INSY 670 - Social Media Analytics:

INFLUENCE VS HOMOPHILY IN REDDIT

By Audrey Delisle, Michelle Barabasz, Sean Clarke, Dhevin DeSilva & Seunghyun Park

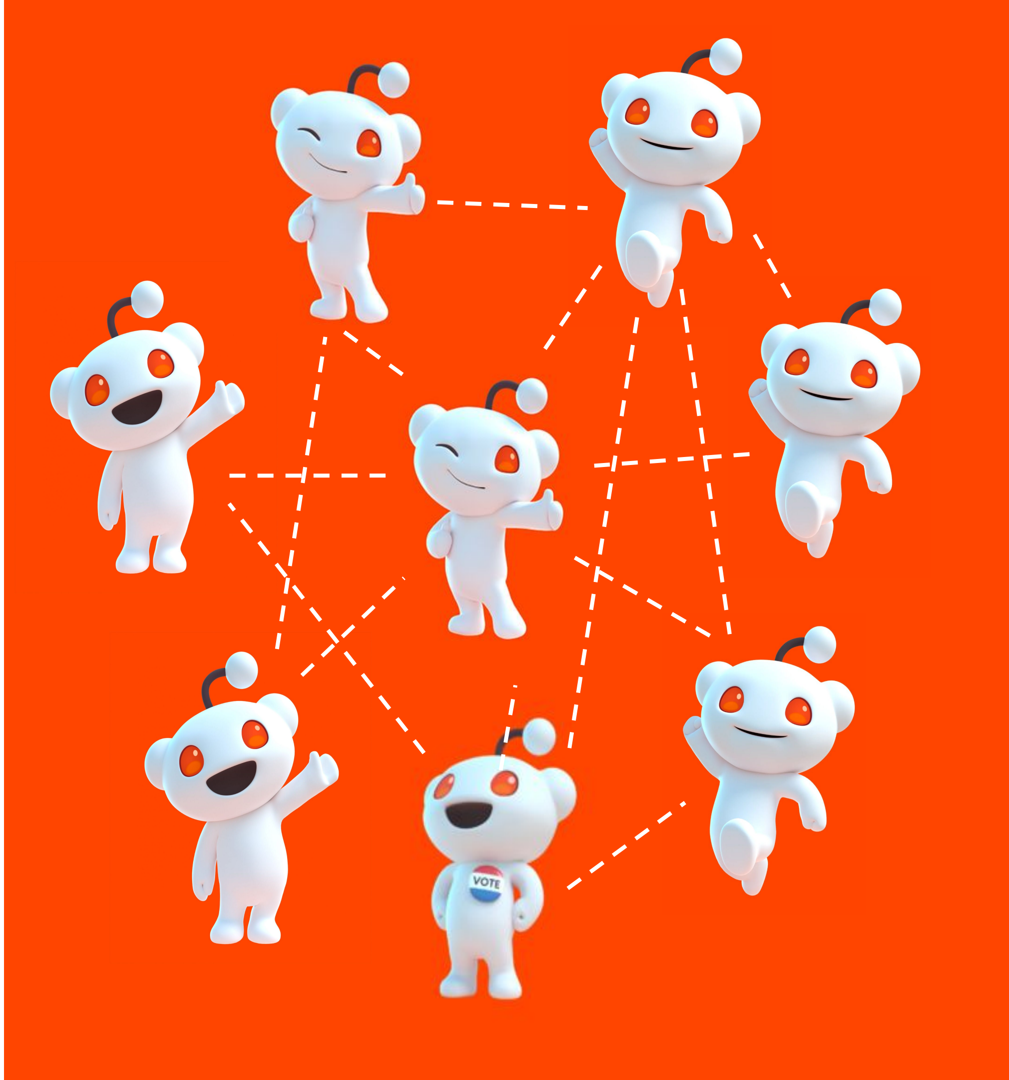


TABLE OF CONTENTS

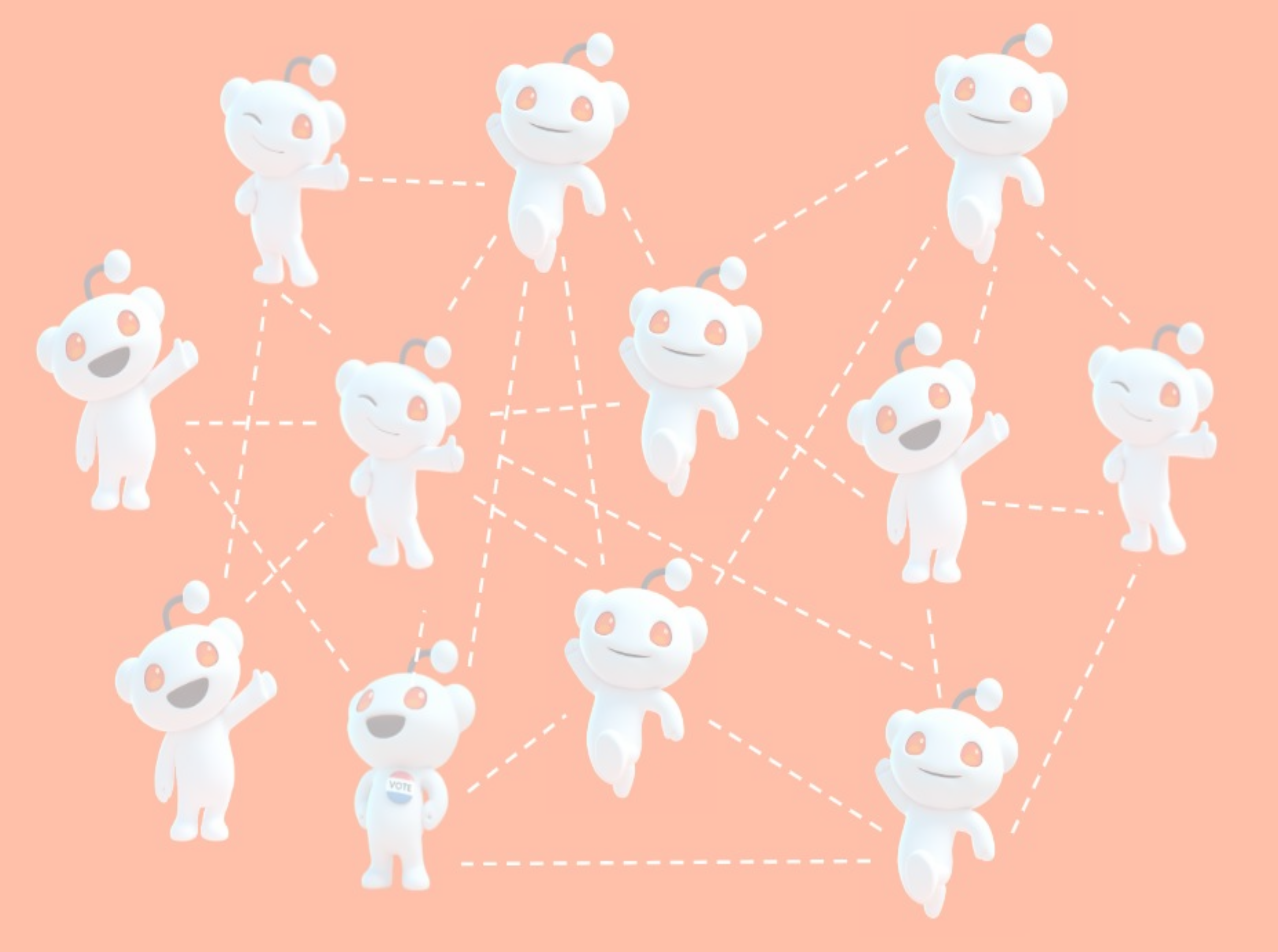
01 BUSINESS SCENARIO

02 DATA COLLECTION

03 DETECTING HOMOPHILY AND INFLUENCE

04 RESULTS & ANALYSIS

05 BUSINESS IMPLICATIONS



01. BUSINESS SCENARIO

