2018 IEEE International Conference on Data Mining Workshops (ICDMW)

Sentiment Prediction in Social Networks

Shengmin Jin
Data Lab, EECS Department
Syracuse University
shengmin@data.syr.edu

Reza Zafarani
Data Lab, EECS Department
Syracuse University
reza@data.syr.edu

***Abstract*—Sentiment analysis research has focused on using text for predicting sentiments without considering the unavoidable peer influence on user emotions and opinions. The lack of large-scale ground-truth data on sentiments of users in social networks has limited research on how predictable sentiments are from social ties. In this paper, using a large-scale dataset on human sentiments, we study sentiment prediction within social networks. We demonstrate that sentiments are predictable using structural properties of social networks alone. With social science and psychology literature, we provide evidence on sentiments being connected to social relationships at four different network levels, starting from the ego-network level and moving up to the whole-network level. We discuss emotional signals that can be captured at each level of social relationships and investigate the importance of structural features on each network levels. We demonstrate that sentiment prediction that solely relies on social network structure can be as (or more) accurate than text-based techniques. For the situations where complete posts and friendship information are difficult to get, we analyze the trade-off between the sentiment prediction performance and the available information. When computational resources are limited, we show that using only four network properties, one can predict sentiments with competitive accuracy. Our findings can be used to (1) validate the peer influence on user sentiments, (2) improve classical text-based sentiment prediction methods, (3) enhance friend recommendation by utilizing sentiments, and (4) help identify personality traits.**

***Index Terms*—Sentiment Prediction, Social Networks**

I. Introduction

Emotions impact different aspects of our daily lives from how we make decisions [1] and learn [2] to our overall health [3]. Social media sites have become the primary online venue for users to express their emotions via positive and negative sentiments. Social media users can express sentiments via blog posts, comments, photos, and likes, among other interactions. Social relationships are central to the formation of sentiments [4], [5]. However, the bulk of research on sentiment prediction has utilized text (in place of network structure) for predicting sentiments. Recent studies have highlighted emotional contagion among friends [6], indicating the possibility of using network structure for sentiment prediction. In this paper, we explore this possibility and investigate sentiment prediction using social relationships. This investigation allows us to answer questions such as: Can we predict an individual's sentiment based on the sentiments of her friends? Are sentiments of users with many friends more predictable? Which types of social relationships or network structures help best predict one's sentiments? We systematically investigate the utility of network structure for sentiment prediction at four different network abstraction levels: the ego-level, the triad-level, the community-level, and the whole network-level. At each network abstraction level, we capture structural properties that we speculate can assist in sentiment prediction.

Ego-Level Analysis. At the ego-level, we investigate whether sentiments expressed by directed (follower/followee) or undirected (friends) connections of a user can help predict her sentiments. At this level, we aim to exploit sentiments expressed by pairs of individuals (i.e., *dyads*) for prediction purposes.

Triad-level Analysis. At the triad-level, we generalize ego-level analysis by investigating whether sentiments expressed by members of the triads (three connected users) that an individual is a part of can help predict her sentiments. Studying sentiments in triads raises the possibility of connecting this study to structural balance [7] and status theory in which triads with signed edges (e.g., friendly/antagonistic relationships) can be used for prediction purposes. We investigate this possibility.

Community-Level Analysis. At the community-level, we explore the possibility of relating one's sentiments to the communities that the user has joined and the sentiments expressed by their members.

Network-Level Analysis. Using the whole-network information, we investigate whether structural properties at the macro (whole network-level) or micro (node-level) level can help predict one's sentiment.

At each network-level, we identify (1) structural properties that help best predict sentiments and (2) how prediction performance varies as more network information becomes available. We make the following contributions:

- We provide evidence on how sentiments and network structural properties are connected at various network levels;
- We demonstrate the feasibility of predicting sentiments by exploiting various network structures and with different levels of information availability;
- We assess the importance of structural information at different network levels for sentiment prediction; and
- By comparing network-based with text-based sentiment prediction methods, we identify (1) cases in which each method performs best and demonstrate (2) the trade-off between network information and text for sentiment prediction.

The rest of the paper is organized as follows. To follow a systematic approach, Section II highlights the natural connections that have been identified between sentiments and social

2375-9259/18/$31.00 ©2018 IEEE
DOI 10.1109/ICDMW.2018.00190

ties, mostly within social sciences. These findings inspire the principal direction of the paper, where we capture the aforementioned connections via machine learning experiments outlined in Section III. We review further related work in Section IV and conclude in Section V with future directions.

II. Links between Sentiments and Network Structure

Social media websites have become important channels through which users can develop their social relationships. Users can befriend or follow other users, and can form communities or join existing ones. Here, using social science literature, we provide evidence on sentiments being connected to social relationships at different network levels. The discussion will mainly focus on the emotional signals that can be captured at each level of social relationships.

