Giving Gold: Understanding Appreciation in Reddit Communities

Julia Mendelsohn[1,2] **Lucy Li**[1,3]

Stanford [1]Computer Science, [2]Linguistics, and [3]Symbolic Systems
CS 224W, Fall 2017
{jmendels, lucy3}@stanford.edu

1 Introduction

In the social media age, our actions are constantly evaluated by other users, many of whom are strangers. While anonymous users are often associated with adversarial behaviors, such as scamming and trolling, they are also capable of positive interactions. Community appreciation, a positive public evaluation of an individual's content within a community, is found in nearly every social network; Facebook "likes", Reddit "upvotes", and Twitter "retweets" for example are all indicators of community appreciation. The goal of this project is to understand what social and linguistic factors contribute to community appreciation behavior, specifically on Reddit, and how these factors differ based on the specific form of appreciation as well as the type of community.

Because of its pervasiveness, community appreciation is an important social phenomenon that can profoundly impact users. For example, the extent to which a user is appreciated in an online community could affect their sense of worth and belongingness, which could have psychological implications. Furthermore, this line of work could shed light on strategies that may be employed in order to resonate with a largely anonymous group, which can have massive industrial, social, and political implications. Additionally, we hope to contribute to advances in information network research. Because social networks curate and suggest content to users based off various measures of a submission's appreciation, information from highly-appreciated contributions can spread more widely and quickly throughout the network. Understanding why content becomes appreciated in the first place is essential for explaining this process of information diffusion.

In this paper, we compare two different appreciation behaviors on the content-aggregation and discussion website Reddit. First, we examine the factors that impact a contribution's **score** (number of upvotes - downvotes). Second, we consider **gilding** as a behavior indicative of community appreciation. Gilding can be viewed as the highest demonstration of appreciation among users; this occurs when one user pays real money ($3.99 per gild) to another user, usually in response to their submission[1]. This gives the contributor a month's worth of "Reddit gold", or premium membership and access to additional website features.

Reddit stands out from most discussion websites with its wide spread of sub-communities, called "subreddits", which differ in topic and style of sharing. We hypothesize that the linguistic and social factors that contribute the most to community appreciation differ across sub-communities. For example, the factors leading to a gilded post in science-related subreddits may be quite different from those in a sports-related subreddit.

In order to examine how community appreciation behavior, specifically gilding, differs across different types of communities, we must first determine what different groupings of communities exist. Thus, an intermediate goal of our study is to create a network of subreddits that appropriately represents the relationships between them, using measures of user overlap and text similarity. Given these graphs, we use community detection to identify groups of related subreddits, which we then use to assess our hypothesis.

In summary, the two research questions below drive our investigation into how appreciation operates on Reddit:

> **RQ1:** What social and linguistic features predict upvoting and gilding behavior?
>
> **RQ2:** How can we extract groups of related subreddits, and how does gilding behavior differ between these groups?

2 Related Work

The study of user interactions on the Internet pulls methods and theories from social psychology, natural language processing, and network science. This project is closely related to prior works on Reddit, many of which aim to predict the extent to which a submission is appreciated by the community.

For example, Althoff et al. (2014) studied the social, linguistic, and temporal factors that contribute to the success of altruistic requests from the

[1]https://www.reddit.com/gilding/

sub-community "Random Acts of Pizza". Satisfying the altruistic requests, where the asker does not offer anything immediately in return, is one form of community appreciation. Their model textual features that captured post length, sentiment, politeness, and reciprocity. They also use topic modeling techniques and vocabulary from Linguistic Inquiry and Word Count (LIWC) lexicons (Pennebaker et al., 2001) to build concise lexicons for five narratives ("money","job","family","craving","student"), which were also included as features. Social features included user "karma" (aggregated upvotes - downvotes over all submissions), if a user posted on that particular subreddit in the past, the user's account's age, and similarity between community members. Finally, the authors accounted for temporal effects. They found that gratitude, reciprocity, evidentiality, status, and narrative type were all significant in predicting the success of an altruistic request.

Although it focuses on another form of appreciation Althoff et al. (2014)'s work is closely related to gilding, since gilding is also an altruistic behavior. We are thus heavily inspired by this work and incorporate some of their features, such as length, time past, sentiment, and similar measures capturing user reputation and loyalty. Although we do not yet include narrative lexicons, we are also inspired by this paper's use of Linguistic Inquiry and Word Count (LIWC) features and include them into our model.

In a slightly earlier work, Lakkaraju et al. (2013) used Reddit image resubmissions to understand how the title, community, and time impact the popularity of a post. They consider three aspects of a post's popularity: score (upvotes - downvotes), engagement (number of comments), and attention (upvotes + downvotes). First, they fit a "community model" to the data, which accounts for the specific community, the time of day of submission, and the number of times the image has been submitted. They subsequently fit a "language model" in order to model the title's specific impact on the popularity of a submission.

A key finding was that a title of a post should match the style of the community to which it is submitted, but should also be distinct enough to be memorable. We draw inspiration from the social and linguistic factors discussed by Lakkaraju et al. (2013), specifically with its emphasis on the impact of stylistic linguistic factors, such as a submission's distinctiveness, on a post's popularity. Furthermore, we note the subtle distinction between Lakkaraju et al. (2013)'s goal of predicting post popularity and our goal of predicting community appreciation. While there is certainly overlap and prior work on popularity prediction is relevant, we consider appreciation to be restricted to positive interactions. Thus, a post's score is indicative of both popularity and appreciation, but engagement and attention are not necessarily measures of appreciation.

In the work most closely related to our own, Jaech et al. (2015) used a support vector machine with lexical and social features to predict a comment's score, which is an appreciation measure that we also study. These features included the comment's informativeness, relevance to the comment thread, the author's reputation, and the structure of replies and discourse history. Instead of predicting a comment's score directly, Jaech et al. (2015) classify comments into buckets that represent the relative ranking of all responses to a given post by score; we adapt this method into our own classification task. The authors focused on six subreddits and found that the factors that influence a high-ranking comment depend on the subreddit. For example, an author's reputation impacts comment score in the "AskScience" community, but not in others. Much of our experimental design and classification task were influenced by Jaech et al. (2015). We extend their work by including features from other related works and by comparing differences in the factors that influence two different appreciation measures, scores and gilds. We also further characterize the intra-community variability in community appreciation by generalizing to groups of related subreddits, rather than individual ones.

Community appreciation has also been studied in other social networks. Hong et al. (2011) predicted if a post on Twitter would be retweeted (a measure of appreciation), by using linguistic, temporal, metadata, and social graph features. Anderson et al. (2012) studied how similarity between two users could affect the evaluation that one user provides of another. By using Wikipedia, Stack Overflow, and Epinions as case studies, the authors found that evaluations tend to be less status-driven when users are more similar to each other. Danescu-Niculescu-Mizil et al. (2009) analyzes helpfulness votes on Amazon.com reviews, and found that the perceived helpfulness of a review depends not only content, but also on how it relates to other reviews of the same product.

These papers present useful features for classification and novel methodologies which have influenced the present work. We hope to further contribute to this area of research by analyzing gilding as an appreciation behavior, as it has to the best of our knowledge, not been studied. We also aim to more broadly understand how different sub-communities relate to each other, and how appreciation behavior varies between groups of related sub-communities.

