Culinary Polarization on Reddit: Analyzing Ethnic and Dietary Preferences in Online Food Communities

Alberto Corti[a] and Lorenzo Lattanzi[b]

[a] *University of Pisa, Department of Physics, a.corti5@studenti.unipi.it, Student ID: 598370*
[b] *University of Pisa, Department of Computer Science, l.lattanzi2@studenti.unipi.it, Student ID:597970*

Abstract

At the time of analysis, **r/food** ranked as the 22nd largest subreddit, boasting a membership of 24 million users. This platform serves as a hub for users to share and discuss their culinary preferences by posting original photos accompanied by descriptive titles of dishes they have prepared or consumed. Leveraging these structured interactions, we explore user culinary preferences through social network analysis techniques. Our study aims to understand how these preferences shape online behaviors and interactions within the community.

1 Introduction

This project explores the relationship between culinary preferences and opinion polarization on Reddit, a major platform for diverse discussions. Specifically, we examine whether users who post about foods from a particular ethnic cuisine or dietary preference (e.g., vegetarian or omnivorous) tend to engage primarily within their own culinary communities, limiting interactions with different cuisines. Additionally, we analyze whether these engagement patterns persist across different types of food-related posts, such as those tagged [I Ate] or [Homemade].

The project is divided into four main parts:

- **Data Collection and Network Construction**: This phase involves extracting data from a Reddit data dump, identifying key entities such as users (nodes) and their interactions (edges). We ensure the dataset includes at least 10,000–15,000 nodes for meaningful analysis. All code, methodologies, and dataset details are documented and shared via the project's GitHub repository.

- **Network Analysis**: We perform an in-depth analysis using Python libraries such as NetworkX or iGraph on the unweighted, undirected version of the graph, benchmarking our findings against synthetic models for comparison.

- **Spreading Task**: We simulate and analyze epidemic spreading on our graph, as well as on Erdős-Rényi and Barabási-Albert models, varying parameters and infection seed placements to assess their impact on diffusion dynamics.

- **Open Question**: This section explores whether culinary interests contribute to polarization in interactions on Reddit. We analyze user interaction patterns to determine if individuals cluster based on shared culinary preferences and whether these trends persist across different post categories, such as [I Ate] and [Homemade].

2 Data Collection and Network Construction

We obtained all posts and comments of the r/food subreddit from the data dump available at `https://the-eye.eu/redarcs/`. The data was converted into csv format, and we extracted the features needed for our analysis. To process the zst files and convert them into csv, we utilized a script available at `https://github.com/Watchful1/PushshiftDumps/blob/master/scripts/to_csv.py`, while all other scripts used in our analysis were manually developed. The extracted features included: author (of the post or comment), ID, parent ID (for comments), score, creation date, post titles, and comment text.

Using the authors who have submitted at least one post as nodes and defining links based on interactions between them (e.g., when Author A comments on Author B's post, and vice versa), we constructed an undirected weighted graph. The weight of a link represented the frequency of these interactions. Due to the large size of the r/food dataset, we restricted our analysis to posts and comments from 2022 to effectively manage computational demands.

3 Network Analysis

In the following section, we analyze the unweighted and undirected version of the graph created from the crawled data, as requested.

3.1 Synthetic Networks Construction

We analyzed several networks of comparable size to the one we constructed and extracted their key characteristics, specifically, we compared our network to Erdős–Rényi and Barabási–Albert models, ensuring the same number of nodes and a similar number of links for a meaningful comparison.

In this comparison, the probability p_{ER} of connecting each pair of nodes in an Erdős–Rényi graph was computed using the following expression:

$$p_{\mathrm{ER}} = \frac{2\mathrm{L}}{\mathrm{N}\cdot(\mathrm{N}-1)} = 3.25\cdot 10^{-4},$$

where N is the number of nodes and L is the number of links of the graph obtained from the crawled data.

The value m_{BA} for the Barabási-Albert graph, referred to as the number of edges to attach to each new node, was chosen as follows:

$$m_{\mathrm{BA}} = \left\lfloor \frac{\mathrm{L}}{\mathrm{N}} \right\rfloor = 3.$$

Graph	Nodes	Links
r/food (2022)	19549	62084
Erdős–Rényi	19549	61776
Barabási-Albert	19549	58638

Table 1: Number of nodes and links for the graphs.

3.2 Degree Distribution Analysis

We calculated several metrics related to the degree distribution as written in the following table.

Graph	Min Degree	Max Degree	Mean Degree
r/food	1	3373	6.35
Erdős–Rényi	0	17	6.32
Barabási-Albert	3	548	6.00

Table 2: Degree metrics.

Notably, the minimum degree in our graph is greater than zero, as there are no isolated nodes. This is expected since our network is constructed from an edge list, whereas an Erdős-Rényi (ER) graph may contain isolated nodes. In contrast, the Barabási-Albert (BA) model enforces a minimum degree of $m_{BA} = 3$, ensuring that each node has at least three connections. Additionally, the maximum degree in the ER graph confirms the absence of hub nodes, and our calculations indicate that the ER graph is in the supercritical regime.

As shown in Figure 1, the degree distribution of the r/food network differs significantly from that of an Erdős-Rényi graph. Unlike the ER model, which follows a Poisson distribution, r/food exhibits a power-law distribution, indicating the presence of highly connected nodes (hubs).

Figure 2 compares the r/food degree distribution with that of a Barabási-Albert model. The two are quite similar, as both follow a power-law distribution and contain hub nodes. However, they differ in key aspects:

- The r/food network contains nodes with fewer than $m_{BA} = 3$ links, whereas the Barabási-Albert model does not.
- The highest-degree nodes in r/food exceed the connectivity predicted by the Barabási-Albert model, suggesting that very active users play a more significant role in the real network than in the synthetic model.

A Poisson fit was applied to the Erdős-Rényi (ER) degree distribution, while a power-law fit was used for the other two distributions. According to theory, the degree distribution of an ER graph follows a Poisson distribution:

$$P(k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

where $\lambda = \langle k \rangle$ is the expected average degree. In our case, we found $\lambda = 6.32$ which is exactly the mean degree calculated from the ER graph.

For the *r/food* and Barabási-Albert (BA) degree distributions, we fitted a power-law function:

$$P(k) \propto k^{-\gamma}$$

where γ represents the exponent of the distribution. The fits yielded values of $\gamma = 2.62$ for *r/food* and $\gamma = 2.81$ for the BA model. Since γ is between 2 and 3, we are in the scale-free regime for the r/food network. However, we observe that the power-law fit does not fully capture the entire shape of the distribution. Specifically, the lower-degree nodes, which are the most frequent, do not align with the other nodes on the log-log plot. This suggests that these nodes deviate from the power-law behavior. According to the theory, we expect $\gamma = 3$ for the Barabási-Albert degree distribution, however, slight discrepancies in the numerical value of γ can arise due to the finite size of the system and statistical fluctuations within the network.