A. Ego Level

Psychological research has provided evidence that individuals are happier when they are with others. When asked by the National Opinion Research Center, "How many close friends would you say you have?" (excluding family members), 38% of individuals reporting five or more friends indicated that they were "very happy" [8]. A recent study corroborates this finding: having more friends on social networks may enhance a user's subjective well-being [9]. Both studies indicate that friendships have a considerable impact on psychological well-being. These findings direct us to look at the most basic social relationship in social networks: friendships, or equivalently, relationships in an undirected ego-network. Inspired by studies of emotional contagion among friends [6], we believe that the degree of happiness among one's friends may also provide evidence of his expressed sentiments. Similarly, due to the existence of emotional contagion among followers/followees [10], we believe that the directed ego-network of follower/followee relationships may also carry information on one's sentiments.

B. Triad Level

Consider the following: Emma and Noah are friends and both positive individuals. Liam is their mutual friend. Is Liam more likely to be a positive person? Similarly, does Emma and Noah both being negative lead to Liam also being negative?

A natural extension to predicting sentiments in the ego network (which often involves two users: ego and a friend) is to predict sentiments in sets of three connected individuals (a triad). Network structures involving three nodes have proven fundamental to understanding social networks, as (1) triads occur frequently due to transitivity *"A friend of my friend is my friend;"* and (2) a closed triad is the simplest complete graph and network motif, in which every pair of nodes is connected indicating a close social tie. In large-scale networks *assortativity* [11] is a common pattern. We have reason to believe that users in the same triad may share sentiments.

We can study the connection between social relationships in triads and user sentiments from different views. From a user's view, if a user is part of many triads with two other positive users, we speculate that the user is more likely to be positive. From the view of a triad, if two members are positive, we may expect the third one to be positive as well.

Similar to structural balance and social status theory, we can consider all the possible ways in which the three users in a triangle can be signed. Differing from these theories, we assign signs to nodes with user sentiments, but not to edges. We speculate that (1) specific configurations of triads are more frequent with respect to the sentiments of their members and that (2) sentiments of individuals that are in such configurations are easier to predict.

Here, we consider triads as the most basic network motif. However, our approach can be extended to higher-order structures or motifs that involve more nodes.

C. Community Level

In social networks, a community is formed when like-minded users become friends and start interacting with each other [11]. Naturally, we speculate users in the same community express similar sentiments. We can hypothesize that the user sentiments is influenced by the most positive or negative users in the community, or the overall community sentiment. On the contrary, we can speculate that users prefer joining communities whose members express similar sentiments.

D. Network Level

Connecting structural properties at the whole-network level with user attributes (including sentiments) is a topic less explored. However, the importance of network-level structure should not be ignored, as it is the only level that provides a global view of the network. At this level, we are not only considering one's friendships, but are also including information from friends of friends, three-hop connections, and even the connectivity at the whole-network level. We analyze the network-level structure from a (1) *macro view*, where the network is formed by linked users with different sentiments. If the way users choose to befriend or follow others is related to their sentiments, we should be able to observe it from this macro view; (2) *micro view*, where each user is embedded in the network differently. Consider a node embedding method [12] that maps nodes into a lower dimensional space. If sentiments are connected to the structural properties of the whole network, a node embedding that preserves structural similarity should result in similar embeddings for nodes expressing similar sentiments.

III. Sentiment Prediction using Network Structure

Our approach to sentiment prediction using social networks aims to identify structural properties that are related to sentiments. Hence, we start by constructing features on each network level that we speculate can help with sentiment prediction. Following feature construction, we conduct experiments to assess the importance of each network level and the validity of our speculations. Finally, we assess the predictive power of using all features. Before we delve into our experiments, we briefly discuss the dataset we use for the experiments in this paper.

A. Experimental Setup

To predict sentiments using network structure, we require data that contains social network information for users as well as their sentiments. As sentiment classification can be subjective and imprecise, exact ground-truth sentiments are preferred. For network information, there is a need for (1) directed and undirected relationships between users and (2) user community memberships. It is preferable that explicit community membership information is provided, as community detection can be subjective [13]. Finally, to compare network-based sentiment prediction with text-based techniques, the data should contain user-generated text such as posts [14]. In our previous study on emotions within social networks [15], we have crawled one such dataset from social networking/blogging site LiveJournal (http://livejournal.com/).

Users on LiveJournal can maintain a blog and can have (1) friends (undirected/mutual relationships), (2) directed relationships (follower/following) and be in (3) explicit communities (create or join a community). When users post, they have the option of reporting their sentiment by selecting a *mood*, which can be selected from a predefined alphabetically-ordered list of 132 common moods such as `happy` or `angry`, or can be entered as free-text. Our dataset contains 10 years of LiveJournal data, including 14,767,000 posts, where each post includes the sentiment directly provided by the user, 1,135,000 friendships, 14,196,000 million follower/followee relations, and the community memberships for all users. The data is publicly available at https://data.syr.edu/get/EmotionPatterns/

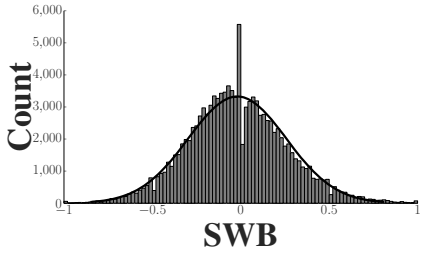
Data Preprocessing. We preprocess the dataset by

- Retaining the posts that have their moods selected from the predefined list provided by LiveJournal. This allows consistency in sentiment analysis and removes meme-type moods. Predefined mood posts account for the majority (85.96%) of posts within our data;
- Excluding infrequent or inactive users by removing users that have fewer than 10 posts; and
- Manually labeling each mood with its polarity (`positive`, `negative`, or `neutral`). After this step, all moods in the datasets are either positive (+), negative (−), or neutral (0).

