Content Similarity Between a User and Their Social Media Network

Ben Thorne, Phong Hoang, William Simon, Charles Winston *

May 9, 2018

Abstract

In the ever expanding world of social media, examining the connections that users make and the contents that they share can provide valuable insights. With the creation of a model to examine the similarity between a user and his social network, a host of possible applications can be realized: personalized trending content, personalized advertising, and social analysis. In this paper, a model using Twitter networks and data to define the similarity of interests between a source user and their network is outlined. By using the entity values of a tweet, the likeness of ideas between users and their network can be measured. This is done mainly by aggregating the profiles of one and one's network separately into entity vectors and comparing the similarity of two vectors using cosine similarity. The paper will analyze four different users: LeBron James, Tyler, the Creator, Donald Trump, and Cardi B. Throughout the paper, we can see that all users have much higher content similarities with their networks than with a random sample. This proves the hypothesis that users develop social networks that share similar interests to their own. Such key fact is really important in predicting whether certain topics will interest a user by analyzing the content of their social network.

1 Statement of the Problem and Goals of the Model

The main question that we try to answer in this paper is: is it possible to accurately determine and predict topics that will interest a specific user based on the content of his network? The paper will model and compare the interests of an individual, his social network, and the world, as a result of the contents they put on their social media. Specifically, Twitter will be the platform that we analyze thanks to its extensive API and content analysis. This model is intended to target the idea of predicting the relevance of content to a user. Since an individual's interests far surpass what content they put out, the addition of a user's social network profile allows for a more holistic view of users interests.

By using the data that Twitter has amassed, a profile of a specific user's interests will be built and then separately compared to the aggregate profiles of those they follow and a random subset of a worldwide sample. This is an attempt to model an individual's real-life social network as a result of their online presence. The expectation is that a user will have a higher rate of shared interests on social media with those they have chosen to follow - similar to how people are friends with those who are similar to them [1]. As the size of the network increases, the chance of the model picking up on events happening in real time increases. Since the users in this network theoretically share similar interests with the source user, this model could possibly lead the way toward predicting personalized trending content as events occur. This would then bring more intricacy to the currently encapsulating idea of personalized content in the marketing world [2].

One curious aspect of this model is how it will begin to resemble the entire Twitter landscape. Instead of merely creating a network based on the first connection between a source user and those they follow, the network can be expanded to users with more degrees of separation. With an expansive enough network, it may be possible to truly see the classic idea of six degrees of separation (Morse). It may also see if people form clusters of similar content, such as groupings of athletes, politicians, musicians, etc. As the model adds degrees of separation from the source user, it is expected that the similarity between the source user and their network will eventually fall off to the worldwide similarity. Ideally, if this model can be expansive enough, an optimum number of degrees of separation will be seen.

*Department of Mathematics, 503 Boston Ave, Tufts University, Medford, MA 02155. {*Benjamin.Thorne, Phong.Hoang, William.Simon, Charles.Winston*} @tufts.edu

Honestly, the accuracy of such a model cannot be measured perfectly. There is no direct measure for whether or not a person is interested in a topic unless that person is asked directly. Even after exposure, it is not guaranteed that a user will mention a topic they are interested in. Additionally, interests change over time and can fluctuate from day to day based on various influences. Consequently, there will be no straightforward measurement of the success of this model, instead, the success will be determined through indirect means and judgments based on the expectations outlined below.

Firstly, it is expected that a higher correlation between the content of a source user and the content of their network will be seen, as opposed to the correlation between the content of the source user and a worldwide sample. This expectation is integral to the success and application of the model. Additionally, it is presumed that the similarity between the source user and their network will not be too high, as this may indicate false data or incorrect analysis.

Secondly, as the degrees of separation of the model are increased, an accurate model should show a falloff of the correlation between a user and their network that approaches the worldwide similarity. This is because increasing the degrees of separation should progress the network toward resembling the entire Twitter landscape. As a secondary result of increasing the degrees of separation, a successful model should be able to notice events happening relatively quickly. With enough users in a network, it can be reasoned that some subset will mention new content that could be of interest to the source user.

Finally, if the built model can be expansive enough, a successful model should show clustering of users that are routinely mentioning similar content. Theoretically, users who routinely show interest in similar topics should have closer connections between each other and develop into clusters in the network. Although this is an expectation of a successful model, based on various limitations such as the Twitter API, computational power, and time, this is not something that is reasonably expected from this model.

