Influence of Topical Interests on Users' Social Networks

Sameep Bagadia[1], Pranav Jindal[2], & Rohit Mundra[3]

Abstract— **The ubiquity and prevalence of online social networks in the recent years has made available vast amounts of data about human predilections and preferences. In the past, works in fields such as social and cultural anthropology have attempted to study the various factors that influence social behavior and relationships. However, most of these works faced the problem of having limited data due to difficulties in data collection and also suffered from biases in the data collection process. In this study, we use Twitter's user base to study the factors that influence or are influenced by friendships on Twitter under the hypothesis that friendships are closely associated with common interests. This allows us to compare and contrast the experimentally found influential interests on Twitter with those expected based on past anthropological and sociological work. The large amount of data available allows us to overcome the shortcomings of using small samples. With the insight gained from this experiment, we – (1) Identify members within a topical interest that have similar user bases, (2) Find different topical interests which appeal to similar user bases and thus perform cross-domain recommendation, (3) Predict links (friendships) for users based on their shared interests in different topical interests.**

I. INTRODUCTION

In the age of social media, understanding the dynamics of social networks such as its evolution and the interactions between its users has gained a lot of importance for a variety of reasons such as knowledge discovery, marketing, understanding social behavior etc. A lot of data has been made available through social media websites such as Google+, Facebook and Twitter, which can be analyzed to draw out interesting conclusions about human social behavior.

In this paper, we analyze a sample of the Twitter network to gain insight into social relationships between users. In particular, we analyze the user similarity based on the topics they follow and correlate it with their friendships. As a result, we draw conclusions about some of the factors that influence or are influenced by friendships on social networks.

In the following sections, we have detailed the questions we are trying to answer, reviewed literature relevant to our project, discussed data collection techniques employed, and briefly touched upon the statistics of the resulting data. We then discuss the social network analysis techniques and algorithms that we apply to our data and discuss our findings and results along with the further work that can be done.

II. TERMINOLOGY

This section contains some of the terminology used throughout the paper:

- **User**: A user is a person with an anonymized and registered Twitter ID. Some Twitter IDs are registered for businesses, celebrities, sports teams, etc. and we do not consider them to be users – instead they are considered to be features.
- **Feature**: A feature is a registered Twitter ID that has a large user base because of its prominence in one or more topical interests. For instance, @KevinHart is a feature because of his prominence in the topical interest "Comedy".
- **Topical Interest**: A topical interest is a genre or set of features. Features within a topical interest are outrightly similar based on the characteristics of the topical interest.
- **Single Topical Interest Graph**: For each topical interest we create a bipartite graph from user nodes to feature nodes belonging to that topical interest.
- **Dual Topical Interests Graph** This graph, defined for a pair of topical interests, is simply the union of the single topical interests graphs for the topical interests in the pair. We use this graph for making cross domain recommendations.

III. MOTIVATION AND OBJECTIVES

In this study, we wish to accomplish the following using Twitter's social graph:

(A) Identify similar features within a topical interest
(B) Identify similar features across topical interests
(C) Identify potential friendships based on their ego network characteristics

By accomplishing the above tasks, we want to recommend links to users at every stage of his/her social networking experience – a system which can do all of the above reasonably well will not only allow a user to grow his/her circles, but also to identify features that interest them.

[1] S. Bagadia is a graduate student of Computer Science, Stanford University, Stanford, CA 94040, USA `sameepb-at-stanford.edu`

[2] P. Jindal is a graduate student of Computer Science, Stanford University, Stanford, CA 94040, USA `pranavj-at-stanford.edu`

[3] R. Mundra is a graduate student of Electrical Engineering, Stanford University, Stanford, CA 94040, USA `rohitm92-at-stanford.edu`

A. Similar Features within Topical Interests

We cluster items within a topical interest based on the similarity of their user bases. This can have several applications like recognizing political parties that have similar ideologies, music artists in similar genres, etc. From a user's standpoint, such an intra-domain recommendation will allow the user to explore similar features within that topical interest.