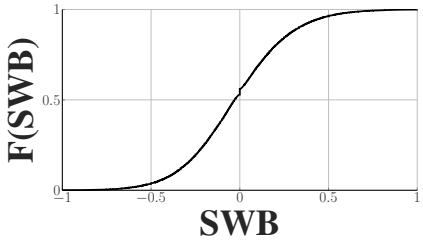
After data preprocessing, we use the previously proposed Subjective Well-Being (SWB) [16] to quantify the sentiments of users. SWB is defined as the fractional difference between the number of positive and negative posts:

$$S(u) = \frac{N_p(u) - N_n(u)}{N_p(u) + N_n(u)}, \tag{1}$$

where $S(u)$ denote the subjective well-being of user u, and $N_p(u)$ and $N_n(u)$ represent the number of positive and negative posts for user u, respectively. In our dataset, the SWB distribution is approximately normal, which can be observed by the normal fit in Figure 1a. The empirical cumulative distribution function (CDF) of the $S(u)$ values (Figure 1b) indicates a slight skew towards users with more negative posts, i.e., $P(S(u) < 0) > 0.5$. Hence, we consider users with



(a) Sentiments Histogram with Normal Fit (Solid Line)



(b) Sentiments Empirical CDF

Fig. 1: User Sentiment Distribution

TABLE I: Positive, Negative, and Neutral Users Distribution

Users	Number	Proportion
Positive (+)	50,705	43.92%
Negative (−)	61,066	52.90%
Neutral (0)	3,673	3.18%
Total	115,444	100.00%

$S(u) > 0$ as positive (+) users, with $S(u) < 0$ as negative (−) users, and with $S(u) = 0$ as neutral (0) individuals.

Table I provides the distribution of users with positive, negative, and neutral sentiments. The majority of users express negative sentiments and negative users are almost 20% higher than positive users. Since neutral users account for only 3% of the population, we remove them from the network and predict sentiments for users that are either positive or negative. Among the remaining users, there are 37 users (21 are positive and 16 are negative) that are not in friendship or follower/followee networks. We also remove these users as they do not carry link information for prediction.

After data preparation, we construct features for prediction. Table II provides our feature set on four network levels: Ego-level, Triad-level, Community-level, and Whole-Network Level. Following feature construction, we assess the effectiveness of network features via the following experiments. In our experiments, we use 10-fold cross validation and logistic regression as our classifier. Finally, we compare our approach with text-based methods, and discuss the trade-off between network information and text.

B. Ego level

We conduct experiments at the ego-level by investigating both undirected and directed ego-networks.

TABLE II: Feature List

Ego-Level (undirected)	# of friends
	# of positive friends
	# of negative friends
Ego-Level (directed)	# of followers
	# of followees
	# of positive followers
	# of negative followers
	# of positive followees
	# of negative followees
Triad-Level (undirected)	# of $(+,+)$ pairs
	# of $(+,-)$ pairs
	# of $(-,-)$ pairs
Triad-Level (directed)	# of $(+,+)$ pairs in non-rotatable triad
	# of $(+,-)$ pairs in non-rotatable triad
	# of $(-,-)$ pairs in non-rotatable triad
	# of $(+,+)$ pairs in rotatable triad
	# of $(+,-)$ pairs in rotatable triad
	# of $(-,-)$ pairs in rotatable triad
	Count of 16 positions
Community-Level	# of communities
	# of positive communities
	# of negative communities
	Fraction of positive communities
	Average SWB of the communities
	Maximum SWB of the communities C_{max}
	Minimum SWB of the communities C_{min}
Network-Level	Kronecker Features
	Unweighted NODE2VEC
	Weighted NODE2VEC

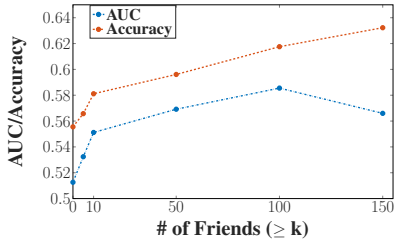


Fig. 2: Sentiment Prediction Performance (Accuracy/AUC) for Users with k or more Friends

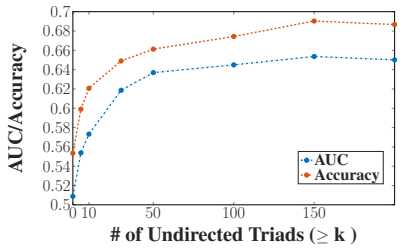


Fig. 3: Sentiment Prediction Performance (Accuracy/AUC) for Users that are in k or more Undirected Triads

Undirected Ego Networks. We speculated that the number of friends and the degree of happiness of friends may help predict one's sentiments. The number of friends of a user is simply the user's degree in the friendship graph, which we include as one feature. To quantify the degree of happiness among the friends, we include the number of positive friends and number of negative friends as two features. With these three features, the prediction accuracy is about 56%. Intuitively, the prediction accuracy must differ for users with a different number of friends, as more structural information becomes available. To verify this speculation, we plot the accuracy and the area-under-the-curve (AUC) for users that have k or more friends in Figure 2. We observe an increasing trend, validating our speculation. The accuracy reaches 63% for users with more than 150 friends. As users with more than 150 friends take up only 1% of our data, we do not study users with more friends.

