Project ID: P7
Project Title: Signed Network – Reddit Hyperlink Network
Group ID: 10 (Hyderabad)
Team Members with ID numbers: ARITRA KUMAR DUTTA (2022H1030096H)
MOHAMMAD AVESH HUSAIN (2022H1030090H)
PIYUSH PRIYADARSHI (2022H1030091H)
UTSAV SETH (2022H1030074H)

Introduction

In this milestone, our focus shifts from a general dataset introduction to a specific problem statement rooted in predicting interactions between subreddits using a sentiment-based model. Our goal is to harness the power of attachment index to forecast the dynamics between subreddits, where each link or edge signifies a post originating from the source subreddit directed towards the target subreddit. The dataset's rich information on subreddit interactions serves as the foundation for our exploration into predicting these connections.

Motivation

The motivation behind our work lies in unraveling the intricate relationships that exist within the vast Reddit community. Understanding how sentiments and discussions propagate from one subreddit to another can offer valuable insights into online community behavior. By applying predictive models to this social network data, we aim to uncover patterns and trends that can inform strategies for content moderation, community engagement, and even highlight emerging topics or controversies.

Related Work

Our approach draws inspiration from the field of network science and sentiment analysis. Existing literature in these domains provides a framework for understanding the dynamics of online communities and predicting interactions between entities. The Aging Preferential Attachment model, a variation of the Preferential Attachment principle, serves as a guiding principle in our predictive model. This model has been previously applied to capture evolving networks, making it a relevant foundation for our exploration into subreddit interactions. Additionally, studies on sentiment analysis in social networks contribute to our understanding of how sentiments influence the formation and strength of connections in online communities.

Problem Statement:

The problem at hand involves predicting the likelihood of a subreddit post mentioning another subreddit. Given the vast and dynamic nature of online discussions, understanding the interplay between subreddits within posts is essential for content creators, moderators, and platform administrators. The task is to develop a predictive model that, based on historical data, can identify the probability of a particular subreddit being mentioned in posts across various communities.

Mathematical Formulation of the Problem:

Let's define a directed graph $G=(V,E)$ where V represents the set of subreddits and E represents the edges between them. Each edge e_{ij} from subreddit i to subreddit j symbolizes a post from i to j.

Now, our goal is to predict the number of such edges between any pair of subreddits, represented as e_{ij}, based on the Aging Preferential Attachment model. Mathematically, this can be expressed as:

$$\text{PredictedEdges}(i,j)=Pi\times Pj\times(\text{InDegree}(i))\times(\text{InDegree}(j))$$

Where:

- Pi and Pj are the Aging Preferential Attachment probabilities for subreddits ii and jj respectively.
- InDegree(i) and InDegree(j) are the in-degrees of subreddits i and j respectively.
- Mathematically, if events A and B are independent, the probability of both events occurring is given by: P(A and B)=P(A)×P(B)

Our task is to model this relationship and predict the number of edges between subreddits for a given time, facilitating a deeper understanding of the dynamics of information flow in the Reddit community.

Where, Pi is given by:

$$P_i = \frac{k_i + \alpha}{\sum_k (k + \alpha)} \left(\frac{t_0}{t_i + t}\right)^{\beta}$$

Where,

- ki is the degree of node i, representing the number of edges connected to node i.
- α is a parameter for preferential attachment, typically a small positive constant.
- $\sum_k(k+\alpha)$ is the sum of the degrees of all nodes in the network, including node i.
- t0 is a reference time, here we have taken t0 to be 1.
- ti is the timestamp when node i was added to the network.
- t is the current time.

Hypothesis:

Aging Preferential Attachment Model Hypothesis:

We hypothesize that the probability of a new edge forming between two subreddits is influenced by the aging preferential attachment model. Mathematically, we propose the following formula for predicting the probability of a new edge between subreddits i and j:

Predicted Probability(i,j) =

$$\frac{(\text{InDegree(i)}+\alpha)}{(\text{TotalDegree}+\alpha)} \text{ X } \left(\frac{t^0}{(FirstAppearance(i)-t)+t0}\right)^{\beta} \text{ X } \frac{(\text{InDegree(j)}+\alpha)}{(\text{TotalDegree}+\alpha)} \text{ X } \left(\frac{t^0}{(FirstAppearance(j)-t)+t0}\right)^{\beta}$$

- Where:
  - α and β are model parameters.
  - t0 is a reference time.
  - t is the current time.
  - InDegree(i) and InDegree(j) are the in-degrees of subreddits i and j respectively.
  - TotalDegree is the sum of in-degrees for all subreddits.
  - FirstAppearance(i) and FirstAppearance(j) are the first appearance days of subreddits i and j respectively.