B. Similar Features Across Topical Interests (Cross-Domain Recommendations)

We also find similarities between features belonging to different topical interests which can expose some surprising hidden relationships like those seen in [3]. From a user's standpoint, the resultant cross-domain recommendation will allow a user to explore a potentially new topical interest likely to interest them. For instance, the Inheritance trilogy might attract the same user base as the Harry Potter movies so a good recommendation to a fan of only one of these would be the other.

C. Predicting ego networks for users

We propose to use the similarity of features that the users follow as a link prediction technique allowing users to grow their social circle by finding friends.

IV. LITERATURE REVIEW

A. Co-Following on Twitter (Garimella & Weber, 2008)

The paper [3] presents an in-depth study of co-following on Twitter and shows that co-following information provides strong signals for various classification tasks.

- The similarity of followers' friends was used to predict a user's preferences and to group popular Twitter users according to their audiences similarities.
- Groups of related Twitter accounts such as politicians, musicians were mapped by looking at their followers friends revealing several interesting facts. For instance, Apple and Puma target a similar, metropolitan audience.
- The work strived to find communities based on a similarity-only based approach that can easily be transferred to domains without any friends-of-friend links.

The paper introduced the concept of 2^{nd} order co-following to provide signals about similarity between nodes; we use SimRank for this purpose, which is a more systematic way of finding similarity between nodes in a network. Also, taking cue from this work, we strive to find ego networks based on a similarity-only based approach which will be useful for link prediction.

B. Factors Influencing the Choice of Friends (Stoyanova, 2008)

This paper [2] presented a valuable study of a real-world social network in Bulgaria. It demonstrated some key features that were important to friendships in the real-world such as ethnicity, religion, and profession. However, we identified that the paper fell short in many aspects – small size of dataset (only 49 respondents), lack of diversity of respondents (44 Bulgarians out of 49 respondents), small set of factors affecting friendships considered, etc. This motivated us to overcome many of these drawbacks and extend aspects of this study to a much larger and more diverse dataset with thorough consideration of every factor that could influence friendships.

V. DATASET

In this section, we describe our dataset and some of the processing steps that were required to make it useful for our experimentation. We also explain some of the statistical properties of the dataset that were taken into consideration during various decision-making steps of our study.

A. Source of Data

We are using the Social Circles Twitter dataset[1] hosted by the SNAP Group at Stanford University for our experiments. The dataset has a large user base with information about the followership for each user and is well suited for our analysis.

B. Structure of Data

The Social Circles dataset provided us with 973 ego nodes. The total number of nodes and edges were 76,269 and 1,336,678 respectively. For an ego-node with Twitter ID 123456, we had the following files:

```
123456.circles
123456.edges
123456.egofeat
123456.feat
123456.featnames
```

The .edges file provided us with Twitter IDs of the user's friends (456123, 998887, etc.) while a combination of the .egofeat file and .featname provided us with information about features for a particular user (@nytimes, @pinkfloyd, etc).

C. Classifying Features into Topical Interests

Once we had information about the features in common between an ego node and his friends, we wanted to categorize the shared features. For instance, a common feature of @nytimes suggests a shared interest in News while a common feature of @pinkfloyd suggests a shared interest in Music.

We wanted features (@nytimes) to be classified as a member of a particular interest (News) not just because the features proclaims to be member of that interest but also because others who proclaim to be members of News acknowledge @nytimes to be a valuable member of News. This idea is strikingly similar to the PageRank algorithm. On Twitter, this PageRank-like ”stamp of approval” can be explicated by followership.

After some search, we identified a Twitter directory service known as ’WeFollow’, which classifies notable Twitter users (i.e. features as per our definition) using an algorithm similar to that described above. Thus, we used a script to request categorization of features in an automated manner. This allowed us classify 176,011 features into 10,220 topical interests. Since many features are prominent in multiple topical interests, topical interests are not disjoint sets.