2 Literature Review

In this section, we will mention some other researches and how they are related to this paper. This project is inspired by [3]. In this paper, the authors stated that trending topics play an important role in spreading emerging issues. Because each person cares about different topics, the goal is to personalize the trending topics based on one's past contents and social network information.

Such paper proposes a method to personalize trending topics. According to the paper, let U = $\{u_1, ..., u_m\}$ is the user set where m represents the number of users, each user has a set of posts $u_i = \{p_{i1}, ..., p_{i|u_i|}\}$, where each post is an attribute vector $p_{ij} \in \mathbb{R}^n$, where n is the number of textual features. $y \in \mathbb{R}^m$ is a label vector for one topic that determines whether a user is interested in such topic or not. $y_i = 1$ if user i is interested in the topic (he mentioned the topic in one of his old post) and $y_i = -1$ otherwise. This denotes the set of social links between users: $a_{ij} = 1$ if i follows j and 0 otherwise. The paper states that an optimal function f where $f(u_i)$ denotes the prediction result of whether user i is interested in such a topic.

The paper suggests that we should first predict the label for a single post p_{ik} of user i with:

$$f(p_{ik}) = \frac{1}{1 + e^{-w.p_{ik} - b}},$$

where b is the model bias, and w is the vector of model parameters. By aggregating the prediction of all posts, the estimation of users can be obtained as:

$$f(u_i) = \frac{\sum_{k=1}^{|u_i|} f(p_{ik}).e^{\alpha f(p_{ik})}}{\sum_{k=1}^{|u_i|} e^{\alpha f(p_{ik})}},$$

where α is the parameter to determine the extent of the softness. So given a label vector y, the cost function over w and b should be minimized:

$$\frac{1}{2}\sum_{i=1}^{m}(y_i - f(u_i))^2 + \frac{\lambda}{2}w^T w + \gamma \left|w\right|_1,$$

where $w^T w$ is the regularization term to avoid over-fitting by penalizing the model complexity and the l_1 norm is used to induce sparsity. According to the assumption that friends are likely to have similar interests, assume E represents friendship between users where $e_{ij} = 1$ if $a_{ij} = 1$ and $a_{ji} = 1$ otherwise 0. Thus, the following regularizer should also be minimized:

$$\sum_{e_{it} \in E} e_{it}(f(u_i) - f(u_t))^2,$$

to penalize the difference between friends. Thus, the problem is to find w and b that minimize the objective function:

$$\frac{1}{2}\sum_{i=1}^{m}(y_i - f(u_i))^2 + \frac{\lambda}{2}w^T w + \gamma \left|w\right|_1 + \frac{\mu}{2}\sum_{e_{it} \in E} e_{it}(f(u_i) - f(u_t))^2.$$

The paper then proposes some methods to optimize the objective function. However, it is of the scope of this analysis.

It is totally understandable that one’s interest will be heavily based on their past contents; but is it really dependent on one's social network as assumed in the paper? In this paper, we will propose a method to examine such assumption by looking at the similarity between one's interest and their network.

3 Assumptions and Justifications

At the heart of our model lays a founded assumption that a persons tweets are good indications of what that person is interested in. This logic follows from the idea that, in the current online world, what you talk about is what you are interested in. This leads to more selective interactions as people will more often choose to interact with those they share similar interests with [1]. Using a selection of tweets to profile a users interests is justified by the fact that Twitter is a platform in which people speak about that which they find important to share with others. Twitter is specifically built to share ideas succinctly - a result of the imposed character limit - therefore it is reasoned that an individual tweet will display the main idea of interest clearly.

A second assumption is that a users Twitter network is a good representation of the content that they are exposed to. Although this assumption is not entirely valid, Twitter is a large influence in social media and a persons online presence. In the last decade, Twitter has grown into one of the largest and most influential social media platforms (Lee). Therefore, given the ease of access to the information it has gathered, Twitter was chosen to be the online platform for this model.

The next assumption that was used in the creation of this model is that the entity values Twitter assigns to a tweet are an accurate representation of the content of a tweet - this breakdown is discussed more thoroughly in the Model Design section of this paper. This function takes attributes of a tweet, such as hashtags and mentions, and puts them into categories. The justification for this containing the most relevant information in a tweet is a result of the way Twitter operates and the key role that these attributes play. Twitter is designed on the basis of trending content and conversations. Although these attributes do not incorporate all of the ideas of a tweet, they should give some representation of the core ideas.

