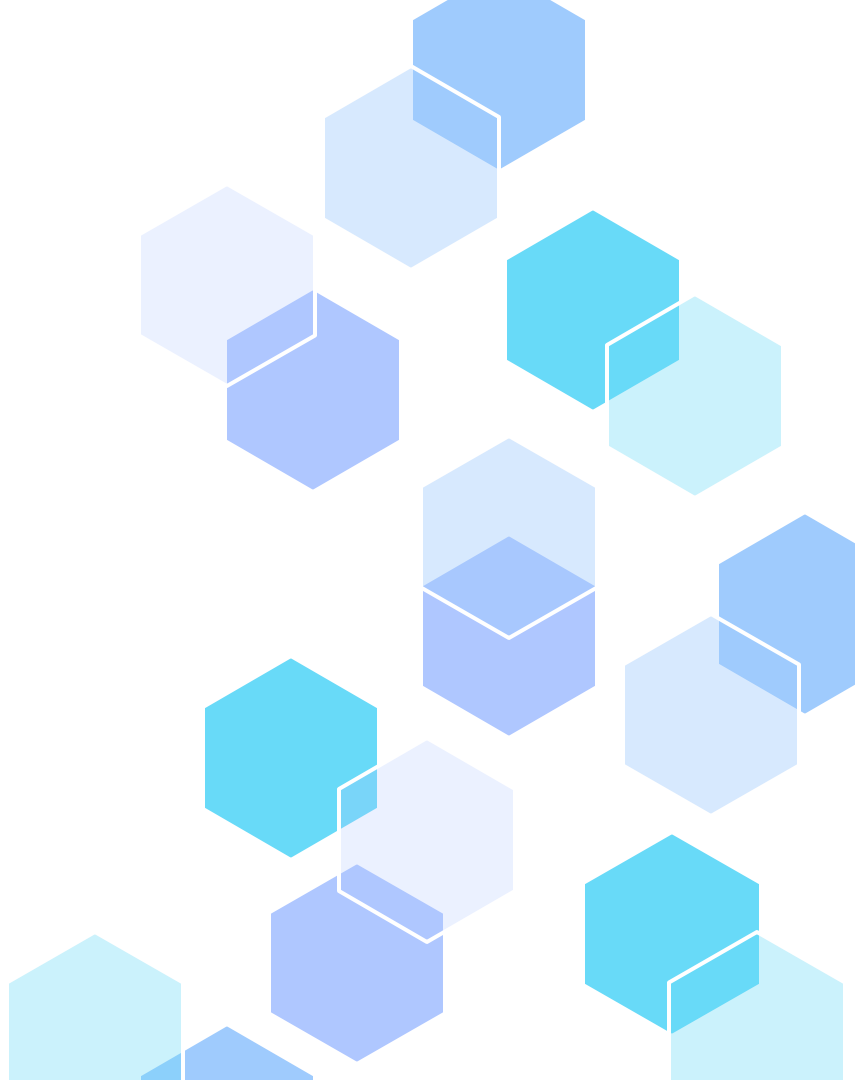
**No. of clusters: 30**



---

# 04

## RS



# Collaborative Filtering

## Item-based

- To train different models, we applied a train and test set division on the cluster with the highest number of recipes.
- We used 2 different algorithms:
  - **SVD** (Singular Value Decomposition)
  - **KNNWithMeans**. For this second algorithm, we used two different configurations, one where it was applied Cosine Similarity and the other the Pearson Correlation.

## User-based

- On the User Based collaborative filtering, we applied the last algorithm from the Item-Based approach (**KNNWithMeans**), with the same two variations, but with a User-based configuration.

# Collaborative filtering

- After training the models, we apply to each pair of member and item it, to predict a rating.
- To obtain recommended items, we indicate a user id and the number of recommendations.
- The recommender sorts the predicted ratings in descending order and selects the ones the user had never rated before.



