Can you convince me?

Shared internet communities and persuasion on Reddit r/ChangeMyView

Roshan Ramkeesoon
eScience Institute
University of Washington
Seattle, WA 98195
roshanr@uw.edu

Deepthi Hedge
eScience Institute
University of Washington
Seattle, WA 98195
dhegde@uw.edu

Geoffrey Li
eScience Institute
University of Washington
Seattle, WA 98195
geoffli@uw.edu

Abstract

The factors that determine the persuasiveness of an argument are difficult to understand and even harder to evaluate. Many factors are intrinsic to the content and tone of the conversation, while other extrinsic factors are better defined by the background and communities of the participants. Prior works have revealed that linguistic elements are predictive of the persuasiveness of an argument. In this work, we show that combining linguistic features with social similarity measurements results in a 1% improvement in accuracy of predicting the persuasiveness of arguments. We find that increased social distance in online communities is weakly associated with increased likelihood of changing a person's views.

1 Introduction

Understanding the subtleties of conversation that can change a person's mind has the ability to influence many fields including advertising, activism, and even suicide prevention. Prior research suggests that context and choice of words can have a large impact on the persuasive outcome [1][5]. In this study, we aim to examine whether having shared communities with another person also has a measurable effect on persuasiveness.

The advent of large online forums such as Reddit, and the methods to process high volumes of textual data can inform our understanding of the psycholinguistics and context in which people are persuaded. Reddit is a popular online discussion forum. It is organized into many smaller forums called "subreddits" which vary in content. Users can post to a subreddit as well as comment on posts and other comments. Posts and comments are organized by how many up or down votes they receive from other users.

In many forms of social media, so-called "echo chambers" are becoming increasingly prevalent, where inherent tribalism echoes and magnifies similar opinions. In certain areas, Reddit offers a unique forum where users willingly leave such echo chambers and engage in discussions that may run counter to their established viewpoints.

In this study, we analyze one such subreddit, r/ChangeMyView (CMV) where users post a point of view they have with the intention of receiving counterarguments from other users. Participants

Preprint. Under review.

in the subreddit post an opinion or viewpoint, and “challengers” are able to reply to the original participant’s post in an attempt to change their view (as the name of subreddit suggests) through debate and conversation. The original poster (OP) has the ability to award "Delta" points to posts which they find insightful. Delta points do not have to indicate a complete reversal of opinion on the topic but rather indicate any degree of change, significance, or importance to the reader of the counterargument. One key benefit of the subreddit is the level of careful moderation: the quality of debate is ensured to be high since irrelevant/unhelpful/"low-effort" comments are removed.

We explored graph-based approaches in this study and found that augmenting features with linguistic richness of comments and the dynamics of users in the user-subreddit space is predictive of the likelihood of persuasion.

2 Related Work

2.1 Persuasion strategies on r/ChangeMyView

In this study [5], the authors attempt to identify linguistic factors that make some arguments more successful than others. They consider aspects like participant entry-order, degree of back-and-forth exchange, and the interplay between the opinion and the counterargument. The authors found that the strongest predictive features include interplay features, including links as evidence, and dissimilarity with the original post in terms of word usage.

The authors primarily are concerned with the linguistic structure of the arguments, instead of reasoning strategies (e.g. how the argument is presented instead of what is being said). They trained logistic regression models with L1-regularization to predict the likelihood of each challenger reply successfully changing the original poster’s mind or not. Comparing models with varying amounts of the features they built, they conclude which features most likely contribute to the success of an argument. They also tried L2 regularization, random forests, and gradient boosted classifiers with no improvement beyond the cross-validation standard error.

We reproduce the results from this paper and use it as our baseline since it leverages the same corpus of Reddit data that we looked at (the r/ChangeMyView subreddit). We consider a setup similar to what is proposed by the authors of the paper. This gives us a reasonable baseline for our proposed method and a fair way to evaluate our own findings. We propose an additional dimension to the feature set, one that leverages the interaction of users within different subreddit communities. The users who participate in CMV are also likely to participate in other subreddits, which proves to be useful in predicting the likelihood of receiving a delta point. We try to address questions such as: Are users who have overlap in shared interests (via common subreddit participation) and are a part of similar communities more susceptible to have their views changed by each other?

