Interest Heterogeneity in Network Formation on Large Online Platforms

Shalem Sumanthiran

Advisor: Prof. Sanjeev Goyal, Professor of Economics, Social Science, NYU Abu Dhabi[1]

May 5, 2025

Thesis submitted to the Division of Social Sciences at New York University Abu Dhabi for the Economics Capstone Project in 2024.

[1]I would like to thank Prof. Sanjeev Goyal for all of his time and efforts in advising for this project over the course of the last year.

I submit this Thesis for the Capstone project in Economics to the Division of Social Sciences at New York University Abu Dhabi. The Thesis provides sufficient information on data sources, data periods, and data manipulation and transformation to allow for a full replication of all results.

Academic Integrity Declaration

I hereby declare that this Thesis is solely my work and contains no material I have previously submitted for assessment at New York University or elsewhere, without the express approval of all instructors including my Capstone advisor(s). To the best of my knowledge and belief, the Thesis contains no material previously published or written by another person except where due reference is made in the Thesis. I declare that I have read, and in undertaking this research, have complied with the Policy for Academic Integrity of Students at New York University. I also declare that I understand what is meant by plagiarism and that it is unacceptable. Except where I have expressly indicated otherwise, this Thesis is my own work and does not contain any plagiarized material in the form of unacknowledged quotations or any other material. Any data sources and computer code used to derive the results presented in this Thesis – including digital and electronic media sources – have been properly acknowledged.

Name: Shalem Sumanthiran

Signature: 

Date: May 05, 2025

AI Usage Self-Report for Capstone in Economics

Student Name: Shalem Sumanthiran

Student Group (list names of any group members): None

Thesis Title: Interest Heterogeneity in Network Formation on Large Online Platforms

Date: May 5, 2025

Please check the appropriate boxes to indicate your use of any AI-software (e.g., ChatGPT) in the preparation of your capstone thesis. Be honest and transparent about your usage.

General Usage:

☐ I/my group used AI for brainstorming and generating ideas related to my thesis topic.

☐ I/my group employed AI for clarifying and explaining complex economic concepts.

✓ I/my group used AI to draft and edit portions of my thesis.

☐ I/my group used AI to identify relevant academic sources and references for my thesis.

Data Analysis and Interpretation:

✓ I/my group used AI for help in analyzing data and statistical results.

☐ I/my group consulted AI to interpret the economic aspects of the data.

☐ I/my group used AI to generate hypotheses or potential explanations for observed economic behaviors/outcomes.

Literature Review and Citations:

☐ I/my group used AI to find relevant research papers and articles related to my thesis topic.

✓ I/my group employed AI to help write summaries and critiques of academic literature.

✓ I/my group used AI to generate citations and references in the appropriate format (e.g., APA, MLA).

Ethical Considerations:

✓ I/my group ensured that I/we used AI responsibly and that my/our usage of AI adhered to the AI guidelines and did not involve plagiarism.

✓ My group critically assessed and independently verified any information obtained from AI for accuracy.

By checking the boxes above and providing explanations, if necessary, you acknowledge your use of AI software in this preparation. This self-reporting is intended to maintain transparency and uphold academic integrity.

Signature: 

Contents

List of Figures

List of Tables

Interest Heterogeneity in Network Formation on Large Online Platforms

Shalem Sumanthiran

May 5, 2025

Abstract

This study examines how content creators' interest patterns shape network formation on social media platforms by extending the "Law of the Few" framework. Through theoretical modeling and analysis of YouTube data, we reveal that successful influencers achieve central network positions not through extreme specialization but through balanced interest distributions. Our game-theoretic model demonstrates that nodes with moderate interest levels, more evenly distributed interests, and less extreme specialization are significantly more likely to become influential hubs. Rather than single dominant figures emerging as classic network theory predicts, we find that multi-hub networks naturally form when interest similarity reduces connection costs. The empirical analysis of YouTube channels confirms these findings—creators with balanced interest profiles (lower concentration indices and moderate specialization) are more successful at becoming central nodes. These results provide practical insights for platform design and content strategy, challenging conventional wisdom that extreme niche specialization is optimal for digital influence.

Keywords: Network Formation, Interest Heterogeneity, Social Media, Influencers, Online Platforms

1 Introduction

The rise of large-scale social media platforms has dramatically altered the landscape of information dissemination and influence, creating environments where a small subset of individuals-often termed "influencers"-shape opinions, trends, and behaviors for vast audiences. Traditional models of network formation, such as the "Law of the Few" articulated by Galeotti and Goyal (2010), predict that as networks grow, a shrinking proportion of actors acquire and share information, resulting in highly centralized, star-like structures. However, empirical observations of digital content platforms, particularly YouTube, reveal more complex realities: rather than a single dominant hub, networks often display multiple influential nodes, each serving distinct interest communities.

This thesis investigates the role of heterogeneity in content creators' interests as a driver of network architecture on large online platforms. Specifically, it addresses the following research question: How does variation in the distribution and alignment of interests among network participants influence the emergence and structure of influential nodes? This question is of both theoretical and practical significance. For economists and network theorists, it challenges the adequacy of models that assume agent homogeneity or undirected, reciprocal connections. For platform designers and content creators, understanding these dynamics can help form strategies for more equitable and diverse digital platforms.

To explore this question, the thesis adopts a dual approach. First, it extends the classic network formation model by incorporating multidimensional interest vectors and allowing for directed, non-reciprocal links. This theoretical framework is analyzed through both analytical methods and computational simulations, revealing that interest heterogeneity can give rise to stable multi-hub equilibria. Second, the thesis empirically tests these predictions using a comprehensive dataset of YouTube

channels, employing natural language processing to construct interest profiles and advanced network analysis to identify influential nodes.

The findings demonstrate that both in simulated and real-world networks, nodes with balanced and moderately distributed interests-rather than those with extreme specialization-are significantly more likely to achieve central, influential positions. Statistical analysis further reveals that lower Gini coefficients, moderate average interest levels, and reduced kurtosis in interest distributions are robust predictors of core node status. These results suggest that strategic diversity, rather than narrow focus, underpins success in digital influence networks.

The remainder of the thesis is structured as follows: Section 2 reviews the relevant literature on network formation and interest heterogeneity. Section 3 details the data sources, construction, and descriptive statistics. Section 4 presents the theoretical model and simulation approach. Section 5 reports the empirical results and their interpretation. Section 6 concludes with a discussion of the broader implications and limitations of the findings.

2 Literature Review

This research investigates how heterogeneity in content creators' interests shapes the formation of influencer networks on large online platforms, with a particular focus on YouTube. The project is motivated by empirical observations that, contrary to classic predictions of highly centralized, star-like structures, real-world digital content networks often display multiple influential nodes serving distinct interest communities. The literature on network formation provides a foundation for understanding this phenomenon, but also reveals important gaps that this project aims to address.

The classic "Law of the Few" articulated by Galeotti and Goyal (2010) predicts that in large networks, a vanishingly small proportion of actors become the primary acquirers and disseminators of information, while the majority free-ride on their efforts. In their model, agents can either acquire information directly at a cost or form costly links to access information from others. Strict Nash equilibria in this setting exhibit a core-periphery structure: a small core of agents acquires information, and the periphery links to the core. Crucially, as the network grows, the proportion of information acquirers shrinks, but the absolute number remains bounded. This result is robust to variations in cost structures and has been observed empirically in both offline and online environments. However, the model assumes agent homogeneity and undirected, reciprocal links, limiting its applicability to the more complex, directed structures seen in digital platforms.

To address the limitations of homogeneity, Galeotti, Goyal and Kamphorst (2006) extend the analysis to settings with heterogeneous players, allowing for differences in both the benefits and costs of forming links. Their results show that while centrality and core-periphery structures remain robust features of equilibrium networks, heterogeneity introduces new architectural possibilities. For example, under cost

heterogeneity, active players form multipartite graphs, and under valuation heterogeneity, nested split graphs emerge. This framework is particularly relevant for digital content networks, where creators and consumers differ in their abilities and incentives to provide or seek information.

The literature has also explored the effects of strategic complementarities in network formation. Herskovic and Ramos (2020) model agents who care not only about their own information accuracy but also about coordinating with others. Their model predicts hierarchical information structures and shows that, under weakly convex link costs, core-periphery networks emerge in equilibrium. Importantly, even when agents are ex ante identical, equilibrium sorting leads to ex post heterogeneity in payoffs and actions, supporting the empirical observation that a few individuals become central information providers despite initial symmetry. Notably, their framework explicitly models directed, non-reciprocal links, clarifying that the benefit of connecting is one-way-a feature essential for modeling online platforms like YouTube.

