When a Friend in Twitter is a Friend in Life

Wei Xie
East China Normal University
linegroup3@gmail.com

Cheng Li
Zhejiang University
wmdty@zju.edu.cn

Feida Zhu
Living Analytics Research Center, Singapore
Management University
fdzhu@smu.edu.sg

Ee-Peng Lim
Living Analytics Research Center, Singapore
Management University
eplim@smu.edu.sg

Xueqing Gong
East China Normal University
xueqing.gong@gmail.com

ABSTRACT

Twitter is a fast-growing online social network service (SNS) where users can "follow" any other user to receive his or her mini-blogs which are called "tweets". In this paper, we study the problem of identifying a user's off-line real-life social community, which we call the user's *Twitter off-line community*, purely from examining Twitter network structure. Based on observations from our user-verified Twitter data and results from previous works, we propose three principles about Twitter off-line communities. Incorporating these principles, we develop a novel algorithm to iteratively discover the Twitter off-line community based on a new way of measuring user closeness. According to ground truth provided by real Twitter users, our results demonstrate the effectiveness of our approach with both high precision and recall in most cases.

INTRODUCTION

One of the fastest-growing social network services (SNS), Twitter has attracted much research interest from social network community. [3, 4] studied the properties of Twitter, such as its topology, geographical features and its power as a new medium of information sharing. A Twitter user can choose to "follow" any other public user, which is a one-way social action requiring no confirmation from the party to be followed. Composed of Twitter users and the follow links among them, we call this online network the Twitter *"follow network"*.

Twitter distinguishes itself from other SNS like Facebook with two unique characteristics. First, as shown in [4], Twitter functions as a mixture of news media and social network combining features from both. Second, mutual consent is not required to establish a follow link. Consequently, a Twitter user can easily gather a huge follow network, much beyond the prediction given by power-law distribution shared by most other social networks [4]. This gives rise to some interesting questions such as: How much does the Twitter follow network reflect a user's off-line real-life social network? Even if two users follow each other, can we conclude that they are friends in person? The eluding social characteristics of Twitter network have so far fogged answers to these important questions. Our aim in this paper is to show that it is indeed possible to identify the portion of the user's Twitter follow network that maps to his or her off-line social life, which we call the user's *Twitter off-line community* or simply *off-line community* for short.

Based on observations from our Twitter data, we put forward three principles in characterizing the off-line community of a user, which are *Mutual Reachability*, *Friendship Retainability* and *Community Affinity*. Incorporating these principles, we propose an algorithm to iteratively grow a user's Twitter off-line community by using random walk with restart to measure user closeness. Our approach uses only the structure of the follow network without probing into the text of tweets or other information like mention or retweet. Interestingly, our results shows that when it comes to relationship mining, a user's follow network could be highly informative.

PROBLEM FORMULATION

Given two Twitter users u and v, we denote as $u \leftarrow v$ if u follows v and $u \leftrightarrow v$ if they mutually follow each other. In this case, v is called u's followee and u is called v's follower. The arrows are assigned in accordance with the direction of the information flow, e.g., all u's tweets will be seen by all those who follow u. A Twitter follow network is represented as a directed graph $G = (V, E)$, such that V is the set of users and $E \subset V \times V$ is the set of all follow links among them. Given two users u and v, denote as $\vec{d}(u, v)$ the directed shortest distance from u to v. Given a target user u, denote as $F^k_{u\leftarrow}$ u's k-hop followee set such that $F^k_{u\leftarrow} = \{v | 0 < \vec{d}(v, u) \leq k, v \in V\}$.

WebSci 2012, June 22–24, 2012, Evanston, Illinois, USA.

Similarly,denote as $F^k_{u\rightarrow}$ u's k-hop follower set such that $F^k_{u\rightarrow} = \{v|0 < \vec{d}(u,v) \leq k, v \in V\}$. Specifically, $F^1_{u\leftarrow}$ is u's followees and $F^1_{u\rightarrow}$ is u's followers. We denote as N^k_u u's *k-hop neighborhood* such that $N^k_u = F^k_{u\leftarrow} \bigcup F^k_{u\rightarrow}$.

Given a user's k-hop neighborhood N^k_u, we call the set of users who are actually in u's off-line social network u's *off-line community* with respect to N^k_u, or simply u's off-line community when the context is clear, which is denoted as C_u. We now give our problem statement.

DEFINITION 1. [**Twitter Off-line Community Discovery**] *For a Twitter user $u \in V$ and a small integer k, given its k-hop neighborhood N^k_u and the follow links among N^k_u, find the u's off-line community.*

OFF-LINE COMMUNITY CHARACTERIZATION

To identify a user's off-line community, we propose three principles: (I)*Mutual Reachability*, (II)*Friendship Retainability* and (III)*Community Affinity*. These principles are based on both our observations from the ground truth provided by real Twitter users and previous work revealing similar insights.