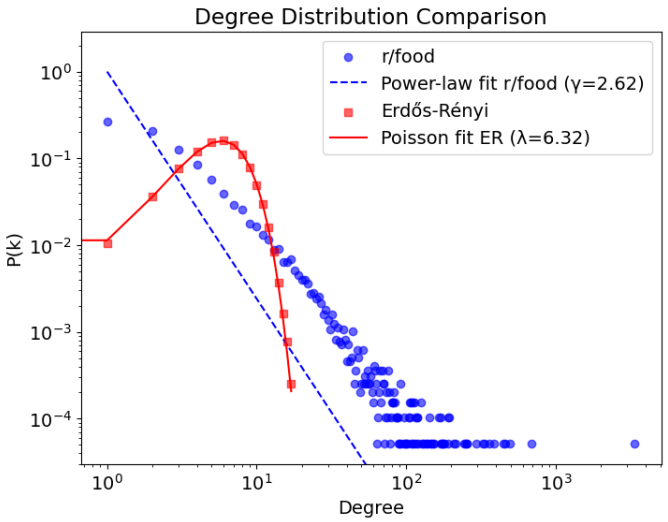

Figure 1: The r/food graph's degree distribution compared to the Erdős-Rényi model's degree distribution, fitted respectively with a powerlaw and a poisson distribution.

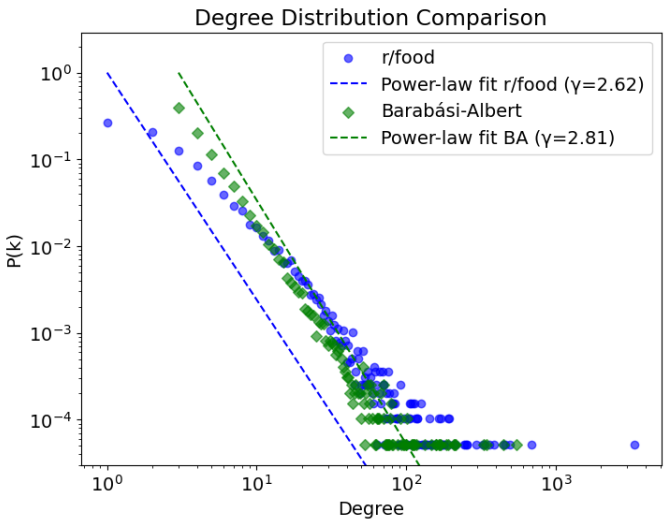

Figure 2: The r/food graph's degree distribution compared to the Barabási-Albert model's degree distribution, fitted with a powerlaw distribution.

3.3 Connected Component Analysis

Our constructed graph contains many connected components, primarily because many users are not very active and do not interact with more connected users. However, we identify a largest

component that includes 91.9% of the nodes in the entire graph. This phenomenon is also observed in the Erdos-Rényi graph with a 99.8% of nodes in the largest component, although, as expected, the Barabási-Albert graph is fully connected due to its construction method.

Graph	Connected Comp.	Nodes in the Largest Component (%)
r/food	1496	91.9%
Erdős–Rényi	38	99.8%
Barabási-Albert	1	100%

Table 3: Connected components metrics.

3.4 Path Analysis

The diameter and the average shortest path were calculated exactly for the three networks. We observe that the presence of hubs tends to reduce the average shortest path compared to the Erdős-Rényi graph, as nodes are generally closer to the hubs. As mentioned earlier in the degree distribution section, hubs play a more significant role in the r/food graph, which is reflected in its lowest average shortest path length among the three graphs analyzed.

Graph	Diameter	Average Shortest Path
r/food	11	3.86
Erdős–Rényi	10	5.56
Barabási-Albert	7	4.46

Table 4: Path length metrics.

To further investigate the nature of the analyzed graphs, we plotted their path length distribution. Rather than computing every possible shortest path explicitly, we used an approximation by sampling a subset of nodes. Specifically, we sampled 10,000 nodes.

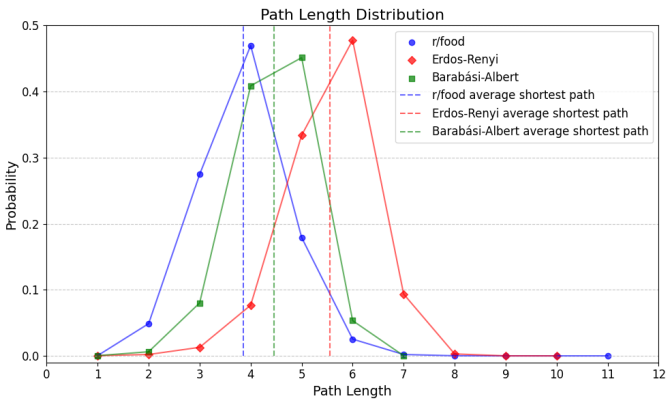

Figure 3: Comparison of the path length distribution of the r/food graph with those from the Barabási-Albert and Erdos-Renyi models.

Each distribution shows how probability varies with path length, revealing distinct structural and topological features of the graphs. The path length distributions of the Erdős-Rényi and Barabási-Albert graphs are symmetric around the average, indicating more uniform node connections. In contrast, the r/food graph has a few nodes significantly farther apart, creating a right-tail in the distribution. However, this tail is small, so the r/food path length distribution remains mostly symmetric but shifted toward shorter distances compared to the other two graphs.

3.5 Clustering Coefficient and Density Analysis

The average clustering coefficient and graph density were computed. We observed that the clustering coefficient of the Erdős-Rényi graph was the lowest, followed by that of the Barabási-Albert graph, which is still an order of magnitude lower than the value computed for the r/food graph as expected from the theory. The densities of the graphs were similar, as they depend solely on the number of nodes and edges. However, since this value is much lower than 1, we can conclude that our graphs are sparse.

Graph	Average Clustering	Density
r/food	$2.34 \cdot 10^{-2}$	$3.25 \cdot 10^{-4}$
Erdős–Rényi	$3.67 \cdot 10^{-4}$	$3.23 \cdot 10^{-4}$
Barabási-Albert	$3.65 \cdot 10^{-3}$	$3.07 \cdot 10^{-4}$

Table 5: Clustering coefficient and density metrics.

We plotted a graph comparing the clustering coefficient to the node degree for the three graphs, with the r/food graph shown here as an example.

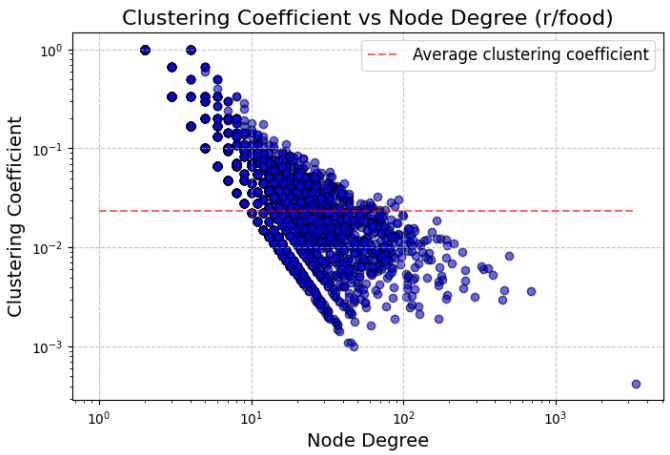

Figure 4: Clustering coefficient vs node degree in the r/food graph.