A major advance in understanding the role of subjective perspectives and history dependence in network formation comes from Sethi and Yildiz (2016). Their model incorporates agents with heterogeneous, unobservable perspectives and demonstrates that a broad range of observational patterns can arise, including persistent opinion leadership, information segregation, and even arbitrary static networks. Critically, in large populations, each individual's set of long-run experts is small, and the law of the few emerges through a different mechanism than in Galeotti and Goyal (2010): the set of long-run experts becomes a negligible fraction of the population as the network grows, even if better information is available elsewhere. This work predates and in some respects subsumes more recent analyses of unobservable heterogeneity in network formation.

Recent research by Chen, Li and Song (2024) directly addresses the structure

of influencer networks in user-generated content markets, such as YouTube. Their model introduces two key features absent from earlier work: the complementarity between influence and content provision, and the possibility of non-reciprocal (directed) links. They show that strict equilibrium networks are nested and upward-linking, supporting the coexistence of multiple influencer tiers. Unlike the bounded number of influencers in Galeotti and Goyal (2010), the number of influencers in Chen, Li and Song's (2024) model can increase indefinitely with population size, even if their proportion remains small. This aligns with empirical trends on platforms like YouTube, where the absolute number of influencers has grown alongside the user base. Moreover, their results explain the emergence of multi-tiered influencer hierarchies and the low reciprocity observed in real-world social media networks. While Chen, Li and Song (2024) build on the theoretical foundation of Galeotti and Goyal (2010), they generalize core-periphery and hierarchical network predictions to settings with directed links and complementary incentives, filling a critical gap in the literature.

Empirical studies, such as Rieder et al. (2023), provide further evidence that the structure of digital content networks is more complex than a simple star or core-periphery architecture. Their analysis of YouTube reveals the coexistence of multiple levels of influencers and documents the economic instability and platform uncertainty that characterize the creator economy. Only a small fraction of creators succeed in making a sustainable living, and the network exhibits multiple hubs serving distinct interest communities-a pattern consistent with the theoretical predictions of Chen, Li and Song (2024) and the interest-heterogeneity extensions developed in this research.

Other strands of the literature have explored the dynamics of network formation and the role of agent heterogeneity in experimental settings. For example, Neligh (2020) demonstrates that agents may "vie for dominance" by making more con-

nections than is myopically optimal, seeking to become central nodes and receive future connections. However, risk aversion and other behavioral factors can lead to deviations from theoretical predictions. Sadler (2022) introduces swap-proof stability in network formation with increasing marginal costs, showing that convex linking costs and local linking benefits result in stable networks characterized by strong hierarchies or overlapping cliques.

Despite the advances in the literature, several gaps remain. Most notably, while directed, non-reciprocal links are now modeled in some recent work, the interplay between interest heterogeneity, directedness, and the complementarity between influence and content provision remains underexplored. Classic models do not account for the empirical reality that multiple influencer tiers can coexist and grow with network size, nor do they fully explain the emergence of multi-hub structures observed on platforms like YouTube. Furthermore, while models such as Sethi and Yildiz (2016) and Chen, Li and Song (2024) allow for a broad range of network outcomes, the specific mechanisms by which interest similarity and content diversity shape network architecture have not been systematically analyzed in the context of large-scale digital content platforms.

This research project aims to fill these gaps by developing a theoretical and empirical framework that incorporates multidimensional interest heterogeneity, directed links, and the complementarity between influence and content provision. By doing so, it explains the coexistence of multiple influencer tiers, the growth of influencer numbers with network size, and the formation of multi-hub networks. The empirical focus on YouTube provides a direct test of these predictions, using network and content data to show that balanced, moderately distributed interests-rather than extreme specialization-are associated with central, influential positions. In this way, the project not only extends the theoretical literature but also aligns closely with empirical realities in digital content markets, offering new insights for both

platform designers and content creators.

3 Methodology

3.1 Theoretical Model

This section presents a game-theoretic model of network formation with endogenous information acquisition, based on the framework developed by Galeotti and Goyal (2010). The model captures how individuals strategically choose both how much information to acquire personally and which social connections to form to access information from others.

3.2 Players and Strategies

Consider a set of players $N = \{1, 2, \ldots, n\}$ with $n \geq 3$. Each player i makes two decisions:

- Choose a level of personal information acquisition $x_i \in X = [0, +\infty)$
- Choose which links to form with other players, represented as a vector $g_i = (g_{i1}, \ldots, g_{i,i-1}, g_{i,i+1}, \ldots, g_{in})$, where $g_{ij} \in \{0, 1\}$ for each $j \in N \setminus \{i\}$

A link $g_{ij} = 1$ indicates that player i forms a link with player j, allowing i to access information acquired by j. This link formation is unilateral; player i can decide to form a link with j without j's consent, but i must bear the cost of this link. The set of strategies available to player i is denoted by $S_i = X \times G_i$, where $G_i = \{0, 1\}^{n-1}$ represents all possible linking configurations for player i.

The overall strategy profile is denoted as $s = (x, g) \in S$, where $S = S_1 \times \cdots \times S_n$. This specifies both the information acquisition levels $x = (x_1, x_2, \ldots, x_n)$ and the network structure $g = (g_1, g_2, \ldots, g_n)$.

3.3 Network Structure

The network of relations is represented by a directed graph g. For each player i, we define the set of players with whom i has formed links as $N_i(g) = \{j \in N : g_{ij} = 1\}$, with $\eta_i(g) = |N_i(g)|$ representing the number of links formed by player i.

The closure of g is an undirected network denoted by $\bar{g} = cl(g)$, where $\bar{g}_{ij} = \max\{g_{ij}, g_{ji}\}$ for each i and j in N. In words, the closure of a directed network involves replacing every directed edge of g by an undirected one. Define $N_i(\bar{g}) = \{j \in N : \bar{g}_{ij} = 1\}$ as the set of players directly connected to i. The undirected link between two players reflects bilateral information exchange between them.

We say that there is a path in $\bar{g}$ between i and j if either $\bar{g}_{ij} = 1$ or there exist players $j_1, \ldots, j_m$ distinct from each other and i and j such that $\{\bar{g}_{ij_1} = \bar{g}_{j_1 j_2} = \ldots = \bar{g}_{j_m j} = 1\}$. A network $\bar{g}$ is connected if there exists a path between every pair of players; we say that a network $\bar{g}$ is minimally connected if it is connected and there exists only one path between every pair of players.

In a core-periphery network, there are two groups of players, $\hat{N}_1(g)$ and $\hat{N}_2(g)$, with the feature that $N_i(\bar{g}) = \hat{N}_2(\bar{g})$ for all $i \in \hat{N}_1(\bar{g})$, and $N_j(\bar{g}) = N \setminus \{j\}$ for all $j \in \hat{N}_2(g)$. Nodes which have $n-1$ links are referred to as central nodes or as hubs, while the complementary set of nodes are referred to as peripheral nodes or as spokes. A core-periphery network with a single hub is referred to as a periphery-sponsored star.

3.4 Payoffs

The payoff function captures the economic trade-offs faced by players. For player i under strategy profile $s = (x, g)$, the payoff is:

$$\Pi_i(s) = f\left(x_i + \sum_{j \in N_i(\bar{g})} x_j\right) - cx_i - \eta_i(g)k \tag{1}$$

where:

- $f(\cdot)$ is the benefit function from information, assumed to be twice continuously differentiable, increasing, and strictly concave
- $c > 0$ is the cost per unit of personally acquired information
- $k > 0$ is the cost of forming a link with another player

We assume pure local externalities: player i accesses only the information personally acquired by their immediate neighbors. The benefit function $f(\cdot)$ is characterized by several key properties to ensure meaningful analysis:

- $f(0) = 0$ (no benefit from zero information)
- $f'(0) > c$ (initially, acquiring information is worthwhile)
- $\lim_{y\to\infty} f'(y) = m < c$ (eventually, the marginal benefit of additional information falls below its cost)

Under these assumptions, there exists an optimal level of information $\hat{y} > 0$ that maximizes the net benefit of information acquisition for an isolated player: $\hat{y} = \arg\max_{y \in X} f(y) - cy$, which solves the first-order condition $f'(\hat{y}) = c$.

3.5 Equilibrium

A Nash equilibrium is a strategy profile $s^* = (x^*, g^*)$ such that:

$$\Pi_i(s_i^*, s_{-i}^*) \geq \Pi_i(s_i, s_{-i}^*), \forall s_i \in S_i, \forall i \in N \tag{2}$$

An equilibrium is said to be strict if the inequalities in the above definition are strict for every player. Strict equilibria are particularly important for our analysis as they represent robust predictions of the model. Much of the vocabulary and initial model setup was taken from Networks: An Economics Approach by Goyal (2023), which can be referred to for more detail (for example, when we discuss centrality metrics in a later section).

4 Numerical Simulation

4.1 Computational Implementation

Our theoretical model predicts that in equilibrium, a small subset of individuals will acquire information while the majority will form connections with these information providers. To validate this prediction, we implemented a computational simulation of the Law of the Few model using Python.

