A Micromodel to Predict Message Propagation for Twitter Users

Prasanth G Rao, Venkatesha M
VTU, Belagavi-590 018, India.
Email: prasanthgrao@gmail.com

Anita Kanavalli
RIT, Bangalore-560 054, India.

P Deepa Shenoy, Venugopal K R
UVCE, Bangalore-560 001, India.

***Abstract*—Twitter is the most popular social platform for broadcasting opinions but the reach of tweets is often non-deterministic. While people believe the world is taking cognizance of their tweets, this seldom is the truth. This paper presents a micro-prediction model for determining message propagation for a user, especially for the non-influential majority. Our framework uses Ego network in predicting message propagation. The work focuses on determining the possible users who would interact and their immediate reach. This is achieved by using Twitter API in a limited manner. We attempt to make a responsive prediction model; simple, stateless and scalable, capable of catering to parallel requests. The simulation results shows a promising 86.4% accuracy for data constituting of 336768 connected users.**

***Index Terms*—Twitter, Ego Network, Message propagation, Decision Trees**

I. Introduction

Social Media has grown enormously from being a platform for information exchange and people networking to becoming an integral part of human lives. Social networks are known to influence personality attributes, trust factors and define the very idea of social capital [1]. People with good repute on networking sites have a chance to better opportunities, both in professional and personal space. The impact of social web mining on Sociology, Governance, and Economics is acknowledged. Enterprises invest considerable resources in creating an engaging narrative for its customer on social platforms. The recent example is the Ice Bucket challenge where a simple activity became a global phenomenon and managed to generate donations worth millions of dollars. Psephologists are successful basing their predictions on results from sentiment analysis of the opinions shared by the users on social platforms. Twitter polls are a common sight in news debates. 83% of the world leaders have a Twitter account. The usage of social networks in news media and governance gives the common citizen a sense of might and inclusion. Despite the widespread usage, reality is quite different and communications are largely one-sided. Comparing an average user to an influential user shows the stark contrast. Rouge users such as spammers, extremists, etc. often use the service for questionable causes but go unnoticed. People in distress too often reach to Twitter for help. In a system with 40% of the data classified as pointless babble, it becomes impossible to determine the impacted users for the given message.

Twitter is the most popular microblogging service with 330 million active users sharing 500 million *tweets* or messages each day. Table I details the recent usage statistics [2]. Each message can be 280 characters long and supports multimedia content. Apart from sharing opinions, users can tweet to any other public users using their address called a handle beginning with @ symbol. Users can endorse any tweets by sharing it(also known as retweeting) or liking a tweet using the favorite feature. User relations are formed by the means of *following* and *friends*. Let us say Jack is a Twitter user and Jill follows him. That makes Jill a friend of Jack. For Jack, Jill becomes a follower. Jill being Jack's follower will receive all his updates. Jack would receive Jill's tweets only if he chooses to follow her. Thereby it might be clear now that the social graph formed in case of Twitter is a directed graph.

TABLE I: Twitter usage statistics

Attribute	Count
Total Users	1.3 billion
Active Users	550 million
Monthly active users	330 million
Verified users	293,027
Bots	23 million
Accounts without followers	391 million
Tweets per day	500 million
Average followers per user	707

This paper is organized as follows. Section II outlines the significant contributions made to the field thus far. Section III introduces the problem and describes the objectives of our work. Section IV details out the solution with insights into implementation. Section V presents the outcome of our execution. Concluding section VI briefly discusses the next steps.

II. Related Work

A. Small World Phenomenon

Milgram [3] first presented the notion of Small World Phenomenon, also known as the Six Degrees of Separation. Milgram created an experiment to determine the shortest paths of acquaintances required to reach each other. The setup involved delivering a letter addressed to a person in Boston. Milgram found the letter changed hands six times to reach the target. Milgrams experiment had two important discoveries. Firstly, it established the existence of a short path between otherwise unrelated people. Secondly, it showed people collaborating within their independent capabilities could deliver

978-1-5386-4855-1/18/$31.00

the letter outside their immediate social circle. The work went on to define many significant research streams.