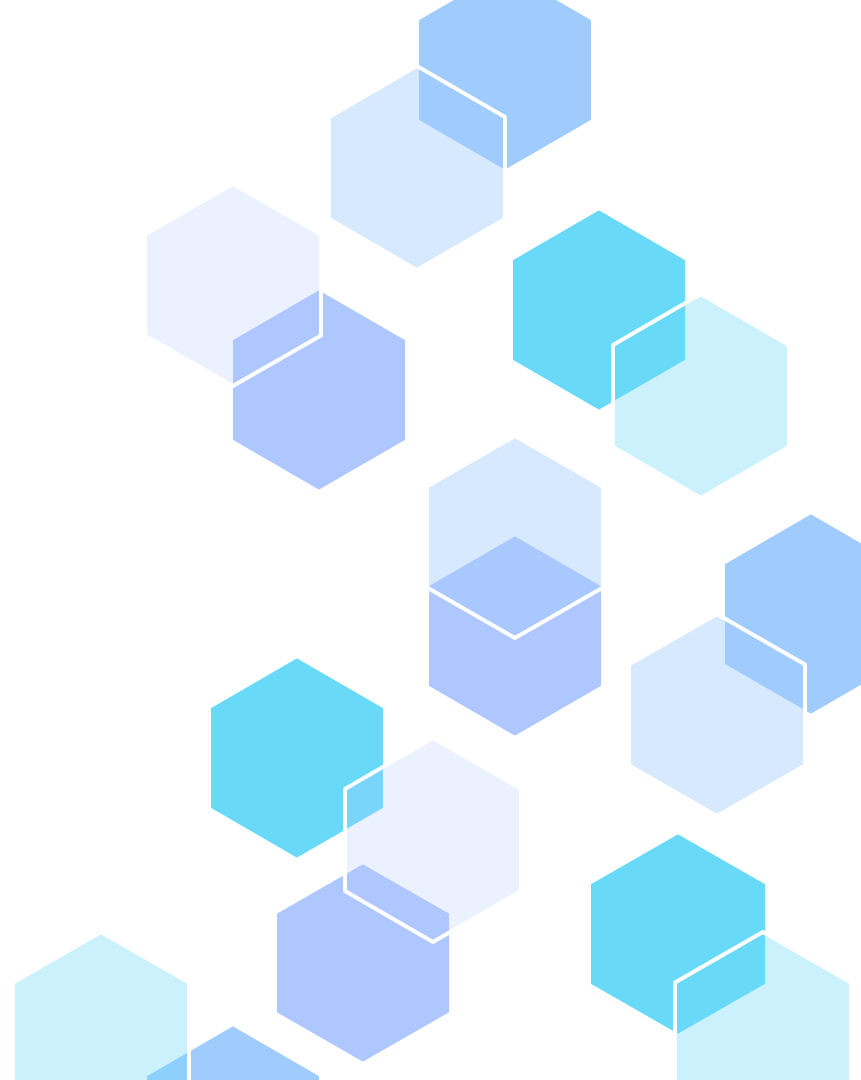
# Content-based filtering

- Based on **recipe's description** rather than relying on user-item interactions or similarities between users.
- Resorted to **TF-IDF vectorizer** used to convert recipe descriptions into numerical feature vectors.
- TF-IDF assigns weights to words based on their frequency in a document relative to their frequency in the whole column.
- Similarity calculated via cosine similarity.
- For each recipe, determine 20 similar recipes.
- **For a given user:**
  - Fetch for each positively rated recipe (>3 in rating) recommended recipes.
  - Return top 10 recipes by rating.

---

05

# Metrics



# Item-based – different datasets

Graph	Division	Model	RMSE	Precision	Recall	F1-score
GSN	Cluster	SVD	0.9239	0.9381	1.0	0.9681
	Total		0.9753	0.9128	1.0	0.9544
BSN	Cluster		1.0048	0.9033	0.9959	0.9474
	Total		0.9993	0.8834	0.9952	0.9359
MGSN	Cluster		0.9728	0.9366	1.0	0.9673
SGSN			1.1030	0.9186	1.0	0.9576
GSN	Cluster	KNN	0.9766	0.9375	0.9899	0.9630
	Total		1.1156	0.9125	0.9782	0.9442
BSN	Cluster		1.1141	0.9011	0.9713	0.9349
	Total		1.1184	0.8853	0.9776	0.9292
MGSN	Cluster		1.0444	0.9362	0.9948	0.9646
SGSN			1.1289	0.9176	0.9873	0.9512

