

TwitterMancer: Predicting User Interactions on Twitter

Konstantinos Sotiropoulos*, John W. Byers*, Polyvios Pratikakis[†] and Charalampos E. Tsourakakis*

*Boston University,[†]University of Crete

‘Email: ksotirop@bu.edu, byers@bu.edu, polyvios@ics.forth.gr, ctsourak@bu.edu

Abstract—This work investigates the interplay between different types of user interactions on Twitter, with respect to predicting missing or unseen interactions. For example, given a set of *retweet* interactions between Twitter users, how accurately can we predict *reply* interactions? Is it more difficult to predict *retweet* or *quote* interactions between a pair of accounts? Also, which features of interaction patterns are most important to enable accurate prediction of specific Twitter interactions?

Our empirical study of Twitter interactions contributes initial answers to these questions. We have crawled an extensive dataset of Twitter accounts and their *follow*, *quote*, *retweet*, *reply* interactions over a period of a month. Using a machine learning framework, we find we can accurately predict many interactions of Twitter users. Interestingly, the most predictive features vary with the user profiles, and are not the same across all users. For example, for a pair of users that interact with a large number of other Twitter users, we find that certain “higher-dimensional” triads, i.e., triads that involve multiple types of interactions, are very informative, whereas for less active Twitter users, certain in-degrees and out-degrees play a major role. Finally, we provide various other insights on Twitter user behavior.

Our code and data are available at <https://github.com/twittermancer/>.

Index Terms—Graph mining, machine learning, social media, social networks

I. INTRODUCTION

Twitter is a microblogging service with more than 330 million monthly active users worldwide, as of the first quarter of 2019. Its unique characteristics have drawn the attention of many researchers, and provide a novel opportunity for understanding human behavior in a public and observable forum at an unprecedented scale. Among many other applications, the Twitter repository of human signals has been used to predict the stock market [1], estimate mortality of heart diseases [2], forecast election outcomes [3], [4], and detect humanitarian crises in real time [5], [6]. Twitter follows the pulse of global society, and therefore studying it from all possible angles is an active area of research. The angle we take here is related to the multiple networks that naturally underlie the Twitter platform. Specifically, over a fixed window of observation, a user u can interact with a user v in more than one way; u may follow, reply, quote, retweet v , or like a tweet of v , or send her a message. These types of interactions naturally define a *multilayer* directed network, with layers corresponding to the types of interactions that occur. We study both the unweighted and weighted version of these networks, but for simplicity, we discuss the unweighted (0/1) case herein. In this setting, the sets of adjacencies (directed edges) across layers are typically

correlated, but also exhibit clear differences. We are interested in the extent to which those differences can be characterized, i.e., differences that are specific to certain layers (e.g., sparsity of a given type of interaction), and differences that relate to the local neighborhoods of users (e.g., graph structure around a celebrity). Difference characterization via relevant feature analysis enables accurate cross-layer prediction; conversely, differences that are hard to characterize makes cross-layer prediction more difficult. In this work we focus on the following question:

Question. *Given two Twitter users u, v , and the Twitter multilayer interactions, can we predict what type of interactions will take place between u, v ?*

For the purpose of answering the above Question, we have crawled a large Twitter dataset, spanning the full month of February 2018. Using this corpus of tweets, we have created a multilayer network with four different layers, corresponding to four different types of interactions respectively: *follow*, *quote*, *reply*, *retweet*. We use this network, together with the temporal information available to us, to attack the following series of problems:

High-dimensional link prediction. Suppose we are given the multilayer Twitter network, except for all interactions between a *pair* u, v of nodes that are known to have interacted. How reliably can we infer whether u, v will *follow*, *reply*, *quote*, or *retweet* each other using the information provided by the rest of the network? Our approach is data-driven, generalizes the seminal work of Liben-Nowell and Kleinberg [7] on link prediction, and follows the established framework of Leskovec, Kleinberg and Huttenlocher [8]. Our main finding is that leveraging information from other types of interactions boosts prediction accuracy significantly, on the order of 9–32%. We observe that typically higher gains are obtained for the sparser interaction layers, and that overall the simplest forms of interaction, like retweeting (just two clicks), are easiest to predict.

Correlation between types of interactions. We perform a detailed study on how certain interactions increase or decrease the likelihood of other interactions. We focus on two types of experiments to assess these interaction correlations. First, for each interaction, we test how the prediction accuracy changes for the standard link prediction formulation when we use

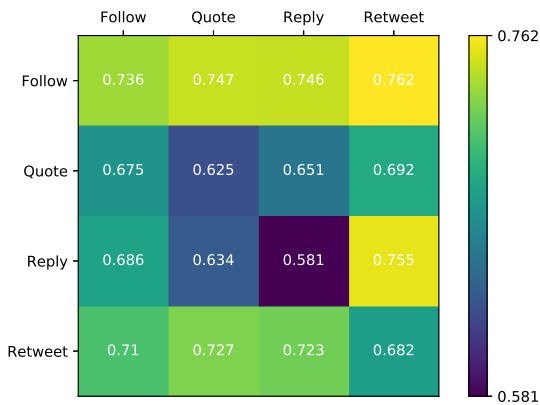


Fig. 1: Predict edges of a type on the horizontal axis, using also information from another type of the vertical axis. Diagonal tiles refer to using only one type to predict this type.

