Data Analytics Practice Opportunity 2025/26

Factors Influencing Information Diffusion and Diffusion Patterns in Social Networks

GUAN Qishu [MBTE/3]
REN Miao [FTEC/2]
LU Yuxuan [SOCI/3]

香港中文大學圖書館
The Chinese University of Hong Kong Library

What Is Interest-Driven Diffusion?

In online social networks, information spreads through diffusion cascades that take two characteristic shapes: broadcast-like (wide and shallow) and viral-like (narrow and deep) (Goel et al., 2015). A central mechanism behind these cascades is interest similarity — users are more likely to reshare content that matches their interests. However, the strength of this effect is not uniform: it varies across diffusion structures and changes systematically as the cascade propagates to deeper layers. Understanding this dynamic is essential for both effective content strategy and responsible misinformation control.

1. Introduction

Each day, users tweet, like and retweet across social media platforms, collectively creating large-scale information cascades (Granovetter, 1978). This study examines how interest similarity between users and content drives information diffusion in social networks. While prior work shows a positive association between similarity and retweeting, we argue that this relationship is dynamic and shaped by diffusion pattern and depth. We also show that achieving deep diffusion requires more than an influential creator; it depends on audience - content fit, repeated exposure, reciprocity and favourable network pathways.

2. Research Questions

- How can we distinguish broadcast - like and viral - like diffusion patterns in real social - network cascades?
- How does interest similarity affect retweet probability under different propagation structures and at different depths?
- Are social - network users a homogeneous mass, or do they form distinct personas with different diffusion mechanisms and scale profiles?
- What do these findings imply for audience selection, beyond simple follower counts, in designing effective and responsible communication strategies?

3. Methodology Overview

Propagation Graphs

For each cascade we build a propagation graph and classify cascades as broadcast - like or viral - like.

Binary Logistic Regression and Descriptive Statistics

We modelled retweeting as a binary outcome and estimated how interest similarity affects retweet probability.

User Clustering

We categorised users by interest similarity, exposure, interaction and redundancy to derive personas with users' distinct diffusion roles and community scale profiles.

4.1 Propagation Graphs: Broadcast vs Viral

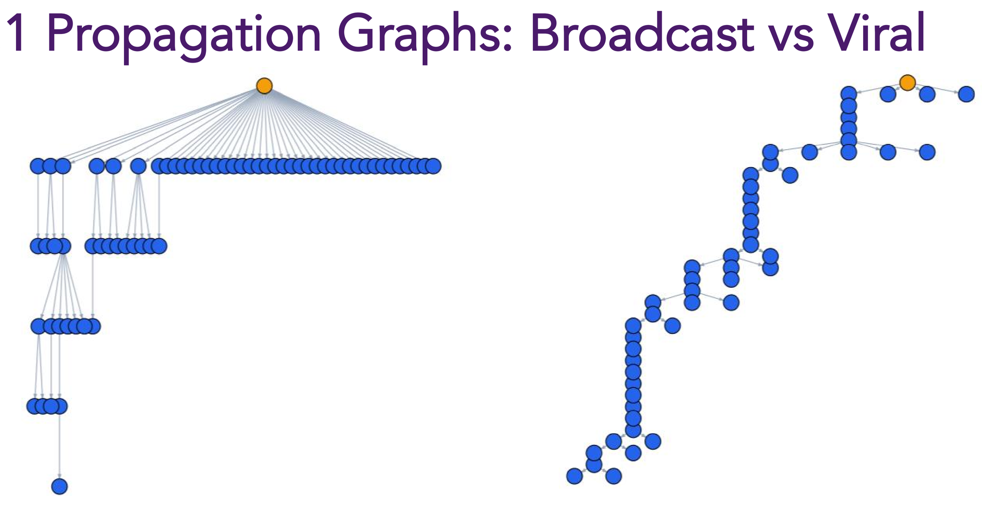


Figure 1. Broadcast-like and Viral-like PropagationTrees

According to the grahpics…

- Broadcast cascades: wider but shallower; dominated by a strong root account and a large first layer of reshares
- Viral cascades: narrower but deeper; mainly driven by audience–content fit, reciprocity and repeated exposure.
- Not all large cascades are viral: some are merely broad one-shot broadcasts without statistcially significant depth.