Lastly, a main goal of this model is to compare a users content to a worldwide average. Based on the computational power that would be needed for this direct comparison, a random sample of users is selected to represent the worldwide average. By using a random sample of users of the same magnitude as the network, a similar scale of ideas is harnessed, allowing for fewer limitations in the comparison.

4 Model Design

In order to analyze content similarities and degrees of separation easily, a network that resembles Twitter followers can be designed. The main invariant of this graph is that directed edges represent a user following another user. (Twitter calls this being friends in its API). This process begins with a source user, and continues through their friends, friends of friends, etc. This network results in two possible representations: a model of the exact network by including edges between every existing node, or only include directed edges from the source user outwards to their friends (for the graph-algorithm fluent, a graph that represents a breadth-first search of the total network). The first model is much better for retrieving general, realistic data about the entire social network, but in this instance, the second approach is taken in order to specifically analyze an individual. This way, each user only appears as a friend once within the model. Figure 1 below shows a representation of how the model constructs a social network, displaying the degrees of separation between a node and the source user.

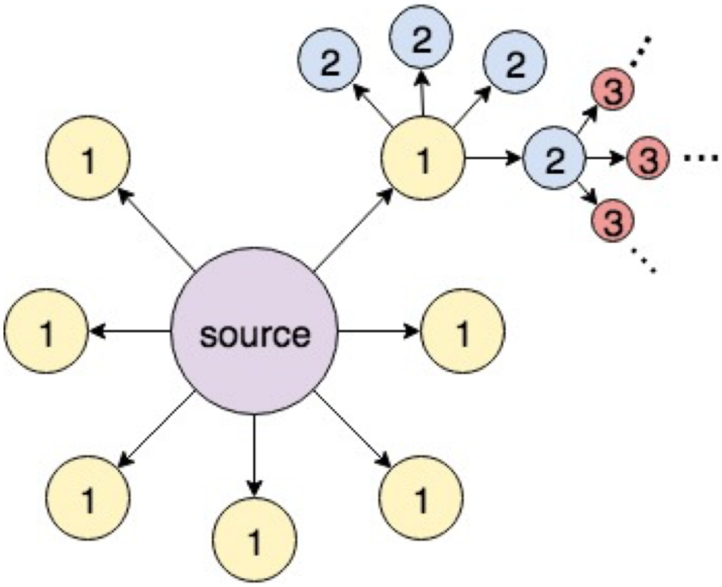

Figure 1: Sample Diagram of a Source User's Social Media Network.

Unfortunately, due to call-rate limitations with the Twitter API, only the first degree of separation and a random sample could be examined (see Strengths and Weaknesses for more). However, just those two comparisons yielded expected results.

Four Twitter users were selected for analysis based on their notoriety, number of friends, total tweets, and areas of influence. Since the API has extensive rate-limits, the less number of friends we use to compute the quicker and more efficient the computation is. This is because the program does not have to wait a time-window when the rate limit is exceeded in more complex cases. Also, it is important that the user is notorious and tweets a lot - an inactive account would result in not entity values and therefore no similarity metrics. Lastly, it is interesting to compare the differences between athletes, politicians, entertainers etc. With this in mind, the users displayed in Table 1 below were chosen.

Name	Handle	Number of friends	Number of tweets	Occupation
LeBron James	@KingJames	181	5,800	athlete
Tyler, the Creator	@tylerthecreator	175	40,100	entertainer
Donald Trump	@realDonaldTrump	46	37,500	businessman / politician
Cardi B	@iamcardib	137	20,100	entertainer

Table 1: List of Chosen Source Users

Now that users have been chosen for the study, the method for determining similarity between tweets must also be determined. The best approach is an application of the cosine similarity of two vectors, which measures how similar two vectors are. The cosine similarity of two vectors technically finds the cosine of the angle between two vectors, and since that value is between 0 and 1, it can be used to determine vectors percent similarity. The smaller the angle, the larger the cosine and the higher the correlation; the larger the angle, the lower the cosine and the lower the correlation. The cosine similarity of two vectors A, B; $A, B \in \mathbb{R}^n$ is:

$$cos(A,B) = \frac{(A,B)}{\|A\|_2 \|B\|_2} = \frac{\sum_1^n A_i B_i}{\sqrt{\sum_1^n A_i^2}\sqrt{\sum_1^n B_i^2}}$$

The cosine similarity becomes useful when a vector space can be defined: the entity vector space. The entity vector space represents a space in which each unique entity mentioned by a user or their network corresponds with a single dimension of the vector space. A users entity vector is defined by the frequency of entities mentioned in their Tweets. A networks entity vector is defined by the frequency of entities mentioned in the networks Tweets. The cosine similarity of these two vectors can be computed to measure the content similarity.

