Group Project: Influence vs Homophily in Reddit

Presented to
Professor Taha Havakhor
TA Jitsama Tanlamai

By
Clarke, Sean - 260706014
DeSilva, Dhevin - 261177497
Delisle, Audrey - 261142504
Barabasz, Michelle - 261152119
Park, Seunghyun- 260686853

INSY 670 - Section 075

McGill University - Desautels Faculty of Management
Thursday April 25th 2024

Table of Contents

Introduction

Reddit recently became a public company, a significant transition that places new emphasis on understanding and optimizing user engagement across its many communities. This project specifically delves into a subreddit of “dogs” as a case study to solve a critical question: do users participate more because of shared interests or due to the influence of their social connections? Addressing this question is important because it helps Reddit better align its content and advertising strategies to the underlying motivations of its users, hence enhancing engagement and commercial opportunities. In this project we aim to develop predictive models, and craft a tool for continuous assessment to provide actionable insights that can guide strategic decisions for Reddit executives.

Data Collection & Analysis

For this project, we extracted submissions and comments from the "DOG" subreddit, as well as from related subreddits such as "puppies," "PuppySmiles," and "Pets." This dataset was sourced from https://the-eye.eu/redarcs/, a website that archives various digital data, including reddit content. Each subreddit has submissions files and comments files. Once the data was collected, we used data cleaning steps such as removing duplicate entries and filtering out irrelevant data. Additionally, we performed standardization of timestamps and user identifiers to facilitate analysis across different datasets. Lastly, our sample size was of 2,094 Reddit users (all users active in 2017 DOG subreddit).

To determine if users A and B have ties, we focused on the 'author' and 'parent_id' columns in the comments dataset. Each comment in the dataset lists the 'author' who made the comment and the 'parent_id', which points to either the original post or another comment to which the comment is replying. By matching the 'author' of a comment with the 'author' of the parent post, we can identify direct interactions or replies between users, indicating a tie. This gave us 190 rows. This makes sense because the formula to calculate the number of inter-user interactions is n*(n-1)/2 = 2,094*2.093/2 = 2,191,371.

Shared interests were then identified by if a user either posted content or commented in any of the the three related subreddits. This shows us that they either do have a general interest for dogs or do not. For each user, we compiled a feature dataframe identifying if they are interested in dogs or not. However, this only gave us a matrix of 20 rows. In order to evaluate against the ties dataframe, we had to make this 190 rows. Hence, we multiplied this dataframe on itself, giving us all combinations of two users and both user interests. Then if both users had 1,1 or 0,0 (meaning they were both present or not present in related subreddits), a 1 was associated with the pair. This meant they had the same interest. If the user pair had 1,0 or 0,1, a 0 was associated for differentiating interests.

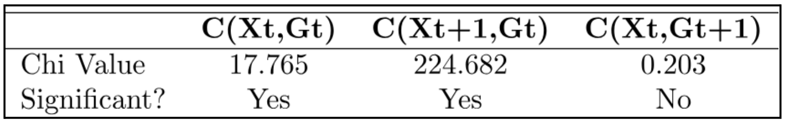
For both ties and shared interests, we segmented the data into two time periods, Time t and Time t+1, to assess changes and developments over time. This ended up being 2017 and 2019. This was done by using the 'created_utc' timestamp into the dataset. We then analyzed the ties and shared interests separately for each time period to track how relationships and interests either initiated, persisted, or dissolved from Time t to Time t+1. At the end, we had 4 dataframes of 2,191,371 rows; 2 for each year (2017, 2019) and two for each topic (ties, features).

Detecting Homophily and Influence

To determine whether the "DOG" subreddit was driven by homophily (like attracts like) or influence (users are swayed by others), we constructed 3 contingency tables for ties and shared interests at two separate time points, time t and time t+1. Each contingency table mapped the presence or absence of ties against shared interests within the community at each time point. From these 3 tables, we calculated three Chi-squared values: C(Xt,Gt) representing the relationship between ties and interests at time t, C(Xt,Gt+1) for ties at time t+1 against interests at time t, and C(Xt,Gt+1) for ties at time t against interests at time t+1.

These Chi-squared tests, with degrees of freedom set at 1 due to the analysis involving two characteristics (ties and interests), helped us examine the significance of the relationships between these variables over time. We then wanted to check if the values were significant, so we mapped the values to a Chi-squared table. In conclusion, two out of the three relationships analyzed over time are statistically significant at the 5% significance level. The results are shown in table 1 below. We then concluded that based on the chi-squared tests, the DOG subreddit shows evidence of influence rather than homophily. We know this because $C(Xt+1,Gt) > C(Xt,Gt)$ or 224.68 > 17.76. This indicates that changes in the network ties from one year to the next are more reflective of the influence users have on each other, rather than users forming ties based on pre-existing shared interests. That being said, since the final Chi squared test is not significant, that is a caveat to take into consideration. With more time, we could have tested this on different subreddits to see if they were significant.

Table 1 : Chi Squared Values and significance at 0.05%



	C(Xt,Gt)	**C(Xt+1,Gt)**	**C(Xt,Gt+1)**
Chi Value	17.765	224.682	0.203
Significant?	Yes	Yes	No

Business Implications

This project holds significant business implications for Reddit, particularly as it transitions into a publicly traded company seeking to maximize shareholder value and operational efficiency. By distinguishing between homophily and influence within various subreddit communities, Reddit can refine its strategies for user engagement, content curation, and advertising.

For subreddits dominated by homophily, where users gather based on shared interests, Reddit can enhance community bonding through personalized content recommendations and targeted community events that reinforce these common interests. This approach not only improves user satisfaction and retention but also makes these communities more attractive for niche advertisers.

On the other hand, in subreddits where influence is more pronounced, like our DOG example, Reddit has the opportunity to leverage influential users or posts to drive discussions and engagement. Marketing strategies in these areas could involve promoting popular content or users more prominently, increasing the visibility of sponsored posts or advertisements that align with the influencers' activities. Additionally, understanding these dynamics helps Reddit to better manage the spread of information, helping to mitigate risks associated with misinformation in influence-driven communities. Furthermore, this insight into community dynamics is important for advertisers on Reddit who can use this data to tailor their advertising strategies, ensuring that their messages are better communicated to the right audiences.

Future State Pipeline

A future direction would be to create an automated pipeline to assess homophily or influence within a subreddit. The system would extract related topics using natural language processing (NLP) and preprocess the data for analysis. Although direct data scraping from Reddit is not feasible, the envisioned process would clean the data and compute chi-squared values to discern whether homophily or influence predominates in a subreddit. This critical insight allows Reddit's business teams to make informed decisions on how to tailor content, engage users, and manage the community.