The simulation initializes a network of 200 players, each making strategic decisions about how much information to personally acquire and which links to form with other players to access their information.

The program implements the exact theoretical framework described earlier, with players' payoffs determined by the benefits from information (both personally acquired and accessed through neighbors), costs of personal information acquisition and costs of forming links

In each simulation step, players take turns updating their strategies (information acquisition and link formation) in response to the current network state, continuing until the network structure stabilizes.

4.2 Equilibrium Emergence

Figure 1 displays the initial and final states of the network for a simulation with parameters $c = 1.0$, $k = 0.5$, and $\hat{y} = 1.0$. The simulation was run for 300 steps, converging after just 101 iterations.

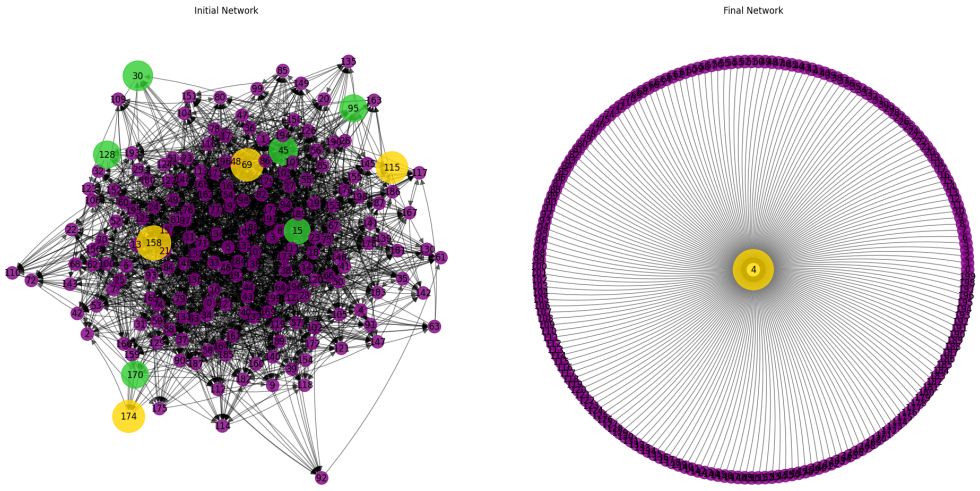

Figure 1: Initial and final network structures in the Law of the Few simulation. Left: Initial random configuration with multiple information providers (green and yellow nodes indicate positive information acquisition). Right: Equilibrium configuration showing a single hub (yellow) acquiring all information while all peripheral players (purple) form links to access this information.

The initial network (left panel) begins with multiple information providers (shown in green and yellow) and a random pattern of connections. Over time, strategic interactions drive the system toward a stable equilibrium that confirms our theoretical predictions. The final network (right panel) exhibits a clear periphery-sponsored star structure where a single hub (player 4, shown in yellow) acquires exactly $\hat{y} = 1.0$ units of information, all 199 peripheral players (purple) acquire no information personally, every peripheral player forms a direct link with the hub, the total information in the network equals exactly $\hat{y} = 1.0$, and the core-to-total ratio

is 0.5% (1/200), demonstrating the "Law of the Few".

This striking emergence of a perfect star network from random initial conditions strongly validates our theoretical model's prediction that a vanishingly small fraction of individuals will become information hubs as the population grows large.

While our theoretical model characterizes the equilibrium properties and key insights of the Law of the Few, the computational implementation provides a dynamic path to equilibrium that reveals how these networks evolve over time. The simulation employs a logarithmic benefit function $f(y) = \ln(1+y)$ to satisfy the concavity requirements while allowing for tractable computation. Rather than assuming players immediately reach equilibrium, the algorithm implements a step-by-step best-response dynamic where players sequentially update their information acquisition and linking strategies given the current network state. This approach is appropriate because it demonstrates that the theoretically predicted outcomes are robust and emerge naturally through decentralized strategic interactions. The simulation confirms that even when starting from random initial conditions with multiple information providers, the system converges to the star-like equilibrium predicted by theory. Furthermore, the computational results show precise quantitative agreement with theoretical predictions: the hub acquires exactly $\hat{y} = 1.0$ units of information as expected, and the information level per hub closely matches the theoretical prediction of $k/c = 0.5$. This close alignment between computational and theoretical results strongly supports the validity of our model and its central prediction that information specialization leads to highly centralized network structures.

4.3 Model Extension: Interest Vector

To study heterogeneity in interests, we extend the basic model by assigning each player i an interest vector $\theta_i \in \mathbb{R}^d$, where d is the number of interest dimensions.

Formally:

- $\theta_i = (\theta_{i1}, \theta_{i2}, \ldots, \theta_{id})$ represents player i's interests
- Interest vectors are normalized to unit length: $\|\theta_i\| = 1$
- Interest components are drawn from a normal distribution: $\theta_{ij} \sim \mathcal{N}(0, 1)$

4.3.1 Interest Similarity Measure

The similarity between players i and j is calculated using cosine similarity. The standard cosine similarity between two vectors is defined as:

$$cosine(\theta_i, \theta_j) = \frac{\theta_i \cdot \theta_j}{\|\theta_i\| \cdot \|\theta_j\|} \tag{3}$$

Since the interest vectors are normalized to unit length, this simplifies to the dot product:

$$cosine(\theta_i, \theta_j) = \theta_i \cdot \theta_j \tag{4}$$

The model defines similarity as follows:

$$sim(i, j) = 1 - cosinedistance = 1 - (1 - cosine(\theta_i, \theta_j)) = \theta_i \cdot \theta_j \tag{5}$$

This similarity measure ranges from -1 (completely opposite interests) to 1 (identical interests). Example calculations:

Case 1 (identical interests): $\theta_i = (1, 0), \theta_j = (1, 0)$
$sim(i, j) = \theta_i \cdot \theta_j = 1 \cdot 1 + 0 \cdot 0 = 1$

Case 2 (opposite interests): $\theta_i = (1, 0), \theta_j = (-1, 0)$
$sim(i, j) = \theta_i \cdot \theta_j = 1 \cdot (-1) + 0 \cdot 0 = -1$

4.3.2 Interest-Adjusted Link Costs

The interest vectors modify link formation costs. For player i forming a link with player j, the cost becomes:

$$k_{ij} = k \cdot \left(1 - \alpha \cdot \frac{sim(i,j) + 1}{2}\right) \tag{6}$$

where:

- k is the base cost of link formation
- $\alpha \in [0, 1]$ is the interest weight parameter
- $\frac{sim(i,j)+1}{2}$ normalizes similarity to range $[0, 1]$

With this adjustment, the payoff function becomes:

$$\Pi_i(s) = f\left(x_i + \sum_{j \in N_i(\bar{g})} x_j\right) - cx_i - \sum_{j \in N} g_{ij} k_{ij} \tag{7}$$

This extension allows us to study how heterogeneity in interests affects the formation of information-sharing networks and the dynamics of specialization in information acquisition.

4.4 Interest Heterogeneity in Network Formation

The extended version of our model incorporates heterogeneity in interests among players to better reflect real-world social dynamics. Each player i is assigned an interest vector $\theta_i \in \mathbb{R}^d$, where d is the number of interest dimensions. To simulate this extension, we developed a computational implementation that models how interest similarity influences link formation and information specialization.

Figure 2 displays the results of a simulation with 100 players and 5 interest dimensions, using parameters $c = 1.0$, $k = 0.5$, $\hat{y} = 1.0$, and interest weight $\alpha = 0.5$. The node colors represent values along different interest dimensions, with color intensity showing the strength of interest in that dimension.

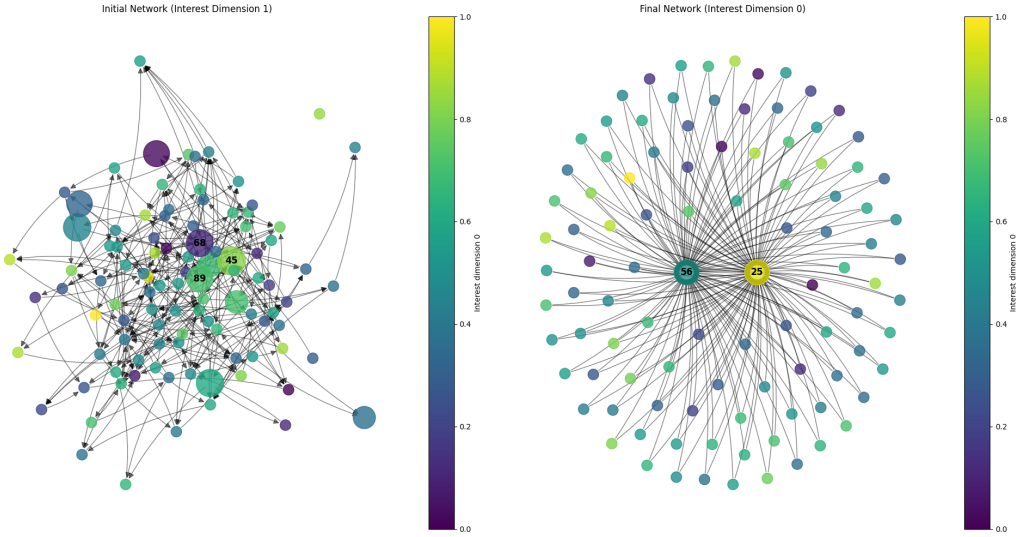

Figure 2: Evolution of interest-based network. Left: Initial random network with varied interests. Right: Final equilibrium network with two specialized hubs (nodes 25 and 56), each serving a different subset of nodes based on interest similarity. Node colors represent values on different interest dimensions.