One criticism of this paper is how the authors chose negative labels for the binary classification task. While the positive examples are clearly labeled by the OP with a Delta, one must define a method for selecting negative examples. The selection of negative examples could drastically alter the results of an experiment. They selected negative examples by choosing the counterargument that had the highest bag of words Jaccard similarity to an argument that received a Delta. Their conceptual goal was to evaluate the differences in how two topically similar arguments were presented. In other words, if two arguments use similar words, they assumed they are likely presenting the same argument in slightly different ways. We felt that this method was weakly validated.

2.2 SimRank - A Measure of Structural-Context Similarity

While many measures of similarity exist, SimRank [4] is a generalizable algorithm used to measure similarity of nodes in a graph that exploits the object-to-object relationships found in many domains. SimRank is an extension of the PageRank algorithm which works by initializing random walkers on a node of interest and measuring the likelihood of seeing the random walkers at other nodes in the graph after many iterations. The random walk also has restarts, meaning there is a probability, governed by a fixed model parameter (β), with which a random walker will restart back at the original point. In practice this is accomplished using power iteration. The primary difference between SimRank and PageRank is that SimRank starts by concentrating all of its PageRank on the starting node.

The basic SimRank equation is defined recursively:

$$s(a,b) = \frac{C}{|I(a)||I(b)|} \sum_{i=1}^{|I(a)|} \sum_{j=1}^{|I(b)|} s(I_i(a), I_j(b))$$

where $I(a)$ is defined as the set of in-neighbor nodes of node a, C is a constant between 0 and 1, and self similarity is 1, $s(a,a) = 1$.

This method works for many types of graphs including bipartite graphs such as the user-subreddit graph we construct. Scores can be thought of as "flowing" from a node to its neighbors. Each iteration propagates scores one step forward along the direction of the edges, until the scores converge. In the case of a bipartite graph, the PageRank score oscillates between the members of the two classes of nodes at each iteration. However, the algorithm will still converge to a stable SimRank score across all nodes in the network.

An important limitation of SimRank is that it only considers similarity in the structural context of a graph, and does not consider other, potentially relevant, similarities. For example, in an Internet network composed of pages and links, SimRank would only measure similarities in the structural context of the graph created by the links, and ignore similarities in the actual context of the page. For this reason, the authors recommend aggregating SimRank with other similarity measurements.

As defined by Goyal et al. [3], in graphs, one can define measures of first-order proximity, which only consider properties of the edge between two nodes, and second-order proximity, which compares the neighborhoods of two nodes. The cosine similarity of the subreddits that two members participate in is a first order proximity measure. The SimRank similarity of two nodes is a second-order proximity measure. In our study, we are interested in comparing the association between various social structural context features with linguistic features.

3 Dataset

Since we were building on the analysis of Tan et al. [5], we leveraged their Reddit r/ChangeMyView dataset. This balanced dataset included 4,546 training examples and 1,046 test examples from r/ChangeMyView spanning 1/1/2013 through 9/1/2015. Each example has the original post, a comment which received a Delta, and a comment which did not receive a Delta.

To examine the association of communities with likelihood of receiving a delta, we needed to construct a graph of Reddit during that time period to observe those user's behavior across Reddit. We constructed a bipartite graph of Reddit from sets of users and subreddits. To build the Reddit graph, we used Google BigQuery which has hosted all Reddit comments and posts from 2009 to 2018. We used Google Cloud Platform (GCP) for storage and compute instances.

We model users, subreddits, and their relationships with a graph $G = (V, E)$ where nodes in V represent the set of users and subreddits, and edges in E represent the set of relationships between users and subreddits. The weighted directed edge <p, q> corresponds to the number of comments that a user p made in subreddit q. The edges are weighted by number of comments (higher activity edges have heavier weights). We created a graph for each 6-month period in our dataset. For the time period in our analysis (1/1/2013 through 9/1/2015), there were: 1.6 billion comments, 10K subreddits with more with 1K comments, 84K unique users on r/ChangeMyView alone, and 5.6 million edges.