3 Data

Our data consists of user-created content on Reddit during May 2015[2]. Our data is separated into two components. One contains posts and their metadata, including information on gilding, score, subreddit, and author. The other contains comments and their metadata, which is similar to posts' metadata but also

[2]We chose this month because it is also found publicly on Kaggle, which would allow others to replicate our work.

includes parent comment and parent post information.

On Reddit, many post submissions contain images and URLs, and as a result tend to contain few words. Comments, however, are more linguistically and socially rich (with many layers of replies), which makes them more suitable for our study. We thus focus our study on only comments on May 2015 submissions.

The majority of subreddits contain very few comments, while a small number of subreddits contain thousands of comments. We decided to select the top 100 subreddits with the most comments, as they contain vastly more information than a randomly selected sample of subreddits. AskReddit is the most active subcommunity, and the remaining 99 subcommunities span a wide variety of topics, including gifs, personalfinance, unitedkingdom, and CasualConversation. One of the top subreddits by number of comments, fatpeoplehate, was banned, and some of its data could not be accessed, so we excluded this subreddit.

To improve retrieval times for different types of data to $O(1)$, we restructured the original dataset to contain nested dictionaries keyed by subreddit and post. We also limited our comments to those on posts also submitted in May 2015 so that we would have access to the entire parent tree of each comment we studied. Our final dataset contains 25,300,306 comments, of which 10,403 are gilded.

4 Methods

4.1 Research Question 1

To identify factors involved in appreciation behavior, we create two comparable binary classification tasks. In one task, we predict whether a comment is gilded or not. In the other, we predict whether a comment has a high or low score (difference between upvotes and downvotes) relative to other comments to the same post. This allows us to see how an interaction unique to Reddit (gilding) differs from an interaction more common across social media platforms (voting). We compare the features that are most predictive of each type of behavior. We divide our features into two larger categories, one focusing on the lexical content of a comment, and another other social and structural aspects of a comment.

Classification Task Setup

We treat gilding as a binary variable for each comment, since an overwhelming majority of gilded comments (94.6%) are gilded only once. Since we hope to understand how gilding differs from voting, we transformed comment score into a binary metric as well. For this, we are mostly interested in what differentiates highly rated comments and those that are not. We split the comments on each post with more than four comments into quartiles, ranking them by score, and classified comments found in the top 25% and bottom 25%. Since we are predicting the relative ranking of comments by score, rather than the comment score itself, we refer to this measurement as **rank**, and the task as rank classification.

The rank classifier had an even ratio of top comments and bottom comments in both the train and test sets, containing 25,000 randomly sampled comments total. For our gilding classifier, we utilized all of our gilded comments and a randomly sampled subset of nongilded comments. We balanced the training set (8842 gilded, 8842 nongilded) so that the model would see enough examples to learn, and its test set was designed to have an unbalanced 1:5 ratio between gilded and nongilded comments (1561 gilded, 7805 nongilded), since there are more nongilded comments in the overall dataset. In total, our gilding model trained and tested on 27,050 comments. Our train and test sets for both gilding and rank classifiers had a 85%-15% split.

We chose linear support vector machine (SVM) and Random Forest as our classifiers because they allow us to clearly identify the impact of each feature. Random Forest provides feature importances based on how features are weighed in decision trees, while SVM provides a more fine-grained differentiation between positively weighted features predictive of one class and negatively weighted ones predictive of the other class. We tuned our Random Forest classifier to have 500 estimators, 5 minimum samples per leaf, and 3 minimum samples to split internal nodes. Our SVM classifier used a penalty parameter of 14, hinge loss, and a stopping tolerance of 0.01. Both the SVM and Random Forest classifiers were implemented in Python with scikit-learn [3]

Lexical Features

We preprocessed our lexical content by setting all words to lowercase and removing non-apostrophe punctuation. In features involving bigrams, we also tokenized comments by sentence so that bigrams are isolated within sentences.

- **Comment length.**
- **LIWC:** These features are normalized counts of words and prefixes from 64 categories that capture several social and psychological dimensions of meaning (Pennebaker et al., 2001; Tausczik and Pennebaker, 2010). These categories include sentiment, emotion, parts of speech, negation, family, pronouns, and cognitive processes[4].
- **Relevance:** This aspect is broken down into two features: a comment's relevance to its parent comment and its relevance to the post and post title. We used Jaccard similarity, treating content as a bag-of-words and removing NLTK English stop words. A similar relevance feature was also used by Jaech et al. (2015).
- **Style:** This is measured as the posterior probability of a comment appearing in its

[3] http://scikit-learn.org/stable/

[4] LIWC is not open-source.

subreddit, treating the comment as a set of words. The probability of a word in each subreddit is the number of comments that word appears in divided by the total number of comments in that subreddit. This feature also parallels one used by Jaech et al. (2015).

- **Distinctiveness:** For each comment, we compute its likelihood using a "common language" corpus, such as the Brown corpus. The lower the likelihood, the more distinctive the comment. This relates to a "memorability" metric described in Danescu-Niculescu-Mizil et al. (2009). Our language model calculated unigram and bigram probabilities with Laplace smoothing, where V is the number of unique words, $c(x)$ is the count of n-gram x, and $sum(c)$ is the total count of all words:

$$p(w_i|w_{i-1}) = \frac{c(w_i, w_{i-1}) + 1}{c(w_{i-1}) + V},$$

$$p(w_i) = \frac{c(w_i) + 1}{sum(c) + V}.$$

Social Features

- **Author's status:** Unfortunately, our data does not contain information from user profiles, so we cannot access the author's "karma" (aggregated score over a user's lifetime). Instead, we estimate an author's status, with the total score over all of the author's May 2015, after excluding the comment whose features are currently being extracted. Because our task involves prediction of a comment's score, it is crucial to our model's integrity that the same comment's score is not used as a feature.
- **User loyalty to community:** We consider two measurements as features of user loyalty to the particular subreddit where the user's comment is found. We take the total number of comments made by the user in that specific subreddit. For the first feature, we normalize by the total number of comments made by the user (this feature will be referred to as *loyalty (user)*). For the second, we normalize by the total number of comments in the subreddit (this feature will be referred to as *loyalty (subreddit)*).
- **Popularity of parent comment:** The score of the comment's parent (the comment or post that the comment being considered directly replies to).
- **Time elapsed since original post:** Because time is measure in seconds, this number is often quite large, so the corresponding feature is divided by a factor of 1000.
- **Distance from original post:** This feature captures the comment's depth in a post's comment tree. Comments that are an immediate reply to the original post have a distance of 1, those that are replies to an immediate reply have a distance of 2, and so forth.

We seek to best replicate factors that exist at the time that a comment is published. We thus intentionally did not include features that rely on information from after the comment's submission, such as the number of replies, as they could bias our classifier unfairly.

4.2 Research Question 2

After looking at gilding behavior overall, we want to understand how it operates on a more local level, using a networks perspective. Previous work such as Jaech et al. (2015); Althoff et al. (2014) examined behavior in individual subreddits, but we are also interested in how connections between subreddits relate to appreciative behavior. We use two types of similarity between subreddits, one based on topic and another on user overlap, to create two different networks. We then partition each of these networks into larger communities, which represent groups of related subreddits.

We posit that the subreddits in each community share commonalities in appreciation behavior. We examine what features tend to be shared across communities and which are not, and hypothesize that our results will indicate that there are distinct cultures of behavior within Reddit.

Creating a Network of Subreddits

To determine how appreciation behavior differs across communities, we must first define how subreddits relate to each other.