Cosine similarity does not rely on the magnitude of the vectors, only the angles, so it is a good way to compare vectors of different sizes and magnitudes effectively. In order for the cosine similarity to be run on the selected users, their networks, and sample tweets, all of the tweet data must be extracted from the Twitter API. The algorithm for retrieving this data is explained below.

Before we detail the algorithm to compute the content similarity between a user and their network, there are some important terms that must be defined.

- *Twitter user*: A single account on Twitter.
- *Screen name*: A unique string identifier for a Twitter user. Chosen by the user.
- *User ID*: A unique integer identifier for a Twitter user. Assigned to user by Twitter.
- *Friend*: A relationship between two users where if user A follows user B, then B is As friend.
- *Tweet Object*: A JSON object representing a single Tweet sent by a user. Among the many fields of a Tweet Object are its entities.
- *Entities Object*: A JSON object contained inside a Tweet Object that provide metadata and additional contextual information about the Tweet. This includes hashtags and user mentions, which are the entities this algorithm is focused on.
- *Hashtag*: A string beginning with the character used as a content tag. Hashtags can be searched on Twitter and used to track trends.
- *User mention*: A string beginning with the @ character that tags a users screen name.
- *Timeline*: A list of Tweet Objects from a single user.

There are several steps to the algorithm for computing the content similarity between a user and their network. The algorithm begins with a given user, and for that user obtains a list of their friends IDs using the API call friends/ids, which takes a Twitter screen name (the source user) and returns a list of their friends Twitter IDs. This friend data is dumped into a JSON file so it can be extracted for future methods in the algorithm. For each friend, a list of 200 of their latest tweets are compiled using the API call statuses/user_timeline, which takes a users screen_name or ID and returns a list of Tweet Objects of their most recent Tweets. A list of all friends of the sources latest 200 Tweets are stored in a JSON file. The source users latest 200 Tweets are also retrieved and stored in a JSON file using the same API method.

Once two separate lists of Tweets are compiled: one for the users most recent Tweets, and one for their networks most recent Tweets, the entity vector space and entity vectors can be defined, and the cosine similarity can be computed. First, it should be noted that the subset of entities used are hashtags and user mentions. To define the entity vector space, all Tweets from both the user and their network are iterated over while keeping a dictionary which maps unique entities mentioned to their dimension in the vector space. Every new entity encountered during the iteration over Tweets is added to the dictionary and assigned the next available dimension. Once the dictionary that maps all unique entities to their dimension is obtained, the individual entity vectors for the user and their network can be computed easily by setting the value of the dimension corresponding to each entity equal to the frequency with which the user or network mentioned that entity. The cosine similarity of these two vectors is computed to measure the similarity of the content of the user to the content of the network.

In order to draw conclusions from these more local results, running cosine similarity between a users Tweets and a random sample of Tweets is helpful. To generate the random sample, the Twitter API method statuses/sample was used to stream recent Tweets posted to Twitter. These Tweets were iterated over to obtain a list of English speaking users, the number of users the same as the number of the source users friends, in order to get an accurate comparison. This random sample of users was used in the above algorithm as if it were the users network.