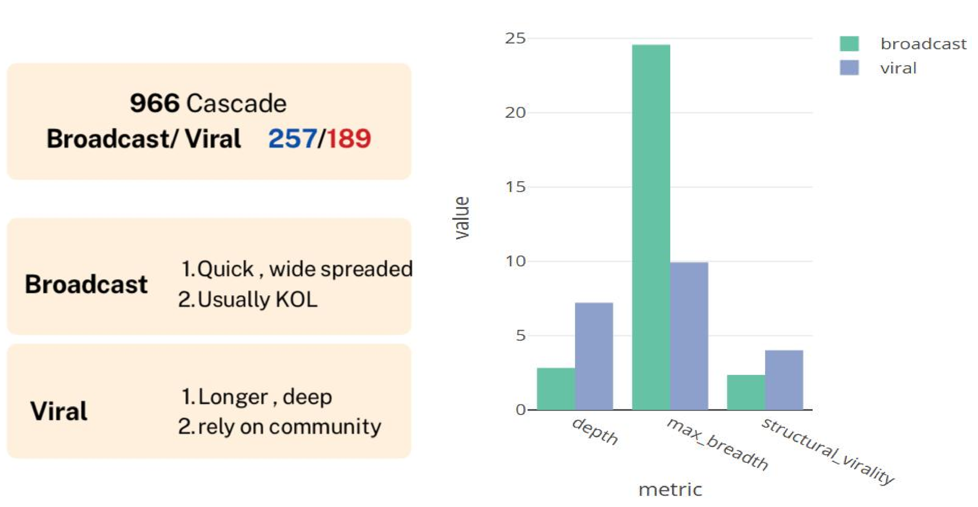


Figure 2. Proportion of different Cascades in the given data

* We excluded graphs with size < 10

4.2. User Clustering Based on Persona Share

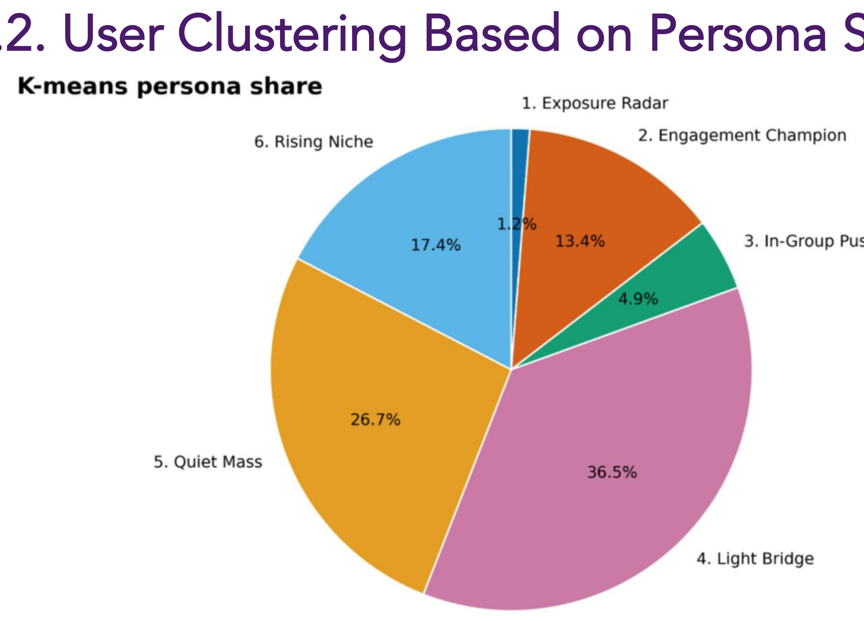


Figure 3. Proportion of Different User Categories

The six personas fit different social media campaign strategies:

- Best for immediate amplification: Cluster 1 (Exposure Radar)
- Best for breakout proxies: Cluster 4 (Bridges)
- Best for deep cascade tail: Cluster 1 (Exposure Radar)

The hints we got from this pie chart:

- The users are not completely homogeneous. Instead, they have different potentials in spreading information.
- Further analyses of the diffusion patern and user's position in the diffusion (i.e. level of depth in the porpagation tree) were required.