Jon Kleinberg [4] presents a decentralized algorithmic interpretation of the small world phenomenon. Kleinberg's model for the small world phenomenon is a k-dimensional matrix of nearest-neighbors. The distance measure is defined between points in the matrix x and y as $d(x,y)^{-k}$. Equation 1 gives the Probability p of the shortest routing path.

$$p(x \leftrightarrow y) = \frac{d(x,y)^{-k}}{H_k(n)} \tag{1}$$

$H_k(n)$ is a normalization constant.

Robert E Hiromoto [5] further explores the Kleinberg's model with random graphs in Neuroanatomical networks elaborating on the complexities and concerns around parallelism. He explores data communication schemes over different topologies to discover the algorithm succeed to avoid random uncertainties to a good extent. The concept of small world phenomenon is extensively applied to Social Networks. Facebook research [6] found mean degree of separation to be at 3.57 for 1.59 billion active users. Masaru Watanabe et al. [7] deduce the mean degree of separation for twitter at 4.59. Noticeably both the numbers are much lesser than 6 signifying the strong interlinking among users.

B. User Graph

Much work has gone into the field of user characterization and discovering user networks. Hughes et al. [8] define the Big-Five personality predictors for social networking sites. The Big-Five consists of five broad personality traits, namely, Neuroticism, Extraversion, Openness, Agreeableness, and Conscientiousness. Natural graphs, such as social networks, email graphs, or instant messaging patterns, have become pervasive through the Internet. These graphs are massive, often containing millions of nodes with billions of edges. Ahmed et al. [9] addresses the issue of factorizing such natural graphs using vertex partitioning algorithms. The algorithms are developed using distributed methods. The partitioned graphs have vertices labeled $owned$ and $borrowed$. The $borrowed$ vertices are shared between graphs and used for convergence and completeness. Ahmed et al. [10] extend this further with regional context. The framework detects regional contexts from regional models and language models and identifies the geographic locations for the information shared on microblogging services. Lin et al. [11] has done extensive work to address the problems of extracting and analyzing communities, but the factors that drive their formation are still not well understood. Papadopoulos et al. [12] have defined explicit and implicit communities, and have discussed the strategies for Scalability. Roth et al. [13] suggest friends with detecting implicit graph based on email exchanges. Graphs are constructed based on users addressed in email exchanges, together with weighting functions for edge priorities, implicit graphs are derived.

C. Ego Networks

Ego networks is a micro-graph outlining a person(ego) and his interactions with the other people($alter$) in his neighborhood. Figure 1 shows the representation as put forward by Arnaboldi and his colleagues [14]. Increasing diameter of circles(also called Dunbar's circles) implies decreasing proximity of the relation. The work establishes the potential of Ego networks in learning the cognitive properties that define human relations in the real world. McAuley and Leskovec [15] use Ego networks to learn implicit social circles on Twitter. The circles are defined based on Features such as hometowns, birthdays, colleagues, political affiliations, etc.

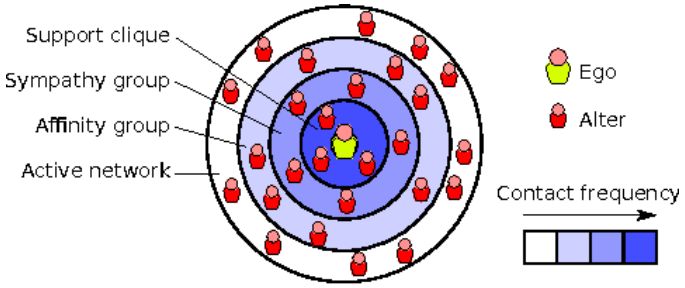


Fig. 1: Ego Network Model

D. Predicting retweets

Retweets can be predicted by analyzing statuses. Paper [16] handles streaming prediction of retweets using time-sensitive models. Authors establish predicting retweets are possible with 202 tweets predicted by two subjects. The first subject predicts with only the text of tweets with an accuracy of 76.2%. The second subject predicts with an accuracy of 73.8% tweets containing social circle information. Finally proposed unsupervised model predicts with an accuracy of 69.3% and supervised model with 82.7%. Chenhao Tan et al. [17] use Topic and Author-Controlled(TAC) pairs to predict retweets based on the wording used in the tweets.