D. *Statistical Properties of the Data*

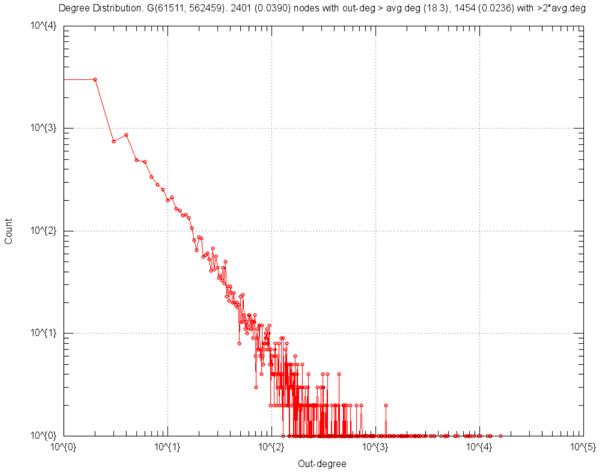

Fig. 1. Degree distribution of Interests

In Figure 1, we see the distribution of the number of users in each of the 10,220 topical interests. Visibly, the distribution follows a power law and thus, most topical interests have very few users in them. Thus, we pruned out the interests with a small user-follower base. Figure 2 shows the distribution of the number of topical interests a user has features present in.

VI. METHODOLOGY

This section describes the methodology, algorithms and network analysis techniques used for different parts of our project.

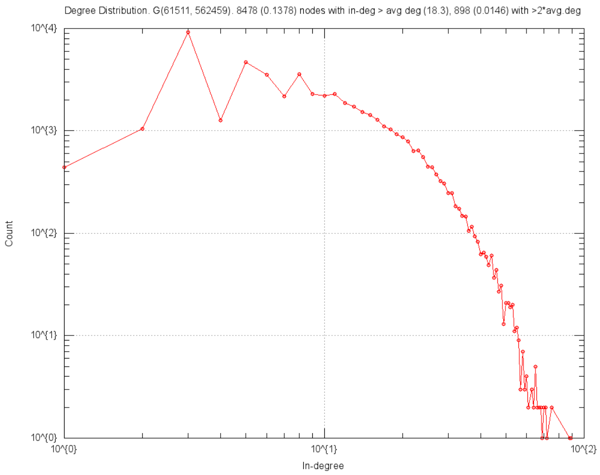

Fig. 2. Degree distribution of Topical Interests

A. *SimRank*

SimRank algorithm calculates similarity between nodes using graph-theoretic model. The basic equation of SimRank [4] is as given below:

$$s(a,b) = \frac{C}{\deg(a) * \deg(b)} \sum_{m \in Nbr(a)} \sum_{n \in Nbr(b)} s(m,n)$$

where $s(a,b)$ denotes SimRank score of nodes a and b, $\deg(a)$ denotes the degree of node a, $Nbr(a)$ denotes the set of adjacent nodes of a. C denotes the decay factor in propagation set to the value of 0.6[5]. As the algorithm is computationally very expensive, we implemented our own version incorporating the following optimizations suitable to our need:

- Instead of storing the entire NxN matrix of values, where N is the number of nodes, we use hashtable from pairs of nodes to the SimRank score to represent the sparse matrix.

- We pruned out nodes with score lower than a threshold $\tau = 0.01$ after each iteration [6]

- Instead of calculating the SimRank scores for each pair of nodes, we propagate the score from each pair of nodes for which value is calculated to its neighbours.

B. *Quantification of Factors Influencing Friendships*

Before attempting link prediction, we evaluated which topical interests are most influential on a user’s social network. This is an important part of the study to better understand what factors are more important for link prediction.

One way to find the topical interests that are most influential on users’ networks is by comparing how well known friendships of the ego nodes are correlated with their SimRank scores calculated on Single Interest Graphs. One way to express this score for a particular interest is:

$$Q_i = \frac{1}{|E|} \sum_{u \in E} \frac{1}{\deg(u)} \sum_{v \in Nbr(u)} s(u, v)$$

In the equations here, E is the set of all ego nodes in the network. To see how well this topical interest finds similarity between every ego node and every other node in the graph (not just neighbors as above), we also calculate:

$$\bar{Q}_i = \frac{1}{|E|} \sum_{u \in E} \frac{1}{|N|} \sum_{v \in N} s(u, v)$$

Both of these scenarios can be seen in Figure 3.

To compare the efficacy of topical interest i's ability to explain user friendships accurately, we evaluate its influence score: $(Q_i/\bar{Q}_i)$.