4.3. User Clustering Based on Scale Profiles

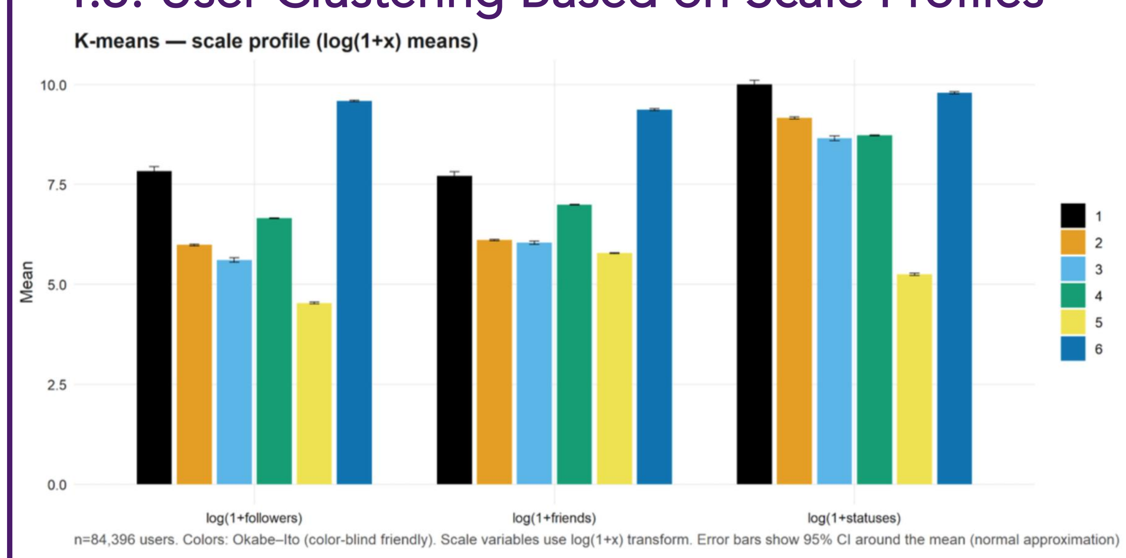


Figure 4. Average Levels of Scale-based Matrics of the 6 Personas

Inspiration from the graph

- User clusters differ not only in diffusion mechanism, but also in user scale and activity
- The differences across personas are driven by…

(1) Account size: follower count, connection degree, historical audience exposure

(2) User activity: engagement frequency

5. Macro-Analysis: Conditional Effect Plots

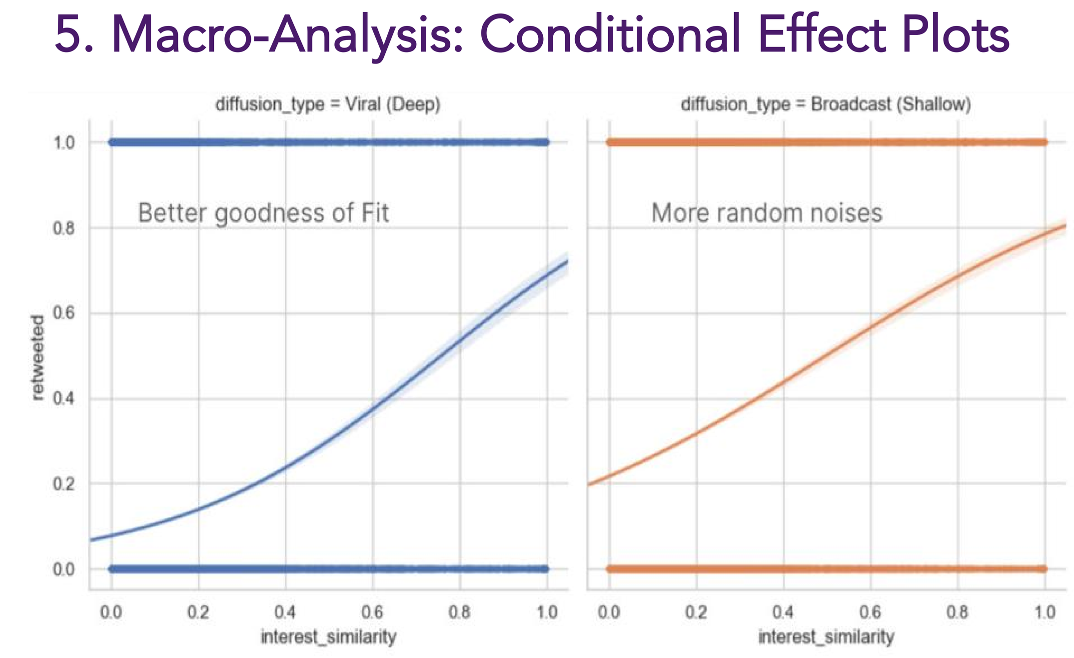


Figure 5. Conditional Effect Of Interest Similarity

We applied binary logistic regression to estimate how interest similarity affects retweet probability within the two patterns respectivelt. The plots show that the long-term effect of increased interest similarity on retweeting probability is stronger in viral than in broadcast cascades at later diffusion stages. Thus, we came up with a new question: Is there a threshold value acting as a boundary between low and high probability of retweeting?