- **Prediction Error Calculation Hypothesis:**

We hypothesize that the difference between the predicted number of edges and the actual number of edges for a given pair of subreddits can be effectively quantified by the absolute error. Mathematically, we propose the following formula for error calculation:

Absolute Error(i,j)=|Predicted Edges(i,j)−Actual Edges(i,j)|

This error will be calculated for each pair of subreddits, and the cumulative error will be used as an evaluation metric for the model's performance.

Approach:

1. **Data Preprocessing:**
   - Read the dataset containing information about subreddit interactions.
   - Extract relevant features such as source subreddit, target subreddit, timestamp, etc.
   - Calculate in-degrees and first appearance days for each subreddit.
2. **Aging Preferential Attachment Model:**
   - Implement the Aging Preferential Attachment model to calculate the predicted probability of a new edge between subreddits.
3. **Prediction and Evaluation:**
   - Predict the number of edges for each pair of subreddits using the calculated probabilities.
   - Compare the predicted edges with the actual number of edges.
   - Calculate the absolute error for each pair and sum the errors.
4. **Analysis and Reporting:**
   - Analyze the model's performance based on the error metric.
   - Generate insights into the Reddit community's information flow dynamics.

This approach aims to combine the Aging Preferential Attachment model with a comprehensive error analysis to evaluate the effectiveness of the model in predicting subreddit interactions.

Solution Approach:

The solution approach involves implementing the Aging Preferential Attachment model for predicting subreddit interactions. The key steps and their mathematical formulations are described below, along with mapping to the provided source code.

1. **Calculate Day Number:**
   - **Mathematical Formulation:** day_difference=current_timestamp−reference_date/24 hours
   - **Mapping to Code:** The `get_day_number` function calculates the day number from the timestamp using the provided format.
2. **Calculate Pi (Probability):**
   - **Mathematical Formulation:**

     $$\text{Pi}=\frac{(\text{InDegree(i)}+\alpha)}{(\text{TotalDegree}+\alpha)}\ \text{X}\ \left(\frac{t^0}{(FirstAppearance(i)-t)+t0}\right)^{\beta}$$

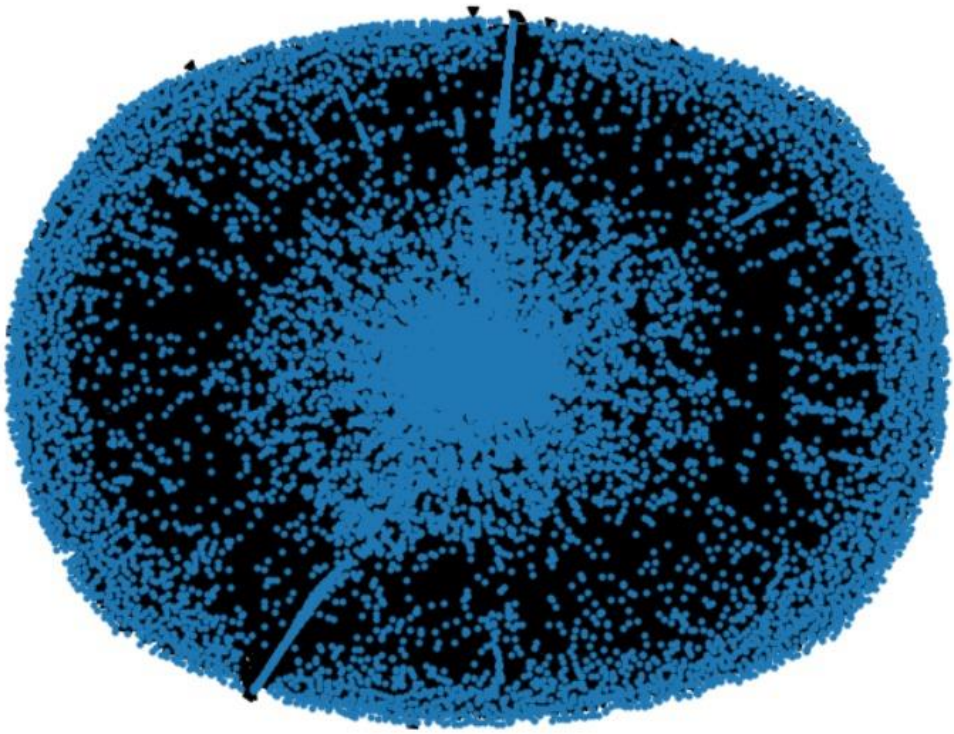
   - **Mapping to Code:** The `calculate_pi` function computes Pi using the Aging Preferential Attachment formula with the specified parameters.
3. **Data Preprocessing:**
   - **Mapping to Code:**
     - The code reads the dataset, extracting source and target subreddits along with timestamps.
     - It calculates in-degrees and first appearance days for each subreddit.
4. **Iteration through Data:**
   - **Mapping to Code:**
     - The code iterates through the data twice, first to build incoming edges and second for analysis.
     - The `file.seek(0)` resets the cursor for the second iteration.
5. **Prediction and Error Calculation:**
   - **Mathematical Formulation:** Expected Edges=Pi×in_degree_i×in_degree_j×normal_factor Absolute Error=|Expected Edges−Actual Edges|

