Determining Topic Homogeneity in Online Sub-communities

Frank Chan (fc249), Peter Li (pl488), Saakshi Singhal (ss992)
Github URL: https://github.com/peteli3/networks-final-project

May 16, 2018

Abstract

The emergence of online social media has created numerous opportunities for individuals to share ideas within communities centered around areas of interest. In the mid-2000s, online forum culture grew rapidly as a result. One such example is Reddit, a bulletin-board-like platform built in 2005 to scale the forum world. At first, content seeded on Reddit was curated and broad. The very first "subreddit" to exist was the entire website. As Reddit later launched user-created subreddits, targeted sub-communities materialized around topics of interest. We aim to model this trend toward specificity using a structure-based analysis on relevance of posts within sub-communities.

1 Introduction

In this paper, we will discuss the methods undertaken to model internal relevance within sub-communities on Reddit. We will be employing elementary Natural Language Processing methods and network structure analysis methods to propose a single empirical measure of closeness within a sub-community. We will then attempt to justify our model by hypothesizing about its function within the context of other works.

To begin, we will clarify some terminology that we will consistently refer to throughout this paper. In Reddit, a sub-community is termed a subreddit, and we will use the two interchangeably throughout the paper. In addition, we will refer to a "thread" as a body of text in an original post concatenated with all the comments. We will be using reasonable pre-processing rules to prepare each thread for topic modeling. We refer to a "post" as any type of text submission to the reddit platform, whether it be as a comment or as an original thread body.

The datasets used in this project are composed of actual Reddit comments. They are represented as *.json files where each file contains all the posts made to Reddit over a single month interval. We then supplement each *.json file with the original thread bodies associated with each thread, scraped directly from reddit.com's API (more on this in the Fabrication section below). In cleaning the data, we decided to remove thread bodies and comments for which the original post is a link, since

(due to their reaction-like nature) comments to that post will not contain enough information to induce a meaningful topic vector. In a similar manner, comments of reaction-like nature will not contribute to the training of the topic model.

We hypothesize that there exists some scalar value that can represent internal topic closeness within a subcommunity. For instance, we expect to see a substantially higher representation of closeness in subreddits with a specific title like r/zuckmemes than in subreddits with a general title like r/politics. An initial naïve representation would be to topic model the subreddit via a method such as LDA and use the sum of vector distances as a metric. However, this metric is no better than a RSS is to regression analysis and does not consider any specific aspects of the subreddit such as bottle-necking expansion or distribution of topic spread. To address these issues, we will break down the target definition of "internal topic closeness" or "tightness" into two major components: connectivity and spread.

We will then attempt to model the relevance structure of a subreddit from our curated dataset by building networks with various structures and implications. We expect to be able to draw mathematical representations of connectivity and spread from these networks. Each such representation of the data provides a different perspective on the same data and lends insight toward the true relevance structure within a subreddit. Using our developed measures of connectivity and spread, we will appropriately penalize or reward strong values and build a hypothesized "closeness" or "tightness" index. We aim to propose a representation that is empirical and can be meaningful as a standalone value.

2 Prior Work

Previous literature has explored the idea of communities in Reddit, focusing on how user interact with different subreddits on the website, as well as what types of posts do well and where they do well.

2.1 What's In A Name (H. Lakkaraju, J. J. McAuley, J. Leskovec, 2013)

What's In a Name by Lakkaraju et. al. is a 2013 paper that looks at the success of resubmitted content on Reddit. In particular, the authors explore images that have been submitted multiple times on multiple Reddit communities and attempt to see what features impact the success of these posts when the content of the image remains the same. In particular they explore the effects of community (which subreddit the post was submitted to), title, and temporal features (time of day the post was submitted) in building a model to predict success. Their findings indicate that while content is an important factor in post success, both the title and community features also play a large role. In particular, the paper's findings show that content that becomes popular in highly visible communities is unlikely to become popular again while content that succeeds in more niche communities can become successful in a different niche community. In addition, the language model proposed in the paper

shows that successful Reddit titles are the ones that are similar to other titles in the community in terms of style and semantics, yet unique in their content in order to capture the interest of other users. This results suggest that each community, or subreddit, has a certain sense of identity that we can further explore. Meanwhile, the same posts can do well in highly visible communities which are not particularly centered around a specific topic, which leads to questioning how homogeneous these communities are.