information from *one* additional layer. Figure 1 displays a heatmap depicting the classification accuracies of a specific (u, v) interaction (row) by leveraging other interactions of that type plus an additional interaction (column). Entries along the diagonal correspond to using no outside interaction types. For example, consider the Reply row. When only Replies are used in prediction, the prediction accuracy is a relatively low 58.1% on the diagonal (as described in Section III, we test against an equal number of edges and non-edges so that 50% is achieved by random guessing). The prediction accuracy is boosted to 75.5% using the additional information from Retweets (rightmost column), but only to 63.4% using Quotes (second column). We observe that for all interactions, side information is always beneficial, as one would anticipate. The most informative other interaction is consistently Retweets, and for Retweets themselves, the most informative interaction is from Quotes (even though that graph layer is significantly sparser than, for example, the follows layer). Finally, our interpretable logistic regression model, presented in Section III, allows us to understand the relative importance of certain types of interactions.

Accuracy of prediction as a function of user profiles. Finally, we provide novel insights on what features are important for accurate prediction for different user profiles. Our findings strongly suggest a dichotomy between two types of pairs of users; pairs of users that engage jointly with many other Twitter users in various combinations of interactions, and those who don't (for details see Section V). For the former, features involving triads turn out to be important features in accurate prediction, complementary to degree-based features that are useful for less well-connected users. We provide a detailed study of how all these features jointly affect the empirical effectiveness of link prediction. Our findings can be seen as a higher-dimensional analog of the importance of triads in the signed link prediction problem [8].

Roadmap. Section II presents closely related work. Sections III, IV, and V present our machine learning framework,

discuss the dataset collection methods we use, and our experimental findings, respectively. Section VI concludes our paper.

II. RELATED WORK

Due to the large volume of work related to Twitter and link prediction, we focus on related literature that lies closest to our work.

Link prediction. The link prediction problem was popularized by the seminal paper of Libel-Nowell and Kleinberg [7]. Since then, link prediction has been studied extensively [9]. Close to our work lies the framework proposed by Leskovec et al. [8] that extended the link prediction problem to graphs with positive and negative interactions between nodes. Their work suggests a machine learning framework that uses local features and a logistic regression classifier to predict the unknown sign of an edge.

Edge sign prediction. Leskovec, Huttenlocher, and Kleinberg [8] formulate the edge sign prediction problem as follows: given a social network $G(V, E, s)$, where $s : E \rightarrow \{-1, +1\}$ is the edge sign function, how well can we predict the sign of a future edge $s(x, y)$ between nodes x, y from the rest of the network? In their influential paper [8], Leskovec et al. proposed a machine learning framework where in practice one of the edges $(x, y) \in E$ is removed from the network, and its sign is predicted using a logistic regression classifier that uses 23 features in total. These features include positive and negative out-degrees $d_{out}^+(x), d_{out}^-(x)$ of node x , positive and negative in-degrees $d_{in}^+(y), d_{in}^-(y)$ of node y , the total out- and in-degrees $d_{out}(x), d_{in}(y)$ of nodes x, y , and the number of common neighbors (forgetting directions of edges) $C(x, y)$ between x, y . The quantity $C(x, y)$ was referred to as the *embeddedness* of the edge $x \rightarrow y$ in [8], and we will follow the same terminology. The rest 16 features correspond to counts of each possible triad configuration between x, y . Recently, Tsourakakis et al. extended this framework by including path-based features [10]. It is worth mentioning that the model the authors propose in [10] is closely related to the famous planted partition model [11], [12], [13].

Prediction on Twitter. A wide variety of prediction problems have been studied on Twitter, due to its unique nature. Petrovic, Osborne and Lavrenko [14] studied the problem of predicting whether a user will retweet a particular tweet, or more generally spread an item of interest. On the same problem, Galuba et al. [15] used a propagation model to find which users are likely to mention certain URLs. Martinčić-Ipšić et al. [16] focus on predicting pairs of hashtags (or words) that will co-occur in future tweets.

Jalili et al. [17] focused on the following link prediction problem: given a set of users who participate both on Twitter and Foursquare, predict links between users at Foursquare by using information from the Twitter network. While experimentally they do not study the inverse questions, their tools can be used to predict links on Twitter from Foursquare. Hristova et al. [18] enrich this framework by the use of random forest classifiers.

Our work is however, the first —to the best of our knowledge— that explores link prediction in the context of different interactions among users on the Twitter network. Abufouda and Zweig [19] use multiple networks to predict which links among users represent actual, real-life links. Our work differs from the bulk of such Twitter-related link prediction problems as we focus on predicting interactions on, and across different Twitter layers.

