Analysis and Prediction of Dyads on Twitter

Isa Inuwa-Dutse, Mark Liptrott, and Yannis Korkontzelos

Edge Hill University, Liverpool, United Kingdom
{dutsei, Mark.Liptrott, Yannis.Korkontzelos}@edgehill.ac.uk

Abstract. Social networks are regarded as useful tools for linking *micro* and *macro* levels of sociological theory by enabling the analysis of various forms of relationships. Studies in social science show how a taxonomy of social relationships is described as a function of closeness among users. The closer the users are, the more cohesive and trustworthy. In Twitter, identifying dyadic ties becomes more challenging due to the flexible and eccentric underlying connection pattern. The openness to connect with anyone results in many unidirectional connections that are socially disconnected and ultimately affects tasks such as clustering and, in turn, the veracity of online content. A major challenge is the lack of substantial numbers of dyads and effective means to identify dyads on Twitter at large scale. In this study, we queried vast amount of *verified* and *unverified* Twitter profiles and retrieve dyadic ties for analysis. In the collected data, 55% and 21% of *unverified* and *verified* profiles, respectively, participate in dyadic ties. We describe the usefulness of dyads in the detection of cohesive group of users and validation of content's trustworthiness and how the incorporation of dyadic ties can eventually improve Twitter analysis. Finally, we develop a dyad prediction model using deep learning methods, as our contribution in making dyadic ties useful.

Keywords: Social networks · Twitter · dyadic tie · clustering · reciprocity

1 Introduction

The growing relevance of online socialisation, facilitated by platforms such as Twitter[1] and Facebook[2], attracts much research interest and questions to be addressed. For a long time, *social networks* has been recognised as a useful tool for linking *micro* and *macro* levels of sociological theory [6]. Many forms of social relationships have been analysed at various levels. It can be argued that understanding social interactions today would be incomplete without taking online social relationships into account. Sufficient understanding of the structural properties of online platforms is considered as a crucial factor in the design of a more *human-centric* future internet [2].

[1] twitter.com
[2] facebook.com

However, the growing complexity and heterogeneity of connections makes the task of identifying social relationships at the micro level more challenging. This identification would be useful in detecting cohesive user groups and improve the veracity of information. In Twitter, the global openness to be able to connect with anyone results in many unidirectional connections that are highly likely to be socially disconnected. This makes it difficult to extract dyads. Thus, dyadic ties are usually overlooked in tasks such as clustering and authentication of online content to be posted by trusted users. According to the *ego network model*, which is based on Dunbar's classification of social relationships[4], a *social support clique* consists of a few users connected fully with the strongest relationship in the network [2]. We opined that the level of trust is stronger among users that share dyadic ties and it is highly unlikely for a user in the group to misuse the network e.g. spread rumour, fake news or spam. However, acquiring large amounts of tweets sufficient to identify such cohesive groups is challenging, time consuming and inefficient.

In this study, we analyse a large collection of dyadic and non-dyadic ties and explore its potential application in online clustering and content veracity or authentication. Our study found that 55% and 21% of *unverified* and *verified* profiles, respectively, are involved in dyadic ties. Despite the large proportion of dyads, a random collection of data from Twitter returns much fewer users with dyadic ties. We analyse the cohesiveness of the cliques, i.e. fully connected groups, in terms of size. Finally, we proposed a deep learning method to approach the prediction of dyadic ties, attempting to avoid the time-consuming search for dyads on Twitter. The model learns how to predict dyads using various features. It achieves promising performance when trained on real data. Employing this strategy to check Twitter users limits the danger of spurious content and allows to collect contents from legitimate ones. This is potentially useful in ensuring cohesive clustering and reducing the proportion of irreverent content in Twitter.

The remaining of this paper is structured as follows. Section 2 reviews related work and presents relevant background information. Section 3 describes our method and Section 4 provides a detailed analysis and discussion of the experimental results. Finally, Section 5 concludes the study and suggests some future work.