2.2 The Impact of Crowds on News Engagement: A Reddit Case Study (B. Horne and S. Adali, 2017)

Another study that looks at the Reddit database is The Impact of Crowds on News Engagement: A Reddit Case Study by Horne and Adali. The study looks at features that affect news popularity on the "world news" subreddit. Perhaps the most compelling find in the paper is that user voting on posts is predicted heavily by various features including content while commenting on posts is based mostly on subjective features. The authors speculate that this indicates that commenting is done by users based mostly on emotion while voting is highly impacted by content. The paper also finds that users who changed the titles of news articles followed patterns for the alterations they made. In particular, the paper found that titles were changed to be more positive and less negative. In addition, titles were often made longer, more informal and more difficult to read. Finally, the authors found that changing the title of a news article slightly increased both the score and number of comments on the post in the subreddit. This paper allows us to gain a better general understanding of the content and popularity of posts on reddit subcommunities and gives us a better understanding of how content is changed slightly as it makes its way around the Reddit community.

2.3 All Who Wander: On the Prevalence and Characteristics of Multi-community Engagement (C. Tan and L. Lee, 2015)

All Who Wander: On the Prevalence and Characteristics of Multi-community Engagement by Tan and Lee looks at three main user behaviors on Reddit. They observe the sequence of communities that users join, the language the use in those communities and the feedback that they receive on their posts. In addition, they use these features to predict the future activity level of users on Reddit. Tan and Lee find that, contrary to their hypothesis, longtime users online still visit new communities and continue to "wander" throughout their lifespan on the application. In addition, the paper finds that for their lifetime on Reddit, users continually adopt the language models employed within the communities they visit. This shows that users do not necessarily settle into one way of writing, but instead adapt based on the community environment they are in. The concept of language adaptation raises natural questions about topic homogeneity across Reddit communities. On one hand, it is natural to predict that topics could be homogeneous due to user language adaptation. On the other hand, the wandering nature of users could indicate communities with a group

of varied topics.

3 Network Models

In this section of our paper, we will discuss the construction behind our network models. Each description will be written generally, but we will provide examples from r/politics January 2010 to emphasize structural characteristics worth considering.

3.1 Topic Modeling and Node Creation

We have chosen Latent Dirichlet Allocation (LDA) [2] on t topics to be the crux of this study since it has been shown to be very effective in picking out topics within training corpora. As such, all analyses we perform from this point on will be based on the effectiveness of LDA and how accurately the probability vector obtained from LDA prediction represents the contents of the testing text. Using this technique for topic discovery, we then represent each thread as its probability vector and create a node in our graph for each thread.

At this point, we have our set of nodes and their associated probability vectors which lie in $\mathbb{R}^t$. As is common in text mining, we will use cosine distance (1 - cosine similarity) for the distance metric between topic vectors for threads since we want a distance metric that ignores the magnitude of vectors. Furthermore, with the following edge creation rules specified below, we be analyzing the resulting unweighted and undirected graphs for various structure implications.

3.2 Threshold Model

Our first model will be based on creating edges within a specific subreddit using a threshold. In particular, nodes represent threads in the subreddit, while edges between nodes represent "closeness" between threads – threads similar enough within a certain threshold are thus connected. Using cosine distance as the metric between vectors, we measure the distance between all pairs of nodes and create an edge between any two nodes u, v if their cosine distance is within a predefined threshold: $\sigma(dist\_matrix)/\delta$ where δ is a chosen parameter. We've specifically chosen to use the standard deviation of the distances since it is a relative measure of spread within each unique network's distribution of cosine distances. In doing so, we "normalize" the influence of topic spread on edge creation across various subreddits.

Initial observations of subreddits modeled using thresholds have consistently resulted in graphs that begin completely disconnected for zero threshold (zero dissimilarity tolerance) and quickly add edges until the entire graph becomes one connected component.

In this manner, the threshold model's growth toward being fully connected is tied to the observed distribution of distances between pairs of nodes in the graph. We previously attempted to measure how quickly the giant component consumes the

disconnected nodes - as a measure of closeness. However, the method provides only a relative (to other graphs) sense of the distribution. As such, it is too shallow a metric to use as a standalone index of subcommunity closeness. Thus, we will consider the Threshold Model only in qualitative analysis of true relevance structure.

3.3 k-NN Model

An alternative and softer construction method for our graph is k-NN. For each node u, we will add outgoing edges to the k closest other nodes by cosine distance. This construction provides a graph in which the minimum degree of any node must be k. When varying k, we expect to observe that the graph starts completely disconnected for $k = 0$ and that as k increases, multiple evenly sized components should emerge, with perhaps sparse connections in between. This model attempts to preserve as much of the true relevance structure as possible since k-NN rewards highly related topic chains by representing them as dense sub-components and penalizes tangentially related topic chains by building them as sparsely connected sub-components.