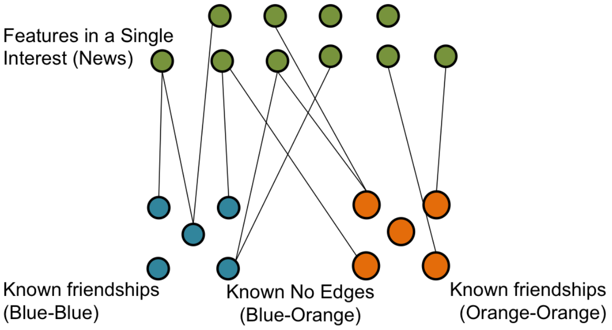

Fig. 3. Single Interest Subgraphs for SimRank

As such, we can rank the top interests by individually selecting the corresponding subgraphs and computing their influence scores $(Q_i/\bar{Q}_i)$.

C. Identifying Similar Features Within A Topical Interest

To identify similar features within a topical interest, we computed SimRank scores for pairs of features in Single Topical Interest Graphs. We then looked at the pairs of features which demonstrated highest similarity within the topical interest.

D. Identifying Similar Features Across Topical Interests

To identify similar features across topical interests, we computed SimRank scores for pairs of features in Dual Topical Interests Graphs. We then looked at the pairs of features across the two topical interests which demonstrated highest similarity.

E. Predicting Ego Networks For Users

Consider an ego node u for which we are interested in recommending friends. We know that any given node v_i can either be a friend of u (edge exists) or not a friend (no edge exists). Our hypothesis is that the existence of a friendship should be correlated with similar interests. Thus, we used the similarity values between users based on different topical interests to generate data on which we apply supervised learning models to recommend friendships. This is explained in greater detail here.

For an ego node u, each of the other nodes v_i correspond to a data-point. We labelled the data-point d_i as 1 if there exists a friendship between u and v_i, and 0 otherwise. A data-point is a vector of $n = 237$ features, where each feature $f_{i,j}$ corresponds to topical interest j and is calculated as follows:

$$f_{i,j} = \frac{s_j(u, v_i)}{\bar{Q}_j}$$

Thus we generated data in the following form:

Label(y)	Interest 1	$\cdots$	Interest n
$1[v_1 \in Nbr(u)]$	$f_{1,1}$	$\cdots$	$f_{1,n}$
$1[v_2 \in Nbr(u)]$	$f_{2,1}$	$\cdots$	$f_{2,n}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$1[v_m \in Nbr(u)]$	$f_{m,1}$	$\cdots$	$f_{m,n}$

where 1[] is the indicator function defined as 1[true] = 1 and 1[false] = 0.

This data can now be used to train classifiers using supervised machine learning algorithms. For the purpose of evaluation, we split the data into training and test sets. From the above table, we sampled 70% of known neighbors of u (y = 1) and an equal number of non-neighbors (y = 0) in the training set. The remaining data-points were put in the test set.

We then used supervised machine learning techniques to train the classifier using training set and used the trained classifier to predict the labels of the test set. For each user that our classifier predicts y = 1, we can recommend that user as a friend to node u.

To evaluate the performance of our recommendation algorithm, we calculated the precision, recall and f-score of our predictions on the test set.

$$\text{precision} = \frac{\sum_i (1[y_i == 1] \wedge 1[y_i^P == 1])}{\sum_i (1[y_i^P == 1])}$$

$$\text{recall} = \frac{\sum_i (1[y_i == 1] \wedge 1[y_i^P == 1])}{\sum_i (1[y_i == 1])}$$

$$\text{f-score} = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

where y is the true labels vector for test set, y^P is the predicted labels vector for test set.

It is worth noting that the classifier will be retrained for each ego node u that we wish to make predictions for. This is because every user has a very different set of interests that affect their friend circles and thus we wanted to have a personalized link-prediction method for each person. For each ego node, along with making friendship recommendations, we also discovered the topical interests which influence their friend circles the most.