2 Related work and background

2.1 Networks and online social networks

Relationships and structural properties in networks have been extensively studied at different levels of granularity and sophistication, ranging from the network structure in microscopic organisms to large and complex networks, such as the internet and social networks [5, 16, 18, 14, 12, 1]. While many properties are common across various networks, social networks show different properties with respect to the degree of correlation and tendency for clustering. The formation of clusters is easier and the correlation degree between users is positive [13].

Homophily, the tendency for humans to connect with people they share commonalities with, is central to humans social interaction [11]. Many aspects of homophily have been studied and have been shown to be positively correlated. The work of [19] reported how users in reciprocal relationship exhibit homophily in terms of topics of discussion and [10] investigated homophily in the context of geolocation and popularity. The manifestation of dyad or reciprocity in social networks is viewed from different perspectives and often with contradicting results. With respect to how a popular user follow other users with proportionate popularity level, [10] reports low-level reciprocity and a high proportion of directed connections in Twitter. However, [19] reports high reciprocity in Twitter by computing the ratio of follower/following. The work of [3] computed the probability of a user reciprocating relationship i.e. by following back, and how users of varying influence on Twitter reciprocate most of their followers. We extend this concept by proposing a formal method to predict the likelihood of reciprocity between users.

2.2 Connection in Twitter

Online social media platforms, such as Twitter and Facebook, enable the empirical quantification and evaluation of social relationships among users at an unprecedented scale. Theories and analytical methods can be validated using real social data. We argue that the presence of random connection among some users on Twitter (see Figure 1) contributes to the limited overall cohesiveness and the growing proportion of fake and spam contents. If a user genuinely engages with other users in a bidirectional means, such opportunity will help in curbing the circulation of irreverent information from unknown sources. Users with dyadic ties are more likely to be genuine, trustworthy and will probably result in a more cohesive clusters.

Connections can be directed or undirected on Twitter. For instance, if A follows B, they are in a directed relationship. If B follows A back, their relationship is undirected and equivalent to a dyadic tie (Figure 2). *How can we predict the likelihood of a dyadic tie between random pairs of users*? Section 3 describes our approach to the problem.

3 Method

In this section, we describe our approach to the problem, our data collection pipeline, experimentation and evaluation benchmark.

Definition 1: **dyadic tie**[3] – a relation R over a set D is *dyadic* iff $aRb = 1, \forall a, b \in D$. In the context of this study, a follows b is a directed relationship. If b follows a back, then the resulting relationship is undirected and is called dyad (see Figure 2).

[3] Dyadic tie, pairwise or binary relations are used interchangeable in this work and are considered synonymous.

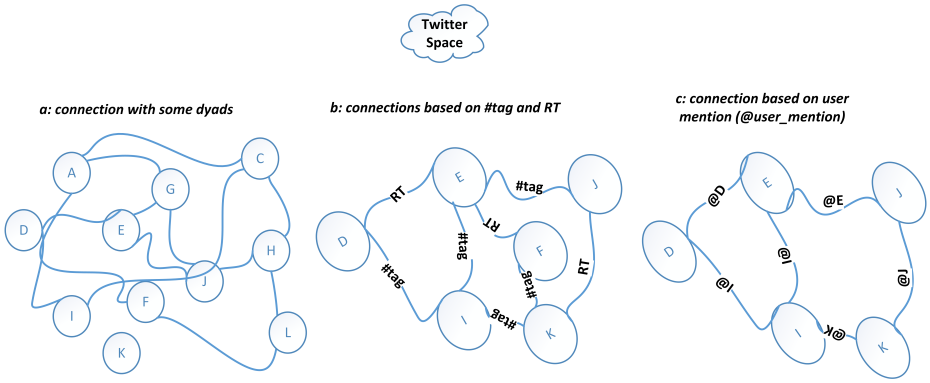

Fig. 1. Connections on Twitter may manifest in different ways such as sharing a link, *re-tweeting (RT)*, using the same or similar *hashtags*, *user mention (@)* or *followership*. Evidently, the connection is porous allowing to connect with many diverse users and limiting the chances of dyads.