Our Data.

In this work, we limit our target users to those accounts operated by real users correspondent to their account screen name. We used both Amazon Mechanical Turk (AMT) and off-line recruiting to hire real Twitter users for our evaluation.

1. **Registration.** Participants in the evaluation first registered using their Twitter accounts.
2. **Filtering.** We then filter these participants by the following factors: (I) whether they are spammers, (II) how long the account has been registered and (III) if the account has been dormant for a long time.
3. **Data Preparation.** For each qualified participant, we crawled his/her 1-hop neighborhood, and presented them in the form of a webpage together with their profile information.
4. **Evaluation.** The participant would then go over them one by one and label each user as either "off-line friend" or "online friend". The participants have been instructed that the off-line users refer to their real-life friends, colleagues, classmates, advisor of your close friends, people with whom you have had off-line real-life interactions with.

The evaluation task turns out to represent a significant effort as each participant would need to manually evaluate 248 users on average. It is to be understood that the private nature of the task and our quality control has constrained the scale of our evaluation. We finally collected 65 successful evaluations.

Principle I: Mutual Reachability.

Real-life common sense suggests that a necessary, though not sufficient, condition for off-line friendship is that information should be able to reach each other between two friends. However, a Twitter follow link between two users only indicates a one-way information flow from the followee to the follower, i.e., if $u \leftarrow v$, while all v's tweets are delivered to u, those of u's are not automatically visible to v. Therefore, one-way follow links do not offer strong evidence for off-line friendship. Our observation from Twitter data reveals that *mutual reachability*, i.e., whether information can flow in both directions between two users, indeed distinguishes off-line friends from online ones. Note that mutual reachability includes not only two-way direct following but more complicated connecting situations as well. Figure 1 shows, for 65 real Twitter users, the distribution of the percentage of online vs off-line friends satisfying mutual reachability in each user's follow network. It is quite clear that, compared against online friends, a much higher percentage of off-line friends are mutually reachable with the target user. Based on these analysis, we propose our first principle as follows.

PRINCIPLE 1. **Mutual Reachability.** *Given a target user u, for any user $v \in C_u$ with respect to N^k_u, v and u should be mutually reachable.*

Principle II: Friendship Retainability.

Dunbars number suggested that the number of people with whom one can maintain stable social relationships has an upper bound [1], the presence of which has also been confirmed in Twitter [2]. Figure 2 shows the numbers of off-line friends of the 65 Twitter users in our ground truth data, which are upper-bounded by the red line for most users. The existence of such an upper-bound means that the larger the size of a user's neighborhood, the smaller the possibility that a random user from the neighborhood shall be this user's off-line friend. It is to note that Principle 2, as defined as follows, is thus naturally incorporated into our algorithm without specifying explicitly the value for σ.

PRINCIPLE 2. **Friendship Retainability.** *Given a target user u, we should have $|C_u| \leq \sigma$ where σ is a upper-bound threshold measuring friendship retainability.*

Principle III: Community Affinity.

The third principle is based on the observation that a user's off-line friends usually group into clusters such that within each cluster members also know each other personally. These clusters corresponds to different communities where the target user has come to know these off-line friends, e.g., high schools, colleges, work places, etc.. Friends within the same community naturally share common friendship, which we call *community affinity*. Similar observations have been reported in real-life social networks under the terms of *homophily* [5]. As such, Principle 3 is useful in identifying those off-line community members who do not have direct two-way follow

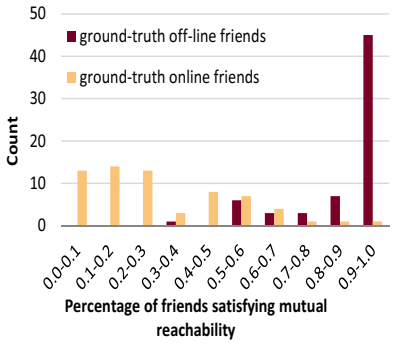

Figure 1. Mutual Reachability.

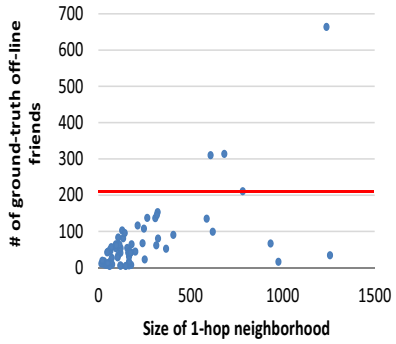

Figure 2. Friendship Retainability.