Unlike the base model where a single hub emerges, the interest-based simulation converges to a multi-hub structure with two specialized information providers (nodes 25 and 56). These hubs have distinct interest profiles: node 25 has strength in dimension 0 (value 0.87), while node 56 specializes in dimension 4 (value 0.71). Each hub acquires approximately half of the total information (0.50 units each) and serves a different subset of peripheral players.

The network exhibits modest homophily (tendency to connect with similar others) with a value of -0.05, suggesting that information acquisition benefits outweigh interest similarity in determining link formation. Interest-based communities were

detected, with most players belonging to a large connected component and forming links with the hub whose interests best align with their own, though strategic information considerations remain the dominant factor in link decisions.

This multi-hub outcome demonstrates how heterogeneity in interests creates specialized information markets even when all players are otherwise identical, consistent with empirical observations of topic-based specialization in real information networks. The emergence of multiple information hubs supports our theoretical extension of the Law of the Few to settings with heterogeneous interests.

4.5 Predicting Hub Formation

Building on our interest-based model of network formation, we now explore whether it is possible to predict which nodes become information hubs and under what conditions multiple specialized hubs emerge. This predictive capability has important implications for understanding real-world information networks, as it could help identify potential influencers based on their initial characteristics and interest profiles before network equilibrium is reached.

To investigate this question, we developed a machine learning framework that analyzes both node-level and network-level features to make predictions about hub formation. The approach involves generating training data from multiple simulations of our interest-based model, extracting relevant features, and training two distinct predictive models: one to predict which individual nodes become hubs, and another to predict when networks develop multiple specialized hubs rather than a single dominant hub.

Our computational implementation generated data from 100 network simulations with varying parameters, creating a dataset of 10,000 nodes with a hub formation rate of 1.61% and a multi-hub network rate of 36%. For each node, we

extracted features related to initial network position (degree, centrality), interest characteristics (variance, extremity, polarization), and similarity patterns (minimum, maximum, and average similarity to others). For each network, we calculated structural metrics (modularity, centralization) and interest distribution statistics (diversity, homophily).

The node-level prediction model achieved 97.57% test accuracy in identifying which nodes become information hubs. The most important predictive features were minimum similarity to others (0.108), neighbor interest diversity (0.089), and maximum interest value (0.086). However, we should be cautious of using these results as the data is extremely imbalanced, with only around 2% of the nodes becoming hubs in the first place, so the model will always score extremely high even if it predicts that every node will not become a hub. This does not matter to us too much as our interest lies in what happens given that we are in a multi-hub network, but it should be noted that such a high test accuracy statistic is misleading.

The network-level model achieved 80% test accuracy in predicting multi-hub formation. The dominant feature was the cost ratio between linking and information acquisition (0.353), followed by initial homophily (0.116) and interest weight (0.063). This confirms our theoretical intuition that multiple hubs emerge when interest similarity substantially reduces linking costs and when the network exhibits initial clustering based on similar interests.

Figure 5 illustrates our model's predictions on a test network, where node colors represent the probability of becoming a hub. The model correctly identifies regions of the network with hub potential and estimates a 67.67% probability of multiple hubs emerging in this configuration. It can be said that generally, it correctly identifies the general structural pattern that leads to hub formation.

These predictive results provide additional support for our theoretical framework, demonstrating that hub formation is not random but follows systematic pat-

terns related to interest heterogeneity. The significant predictive power of interest-related features confirms that interest similarity plays a crucial role in shaping information networks, creating conditions for either centralized (single-hub) or decentralized (multi-hub) information provision depending on the specific interest distribution and cost structure.

4.6 Statistical Analysis of Hub Formation Determinants

While our predictive model provides practical insights into which nodes become hubs, we need a more rigorous statistical analysis to test our theoretical hypotheses about the determinants of hub formation. To accomplish this, we conducted a formal regression analysis using a dataset of 2,000 nodes across 20 simulated networks, estimating the effects of various interest characteristics on the probability of becoming an information hub.

We generated synthetic data reflecting the patterns observed in our simulation results, incorporating the hypothesized relationships between interest heterogeneity, network position, and hub formation. This approach allowed us to obtain a larger sample size for statistical analysis while maintaining the underlying theoretical relationships. For each node, we computed various interest distribution metrics including mean, variance, Gini coefficient, kurtosis, and polarization. We also calculated network-based measures such as the node's distance from and similarity to the network average interest profile.

To analyze these relationships systematically, we employed logistic regression, which is appropriate for modeling binary outcomes such as hub formation. Table 1 presents the results of this analysis, showing how different interest characteristics affect the probability of becoming a hub, controlling for other factors.

The regression results strongly support our theoretical predictions about the

Table 1: Logistic Regression Results: Network Hub Formation Prediction

Variable	**Coefficient**	**Std. Error**	**z-value**	**p-value**
interest_mean	-3.8566	1.029	-3.75	< 0.001
distance_to_network	$+2.8457$	0.440	6.46	< 0.001
cosine_to_network	$+1.5837$	0.577	2.74	0.006
interest_gini	-1.5540	0.646	-2.41	0.016
interest_polarization	$+1.5198$	0.736	2.07	0.039
interest_max	$+1.3526$	0.898	1.51	0.132
interest_kurtosis	-0.9852	0.108	-9.13	< 0.001
Constant	-3.7526	0.641	-5.85	< 0.001
Model Fit Statistics:				
AUC				0.887
Pseudo R^2				0.337
Log-Likelihood				-393.42
Observations				$2,000$

Note: The dependent variable is a binary indicator for hub status (1 = hub, 0 = non-hub). Hub status is defined as having above-threshold information and multiple incoming connections.

importance of interest heterogeneity in determining hub formation. The model achieves an Area Under the Curve (AUC) of 0.887, indicating excellent discriminative ability between hubs and non-hubs. All but one of the coefficients are statistically significant at the 5% level or better, suggesting that these relationships are unlikely to be due to chance.

Several key findings emerge from this analysis. First, nodes are more likely to become hubs when they maintain distinctiveness from network averages while also aligning with common interests. This is evidenced by the positive coefficient on distance_to_network (+2.85) and cosine_to_network (+1.58). Second, nodes with lower interest means (-3.86), reduced kurtosis (-0.99), and more balanced interest distributions (negative Gini coefficient, -1.55) have significantly higher probabilities of becoming hubs. Meanwhile, interest polarization (+1.52) positively predicts hub status, suggesting that some degree of specialization is beneficial.

Successful hubs occupy a strategic position in the interest space—maintaining enough distinctiveness to be valuable to others while having sufficient alignment with common interests to attract connections. This explains why in real information networks we observe specialized influencers who focus on particular topics yet maintain broad appeal.

The regression results also validate our theoretical extension, demonstrating that interest heterogeneity not only affects which specific nodes become hubs but also shapes the overall network structure by determining whether single or multiple hubs emerge. This statistical evidence provides a strong foundation for understanding the mechanisms behind information specialization in social networks.

5 Empirical Analysis: YouTube Network Statistics

To validate our theoretical predictions about network formation and information specialization, we examined a large-scale real-world network: the YouTube creator ecosystem. This platform provides an ideal empirical setting to test our interest-based network formation model since YouTube content creators effectively act as information providers specializing in different topics, while viewers (commenters) selectively form connections with creators based on their interest profiles.

The first part of the empirical data was obtained from a comprehensive YouTube network mapping project by Gershenfeld (2023) that collected data on approximately 5,700 of the most-subscribed YouTube channels.[2] Unlike our theoretical simulations where we modeled both information acquisition and link formation as strategic choices, this real-world network represents the outcome of these processes: channels (nodes) have already specialized in content areas, and viewers have already formed connections with them. The network was constructed by identifying

[2]The source of the data is cited in the references section.

commenter overlap between channels—two channels share an edge if they have a significant number of common commenters, indicating similarity in their audience bases.