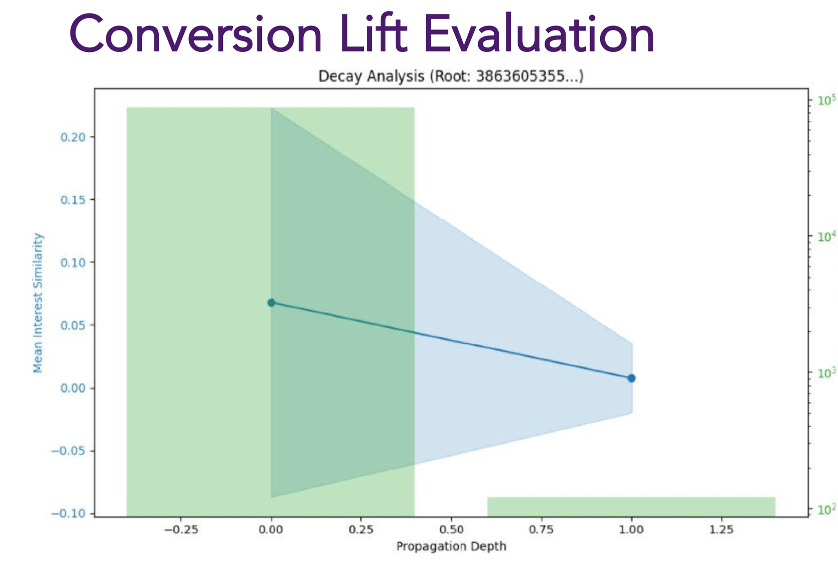
7. Limitations and Reflections

Our data allow us to explain mechanisms behind deep versus shallow diffusion, but they are not sufficient to predict exactly which specific cascade will go viral.

The dataset did not directly answer "what is the exact best timing to promote". Instead of obtaining an optical time value, we proved that longer time spans and repeated exposure favour deep diffusion.

Future work could integrate richer temporal features, content signals and other interventions to improve viral prediction and timing recommendations.

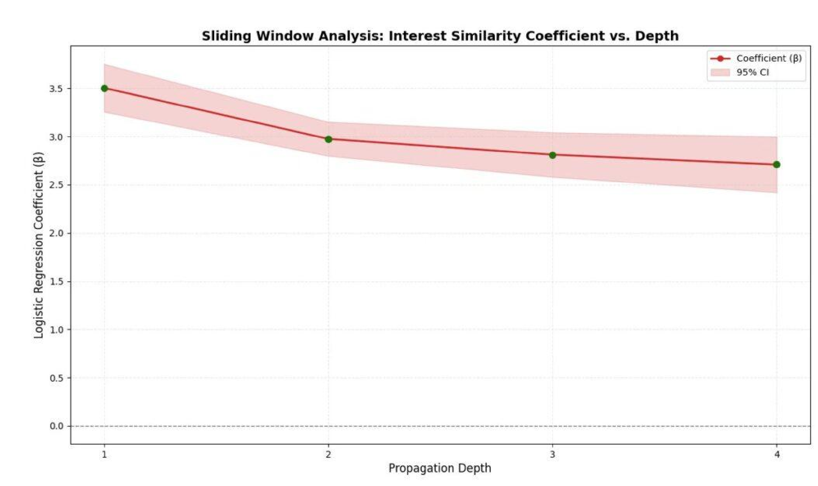
6. Micro-Analysis, Sliding Window Analysis & Conversion Lift Evaluation



	Average interest similarity	What drives the retweet?
Depth (0-1)	0.07	Homogeneity
Depth (1-2)	0.005	(Potentially) Content of the post