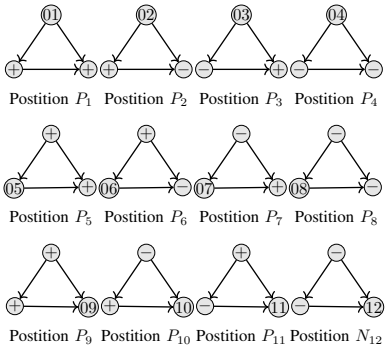
Directed Ego Networks. We extend our study to directed networks where follower/followee types of edges are present. As edges are directed, we double the features we use for undirected networks: # of followers, # of positive followers, # of negative followers, # of users followed, # of positive users followed, and # of negative users followed. With these six features, we obtain an accuracy of 59%, and reach an accuracy of 63% with users following more than 100 users.

C. Triad level

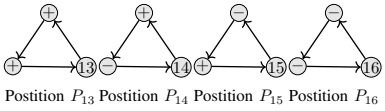
Similar to our ego-level analysis, we study triads in both undirected and directed networks.

Triads in Undirected Networks. In undirected networks, we use the sentiments of two friends, to predict the sentiment of their common friend. For each user, we take all the friendship triads he or she is in, and we group the other two friends into three categories based on their sentiments: $(+,+)$, $(+,-)$, and $(-,-)$. We count the number of friend pairs in each category and include them as three features for the user. We predict sentiments using only these three features and obtain an accuracy of 55.34%. As users that participate in more triads provide more structural information, we plot the accuracy and AUC for users that participate in k or more triads in Figure 3. We observe that the accuracy and AUC increase, and the accuracy reaches 69% users in more than 150 triads.

Triads in Directed Networks. For directed networks, we construct similar features by counting the number of ways in which two friends of a user can be connected via directed links (6 features). In directed networks, we can connect our study to status theory. Status theory is a theory of signed link formation within social psychology, and it can be summarized as *"If u has a higher status than v and v has a higher status than w, then u should have a higher status than w."* We consider all possible (1) edge directions and (2) sentiments of the other two users, which leads to 16 different positions a user can take within directed rotatable/non-rotatable triads (see figures 4a and 4b). Participation in each position may or may not provide evidence on the sign of a user's sentiment. For example, if one speculates that users follow others that are happier than themselves, then a user is more likely to be a positive user if he or she appears frequently in the positions such as P_9, P_{10} and P_{11}. On the contrary, if a user appears frequently in the positions like P_3, P_4 and P_8, he or she is more likely to be a negative one. We count the number of times the user is in each position among all the directed triads that user is a member of and get 16 features. Table III provides the accuracy: 56% and the AUC: 52%. The performance does not improve greatly as the number of triads of which the user is a member increases. The results show that the predictive power of these 16+6 = 22 features are limited.



(a) 12 Possible Positions for a User in Non-Rotatable Triads



(b) 4 Possible Positions for a User in Rotatable Triads

Fig. 4: Positions in Directed Signed Triads

TABLE III: Accuracy with Directed Features

Minimum Triads	# of Users	Accuracy	AUC
0	111,734	56.24%	52.18%
10	64,280	55.50%	55.45%
50	40,006	54.88%	52.93%
100	29,372	55.94%	52.92%
200	19,641	56.91%	52.28%

D. Community Level

We investigate whether we can predict one's sentiment using the sentiments of other community members.

For every community that the user u is in, we calculate the average SWB of the other users and denote it as the SWB of a community with respect to u, i.e., $S_u = \frac{1}{|C|-1}\sum_{v \in C-\{u\}} S(v)$, where C is the community and $|C|$ is its size. For each user, we include the average, maximum, and minimum of S_u values of all communities that the user has joined, as features.

Research has shown that happy people have rich and satisfying social relationships [17], so we take # of communities, # of positive communities, # of negative communities and fraction of positive communities as features (from which negative fractions can be computed), too. With these 7 features, we predict with an accuracy of 58.17% and an AUC of 57.13%.

E. Network Level

On the Network-Level, to capture a macro view of the network, we use a generative network model; to obtain a micro view, a network embedding technique is utilized.

To select a proper generative model for our dataset, we seek patterns in our dataset that can lead us towards the appropriate model. Real-world social networks often exhibit a core-periphery structure[32], where they consist of a dense cohesive core and a sparse, loosely connected periphery. Previous studies show that in social networks, generally users with positive emotions form the core of the network; users with negative emotions form its periphery [15]. It leads us to utilize stochastic Kronecker graphs as our generative model.