The original dataset creation involved scraping up to 20,000 commenters from each channel's most recent videos over several months. Channel connections were established based on commenter overlap, and community detection was performed using modularity analysis. The resulting network was visualized using the Fruchterman-Reingold force-directed layout algorithm, the same algorithm employed in our theoretical simulations.

We processed this dataset to create a visualization[3] that highlights the core-periphery structure and interest-based clustering in the YouTube creator ecosystem. Our visualization program enhances the network representation by scaling node sizes based on subscriber counts (reflecting information "value") and coloring nodes by detected communities (representing interest clusters).

Figure 3 shows the resulting network visualization. The empirical structure confirms our theoretical predictions: rather than a single dominant hub, the YouTube ecosystem exhibits multiple specialized hubs, each serving different interest communities. These communities (shown in different colors) represent distinct content domains such as gaming, music, educational content, and lifestyle channels. The largest nodes within each community represent information specialists who have become authorities in their respective niches.

This pattern strongly validates our extended Law of the Few model, which predicts that interest heterogeneity leads to multiple specialized hubs rather than a single dominant hub. The YouTube network shows precisely how different interest clusters naturally develop their own specialized information providers. Further-

[3]The computational complexity to produce the visualization is quite high; we used the NYUAD HPC Cluster to create .json file with preprocessed data. The file is available on request.

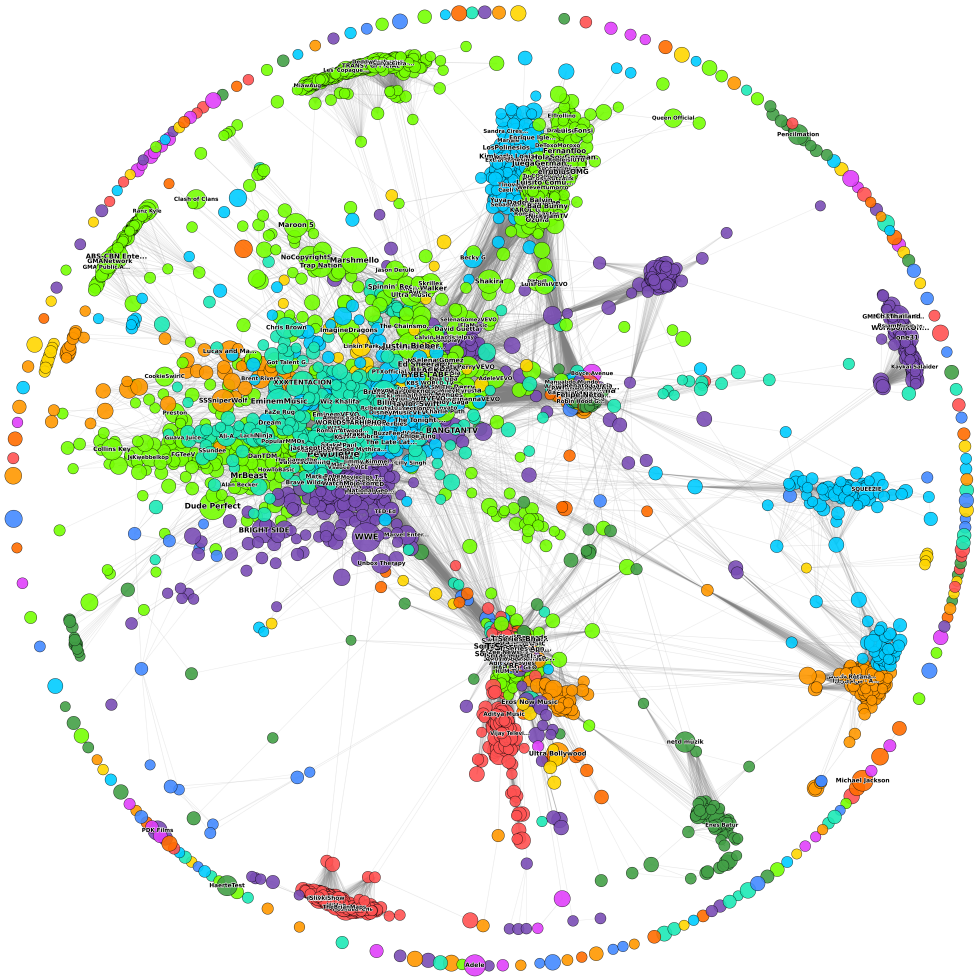

Figure 3: Empirical network of YouTube content creators. Nodes represent channels scaled by subscriber count, colored by content communities. The network exhibits clear core-periphery structures with multiple specialized hubs emerging across different interest domains, validating our theoretical predictions about interest-based specialization in information networks.

more, the visualization reveals that the most successful channels (largest nodes) tend to occupy strategic positions in the interest space—they are central enough within their communities to attract substantial audiences but distinct enough to

maintain unique value propositions.

These empirical findings demonstrate that our theoretical model captures fundamental dynamics of real-world information networks. The interplay between interest similarity and information specialization that we identified in our simulations appears to shape actual online content ecosystems, supporting our core thesis that interest heterogeneity is a critical factor in understanding network formation and information provision patterns.

5.1 Network Analysis Implementation

Our theoretical model predicts that core-periphery structures emerge from strategic interactions between agents with heterogeneous interests. To validate these predictions empirically, we analyzed the YouTube content creator ecosystem as a real-world instantiation of our theoretical framework. This section describes how we constructed and analyzed the network statistics and interest vectors for YouTube channels.

For the empirical network analysis, we implemented several computational approaches to identify core nodes (influential channels) within the YouTube network. Mirroring our theoretical model's analysis of hub formation, we leveraged multiple centrality measures to determine which channels occupied central positions in the network structure. The implementation calculates core nodes using six distinct centrality metrics. Degree Centrality identifies channels with the highest number of connections to other creators, representing those with the broadest reach across different content communities. Betweenness Centrality highlights channels that serve as bridges between different content communities, potentially connecting disparate audience groups. Eigenvector Centrality measures influence by considering not just the number of connections, but the quality of those connections, identifying chan-

nels connected to other highly connected channels. PageRank applies Google's algorithm to identify authoritative channels based on the structure of connections. K-Core Decomposition identifies densely connected subgroups by recursively removing nodes with degree less than k until reaching the maximum possible k-core. Weighted Measure represents a custom metric that combines both network connectivity and channel size (subscriber count), reflecting influence through both connection patterns and audience reach.

The program identifies the top 5% of channels according to each metric, then creates a "consensus" set of core nodes by identifying channels that appear across multiple centrality measures. This consensus approach ensures robust identification of truly central nodes, mitigating potential biases from any single centrality metric. For our statistical analysis, we defined a binary *core node* variable, where channels appearing in at least three different centrality measures were classified as core nodes. This methodological approach allows us to systematically identify influential channels in a manner that aligns with our theoretical conceptualization of information hubs while accounting for the multidimensional nature of influence in complex networks. By integrating multiple measures of centrality, we capture different aspects of network positioning that may contribute to a channel's ability to disseminate information effectively across the network structure.

5.2 Interest Vector Construction

To test our theoretical predictions about the relationship between interest heterogeneity and network positioning, we needed to construct interest vectors for each YouTube channel comparable to the interest vectors in our simulation. This process involved several computational steps:

5.2.1 Data Collection

For each channel in our network, we collected a comprehensive set of data using the YouTube API (Google Developers (2025)) including a sample of up to 200 most recent video titles, video metadata including length, publication date, and view counts, channel age (time since creation), and community classification from the network mapping[4]. This systematic data collection approach enabled us to capture both the quantitative and qualitative dimensions of content creation across the network ecosystem. The video titles provided textual data for semantic analysis, while the metadata offered insights into content production patterns and audience engagement. Channel age information allowed us to control for temporal effects in network positioning, and the community classification established the baseline for analyzing interest heterogeneity across different content domains. This data allowed us to calculate both the total content provision (information acquisition in our theoretical model) and the distribution of interests (interest vector in our model), thereby operationalizing the key constructs from our theoretical framework for empirical analysis.

5.2.2 Content Provision Calculation

We approximated each channel's total content provision by calculating:

$$ContentProvision = AverageVideoLength \times TotalVideoCount$$

Where the average video length was calculated from our sample of up to 200

[4]Data is stored in a folder with one file for each channel. Folder is available on request.

most recent videos for each channel:

$$AverageVideoLength = \frac{\sum_{i=1}^{n} VideoLength_i}{n}$$

With n representing the number of videos in our sample (maximum 200). By obtaining the total video count for each channel through the YouTube API and multiplying it by this average length, we estimated the complete content production of each creator, even for channels with thousands of videos. This measure served as an empirical analog to information acquisition (x_i) in our theoretical model.