Concerns surrounding privacy violations, data misuse and accusations of content profiling for illegitimate causes have made social data much less accessible. All social networking sites including Twitter enforce strict security measures and publishes limited data via highly monitored API implementations. Lots of research goes around identifying influential users in blogging [18], [19] and microblogging forums. However, we must recognize that influential users are quick to get their accounts verified to avoid damages from fake accounts and legal implications. Twitter has 239K verified users against 330 million accounts [20] i.e., less than 0.001% for all the users. With limited data on disposal, qualitative research in social media should be capable of determining actionable tweets and users among a heap of forwards, spams and other tweets employing light and stateless algorithms. Our work is an attempt in this direction. We use Ego networks instead of status analysis to predict user retweets. The approach is expected to give us shorter computing times with similar results.

III. Problem Definition

Given a Twitter user, devise a micro-prediction model working with limited data calls to Twitter API. The objective

TABLE II: Rate limits per 15 minutes window

Endpoint	Resource family	Requests / window (user auth)
GET followers/list	followers	15
GET users/lookup	users	900
GET statuses/lookup	statuses	900
OVERALL	ALL	900

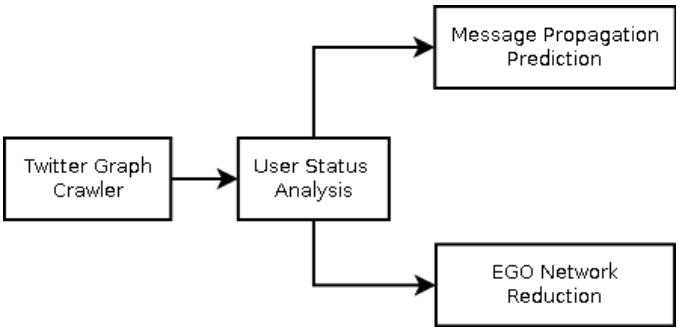


Fig. 2: Architecture of Message Propagation Predictor

of the model is as follows.

1) Predict retweet probability for the user
2) Plot possible reduced Ego network for message propagation

Table II outlines the APIs consumed from Twitter with their respective rate limits. We fetch the required data for analysis in a single window. The micro-prediction framework will be scalable and stateless to ensure the framework can run at any number of parallel instance with any volume of data. We restrict our analysis to the overall graph structure and observable user history. This restriction is made keeping in mind the rate limits enforced by Twitter. We have also dropped features such #hashtags and other such semantic characteristics from the current ambit of prediction.

IV. Methodology

Figure 2 outlines the design of micromodel prediction framework implemented. There are four components as seen.

- *Twitter Graph Crawler*: Collects the user and friends information with breadth-first traversing. The approach ensures data is read as an Ego network with relevant neighborhood information. We read users till Level 2, with level 0 being the Ego user.
- *User Status Analysis*: Derive statistical measures for the tweets/statuses for each user in the graph. Compute average following and friends for followers, turnaround time per tweet and retweets per tweet. Follower information serves as inputs to the prediction model, and remaining data is used for output validation.
- *Message Propagation Prediction*: Decision tree based classifier to predict if the user will receive retweets.
- *Ego Network Reduction*: Determine users for which Ego network for level larger than 1 exist and generate weighted Ego network. Prune the connections based on turnaround time and retweet probability. The approach is retrospective.

The data is generated in a linear order to suit rule-based classifiers such as Decision Trees. This would also ensure shorter computation times. The known delay in processing times is attributed to the rate limits enforced by Twitter. We take the friends, followers, status count directly from the Twitter API response. From followers of user, we determine the simple average of followers and friends as shown in equation 2.

$$\overline{x} = \frac{\sum_1^n x_i}{n} \tag{2}$$

From the status feed, sample of recent 10($= n$) tweets are taken and the interval for create date, d between the tweets are computed as shown in equation 3.

$$\overline{d} = \frac{\sum_1^{n-1} (d_{i+1} - di)}{n-1} \tag{3}$$

We define the Ego networks as a graph $G(V, E)$, where V is the weighted node defined as $V(\overline{d}, \overline{retweet})$. E denotes $following$ relations. The graph is a temporal graph with the timescale defined by d. At any given distinct d', the set of $V'(\overline{d} > d')$ represented the probable users to respond. Thus V'/V gives the probability of retweet.