# Collaborative filtering – Top Cluster

Models	RMSE	Precision	Recall	F1-score
Item-based SVD	0.9239	0.9381	1.0	0.9681
Item-based KNN Cosine	0.9766	0.9378	0.9949	0.9655
Item-based KNN Pearson	0.9766	0.9378	0.9949	0.9655
User-based KNN Cosine	0.9869	0.9375	0.9899	0.9630
User-based KNN Pearson	0.9863	0.9375	0.9900	0.9630

# Item-based CF

SVD

	title	description	average_rating
380092	Pumpkin Cream Cheese Roll	Sundance Inn Bed & Breakfast, New Braunfels, T...	4.88
273673	Strawberry Coffee Cake	Good for a special occasion breakfast.	4.82
41211	Strawberries & Cream Bread (Strawberry or Blue...	A wonderful recipe using fresh strawberries fr...	4.80
92949	Nestle Toll House Chocolate Chip Pan Cookie	This recipe was in my paper recently and I mad...	4.72
323331	Legal Seafood Style Baked Scallops	Whenever we get back to the Boston area, it's ...	4.70
396146	World's Best Cinnamon Raisin Bread (Not Bread ...	This is the best Cinnamon Raisin Bread I've ev...	4.63
170392	Buttermilk Jalapeno Cornbread	NaN	4.63
87555	Quick Yellow Cake	You can mix this up in a hurry.	4.43
82180	Magnolia Bakery's Vanilla Birthday Cake and Fr...	These comments are from their cookbook: \r\n\r...	4.40
506617	Quick Cinnamon Rolls - No Yeast	These wonderful cinnamon rolls are on the tabl...	3.99

KNN Cosine

	title	description	average_rating
139181	Why-I-Joined-Zaar Carrot Cake	On October 2, 2001, I discovered the Recipezaa...	4.87
184572	Amish White Bread	This is my favorite home baked bread recipe. I...	4.84
187192	Pecan Pie	A wonderfully sweet pecan pie that is perfect....	4.82
330669	Ho Ho Cake	This is like a giant Hostess Ho Ho. The creme ...	4.82
434869	The Best Apple Pie Muffins Ever	I got this recipe from a friend. I make thes...	4.81
242558	Chocolate Syrup	This homemade chocolate syrup is just as good ...	4.81
361946	Hershey's Chocolate Cake With Frosting	One night I was craving chocolate cake, but we...	4.80
249131	My Favorite Chocolate Chip Cookies	Get a glass of cold milk ready! This has been ...	4.69
320636	Sugar Cookie Icing	This recipe is quick, easy, dries hard and shi...	4.56
371112	The Ultimate Lemon Meringue Pie	I made this pie this weekend. It's adapted fro...	4.54

# Item-based CF

KNN Pearson

	title	description	average_rating
139181	Why-I-Joined-Zaar Carrot Cake	On October 2, 2001, I discovered the Recipezaa...	4.87
184572	Amish White Bread	This is my favorite home baked bread recipe. I...	4.84
187192	Pecan Pie	A wonderfully sweet pecan pie that is perfect....	4.82
330669	Ho Ho Cake	This is like a giant Hostess Ho Ho. The creme ...	4.82
434869	The Best Apple Pie Muffins Ever	I got this recipe from a friend. I make thes...	4.81
242558	Chocolate Syrup	This homemade chocolate syrup is just as good ...	4.81
361946	Hershey's Chocolate Cake With Frosting	One night I was craving chocolate cake, but we...	4.80
249131	My Favorite Chocolate Chip Cookies	Get a glass of cold milk ready! This has been ...	4.69
320636	Sugar Cookie Icing	This recipe is quick, easy, dries hard and shi...	4.56
371112	The Ultimate Lemon Meringue Pie	I made this pie this weekend. It's adapted fro...	4.54