III. PROPOSED FRAMEWORK

In this work we focus on the following question that extends the seminal formulation of Leskovec, Kleinberg, and Huttenlocher [8] on predicting signed edges in online social networks, and more generally research work on link prediction [7]:

Problem 1. *Given the Twitter graph containing user accounts and their pairwise follow, reply, retweet, and quote interactions, and a pair of user accounts $\{u, v\}$, how accurately can we predict whether u will follow, reply, retweet, or quote v ?*

Understanding Problem 1 will contribute further towards a better understanding of user behavior on Twitter, and may lead to detecting correlations between types of interactions that will be useful for anomaly detection among others. We model the input dataset as a directed, multi-label, multigraph $G = (V, E, I, \ell_E)$. Specifically, the node and edge sets V, E correspond to the set of Twitter user accounts, and the interactions among them, respectively. Different types of interactions are modeled by the label function, i.e., $\ell_E : E \rightarrow \mathcal{I}$ is the function that labels each edge according to the set \mathcal{I} of all possible interactions. Here, we consider $\mathcal{I} = \{\text{follow, quote, reply, retweet}\}$. Our framework naturally extends to larger sets of interactions and also weighted graphs, i.e., graphs where each edge is associated with the counts of interactions.

A. A Machine Learning Framework

The task of predicting a missing edge on a graph can be thought of as the following classification problem: given a pair of users (u, v) and an interaction type i , we are trying to learn a function f that returns 1 if an edge (u, v) with label i is present on the graph, and -1 otherwise. To tackle our problem, while retaining interpretability of results, we use a simple logistic regression framework. We use the term *embeddedness* —as used also by Leskovec et al. [8]— for an edge (u, v) as the quantity $|\{t \mid (u, t) \text{ and } (t, v) \in E\}|$, the number of common neighbors between $\{u, v\}$.

Features: As Twitter graphs are typically on the order of millions of users, we use local features that are computationally efficient to extract. This approach also mirrors a local view that Twitter users usually have (e.g., on their timeline) when deciding to make an action (follow another user, reply to a tweet, etc.). We build on features already used in relevant related work, while also incorporating new feature sets that capture the interplay between different types of interactions.

The first set of features that we use aims at capturing the propensity of users to interact with other users more or less often. To capture the breadth and relative frequency of activity of each user, we define features based on the degrees of corresponding nodes in the interaction graph. As Twitter is inherently a directed network, and since we are concerned with inferring the directionality of interaction (u, v) , we use the following 10 directed *degree features*: We use the *out-degree* of user u for each of the interaction types in \mathcal{I} (4 features), the *in-degree* of user v for those types (4 features), as well as counts of the *number* of different interaction types u initiated and v received, respectively (2 features).

The second set of features we use considers a common neighbor t of u and v (as counted when computing the embeddedness of (u, v)), and identifies all possible ways in which t had an interaction with both u and v . For this set of features, we consider each interaction type separately and retain the directionality of edges. We have, thus, 3×3 possible triads for each type, times 4 interaction types, yielding 36 features in total (see Fig. 2).

For the final set of features, we again employ triads as above, but this time we make use of the interplay between different types of interactions, e.g., for a common neighbor t , if v retweets something that t posted, and u replies on t , then it is likely that u follows v . To keep the cardinality of this set of features manageable and to avoid overfitting, we drop the directionality of the edges and use pairs of different interactions. Thus, we have again 3×3 different triads for each one of the $\binom{4}{2}$ pairs, yielding 54 additional features (see Fig. 3).

As our predictions are directed ($u \rightarrow v$), we will use the following notation for the feature names whenever we refer to them: $Out(i)$ will refer to the out-degree of u at layer (interaction) i and respectively $In(i)$ for the in-degree of v . We will use the first letter of interactions¹ to refer to every layer i . $Total(u)$ and $Total(v)$ will refer to the number of different layers u was an initiator and v a receiver. Lastly, the notation for the triadic features is the one we describe in Figures 2 and 3.

Methodology. Using these 100 features, we train a logistic regression model of the form:

$$\Pr[e \in E|x] = \frac{1}{1 + e^{-(b + \langle w, x \rangle)}} \quad (1)$$

where x is a vector representing the 100 features for a sample, while b is the intercept and w is the vector of coefficients that we want to learn.

For each type of interaction (e.g., reply), we randomly sample an equal number of edges where there was an interaction of this type and where there was not. Therefore, our datasets will be balanced, providing a baseline of how much more accurately we can predict over random guessing. We use 10-fold cross-validation: 10 disjoint folds, where within each fold, 90% of the edges will be used for training and the remaining 10% for validation.

¹We distinguish between reply and retweets, by using r and rt , respectively.

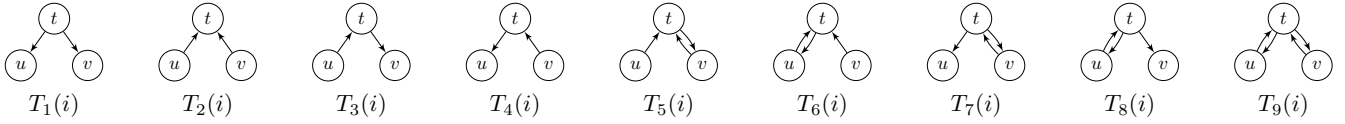


Fig. 2: Different types of triads involving only one interaction type $i \in \mathcal{I}$, the set of all possible interactions, while taking into account edge direction.