All the graphs exhibit a decreasing trend in the clustering coefficient as the node degree increases. However, it is evident that the rate at which the clustering coefficient decreases is slower in the r/food graph. The trend we observed indicates that the clustering at each node degree is higher in the r/food graph compared to synthetic networks. This could be attributed to users in the r/food graph forming tighter clusters with one another, possibly driven by shared interests or commonalities between interacting users.

3.6 Centrality Analysis

Different centrality metrics were computed, including degree centrality, betweenness centrality, closeness centrality, eigenvector centrality, PageRank centrality, and Katz centrality. These metrics share common characteristics influenced by the graph structure and the presence (or absence) of hubs:

- A similar average centrality score across the three graphs, with some exceptions for distance-based scores and eigenvector centrality.

- A consistently higher maximum centrality score in the r/food graph, followed by the Barabási-Albert and Erdős-Rényi graphs, reflecting the prominence of hubs.
- A variable minimum centrality score, strongly dependent on the presence of isolated nodes and the specific metric definition.

Graph	Avg deg. centr.	Max deg. centr.	Min deg. centr.
r/food	$3.25 \cdot 10^{-4}$	0.173	$5.12 \cdot 10^{-5}$
Erdős–Rényi	$3.23 \cdot 10^{-4}$	$8.70 \cdot 10^{-4}$	0
Barabási-Albert	$3.07 \cdot 10^{-4}$	$2.80 \cdot 10^{-2}$	$1.53 \cdot 10^{-4}$

Table 6: Degree centrality metrics.

The betweenness centrality and eigenvector centrality were not computed explicitly due to the high computational cost. Instead, they were approximated: betweenness centrality was estimated by randomly sampling the graph rather than calculating every possible path, and eigenvector centrality was approximated by setting a tolerance value to halt the calculation once further changes in the average value became negligible. The closeness centrality, on the other hand, was computed explicitly on the largest component of each graph. Every other centrality score was computed exactly.

Graph	Avg betweenness	Max betweenness	Min betweenness
r/food (2022)	$1.24 \cdot 10^{-4}$	0.309	0
Erdős–Rényi	$2.33 \cdot 10^{-4}$	$2.18 \cdot 10^{-3}$	0
Barabási-Albert	$1.77 \cdot 10^{-4}$	0.135	$4.38 \cdot 10^{-7}$

Table 7: Betweenness centrality metrics.

Graph	Avg closeness	Max closeness	Min closeness
r/food (2022)	0.263	0.441	0.139
Erdős–Rényi	0.180	0.207	0.133
Barabási-Albert	0.226	0.359	0.178

Table 8: Closeness centrality metrics.

Betweenness and closeness centrality are distance-based metrics, directly reflecting the patterns observed in the path analysis section, where the presence and significance of hubs influence connectivity across the analyzed graphs.

Graph	Avg eigenvector	Max eigenvector	Min eigenvector
r/food (2022)	$2.86 \cdot 10^{-3}$	0.673	$4.63 \cdot 10^{-10}$
Erdős–Rényi	$6.37 \cdot 10^{-3}$	$3.03 \cdot 10^{-2}$	$1.18 \cdot 10^{-4}$
Barabási-Albert	$2.37 \cdot 10^{-3}$	0.552	$1.37 \cdot 10^{-5}$

Table 9: Eigenvector centrality metrics.

Eigenvector centrality assigns a score to each node based on the idea that a node is more important if it connects to other important nodes. The significantly higher average eigenvector centrality may result from the degree correlation and assortativity of the graphs, where well-connected nodes in the r/food and Barabási-Albert graphs tend to link with low-degree nodes, unlike in the Erdős-Rényi graph. This result is seen in detail in the next section.

3.7 Assortativity and Degree Correlation

Finally, we computed the assortativity and plotted the degree correlation for the three graphs. As expected, the r/food and Barabási-Albert graphs show negative assortativity values, indicating a disassortative structure where high-degree nodes tend to connect with low-degree nodes. In contrast, the Erdős-Rényi graph has a positive assortativity, meaning it is assortative, with high-degree nodes more likely to link to other high-degree nodes.

Graph	Avg page rank	Max page rank	Min page rank
r/food (2022)	$5.57 \cdot 10^{-5}$	$3.21 \cdot 10^{-2}$	$1.39 \cdot 10^{-5}$
Erdős–Rényi	$5.13 \cdot 10^{-5}$	$1.16 \cdot 10^{-4}$	$1.39 \cdot 10^{-5}$
Barabási-Albert	$5.12 \cdot 10^{-5}$	$3.97 \cdot 10^{-3}$	$2.62 \cdot 10^{-5}$

Table 10: Page rank centrality metrics.

Graph	Avg Katz	Max Katz	Min Katz
r/food (2022)	$6.76 \cdot 10^{-3}$	0.329	$5.55 \cdot 10^{-3}$
Erdős–Rényi	$7.16 \cdot 10^{-3}$	$7.94 \cdot 10^{-3}$	$6.77 \cdot 10^{-3}$
Barabási-Albert	$7.11 \cdot 10^{-3}$	$5.07 \cdot 10^{-2}$	$6.81 \cdot 10^{-3}$

Table 11: Katz centrality metrics.

Graph	Assortativity
r/food (2022)	$-5.20 \cdot 10^{-2}$
Erdős–Rényi	$5.58 \cdot 10^{-3}$
Barabási-Albert	$-2.98 \cdot 10^{-2}$

Table 12: Assortativity.

The assortativity score is reflected in the slope of the degree correlation plot. A negative coefficient (as in the r/food and Barabási-Albert graphs) indicates that the average neighbor degree decreases as node degree increases. Conversely, a positive assortativity score (as in the Erdős-Rényi graph) signifies that the average neighbor degree increases with node degree.

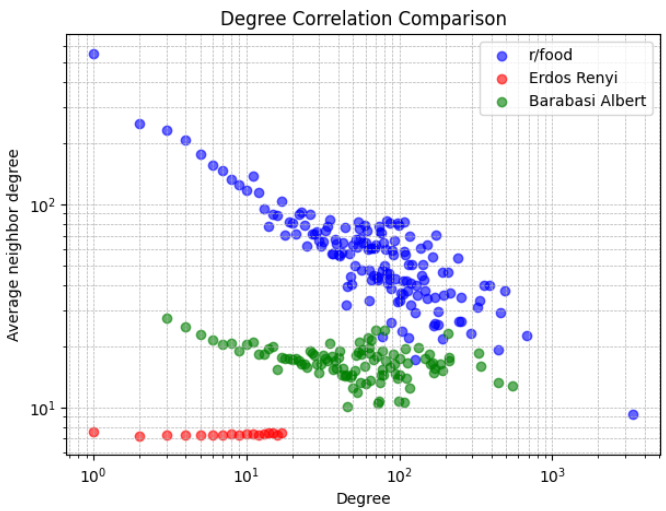

Figure 5: Degree correlation comparison.