In this manner, edges surface as a result of the increasing tolerance of dissimilarity which slowly connects dense pseduo-regular components until the entire graph is connected. In other words, the k-NN construction's growth toward connectivity is tied to the true underlying relevance structure. Thus, choosing the smallest integer k needed that connects the k-NN graph (call it k^*) also gives us a relative measure of closeness. These qualities and expectations for the k-NN Model make it a more suitable candidate for our structural analysis than the Threshold Model.

3.4 Fabrication

The data for our network model was gathered by parsing from two sources. The first source was an archived collection of Reddit comments during specific months of history. For our purpose, we used the January 2010 dataset, consisting of all Reddit comments made during this month. This data was structured as an unordered file with each line containing a JSON object that represented a comment, complete with id, timestamp, score, and thread id, and other details. Given a list of particular subreddits that we wished to observe, we had to read every line of the file to filter out all the other subreddit comments and sort the comments into separate files per subreddit of interest. In order to rehydrate the original post, we made a request to reddit.com using the thread id of the comments.

Sifting through the comment file and rehydrating the original posts for each thread requires significant computational time. In order to retrieve the data required in a reasonable amount of time, the process of reading through the comment file was parallelized. The comments file itself was divided up into smaller files, while a pool of worker processes filtered files and accumulated results for each thread id of each subreddit of interest, collecting the text body, net score, and number of comments. The results were then aggregated by the master process into a single file to be used later to build the network. Requests to reddit.com were batched, with each request

containing approximately 100 thread ids to retrieve. This was based on the fact that Reddit limits requests to only returning 100 or less results. Any request that asks for more will be divided into multiple requests. Even then it was necessary to add a wait time in between requests due to rate limiting. By registering our data collection script with Reddit, we were granted API access with a rate limit of 60 requests per minute, or 1 request per second. As a result, 6000 posts were hydrated per minute, a very reasonable rate given the sizes of the networks we were trying to create.

4 Mathematical Models

In this section, we will discuss the major points of consideration in our proposed closeness index and how we will fit them together. All metrics discussed below are obtained using the k-NN Network Model with $k = k^*$.

4.1 Spread

One crucial component of closeness is to measure the outward spread of topics within the subcommunity. The main metric we want is the edge distance (or number of hops) between nodes u, v with the highest cosine distance of all pairs of nodes. Ideally, the smaller this value the better since it would be indicative of topic "tightness." In the context of the problem, we are evaluating spread as how well the graph connects the two furthest apart topic vectors. In the worst case, the spread is the diameter which shows us that there is a substantial amount of topics between u and v; indicative of poor closeness of topic within the network model. In the best case, u and v are neighbors; indicative of well-contained topic spread.

In the context of extreme examples, a complete k-NN Model graph with n nodes and $k = k^*$ (in which we know the topics are highly related) has all $u - v$ distances as 1 since the diameter is also 1. On the other hand, the worst possible case of topic spread - a linear graph of n nodes long where the two ends correspond to u and v - has a $u - v$ distance of $n - 1$, which is large.

4.2 Connectivity

We explore the structure of the network generated in an effort to understand how homogeneous the subreddit community is. We expect that a subreddit that is homogeneous will have a high expansion. That is, the smallest bottleneck in the k-NN Model graph G for the subreddit should still be quite large, indicating that the worst connected component is still decently connected. In the context of our experiment, this would mean that even the least relevant topic chain is still decently relevant to the other topic chains. On the other hand, a subreddit that has multiple topic chains may exhibit smaller expansion. A clear case of this would be the graph in Figure 2.

Unfortunately, node expansion in a graph is computationally infeasible to calculate, but can be approximated. From the k-NN graph G, we are able to 2-approximate the worst-case expansion of our relevance network using the second smallest eigenvalue

of G's normalized Laplacian matrix (written as $\lambda_2[L_G]$), as per Cheeger's Inequality [1]:

$$\lambda_2[L_G]/2 \leq \phi_G \leq \sqrt{2\lambda_2[L_G]}$$

4.3 Closeness Index

In considering all the points mentioned above, we propose a closeness index that is based on the 2-approximated expansion of the k-NN Model and heavily penalizes high spread. More concretely, we have, for a subreddit with n threads, and associated k-NN Model G:

$$\text{closeness} = \frac{1}{\text{edge-dist}_G(a,b)} * \frac{\phi_G}{\phi_{\text{complete-graph}(n)}}$$

Where a, b satisfies:

$$\underset{(a,b)\in V}{\text{argmax}} \big(\text{cosine-dist}(a,b)\big)$$

And $\phi_G, \phi_{\text{complete-graph}(n)}$ are 2-approximated lower bound node expansions for the k-NN Model G and the complete graph with n nodes respectively, obtained from Cheeger's Inequality.