To create a network based on topic similarity, we measure word usage in post titles. We treat each subreddit as a collection of post titles from posts created during May 2015. For each subreddit, we create vectors with each element representing a term weighted by term frequency-inverse document frequency (tf-idf) with smoothing, removing stop words:

$$w_{td} = (1 + \log tf_{td}) \log(1 + N/df_t),$$

where t is a term, d is a document, tf is term frequency, df is document frequency, and N is the total number of subreddits. These vectors are normalized and reduced to 50 dimensions using latent semantic analysis (Landauer et al., 1998). The similarity between two subreddits' vectors X and Y is represented as $\cos(X, Y)$, which is then used to weight edges in the subreddit topics network.

To create another network of subreddits based on commenter overlap, we weight our edges based on the Jaccard similarity coefficient between the commenter sets, X and Y, of two subreddits:

$$J(X, Y) = \frac{|X \cap Y|}{|X \cup Y|}.$$

We use overlapping commenters to create this user overlap graph rather than subscriber data because we did not have access to subscriber data from this time

Metric	Classifier	Accuracy	Precision	Recall	F1
Gild	RF	0.7685	**0.6636**	**0.7499**	**0.6791**
	SVM	**0.7757**	0.6355	0.6732	0.6478
Rank	RF	0.6493	0.6558	0.6476	0.6440
	SVM	0.616	0.6272	0.6135	0.6043

Table 1: Random Forest and SVM performance on classifying comment rank and gild, with the highest value in each column in bold.

Classifier	Features	Accuracy	F1
RF	Lexical	0.6970	0.5944
	Social	0.6991	0.6181
	Full	**0.7685**	**0.6791**
SVM	Lexical	0.7531	0.6091
	Social	0.7552	0.5931
	Full	**0.7757**	**0.6478**

Table 2: Ablation by feature type for SVM and Random Forest classifiers on gilded task

period. Furthermore, using comment author data specifically from May 2015 gives us a better sense of who is active in the community in the time period we are studying.

Community Detection

We use the Louvain Method for community detection to identify communities in our user-based and topic-based networks (Blondel et al., 2008)[5]. This method optimizes modularity,

$$Q(y) = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{d_i d_j}{2m} \right] I_{y_i = y_j}$$

where y_i is the assignment of a node i to a community, m is the number of edges, A is the adjacency matrix of the graph, and d_i is the degree of node i.

This method is a greedy one, operating by first optimizing modularity locally, creating small communities. It then represents these communities as nodes in a new network, and repeats the process until modularity is not increased further.

Behavior Comparison

Once we identify communities of related subreddits, we run Random Forest and SVM classifiers with the same features as described in section 4.1 on each community. We retune our classifiers to these smaller datasets and changed some hyperparameters so that our SVM has a penalty parameter of 8 and stopping tolerance of 0.05, and our Random Forest has 200 estimators. The minimum number of gilded comments in a community was 900, and we set each dataset for each community's classifier to be the same size. Each training set is balanced with 765 gilded comments and 765 nongilded comments, and each test set has a 1:5 ratio with 135 gilded comments and 675 nongilded comments. This train-test split, like before, is 85%-15%.

We then focus on the topic-similarity network's communities and analyze their feature weights by highlighting commonalities and differences.

5 Experiment 1 (RQ1)

5.1 Results

Table 1 shows performance for our two appreciation types, each modeled by both Random Forest (RF) and SVM. While the SVM outperforms RF in accuracy in gild classification, RF outperforms SVM with respect to F1 in the gild classification task and in both accuracy and F1 in the rank task. Both models have a higher overall performance in the gild classification than rank classification task.

In order to determine the extent to which the lexical and social feature groups impacted the classifiers' performances, we ran an ablation test, the results of which are shown in Table 2. We only show the ablation study results for gilding, but the trend was similar for upvoting (rank classification). Only including either lexical or social features allows our models to perform better than random, but are more powerful when combined in a single model.

The most important features for each behavior are shown in Tables 3 and 4, respectively. There is considerable overlap between important features in rank and gild classification, particularly in the Random Forest model. With this model, the only difference is that distance from the original post is a more important feature for gild classification, while cognitive mechanisms (a category with words such as "know" or "cause") is more important for rank.

A more drastic difference in features' importances emerges from the SVM classifier's feature weights, shown in Table 4. For the rank task, a more positive weight indicates that a feature is more predictive of the comment being highly ranked in score relative to other comments responding to the same post. Similarly for gild classification, a more positive weight indicates that a feature favors a comment being gilded. Note that the most positively weighted features for gild classification are more highly weighted than those for rank classification.

5.2 Discussion

Although considerably better than random for binary classification, the rank and gild classifiers both leave

[5]We also tried the Girvan-Newman algorithm (Girvan and Newman, 2002). However, the implementation we used resulted in a slower runtime and there did not seem to be a single peak in modularity that would allow us to select the optimal number of clusters.

Rank		Gild	
Feature	Importance	Feature	Importance
time elapsed	0.1413	time elapsed	0.1031
status	0.0501	parent popularity	0.0603
user loyalty	0.0405	user loyalty	0.0585
style	0.0403	style	0.0556
unigram distinctiveness	0.0347	unigram distinctiveness	0.0533
bigram distinctiveness	0.0305	status	0.0529
function words (LIWC)	0.0242	bigram distinctiveness	0.0414
parent popularity	0.0218	comment length	0.0368
comment length	0.0205	distance from post	0.0218
cognitive mechanisms (LIWC)	0.0203	function words (LIWC)	0.0187

Table 3: Random Forest highest feature importances for rank and gild classification

Rank		Gild	
Feature	Importance	Feature	Importance
Most Positive			
bigram distinctiveness	2.6426	distance from post	15.1513
death (LIWC)	2.1932	humans (LIWC)	5.2486
religion (LIWC)	2.0674	status	4.3129
filler words (LIWC)	1.939	family (LIWC)	4.2987
unigram distinctiveness	1.8842	health (LIWC)	4.2047
Most Negative			
negation (LIWC)	-1.2686	hear (LIWC)	-2.169
parent popularity	-2.4969	post relevance	-3.2106
comment length	-5.3841	unigram distinctiveness	-5.8211
style	-8.0313	style	-19.1528
time elapsed	-66.8298	time elapsed	-36.2034

Table 4: SVM highest magnitude feature weights for rank and gild classification.

room for improvement. One issue may be that we simply did not give the models enough data, which hinders their ability to generalize. For computational feasibility, we limited the amount of data to one month, and we sampled a small number of comments for rank classification so the overall size of the train and test sets were similar for both tasks. In future work, we will ensure that we have enough gilded comments by using data from a longer period of time. More data would also allow us to include more features.

Both gild and rank classification are non-trivial tasks. Since gilding is so rare relative to other comments (only about 0.04% of May 2015 comments are gilded), there are a plethora of comments that may be highly appreciated, but are not gilded. Gilding depends on only one user's action, and users may appreciate different aspects of comments. So, the factors that drive gilding may be difficult to characterize on a large scale, across entire communities.

Rank classification also presents challenges. We classified rank instead of score directly in order to normalize for a post's popularity. However, it is possible that all comments responding to the same post have a similar score. For example, many posts have comments that all have scores of 0 and 1. Even after splitting rank into quartiles (arbitrarily breaking ties) and only predicting the top and bottom quartiles, the distinctions may be extremely subtle or even nonexistent in some cases.