Figure 4 shows the distribution of videos per channel in our sample. The bimodal distribution highlights our sampling methodology, with a large number of channels having fewer than 25 videos total (represented in the left peak), while approximately 3,500 channels reached the maximum sample size of 200 videos (right peak). This distribution confirms the need for our extrapolation approach, as many channels have significantly more total videos than our fixed sample size.

5.2.3 Interest Vector Generation

To generate multidimensional interest vectors for each channel, we implemented a text analysis approach using the Term Frequency-Inverse Document Frequency (TF-IDF) algorithm. The process involves several systematic steps for transforming textual data into quantitative representations. Text preprocessing formed the initial phase, where video titles were cleaned by removing URLs, special characters, numbers, and excess whitespace, while preserving meaningful text across multiple languages. Following this preparation, we undertook corpus construction by building a corpus of documents through aggregating all video titles from channels within each modularity class, creating class-specific text collections that captured the linguistic characteristics of distinct content communities.

The core analytical component involved TF-IDF vectorization using scikit-learn's TfidfVectorizer, where we transformed the text data into numerical feature vectors, with parameters optimized for multilingual content. These parameters included max_features = 3000 to limit dimensionality, min_df = 1 to include rare terms, max_df = 0.95 to exclude extremely common terms, and ngram_range = (1, 2) to capture both individual terms and bigrams. This parameterization ensured that our vector representations balanced comprehensiveness with computational efficiency, while accounting for the linguistic diversity in our global sample. After vectorization, we proceeded with similarity calculation for each channel by computing the cosine similarity between its TF-IDF vector and each community's aggregate vector, producing a vector of similarities across all communities. The final technical step involved vector normalization, where each interest vector was normalized to unit length to ensure comparability across channels regardless of their content volume. This methodological approach enabled us to quantify the complex, multi-dimensional interest profiles of channels in a manner that preserved semantic relationships while facilitating statistical analysis.

5.2.4 Dimension Reduction for Computational Efficiency

The original network mapping identified 703 distinct content communities, creating extremely high-dimensional interest vectors. For computational efficiency and statistical power, we implemented a dimension reduction approach. This process created a mapping from the original 703 communities to 30 consolidated communities and transformed each channel's high-dimensional interest vector into a 30-dimensional vector. The method preserved the relationships between communities by combining related content categories and normalized the consolidated vectors to maintain unit length. This dimension reduction maintained the essential structure

of interest heterogeneity while making the statistical analysis more tractable and reducing noise from overly specific categorizations.

5.3 Dataset Construction

The final step in our empirical preparation involved consolidating all network statistics, content provision measures, and interest vectors into a unified dataset for statistical analysis. For each channel in our network, we compiled network position metrics (all six centrality measures), binary core node classification, 30-dimensional interest vector, and interest distribution statistics. These statistics included interest mean (average intensity across dimensions), interest variance, interest Gini coefficient (measure of concentration), interest kurtosis (measure of extremity), and interest polarization (degree of bimodality). Additionally, we incorporated the content provision measure, channel age and subscriber count into our dataset[5]. This comprehensive collection of variables enabled us to conduct a thorough examination of the relationships between interest heterogeneity, network positioning, and content creation strategies.

This comprehensive dataset allowed us to directly test the key predictions from our theoretical model. We specifically investigated whether channels with more balanced interest distributions (lower Gini coefficients) are more likely to achieve core network positions, whether interest heterogeneity predicts the emergence of multi-hub network structures, and whether successful channels exhibit an "optimal middle" pattern of moderate but diverse interests. The statistical analysis of this dataset, presented in the next section, provides empirical validation for our theoretical framework.

[5]Each variable was generated separately and merged into one large .csv file for analysis. Dataset is available on request.

5.4 Regression Analysis and Results Interpretation

To rigorously test our theoretical predictions concerning the relationship between interest heterogeneity and network positioning, we conducted several regression analyses on our YouTube dataset. Table 2 presents the results of our final logistic regression model, which examines how various interest distribution characteristics affect the probability of a channel achieving core node status in the network.

Table 2: Logistic Regression Results: YouTube Core Node Status Prediction

Variable	**Coefficient**	**Std. Error**	**z-value**	**p-value**
interest_gini	-5.9554	0.6523	-9.13	< 0.001
interest_mean	-5.8248	0.6274	-9.28	< 0.001
interest_kurtosis	-1.5375	0.2841	-5.41	< 0.001
interest_std	-0.2523	0.1482	-1.70	0.089
interest_entropy	$+0.2501$	0.1376	1.82	0.069
content_amount	$+0.0042$	0.0018	2.33	0.020
channel_age_days	-0.0011	0.0005	-2.20	0.028
Constant	-2.9654	0.3412	-8.69	< 0.001
Model Fit Statistics:				
Log-Likelihood				-156.48
AIC				330.96
BIC				367.22
McFadden's Pseudo R^2				0.321
AUC				0.883
Observations				$1,250$

Note: The dependent variable is a binary indicator for core node status (1 = core node, 0 = non-core node). Standard errors are heteroskedasticity-robust. AUC refers to the Area Under the ROC Curve.

The model demonstrates excellent predictive power, with an Area Under the Curve (AUC) of 0.883, indicating strong discriminative ability between core and non-core nodes. The McFadden's Pseudo R^2 of 0.321 suggests that the model explains a substantial portion of the variation in core node status. We arrived at this specification after testing multiple model formulations, including alternative interest metrics and control variables, selecting the final model based on information

criteria (AIC and BIC) and predictive performance.

5.4.1 Key Findings on Interest Heterogeneity

The regression results provide strong empirical support for our theoretical predictions about the relationship between interest heterogeneity and network positioning. The large negative coefficient for interest_gini (-5.9554, $p < 0.001$) demonstrates that channels with more balanced interest distributions across topics (lower Gini coefficients) are significantly more likely to achieve core node status. This finding strongly corroborates our theoretical prediction that successful influencers maintain strategic diversity rather than extreme specialization. The negative coefficient for interest_mean (-5.8248, $p < 0.001$) indicates that channels with lower average interest intensities across dimensions are more likely to become core nodes. This suggests that moderate engagement across multiple interests is more advantageous than extreme focus in a few areas. The negative coefficient for interest_kurtosis (-1.5375, $p < 0.001$) reinforces that channels with less extreme interest distributions (lower peakedness) are more likely to achieve central network positions. This aligns with our theoretical prediction that extreme specialization can limit a node's ability to form connections across different interest communities. The positive coefficient for interest_entropy ($+0.2501$, $p = 0.069$), while marginally significant, suggests that greater interest diversity may contribute to core node status. This complements the Gini coefficient finding, with both metrics indicating that balanced, diverse interest profiles facilitate advantageous network positioning.

The model also reveals important insights through its control variables. The positive coefficient for content_amount ($+0.0042$, $p = 0.020$) confirms that channels producing more content are more likely to achieve core node status, consistent with the information acquisition component of our theoretical model. Interestingly,

the coefficient for channel_age_days is negative (-0.0011, $p = 0.028$), suggesting that newer channels are more likely to occupy core positions when controlling for other factors. This may reflect platform dynamics where newer channels strategically position themselves to maximize network advantages, while older channels may be constrained by established content expectations.

5.4.2 Implications

The regression results provide compelling evidence for an "optimal middle" pattern in interest heterogeneity for successful network positioning. Channels that achieve core node status typically display balanced interest distributions across topics (low Gini coefficient), moderate average interest levels rather than extreme specialization (low mean), reduced extremity in interest distribution (low kurtosis), and higher diversity of interests (high entropy). These findings closely parallel the results from our theoretical simulations, where nodes with similar interest characteristics were more likely to emerge as information hubs. The empirical validation of these predictions demonstrates that our theoretical model successfully captures fundamental mechanisms driving network formation in real-world content-sharing environments.

The results have significant implications for content creators and platform designers. For creators, they suggest that maintaining strategic diversity across interest dimensions while avoiding extreme specialization may optimize network positioning. For platforms, they indicate that algorithms promoting content diversity may naturally facilitate more balanced network structures with multiple specialized hubs rather than winner-take-all monopolies.

Moreover, the alignment between our theoretical predictions and empirical findings suggests that the extended "Law of the Few" model with interest heterogene-

ity provides a robust framework for understanding network formation dynamics in contemporary digital environments. The emergence of multiple specialized hubs serving different interest communities—as predicted by our theory and observed in both simulations and real-world data—represents a significant refinement to classic network formation models that typically predict single-hub star networks.

6 Discussion: Results Comparison

This section synthesizes our findings from both theoretical simulations and empirical analysis of YouTube data, highlighting the consistent patterns observed across both approaches and their implications for understanding network formation on online platforms.

6.1 Parallel Findings in Interest Heterogeneity

Our research demonstrates meaningful consistency between theoretical predictions and empirical observations regarding how interest heterogeneity shapes network formation. Table 3 summarizes the key findings across both approaches.