In interpreting the value, we can first view the components separately. The penalty term for spread is a value that will range from $(0, 1]$ since in the worst case, n is very large and the a, b edge distance is $n - 1$, driving the penalty term down close to 0. In the best case with a dense relevance model, a, b edge distance is 1, resulting in a penalty term of 1 which does not affect the base value.

On the base term, we know the worst possible case of a linear graph exhibits an expansion of $1/(n-1)$ (which shrinks to 0 for large n) since the bottle-necking subset is composed of all nodes in the linear graph minus a single node on either end. Finally, we include the denominator of $\phi_{\text{complete-graph}(n)}$ since the complete graph with n nodes provides the best possible expansion of any graph with n nodes. This denominator normalizes our graph expansion term, forcing the base term to be within the range $(0, 1]$. Finally, we now have that the overall closeness index will be between 0 and 1 where values close to 0 represent bad closeness of topic and values close to 1 represent high closeness of topic.

5 Experimentation

We chose r/politics from January 2010 since the graph size is large enough (around 9,000 threads) to draw significant distributions from while still being computationally feasible to work with. Additional parameters we used in the following modeling experiments include: 20 passes for LDA training, 5 topics per subreddit, and 5 words per topic.

Topic 0: know right www news really
Topic 1: obama party republicans vote democrats
Topic 2: corporations government speech rights law
Topic 3: war military world country bush
Topic 4: government money tax pay taxes

5.1 Threshold Model

One way we evaluated our Threshold Model was by building a network using the r/politics subreddit and analyzing the structure of the result. For the cosine distance threshold, we used the standard deviation of all distances, divided by 5. This resulted in a sizable network with under 9,000 nodes and nodes having an average degree of 1100 for a total of approximately 5 million edges.

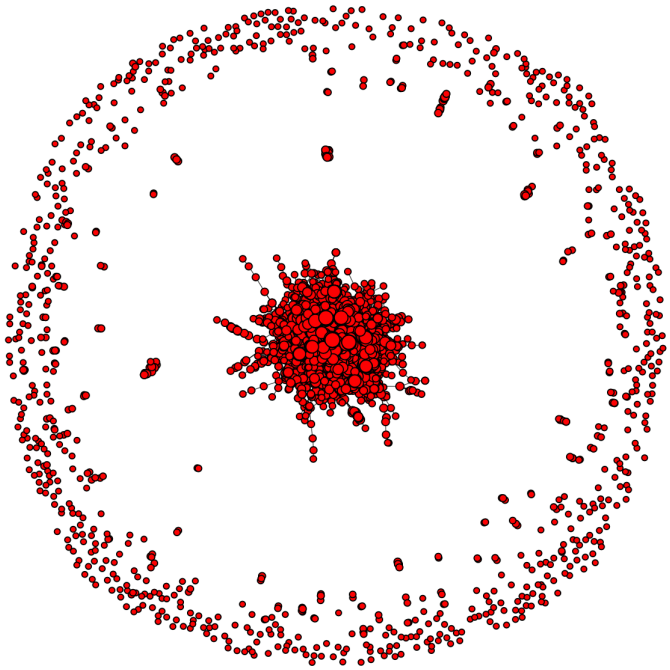

Figure 1: *January 2010 r/politics Threshold model with threshold value = st.dev/5*

We attempted to gain insight into which topics were most prevalent in the subreddit by observing the degree distribution for nodes in the graph. By graphing the distribution for all node degrees, as shown in Figure 3, it became apparent that there were five different curves in the graph, representative of the five different topics generated by the model. The x-axis in the figure represents the degree. We see a large distribution of nodes having a large degree in cluster 5. When taking the mean of the topic vectors that contribute to this cluster, it is apparent the cluster is heavily weighted towards Topic 0. This suggests that Topic 0 is the most prevalent in the subreddit, followed by Topic 1, Topic 3, Topic 4, and Topic 2. The outlier in the graph at the top represents nodes that have no topic, with a vector that weights all the topics equally. Many of the threads had simply too little text to identify a topic, but were

difficult to filter out generally without removing threads that did have meaningful text.