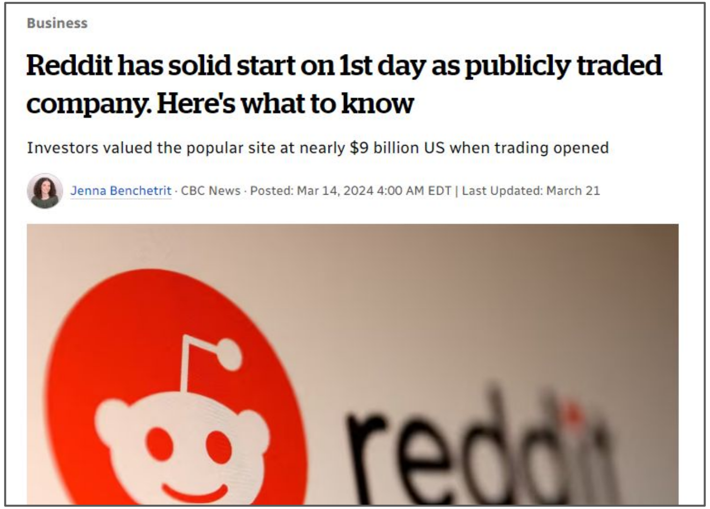
Business Scenario

- In March of 2024, the social media platform known as Reddit went public

- Understanding influence (your friends take up your interests) vs homophily (you become friends with people having the same interests) becomes crucial

- Meeting investor expectations, advertising targeting, content and community management, risk mitigation, etc.