# User-based CF

KNN Cosine

	title	description	average_rating
111626	Linda's Cheesecake-Stuffed Strawberries	OMG.. These are like heaven in your mouth! Ta...	4.91
228943	Intensely Chocolate Cocoa Brownies	A rich brownie that is mixed in a sauce pan on...	4.86
434869	The Best Apple Pie Muffins Ever	I got this recipe from a friend. I make thes...	4.81
243084	Whole Wheat Pancakes	These are SO delicious - the honey & whole whe...	4.80
49800	Sue B's Chocolate Cake	I adopted this recipe. I had prepared it for ...	4.80
182255	Unknownchef86's Very Best Dinner Rolls	These rolls are very soft and light. They make...	4.76
208219	Peanut Butter Balls	My mom helped me post this recipe. These candi...	4.76
296793	Banana-Chocolate Chip Muffins	WOW! Get this...a very low-fat muffin recipe t...	4.69
298216	Blueberry-Oatmeal Muffins	NaN	4.66
182190	Lemon Bars	Don't blame me if you get addicted to these bars!	4.55

KNN Pearson

	title	description	average_rating
111626	Linda's Cheesecake-Stuffed Strawberries	OMG.. These are like heaven in your mouth! Ta...	4.91
228943	Intensely Chocolate Cocoa Brownies	A rich brownie that is mixed in a sauce pan on...	4.86
187192	Pecan Pie	A wonderfully sweet pecan pie that is perfect....	4.82
434869	The Best Apple Pie Muffins Ever	I got this recipe from a friend. I make thes...	4.81
243084	Whole Wheat Pancakes	These are SO delicious - the honey & whole whe...	4.80
49800	Sue B's Chocolate Cake	I adopted this recipe. I had prepared it for ...	4.80
182255	Unknownchef86's Very Best Dinner Rolls	These rolls are very soft and light. They make...	4.76
296793	Banana-Chocolate Chip Muffins	WOW! Get this...a very low-fat muffin recipe t...	4.69
298216	Blueberry-Oatmeal Muffins	NaN	4.66
182190	Lemon Bars	Don't blame me if you get addicted to these bars!	4.55

# Content-based filtering

## Whole Network

	title	description	average_rating
90522	Big Grandma's Best Peanut Butter Cookies	This is my husband's grandmother's recipe that...	4.86
89747	Soft, Spicy, Heavenly Ginger Cookies	I found the original recipe for these cookies ...	4.85
79666	Lisa's Swirled Chocolate Chip Cookies	This is a chocolate chip cookie recipe that my...	4.84
236714	Pulled Pork (Crock Pot)	I found this pulled pork recipe years ago, and...	4.83
214348	Oreo Truffles	These are so easy to make and so easy to eat! ...	4.81
159654	World's Best Butter Cookies	I tested 8 different butter cookie recipes, lo...	4.81
52627	Meatloaf Barbecue Style	This is our favorite meatloaf recipe. It's eas...	4.74
323678	Flourless Peanut Butter Cookies	This flourless peanut butter cookie recipe is ...	4.73
58863	Fudge Crinkles (A Great 4 Ingredient Cake Mix ...	These are chewy, fudgy, SUPER EASY cookies tha...	4.72
298110	Granny's Sugar Cookies	I make sugar cookies for almost every holiday....	4.70

## Top Cluster

	title	description	average_rating
200676	Basic Machine French Bread	This is my winner with Tammiev and our friends...	4.86
90522	Big Grandma's Best Peanut Butter Cookies	This is my husband's grandmother's recipe that...	4.86
89747	Soft, Spicy, Heavenly Ginger Cookies	I found the original recipe for these cookies ...	4.85
79666	Lisa's Swirled Chocolate Chip Cookies	This is a chocolate chip cookie recipe that my...	4.84
385241	Deep Dark Chocolate Cake	Here it is -- the best chocolate cake ever! Pe...	4.83
159654	World's Best Butter Cookies	I tested 8 different butter cookie recipes, lo...	4.81
148228	The Ultimate Strawberry Shortcake	The perfect summer dessert. Not your ordinary ...	4.77
176189	Rice Pudding	My husband's favorite!	4.77
169225	Soft Chocolate Chip Cookies	These are my kids' favorite chocolate chip coo...	4.76
154692	Healthy Honey Oatmeal Cookies	These are easy to make! and healthy for you to...	4.76



# Result Analysis

- SVD showed better results than KNN.
- As expected, item-based CF had slightly better results, given the clusters were based on item similarities.
- Both Content-based and Collaborative filtering with a cluster instead of the whole network bring more specific and accurate results, increasing also the ratings.