Figure 6 & Table 1. Decay Of Average Interest Similarity Across Depth

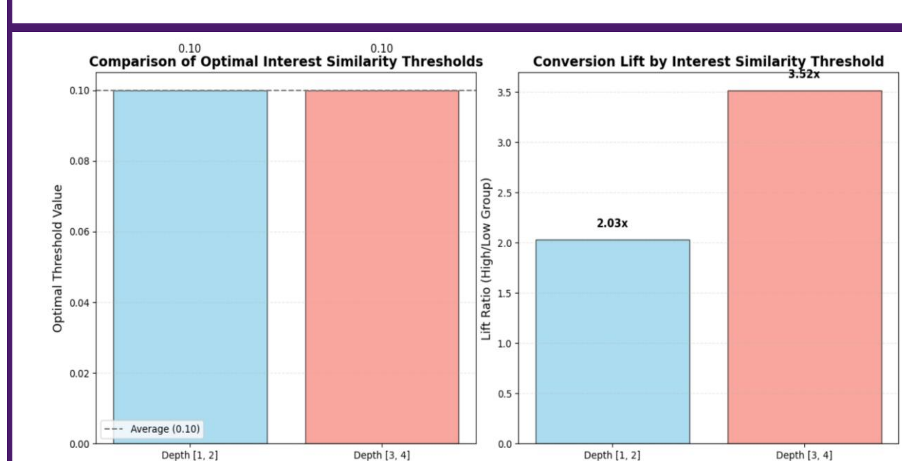
We first analysed the change in average interest similarity by depth. The downward-sloping curve and the reduced blue area indicate a decline in retweeters. This suggests the decreasing importance of user heterogeneity at greater depths.



	Features
Trend of the coefficient	Decreases by approximately 78.6%
Marginal impact of interest similarity on retweeting	Midly decreasing

Figure 7 & Table 2. Analysis of Regression Coefficient Of Interest Similarity by depth

We then applied a sliding window analysis across different ranges of depths (1–2, 2–3, 3–4). All coefficients of interest similarity are positive, confirming that similarity always helps. Yet the coefficients decreased with depth, indicating that the marginal impact of interest matching weakens in deeper layers, where user's interest in the content itself might have become the main driver.



	Optimal similarity threshold	Conversion lift above the threshold
Depth: 1-2	0.1	2.03 times higher than before
Depth: 3-4	0.1	3.52 times higher than before

- **Primary** diffusion: social relationship > interest similarity
- **Secondary** diffusion: interest-driven diffusion becomes scarce

Figure 8 & Table 3. Evaluating the Conversion Lift Above the Threshold at two stages: Priamry Diffusion (1-2) and Secondary Diffusion (3-4)

Despite discovering similar thresholds between the primary and secondary stage of diffusion, we noticed that secondary diffusion exhibits a higher conversion lift above the threshold. A soceital explanation would be: when users have moved across the interest similarity threshold, postings from neighbouring social relationships appear more frequently. Thus, the information they receive tend to be more trustworthy. The interest-community-based nature and scarcity of interest similarity makes the conversion from willingness to the action of retweeting more possible.

Key Findings

① User profiling is important for precision marketing.

- User activity, user scale, user persona, and diffusion metructure all matter.

② The effect of interest similarity between users and social relationships on retweeting is dynamic

- At smaller depths: Interest similartity determines the willingness to retweet
- At larger depths (especially above the threshold): Interest similarity determines how likely those users with the willingness to retweet will actually do so.

Theoretical Conclusions

- Distinguishes viral and broadcast diffusion by linking cascade shape to micro-mechanisms of interest similarity, exposure, and structure.
- Bridges threshold contagion and complex contagion theories with large-scale platform data, connecting Granovetter / Centola / Ugander frameworks to real Twitter-like cascades.
- Reveals a phase-based propagation pathway: a primary stage governed by social relationships and a secondary stage governed by individual interest.

Societal Effects

- Enables responsible virality control by identifying user types and structural positions that drive extremely deep cascades (Ugander et al., 2012).
- Improves targeting for public-interest communication — selecting better seed users for health, crisis and civic campaigns (Centola, 2010).
- Enhances user experience by using similarity and exposure thresholds to reduce spammy over-exposure and irrelevant pushes.

Bibliography

[1] Centola, D. (2010). The spread of behaviour in an online social network experiment.
[2] Goel, S., et al. (2016). The structural virality of online diffusion.
[3] Granovetter, M. (1978). Threshold models of collective behaviour.
[4] Ugander, J., et al. (2012). Structural diversity in social contagion.

Acknowledgement

This project was conducted as part of the Data Analytics Practice Opportunity at The Chinese University of Hong Kong. We thank our supervisors and CUHK Library for their support. We greatly appreciate the author of the source paper (https://doi.org/10.1016/j.tele.2021.101623), Professor Hai LIANG, for giving us the access to the insightful research results.

If you would like to reuse the data, please contact at data@cuhk.edu.hk.

More About The Project

For details, please visit our project website: https://dsprojects.lib.cuhk.edu.hk/en/projects/information-diffusion/introduction/

Scan the QR code on the top right corner to learn more about the source paper and our project outcomes.