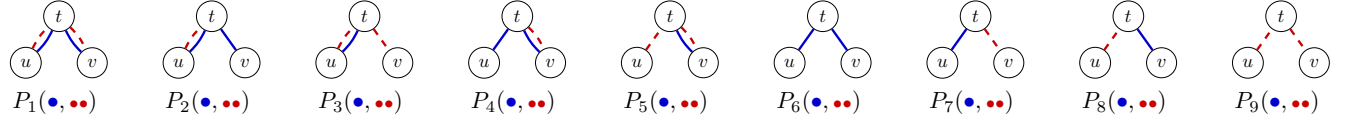


Fig. 3: Different types of triads involving the interplay between pairs of interactions $\bullet, \bullet\bullet \in \mathcal{I}$, the set of all possible interactions. Whenever a solid or dashed edge is absent, they did not have this type of interaction directly.

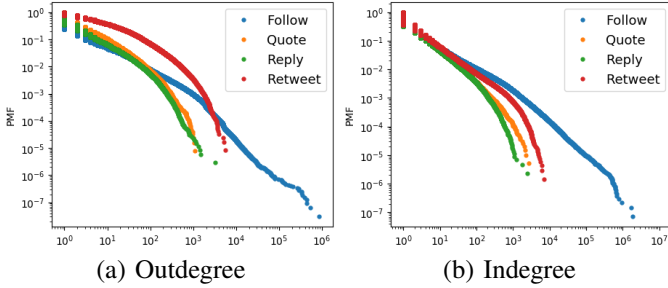


Fig. 4: Degree distributions on log-log scale

IV. DATASET COLLECTION

We used an open-source Twitter API crawler to monitor Twitter traffic generated during February 2018 [20]. The crawler targets the Greek-speaking users of Twitter, and performs a near-total crawl of all tweets by the selected users. Focusing the crawler in such a way produces a dense sample of a localized part of the Twitter graph, instead of a sparse random sample of the whole graph, as the language — or alphabet— barrier facilitates recognizing interesting users with high probability of locality. A similar technique has been applied in the past for the Korean-speaking part of Twitter [21].

The resulting dataset contains 21 million tweets, of which 9.8 million are in Greek. There are 204 million follow relations among users that were observed before the start of February 2018, and 33 million additional follow relations crawled during February 2018 proper. We obtain user information on 13 million unique users, and we classify 340 thousand of them as Greek-speaking, using the conservative rule of thumb of having posted more than 100 tweets, at least 20% of which are in Greek. The dataset contains many more Greek-speaking users than those classified marked as Greek-speaking, because (i) many accounts have not yet posted enough tweets for our heuristic to label them as Greek-speaking or not, and (ii) even if they have shown zero evidence of tweeting in Greek, they are followed-by or are following Greek speakers. From this Twitter dataset we extract four graphs, namely Follow, Quote,

Reply, and Retweet, described below. Figures 4(a) and 4(b) show their respective out-degree and in-degree distributions, in log-log scale.

The **Follow Graph** is the directed graph of the follow relation among users. The crawler uses the Twitter API to periodically scan all tracked users for their lists of friends and followers. Newly discovered users are given priority in scanning, but after the first scan of friends and followers, users are revisited in a FIFO order that requires several months to cycle through. Moreover, the Twitter API does not date the follow edges.² Due to the long time interval between crawls of the friend and follower lists, we construct a static follow graph without time information for all edges crawled before February 2018, and daily graphs for the follow edges crawled during the month.

The **Quote Graph** is the directed, weighted graph of quote retweets; these are tweets that include the URL of another tweet in their text, along with commentary text. These are rendered by most Twitter clients to include a box of the quoted tweet within the box of the quoting tweet. A weighted edge (u_1, u_2, w) indicates that there are w quote-retweets by user u_1 that quote tweets posted by user u_2 . As with all other dated relations, we consider the edge to have the date of the quote, not the original post, and extract daily aggregates for all of February 2018.

The **Reply Graph** is the directed, weighted graph where an edge (u_1, u_2, w) indicates that user u_1 has posted w tweets that directly reply to tweets posted by user u_2 . Since tweet objects returned by the Twitter API are dated, this graph is also dynamic, and we compute separate reply graphs for each day of February 2018.

The **Retweet Graph** is the directed, weighted graph where an edge (u_1, u_2, w) indicates that user u_1 has retweeted w tweets originally posted by user u_2 . This graph is also dynamic, as retweets are dated. Similarly to the Reply Graph, we compute separate retweet graphs for each day of February 2018.

²It is sometimes possible to infer approximately when a follow edge was added [22].

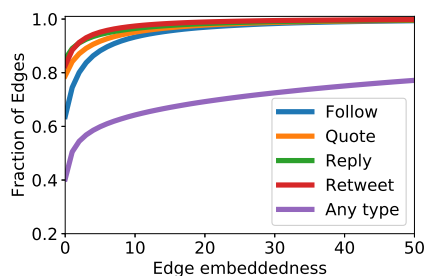


Fig. 5: Empirical CDF of Edge Embeddedness. Observe that most edges are contained in few triangles across all interaction types.