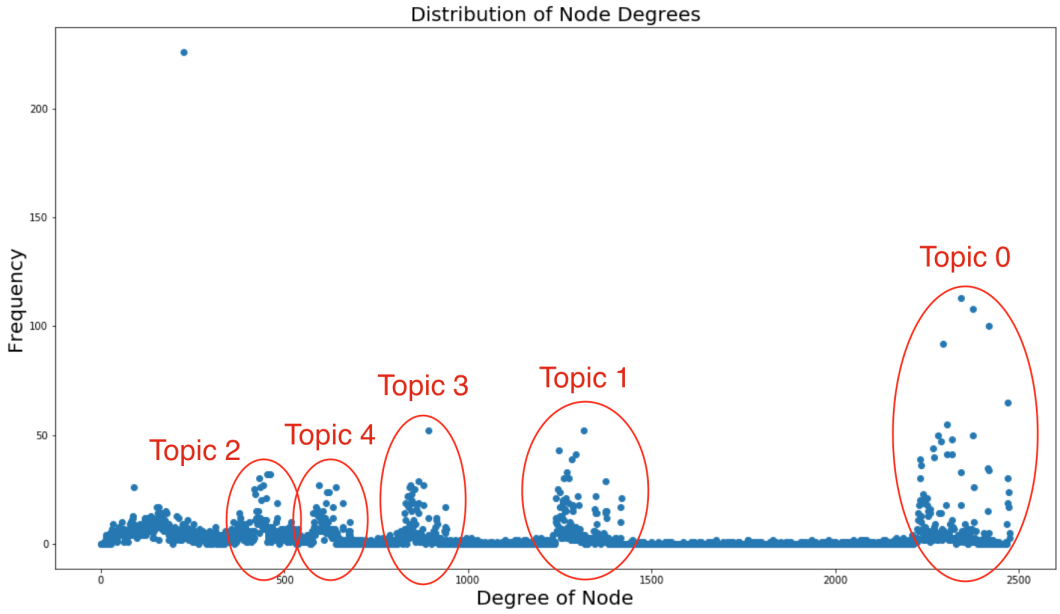

Figure 2: *Distribution of node degree found in the Threshold Model*

However, if we look at the words that Topic 0 consists of, it is difficult to really label the topic with an overarching theme. On the other hand, the other topics are very distinct and specific, having to do with various facets of government and politics. As a result, we are inclined to believe Topic 0 fits as a sort of "other" category, one that is general enough for posts that don't fit the other topics to gravitate towards. A lot of these posts in the category can be considered reactions to new or commentary that happens not to use any indicative words. While it is difficult to distinguish whether or not each post that fit into this "other" should be considered off-topic in the subreddit or not, it is important to note that there exists a large portion of the network that is seemingly unrelated.

5.2 k-NN Model

For the r/politics posts from January 2010, we've also built the k-NN Model for structural analysis (pictured below in Figure 3). Qualitatively, we can tell there are some highly sparse cuts, which is indicative of poor expansion and representative of the data given the background knowledge that r/politics is highly broad. From our structural analysis, we've observed a $\lambda_2[L_G] = 0.0006579811554866333$ which is small and expected due to the highly sparse cuts. Next, we obtain a spread measure (reference section 4.1) of 19 which is relatively large compared to the diameter 28. All of these considerations would lead to a poor closeness rating since it means our topics within the network are not too well connected and span out in a significant manner.

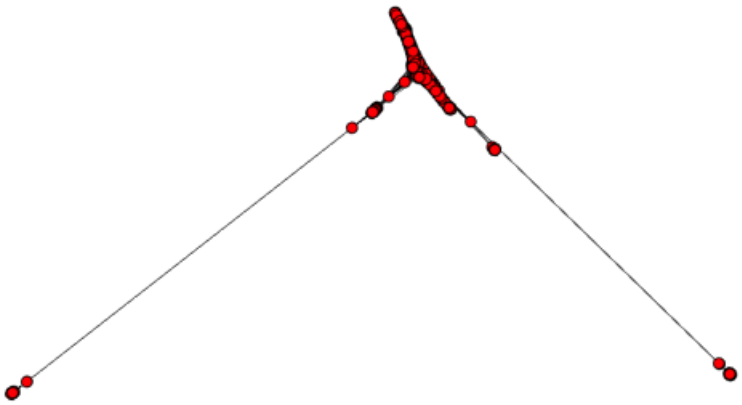

Figure 3: *January 2010 r/politics k-NN Model with* $k^* = 23$ *in spectral layout to emphasize sparse cuts*