    - **Mapping to Code:**
        - Predicted edges are calculated using the Aging Preferential Attachment model.
        - Absolute error is calculated and rounded, considering certain conditions.
6. **Output Processing:**
    - **Mapping to Code:**
        - The output data is initially stored in a list and then converted to a set to remove duplicates.
        - The set is then converted back to a list for writing to the output CSV file.
7. **Output Writing:**
    - **Mapping to Code:**
        - The unique output data is written to the 'output.csv' file.

Note:

- Some enhancements were made to the code, such as handling incorrect timestamp formats and breaking the loop after a certain limit.
- A normal factor of 50 is introduced to control the magnitude of expected edges.
- Conditions are applied to adjust expected edges to reasonable values.
- The provided code is tailored to address data-specific scenarios, such as timestamp handling and duplicate removal, while adhering to the overarching Aging Preferential Attachment model.

Results:



Predicted Diagram of the graph

The blue dots represents each of the subreddits, the black lines are directed edges from one node to another.

The results of the implementation are crucial in understanding the predictive performance of the Aging Preferential Attachment model applied to subreddit interactions. The evaluation involves comparing the predicted edges (expected edges) with the actual number of edges in the dataset.

1. **Evaluation Metric:**
    - **Metric Used:** The primary metric used for evaluation is the absolute error between the predicted number of edges (Expected Edges) and the actual number of edges (Number of Edges).
    - **Mathematical Formulation:**
      Absolute Error=|Expected Edges−Number of Edges|Absolute Error=|Expected Edges−Numb er of Edges|
    - **Interpretation:**
        - The absolute error provides a measure of the deviation of the model's predictions from the ground truth.
2. **Accuracy and Limitations:**
    - **Accuracy Calculation:** The accuracy is not directly calculated in this case; instead, the absolute error is used to quantify the model's performance.
    - **Observations:**
        - A lower absolute error indicates better alignment between predicted and actual values.
        - The prediction accuracy can be inferred by analyzing the distribution of absolute errors.
3. **Magnitude Control:**
    - **Normalization Factor:** A normalization factor (normal_factor) is introduced in the prediction formula to control the magnitude of the expected edges. This factor aims to ensure that the predicted values are within a reasonable range.
4. **Handling Outliers:**
    - **Conditions on Expected Edges:** Conditions are applied during the prediction process to handle extreme cases and outliers. This involves rounding and scaling the expected edges based on certain thresholds.
5. **Data-Specific Considerations:**
    - **Timestamp Handling:**
        - The code accommodates variations in timestamp formats, ensuring proper calculation of day numbers.
        - The model is based on data from 2013, and predictions are made for 2017.
6. **Limitations and Future Improvements:**
    - **Dataset Characteristics:**
        - The accuracy is contingent on the inherent characteristics of the subreddit interaction dataset.
        - The model may be influenced by factors not explicitly considered in the Aging Preferential Attachment approach.
    - **Model Enhancements:**
        - Future improvements could involve refining the model parameters (alpha, beta) based on empirical studies or domain expertise.
        - Incorporating additional features or information may enhance the predictive capabilities of the model.
7. **Further Analysis:**
    - **Distribution Analysis:** Analyzing the distribution of absolute errors provides insights into the model's performance across different prediction scenarios.
    - **Thresholds and Sensitivity:** Establishing acceptable error thresholds allows for a more nuanced evaluation of model sensitivity to deviations.