- Allows for monetization and management of the platform



Business

Reddit has solid start on 1st day as publicly traded company. Here's what to know

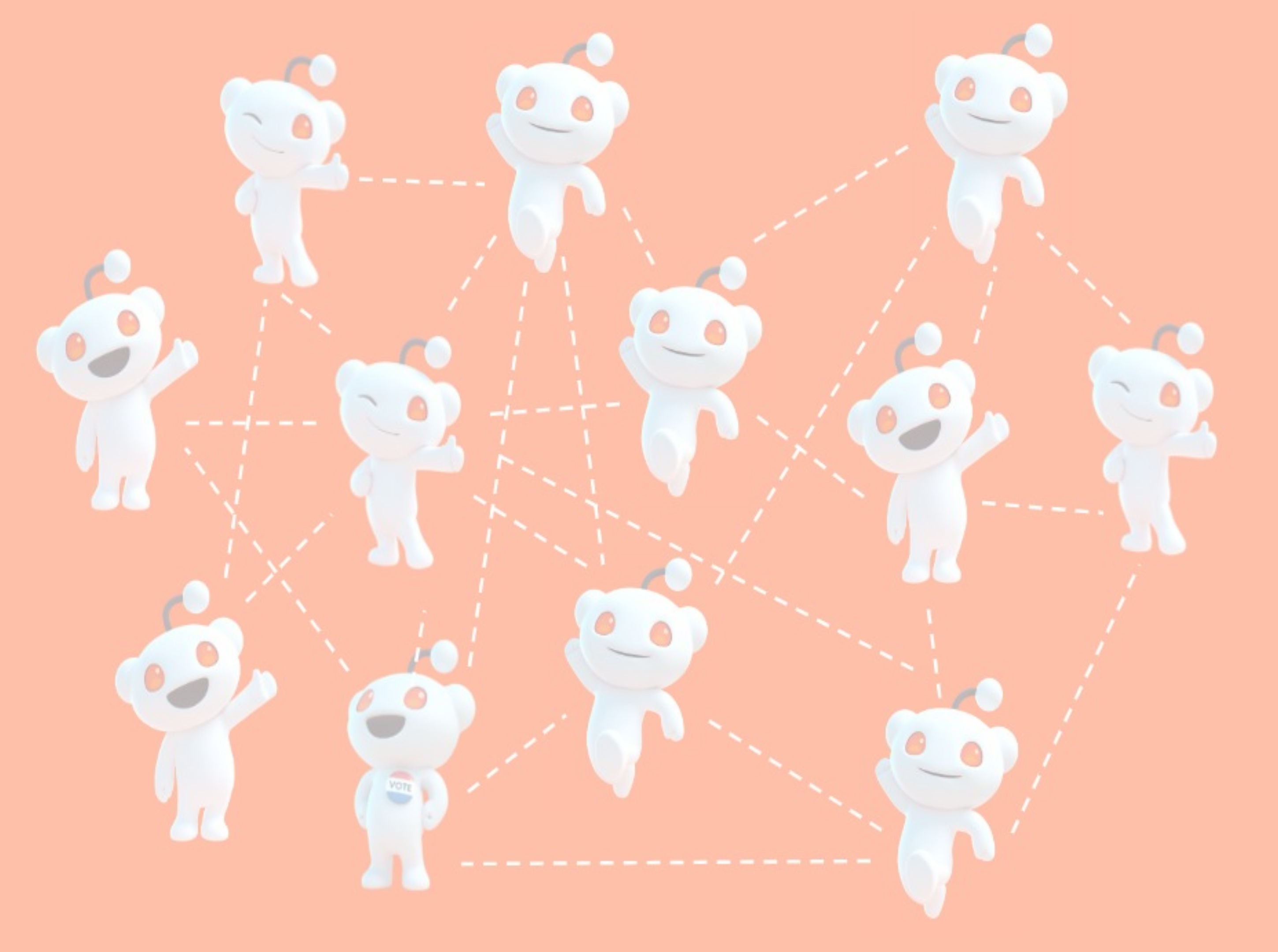
Investors valued the popular site at nearly $9 billion US when trading opened

Jenna Benchetrit · CBC News · Posted: Mar 14, 2024 4:00 AM EDT | Last Updated: March 21

Project Goals

The project aims to investigate the **nature of user interactions** within Reddit's diverse communities, focusing on whether these interactions are driven more by **homophily or influence**. This understanding will be pivotal for Reddit's business strategy and operational focus post-IPO. For the sake of the project, we will use subreddits pertaining to dogs.

1. **Quantitative Analysis of User Behavior:** Analyze interaction data (comments, posts, upvotes) and presence in other subreddits to determine whether users are more likely to engage due to shared interests or influential posts/users.
2. **Develop Predictive Models:** Use Chi-squared tests to predict whether a new subreddit will likely develop as a homophily-driven or influence-driven community.
3. **Develop a Pipeline:** Develop a scalable and reusable tool that can be used by Reddit administrators and marketers to continuously analyze and monitor the social dynamics of any subreddit.



02. DATA COLLECTION

Data Source

- Reddit data was downloaded from: https://the-eye.eu/redarcs/
- Main subreddit we are analyzing is: “DOG” (comments.zst and submissions.zst)
- In order to see if people have same interests, we also look at related subreddits: “puppies”, “Pets”, “PuppySmiles”

Data Cleaning

Loading Data:
The data is loaded from compressed JSON files. Various datasets such as `DOG_comments`, `DOG_submissions`, and others from different categories are ingested using a custom function to decompress and read the files into pandas DataFrames.