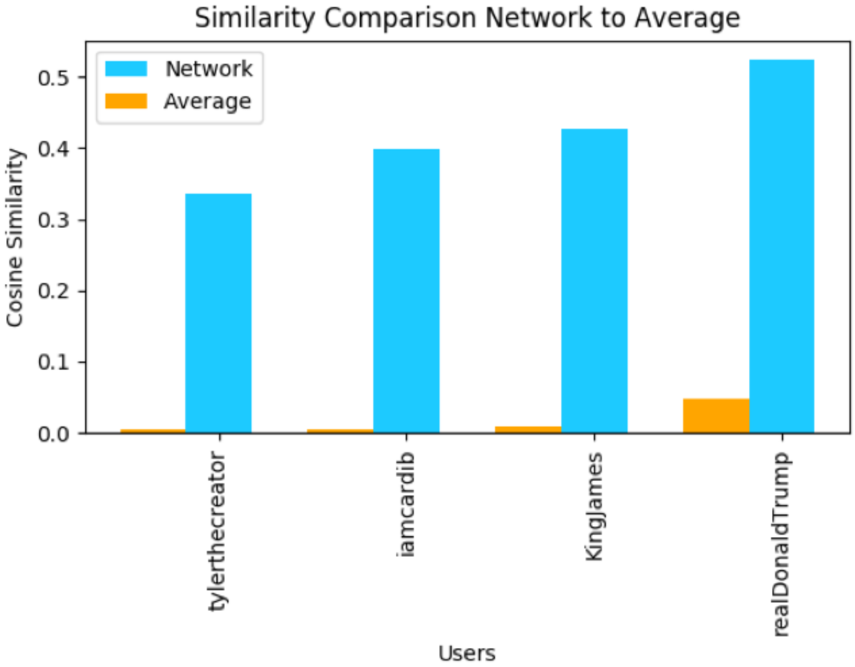

Figure 2: Similarity between users and network

5 Model Testing and Results

The following table shows the results obtained after measuring the cosine similarities of each source user against their network and a random sample.

Handle	Similarity to Network	Similarity to Random Sample
@KingJames	0.427	0.00805
@tylerthecreator	0.337	0.00486
@realDonaldTrump	0.525	0.0479
@iamcardib	0.400	0.0033

Table 2: Cosine similarity metrics between each user and their network and each user and the random sample

The cosine-similarity data is visualized in figure 2.

Clearly, the similarity between a user and their network is high, and is far higher than the similarity between a user and a random sample.

Furthermore, the top 10 entities mentioned by each user are compared against the results for those same entities in the network, and vice versa. In the following bar graphs, the top mentioned entities are plotted by percent of total entities mentioned by either the user or the network.