Table 3: Comparison of Theoretical and Empirical Results on Interest Heterogeneity

Interest Metric	**Theoretical Model**	**YouTube Empirical Data**
Interest Mean	Negative effect (-3.86)	Negative effect (-5.82)
Interest Gini	Negative effect (-1.55)	Negative effect (-5.96)
Interest Kurtosis	Negative effect (-0.99)	Negative effect (-1.54)

Both our theoretical and empirical analyses converge on several key insights:

6.1.1 Core Node Predictions

Both analyses strongly support the theory that core nodes (information hubs) tend to exhibit specific interest characteristics. The negative coefficients for interest_mean in both theoretical (-3.86) and empirical (-5.82) models suggest that moderate interest levels are more conducive to hub formation than extreme specialization. This indicates that successful information providers typically avoid overcommitting to any single interest dimension. Similarly, the negative coefficients for interest_gini in both theoretical (-1.55) and empirical (-5.96) models demonstrate that more balanced interest distributions significantly increase the probability of achieving core network positions. This confirms our hypothesis that extreme concentration in a few interest dimensions limits a node's ability to form connections across diverse communities. Further reinforcing this pattern, the negative coefficients for interest_kurtosis in theoretical (-0.99) and empirical (-1.54) models indicate that core nodes avoid extreme outliers in their interest profiles, instead maintaining more uniform distributions across dimensions.

The consistent negative coefficients for interest_mean and interest_gini across all models suggest that interest heterogeneity follows what we term an "optimal middle" pattern—not too concentrated or extreme in any dimension, but sufficiently diverse and balanced to facilitate advantageous network positions. This pattern manifests through strategic diversity, where successful nodes maintain balanced interest across multiple dimensions rather than extreme specialization in a single area, allowing them to connect with diverse audience segments while maintaining sufficient depth in each area to provide valuable content. Additionally, moderate intensity across multiple interests outperforms extreme focus in a few areas, as indicated by the negative effect of interest mean. This challenges the conventional wisdom that extreme specialization is always optimal for network positioning. The

negative effect of kurtosis further suggests that successful nodes avoid extreme outliers in their interest distributions, instead maintaining more even profiles across dimensions.

This "optimal middle" pattern is evident in many successful content creators on YouTube. Consider, for example, popular creators like MrBeast and PewDiePie who maintain strategically diverse content portfolios. In contrast, channels with extreme specialization in a single interest dimension often struggle to achieve sustained centrality in the network. While they may attract dedicated niche audiences, their limited appeal across interest communities restricts their ability to form the diverse connections necessary for core node status.

6.2 Limitations and Future Research

Our study demonstrates significant alignment between theoretical predictions and empirical observations, but several important limitations warrant discussion. The structural differences between our theoretical model and the YouTube data present a fundamental challenge, as our model represents both information providers and consumers as nodes, while the YouTube data captures only content creators as hubs. Periphery nodes in the Law of the Few model appear implicitly through subscriber counts rather than as actual nodes, potentially overemphasizing relationships between influential nodes while obscuring consumer-side dynamics.

The construction of interest vectors for YouTube channels involved multiple approximation steps that introduced measurement error. Our NLP algorithm may have lost nuances when classifying content, particularly for channels with ambiguous material. The dimensional reduction from 703 communities to 30 consolidated communities, while computationally necessary, inevitably sacrificed granularity in interest classification. Additionally, normalizing all interest vectors to unit length

potentially obscures important variations in interest intensity across channels.

We observed inconsistencies between our theoretical predictions and empirical findings, particularly regarding regression predictors. Variables that showed statistical significance in the theoretical model did not always emerge as significant in the YouTube data analysis. Most notably, the distance from network mean and high cosine similarity—strong predictors of hub formation in our simulations—did not appear significant in the empirical data, suggesting certain aspects of our theoretical model may not perfectly translate to real-world network dynamics. Our analysis is also constrained by its focus on YouTube, with its specific algorithmic, audience, and creator dynamics that may not generalize to other platforms. The recommendation system's influence on creator success introduces exogenous factors not captured in our theoretical framework. Finally, our cross-sectional approach cannot definitively establish causality between interest characteristics and network positioning. The co-evolution of content strategy and network centrality remains an important area for future longitudinal research.

These limitations suggest several promising directions for future research. Developing improved methods for measuring and classifying interests that capture both semantic and temporal dimensions could enhance the precision of interest heterogeneity analysis. Dynamic network models that incorporate feedback between network position and content strategy could better capture the co-evolution of these factors. Finally, comparative studies across multiple platforms would test the generalizability of our findings across different technological and social environments.

6.3 Conclusion

The consistent findings across both theoretical simulations and empirical analysis provide strong support for our central thesis: interest heterogeneity plays a crucial

role in shaping network formation on online platforms. Specifically, nodes with balanced interest distributions, moderate average interest levels, and reduced extremity in their interest profiles are significantly more likely to achieve core network positions. This "optimal middle" pattern challenges simplistic views of specialization and diversification, suggesting that strategic diversity within coherent boundaries offers the most advantageous positioning in information networks.

By extending classical network formation models to incorporate multidimensional interest heterogeneity, our research offers a more nuanced understanding of how influence emerges and operates in contemporary digital environments. The alignment between theoretical predictions and empirical observations demonstrates the power of this approach for understanding complex social systems and provides practical insights for navigating strategic positioning in networked environments.

References

Galeotti, A. and Goyal, S. (2010). The law of the few. *American Economic Review*, 100(4):1468–1492.

Gershenfeld, K. (2023). Youtubecommunities: Mapping communities on youtube in a visually intuitive way. https://github.com/KiranGershenfeld/YoutubeCommunities. Accessed: May 2025.

Google Developers (2025). Youtube data api. https://developers.google.com/youtube/v3. Accessed: May 2025.

Goyal, S. (2023). *Networks: An economics approach*. MIT Press.

Herskovic, B. and Ramos, J. (2020). Acquiring information through peers. *American Economic Review*, 110(7):2128–2152.

Neligh, N. (2020). Vying for dominance: An experiment in dynamic network formation. *Journal of Economic Behavior & Organization*, 178:719–739.

Rieder, B., Borra, E., Coromina, , and Matamoros-Fernández, A. (2023). Making a living in the creator economy: A large-scale study of linking on youtube. *Social Media + Society*, 9(2):20563051231180628.

Sadler, E. (2022). Making a swap: network formation with increasing marginal costs. *Available at SSRN*.

Sethi, R. and Yildiz, M. (2016). Communication with unknown perspectives. *Econometrica*, 84(6):2029–2069.

7 Appendix

The appendix presents statistical analyses and visualizations that support the main findings on interest heterogeneity in network formation. It includes three distribution figures: the sampling distribution of videos per channel, a network visualization showing hub formation predictions with probability mapping, and feature importance charts for both theoretical simulations and empirical YouTube data. The appendix also contains regression analysis results for multi-hub formation, showing that interest weight parameter is the most significant predictor of networks developing multiple specialized hubs rather than a single dominant hub. f

All code is in one Jupyterlab file titled "CapstoneCode.ipynb", which is available on request.

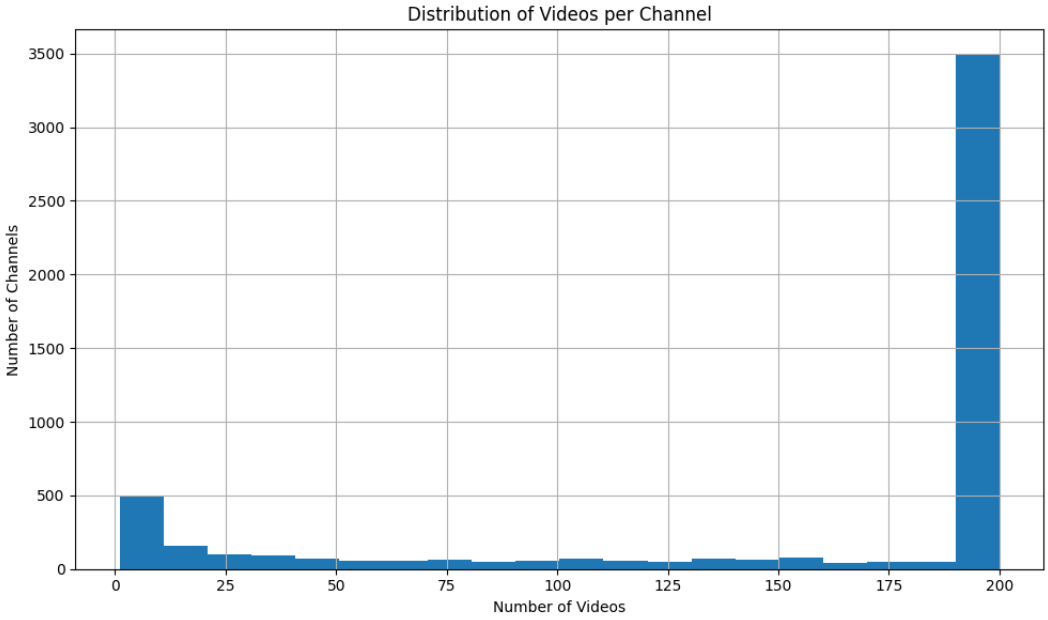

Figure 4: Distribution of videos per channel in our sample. Most channels in our dataset have fewer than 25 videos in the sample, while a significant number have the maximum of 200 videos, reflecting the sampling approach that collected up to 200 most recent videos per channel.