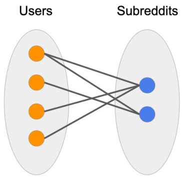


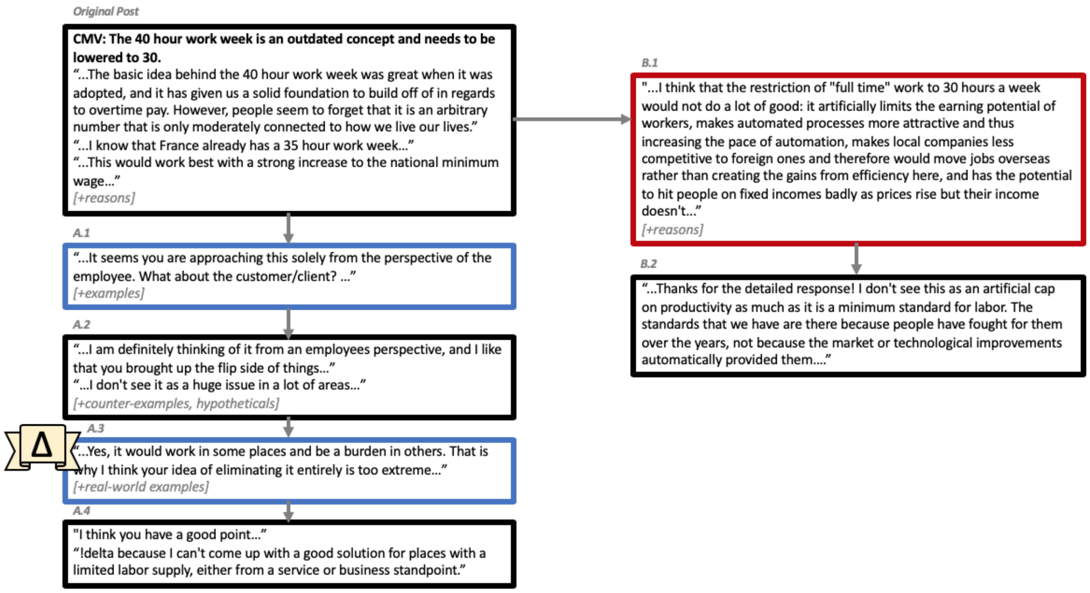
Figure 1: The relationships between users and subreddits on Reddit forms a bipartite graph. In our study, we weight edges based on the number of comments a user has made in the given subreddit in a 6 month period.

4 Problem Setup and Analytic Approach

Using the ChangeMyView subreddit, we will analyze linguistic and community based factors related to the likelihood of a user or comment receiving a Delta.

4.1 Problem Setup and Creating Labels

We plan to create a binary classification prediction model to evaluate the likelihood of a challenger being awarded a Delta. Our general approach is to define a similarity metric between two Reddit users that embeds the social similarity between these two users.



Original Post

CMV: The 40 hour work week is an outdated concept and needs to be lowered to 30.
"...The basic idea behind the 40 hour work week was great when it was adopted, and it has given us a solid foundation to build off of in regards to overtime pay. However, people seem to forget that it is an arbitrary number that is only moderately connected to how we live our lives."
"...I know that France already has a 35 hour work week..."
"...This would work best with a strong increase to the national minimum wage..."
[+reasons]

A.1

"...It seems you are approaching this solely from the perspective of the employee. What about the customer/client? ..."
[+examples]

A.2

"...I am definitely thinking of it from an employees perspective, and I like that you brought up the flip side of things..."
"...I don't see it as a huge issue in a lot of areas..."
[+counter-examples, hypotheticals]

A.3

"...Yes, it would work in some places and be a burden in others. That is why I think your idea of eliminating it entirely is too extreme..."
[+real-world examples]

A.4

"I think you have a good point..."
"!delta because I can't come up with a good solution for places with a limited labor supply, either from a service or business standpoint."