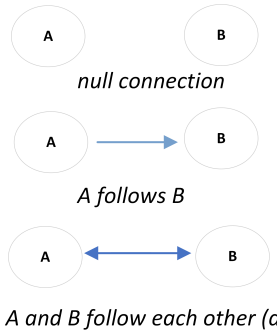

Fig. 2. Possible relationships between two users A and B in Twitter: no relationship (*null connection*), directed relationship ($A \longrightarrow B$) and dyadic or pairwise relationship ($A \longleftrightarrow B$).

3.1 Dataset: collection and training features

We collect data using the Twitter API based on a collection criteria that satisfies the definition of dyadic tie (definition 1). We begin with 4022 seed users[4] from *verified* and *unverified* accounts. A collection crawler is designed to search the profile of each user's network G (i.e. *their list of friends and followers*) to determine whether both users are friends and followers of each other. Essentially, for each user network, $G = \{u|\exists u' \in G\}$ such that $u \cap u' = 1$, i.e. dyadic tie. Table 1 shows basic statistics of users visited by the collection crawler. In particular,

[4] These are genuine users devoid of spammers or social bots collected based on the SPD filtering technique [8]

Table 1. Basic statistics about the data. Many users have to be visited in the *unverified user category* due to the high proportion of $1-edge$ or directed connections in the network, which can be explained by many followers not being followed back on Twitter i.e. $\exists a, b \in D, a \longrightarrow b = 1$ and $b \longrightarrow a = 0$

Category	Seed Size	Visited Users	Retrieved	Remark
Unverified dyads	2,023	13,409,661	8,715	utilised for prediction
Verified dyads	1,999	3,893,075	–	not used for prediction
1-edge and null tie	1,700	–	7,014	utilised for prediction

it shows the counts of directed $(1-edge)$ connections and dyadic ties. Similarly, Figure 4 summarises dyads in *verified* and *unverified* user categories.

To train the prediction model, the following feature groups have been considered:

- **Network features f_n:** followers, friends, account category
- **Text feature f_t:** account description

Features consist of a rich set of meta-data information describing users based on their behaviour and the textual part of their account description.

We use a *Convolutional Neural Network (CNN)* to extract textual features. This is essential because if the users comprising a potential dyad have conflicting ideologies expressed in their profile descriptions, the likelihood of dyadic tie is minimal. According to the collected data (Table 1) and insights from our empirical analysis, we can estimate the likelihood of dyadic tie between users. B is likely to follow A back:

- if A and B are both in the unverified user's category
- if both A and B have low or relatively large number of followers or network size, i.e. based on the average of those metrics in the users' categories
- if A has more followers than B or if A is a verified user.

The opposite of the above statements holds for verified users.

3.2 Prediction pipeline

The set of network and text features $F = \{f_n, f_t\}$ for training our model was introduced in the previous subsection. Among other intrinsic factors, these are the likely features a user can easily access in making a decision to follow back a request or not. Each user U_i is represented by the following vector of reciprocal relationships $U_r^i = [u_{i,j}, u_{i,k}, ..., u_{i,n}]$ where users $j, k..n$ have dyadic ties with user u_i. Features from the *account description text* are learned by applying a CNN on the *n-dimensional* embedding of tokens[5] in text. CNNs has been applied to various domains and many successful studies in NLP have used them [9, 20, 17]. In this study, the CNN is used as a textual feature extraction engine (Figure 3), whose output is encoded using *Long Short-Term Memory (LSTM)*. The encoded vector is merged with the main features for training the prediction model.