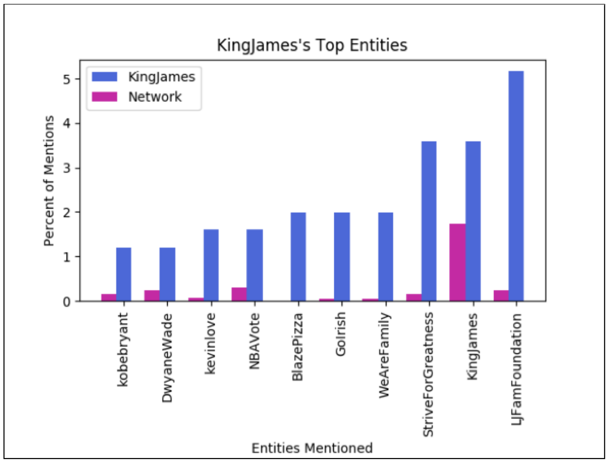

Figure 3: LeBron James Entities

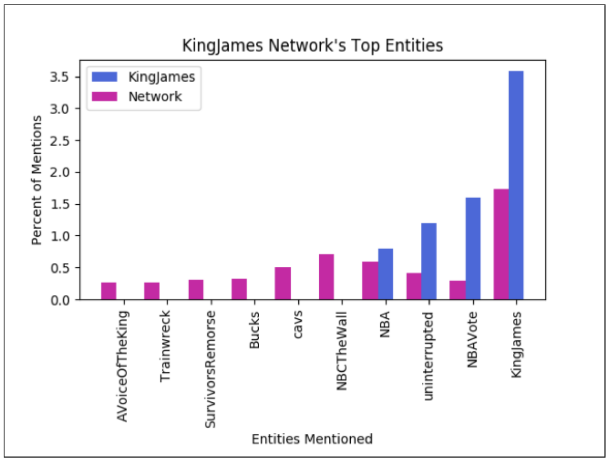

Figure 4: LeBron James Network Entities

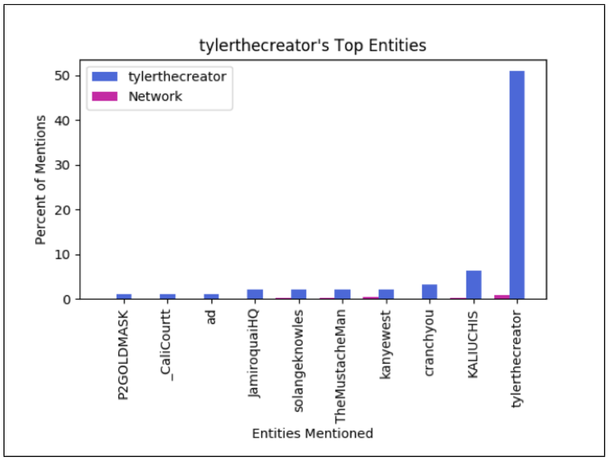

Figure 5: Tyler, the Creator Entities

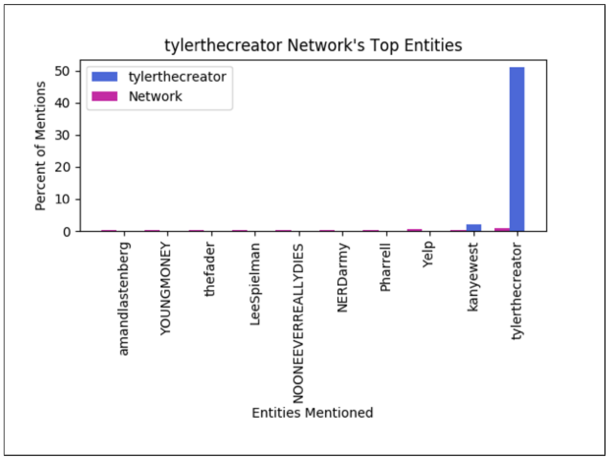

Figure 6: Tyler, the Creator Network Entities

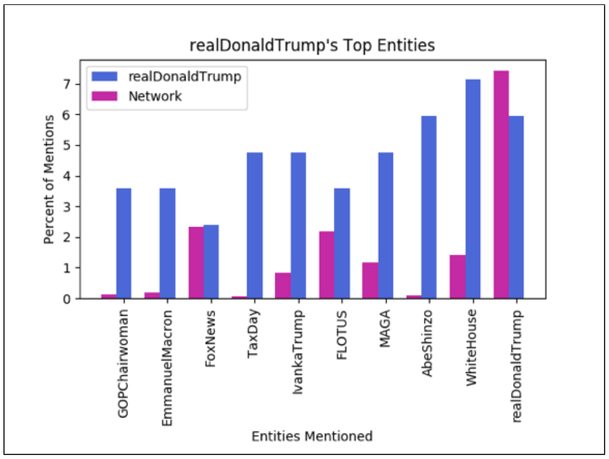

Figure 7: Donald Trump Entities

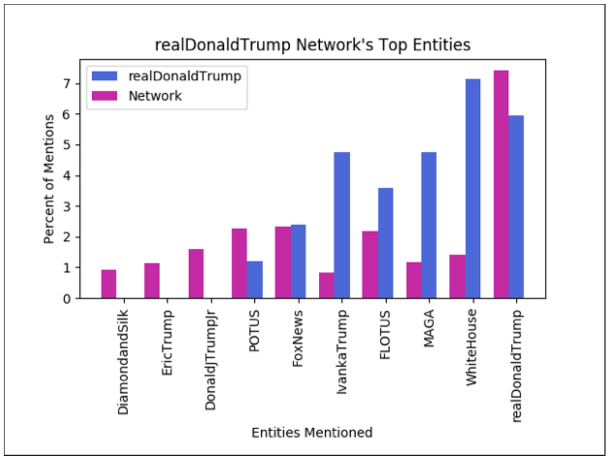

Figure 8: Donald Trump Network Entities

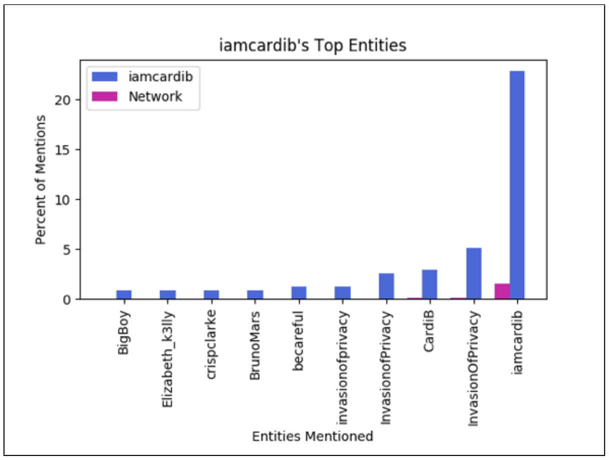

Figure 9: Cardi B Entities

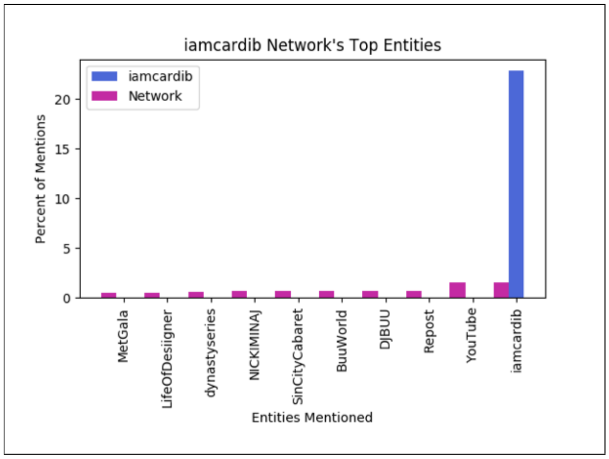

Figure 10: Cardi B Network Entities

In most cases, a user's own handle is one of, if not the most, popular entity in their tweets. This makes sense because of Twitters method of replying to tweets: when replying to someones tweet, or a thread of tweets, their handle appears as an entity. It is likely that these users engage in threads on their own feed, especially Tyler, the Creator and Cardi B. This can be seen as a weakness because it seems obvious that a user and their network should mention the users screen name frequently. The following table shows the results if the source user's screen name is eliminated as a counted entity:

Handle	Similarity to Network
@KingJames	0.305
@tylerthecreator	0.166
@realDonaldTrump	0.449
@iamcardib	0.033

Table 3: Similarity Metrics Eliminating Source User Screen Name as Entity

The table shows that in general, even if the users screen name is eliminated, the cosine similarity is still fairly high, so our measure does not rely too heavily on a user's own screen name mentions.

Another takeaway from this data has to do with the notoriety of the subjects. Tyler, the Creator, Cardi B, and LeBron James all have similar cosine similarities with the random sample - Trumps is ten times that. Even though all four of the subjects are high-profile individuals, the content surrounding Trump in the news far surpasses that of the other users. Additionally, Trump's Twitter following has greatly increased as a result of his election campaign and presidential status. As a result, it logically follows that Trump has the most entity similarity out of the sample. However, even with Trump's higher similarity metric, the cosine-similarity trend appears to converge to 0 as the degrees of separation from the source user are increased. This can be confirmed by examining additional degrees of separation which, due to API-call rate limitations, was unfortunately out of the scope of this analysis.

In any case, it is found that a user's tweets have much more in common with their network's tweets than a random subset of live tweets, with varying degrees of difference. Donald Trump's content has the closest relationship to his network, which follows from the amount of news surrounding Trump's Twitter presence.