B.1

"...I think that the restriction of "full time" work to 30 hours a week would not do a lot of good: it artificially limits the earning potential of workers, makes automated processes more attractive and thus increasing the pace of automation, makes local companies less competitive to foreign ones and therefore would move jobs overseas rather than creating the gains from efficiency here, and has the potential to hit people on fixed incomes badly as prices rise but their income doesn't..."
[+reasons]

B.2

"...Thanks for the detailed response! I don't see this as an artificial cap on productivity as much as it is a minimum standard for labor. The standards that we have are there because people have fought for them over the years, not because the market or technological improvements automatically provided them...."

Figure 2: Example of two CMV discussion threads, one successful and one not. Colors denote unique users.

For simplicity, we are limiting the cases we examine to interactions between OP and other commenters, and ignoring all interactions between commenters.

We will use the same positive and negative labels for our prediction model as Tan et al. First, Tan et al. filters the set of r/ChangeMyView posts for quality to have a minimum number of comments and unique commenters in the conversation. Challengers that were awarded a Delta by OP are considered to be the positive case. Challengers that were not awarded a Delta by OP but have a similar response, as measured by Jaccard similarity of the words used, to the successful challenger are considered to be in the negative case. Figure 2 above shows an example of a CMV post. We would label the blue challenger in discussion tree A as a positive case, since the blue challenger is successful in changing the view of the OP after a few back-and-forth replies, and the OP awards the challenger a Delta in reply A.4. The challenger highlighted in red is unsuccessful. By only accepting positive and negative cases in pairs, we ensure class balance in our dataset.

4.2 Cosine Similarity

The set of subreddits in which a user is an active participant provides one way of denoting that user's communities. If two users participate in similar subreddits, it may indicate they have similar interests and views. We are interested in evaluating if shared subreddits is associated with persuasiveness in conversation. Moreover, we are interested in using cosine similarity as a baseline similarity measurement to compare against SimRank, a method which compares the neighborhood of two users in a graph.

Let U represent the set of all users

$$U = \{u_1, u_2, ..., u_m\}$$

Let S represent the set of all subreddits

$$S = \{s_1, s_2, ..., s_n\}$$

Define z_i the subreddit embedding for each user u_i

$$z_i = [1_{u_i \in s_j}]_{j=1}^{n}$$

$$cosine - similarity(u_1, u_2) = \frac{z_1 \cdot z_2}{\|z_1\|\|z_2\|}$$

Equivalently:

$$cosine - similarity(u_1, u_2) = \frac{\#\ shared\ subreddits}{(\#\ subreddits\ u_1\ is\ in)(\#subreddits\ u_2\ is\ in)}$$

4.3 SimRank

By applying the SimRank algorithm to a bipartite Reddit graph (Figure 1), we can generate similarities for each user pair that compares the neighborhood of users and subreddits local to each user. We divided the time span of our dataset into smaller groups of 6 months to account for changes in the graph with time. Since each graph has ~2 million edges, we implemented SimRank in Spark in a memory efficient manner by reading the adjacency matrix into memory one row at a time at each iteration.

We used a standard teleportation probability of $\beta = 0.8$ and a short random-walk length of 10 to keep the neighbourhood local in our implementation. As SimRank scores oscillates between the two sets of a bipartite graph before convergence, we chose an even number for the walk length since we want our SimRank scores to be more heavily weighted on the users side of our bipartite graph.

Previously, we considered using other graph-embedding techniques such as DeepWalk and Node2Vec. However, we found SimRank to be better aligned with what we want to achieve since it gives a similarity score between each node pair as opposed to a round about way of getting similarity using latent representations.

4.4 Linguistic Features

We would like to see if augmenting the linguistic feature set with features pertaining to shared communities helps improve our predictions of awarding delta. We do so by extracting the same features that Tan et al. [5] mentioned to be associated with awarding a Delta point:

1. Word category-based features. This includes features around the purpose of words used in the argument: definite vs. indefinite articles used, first-person vs second-person pronouns, number of links, number of quotations, number of examples, etc.

2. Word score-based features. Word choice in the argument can have a significant impact on how the argument is received. Accordingly, the authors scored each word in the argument based on categories like arousal, concreteness, dominance, and valence.