Initial Preprocessing:
   - Removing unnecessary or irrelevant columns.
   - Handling missing values by either filling them with default values or removing the rows/columns.
   - Normalizing text data, possibly involving transforming all text to a consistent case (e.g., lowercasing), removing special characters, or other text normalization techniques.
- Sample size: 2,094 users

Setting up Data Frames

- Tie data frame and feature data frame were set up
- Tie dataframe associated each user to each other; giving a 1 if there were any comments between both in the DOG subreddit, and a 0 if there were no comments

- Tie data frame was 2,191,371 rows
- Feature dataframe gave each user a 1 if they were active in related subreddits and 0 if not (post or comment in PuppySmiles, Pets, or Puppies)
- Feature dataframe was only 2,094 rows
- Feature dataframe was multiplied by itself to get a combination of features for each pair of users
- If both users had 1,1 or 0,0, a 1 was associated = same interest (either they both like or dislike dogs)
- If both users had 1,0 or 0,1 a 0 was associated = different interests (either one likes or doesn't like dogs)
- This gave a new dataframe of 2,191,371 rows

- We finished with 4 dataframes: one for each year (2017, 2019) and one for each topic (ties and features)

Setting up Data Frames

Tie Data Frame

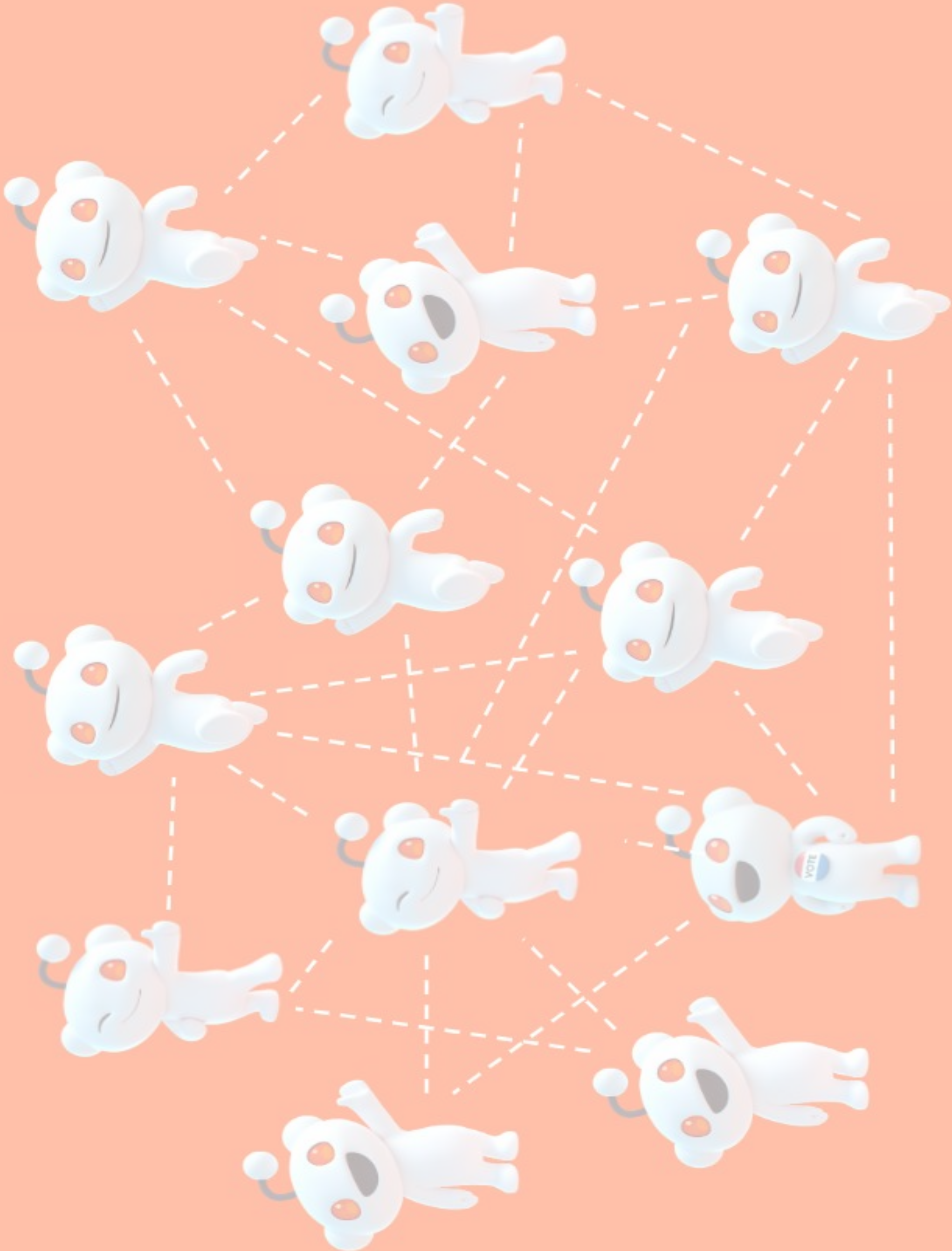
Parent user	Child user	Comment?
a	b	1
a	c	0
a	d	0
a	e	1
...		
2,191,371 rows		

Feature Data Frame

User node	Active in related subreddits?
a	1
b	0
c	1
d	1
...	
2,094 rows	

User node x	User node y	X active in related subreddits?	Y active in related subreddits?	Shared common interest?
a	b	1	1	1
c	d	0	0	1
a	c	1	0	0
d	c	0	1	0
...				
2,191,371 rows				