4 Spreading Task

This paper simulates and analyzes epidemic spreading on three different network structures: a real-world network crawled from a subreddit about food, an Erdős-Rényi (ER) random network, and a Barabási-Albert (BA) scale-free network. The crawled network represents an online social interaction graph, where nodes correspond to users, and edges indicate comments under other user's

posts. The ER and BA models serve as synthetic benchmarks to compare random and scale-free structural properties. Using the NDlib Python library, we implement four epidemiological models: SI, SIS, SIR, and the Threshold model, to study how different topologies influence information or disease propagation. Key parameters such as infection rate β, recovery rate γ, and activation threshold are varied to observe their effects. The results are visualized through plots showing the proportion of infected nodes over time, revealing how network structure impacts epidemic dynamics. To better distinguish spreading behaviors across different networks, adjustments are made in the simulations, including selecting a smaller proportion of initially infected nodes and reducing transmission rates. These changes prevent overly rapid saturation, allowing clearer observation of how each network's topology influences diffusion patterns. Below, we provide a comprehensive analysis of each diffusion model and how epidemic curves change with parameter variations.

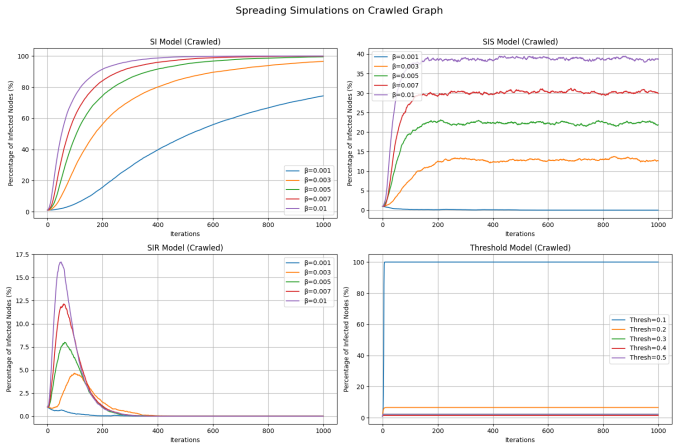

Figure 6: Infection spreading in the crawled graph with variable parameters.

4.1 Metodology

4.1.1 Network Construction

- **Crawled Graph** Represents a real-world network with unique topological properties derived from empirical data.
- **Erdős-Rényi (ER) Graph:** A synthetic network generated using the $nx.gnm\_random\_graph()$ function. It has a uniform connection probability between nodes, leading to a homogeneous structure.
- **Barabási-Albert (BA) Graph:** A scale-free network generated using $nx.barabasi\_albert\_graph()$, where a few highly connected hub nodes dominate the structure.

More precisely, Figures 6-7-8 correspond to the results obtained by varying the parameters of the models. Figures 9-10-11 instead correspond to the variation of the initial condition (seed). Finally, Figures 12-13-14 show results obtained using the random and high-degree strategy.

4.1.2 Simulation Parameters

- β (Infection Rate): [0.001, 0.003, 0.005, 0.007, 0.01] (lower values slow the spread).
- γ (Recovery Rate for SIS and SIR models): [0.01, 0.03, 0.05, 0.07, 0.1].
- Threshold Values for Threshold Model: [0.1, 0.2, 0.3, 0.4, 0.5].
- Initial Infected Nodes: Only 1% of the network is initially infected (randomly chosen).
- Number of Iterations: 1000 iterations for each simulation to track infection evolution.

4.1.3 Results & Analysis

Below we show the results obtained:

Crawled Social Network:

- SI Model (Top Left): The number of infected nodes increases over time, approaching saturation. Higher β (transmission rate) leads to faster infection spread. The curves show different β values, where β=0.01 causes the quickest infection spread, while β=0.001 is the slowest. Since SI lacks recovery, all nodes eventually become infected.
- SIS Model (Top Right): Initially, infections rise rapidly, but they stabilize instead of infecting the entire network. This stabilization occurs because recovered nodes can become susceptible again. Higher β values (e.g., 0.01) result in a higher equilibrium of infected nodes.
- SIR Model (Bottom Left): The infection spreads initially but eventually declines as nodes recover permanently. The peak infection occurs early before nodes transition to the recovered state. Higher β values cause a larger peak but still lead to eventual decline.
- Threshold Model (Bottom Right): When the infection threshold is low (e.g., 0.1), nearly all nodes get infected quickly. Higher thresholds (e.g., 0.5) drastically reduce the number of infected nodes. The infection spread is fast for lower thresholds, while for high thresholds, it remains contained.

Erdős-Rényi Graph:

- SI Model (Top Left): Similar to the Crawled Graph, all nodes eventually become infected. The spread is slightly faster compared to the Crawled Graph, suggesting a more uniform connectivity.
- SIS Model (Top Right): Compared to the Crawled Graph, SIS infections stabilize at a lower level. A higher β (e.g., 0.01) still sustains a significant infection level, but lower β values (e.g., 0.001) result in minimal infections. The structure of Erdős-Rényi makes it less efficient for sustaining long-term infections.
- SIR Model (Bottom Left): Infections rise and then decay faster than in the Crawled Graph. The peak number of infections is lower compared to the previous graph, showing that random networks are less prone to sustaining outbreaks.

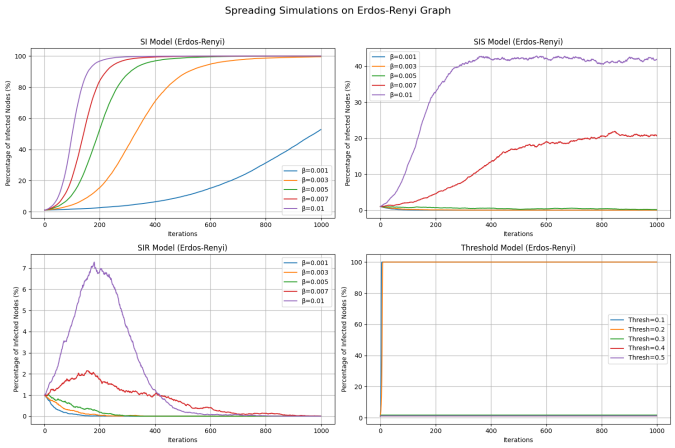

Figure 7: Infection spreading in the Erdos-Renyi graph with variable parameters.

- Threshold Model (Bottom Right): For low thresholds (e.g., 0.1), infections quickly dominate the network. Higher thresholds (e.g., 0.4) significantly prevent spread. Erdős-Rényi graphs exhibit a similar infection pattern to the Crawled graph but with faster stabilization.

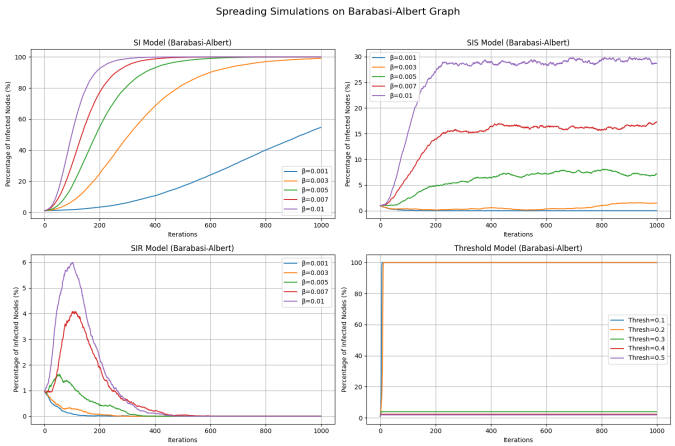

Figure 8: Infection spreading in the Barabasi-Albert graph with variable parameters.