In addition, both upvoting and gilding behaviors depend on factors that are not included in our models. For example, both of these behaviors could, depend on factors such as what other content is on Reddit that day, how many users are active when the comment is submitted, and other macro-level features. It would be interesting to incorporate these into future work.

Another reason for low classifier performance is the importance of context in community appreciation. Both gilding and upvoting behaviors are likely to reflect the current context of the world, and are deeply embedded in the common knowledge held among community members. Some strategies that may be important for community appreciation, such as humor and sarcasm, operate successfully within this body of common knowledge between users. However, these strategies are difficult to detect automatically with relatively simple features and models. Below are two examples of gilded posts that rely on common knowledge and humor. One requires knowing cultural norms around dating, and another is only humorous if one knows who Oprah is.

"I'm a woman who loves making the first move for two reasons. 1.) Because I believe if someone wants something, they should go after it. Don't pussydick around and wait for it to come to you. You're never going to get anywhere in life if you sit around on your ass waiting for things to happen. And 2.) because I find that it weeds out the guys who are

emasculated by strong women. I don't have time for that shit, and making the first move does a good job of getting rid of the idiots that are offended when I want to buy them dinner." - AskReddit.

"Different people fear different things. For me, its when someone gets up real close to me, puts their hand on my shoulder and says, 'Have you accepted Oprah Winfrey as your personal Lord and Savior?'" - AdviceAnimals.

As noted in the results section, there is considerable overlap in the Random Forest models' important features. This overlap in features is unsurprising considering upvoting and gilding are both forms of positive feedback. In Figure 1, we see that gilded comments tend to be in the top 50% in terms of score, while nongilded comments tend to be in the bottom 50%.

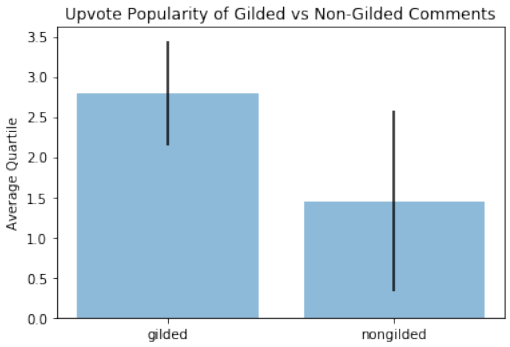

Figure 1: The average rank quartile of gilded and nongilded comments. The higher the y-axis value, the higher the rank. Error bars represent standard deviation.

The Random Forest feature importances suggest that many social features, such as elapsed time, user loyalty, and status influence both upvoting and gilding behavior. However, the relative difference in feature weights indicate that some features that are predictive of one appreciation behavior are not equally predictive of the other, which suggests that gilding and upvoting behaviors are not identical. The most highly weighted features from the SVM model, which is more sensitive to features with smaller signals, further emphasizes this observation.

The positively weighted LIWC features from the SVM model in gild classification are particularly interesting. These features include "humans" (which includes 'baby' and 'boy'), "family" (which includes 'daughter' and 'husband'), and "health" (which includes 'clinic' and 'pill'). Together, these features suggest that gilded posts tend focus on people's personal life stories, which are likely to invoke these features. Although not in the top five, the "death", "religion", and "home" LIWC features are also highly weighted (with weights > 2.5). If we take these LIWC features to be represented as topics discussed in the comments, it further supports the idea that gilded comments focus more on topics related to people's personal lives, such as relationships and issues in the private sphere. Personal pronouns (such as "I" and "we"), which are also likely commonly used in personal narratives, are also a highly weighted category. In the gilded example below, gilding is likely a demonstration of support in response to one's personal story.

"Losing a parent as a teenager. My mom went through three bouts with cancer in her life, once before I was born, once when I was about eight, and one final time when I was about twelve. Watching the parent you've always looked up to as strong and untouchable, slowly waste away in a bed really puts life into perspective. I'm glad that she doesn't have to suffer anymore, and I do think that the experience has made me a stronger person, but damn I miss my mom." - AskReddit.

The most positively weighted features from the SVM rank classifier have considerably lower weights than those from the gild classifier. Perhaps this means that, unlike with gilded posts, there aren't any features that are very strongly indicative of a highly ranked comment. One interesting observation is that both unigram and bigram distinctiveness are among the five most positive features for rank classification. This could suggest that in order to be highly ranked, it is important for to be unique relative to other comments. Like with the gild classifier, the "death" and "religion" LIWC features are among the top most positive features. This could suggest more personal stories, but it is also possible that the words in these categories are invoked in more ideological discussions or in response to current events.

Finally, we note that our results may be skewed by subreddits that dominate the number of gilded comments, such as AskReddit. Thus, our analysis may be unbalanced by these subreddits' cultures and main topics of interest. In order to account for this skewed data, we need to analyze the classifiers and their feature importances when run on different types of communities within Reddit. We thus turn to our second experiment, where we find groups of related subreddits, and extract the most important features for each group.

6 Experiment 2 (RQ2)

6.1 Results

Networks and Community Detection

Visualizations of our networks can be found in Appendix A.

In our user-overlap network, the maximum Jaccard similarity coefficient is 0.2476, which means that at most 24.76% of the users of two subreddits are in both. The user-overlap network contains a large connected component that includes some of the most popular subreddits, such as "AskReddit", "funny", and "pics". This connected component contains subreddits that vary in topic, community cultures, and formats. For example, "gifs" and "pics" rely on visuals, while "todayilearned" and "worldnews" tend to be based

	User	Topic
1	2007scape Android DestinyTheGame DotA2 EliteDangerous Eve Fireteams Games GlobalOffensive GlobalOffensiveTrade Guildwars2 KotakuInAction PS4 Smite TumblrInAction amiibo anime bloodborne buildapc csgobetting electronic_cigarette ffxiv fivenightsatfreddys hearthstone heroesofthestorm leagueoflegends magicTCG newsokur pcmasterrace pokemontrades smashbros thebutton whowouldwin witcher wow xboxone	2007scape Android DestinyTheGame DotA2 EliteDangerous Eve FIFA Fireteams Games GlobalOffensive GlobalOffensiveTrade Guildwars2 PS4 Smite amiibo anime bloodborne buildapc cars csgobetting electronic_cigarette ffxiv fivenightsatfreddys gaming hearthstone heroesofthestorm leagueoflegends magicTCG motorcycles newsokur pcmasterrace pokemontrades smashbros thebutton witcher wow xboxone
2	AdviceAnimals AskReddit IAmA Music Showerthoughts WTF aww explainlikeimfive funny gaming gifs mildlyinteresting movies news nottheonion pics television tifu todayilearned trees videos worldnews	TrollXChromosomes AdviceAnimals TwoXChromosomes AskMen AskWomen TumblrInAction CasualConversation relationships Fitness KotakuInAction tifu OkCupid gonewild Random_Acts_Of_Amazon SubredditDrama
3	AskMen AskWomen Bitcoin CasualConversation Christianity Fitness OkCupid Random_Acts_Of_Amazon SubredditDrama TrollXChromosomes TwoXChromosomes atheism canada conspiracy fatpeoplehate gonewild india personalfinance politics relationships rupaulsdragrace science technology	AskReddit Bitcoin Christianity IAmA Music Showerthoughts WTF atheism aww canada conspiracy europe explainlikeimfive funny gifs hiphopheads india mildlyinteresting movies news nottheonion personalfinance pics politics science technology television todayilearned trees ukpolitics unitedkingdom videos whowouldwin worldnews
4	motorcycles nba asoiaf CFB baseball cars FIFA nfl europe soccer hiphopheads formula1 MMA gameofthrones survivor ukpolitics unitedkingdom SquaredCircle hockey	nba asoiaf CFB nfl baseball gameofthrones rupaulsdragrace soccer survivor formula1 MMA hockey SquaredCircle

Table 5: The communities found in the user overlap network and the topic similarity network.

on text information. One explanation for this trend is that in the past, Reddit had a set of default subreddits that users are automatically subscribed to when they first join the site[6]. The relatively small Jaccard similarity is because we measure user overlap by commenting activity, not passive subscription of users. There is some indication of user overlap based on topic, such as highly weighted edges between the sports-related subreddits "nfl", "hockey", "nba", "CFB", and "basketball".