5.3 Closeness

From the case study presented above with r/politics January 2010, we obtain a closeness index:

$$\text{closeness} = \frac{1}{\text{edge-dist}_G(a,b)} * \frac{\phi_G}{\phi_{\text{complete-graph}(n)}} \tag{1}$$

$$= \frac{1}{19} * \frac{0.0006579811554866333 * 0.5}{1.0001167269754219 * 0.5} \tag{2}$$

$$= 0.000034627 \tag{3}$$

Which is very small, indicating that r/politics has poor topic closeness.

In comparison, r/bicycling, as a smaller subreddit that also isn't relatively dense, has a closeness index of 0.08504941515215583 and with the following k-NN network:

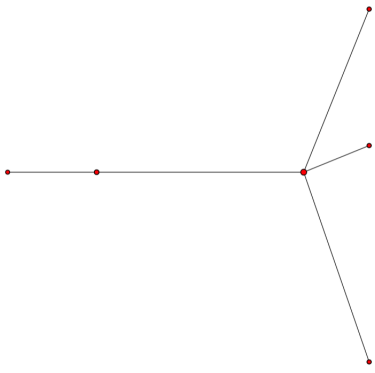

Figure 4: *January 2010 r/bicycling k-NN Model with* $k^* = 2$ *in spectral layout to emphasize sparse cuts*

6 Weaknesses

Our networks are built using only data scraped from Reddit as training data for the LDA topic modeling algorithm. As a result, our analyses based on these topic models is only as accurate as LDA was when discovering the topics. This dependency is a large weakness of our model in some specific subreddits where the content of the thread points largely to external sources. An example of this would be r/politics, in which many of the original posts are simply links to new articles. We mentioned earlier in the paper that URLs are ignored as part of preprocessing our data, and as a result we can only infer the topic of the original post based on the comments left on the thread.

7 Future Work

7.1 Image-Based Topic Modeling

Due to the time constraint of the project and the difficulty of the subproblem involved, we currently don't have infrastructure to support analysis of topics from images. As of right now, we are focusing on text-based posts. In order to add and independently consider imagery for topic modeling and prediction, we could use a form of image annotation and then use topic modeling algorithm on these annotations. While this method might be promising, the overhead cost of image annotation is significantly high. Recent work has looked into using learning methods to automate image annotation. Yavlinsky et. al. looked into the problem of automated image annotation using features such as global color and texture distributions. The paper's results are significant and show that even with simple models, it is possible to achieve decent image annotation.

7.2 k^* Reconciliation

A current weakness of the project is that k^* might vary across different subreddits since it is related to the size of the subreddit as well as the true underlying relevance structure. Future extensions of our work can explore how to resolve that difference by normalizing out the effect of k^* on all measures extracted from k-NN model. Smaller k^* values indicate a tighter true relevance structure as mentioned previously in this paper. A proposed method to reconcile the differences in k^* would be to compare the all other measures against k and keep k as a free variable.

7.3 Multicommunity Model & Modularity

Also due to constraints on resources, we were unable to further explore implications of topic homogeneity on models that include multiple subcommunities. A proposed experiment would be to graph out multiple subreddits using a k-NN method and proceeding by computing the modularity of subsets (where each subset is composed of all threads within an individual subreddit) within the context of the entire relevance network. An ideal Reddit relevance network would have very high modularity

for all subsets (or might even be disconnected) since our relevance models should theoretically preserve the true relevance structure well.

References

[1] Alon, Noga. *Eigenvalues and Expanders*. Combinatorica 6 (2), 1985.

[2] Blei, David M.; Ng Andrew Y.; Jordan Michael I. *Latent Dirichlet Allocation*. The Journal of Machine Learning Research 3, 2003.

[3] Tan, Chenhao; Lee, Lillian. *All Who Wander: On the Prevalence and Characteristics of Multi-community Engagement*. Proceedings of WWW, 1056, 2015.

[4] Horne, B. D.; Adali, S. *The impact of crowds on news engagement: A reddit case study*. arXiv preprint arXiv:1703.10570., 2017.

[5] Yavlinsky, A.; Schofield, E.; Ruger, S.. *Automated image annotation using global features and robust nonparametric density estimation*. Image and video retrieval, 593, 2005.