3. Entire argument features. This includes features in terms of how the argument was structured, such as the number of sentences vs. number of paragraphs, word entropy, and readability (using Flesch-Kincaid grade levels).

4. Markdown formatting. Since Reddit comments employ Markdown conventions, we are able to recover the formatting choices the challenger used, including italicized and bolded words, and the number of bulleted lists.

4.5 Predicting Likelihood of Awarding Deltas

We first want to test the hypothesis that cosine similarity or SimRank is associated with likelihood of OP awarding a Delta point. We will do this by using Welch's robust two-sample T-test.

We will then compare the different similarity measurement's ability to predict likelihood of OP awarding a Delta point by evaluating the performance of logistic regression models trained on each similarity measurement. We will evaluate how each model performs on a holdout set as an indication for the relative improvement in predictive power. In addition, we will compare the performance of

Feature Name	Significance
words	****
Word category based features	
definite articles	****
indefinite articles	****
positive words	****
negative words	****
1st person pronouns	****
1st person plural pronouns	****
2nd person pronouns	****
links	****
.com links	****
frac. of links	****
frac. of .com links	****
frac. of definite articles	*
frac. of positive words	*
.edu links	*
PDF links	*
quotations	*
examples	*
question marks	*
Word score based features	
Arousal	*
Valence	*
Dominance	
Entire argument features	
Word entropy	****
sentences	****
Type-token ratio	****
paragraphs	****
Flesch-Kincaid grade levels	****
Markdown formatting	
italics	
bolds	
Numbered words	
Interplay features	
Reply frac. in all	****
Reply frac. in content	****
OP frac. in stopwords	****
Reply frac. in stopwords	****
OP frac. in all	****
common in all	****
Jaccard in content	****
Jaccard in stopwords	****
common in content	****
OP frac. in content	*
Jaccard in all	*

Table 1: Linguistic features and their significance levels indicated by * (More number of * indicates lower p-value)

models trained using user community-based similarity measures along with the linguistic embeddings described in Tan et al. As a result, we can discuss the difference we observe by adding community features in predicting if a person is likely to be convinced in online forums.

5 Results and Findings

5.1 Validation of distance measures

We examine how the metric is distributed over the dataset and validate that it measures an intuitive notion of user similarity: number of shared subreddits. Most users share few subreddits with other users. This implies sparsity in the graph and may be the result of local communities on Reddit. The sparsity is also seen in the high cosine distances and low SimRank similarity between most users.

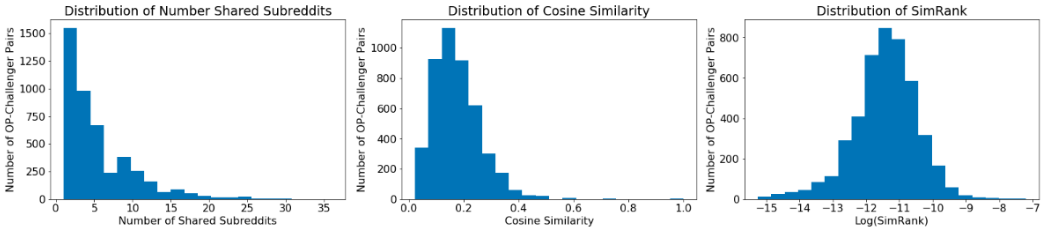


Figure 3: (a) Distribution of number of shared subreddits, (b) cosine similarity , and (c) log(SimRank) score. All plots were created using all user pairs in the Reddit r/ChangeMyView dataset generated by Tan et al. [2]

We expect that as two users are active in more shared subreddits, on average, any measure of user similarity should increase. We notice a small but direct relationship between number of subreddits and both similarity measures (Figure 4). In addition, we see a direct but high variance relationship between cosine similarity and SimRank similarity.