03. INFLUENCE VS HOMOPHILY



Contingency Table at time Xt,Gt

X = 2017

G = 2017

	a & b have same interest	a & b don’t have same interest	
a & b have ties	489	189	678
a & b don’t have ties	1,725,109	465,584	2,190,693
	1,725,598	465,773	2,191,371

n (total users analyzed) = 2,094

Total pairs (N) = n x (n-1) / 2

= 2,094 x (2,093) /2

= 2,191,371

Contingency Table at time Xt,Gt+1

X = 2017

G = 2019

	a & b have same interest	a & b don't have same interest	
a & b have ties	6	1	7
a & b don't have ties	1,725,592	465,772	2,191,364
	1,725,598	465,773	2,191,371

n (total users analyzed) = 2,094

Total pairs (N) = n x (n-1) / 2
= 2,094 x (2,093) /2
= 2,191,371

Contingency Table at time Xt+1,Gt

X = 2019

G = 2017

	a & b have same interest	a & b don’t have same interest	
a & b have ties	583	95	678
a & b don’t have ties	2,114,679	76,014	2,190,693
	2,115,262	76,109	2,191,371

n (total users analyzed) = 2,094

Total pairs (N) = n x (n-1) / 2
= 2,094 x (2,093) /2
= 2,191,371

Calculating Chi-Square Values

1

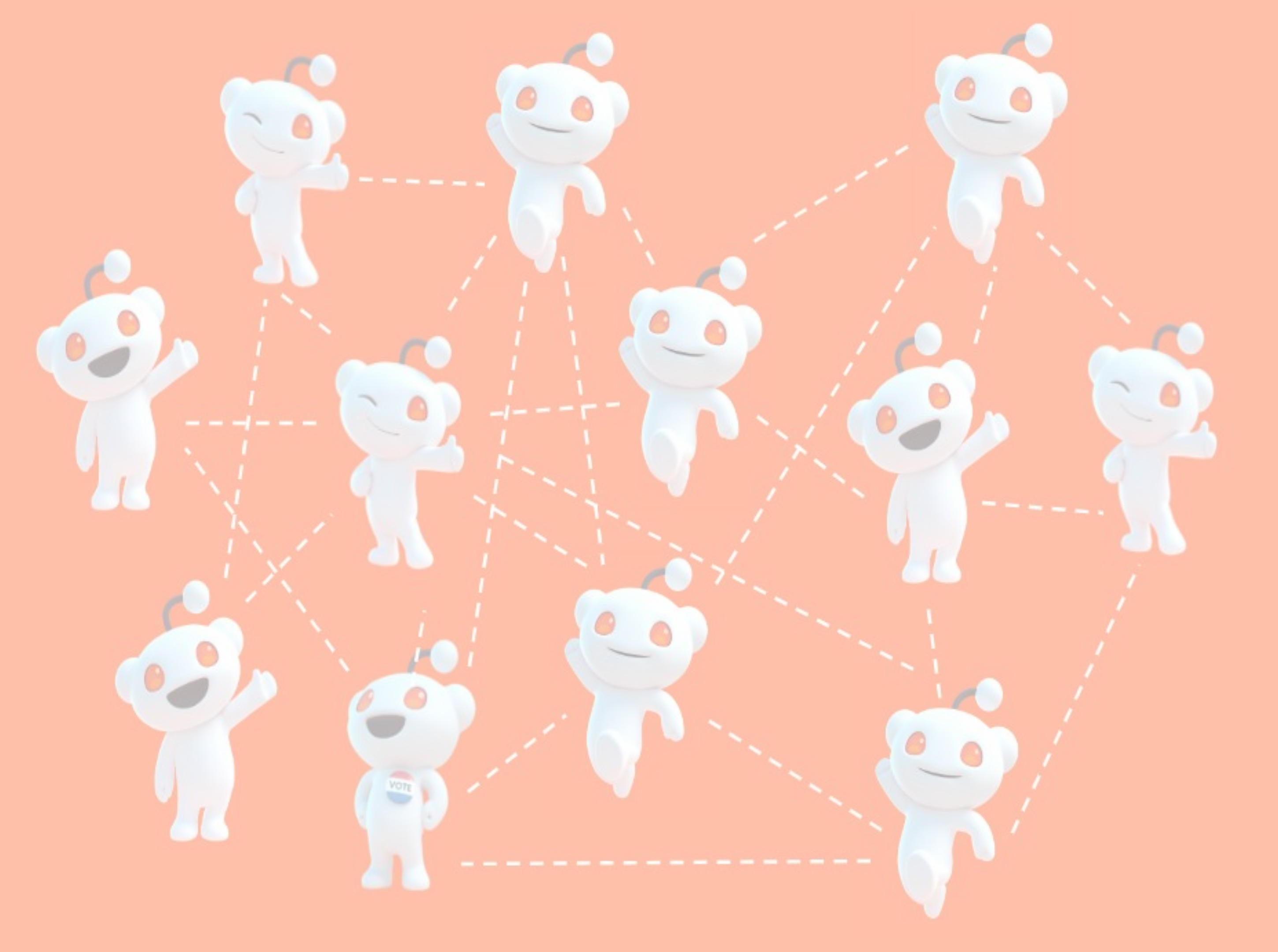
C(X_t,G_t) = x^2 = Chi Squared = N (ad - bc)^2 / [(a+b) (c+d) (a+c) (b+d)]
= 2191371 (489*465584 - 189*1725109)^2 / [(489+189) (1725109+465584) (489+1725109) (189+465584)]= 17.765