VII. FINDINGS & DISCUSSION

A. *Interests with Highest Influence Scores*

We applied SimRank on Single Topical Interest Graphs and used the quantification technique described in section VI-B to find Influence Scores of the top 20 topical interests. The results are in Table I

TABLE I

INFLUENCE SCORES

	Topical Interest	**Influence Score**
1	Design	135.62
2	Science	70.18
3	Music	59.29
4	Writer	58.43
5	Sports	45.08
6	Web	41.84
7	Entrepreneur	41.72
8	Blogger	39.74
9	Videogames	37.15
10	Gaming	32.58
11	Media	34.79
12	Games	32.58
13	Geek	30.04
14	TV	25.39
15	Entertainment	18.90
16	News	18.50
17	Celebrity	15.50
18	Actor	9.89
19	Comedy	5.68
20	Politics	3.81

It is interesting to see that topical interests in Table I are what one would often expect friendship to be correlated with, e.g. shared interest in politics, celebrities, sports, etc. However, it is even more interesting to observe the qualitative differences between the high influence score interests and the low influence score interests. We see that extremely broad categories that most people have interest in (such as Politics and Comedy) do not have a very high influence score. On the contrary, more specific interests (such as Science and Design) have much higher influence scores. This result has a very natural interpretation: factors that most influence one's friendships are those that are more specific and local to his/her ego network such as the professional activities they engage in or in the sports team they follow (geographically local).

B. *Similarity Within Topical Interests*

We ran SimRank on all Single Topical Interest Graphs to find similarities between features within each topical interest. These results can be used for intra-domain recommendations and clustering features. Some applications are finding genres within music, finding political parties with similar ideologies, etc.

A sample of our results are shown in Table II. It is noticeable that we can successfully identify similar features based on just their followership and users following only one of the two features are very likely to be interested in the other.

TABLE II

FEATURE SIMILARITY USING SIMRANK

	Interest	**Feature 1**	**Feature 2**	**SimRank Score**
1	Science	@BrookhavenLab (Brookhaven National Laboratory)	@argonne (Argonne National Laboratory)	0.278
2	Science	@kylejasmin (PhD student in cognitive neuroscience at UCL and NIMH)	@marcoiacoboni (UCLA Professor and neuroscientist)	0.191
3	Science	@alandove (TWiV co-host)	@profvrr (TWiV host)	0.182
4	Sports	@McLaren_eShop (e-shop for McLaren)	@thefifthdriver (official account of McLaren Mercedes Formula 1 Team)	0.296
5	Sports	@TaylorMadeGolf (Golf product company)	@Golfsmith (Golf Store)	0.200
6	Comedy	@ComedianSpank (Writer known for Kevin Hart:Let Me Explain)	@KevinHart (Comedian)	0.231

C. *Similarity Across Topical Interests (Cross-domain Recommendations)*

Using the Dual Topical Interests Graph for any pair of interests we can measure the the similarities between features within the two topical interests and thus use the scores for cross-domain recommendation. Some of the results, described in Table III are clearly able to capture how features are similar across interests. For example, people who are interested in sports and follow sports-persons also tend to watch sport-shows on TV.
We also capture some surprising similarities in our experiments. (See rows 6,7 in Table III)

D. *Predicting Users' Ego Networks*

Using a Random Forest classifier for predicting the ego-network of users (see section VI-E), we are able to recommend links with good precision/recall scores. A sample of our results can be seen in Table IV. It is once

TABLE III

Cross Domain Feature Similarity using SimRank

	Topical Interest Pair	Feature 1	Feature 2	SimRank Score
1	Actor-TV	@humphreybogart (Famous American Screen Actor in 1940s)	@oldfilmsflicker (Old movies)	0.104
2	Sports-Science	@wingoz (NFL Live host)	@sport_science (ESPN show blows the lid off the mysteries of sports)	0.052
3	TV-Sports	@espn_firsttake	@shawnemerriman (3x Pro Bowler)	0.247
4	TV-Sports	@rajmathai (Weeknight News Anchor + Sports Director)	@sfgiants (American professional baseball franchise)	0.088
5	Technology-TV	@science	@fantasysf (Fantasy and science fiction news)	0.132
6	Politics-Comics	@thedemocrats	@guardsofdagma (Comic book from @tcwnyc called bizarre and unnerving)	0.034
7	Politics-Comedy	@obamanews	@gary_c_king (crime author and serial killers expert Gary C. King)	0.069

again interesting to note that for many nodes, the most influential topical interests used to predict friends and non-friends are very specific and local interests that are thematically related. For instance for NID xx840, we see that the most influential topical interests are Music, Fashion, Food, Art, Shopping and Style (all culturally themed topical interests). Similarly, for NID xx615, we see that the most influential topical interests are Media, Gaming, Esports, Sports, Videogames, and Football.