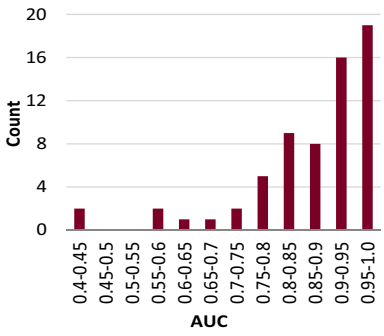

Figure 3. Community Affinity.

links with the target user yet do have strong connections with other off-line community members. The following experiment further illustrates the principle. For each of our 65 Twitter target user u, we examine each user v in u's neighborhood, and count how many off-line friends of u have direct two-way follow links with v. We rank them by the count and compute AUC(Area Under ROC Curve) of the rank list based on the ground-truth off-line friends of u. Figure 3 shows that for most users (52 out of 65), the AUC value is greater than 0.8. This means off-line friends indeed share more direct two-way follow links with other off-line friends, exhibiting much stronger community affinity than online friends. Here we use direct two-way follow links as an indication of greater connection strength.

PRINCIPLE 3. **Community Affinity.** *Given a target user u, for a user $v \in N_u^k$, let $S = \{w | w \in C_u \bigcap N_v^k\}$, the larger the cardinality of S, the more likely we have $v \in C_u$ with respect to N_u^k.*

ALGORITHM

With incorporating the three principles, we propose our algorithm based on the idea of random walk with restart(RWR). It is defined in [6] with the following equation.

$$\vec{r_i} = (1-c)\tilde{W}\vec{r_i} + c\vec{e_i} \tag{1}$$

In our problem setting, given the Twitter network $G = (V, E)$, a target user $u \in V$ and a number k, we focus on G's subgraph G_u^k induced by $N_u^k \bigcup \{u\}$, which is simplified as G_u when k is fixed. A probability transition matrix W is defined for $V(G_u)$ such that, for two nodes $v, w \in V(G_u)$, the entry $W(v, w)$ denotes the probability of v transmitting to w at any step. In accordance with Principle (II), we define $W(v, w)$ as

$$W(v,w) = \begin{cases} \frac{1}{|F_{v\rightarrow}^1|} & \text{if } w \in F_{v\rightarrow}^1 \\ 0 & \text{if } w \notin F_{v\rightarrow}^1 \end{cases} \tag{2}$$

In Equation 1, $\tilde{W}$ is the transpose of the probability transition matrix W as defined above. $\vec{e_i}$ is the starting indicator vector such that $e_{i,i} = 1$ and $e_{i,j} = 0$ where $i \neq j$. $\vec{r_i}$ is the probability vector for node i such that $r_{i,j}$ is the probability of transmitting to node j from i. c is restart probability. It has been shown that $\vec{r_i}$ can be computed iteratively and it finally converges[6]. When it converges, the steady-state probability vector $\vec{r_i}$ reflects the bandwidth of information flow originated from user i to user j for every $j \in V(G_u)$. We use this steady-state probability to define the *closeness score* $c_{i,j}$ for two users i and j:

$$c_{i,j} = r_{i,j} * r_{j,i} \tag{3}$$

The closeness score thus defined satisfies Principle (I). We next explore how to take advantage of the off-line community to identify other unknown members, implementing Principle (III). The idea is to discover the off-line community iteratively, adding new members into the known set in each round. For that purpose, we introduce an auxiliary dummy node, $\hat{v}$, to provide a threshold to cut the new off-line community boundary for each round. $\hat{v}$ is constructed as a virtual node such that (I). $\hat{v}$ and the target user u follow each other, i.e., $\hat{v} \in F_{u\leftarrow}^1 \bigcap F_{u\rightarrow}^1$, (II). $\hat{v}$ only associates with u, i.e., for each $v \in (N_u^k \setminus \{u\}), \hat{v} \notin (F_{v\leftarrow}^1 \bigcup F_{v\rightarrow}^1)$, and (III). the number of followers of $\hat{v}$ is set to be the median of the number of followers of all users in u's k-hop network with the hub users excluded, i.e., $|F_{\hat{v}\rightarrow}^1| = median_{v \in (N_u^k \setminus \mathbb{H})}\{|F_{v\rightarrow}^1|\}$. Hub users, denoted as $\mathbb{H}$, refer to those accounts with more than 2000 followers, which typically belong to celebrities, news media, etc. The dummy node is defined in such a way as to set the lower-bound case for an off-line friend. It simulates the scenario in which the target user u finds by chance this random user $\hat{v}$ who has no connections with u's off-line community. Finding him/her interesting, u follows $\hat{v}$, who then also follows back somehow. As such, $\hat{v}$ represents a connection to u almost as weak as any off-line real-life friend should be.