2

C(X_t,G_{t+1}) = x^2 = Chi Squared = N (ad - bc)^2 / [(a+b) (c+d) (a+c) (b+d)]
= 2191371 (6*465772- 1*1725592)^2 / [(6+1) (1725592+465772) (6+1725592) (1+465772)] = 0.203

3

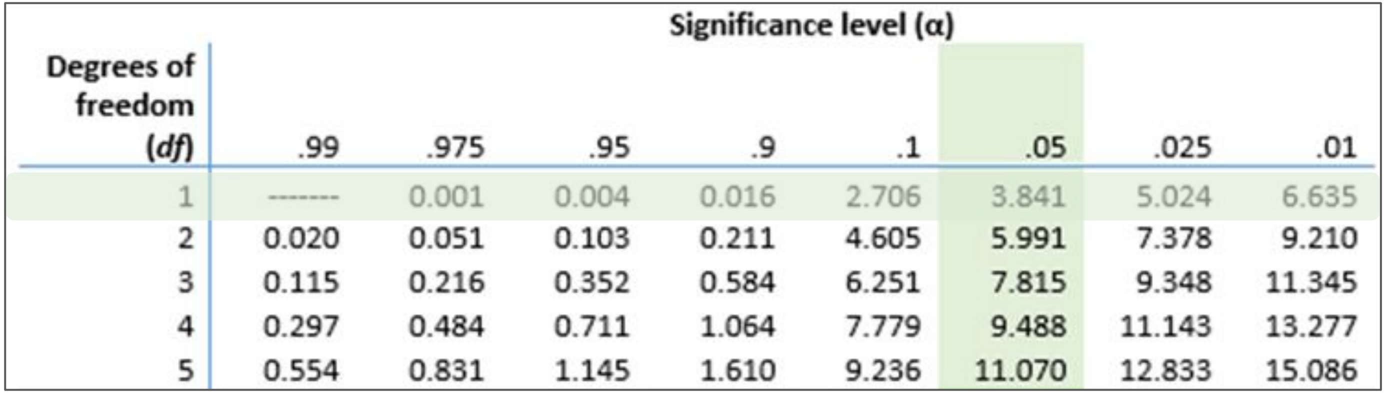
C(X_{t+1},G_t) = x^2 = Chi Squared = N (ad - bc)^2 / [(a+b) (c+d) (a+c) (b+d)]
=2191371 (583*76014 - 95*2114679)^2 / [(583+95) (2114679+76014) (583+2114679) (95+76014)] = 224.682



04. RESULTS & ANALYSIS

Significance of Chi-Square Values

df = number of attributes - 1 = 2-1 =1



Degrees of freedom (*df*)	Significance level (α)							
	.99	.975	.95	.9	.1	.05	.025	.01
1	-------	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086

	$C(X_t, G_t)$	$C(X_{t+1}, G_t)$	$C(X_t, G_{t+1})$
Value	17.765	224.682	0.203
Significant?	Yes	Yes	No

Homophily or Influence?

Homophily

$$C\,(X_t, G_{t+1}) > C(X_t, G_t)$$

If homophily effect is present, autocorrelation will increase when we consider link changes from t to t + 1