V. RESULTS

The data used in the experiments span the period of a month, namely all interactions that occurred in February 2018, and the induced follow relationships. Table I summarizes our dataset and the pairwise overlap between sets of nodes and edges appearing in multiple layers, measured by the *overlap coefficient* between two sets X and Y : $\frac{|X \cap Y|}{\min(|X|, |Y|)}$. We see that those Twitter users involved in retweets overlap the most with all other types of interactions.

A different set of important statistics on our layered graph relate to edge embeddedness. Fig. 5 depicts the empirical CDF of (undirected) edge embeddedness for each of the four layers of the graphs. The fifth CDF depicts edge embeddedness for the induced graph that has an undirected edge (u, v) if u and v participated in any type of interaction. For quotes, replies, and retweets, around 80% of edges have zero embeddedness; for follows, this percentage drops to 63%, and when any type of interaction is considered, only 40% of edges have zero embeddedness. Conversely, more than 20% of edges have any-type embeddedness exceeding a value of 50. Higher embeddedness reflects higher adherence to triadic closure, and we see that the follows interaction has the strongest individual effect. The much higher values of embeddedness when considering multiple types of interactions additionally demonstrate the promise of using specific triadic features in prediction.

One hand helping another. Our first experiment focuses on the standard link prediction problem, to test what kind of benefits in terms of prediction accuracy can be obtained by leveraging additional interactions of other types. Intuitively, we expect that leveraging information from additional Twitter layers should strictly improve link prediction. Consider for instance, a reply interaction. Frequently, replies are correlated with other interactions, e.g., two users who follow each other may have first retweeted the same tweet before one replies to the other. We verify this intuition experimentally. As motivated previously in Section I (see Figure 1), we quantify the improvements in performance by leveraging information from one additional layer. As in all subsequent experiments, we perform 10-fold cross validation, and report the *classification*

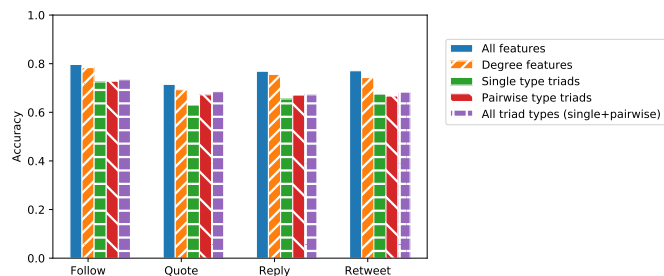


Fig. 6: Bar chart of accuracy prediction using different set of features

accuracy, defined to be the fraction of correct predictions. Table II reports mean accuracy over the 10 folds of cross-validation using no extra layers (first column), and all extra interaction layers (second column). The third column shows the relative improvement we obtain when we add all extra layers as part of the training input, clear illustration of the benefits of the proposed framework. It is worth mentioning that for each day, the accuracies we observe over the 10-folds are well concentrated around their mean. They never exceed 10^{-3} across all days and types of interactions, so we omit reporting them.

Which layer helps most? Suppose we want to predict links on a given layer. We now ask, which of the other layers adds most information to the link prediction classifier? This question is important for two reasons. First it provides significant insights on how Twitter users behave. Secondly, when computational resources are scarce, then one may want to leverage information from only one additional layer. As we can observe from the heat map in Figure 1, for all types of interactions, retweets help most in boosting accuracy. But, for predicting retweets themselves, quotes form the most informative layer of interactions. We assume that this is because quotes are essentially a special type of retweet (even if they are regarded as different by Twitter), where users not only retweet the original tweet, but also add their comment.

Overall accuracy and weights of features. Table III and Figure 6 summarize our results for the classification accuracy when using different sets of features. As we can see, our framework achieves an accuracy ranging from 71.5% for the *quote* type to roughly 80% for the *follow* type of interaction. For all types of features, the degree features are the ones that achieve the best performance. However, as we will see in the following, triadic features matter a lot when triads exist in greater abundance.

Figure 7 displays the features with the highest absolute value (positive and negative) of their learned coefficient weights. As the range of each feature differs significantly from one another, we have standardized them (with a mean value of 0 and standard deviation of 1) and we also used ℓ_2 -norm regularized logistic regression to obtain sparse solutions. From this figure, we see that the degree features are the ones that play the most important role in predicting interactions.

TABLE I: Number of users, directed edges and fraction of common interacting nodes and edges using overlap coefficients (Feb 1–28, 2018)

All types	Nodes	Edges	Node Overlap Coefficient				Edge Overlap Coefficient			
	1,125,044	5,000,833	F	Q	R	RT	F	Q	R	RT
Follow(F)	143,453	1,082,997	1				1			
Quote(Q)	271,824	666,820	0.44	1			0.17	1		
Reply(R)	530,956	1,259,970	0.58	0.57	1		0.21	0.10	1	
Retweet(RT)	762,459	3,501,240	0.90	0.63	0.42	1	0.85	0.23	0.19	1

TABLE II: Accuracy and relative improvement from using additional layers

Interaction Type	Accuracy (one layer)	Accuracy (all layers)	Relative improvement
Follow	0.736	0.797	8.29%
Quote	0.626	0.715	14.22%
Reply	0.581	0.769	32.36%
Retweet	0.682	0.771	13.05%

TABLE III: Accuracy of prediction using different set of features

Type	All	Degree	Single triads	Pairwise triads	All triads
Follow	0.797	0.784	0.725	0.728	0.734
Quote	0.715	0.693	0.630	0.672	0.685
Reply	0.769	0.756	0.655	0.671	0.672
Retweet	0.771	0.743	0.675	0.667	0.683

Observe that for predicting retweets uses mostly features from the same layer, and appears to be negatively correlated with other types of interactions.