TABLE IV: Prediction Performance with Kronecker Features

Features	Accuracy	AUC
One-Hop Graph	53.84%	53.90%
Two-Hops Graph	53.02%	53.40%
Graph minus One-Hop	56.20%	48.70%
Combined	55.04%	51.41%

Kronecker Features. Stochastic Kronecker graph is a generative model that can capture the core-periphery property of real-world networks using Kronecker graph product [18]. In brief, given an adjacency matrix $A \in \mathbb{R}^{n^k \times n^k}$ of a graph, stochastic Kronecker graph model aims to learn a small probability matrix $P \in \mathbb{R}^{n \times n}$ (n between 2-5 typically). This small matrix is known as the *Kronecker initiator matrix*, and the k^{th} Kronecker power of P (i.e., $P^{\otimes k} = \underbrace{P \otimes P \cdots \otimes P}_{k\ times}$) is most likely to have generated A, i.e., $P(A|P)$ is maximized (for further details refer to Ref. [18]). The KRONFIT algorithm can estimate the Kronecker initiator matrix for a real-world graph using the maximum likelihood principle in linear time. Using KRONFIT, we set n equal to 2, and we fit a 2×2 initiator matrix $I = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$ to our graphs. Given a 2×2 initiator, one can interpret it as a recursive expansion of two groups into sub-groups. In a network exhibiting a core-periphery structure, a represents the core strength and is large; by contrast, d represents the periphery and is small. In an undirected network, which has a symmetric adjacency matrix, the Kronecker initiator is also symmetric, i.e., $b = c$.

For each user (i.e., node) we select three types of induced subgraphs: (i) One-Hop (all the nodes that are within one-hop of the user), (ii) Two-Hops and (iii) the whole graph minus One-Hop. For each induced subgraph, we estimate a 2×2 initiator matrix $\left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$ using KRONFIT and use a, b and d as 3 features. Table IV shows the result with features from different induced graphs and the combination of all 9 (=3×3) features. The accuracy varies from 53% to 56%. The predictive power is limited, and we believe the result can be due to the following: (1) One-Hop and Two-Hops induced subgraphs are all generated with the node as the start or center node, which enhances the core-periphery property even when the node does not carry much core strength in the original graph; (2) For the whole graph minus one-hop, our goal was to investigate whether removing one node and its neighbors will take core strength away. It appears that this is not the case, and one-hop's coverage is small compared to the whole graph.

For a micro view of the network, we use NODE2VEC as a node embedding technique that utilizes whole-network information to generate node embeddings.

NODE2VEC Features. NODE2VEC is a framework that learns a mapping of nodes to a low-dimensional space of features that maximizes the likelihood of preserving network neighborhoods of nodes [19]. It is based on a biased random walk procedure for sampling network neighborhoods, and aims to learn a node embedding that maximizes the log-probability

TABLE V: Prediction Performance with NODE2VEC Features

NODE2VEC Feature Type	Accuracy	AUC
Unweighted	56.10%	57.70%
Weighted	60.29%	60.68%
Combined	62.81%	62.59%

TABLE VI: Accuracy with All Features

Minimum # of Friends	Accuracy	AUC
0	60.24%	57.80%
10	62.06%	60.08%
50	67.40%	64.95%
100	67.84%	66.06%

TABLE VII: Accuracy with Combinations of Four Levels

Combinations	Accuracy	AUC
All Features	67.11%	64.95%
Ego + Triad	67.16%	65.12%
Ego + Community	59.46%	56.77%
Ego + Network	59.51%	56.92%
Triad + Community	67.79%	65.64%
Triad + Network	**67.99%**	**66.31%**
Community + Network	61.17%	59.97%
Ego + Triad + Community	67.06%	64.88%
Ego + Triad + Network	67.45%	65.35%
Ego + Community + Network	59.51%	56.90%
Triad + Community + Network	67.50%	65.34%

of the observations. One can choose the dimensionality of the feature space by parameter d, and adjust the sampling strategy with parameter parameters p and q. We apply the algorithm on our friendship network with $d = 128$, $p = q = 0.25$. Thus for each node, we can get 128 features from NODE2VEC.

NODE2VEC supports weighted graphs too. The weights are applied to the sampling strategy only. In our case, the edge weight can be the tie strength of two users, which is related to the sentiment of the users. A previous study of core-periphery property [20] shows that if one can denote the levels of coreness for both nodes i and j as values between 0 to 1, then the product of their coreness can quantify their tie strength. Thus, we use the previously proposed *emotional coreness* and set *emotional tie-strength* [15] as the weights. Emotional coreness for user u is defined as $e_u = (S(u)+1)/2$, which is a bijection that rescales SWB of a user from $[-1, 1]$ to $[0, 1]$, maintains its ordering, and still follows the same normal distribution. After this mapping, emotional coreness of negative users lies in $[0, 0.5)$ and that of positive users is in $(0.5, 1]$. Once emotional coreness is computed, we can compute *emotional tie-strength* e_{ij} between users i and j as $e_{ij} = e_i \cdot e_j$. Note that although we use the user SWB in the model, it is only used in the sampling strategy and does not leak user sentiment information via learned features.