The topic-similarity network (based on word usage in posts) also features a large connected component, as well as a smaller connected component that includes game-related subreddits, such as "leagueoflegends" and "GlobalOffensive". The similarities between the large connected components for these two networks suggest that there is a relationship between user overlap and topic similarity. In Figure 2, we see that this correlation is quite high, so subreddits with similar topics tend to share users.

These similarities also extend to the communities we detect on each network. The Louvain method for community detection yielded four communities for each network, shown in Table 5. The modularity of the partitions in the user-overlap graph was 0.1283, while the modularity of the topic-similarity graph was 0.0556. In both networks, there appears to be a gaming community (Community 1), consisting of subreddits such as "xboxone", "gaming", and "PS4". Another community (Community 4) seems to be more focused on sports, with subreddits such as "nba","baseball", and "soccer". The other communities, 2 and 3, are more mixed for both subreddits, but could potentially be characterized as information and lifestyle/advice subreddits. Communities 2 and 3 for both networks, where the motivation for the subreddits' relations is less clear, also tend to include the default subreddits.

[6]We found a list corresponding to March 2015, which is relatively close in time to the data we focus on from May 2015.

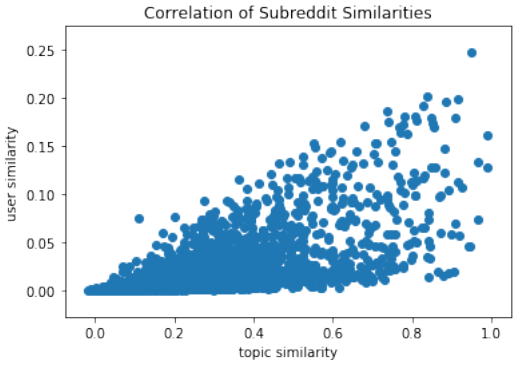

Figure 2: Each point represents two subreddits, and the x-axis value is their topic similarity and y-axis value is user similarity. The Pearson correlation between the two types of similarities is 0.6530 ($p = 0.0$).

Community Type	Classifier	Accuracy	F1
User Overlap	RF	0.7284	0.6298
	SVM	0.7293	0.6069
Topic Similarity	RF	**0.7309**	**0.6351**
	SVM	0.7133	0.5903

Table 6: Accuracy and F1 scores aggregated over communities based on topic similarity and user overlap for Random Forest and SVM classifiers predicting which comments are gilded.

Community Gilding

Even though their modularity differences indicate that it is more difficult to cluster the topic-similarity graph than the user-overlap graph, the topic communities had similar SVM and Random Forest gilding classifier performance, shown in Table 6. For each graph, the average accuracy and F1 score are computed over all four communities.

Because the two different graphs yielded communities with similar classifier performance, we arbitrarily chose to further analyze the features

Community	Acc	F1	Positive Features	Negative Features
1	0.7605	0.6092	comment length, home*, distance from post, work*, parent popularity, humans*, biological processes*, social*	unigram distinctiveness, time elapsed, bigram distinctiveness, health*, achievement*, leisure*, hear*
2	0.6778	0.5718	comment length, articles*, inclusive*, humans*, distance from post, social*, religion*, adverbs*	time elapsed, leisure*, style, friends*, work*, we*, conjunction*, achievement*
3	0.7062	0.6089	status, comment length, health*, adverbs*, perceptual processes*, anger*, swear*, nonfluencies*	time elapsed, style, biological processes*, anxiety*, post relevance*, positive emotion*, assents*, hear*
4	0.7086	0.5717	present tense*, death*, status, home*, comment length, ingestion*, you*, parent popularity	style, time elapsed, see*, filler*, sad*, verbs*, nonfluencies*, conjunctions*

Table 7: Accuracy, F1, and most important features for each of the four superreddits found in the topic-similarity network from an SVM classifier predicting if a post is gilded. Features marked with a * are LIWC features.

corresponding to the model run on the topic-similarity graph's communities. Conducting a similar analysis for the user-overlap graph may provide interesting insights, but is left for future work. Furthermore, for this project, we only analyze the SVM classifier's most important features, since the positive and negative weights provide us with more nuance than those of the Random Forest classifier.

Table 7 breaks down the features determined to be positively weighted and negatively weighted for the communities with similar topics. Comment length emerges a highly weighted positive feature for all communities and time elapsed is highly negatively weighted for all communities. There is more variance in the LIWC features (shown in Table 7 with asterisks), with many LIWC features being highly positive or negative for just one or two communities.

6.2 Discussion

The subreddit graphs produced by user overlap and by topic similarity look relatively similar, and most of the highly-weighted edges between subreddits make sense intuitively given our knowledge about the subreddits. For example, in Figure 4 (in Appendix), highly weighted edges exist between gaming related subreddits, and others exist between politics and news subreddits. It is difficult to evaluate the quality of these graphs, as there is no objective evaluation metric.

The community detection results partially correlate with our intuitions as well, but are also some surprising results from both graphs. For example, "unitedkingdom" and "ukpolitics" are both found in a community with mostly sports subreddits in the user overlap graph. Similarly, "electronic_cigarette" and "cars" are found in a community with mostly game-related subreddits in the topic similarity graph. Furthermore, there is not a very clear way to characterize communities 2 and 3 in either graph. This noise may arise in the user overlap graph because many of the subreddits are default ones for all users. Furthermore, the subreddits in communities 2 and 3 tend to be very diverse in potential topics (such as "AskReddit"), so creating edges based on word usage in post titles could be noisy and lead to communities that are not necessarily intuitive.

We used our classifiers from Experiment 1 to predict gilded posts in each of the four communities in each graph. We found that the average accuracy and F1 over all communities in both graphs was slightly lower than the accuracy and F1 from the Experiment 1 classifiers, which classified all of the data at once. It is possible that this slight decrease in performance is due to a smaller dataset, since there are only 900 total gilded comments that were selected from each community (900 was the minimum number, so we balanced it over all communities). These 900 gilded comments were then further split between the training set and the test set. Since this classification task is the same as that discussed in Experiment 1, the same issues are still relevant here.

We then extracted the SVM classifier's most important features when run on the topic-similarity graph's communities, and found that comment length is positively weighted for all four communities, while time elapsed is negatively weighted for all communities. This suggests that long comments submitted shortly after the original post are more likely to be gilded. The most interesting differences between communities are in the positively weighted LIWC features. For Community 1 (mostly related to gaming), "home", "work", "humans", and "social" (words related to interactions, like "friend" and "talk") are highly positively weighted. This can be seen in the gilded example below:

"Nobody is forcing you not to have fun. Its not about the gp/hr itself, its about the max amount of cash you can make period. 4 hours at a fun boss with your friends at 500 kgp/hr > 2 hours of grinding zulrah.. any day of the week. Why? Because youre having fun. 12 have fun on this game stop looking at it from not doing zulrah and start looking at it from how can I have fun and interact with my mates on rs" - 2007scape.