6 Strengths and Weaknesses

The results of this model show a significant number of strengths that allow for both current and future analysis, but also present some limitations and weaknesses. These will be discussed below.

The first important strength of this model is its dynamic nature. Constructed using current time data, the results of this model will change consistently as users create new content. This allows for the model to give practical and relevant information about users currently and keep the results from being dependent on static sources of data.

A second strength of this model is that the entity similarity metrics are found for both close connectivity and no connectivity. This allows the model to examine both ends of the spectrum to gauge how the metrics compare to each other. Because of this, the hypothesis founded to this paper - that the content of a users network will be more similar to their own as opposed to a random sampling - can be argued.

The third notable strength of this model is its generality and polymorphic programming. Although the results in this paper do not encapsulate the entirety of the content of a user's network, this model can be easily adapted to fit new ideas. As noted before, this model could be readily expanded to show the trending of the similarity coefficient as a function of the degrees of separation. Another possible examination that could be made with a slight expansion of this model would be to see the interconnectivity of the graph. If further developed, showing connections between users in either direction and showing clusters of similar users might lead to further insights. These are just a few possible further adaptations that could be made with ease.

Unfortunately, although this model has many strengths, it is severely limited by some of its weaknesses.

The biggest current limitation of the model results directly from the Twitter API. Twitter enforces a maximum on the number of calls that can be made every fifteen minutes to request the dataset that is required. This limitation bottlenecks the speed of the program and makes it unrealistic for use in many applications. First, this hampers the model's ability to offer personalized trending content since the average number of user profiles that can be requested every minute is one. This also prevents the model from being expanded past the first degree of separation in a realistic time frame. It is estimated that to examine just the second degree network could take up to twelve hours per user. This puts a severe limitation on the current model - although it is possible that this could be remedied with Twitter's consent.