Barabási-Albert Graph:

- SI Model (Top Left): Infection spreads much faster than in both the Crawled and Erdős-Rényi networks. Highly connected nodes (hubs) accelerate the infection process. The curves show that even with low β, the entire network becomes infected quickly.
- SIS Model (Top Right): Compared to the previous graphs, infections stabilize at a higher equilibrium. The presence of highly connected nodes means reinfection is more frequent, making it difficult for infections to die out. β=0.01 results in a persistent high number of infections.
- SIR Model (Bottom Left): The infection peaks more sharply than in the other graphs. Due to the scale-free structure, the epidemic dies out more quickly after the peak because hubs become recovered.
- Threshold Model (Bottom Right): Even at high thresholds (e.g., 0.3), the infection spreads widely due to hub nodes. The infection reaches saturation for low thresholds.

4.2 Varying the infection seeds

In this simulation, we aim to investigate how infection spreads across various network topologies using different epidemic models. To ensure computational efficiency and meaningful analysis, we selected a set of fixed and variable parameters.

Fixed Parameters:

Beta (β) and Gamma (γ) are parameters that govern the transmission and recovery dynamics within the epidemic models. In this simulation, both β and γ are kept fixed at values of $\beta = 0.005$ and $\gamma = 0.03$ across all models and network types. The rationale behind fixing these parameters is to reduce computational complexity and allow us to focus on how network structure and initial conditions affect the spread of infection. By maintaining fixed infection and recovery rates, we eliminate the need to explore variations in these parameters, which would significantly increase computational costs and complexity. This approach simplifies the analysis, enabling a more efficient comparison across different models and network configurations.

In the Threshold model, each node has a fixed activation threshold that determines when it becomes infected based on the activity of its neighbors. A node becomes infected if the fraction of its infected neighbors exceeds its threshold. In this simulation, the threshold is fixed for each node at 0.3, meaning that a node will become infected if at least 30% of its neighbors are infected. By fixing the threshold across all nodes, we ensure that the activation conditions remain consistent and focus the analysis on the role of network structure and initial conditions rather than varying individual node behavior. This allows us to analyze how network topology and initial infection size influence the spread of the infection in a system where node thresholds are uniform.

Varying Initial Conditions:

One of the key variables in this simulation is the seed size, which defines the initial proportion of nodes that are infected at the start of the simulation. By varying the seed size, we analyze how the initial number of infected nodes impacts the spread and diffusion of the infection. Seed sizes in this simulation range from 0.5% to 10% of the total number of nodes, providing a range of initial infection conditions. Larger seed sizes are expected to cause faster and more widespread infections, while smaller seed sizes may lead to more erratic or stochastic infection dynamics.

4.2.1 Seed plots

4.2.2 Seed random and high degree strategy plots

In this part of the simulation, we have fixed both the number of parameters and the initial seed. However, we explore two different strategies: random and high-degree. Below we report the results.

Crawled Social Network

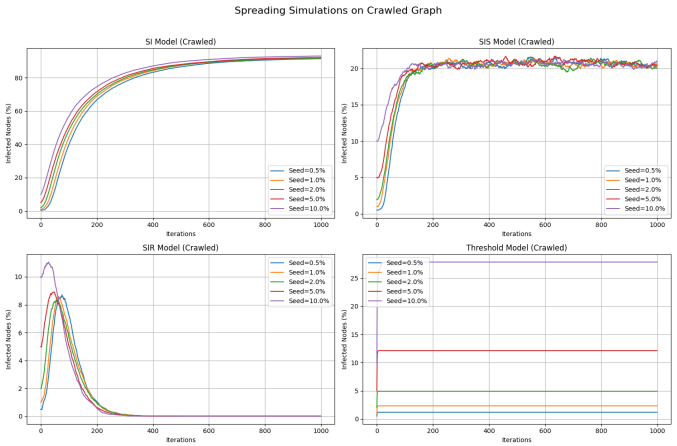

Figure 9: Seed variations in the crawled graph with fixed parameters.

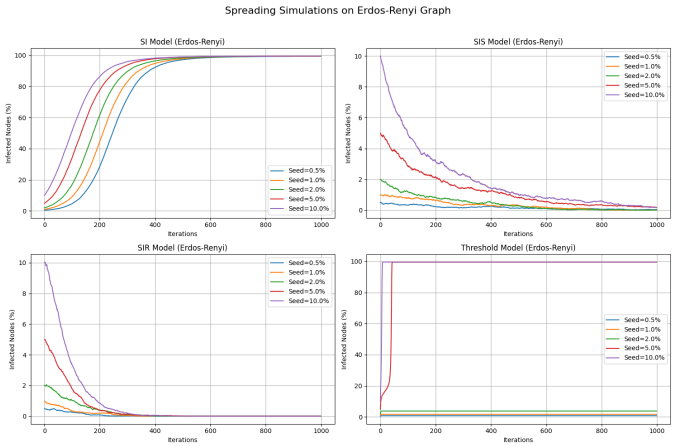

Figure 10: Seed variations in the Erdős-Rényi graph with fixed parameters.

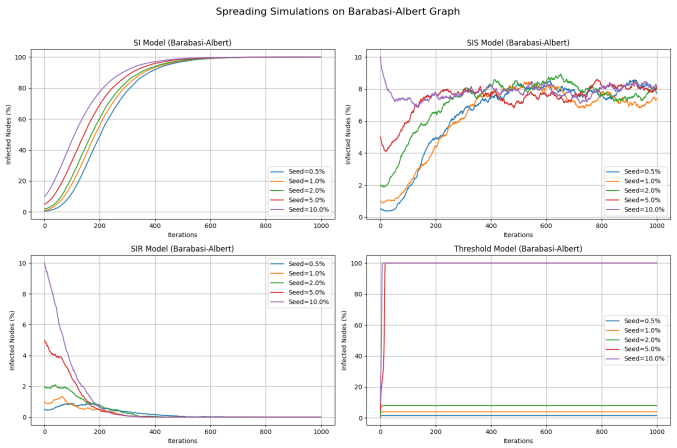

Figure 11: Seed variations in the Barabasi-Albert graph with fixed parameters.

- SI Model (Top Left): Infections grow exponentially until all nodes are infected. High-degree node selection leads to a slightly faster spread.

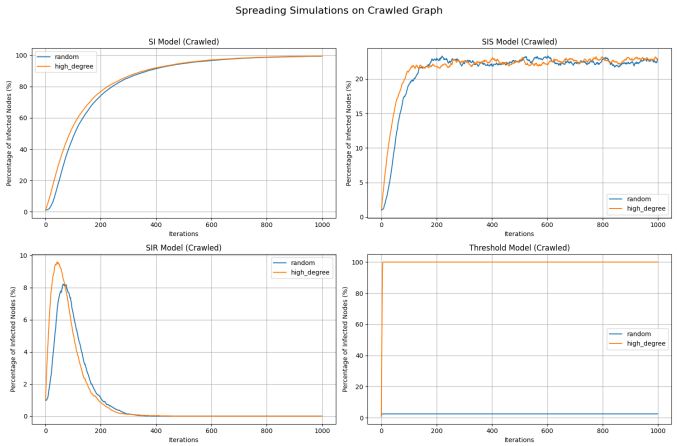

Figure 12: Seed strategy in the rawled graph with fixed parameters and seed.