These feature weights could indicate that within

gaming subreddits, comments that emphasize the author's personal life and relationships with other people tend to be more likely to be gilded. However, more research needs to be done in the future to confirm this finding.

In Community 4 (mostly sports related), the present tense and second-person pronouns (such as 'you') are predictive of gilded comments. These linguistic features are associated with active narrations, when one is caught in the drama of the moment. These features are also very reminiscent of sports commentary, which consists both of narrating events in the present tense and asking/criticizing a fellow expert about their perspective (by using second-person pronouns). Further work needs to be done to confirm this result, though we did find a few examples in our data, including the following gilded comment in the "soccer" subreddit, which use the present tense frequently:

"I think that in that moment at the start when he freezes, he's actually calculating all possible plays for the next 30 seconds, considering grass length, humidity and the NASDAQ index."

In addition, the "death" LIWC feature is positively weighted for Community 4. Metaphors invoking death are also common in sports discussions (i.e. "they killed it on the field"), which may account for most instances of death-related words in this community. A gilded example that uses a death metaphor is below:

"This was also the time of the year when the As were killing everyone in run differential. Something something counting chickens" - baseball.

It is more difficult to analyze Communities 2 and 3, but it appears that Community 3 (mostly information-sharing) appreciates comments with angry words and swear words. Perhaps outrage is valued in these subreddits, and thus comments with these signals are more likely to be gilded. This would not be surprising, since Community 3 contains a lot of politics-related subreddits, where anger tends to thrive.

7 Limitations and Future Work

As mentioned in Experiment 1's discussion section, both gild classification and rank classification are difficult tasks. We hope to improve our models' performance, so that the most highly weighted features would give us more accurate signals as to what factors are important and relevant for gilding and upvoting behaviors. We plan to improve our classifiers for both tasks by using using more data (particularly to get more gilded examples) and incorporating feature selection techniques. Other types of classifiers, such as Naive Bayes, may also perform better than SVM and Random Forest.

In order to obtain a richer interpretation of appreciation behavior, we hope to include linguistic features that capture syntactic properties of comments, as well as higher-level pragmatic phenomena, such as humor and sarcasm. Furthermore, we plan to include more ways of capturing topics within comments. Initially, we included as features normalized counts of words from 194 categories created by Fast et al. (2016). These categories involve not only emotions, but also topics such as everyday objects and activities, such as clothing, tourism, celebration, and technology[7]. However, likely due to the low signals resulting from our small dataset, these features added a considerable amount of noise, especially to our SVM classifiers, and decreased performance. We may incorporate these features again on a larger dataset, or use topic modeling techniques such as Latent Dirichlet Allocation (LDA) to get topic features (Blei et al., 2003).

In our second experiment, we use two separate metrics (commenter overlap and post title similarity) to create networks of related subreddits. Perhaps it would be useful to combine these metrics, or use different ones, to better capture subreddit relationships. Another limitation of our network is that we create edges between all pairs of subreddits with nonzero user overlap or text similarity; this may force every subreddit to be placed into a community, even if they are only tangentially related to the other subreddits in their community. In future work, we could add edges only if the similarity between the two subreddit nodes passes a certain threshold. A network created with these changes may then yield more intuitive communities and possibly increase the modularity.

Finally, we aim to characterize appreciation behavior across Reddit, but for efficiency, we only studied 100 top subreddits from May 2015. It is possible that the community structure of these most popular subreddits is not reflective of Reddit as a whole, and we are missing a long tail of smaller subreddits with fewer comments. In the future, we hope to extend our analysis beyond the most popular subreddits.

8 Conclusion

We first used machine learning techniques to compare two appreciation behaviors, gilding and upvoting, across Reddit's top hundred subreddits. By creating a network of subreddits and using community detection, we examined how gilding varies between groups of related subreddits. We showed that gilding and upvoting are related but distinct appreciation behaviors. These results suggest that if a user wants to be gilded, their comment should be long and memorable, emphasize a personal story, and submitted shortly after a post is made. However, where one comments also matters. Differences in important features between communities of subreddits suggest that distinct cultures within Reddit impact gilding behavior, and further characterizing these cultures is a likely line of future work.

[7]The full listing of categories and words can be found in this tab-delimited file: empath Github

9 Acknowledgments

We would like to thank Will Hamilton for the Reddit data, Tim Althoff for suggesting this project, Dan Jurafsky for log odds calculations, and David Jurgens for our LIWC lexicon. We would also like to thank Silviana Ciurea Ilcus for her helpful feedback.

References

Tim Althoff, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2014. How to ask for a favor: A case study on the success of altruistic requests. *CoRR* abs/1405.3282. http://arxiv.org/abs/1405.3282.

Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. 2012. Effects of user similarity in social media. In *Proceedings of the fifth ACM international conference on Web search and data mining*. ACM, pages 703–712.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3(Jan):993–1022.

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* 2008(10):P10008.

Cristian Danescu-Niculescu-Mizil, Gueorgi Kossinets, Jon Kleinberg, and Lillian Lee. 2009. How opinions are received by online communities: a case study on amazon. com helpfulness votes. In *Proceedings of the 18th international conference on World wide web*. ACM, pages 141–150.

Ethan Fast, Binbin Chen, and Michael S Bernstein. 2016. Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, pages 4647–4657.

Michelle Girvan and Mark EJ Newman. 2002. Community structure in social and biological networks. *Proceedings of the national academy of sciences* 99(12):7821–7826.

Liangjie Hong, Ovidiu Dan, and Brian D Davison. 2011. Predicting popular messages in twitter. In *Proceedings of the 20th international conference companion on World wide web*. ACM, pages 57–58.

Aaron Jaech, Victoria Zayats, Hao Fang, Mari Ostendorf, and Hannaneh Hajishirzi. 2015. Talking to the crowd: What do people react to in online discussions? *arXiv preprint arXiv:1507.02205* .

Himabindu Lakkaraju, Julian J McAuley, and Jure Leskovec. 2013. What's in a name? understanding the interplay between titles, content, and communities in social media. *ICWSM* 1(2):3.

Thomas K Landauer, Peter W Foltz, and Darrell Laham. 1998. An introduction to latent semantic analysis. *Discourse processes* 25(2-3):259–284.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates* 71(2001):2001.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology* 29(1):24–54.

A Networks

In the following pages, we visualize our topic-similarity and user-similarity networks and the communities within them.