We use J4.8 classifier, an improvement over C4.5 decision tree, and Random Forest for predicting retweets. Random Forest [21] uses a combination of decision trees in predicting the outcomes by independently voting the most popular class as output. C4.5 [22] works towards minimizing Information Entropy $H(T)$ while maximizing Information Gain $IG(T, a)$.

$$H(x) = -K \sum_{i=1}^{n} P(x_i) \log P(x_i) \tag{4}$$

$$IG(T, a) = H(T) - \sum_{v \in vals(a)} \frac{\{x \in T | x_a = v\}}{|T|} \cdot H(\{x \in T | x_a = v\}) \tag{5}$$

$P(x)$ is the Probability function. T denotes the training set $T(x, c) = (x_0, x_1, ..., x_n, c)$ with x the attributes and c the class label.

A total of 336768 connected users were analyzed, 243 users had Ego network available till level 2. To avoid human intervention, we use level 2 nodes to determine the expected classifier output.

V. Result

Dataset generation is implemented entirely in Java threads. Weka [23] is used to implement classifier and Graph analysis is done using Gephi [24].

A. Dataset

The data required for our tests was downloaded using a custom graph crawler invoking a series of Twitter API methods. The crawler executed for 4 days collecting roughly 40GB data comprising of users and tweets. Data set details

TABLE III: Input Twitter Data

Entity	Total Instances
Users(V)	336768
Relations(E)	395504
Statuses	264298

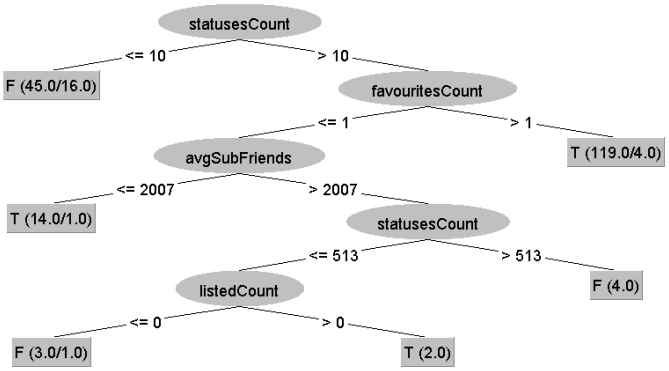


Fig. 3: J4.8 Decision Tree

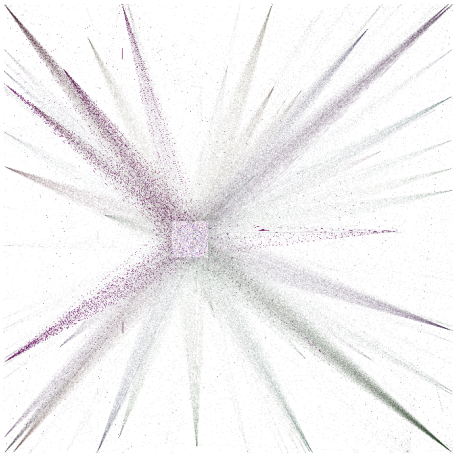
are outlined in table III. To optimize the API invocations, we have assumed users with more than 5000 followers as a local influencer and they are excluded from further probing. Similarly users following over 5000 accounts are excluded as their likelihood to respond to tweets are less. Upon constructing the Ego networks, we are left with 243 users having relations with about 33600 users and 395504 connections. Dataset description for the classifier is detailed in table IV.