0.203 < 17.765

Influence

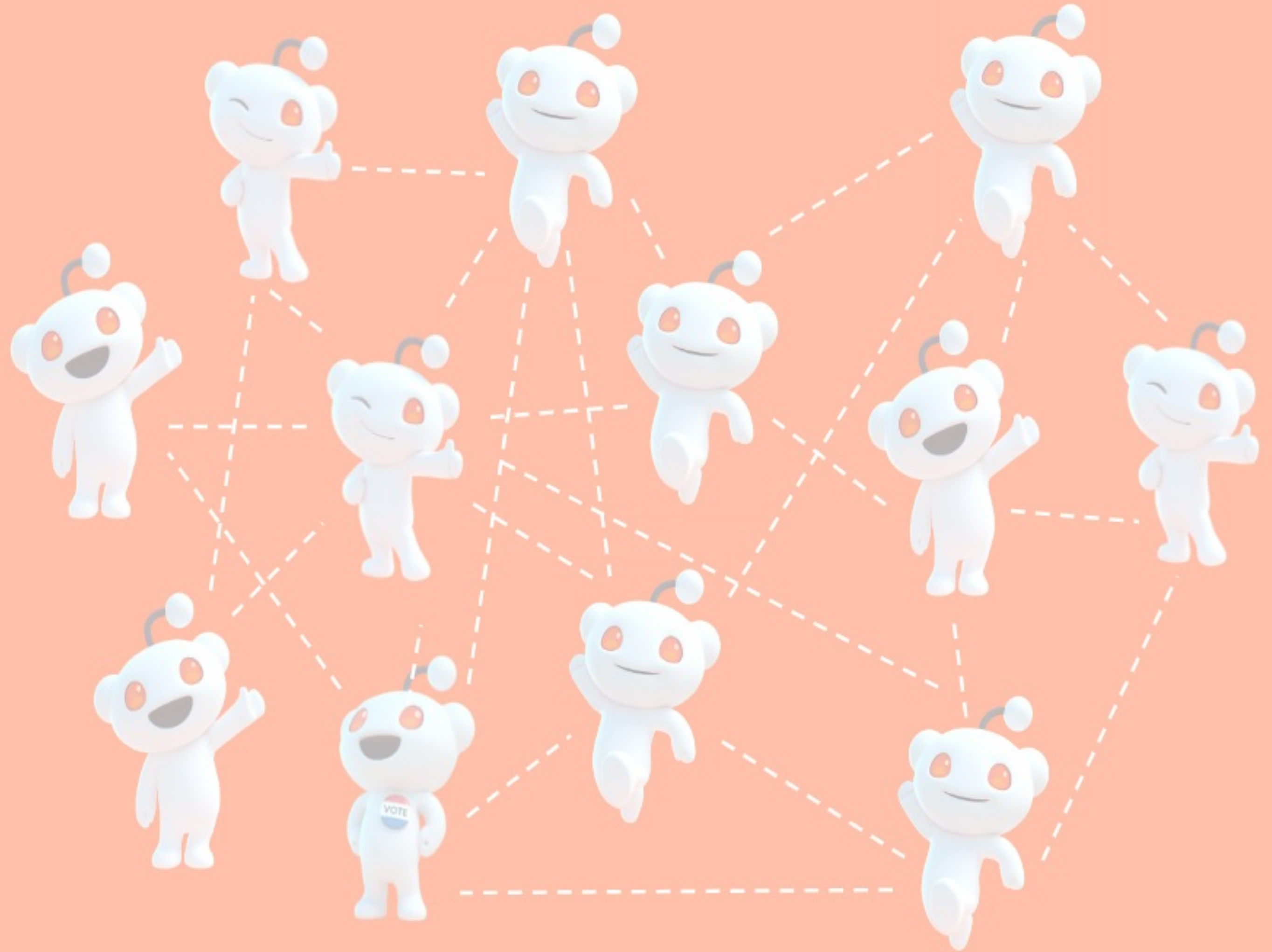
$$C\,(X_{t+1}, G_t) > C(X_t, G_t)$$

If influence effect is present, autocorrelation will increase when we consider attribute changes from t to t + 1

224.682 > 17.765

Conclusion : DOG subreddit shows Influence

05. BUSINESS IMPLICATIONS



Business Implications

Homophily-driven Subreddits:

- **Personalized Content Recommendations:** Implement algorithms to suggest personalized content based on users' interests and preferences within niche communities.
- **Targeted Community Events:** Organize virtual or physical events tailored to specific interests within communities.
- **Community Bonding Initiatives:** Encourage user-generated content competitions or challenges related to common interests.
- **Niche Advertising Opportunities:** Offer targeted advertising options for niche products or services relevant to the interests of specific subreddits.

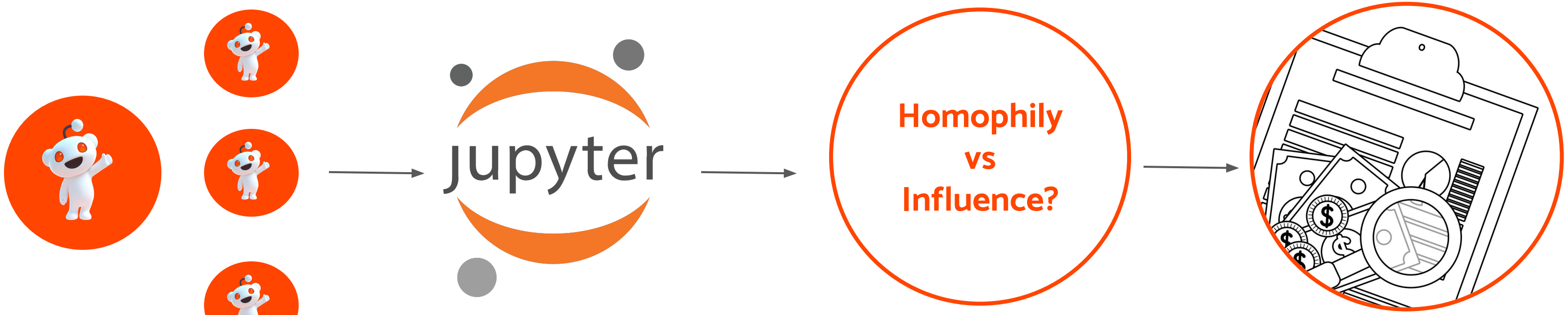
Influence-driven Subreddits:

- **Promotion of Popular Content and Users:** Highlight trending posts or influential users within the subreddit to increase engagement and visibility.
- **Engagement Campaigns:** Launch engagement campaigns around influential user-generated content to encourage broader participation and discussion. Partner with influencers to create exclusive content or events that drive user engagement and interaction.
- **Targeted Marketing Strategies:** Offer targeted advertising packages that leverage the influence of specific users or posts to reach relevant audiences.
- **Information Management and Moderation:** Develop tools and algorithms to identify and mitigate the spread of misinformation or harmful content within influence-driven communities.