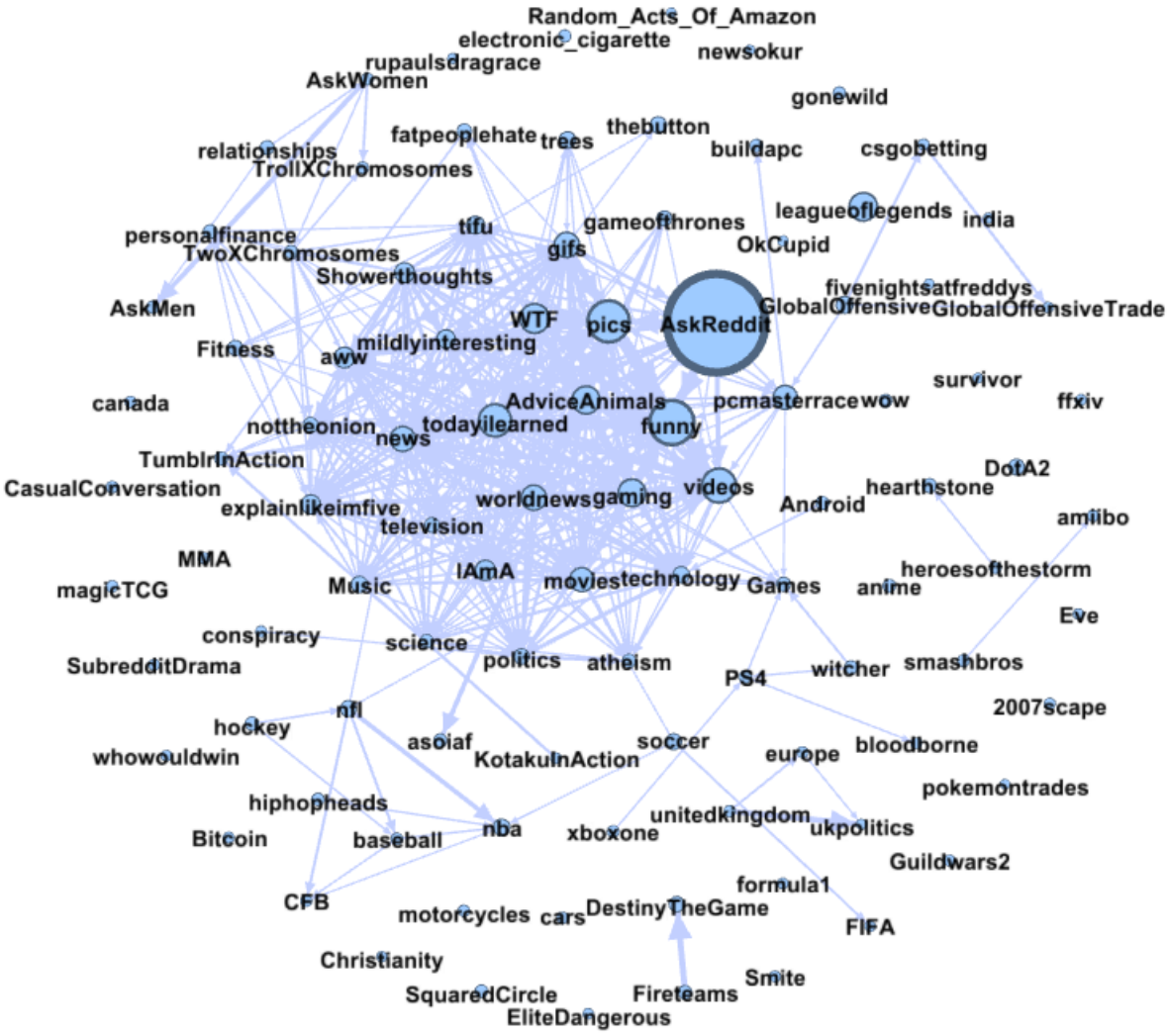

Figure 3: Networks based on subreddit user overlap. Edges connect subreddits with a Jaccard similarity threshold above 0.04. The size of a node corresponds to the total number of commenters in that subreddit.

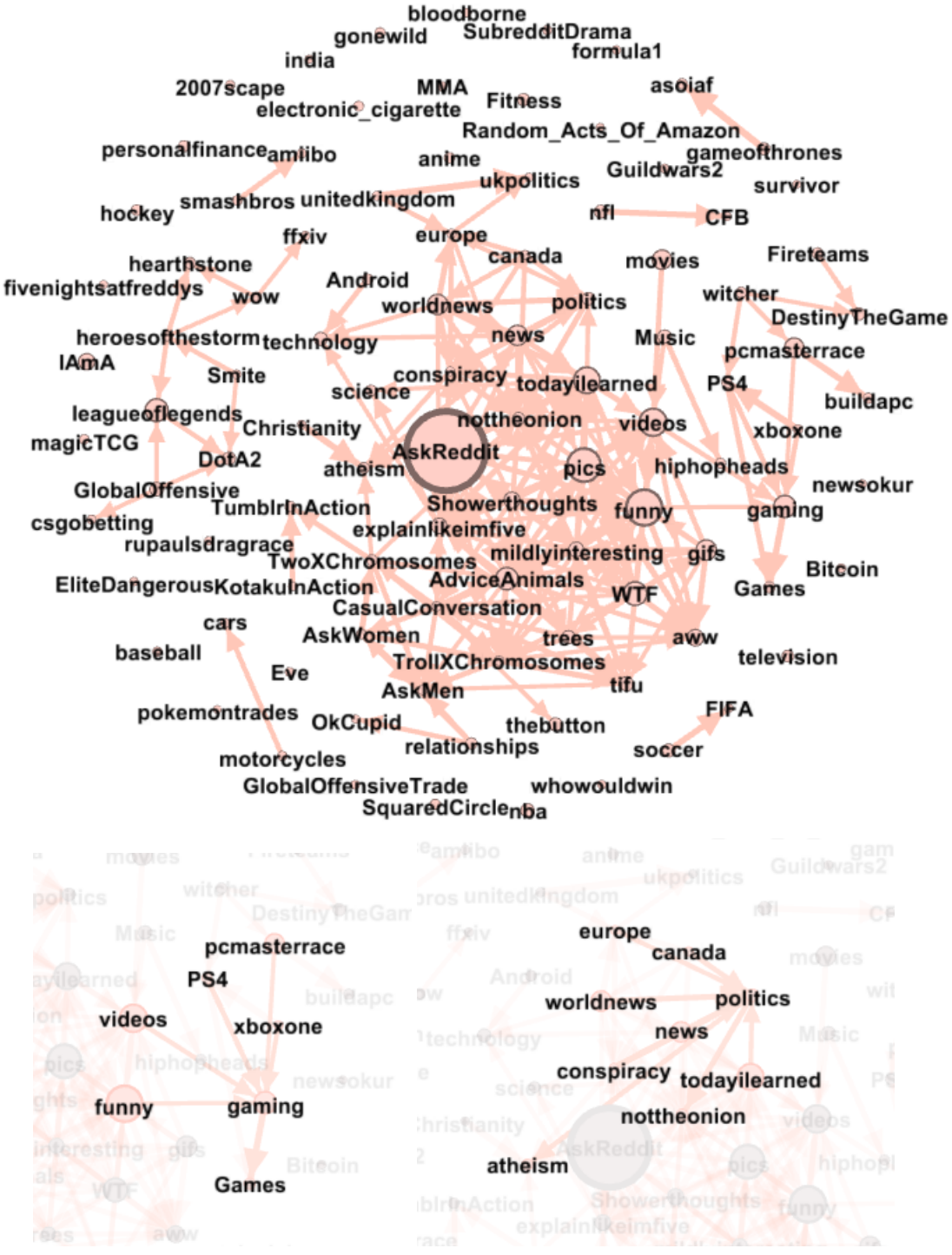

Figure 4: Subreddit network based on word usage in post titles, with edge weights thresholded at a cosine of tf-idf vectors of 0.6. Also shown are the neighbors of gaming and politics.