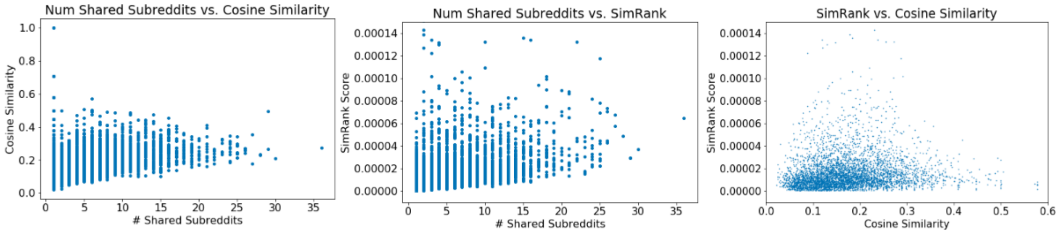


Figure 4: (a) Cosine similarity versus number of shared subreddits. (b) SimRank score versus number of shared subreddits. (c) SimRank score versus cosine similarity of users. All plots were created using all user pairs in the Reddit r/ChangeMyView dataset generated by Tan et al. [2]

5.2 T-tests

	t-statistic	p-value	significance
Cosine Similarity	-0.60	0.55	
SimRank	-2.04	0.09	*

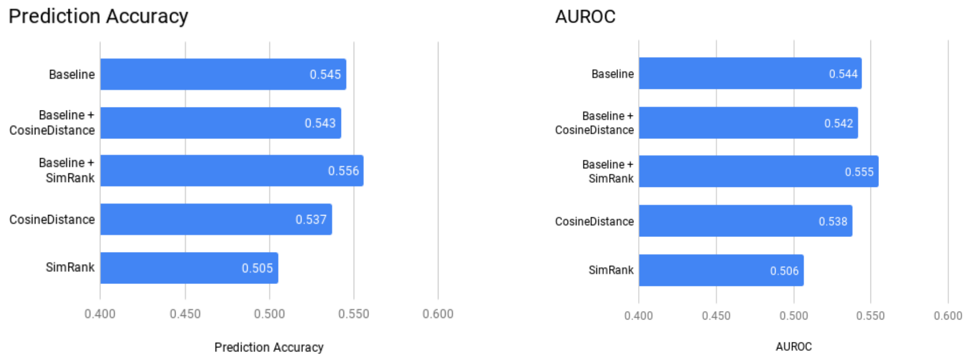
Table 2: Results of unequal variance t-test for first-order and second-order proximity measures.

Table 2 above shows the result of our two-sample Welch's robust t-tests for the first order proximity measure (Cosine Similarity) and second order proximity measure (SimRank), comparing successful Challenger posts and unsuccessful Challenger posts.

The negative t-statistics for both measures indicate that the mean proximity measures were higher for the unsuccessful group than the successful group. Thus, we see that on average, OPs are more likely to award Deltas to Challengers that are less similar to them in the Reddit graph.

The SimRank t-statistic is weakly significant (at $\alpha = 0.10$), and the Cosine Similarity t-statistic is not significant.

5.3 Binary Classification Prediction Model



(a) Prediction Accuracy

(b) AUROC

Figure 5: Performance of L2-regularized logistic regression classification models on holdout set. Baseline model is trained using only the linguistic features from Tan et al. [2]

We trained five L2-regularized logistic regression classifiers with optimal regularization parameters found using 5-fold cross-validation. The performance on the holdout set is shown in Figure 5.

The Baseline model refers to the a model trained with the linguistic features that Tan et al. developed. We were able to reproduce their highest-performing model's AUROC (54%) almost exactly. As the results of the two-sample t-tests suggested, SimRank was able to boost prediction accuracy and AUROC when trained with the Baseline features. Cosine Distance, however, did not improve the Baseline model.

When trained on Cosine Distance and SimRank independently (the fourth and fifth models in Figure 5), the Cosine Distance model performed better than SimRank.

6 Discussion

Since we could not establish statistical significance (p<0.05) on the t-test of cosine similarity and SimRank on receiving a Delta point, we cannot establish a strong association between social structural context and likelihood of receiving a Delta. However, since SimRank has more statistically significant association with whether or not a commenter receives a delta point (p=0.09) than cosine similarity and since the logistic regression classifier trained on linguistic features and SimRank scores outperforms all other models, it appears the neighborhood of users is more important in determining the likelihood of receiving a delta point than just the cosine similarity of shared subreddits between two users. The improved prediction accuracy of the model that uses both linguistic features and SimRank over the models that use only the individual features implies SimRank embeds useful information in predicting likelihood of receiving a Delta that is not captured from linguistic features alone.