On a high level, the algorithm works in iterations as follows. Given a target user u, compute the closeness score between u and all the other users as well as $\hat{v}$. A ranking list of all the users together with $\hat{v}$ in decreasing order of the closeness score is thus generated. All the users ranked before $\hat{v}$ are identified as off-line com-

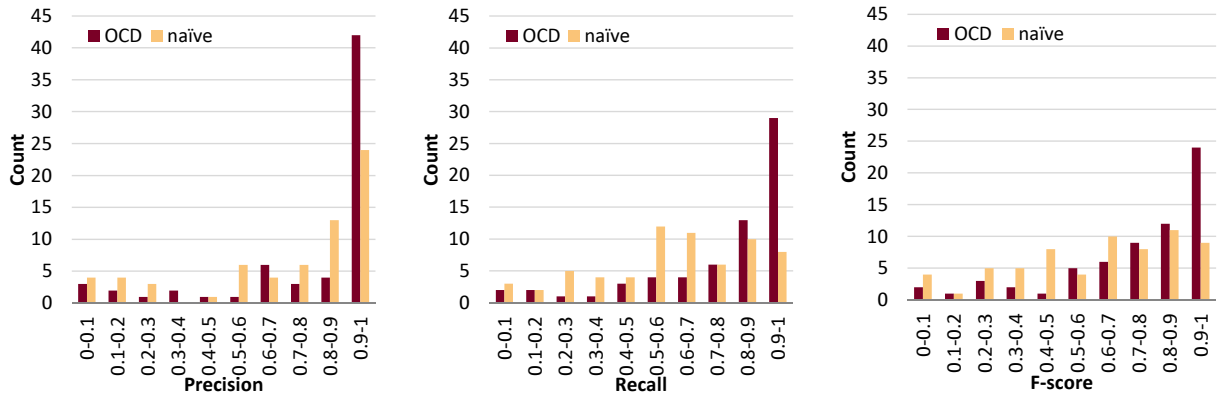

Figure 4. Comparison on distribution of precision, recall and F-score.

munity members, which ends the current iteration. In the next iteration, the key point is that we now treat the whole off-line community identified so far as one virtual user node $\tilde{u}$. Instead of computing the closeness score between u and all the rest users, this time we compute the closeness score between $\tilde{u}$ and every other user. From the ranking list thus generated, if any user jumps ahead of $\hat{v}$ in this iteration, the user will be added to the off-line community of u, which ends this iteration. So on and so forth. As the virtual user node $\tilde{u}$ is actually a set, we now define closeness score between a user node i and a set S as follows.

$$r_{i,S} = \sum_{j \in S} r_{i,j}, r_{S,i} = \sum_{j \in S} r_{j,i} \tag{4}$$

$$c_{i,S} = c_{S,i} = r_{i,S} * r_{S,i} \tag{5}$$

EXPERIMENTAL STUDY

One naive method to identify the off-line community of a target user u is to find the set of users who have direct two-way follow links with u, i.e., they and u follow each other. Do direct two-way follow links provide good indication for off-line real-world friendship? Our experiments suggest that these links are not sufficient. In Figure 4 we show the comparison on the distribution (among the 65 user evaluations)of precision, recall and F score between our algorithm *OCD* and the naive algorithm. In general our solution outperforms the naive solution by a large margin. To conduct more detailed comparison between the two methods,we compute the difference of precision and recall between two solutions for each user. For most users, our solution outperforms the naive solution for both precision and recall. There is only one single case in which our algorithm is prevailed for both precision and recall.

Besides identifying a off-line community through iterations, our algorithm also generates a closeness ranking of all users in the follow network for the target user. Compared against the off-line community found by a clear-cut threshold, this ranking in many cases could be just as useful. For example, when recommending users you have not yet follow, recommending those ranked high in this ranking could be safe. The ranking is based on the closeness score computation in our algorithm. We evaluate the ranking by computing their AUC value for each users. The results shows that more than 60% users' AUC values are greater than 0.9 and more than 80% users' AUC values are greater than 0.8. Moreover, we found that the result will be better, if iteration information, e.g., in which iteration the user is identified, is incorporated into the comparison.

CONCLUSION

In this paper, we proposed the problem of identifying a user's Twitter off-line community. We put forward three principles to characterize off-line community members. Based on these principles, we developed an algorithm to iteratively discover the off-line community by random walk with restart. Results manually evaluated by real Twitter users are shown to illustrate both the effectiveness and the robustness of our algorithm.

Acknowledgements

This research is supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office. This work is also partially supported by National Science Foundation of China under grant numbers 60803022.

REFERENCES

1. R. Dunbar. The social brain hypothesis. *brain*, 9:10.
2. B. Goncalves, N. Perra, and A. Vespignani. Validation of dunbar's number in twitter conversations. *Arxiv preprint arXiv:1105.5170*, 2011.
3. A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65. ACM, 2007.
4. H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *WWW*, 2010.
5. M. McPherson, L. Smith-Lovin, and J. Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 2001.
6. J. Pan, H. Yang, C. Faloutsos, and P. Duygulu. Automatic multimedia cross-modal correlation discovery. In *SIGKDD*, 2004.