Table V provides the prediction results using NODE2VEC features for unweighted and weighted graphs, and the combination of these features. The features learned from the unweighted graph can reach an accuracy of 56% and an AUC of 57%, and features from weighted graph can reach an accuracy and AUC of about 60%. Combined features can achieve an accuracy and AUC of approximately 62%.

TABLE VIII: Top 4 Selected Features via Logistic Regression

Feature #	Feature
1	# of positive followers
2	# of negative followers
3	# of $(-,-)$ pairs (undirected)
4	# of $(+,-)$ pairs (undirected)

TABLE IX: Choice of Features for Sentiment Prediction

Minimum # of Friends	Accuracy	AUC
0	57.97%	54.80%
10	61.55%	58.98%
50	67.16%	64.71%
100	69.19%	67.03%
200	70.00%	67.14%

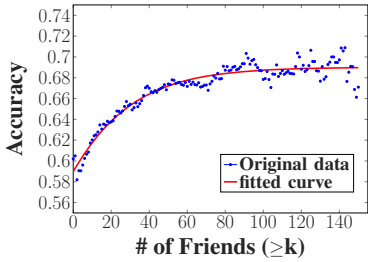


Fig. 5: Performance Improvement with All Features with respect to the Number of Friends

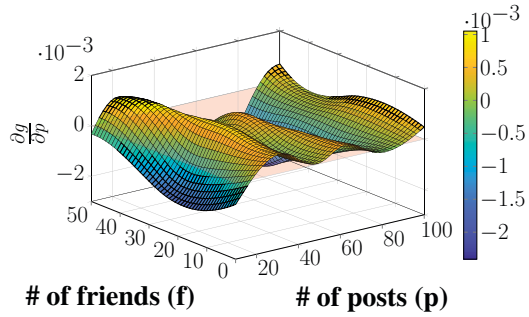
F. Combining All Network Levels

We combine features corresponding to all network levels for sentiment prediction. Combining all the $(3+6+9+6+16+7+9+128+128 = 306)$ features, we obtain the prediction results in Table VI. The accuracy is 60.24%. The accuracy increases with the minimum number of friends a user has, and reaches about 67% for those with more than 50 friends. In fact, a logistic S-curve, $f(x) = 4.054/(5.874 + e^{-0.03696x})$, fits the plot in Figure 5 with $R^2 = 0.92$. For users with more than 50 friends, we also obtain the accuracy with combinations of the four levels of features in Table VII. The results indicate that the combinations including Triad-level features outperform.

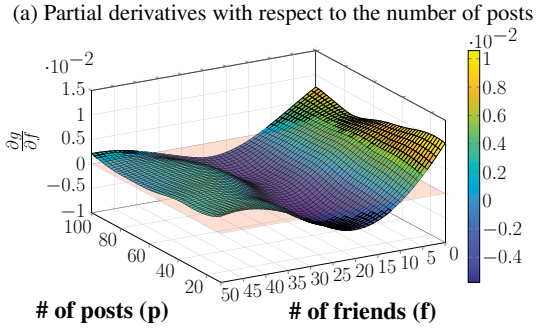
G. Choice of Features and Learning Algorithms

In this section, we identify most informative features for sentiment prediction and assess the impact that our choice of classifier had on our experimental results.

Informative Features. Our goal here is to identify features that are (1) easy to generate and (2) informative for sentiment prediction. Hence, we do not consider network level features, as they are extremely time-consuming to generate and hard to interpret. Most informative features can be identified by standard feature selection techniques such as Information Gain, χ^2, among others. Here, we use logistic regression coefficients for feature importance analysis and ranking the remaining 41 structural features. Table VIII shows the top four features and Table IX shows the prediction result using only these features. We observe that the performance using features selected in terms of accuracy/AUC is very close to the result by using all the features. The performance is especially close for users with more than 50 or 100 friends, where predictions can be up to 69% accurate.



(a) Partial derivatives with respect to the number of posts



(b) Partial derivatives with respect to number of friends

Fig. 6: Change in Prediction Performance Surface with Additional Information (Posts/Friends)

TABLE X: Choice of Learners for Sentiment Prediction

	Naive Bayes		SVM	
Minimum # of Friends	**Accuracy**	**AUC**	**Accuracy**	**AUC**
0	59.31%	57.27%	58.96%	57.41%
10	54.30%	56.94%	57.00%	53.28%
50	55.83%	57.91%	56.18%	50.02%
100	53.37%	56.30%	57.03%	50.00%
200	53.33%	53.61%	60.00%	50.00%

Choice of Learning Algorithm. In our experiments we used logistic regression for sentiment prediction. For evaluating the learning bias, we compared our performance with some basic learning algorithms such as Naive Bayes and the SVM. These classifiers have different learning biases, and we expect to observe different performances for the sentiment prediction task. Table X provides the prediction results. As seen in the table, results are not significantly different among these methods. This observation indicates that when sufficient network information is available in features, sentiment prediction using structural features is reasonably accurate and not sensitive to the choice of learning algorithm. Overall, logistic regression performs slightly better, especially for users with more friends.