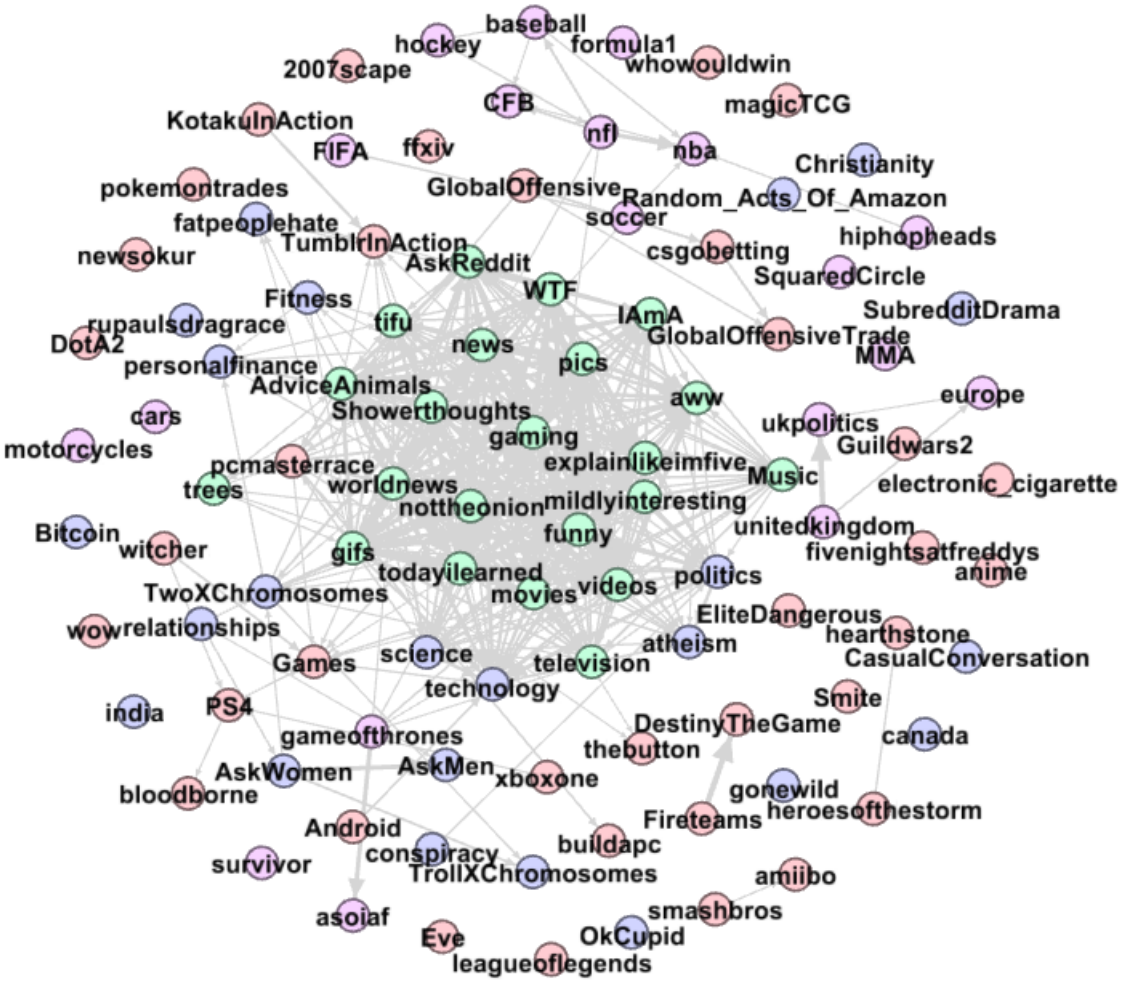

Figure 5: The same network as Figure 3, with nodes colored by community membership. Using the numbering in Table 5, green is community 2, red is 2, blue is 3, and purple is 4.

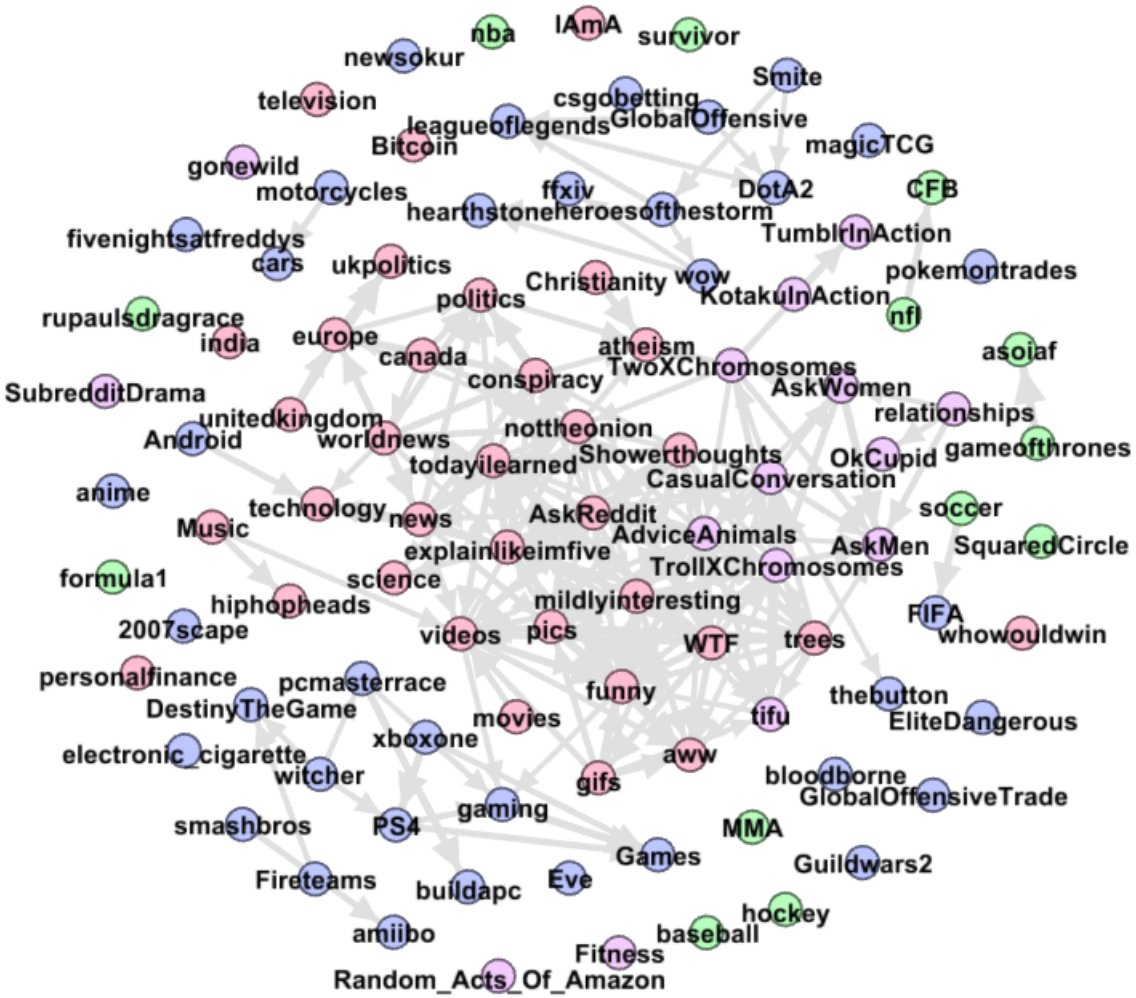

Figure 6: The same network as Figure 4, with nodes colored by community membership. Blue is community 1, purple is 2, pink is 3, and green is 4.