B. Retweet Prediction

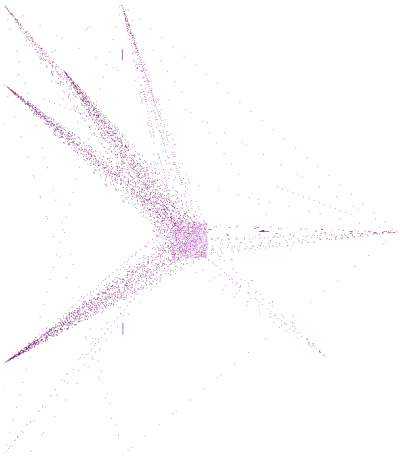
Table V captures the classifier performance. Overall prediction accuracy with J4.8 is 80.7% and success rate for predicting actual retweets is 86.4%. The algorithm does not particularly do well for negative cases with an accuracy of 60%. However, our objective of determining a retweet is well addressed with the current approach. J4.8 decision tree model is shown in figure 3. We see the attribute $avgSubFriends$ in decision tree signifying the importance of connected followers in getting retweets. Random Forest fares better with 82.7% but is overtrained to identify positive retweets. For users with no retweets, the algorithm has accuracy of only 47.5%.

C. Ego Networks

Figure 4 shows the Ego networks for user @*prasanthgrao*. Figure 4a shows all the users following @*prasanthgrao* till level 2. The distance of any node from center is inversely proportional to the $\overline{retweet}$. Figure 4b shows pruning outcome for $\overline{d} = 1day$. Total users are seen reduced to 10%. The number of edges or $following$ reduces to 1.86%, which also denotes the probability of retweet. The result confirms our assumption that Ego network of depth 1 provides the necessary information for our prediction model.



(a) Ego Network for user prasanthgrao. Level 2. 24635 users



(b) Ego Network after pruning. 2137 users

Fig. 4: Ego Network for user @*prasanthgrao*

VI. Conclusion and Future Work

The present work predicts the probability of a user getting retweets and identifies the possible paths for the interactions with limited data. We have only considered observable statistical characteristics. The available dataset does not perform well for predicting true negatives. Twitter APIs allow only a maximum of 900 requests every 15 minutes. Thus the complete required dataset comes with a time lag. Twitter also provides streaming APIs for real-time entries, but the response is a sampled output instead of the entire status set. Therefore a heuristic self-learning solution centered around Fuzzy or Rough set approach can help us make a real-time streaming based micro-prediction model working with a highly limited data source.

References

[1] S. Valenzuela, N. Park, and K. F. Kee, "Is there social capital in a social network site? facebook use and college students' life satisfaction,

TABLE IV: Dataset description

Attribute	Data Type	Description
screenName	String	Twitter handle/account
followersCount	long	Accounts following user
friendsCount	long	Accounts followed by user
statusesCount	long	Tweets published by user
favouritesCount	long	Total favorites
listedCount	long	Number of lists the user features in
tweetInternal	long	Average interval between tweets
avgSubFollowers	long	Average of accounts following user's followers
avgSubFriends	long	Average of accounts followed by user's followers
hasRetweets	boolean	Expected output

TABLE V: Detailed Accuracy for Classifiers

Algorithm	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.864	0.400	0.888	0.864	0.876	0.448	0.643	0.816	**T**
J4.8	0.600	0.136	0.545	0.600	0.571	0.448	0.643	0.366	**F**
	0.807	**0.344**	**0.815**	**0.807**	**0.811**	**0.448**	**0.643**	**0.720**	**Overall**
	0.918	0.525	0.865	0.918	0.891	0.434	0.856	0.949	**T**
Random Forest	0.475	0.082	0.613	0.475	0.535	0.434	0.856	0.561	**F**
	0.824	**0.430**	**0.811**	**0.824**	**0.815**	**0.434**	**0.856**	**0.866**	**Overall**

trust, and participation," *Journal of computer-mediated communication*, vol. 14, no. 4, pp. 875–901, 2009.

[2] K. Smith. (2017) 44 incredible and interesting twitter statistics. [Online]. Available: https://www.brandwatch.com/blog/44-twitter-stats/

[3] J. Travers and S. Milgram, "The small world problem," *Psychology Today*, vol. 1, no. 1, pp. 61–67, 1967.