- SIS Model (Top Right): The infection stabilizes at an equilibrium state due to continuous reinfection. The equilibrium level is higher with high-degree targeting.
- SIR Model (Bottom Left): Infection peaks early and then declines as nodes recover permanently. The high-degree strategy initially leads to a higher infection peak.
- Threshold Model (Bottom Right): When the threshold is low (e.g., 0.1), nearly all nodes become infected. At higher thresholds (e.g., 0.5), the infection remains controlled.

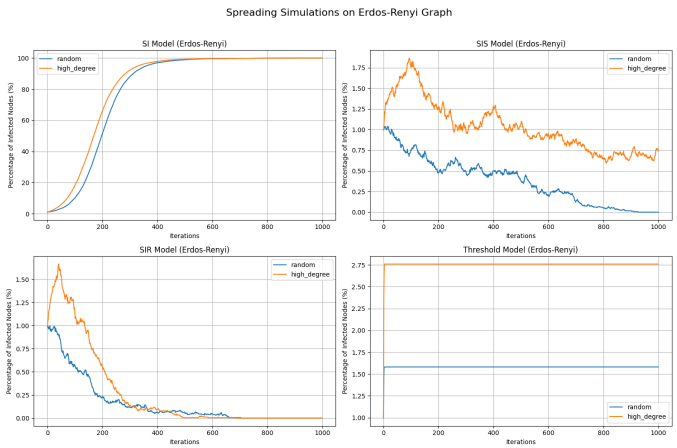

Figure 13: Seed strategy in the Erdős-Rényi graph with fixed parameters and seed.

Erdős-Rényi Random Network

- SI Model: Infection progresses similarly in both strategies due to homogeneous connectivity.
- SIS Model: The infection equilibrium is lower than in scale-free networks, as no hubs dominate the spreading process.
- SIR Model: The infection spreads more uniformly, peaking lower than in Barabási-Albert networks.

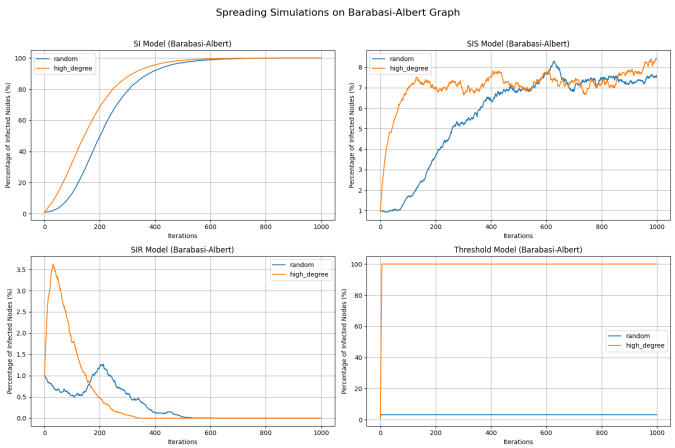

Figure 14: Seed strategy in the Barabasi-Albert graph with fixed parameters and seed.

- Threshold Model: Higher thresholds significantly limit the infection, as no single node plays a dominant spreading role.

Barabási-Albert Scale-Free Network

- SI Model: Infection spreads rapidly due to the presence of hubs, with high-degree targeting accelerating the spread.
- SIS Model: Infection levels stabilize higher than in the Crawled network since reinfection occurs more efficiently in scale-free networks.
- SIR Model: The peak infection is sharp, as hub nodes accelerate the early-stage spread before recovery takes over.
- Threshold Model: High-degree node targeting results in full infection spread, whereas random selection contains the infection at high thresholds.

Summary: Our analysis demonstrates that network topology significantly affects epidemic dynamics. Scale-free networks (Barabási-Albert) are highly vulnerable to epidemics due to hub nodes, whereas random networks (Erdős-Rényi) exhibit more uniform and slower spreading patterns. The choice of infection strategy (random vs. high-degree) further influences the extent and speed of epidemic spread.

4.3 Spreading Task Conclusion

Our analysis of epidemic spreading across different network topologies: Crawled, Erdős-Rényi (ER), and Barabási-Albert (BA). Demonstrates how network structure fundamentally shapes diffusion dynamics.

The BA network, due to its scale-free nature and presence of highly connected hubs, exhibited the fastest spread, as infections quickly reach a large portion of the network. The ER network, with its more uniform and random connectivity, displayed the slowest diffusion, since the lack of dominant hubs limited rapid transmission pathways. The Crawled network, representing a real-world social structure, exhibited an intermediate spreading speed, influenced by community structures and local clustering effects that create bottlenecks in transmission.

A key takeaway is the role of hubs in accelerating diffusion in BA and Crawled networks. However, this reliance on hubs also introduces vulnerability, removing these influential nodes can significantly disrupt the spreading process. In contrast, the ER network is more resilient to node removal, as its uniform structure prevents the over-reliance on a few key nodes. Additionally, real-world networks emphasize the importance of community structure, where diffusion may remain localized before spreading globally.

Ultimately, these findings reinforce the critical role of network topology in understanding how information, epidemics, or influence propagate in complex systems. While scale-free networks enable rapid diffusion, they also introduce fragility. On the other hand, random networks, despite slower spread, offer greater robustness.

5 Open Question: What Do the Culinary Preferencies Say About the User's Behaviour?

5.1 Posts Classification and Node Labeling

To analyze how culinary preferences shape human behavior online, we first needed to categorize users based on the content they posted. This was accomplished by classifying post descriptions through an analysis of the tags required for posting within the subreddit (each post must include one of the tags [i ate]/[homemade] in its description) or by directly considering the content of the titles with the help of a large language model (LLM).

Specifically, we utilized the Llama 3.2-1B-Instruct model (`https://huggingface.co/meta-llama/Llama-3.2-1B-Instruct`), integrating it into our study through a custom Python script. This tool allowed us to classify recipes into categories such as vegetarian/omnivorous, sweet/savory, and the most probable ethnicity, enabling further analysis. Since our primary goal was not to develop the most accurate machine learning model but rather to explore culinary trends, we relied on the model's baseline accuracy and proceeded without additional refinements.

Once all the posts were classified, we proceeded with labeling the nodes using a majority rule. Specifically, if the majority of posts created by a user were classified in the same way, we assigned that classification to the user hence labeling the corresponding node in this manner. In cases of a tie, the node was labeled as "unknown" (the same applied to posts that were misclassified). We then selected the largest component of our graph and performed label propagation to classify the unknown nodes (fewer than 10% of the nodes were labeled as "unknown") and finally we conducted various analyses on the fully labeled network.

5.2 Polarization Analysis on the Undirected Graph

We began calculating the homofily coefficient and the Weighted homophily coefficient in each of our classifications, yelding the following results:

Classification	Homophily	Weihgted Homophily
Vegetarian/Omnivorous	0.61	0.63
Sweet/Savory	0.87	0.87
[i ate]/[homemade]	0.74	0.74
Ethnicity	0.43	0.47

Table 13: Homophily and weighted homophily calculated on the largest component of the unweighted undirected labeled network.

Certain categories exhibit strong homophily, indicating that interactions within these categories are highly polarized. Specifically, nodes with the same label are more likely to connect with each other than with nodes of different labels. Interestingly, we observe that including the weight of the edges does not significantly alter the homophily coefficient, suggesting that the overall pattern of interaction remains consistent regardless of edge weights.