# Item-based SVD Results

## Selected User's ratings

	title	ingredients	average_rating
335495	Chicken Tortilla Soup II	{': [{"carrots, diced', '8 time(s) ounces', 'celery', '8 time(s) ounces', 'onions, diced', '8 time(s) ounces', 'garlic powder or 1 \$template2\$, diced', '0.5 time(s) teaspoon', 'salt', '0.125 time(s) teaspoon', 'pepper', '0.25 time(s) teaspoon', 'corn oil', '1 time(s) tablespoon', 'chicken broth', '4 time(s) (15 ounce) cans', 'tomatoes, diced (optional)', '1 time(s) (15 ounce) can', 'Rotel tomatoes & chilies, diced', '1 time(s) (10 ounce) can', 'I use McCormicks', '1 time(s) (1 1/4-1 1/2 ounce) packet taco seasoning', 'corn, tortillas (NOT FLOUR, cut into small pieces, about 1-in x1-in)', '10 time(s) (8 inch)', 'chicken meat, poached, diced', '12 time(s) ounces', 'milk or 1 cup \$template2\$', '1 time(s) cup', 'blend cheese, shredded', '12 time(s) ounces monterey jack cheese or 12 ounces Mexican', 'corn tortilla chips, broken into small pieces for garnish', 'time(s) '}]}	4.86
338497	Southwestern Baked Spaghetti	{': [{"uncooked spaghetti', '8 time(s) ounces', 'milk', '0.5 time(s) cup', 'egg', '1 time(s)', 'ground beef', '1 time(s) lb', 'medium onion, chopped', '1 time(s)', 'medium green bell pepper, chopped', '1 time(s)', 'garlic cloves, minced', '2 time(s)', 'chili powder', '1 time(s) teaspoon', 'cumin', '0.5 time(s) teaspoon', 'oregano', '0.5 time(s) teaspoon', 'salt', '0.5 time(s) teaspoon', 'pepper', '0.25 time(s) teaspoon', 'tomato sauce', '2 time(s) (8 ounce) cans', 'shredded cheddar cheese', '0.5 time(s) cup', 'shredded monterey jack cheese', '0.5 time(s) cup '}]}	4.64
476454	Turn Your Crock Pot Into a Smokehouse Chicken (Smoked Chicken)	{': [{"whole chicken', '1 time(s)', 'liquid smoke', '0.25 time(s) cup '}]}	4.32
199328	Garlic Green Beans	{': [{"fresh green beans', '1 time(s) lb', 'water', '2 time(s) cups', 'cloves garlic, minced', '3 time(s)', 'butter', '3 time(s) tablespoons', 'salt', '0.5 time(s) teaspoon', 'pepper', '0.125 time(s) teaspoon '}]}	4.81

**We identified the following ingredients as relevant:**

Chicken, tortilla, chips, tomatoes, cheese, spaghetti pasta, pepper  
tomato sauce, ground beef, green beans

# Item-based SVD Results

Based on the presence of the previously identified ingredients, we decided whether a recipe should be considered relevant or not.