H. Comparison with Text-based Methods

We compare sentiment prediction based on network structure with text-based sentiment prediction methods. We choose *Stanford CoreNLP sentiment* [21] as a representative text-based sentiment prediction tool. Stanford CoreNLP is based on Recursive Neural Tensor Networks and the Stanford Sentiment Treebank. It classifies every sentence into five sentiment classes: {`Very negative`, `Negative`, `Neutral`, `Positive`, `Very positive`}. We represent these five classes as sentiment scores {-2, -1, 0, 1, 2}. For each post, we average the sentiment scores of all the sentences in the post. If this average value is greater than 0 (i.e., above `Neutral`), we consider the post positive; If it is less than 0, we consider the post as negative; Finally, if it is zero, we denote the post as a neutral post. After assigning sentiments to posts, we calculate the new (text-based) SWB of each user to predict the user's general sentiments. As the sentiment classification method is computationally expensive, we sampled 1,700 users with about 350,000 posts as the test data. Table XI provides the accuracy rates, indicating an accuracy rate of around 55%. We notice that as the minimum number of posts that a user has increases, the accuracy and AUC slightly increase.

TABLE XI: Accuracy with Text-based Methods

Minimum # of Posts	**Accuracy**	**AUC**
10	54.99%	57.05%
50	55.57%	58.44%
100	56.92%	60.37%
200	56.08%	61.41%

Network Information versus Text Trade-off. Previous experiments demonstrate that sentiment prediction performance is closely related to the amount of data available, i.e., number of friends for prediction using network structure and number of posts for prediction using text. However, in reality, it is not straightforward to obtain one's complete posts and friendship information due to limitations imposed by site APIs or other privacy concerns. Hence, in this section, we analyze the trade-off between the sentiment prediction performance and the information that is available. We model prediction accuracy ACC as a function $g(.,.)$ of the number of friends and the number of posts that we have available for a user, *i.e.* $ACC = g(p, f)$, where ACC is the accuracy, p is the number of posts, and f is the number of friends. We ask the following question: given a user with p posts and f friends, what is the accuracy gain that we can expect by having Δp more of her posts or Δf more of her friends? To determine this gain, we should look at $\frac{g(p+\Delta p,f)-g(p,f)}{\Delta p}$ and $\frac{g(p,f+\Delta f)-g(p,f)}{\Delta f}$, and when $\Delta p \to 0$ and $\Delta f \to 0$, they turn into partial derivatives of the accuracy surface with respect to friends and posts: $\frac{\partial g}{\partial p}$ and $\frac{\partial g}{\partial f}$. For instance, if both partial derivatives are positive at point (x, y), then it means that getting more posts or friends for users with x posts and y friends will help improve the accuracy. Figure 6a shows the partial derivatives of the accuracy surface of the text-based method with respect to the number of posts. From the figure, we have a few observations: (1) For users with very few posts and few friends, getting more posts does not help; (2) For users with many friends and few posts, more posts can help; (3) For users with many posts, more posts lead to accuracy gain. Similarly, Figure 6b depicts the partial derivatives of the accuracy surface of the network-based method with respect to number of friends on the same sample dataset. We observe that for users with a few friends or many friends, more friendship information can improve the prediction accuracy, while the posts information has very limited impact. To compare the predictive power of posts and

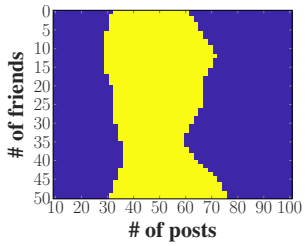


Fig. 7: Comparing prediction improvements with more posts versus more friends. When yellow, additional posts help more than friends and when blue, friends help more.

friendship information, we should look at the relation between $\frac{\partial g}{\partial p}$ and $\frac{\partial g}{\partial f}$ at each point (x, y). In Figure 7, the area where $\frac{\partial g}{\partial p} > \frac{\partial g}{\partial f}$ is yellow, and the area where $\frac{\partial g}{\partial p} < \frac{\partial g}{\partial f}$ is blue. The space clearly splits into three parts, which indicates that for users with few posts or many posts, friendship information is more useful than getting more posts; on the other hand, for users with some but not many posts, more posts are preferred. These findings enable informed decisions under information collection constraints (e.g., API limits).

IV. Additional Related Work

Through our findings, we believe that our methods can be closely linked to the following areas of research.

I. Sentiment Propagation. Recently, Coviello et al. [6] and Zafarani et al. [22] have studied emotional contagion and sentiment propagation in social networks. Here, we do not have access to causal information on influence or propagation with respect to sentiments; however, our prediction results may indicate the existence of such kind of propagations.

II. Signed Networks. Signed networks have been connected to the classical theory of structural balance and theory of status [23]. Leskovec et al. [24] have shown that edge signs are predictable in signed social networks. Specifically, signed networks have been used to study person-to-person sentiments and how individuals evaluate others, e.g., friends or foes [25]. Here, we look at nodes in social networks that carry sentiment, as opposed to edges in previous studies, and predict the sign of the nodes. Hence, our study complements previous studies.