The second most impactful weakness of the model is the way Twitter creates entities from a tweet. The Twitter API assigns entity values to a tweet as a result of two attributes: mentions and hashtags. Although this can determine some key attributes of a tweet, it is far from extensive. Given memory limitations, this is necessary in the current version of the model, however, with more storage space, natural language processing could be run on the text of each tweet in order to break it down into phrases, ideas, and qualities. The resulting dataset could then be grouped into categories based on similarities. For example, this could allow the model to group together all tweets that talk about the NBA into one group. As a secondary consequence of this current implementation, users who do not make use of hashtags and mentions will not have entity values that align with their true interests. In order for the possible applications of this model to be realized, it is likely that this weakness will need to be overcome.

7 Applications

Thanks to the exponential growth of both the internet and social media, not only will the results of this model become more accurate, but the possible applications will become more valuable as well. To reiterate, this model intends to identify the potentially interesting topics to a user based on the topics that interest their friends and their social media feed. The results of this model prove the hypothesis that the topics of interest of a person's social network are more likely to be interesting to the source user than a random sampling of worldwide interests. Using the current version of the model and future alterations, many applications are possible, few of which are outlined below.

A more direct application that has been alluded to already, is the idea of using this model to create personalized trending content for a user. Given a large enough network of friends, this model should pick up on current events or trends based on real-time monitoring of a user's network's tweets. This trending content could then show up as a notification to the source user to present an idea that might be of interest to them. It is clear that predicting trending content is a valuable application since Facebook, Twitter, YouTube, Reddit, and other social media platforms all have some sort of implementation. In an era where the value of personalization is ever increasing, applying this to trending content could prove to be a valuable application (Hyken).

A secondary application that is slightly more abstract in relation to this model is the idea of personalized marketing. It has now become an afterthought in the current digital world that web browsers and web sites keep track of sites that a user visits and products that they look at in order to market similar items. Likewise, this model - with some alterations - could be applied to a personalized marketing algorithm. Using this model, advertisements would be influenced by a user's social network's aggregate profile, creating an abstraction that adds breadth to the recommended products. This could have the added benefit of making advertisements feel less invasive to users, while also marketing products and companies a user may not have even been exposed to yet.

Clearly, whether with the current version of the model presented in this paper, or one with some alterations, applications for this model are both valuable and widespread in the current era of online social media presence.

8 Conclusions

The goal of this model was to answer the question: is it possible to accurately determine and predict topics that will interest a specific user based on the content of their social network. This paper aims to detail a proposed model to measure this similarity and highlight possible applications and further improvements. To answer this question, Twitter was used as the data source to both create and measure a user's social network and the content it produces. Twitter was used because of its easily accessible API, relevance in online social media, large dataset, and predetermined content analysis. In this model, the content that a user created was measured and separately compared to the aggregate content of their network and a random worldwide sampling. The metric used to determine this similarity was the cosine similarity between the entity vector of a user and the aggregate entity vector of their network.

What was seen is that all analyzed users had significantly higher content similarity with their network than with a worldwide sampling. Although this model was far from exhaustive - limitations of the Twitter API restricted the size of the networks that could be analyzed and how much content could be deciphered from a tweet - it proves the concept that a user's social network can be used to predict whether a topic will be of interest to them with higher accuracy.

9 Code

Visit https://github.com/charlesrwinston/content_similarity for code that ran the experiments.

References Cited

[1] Brueck, *Scientists say they can predict who you're friends with based on brain patterns alone*, Business Insider (2018).

[2] Hyken, *Recommended Just For You: The Power Of Personalization*, Forbes (2017).

[3] Liang Wu, Xia Hu, Huan Liu, *Early Identification of Personalized Trending Topics in Microblogging*, ICWSM (2017).

[4] Morse, Gardiner, *The Science Behind Six Degree*, Harvard Business Review (2003).

[5] Lee, Dave, *How Twitter changed the world, hashtag-by-hashtag*, BBC (2013).

Participation

Charles: Interacted with the Twitter API and wrote the majority of the code to compute results.

William: Theorized about the logistics of the project and put a lot of the work into writing.

Ben: Facilitated and helped with bringing everything together with the results, paper, and presentation.

Phong: Researched other reports with similar goals/findings, wrote and edited the paper in LATEX.

We were all involved in all aspects; this is a rough, general categorization.