Certain triadic features (see Figures 2 and 3 for the id-encoding) —which are very important in terms of interpretability, as they can explain patterns of interactions among users—also play an important role. For follow and retweet types, we observe that transitive closure, $T_3(rt)$, and hierarchy, $T_5(f)$ and $T_5(rt)$, can explain the existence of an edge of this type. While for quotes, two users that have quoted a common user tend to also have a retweet relationship between them, $P_8(q, rt)$. We note that these are some first findings of our work, as the task of understanding user behavior on twitter is much broader, and is an interesting open direction.

Are degrees or triads more informative features?

As triadic features constitute a key part of our framework, it is important to understand when they provide crucial information. Intuitively, we expect that the prediction accuracy should increase as the embeddedness of (u, v) increases, simply because these features make use of an intermediate node t in order to predict an interaction between u and v . On the other hand, logistic regression coefficients imply that degree features are more important than triads. This naturally brings the question whether triads or degrees are more important features? The answer is enlightening, and we explain it in detail in the following:

When a pair of nodes has a large embeddedness value, then triads are more informative. However, logistic regression coefficients indicate degree-based features are more important simply because most

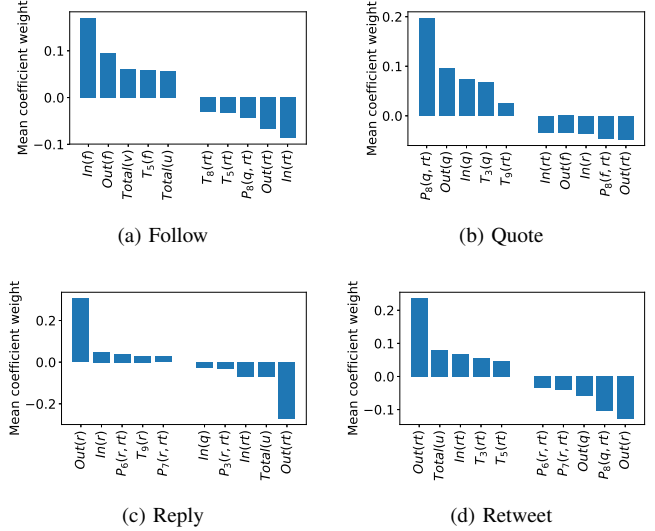


Fig. 7: Highest (in absolute value) logistic regression coefficient weights

of the interactions have zero or very few common neighbors (as we saw in Figure 5).

To consider how prediction accuracy varies by embeddedness, we revisit the broadest form of embeddedness: the number of common neighbors in the undirected graph across all interaction types, i.e., computing embeddedness without regard to directionality of edges or layers. We noted that this notion of full embeddedness, as also depicted in Figure 5, necessarily has higher values of embeddedness than any of the layer-specific measures.

How embeddedness relates to accuracy. Initially, in Fig. 8, we depict for all pair of users grouped by their embeddedness value, the fraction of those pairs for which an edge exists between them. Figure 9 depicts prediction accuracy as we have grouped the results according to the embeddedness of an edge. We find that the task of predicting Quotes and Retweets becomes easier as the embeddedness of an edge increases. Interestingly, for the Follow interaction, we observe that the accuracy actually decreases for the smallest values of embeddedness (with a minimum at 7), which is followed by a steady increase later. This can be explained by the fact that for follows, the absence of any triad between a pair of nodes, is a very strong indication of the absence of a follow relationship among these two users. As we can see in Fig. 8 for only 20% of pairs of zero embeddedness there is a follow relationship

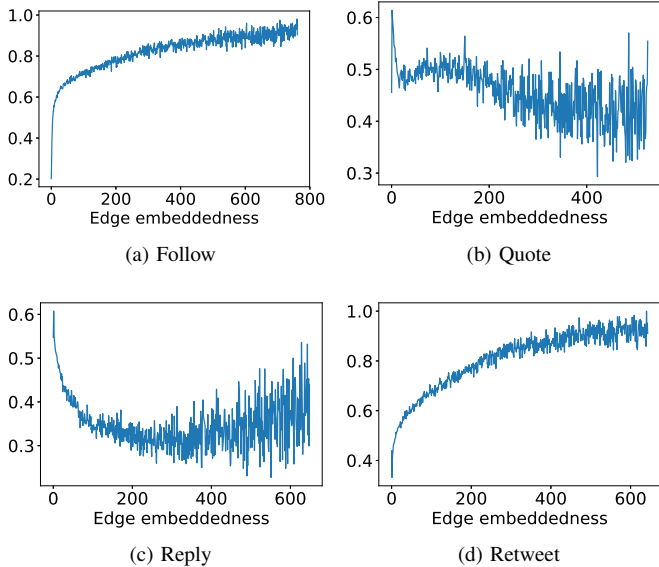


Fig. 8: For all pairs of users that have a certain embeddedness value (x-axis), we calculate the fraction of those pairs that also have an edge between them (y-axis)