We then proceeded by calculating the Newman's assortativity and modularity of our graph, yielding the following results:

Classification	Categorical Assortativity	Modularity
Vegetarian/Omnivorous	0.123	0.092
Sweet/Savory	0.088	0.020
[i ate]/[homemade]	0.056	0.024
Ethnicity	0.064	0.092

Table 14: Assortativity and modularity calculated on the largest component of the unweighted undirected labeled network.

Newman's assortativity coefficient shows that each node classification results in an assortative network, indicating that interactions between similar users are more common. However, these coefficients are close to zero, suggesting a structure similar to a random network, with a maximum value of 0.123 for the vegetarian/omnivorous classification. Similarly, the modularity values are also near zero, with a maximum value of 0.092, implying a weak community structure based on our classifications. However, the slightly higher assortativity and modularity score in the vegetarian/omnivorous classification suggests that this distinction may be more divisive than the others. For instance, we might think that vegetarians are unlikely to engage with posts about meat, whereas omnivores may still explore vegetarian recipes. This idea will be further examined using a directed graph.

An intriguing aspect is the combination of high homophily and low assortativity. High homophily indicates that most connections occur between nodes with the same label, while low assortativity suggests that there is little correlation between the labels of connected nodes. This could imply that the network does not exhibit a strong preference for connecting similar nodes. Together, these two factors may indicate that the network exhibits homogeneous groups that are weakly connected to each other, suggesting a fragmented structure at the global level.

5.3 Correlation Between Degree and Number of Posts

We aimed to investigate whether the primary cause of the high connectivity of hubs can be explained by their number of posts. To do this, we first analyzed the correlation between the number of posts and the node degree for each author, finding a slight positive correlation with a Pearson coefficient of 0.317. While this suggests a general trend, it does not hold universally. For instance, the author with the highest degree has only 33 posts, which is an order of magnitude fewer than the most prolific posters. This indicates that the author's high connectivity is likely driven more by their level of interaction with other posts rather than the sheer number of posts they have made. This insight highlights the need to distinguish interactions based on their direction, as we will explore in the next section.

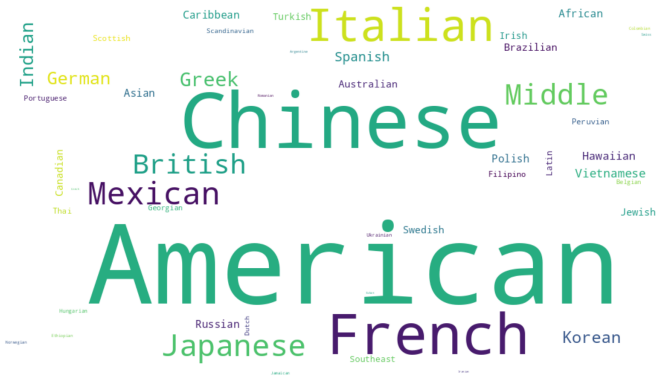

Figure 15: A word cloud representing ethnic classifications, where the word size corresponds to the number of nodes with that classification.

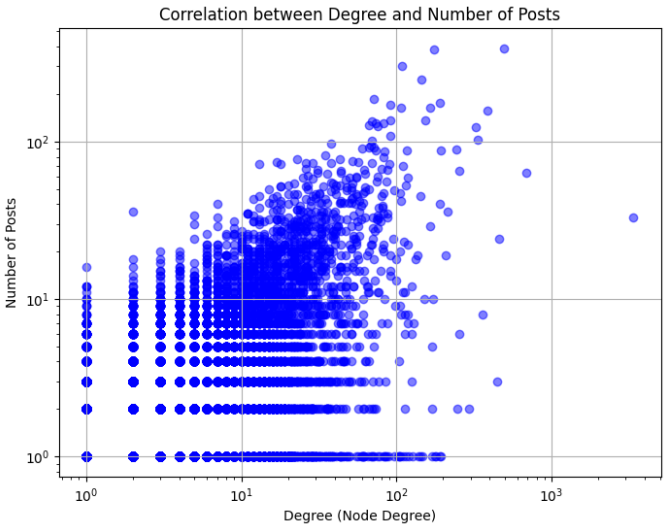

Figure 16: Scatter plot of node degree versus the number of posts.

5.4 Analysis on the Directed Graph

To further investigate the factors contributing to the connectivity, we differentiated links based on a high number of posts versus a high number of interactions. To achieve this, we constructed a directed graph where an edge from author A to author B (A $\rightarrow$ B) signifies that A commented on B's post.

We analyzed the number of connections within the graph, categorizing nodes according to all previously defined classifications. Our findings reveal that Veg$\rightarrow$Veg edges outnumber Veg$\rightarrow$Omn edges, suggesting a stronger tendency for interaction within the vegetarian community. Conversely, Omn$\rightarrow$Omn and Omn$\rightarrow$Veg edges occur at nearly equal frequencies, indicating no clear preference for internal interaction within the omnivorous group.

Type of interaction	Veg$\rightarrow$Omn	Omn$\rightarrow$Veg	Veg$\rightarrow$Veg	Omn$\rightarrow$Omn
Links	11998	11904	25744	11002
Links (scaled)	1.09	1.70	2.35	1.57

Table 15: Number of directed links counted by the type of interaction for the vegetarian omnivorous classification.

In the sweet/savory classification, we observe a highly asymmetric pattern: Sav→Sav interactions dominate, while links directed toward sweet nodes are significantly fewer. This indicates that the majority of interactions are concentrated on nodes labeled as savory.

Type of interaction	Swe→Sav	Sav→Swe	Swe→Swe	Sav→Sav
Links	2098	4389	766	53395
Links (scaled)	1.26	0.27	0.46	3.27

Table 16: Number of directed links counted by the type of interaction for the sweet/savory classification.

Similarly, we observe that the number of edges directed toward [homemade] nodes is significantly higher than those leading to [i ate] nodes.

Type of interaction	Ate→Hom	Hom→Ate	Ate→Ate	Hom→Hom
Links	8426	7531	2572	42119
Links (scaled)	2.44	0.52	0.74	2.90

Table 17: Number of directed links counted by the type of interaction for the sweet/savory classification.

Since the ethnic classification presented more than 50 different labels we plotted the data in the following graph.

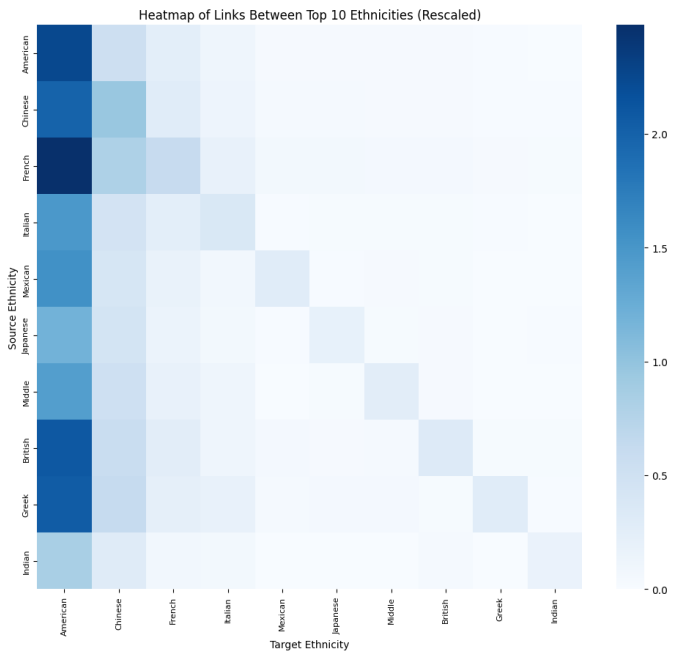

Figure 17: Heatmap depicting the number of links between two different ethnic groups, representing the top 10 ethnic groups sorted by their number of nodes, and normalized by the number of nodes in the source subgraph.