[5] We utilise Glove word embeddings [15], pre-trained on tweet collections

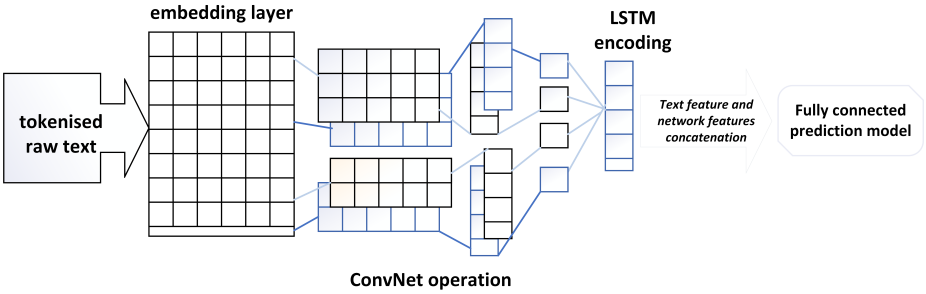

Fig. 3. The *embedding layer* accepts a tokenised *description* text and encodes each token in a dense 100-dimensional vectors to be used by the *ConvNet* part. Finally, the output is transformed by the *LSTM* layer to a lightweight vector that is merged with the *network features* for training.

4 Dyads: Results Analysis

In this section, we present and discuss the results from our study.

4.1 Network topology

Firstly, we analyse the proportions of relationships among users in the data. There exist high proportions of *null connections* and *1-edge connections* as seen in the huge number of *visited users* relative to the *dyads size* in Table 1 and Figure 4. Subsequent analysis and discussion will focus on the *unverified category* since it constitutes ordinary users engaging with each other.

Proportion of nodes and reciprocity *Verified users* have more network neighbours than their *unverified counterparts*, but there is higher proportion of dyadic ties in the *unverified category*, as shown in Figure 4.

4.2 Automatic detection of dyads

Noting the porosity and flexibility in the definition of connections on Twitter and the lack of real connectivity (Figure 1), dyadic ties are rare and difficult to identify in large scale, due to the *curse of dimensionality*. In this section, our aim is to address *the prediction of the likelihood that user A who follows user B will be followed back.* Using a large amount of relevant data, the problem can be modelled as a binary classification task. Given two users A and B with one edge connection between them, the goal is to predict whether pairwise relationship will be established. We build a deep learning classifier that predicts the probability of dyadic tie between two users on Twitter and then we compare with actual dyads collected for evaluation. Figures 6 and 7 show some results from the prediction model.

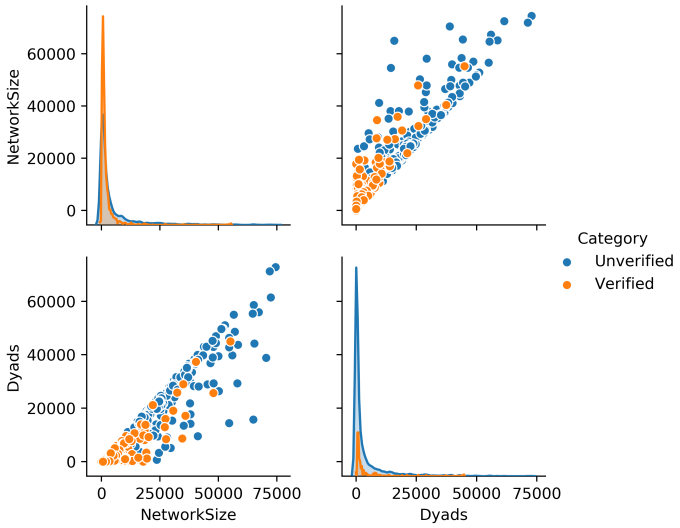

Fig. 4. Proportion of dyadic ties and network size in the data. The *verified* category exhibit larger network sizes but fewer dyads in comparison to the *unverified* category.