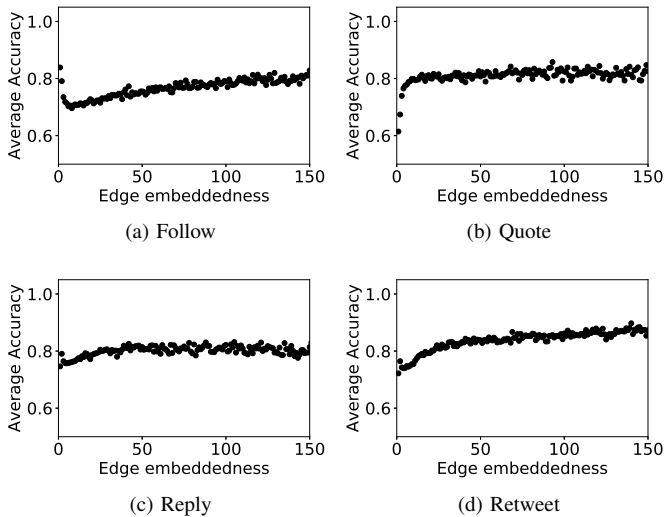


Fig. 9: Average accuracy for edges of certain embeddedness

between them.

When do triads start being more informative than degree? Our final set of experiments test the efficiency of all of our features, both degree and triadic, as we restrict attention to subgraphs whose edges all exceed an embeddedness threshold. On the x-axis of Fig. 10, we vary the threshold value for edge embeddedness in order to include it in our dataset, varying it from 0 to 30. On the y-axis, we plot average link prediction accuracy, with three curves for predictions using only degree features, only triadic features, and all features, respectively.

In general, prediction accuracy follows observations we

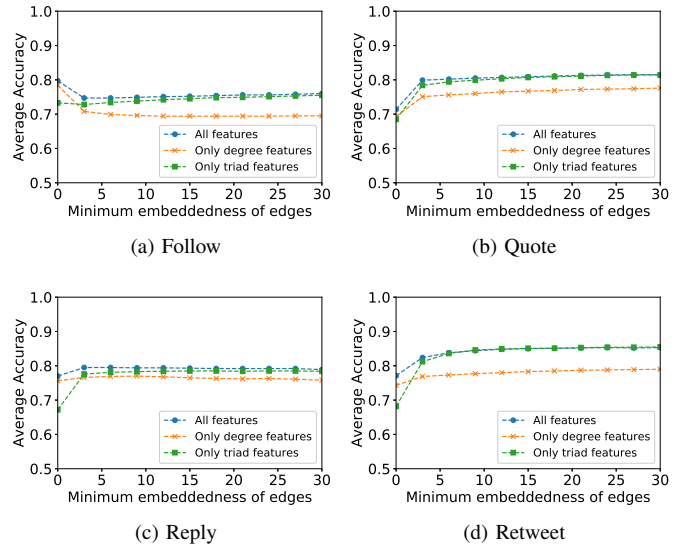


Fig. 10: Prediction accuracy vs. embeddedness threshold. The figures show the prediction accuracy when using only edges above a certain embeddedness threshold (x-axis)

made before. However, a salient and interesting difference is the fact that, while the degree features were the ones that were leading to higher accuracy previously, now, when every edge in our dataset meets an embeddedness threshold, the triadic features are those that become crucial in predicting edges. These observations lead us to the following conclusion: while degree features are important in predicting edges when the two endpoints have few (or no) neighbors in common, the triadic features act complementarily by improving predictions for edges with higher embeddedness. This agrees with existing findings for edge sign prediction [8], [10].

VI. CONCLUSION

Open problems. Our work opens numerous interesting questions in a range of application domains, including two we consider here: in graph anomaly detection, and in approximate graph inference. In the first direction, can we use existing algorithms [23], [24] to locate “anomalous” higher-dimensional subgraphs, e.g., k -cliques for small k , or other observed motifs, and detect subsets of nodes that are dense in these rare subgraphs? In another direction, we note that rate-limiting of requests to the Twitter API is not specific to our work, but exemplifies a challenge in measurement where conducting probes incurs a measurable cost. In this setting, maximizing the utility of a set of measurements that is feasible in a cost or time budget becomes paramount, especially when there is significant correlation and structure across measurements. We view this as especially relevant in scenarios in predictive analytics, where the objective function hinges on prediction accuracy of future queries (such as link predictions) that arrive as an online request stream, not known a priori.

Also, can we use social theories along the lines of [8] to explain how Twitter users react, and which modalities of

interaction they select? Finally, are our findings consistent across other subpopulations of users, e.g., those using either other common languages or forming subcommunities around different shared interests?