Precision @ 5: 0.8

Precision @ 10: 0.8

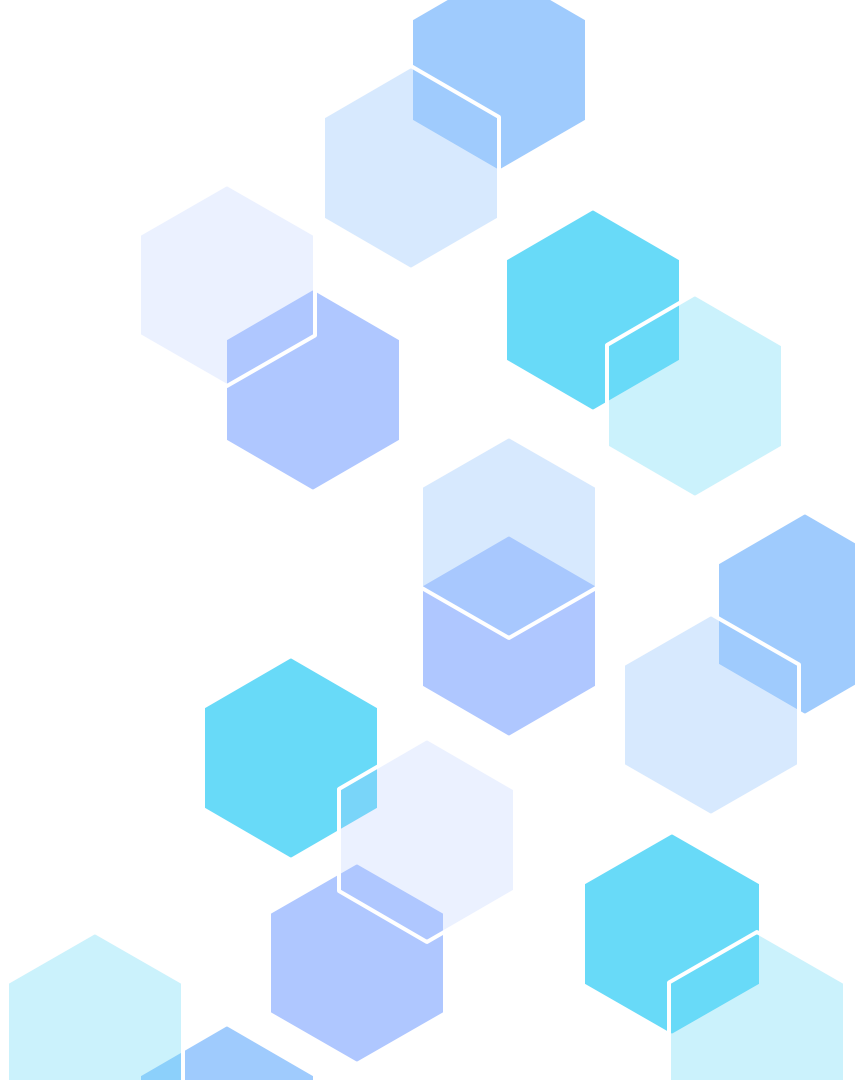
Precision @ 15: 0.87

	title	relevant
182811	Chicken Parmesan	1
410359	Teresa's Veal -or-Chicken Piccata	1
262240	Southwestern Stuffed Bell Peppers	1
464665	Kittencal's Easy and Delicious Ranch-Parmesan Chicken	1
162494	Hummus	0
88159	Kittencal's Pizza Sauce	1
94246	Bev's Spaghetti Sauce	1
124705	Charlie's Famous Chicken Salad With Grapes	0
401044	Oven-Fried Garlic Chicken	1
273500	Simple Baked Chicken Drumsticks	1
324319	Granny's Slow Cooker Vegetarian Chili	1
482131	Awesome Bacon-Tomato Dip	1
492983	Stove Top Tamale Pie	1
111366	Skylike Chili - Skyline Chili Copycat	1
316441	Creamy Burrito Casserole	1
349279	Mexican Stuffed Shells (Oamc)	1
55160	Baked Balsamic Chicken	0
184099	Hamburger Noodle Bake	1
488976	Pepsi Pork Roast	0
288022	Potato Kielbasa Skillet	1

---

06

# Conclusion



# Conclusion

- The Collaborative Filtering Item-based approach seemed to be the best one across all datasets, specially the SVD model.
- The method selection for the weight of the SN edges was adequate, having also good results with weight  $\geq 0.01$ .
- Using the top cluster we achieved better results than the whole network, so we were able to improve the performance by doing it.





---

# Thanks!