Although the performance in Figure 6 is good, it is unstable and seems to be prone to overfitting, noting the proportional relationship between the training accuracy and the validation loss, i.e. both are increasing. We increase the training epochs to 200 and add more layers to the network for stability (Figure 7). There is room for improvement when using larger amounts of data and historical tweets from users.

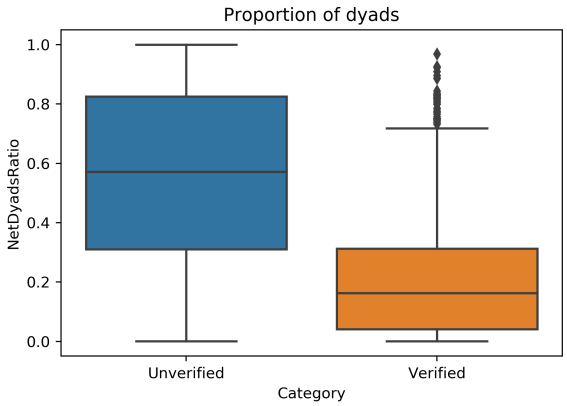

Fig. 5. Dyads proportions in *verified* and *unverified* profiles

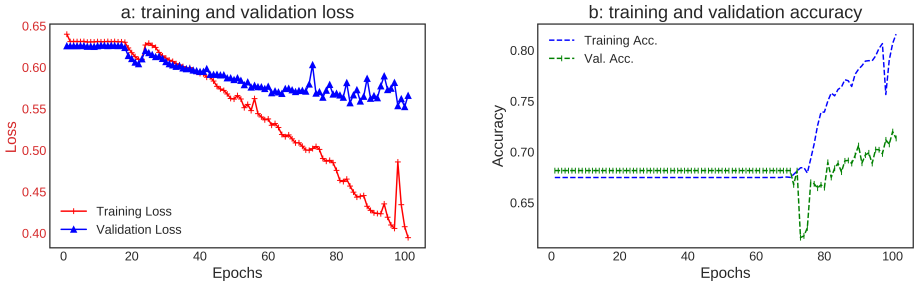

Fig. 6. Dyads prediction performance

4.3 Utility of dyads in clustering and content veracity

Phenomena in real life are associated with numerous network structures and embedded communities. The social media ecosystem enables various forms of interactions among diverse users at various levels. The high dominance of online content from influential users in Twitter makes it difficult to detect low level communities of average users [7]. Low-level communities are a better reflection of true connectivity with strong social cohesion. In this study, we show how the dyadic tie is widespread among users in the unverified category; regarded as proxy for average users on Twitter.

A user with many dyadic ties can be a resourceful representation of a microcosm in Twitter. Such a user can be regarded as a differential entity for deriving new set of related users. For simplicity, if U^3 denotes a user with many dyads of order 3, $3U^2$ and $6U$ are directly related to the user. The constants relate to the size of the users network and the powers the closeness/relatedness to the original user. Such a group can be viewed as a form of *microcosm* in Twitter that can be exploited in various tasks such as clustering.

Fig. 7. Performance of the proposed model on the training and the validation set. The performance remains stable after the first 100 epochs.

In the context of content veracity, a *microcosm* can serve as a unit of analysing groups of users with common online trait, by studying the aspect of homophily. Going by the old adage, *birds of a feather flock together*. A user who spreads rumours or spam content is likely to be strongly connected with similar users. Our future work will explore these aspects from the perspective of dyadic ties.

5 Conclusion

Many relevant theories on various networks and social networks have been proposed and validated analytically or experimentally. Modern social media platforms, such as Twitter and Facebook, enable the empirical quantification and evaluation of social relationships among users at an unprecedented scale. Social network theories and analytical solutions can now be tested using real social data. We conducted an empirical analysis to understand dyads on Twitter, where connections among users are porous, and the composition of communities/groups is not cohesive enough. We collected and curated the first large size dataset that consists of pairwise users. Deeper insight into the underlying mechanisms in dyadic ties on Twitter will be beneficial to many tweet processing tasks. We demonstrated how the recognition of dyads can improve clustering and content validation tasks.