REFERENCES

- [1] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Journal of Computational Science*, vol. 2, no. 1, pp. 1–8, 2011.
- [2] J. C. Eichstaedt, H. A. Schwartz, M. L. Kern, G. Park, D. R. Labarthe, R. M. Merchant, S. Jha, M. Agrawal, L. A. Dziurzynski, M. Sap *et al.*, "Psychological language on twitter predicts county-level heart disease mortality," *Psychological Science*, vol. 26, no. 2, pp. 159–169, 2015.
- [3] M. Hürlimann, B. Davis, K. Cortis, A. Freitas, S. Handschuh, and S. Fernández, "A twitter sentiment gold standard for the brexit referendum," in *SEMANTICS*, 2016, pp. 193–196.
- [4] G. Enli, "Twitter as arena for the authentic outsider: Exploring the social media campaigns of Trump and Clinton in the 2016 US presidential election," *European Journal of Communication*, vol. 32, no. 1, pp. 50–61, 2017.
- [5] M. Mendoza, B. Poblete, and C. Castillo, "Twitter under crisis: Can we trust what we RT?" in *Proceedings of the First Workshop on Social Media Analytics*. ACM, 2010, pp. 71–79.
- [6] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: Real-time event detection by social sensors," in *Proceedings of the 19th International Conference on World Wide Web*. ACM, 2010, pp. 851–860.
- [7] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," *Journal of the American Society for Information Science and Technology*, vol. 58, no. 7, pp. 1019–1031, 2007.
- [8] J. Leskovec, D. Huttenlocher, and J. Kleinberg, "Predicting positive and negative links in online social networks," in *Proceedings of the 19th International Conference on World Wide Web*. ACM, 2010, pp. 641–650.
- [9] M. Al Hasan and M. J. Zaki, "A survey of link prediction in social networks," in *Social Network Data Analytics*. Springer, 2011, pp. 243–275.
- [10] C. E. Tsourakakis, M. Mitzenmacher, K. G. Larsen, J. Błasiok, B. Lawson, P. Nakkiran, and V. Nakos, "Predicting positive and negative links with noisy queries: Theory & practice," *arXiv preprint arXiv:1709.07308*, 2017.
- [11] E. Abbe and C. Sandon, "Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery," in *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*. IEEE, 2015, pp. 670–688.
- [12] B. Hajek, Y. Wu, and J. Xu, "Achieving exact cluster recovery threshold via semidefinite programming," *IEEE Transactions on Information Theory*, vol. 62, no. 5, pp. 2788–2797, 2016.
- [13] —, "Achieving exact cluster recovery threshold via semidefinite programming: Extensions," *IEEE Transactions on Information Theory*, vol. 62, no. 10, pp. 5918–5937, 2016.
- [14] S. Petrovic, M. Osborne, and V. Lavrenko, "Rt to win! Predicting message propagation in twitter," in *Fifth International AAAI Conference on Weblogs and Social Media*, 2011.
- [15] W. Galuba, K. Aberer, D. Chakraborty, Z. Despotovic, and W. Kellerer, "Outtweeting the twitterers-predicting information cascades in microblogs," *WOSN*, vol. 10, pp. 3–11, 2010.
- [16] S. Martinčić-Ipšić, E. Močibob, and M. Perc, "Link prediction on twitter," *PloS one*, vol. 12, no. 7, p. e0181079, 2017.
- [17] M. Jalili, Y. Orouskhani, M. Asgari, N. Alipourfard, and M. Perc, "Link prediction in multiplex online social networks," *Royal Society Open Science*, vol. 4, no. 2, p. 160863, 2017.
- [18] D. Hristova, A. Noulas, C. Brown, M. Musolesi, and C. Mascolo, "A multilayer approach to multiplexity and link prediction in online geo-social networks," *EPJ Data Science*, vol. 5, no. 1, p. 24, 2016.
- [19] M. Abufouda and K. A. Zweig, "Are we really friends?: Link assessment in social networks using multiple associated interaction networks," in *Proceedings of the 24th International Conference on World Wide Web*. ACM, 2015, pp. 771–776.
- [20] P. Pratikakis, "twAwler: A lightweight twitter crawler," *arXiv preprint arXiv:1804.07748*, 2018.
- [21] H. Kwak, H. Chun, and S. Moon, "Fragile online relationship: A first look at unfollow dynamics in twitter," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2011, pp. 1091–1100.
- [22] B. Meeder, B. Karrer, A. Sayedi, R. Ravi, C. Borgs, and J. Chayes, "We know who you followed last summer: Inferring social link creation times in twitter," in *Proceedings of the 20th International Conference on World Wide Web*. ACM, 2011, pp. 517–526.
- [23] M. Mitzenmacher, J. Pachocki, R. Peng, C. Tsourakakis, and S. C. Xu, "Scalable large near-clique detection in large-scale networks via sampling," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015, pp. 815–824.
- [24] C. Tsourakakis, "The k-clique densest subgraph problem," in *Proceedings of the 24th International Conference on World Wide Web*. ACM, 2015, pp. 1122–1132.