Both cosine similarity and SimRank showed that challengers who are less similar to the OP have a higher likelihood of convincing the OP. This could mean that users who are more distant in the Reddit graph are more likely to propose counterarguments which are sufficiently distant from the set of counterarguments the OP has already heard. Future research into the content of the comments is necessary to determine the reason for this relationship.

While our study examined a linear relationship between user similarity and likelihood of receiving a Delta, in future research we would like to examine other types of relationships. For example, there may exist a peak distance associated with likelihood of receiving a delta. A challenger who is very similar to the OP might present counterarguments the OP is familiar with but finds unconvincing, hence the r/ChangeMyView post. But a user who is very different from OP might produce counterarguments that are too different for the OP to accept. Users who are in between may produce "sweet spot" arguments that can help change the OP's mind. Another way to investigate this phenomena would be to investigate cliques using hierarchical clustering and examine if a challenger from a neighboring cluster to the OP is more likely to persuade the OP than a challenger in the OP's cluster or a challenger very far from the OP's cluster.

In our study, we found that a challenge in defining a similarity measure for the Reddit r/ChangeMyView dataset is the sparsity of shared subreddits. As shown in Figure 3a, most users share 0 or 1 subreddit with other users. However, users may be active in a variety of other subreddits. As a result, the cosine similarity shows high variability based on the number of subreddits in which each user is active. In a future study, we may wish to explore other first order similarity measures that are robust to high fluctuations in the number of subreddits in which users are active.

Another key challenge is defining the negative cases to compare against the Delta awarded cases. For the purposes of comparing our results with the results of Tan et al.'s linguistic models, we used their definition of negative cases which attempted to control for the argument being made in order to isolate the effect of psycholinguistics. However, the community effects we are interested in studying may manifest themselves not just in how counterarguments are presented, but in the content of the counterarguments. In future studies, we wish to experiment with other selection methods for negative cases to analyze the social network factors related to how arguments are formed.

A limitation of the SimRank algorithm is the difficulty validating the choice of hyperparameters. We attempted to do so by seeing if there was a positive relationship between SimRank and number of shared subreddits: an intuitive measure of user similarity (Figure 4). In future research, we would like to investigate the number of steps each random walk takes and the way by which edges are weighted. We would also like to investigate SimRank extensions including SimRank++[2] which remedies the issue of "zero-similarity" a node may have with all nodes that are further than the number of steps away from it.

7 Conclusion

In this study, we explored a graph-based approach to see if OP-Challenger relationships via shared subreddit spaces are associated with likelihood of receiving a Delta. We observed that augmenting linguistic features with knowledge about the position of the user in a social graph improves prediction accuracy by 1%. We find that increased social distance is weakly associated with likelihood of changing another person's view. If pursued further, this result could verify the potential danger of echo-chambers. We believe graph-based approaches are a promising direction and may help model user background and social biases, which are hard to capture by linguistic context alone.

References

[1] Althoff, T. Clark, K., & Leskovec, J. (2016). Large-scale Analysis of Counseling Conversations: An Application of Natural Language Processing to Mental Health, *Transactions of the Association for Computational Linguistics.*

[2] Antonellis, Ioannis, Hector Garcia Molina, and Chi Chao Chang. "SimRank++: query rewriting through link analysis of the click graph." Proceedings of the VLDB Endowment 1.1 (2008): 408-421.

[3] Goyal, Palash, and Emilio Ferrara. "Graph embedding techniques, applications, and performance: A survey." Knowledge-Based Systems 151 (2018): 78-94.

[4] Jeh, Glen, and Jennifer Widom. "SimRank: a measure of structural-context similarity." Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2002.

[5] Tan, C., Niculae, V., Danescu-Niculescu-Mizil, C., & Lee, L. (2016, April). Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In Proceedings of the 25th international conference on world wide web (pp. 613-624), *International World Wide Web Conferences Steering Committee.*