Due to the challenging and time-consuming task of collecting dyads on Twitter, we proposed an effective deep learning prediction method that returns the likelihood of two users engaging in pairwise relationship. The fundamental conclusion is that *dyadic ties* can be predicted (if pair of users are socially active) with good performance and will enable identification of cohesive groups of users on Twitter. This strategy can be applied in detection of cohesive communities of users on Twitter, among other benefits. Employing this strategy can limit the danger of spurious content and allows to collect content from legitimate users. We plan to extend this work to include triads and model how transitive users can be predicted using the concept of *transitivity*.

Acknowledgements

This research work is part of the CROSSMINER Project, which has received funding from the European Unions Horizon 2020 Research and Innovation Programme under grant agreement No. 732223.

References

1. Albert, R., Barabási, A.L.: Statistical mechanics of complex networks. Reviews of modern physics **74**(1), 47 (2002)
2. Arnaboldi, V., Guazzini, A., Passarella, A.: Egocentric online social networks: Analysis of key features and prediction of tie strength in facebook. Computer Communications **36**(10-11), 1130–1144 (2013)

3. Cha, M., Benevenuto, F., Haddadi, H., Gummadi, K.: The world of connections and information flow in twitter. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans **42**(4), 991–998 (2012)
4. Dunbar, R.I.: The social brain hypothesis. Evolutionary Anthropology: Issues, News, and Reviews: Issues, News, and Reviews **6**(5), 178–190 (1998)
5. Erdős, P., Rényi, A.: On the evolution of random graphs. Publ. Math. Inst. Hung. Acad. Sci **5**(1), 17–60 (1960)
6. Granovetter, M.S.: The strength of weak ties. In: Social networks, pp. 347–367. Elsevier (1977)
7. Inuwa-Dutse, I.: Modelling formation of online temporal communities. In: Companion of the The Web Conference 2018 on The Web Conference 2018. pp. 867–871. International World Wide Web Conferences Steering Committee (2018)
8. Inuwa-Dutse, I., Liptrott, M., Korkontzelos, I.: Detection of spam-posting accounts on twitter. Neurocomputing **315**, 496–511 (2018)
9. Kim, Y.: Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882 (2014)
10. Kwak, H., Lee, C., Park, H., Moon, S.: What is twitter, a social network or a news media? In: Proceedings of the 19th international conference on World wide web. pp. 591–600. AcM (2010)
11. McPherson, M., Smith-Lovin, L., Cook, J.M.: Birds of a feather: Homophily in social networks. Annual review of sociology **27**(1), 415–444 (2001)
12. Newman, M.E.: Assortative mixing in networks. Physical review letters **89**(20), 208701 (2002)
13. Newman, M.E., Park, J.: Why social networks are different from other types of networks. Physical review E **68**(3), 036122 (2003)
14. Pastor-Satorras, R., Vázquez, A., Vespignani, A.: Dynamical and correlation properties of the internet. Physical review letters **87**(25), 258701 (2001)
15. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014)
16. Scott, J.: Social network analysis. Sociology **22**(1), 109–127 (1988)
17. Wang, W.Y.: " liar, liar pants on fire": A new benchmark dataset for fake news detection. arXiv preprint arXiv:1705.00648 (2017)
18. Watts, D.J., Strogatz, S.H.: Collective dynamics of small-worldnetworks. nature **393**(6684), 440 (1998)
19. Weng, J., Lim, E.P., Jiang, J., He, Q.: Twitterrank: finding topic-sensitive influential twitterers. In: Proceedings of the third ACM international conference on Web search and data mining. pp. 261–270. ACM (2010)
20. Zhang, X., Zhao, J., LeCun, Y.: Character-level convolutional networks for text classification. In: Advances in neural information processing systems. pp. 649–657 (2015)