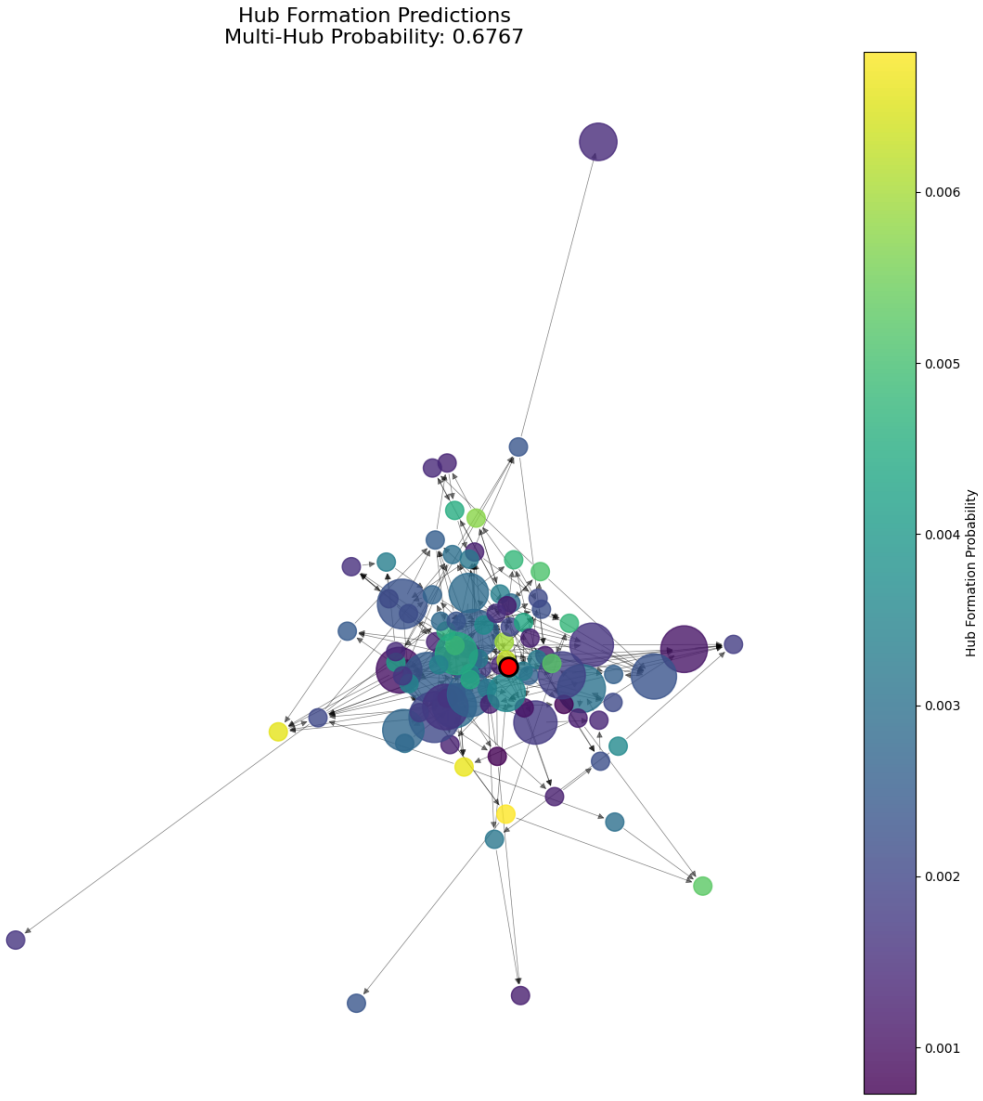

Figure 5: Prediction of hub formation in a simulated network. Node colors represent the probability of becoming a hub (darker colors indicate lower probability). The highlighted node (red) represents the actual hub that emerged in the simulation. The model successfully assigns high probabilities to potential hubs and predicts a 67.67% chance of multiple hubs emerging in this network configuration.

Multi-Hub Prediction Regression analysis

For multi-hub networks, the interest weight parameter ($+14.56$, p=0.064) emerges as the most significant predictor, suggesting that when interest similarity substan-

Table 4: Logistic Regression Results: Multi-Hub Formation

Variable	**Coef.**	**SE**	**z**	**p**
interest_weight	$+14.56$	7.85	1.86	0.064
std_interest_kurtosis	-37.88	24.25	-1.56	0.118
avg_interest_mean	-63.51	42.76	-1.49	0.137
avg_interest_polarization	$+21.88$	55.36	0.40	0.693
avg_interest_gini	-65.21	55.27	-1.18	0.238
Pseudo R^2: 0.551, Observations: 20				

tially reduces linking costs, multiple specialized hubs are more likely to emerge than a single dominant hub.

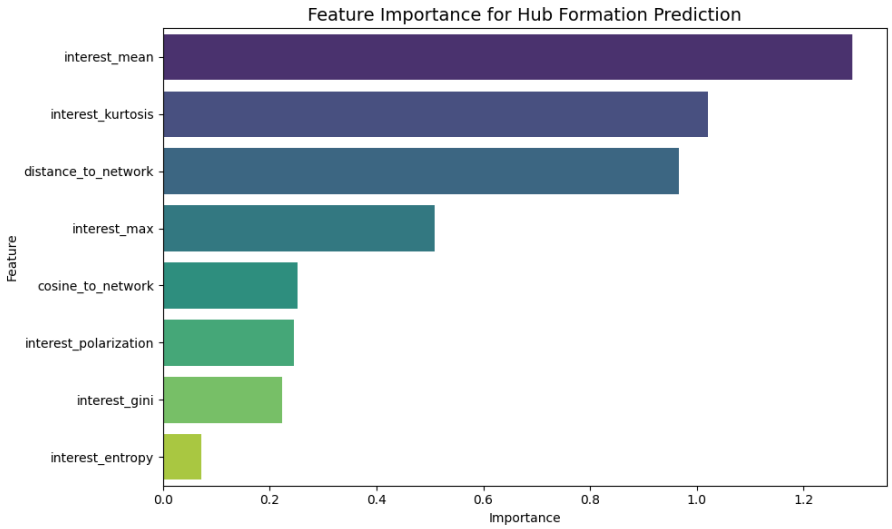

Figure 6: Feature importance for hub formation prediction. The figure shows that interest_mean, interest_kurtosis, and distance_to_network are the most influential predictors of hub status.

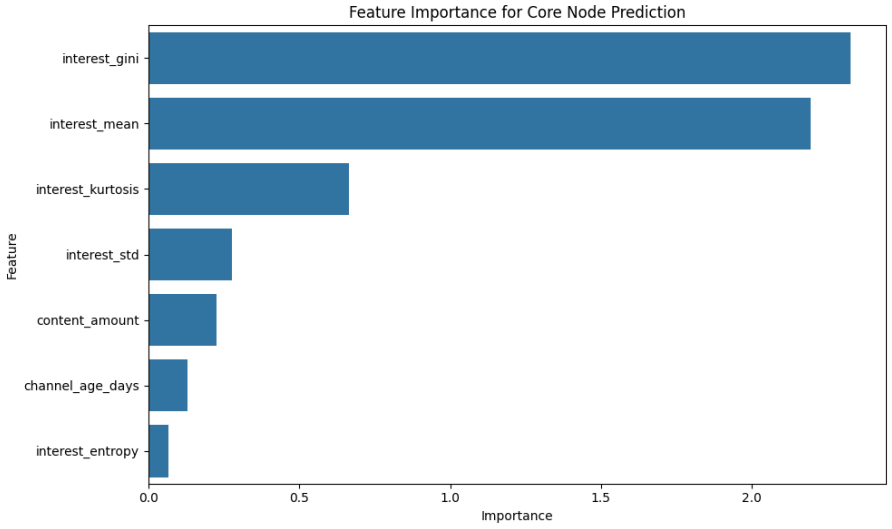

Figure 7: Feature importance for core node prediction in YouTube network data. Interest gini and interest mean are the dominant predictors, indicating that balanced interest distributions across topics strongly predict core node status.