This figure illustrates two key observations:

- The level of interaction increases with the number of target nodes, as indicated by the color fading from left to right.
- The darker color along the diagonal emphasizes that connections within the same ethnic group are more prevalent.

5.5 Community Detection

To explore the structure of our network, we applied Louvain community detection, a modularity-based clustering algorithm that identifies groups of nodes with higher internal connectivity relative to the rest of the network. Given the high degree of homophily in our network, where nodes tend to connect with others of the same type and the low assortativity score, it was expected that our algorithm would reveal communities dominated by a single classification of users. The results confirmed this hypothesis.

The 35 detected communities are primarily characterized by an overwhelming majority of nodes of a single type, reinforcing the idea that user interactions are highly homogeneous within each community. However, a key insight from our analysis is that this homogeneity is not necessarily a consequence of a strong internal preference for the same classification but is instead driven by the overall numerical distribution of users in the network. In other words, the communities reflect the relative abundance of different user classifications in the dataset.

This explains why no community is primarily composed of a classification that is globally underrepresented. The sheer numerical dominance of certain classifications naturally leads to their prevalence within the detected communities. Additionally, the scattered nature of our network contributes to the overlapping visual representation of communities in the plotted network. Since the communities are numerous and not always spatially well-separated, they appear to overlap in the visualization, even though their internal composition remains distinct.

In summary, our community detection analysis reveals a network with clear structural segregation, where each detected community is dominated by a single classification of users. The composition of these communities is primarily dictated by the distribution of users across classifications, rather than an inherent tendency of the algorithm to separate specific groups. This insight aligns with our expectations and further validates the observed homophily in the network. Below we show the results obtained

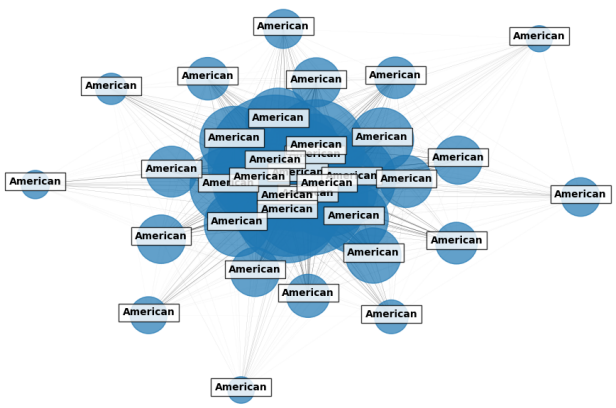

Figure 18: Graph showing each detected community is represented as a single node.The size of each node corresponds to the number of users in that community. Edges between nodes represent inter-community connections, with their thickness proportional to the number of links between the communities. The majority label (most common classification) in each community is displayed as text next to its node.

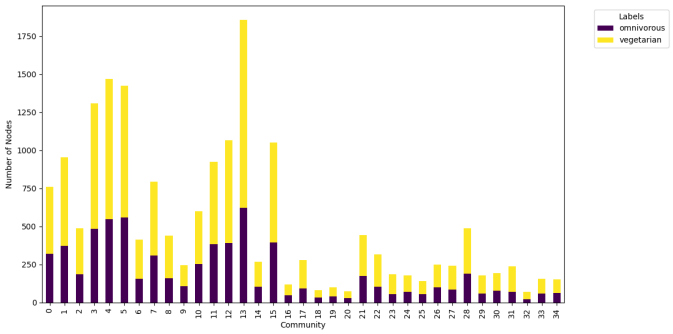

Figure 19: Stacked bar chart showing the distribution of Vegetarian and Omnivorous labels.

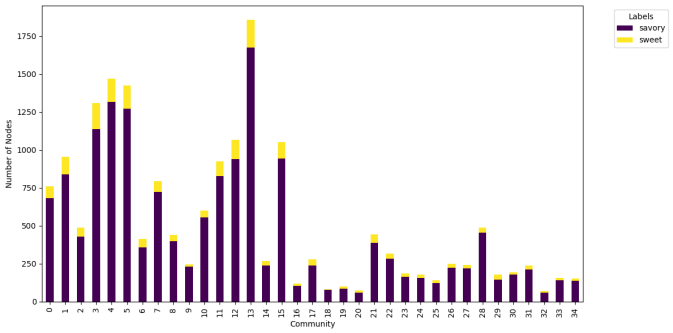

Figure 20: Stacked bar chart showing the distribution of Savory and Sweet labels.

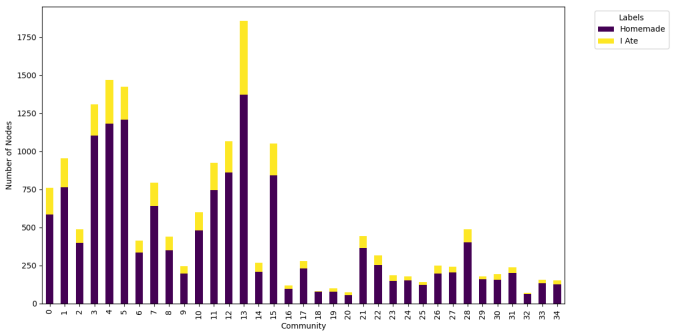

Figure 21: Stacked bar chart showing the distribution of Homemade and I Ate labels.

6 Conclusion

Our research began with the idea that polarization phenomena, often observed in online political debates, could also be found in another widely discussed area: food.

High homophily values reveal that most connections occur between nodes with the same label. However, the low assortativity indicates a lack of strong global preference for connecting similar nodes. This combination suggests the presence of homogeneous groups that are weakly interconnected, pointing to a fragmented network structure at the global level. In summary, while distinct clusters of similar nodes exist, they remain sparsely connected to one another.

Upon further reflection, we realized that directed interactions might offer more meaningful insights, prompting us to shift our approach and analyze the behavior of specific groups. Our data reveal that the probability of interacting with a certain type of post is partly influenced by its sheer popularity, as evidenced by the high number of links targeting the most popular tastes in the graph. However, as shown in the heatmap of links between the top 10 ethnic groups, there is generally a higher likelihood of interaction with nodes that share similar tastes.

Finally, we applied a community detection algorithm to better understand the network's structure. The results revealed multiple homogeneous communities, with the most prevalent classification consistently dominating in terms of node count. This outcome aligns well with our expectations and prior observations.

We can conclude that, although there isn't a strong community-based structure of users with similar interests (based on the classifications considered), the polarization of interactions remains evident. However, the intensity of this polarization varies depending on the most prevalent types of posts observed.