[4] J. Kleinberg, "The small-world phenomenon: An algorithmic perspective," in *Proceedings of the 32nd Annual ACM symposium on Theory of computing*. ACM, 2000, pp. 163–170.

[5] R. E. Hiromoto, "Parallelism and complexity of a small-world network model," *International Journal of Computing*, vol. 15, no. 2, pp. 72–83, 2016.

[6] S. Bhagat, M. Burke, C. Diuk, I. O. Filiz, and S. Edunov. (2016) Three and a half degrees of separation. [Online]. Available: https://research.fb.com/three-and-a-half-degrees-of-separation/

[7] M. Watanabe and T. Suzumura, "How social network is evolving?: a preliminary study on billion-scale twitter network," in *Proceedings of the 22nd International Conference on World Wide Web*. ACM, 2013, pp. 531–534.

[8] D. J. Hughes, M. Rowe, M. Batey, and A. Lee, "A tale of two sites: Twitter vs. facebook and the personality predictors of social media usage," *Computers in Human Behavior*, vol. 28, no. 2, pp. 561–569, 2012.

[9] A. Ahmed, N. Shervashidze, S. Narayanamurthy, V. Josifovski, and A. J. Smola, "Distributed large-scale natural graph factorization," in *Proceedings of the 22nd International Conference on World Wide Web*. ACM, 2013, pp. 37–48.

[10] A. Ahmed, L. Hong, and A. J. Smola, "Hierarchical geographical modeling of user locations from social media posts," in *Proceedings of the 22nd International conference on World Wide Web*. ACM, 2013, pp. 25–36.

[11] Y.-R. Lin, J. Sun, H. Sundaram, A. Kelliher, P. Castro, and R. Konuru, "Community discovery via metagraph factorization," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 5, no. 3, p. 17, 2011.

[12] S. Papadopoulos, Y. Kompatsiaris, A. Vakali, and P. Spyridonos, "Community detection in social media," *Data Mining and Knowledge Discovery*, vol. 24, no. 3, pp. 515–554, 2012.

[13] M. Roth, A. Ben-David, D. Deutscher, G. Flysher, I. Horn, A. Leichtberg, N. Leiser, Y. Matias, and R. Merom, "Suggesting friends using the implicit social graph," in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2010, pp. 233–242.

[14] V. Arnaboldi, M. Conti, A. Passarella, and F. Pezzoni, "Ego networks in twitter: an experimental analysis," in *IEEE International Conference on Computer Communications Workshops*. IEEE, 2013, pp. 229–234.

[15] J. Leskovec and J. J. Mcauley, "Learning to discover social circles in ego networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 539–547.

[16] S. Petrovic, M. Osborne, and V. Lavrenko, "RT to win! predicting message propagation in twitter." *ICWSM*, vol. 11, pp. 586–589, 2011.

[17] C. Tan, L. Lee, and B. Pang, "The effect of wording on message propagation: Topic-and author-controlled natural experiments on twitter," *arXiv preprint arXiv:1405.1438*, 2014.

[18] G. Vasanthakumar, R. Priyanka, K. V. Raj, S. Bhavani, B. A. Rani, P. D. Shenoy, and K. Venugopal, "PTMIB: Profiling top most influential blogger using content based data mining approach," in *International Conference on Data Science and Engineering (ICDSE)*. IEEE, 2016, pp. 1–6.

[19] G. Vasanthakumar, P. D. Shenoy, and K. Venugopal, "PFU: Profiling forum users in online social networks, a knowledge driven data mining approach," in *IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE)*. IEEE, 2015, pp. 57–60.

[20] Twitter. (2018) Twitter verified. [Online]. Available: https://twitter.com/verified/following

[21] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[22] J. R. Quinlan, "C4. 5: programs for machine learning," 2014.

[23] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[24] M. Bastian, S. Heymann, M. Jacomy *et al.*, "Gephi: an open source software for exploring and manipulating networks," *ICWSM*, vol. 8, pp. 361–362, 2009.