V. Conclusions and Discussion

We have investigated the utility of the social information at the ego, triad, community, and the whole-network level for sentiment prediction. Our study shows that using structural properties alone sentiments are reasonably predictable. We have identified most informative features, showing that when computational resources are limited, by using only four network properties one can predict sentiments with reasonable accuracy. We compared this approach with text-based methods and show that it can be as, or more, accurate than text-based techniques. For the situations where complete posts and friendship information are difficult to obtain, we analyze the trade-off between the sentiment prediction performance and the available information. Our findings can be used for (1) enhancing classical sentiment prediction methods that use text or (2) friend recommendation. Our results show that sentiments play a significant role in the formation of friendships and the network, which suggests the possibility of recommending friends that express similar sentiments.

References

[1] N. Schwarz, "Emotion, cognition, and decision making," *Cognition & Emotion*, vol. 14, no. 4, pp. 433–440, 2000.
[2] G. H. Bower, "How might emotions affect learning," *The handbook of emotion and memory: Research and theory*, vol. 3, p. 31, 1992.
[3] M. Macht, "How emotions affect eating: a five-way model," *Appetite*, vol. 50, no. 1, pp. 1–11, 2008.
[4] K. Oatley, D. Keltner, and J. M. Jenkins, *Understanding emotions*. Blackwell publishing, 2006.
[5] I. Burkitt, "Social relationships and emotions," *Sociology*, vol. 31, no. 1, pp. 37–55, 1997.
[6] L. Coviello, Y. Sohn, A. D. Kramer, C. Marlow, M. Franceschetti, N. A. Christakis, and J. H. Fowler, "Detecting emotional contagion in massive social networks," *PloS one*, vol. 9, no. 3, p. e90315, 2014.
[7] D. Cartwright and F. Harary, "Structural balance: a generalization of heider's theory." *Psychological review*, vol. 63, no. 5, p. 277, 1956.
[8] D. G. Myers, "The funds, friends, and faith of happy people." *American psychologist*, vol. 55, no. 1, p. 56, 2000.
[9] J. Kim and J.-E. R. Lee, "The facebook paths to happiness: Effects of the number of facebook friends and self-presentation on subjective well-being," *CyberPsychology, behavior, and social networking*, vol. 14, no. 6, pp. 359–364, 2011.
[10] V. A. Visser, D. van Knippenberg, G. A. van Kleef, and B. Wisse, "How leader displays of happiness and sadness influence follower performance: Emotional contagion and creative versus analytical performance," *The Leadership Quarterly*, vol. 24, no. 1, pp. 172–188, 2013.
[11] M. Newman, *Networks: an introduction*. Oxford university press, 2010.
[12] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Advances in neural information processing systems*, 2002, pp. 585–591.
[13] S. Fortunato, "Community detection in graphs," *Physics reports*, vol. 486, no. 3, pp. 75–174, 2010.
[14] B. Liu, "Sentiment analysis and opinion mining," *Synthesis lectures on human language technologies*, vol. 5, no. 1, pp. 1–167, 2012.
[15] S. Jin and R. Zafarani, "Emotions in social networks: Distributions, patterns, and models," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 2017, pp. 1907–1916.
[16] J. Bollen, B. Gonçalves, G. Ruan, and H. Mao, "Happiness is assortative in online social networks," *Artificial life*, vol. 17, no. 3, pp. 237–251, 2011.
[17] E. Diener and M. E. Seligman, "Very happy people," *Psychological science*, vol. 13, no. 1, pp. 81–84, 2002.
[18] J. Leskovec, D. Chakrabarti, J. Kleinberg, C. Faloutsos, and Z. Ghahramani, "Kronecker graphs: An approach to modeling networks," *Journal of Machine Learning Research*, vol. 11, no. Feb, pp. 985–1042, 2010.
[19] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in *Proc. of the SIGKDD conference*, 2016, pp. 855–864.
[20] S. P. Borgatti and M. G. Everett, "Models of core/periphery structures," *Social networks*, vol. 21, no. 4, pp. 375–395, 2000.
[21] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proceedings of the 2013 conference on empirical methods in natural language processing*, 2013, pp. 1631–1642.
[22] R. Zafarani, W. D. Cole, and H. Liu, "Sentiment propagation in social networks: a case study in livejournal," in *International Conference on Social Computing, Behavioral Modeling, and Prediction*. Springer, 2010, pp. 413–420.
[23] J. Leskovec, D. Huttenlocher, and J. Kleinberg, "Signed networks in social media," in *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 2010, pp. 1361–1370.
[24] ——, "Predicting positive and negative links in online social networks," in *Proceedings of the WWW conference*. ACM, 2010, pp. 641–650.
[25] R. West, H. S. Paskov, J. Leskovec, and C. Potts, "Exploiting social network structure for person-to-person sentiment analysis," *arXiv preprint arXiv:1409.2450*, 2014.