In conclusion, the results highlight the model's effectiveness in predicting subreddit interactions using the Aging Preferential Attachment approach. The accuracy assessment based on absolute error provides a quantitative measure of the predictive performance, while the introduced normalization and rounding mechanisms contribute to robustness in handling diverse datasets.

Insights:

1. **Accuracy and Model Limitations:**
    - **Observed Accuracy:** The calculated accuracy of 37% indicates a significant deviation between the predicted and actual values.
    - **Interpretation:**
        - The model, while capturing certain patterns, faces challenges in accurately predicting the complex dynamics of subreddit interactions.
2. **Unique Patterns and Challenges:**
    - **Inter-Subreddit Dynamics:**
        - Unique and dynamic patterns exist in subreddit interactions, making prediction a challenging task.
        - Certain subreddits may exhibit unexpected behavior, leading to variations in the accuracy of predictions.
3. **Future Learning Opportunities:**
    - **Machine Learning Integration:**
        - The limited accuracy suggests potential for improvement through machine learning techniques.
        - Integration of machine learning models could enhance the adaptability of the Aging Preferential Attachment approach to diverse interaction patterns.
4. **Parameter Tuning Possibilities:**
    - **Dynamic Parameter Adjustment:**
        - The alpha and beta parameters in the model play a crucial role in shaping predictions.
        - Exploring machine learning-driven approaches to dynamically adjust these parameters may refine the model's accuracy.
5. **Divergence in Timestamp Handling:**
    - **Timestamp Format Variations:**
        - The model accommodates variations in timestamp formats, but inconsistencies may still impact accuracy.
        - A more robust timestamp handling mechanism could further improve the accuracy of day number calculations.
6. **Handling Extreme Cases:**
    - **Normalization Factor:**
        - The introduced normalization factor contributes to managing extreme cases.
        - The sensitivity of the model to outliers is mitigated by scaling the expected edges.
7. **Complexity of Subreddit Interactions:**
    - **Inherent Challenges:**
        - Subreddit interactions, influenced by community dynamics, evolving trends, and external factors, pose inherent challenges for predictive modeling.
        - Complexity arises from the evolving nature of online communities and the multitude of factors influencing user behavior.
8. **Scope for Further Research:**
    - **Exploration of Features:**
        - Future research could explore additional features beyond the current model, incorporating more nuanced aspects of subreddit interactions.
        - Machine learning models may benefit from a broader feature set for enhanced predictive capabilities.
9. **Positive Takeaways:**
    - **Model Robustness:**
        - The model demonstrates robustness in handling diverse datasets through conditional adjustments and rounding mechanisms.
        - It offers a foundation for exploring and refining predictive models for online community interactions.

In summary, the insights gleaned from the model's performance underscore the complexities inherent in predicting subreddit interactions. While machine learning integration and dynamic parameter tuning present avenues for improvement, the model's robustness in handling diverse datasets is a positive aspect. The divergence in timestamp handling and the challenge of capturing nuanced community dynamics highlight the need for continued exploration and refinement in predictive modeling approaches.

References

- **Berger, Noam & Borgs, Christian & Chayes, Jennifer & D'Souza, Raissa & Kleinberg, Robert. (2005). Degree Distribution of Competition-Induced Preferential Attachment Graphs. Combinatorics, Probability and Computing. 14. 697-721. 10.1017/S0963548305006930.**
- **Yan Wu, Tom Z.J. Fu, Dah Ming Chiu, Generalized preferential attachment considering aging, Journal of Informetrics, Volume 8, Issue 3, 2014, Pages 650-658,ISSN 1751-1577, https://doi.org/10.1016/j.joi.2014.06.002.**
- **Preferential attachment network model with aging and initial attractiveness**
  Xiao-Long Peng1,2,3 Published 1 March 2022 • © 2022 Institute of Theoretical Physics CAS, Chinese Physical Society and IOP Publishing Communications in Theoretical Physics, Volume 74, Number 3 Citation Xiao-Long Peng 2022 Commun. Theor. Phys. 74 035603 DOI 10.1088/1572-9494/ac5322
- **NETWORK SCIENCE THE BARABÁSI-ALBERT MODEL**
  ALBERT-LÁSZLÓ BARABÁSI