Future Directions

- **Leveraging NLP for Subreddit Discovery:** Implement an NLP-driven system to analyze subreddit titles and recommend three related subreddit topics, enhancing user engagement and content discoverability.
- **Multimodal Data Analysis:** Incorporate other types of data, such as images, videos, or user-generated content, to gain a more comprehensive understanding of user behavior and community engagement. This could provide deeper insights into the factors driving user interactions and content preferences.
- **Temporal Analysis:** Extend the analysis to understand how user behavior and community dynamics evolve over time. This could involve tracking changes in user interactions, identifying emerging trends, and predicting future shifts in subreddit dynamics.

Future State: Pipeline Network



Homophily vs Influence?

Step 1: Reddit picks a subreddit topic and it gets scraped. We do NLP on the subreddit title to find 3 other related subreddits.

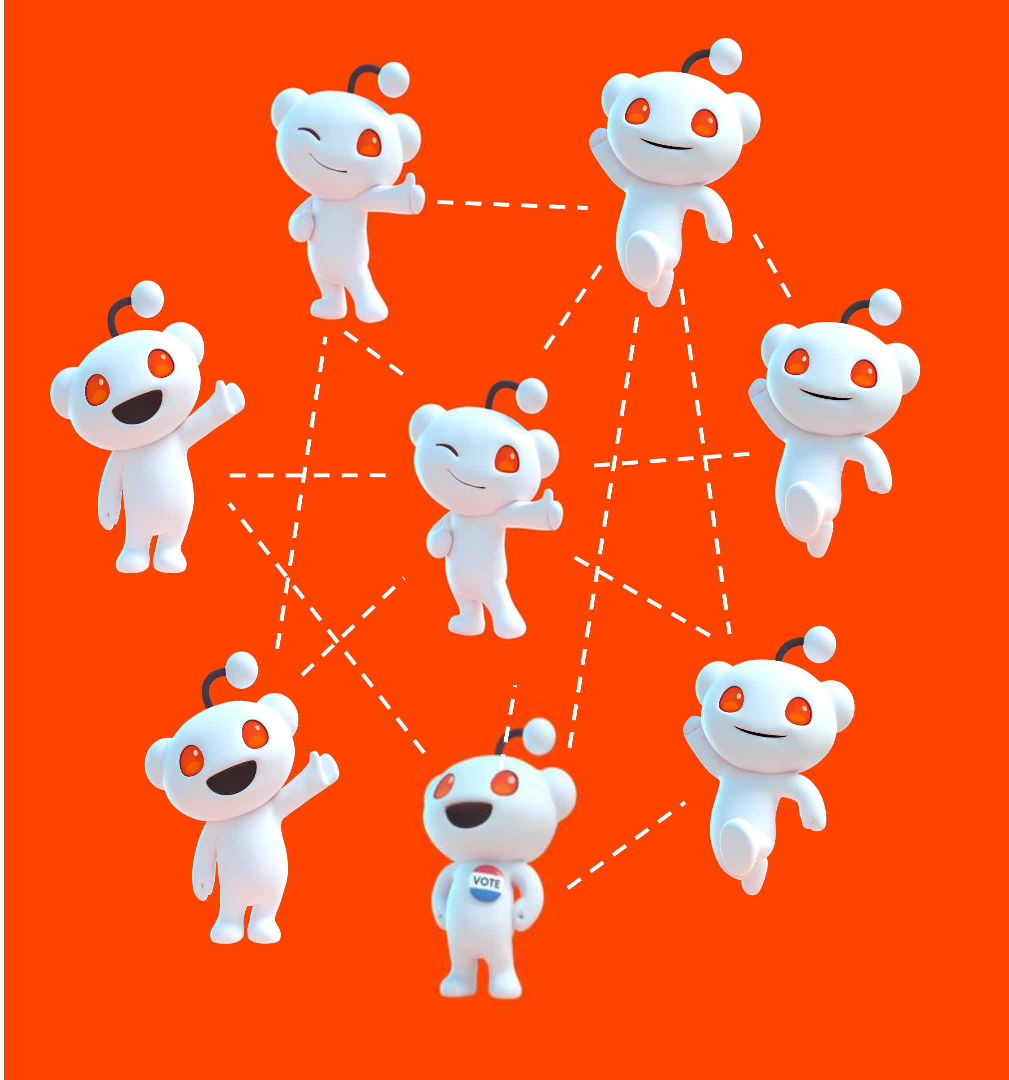
Step 2: Subreddit submissions and comments get parsed in our code, extracting X^2 values at t and t + 1 (+1 year).

Step 3: Our code outputs whether this subreddit is a network of homophily or influence.

Step 4: Reddit takes this information and structures their advertising plan to enhance monetization.

THANK YOU

If you have any questions, feel free to ask



RESSOURCES

- https://www.cbc.ca/news/business/reddit-ipo-explainer-1.7140286
- https://www.scribbr.com/statistics/chi-square-distribution-table/
- https://www.ling.upenn.edu/~clight/chisquared.htm