Thus, we see that using similarity between users (based on their topical interests), we can predict links with reasonably high precision and recall. This technique can be employed for link prediction in social networking websites as a supplement to methods based on user networks to allow users to find friends and grow their networks.

VIII. Next Steps & Conclusion

In this study, we make some interesting observations about human behavior on social networking websites. We find that users tend to have common followership with their friends and certain topical interests are more correlated (and maybe influential) with social circles. More specifically, we see that shared interests relating to careers, hobbies, geographical location, etc. (i.e. localized interests) are better predictors of friendships than are interests relating to politics, news, TV, etc. (i.e. global interests).

TABLE IV

Predicting Ego Networks

NID	Prec.	Recall	F1	Topical Interests Ranked
xx656	0.173	0.319	0.224	Boston, Geek, Humor, Artist, Blogger, Tv, Writer
xx469	0.121	0.333	0.177	Gaming, Comedy, Art, Comics, Tech
xx451	0.074	0.786	0.135	Tech, sports, Podcast, Comedy, Entertainment
xx840	0.089	0.267	0.135	Music, Fashion, Food, Art, Shopping, Style
xx007	0.079	0.338	0.129	Artist, Videogames
xx694	0.058	0.622	0.106	Sports, Gaming, Media, Videogames, Esports, Gamedesign, Gamesindustry
xx297	0.053	0.31	0.091	Vegan, Blogger, Charity, Editor, Education, Journalist, News, Writer
xx842	0.043	0.514	0.08	Podcast, Sports, Baseball, Blogger, Entrepreneur, Gamedesign, Gaming, Nba
xx615	0.04	0.653	0.076	Media , Gaming, Esports, Sports, Videogames, Football
xx786	0.035	0.595	0.067	Government, Museum, News, Blogger, Development, Programming, Women

Using this insight, we proceeded to use supervised learning techniques on similarity-based features to predict and reconstruct users' ego networks with reasonably high precision and recall. The success of this similarity-based approach confirms our initial hypothesis that friends on social networking websites such as Twitter have similar followership and share more common interests when compared to a random person on the network.

We were also able to use the similarity-based approach to successfully identify features similar within and across topical interests. This technique can be used to make intra-domain and cross-domain recommendations to users to improve their experience on the social networking platform.

We envision that our similar based approach can be augmented with current state of the art methods for link prediction on real-world networks.

REFERENCES

[1] J. McAuley and J. Leskovec. Learning to Discover Social Circles in Ego Networks. NIPS, 2012.

[2] S. Y. Stoyanova, Factors influencing the choice of friends - analysis of bulgarian friendship networks, vol. VIII, pp. 93109, 2008.

[3] V. R. K. Garimella and I. Weber, Co-following on Twitter, CoRR, vol. abs/1407.0791, 2014. [Online]. Available: http://arxiv.org/abs/1407.0791

[4] G. Jeh and J. Widom, Simrank: A measure of structural-context similarity, in Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ser. KDD 02. New York, NY, USA: ACM, 2002, pp. 538543. [Online]. Available: http://doi.acm.org/10.1145/775047.775126

[5] D. Lizorkin, P. Velikhov, M. Grinev and D. Turdakov. Accuracy Estimate and Optimization Techniques for SimRank Computation. In VLDB '08: Proceedings of the 34th International Conference on Very Large Data Bases, pages 422–433.

[6] Weiren Yu; Xuemin Lin; Jiajin Le, "A Space and Time Efficient Algorithm for SimRank Computation," Web Conference (APWEB), 2010 12th International Asia-Pacific , vol., no., pp.164,170, 6-8 April 2010