When a Friend in Twitter is a Friend in Life

Feida Zhu
Twitter is a unique social platform

1. The “follow” links are established without mutual consent.
   - An explosion of social links
     - Everyone has a large number of followers and followees.
   - A huge number of tweets are generated everyday. (175 million tweets/daily)
   - A further shrinkage of the network diameter
     - Information diffusion is much faster

2. It is a mixture of social network and news media
   - H.Kwak et al  WWW 2010
Follow network = real-life social network?

How much of this follow network reflects a user’s real-life offline social network?

- Mutual follow links do not necessarily indicate real-life interaction.
- The number of followees and followers varies significantly.

Our Problem: Given a Twitter/Weibo follow network of a target user, identify the user’s offline community by examining the follow linkage alone.
Why do we care

1. More accurate and robust user interest modeling
2. Social profile integration from different platforms
3. Spam, Zombie account detection
4. Business competitive analysis
5. Relationship understanding
6. Finer granularity Information/influence diffusion study
Principle I: Mutual Reachability

Information should be able to flow in both directions within a small distance between real-life friends.

- Ground-truth off-line friends
- Ground-truth online friends

Percentage of friends satisfying mutual reachability
Principle II: Friendship Retainability

The size of a user’s offline community has an upper-bound threshold $\sigma$ related to Dunbar’s number.
5. EXPERIMENTAL STUDY

An implementation of our algorithm as a demo system – TwiCube – is publicly available.

5.1 Case Study

We now present a case study on a real user X who participated in our evaluation. X has 107 followers and follows 385 other users. Figure 6 illustrates the discovery of his core community in a total of 4 iterations each indicated by a different color. In summary, 34 users are identified in Iteration 1, 19 in Iteration 2, 3 in Iteration 3 and only one user in the last iteration. The precision and recall for this result of X’s core community is 0.8947 and 0.9807 respectively. It can be observed from Figure 6 that there is a dense clusters of core community members heavily linked among one another (lower left to X) and another such cluster of non-core-community users similarly linked (upper right to X). This shows that approaches based on dense subgraph mining or structural clustering would have a hard time in distinguishing between these two similarly-structured communities and, consequently, identifying the true core community. In fact, this cluster of non-core-community users consists of media, business and active Twitter users sharing similar interests and topics, which is a good indicator of those of X’s own.

In Figure 6, we pick out two particular users, magnify their follow links with X and present them in two cases (a) and (b) (marked by arrows in the figure). In (a), we show the follow network between X and a non-core-community user "tuniu", which is a travel business. Note that although X and this business node directly follow each other, satisfying our Principle 1, this node is still correctly excluded from the core community by our algorithm. This is mainly because it connects mostly with other non-core-community users by follow links, exhibiting weak core community affinity with X. This case would fail the naive approach trying to identify core community members by two-way follow links.

In (b), we show the follow networks between X and a core community member Y, who is discovered in Iteration 3. In this case, X follows Y but Y does not follow X. Moreover, it is not until more core community members have been identified at Iteration 1 and 2 that Y’s sophisticated connections with the core community are revealed. In this tricky case, by unleashing the power of iterated core community identification, our algorithm is still able to correctly identify Y.

5.2 Effectiveness

One naive method to identify the core 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 7 we show the comparison on the distribution (among the 65 user evaluations) of precision, recall and F score between our algorithm CCD 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, let’s take a closer examination at each user. We compute the difference of precision and recall between two solutions for each user. In Figure 8, each point represents one user and the coordinate is defined as \((P_{CCD} - P_{naive}, R_{CCD} - R_{naive})\) where \(P_{CCD}\) and \(R_{CCD}\) is the precision and recall of our algorithm respectively, \(P_{naive}\) and \(R_{naive}\) is the precision and recall of the naive approach respectively. The result shows that for most users, our solution outperforms the naive solution for both precision and recall. In particular, in two cases, the difference is even close to 1. There is only one single case in which our algorithm is prevailed for both precision and recall.

Principle III: Community Affinity

A user’s off-line friends usually group into clusters such that within each cluster members know each other.
Our approach

- Random Walk with Restart
  \[ \vec{r}_i = (1 - c) \vec{W} \vec{r}_i + c \vec{e}_i \]
- Closeness Score
  \[ C_{i,j} = r_{i,j} \times r_{j,i} \]
- Iterative Off-line Community Discovery
  - Off-line community is discovered by iterations.
  - A virtual user node is used as the threshold to cut for each iteration.
5. EXPERIMENTAL STUDY

An implementation of our algorithm as a demo system – TwiCube – is publicly available.

5.1 Case Study

We now present a case study on a real user X who participated in our evaluation. X has 107 followers and follows 385 other users. Figure 6 illustrates the discovery of his core community in a total of 4 iterations each indicated by a different color. In summary, 34 users are identified in Iteration 1, 19 in Iteration 2, 3 in Iteration 3 and only one user in the last iteration. The precision and recall for this result of X's core community is 0.8947 and 0.9807 respectively. It can be observed from Figure 6 that there is a dense cluster of core community members heavily linked among one another (lower left to X) and another such cluster of non-core-community users similarly linked (upper right to X). This shows that approaches based on dense subgraph mining or structural clustering would have a hard time in distinguishing between these two similarly-structured communities and, consequently, identifying the true core community. In fact, this cluster of non-core-community users consists of media, business and active Twitter users sharing similar interests and topics, which is a good indicator of those of X's own.

In Figure 6, we pick out two particular users, magnify their follow links with X and present them in two cases (a) and (b) (marked by arrows in the figure). In (a), we show the follow network between X and a non-core-community user "tuniu", which is a travel business. Note that although X and this business node directly follow each other, satisfying our Principle 1, this node is still correctly excluded from the core community by our algorithm. This is mainly because it connects mostly with other non-core-community users by follow links, exhibiting weak core community affinity with X. This case would fail the naive approach trying to identify core community members by two-way follow links.

In (b), we show the follow networks between X and a core community member Y, who is discovered in Iteration 3. In this case, X follows Y but Y does not follow X. Moreover, it is not until more core community members have been identified at Iteration 1 and 2 that Y's sophisticated connections with the core community are revealed. In this tricky case, by unleashing the power of iterated core community identification, our algorithm is still able to correctly identify Y.

5.2 Effectiveness

One naive method to identify the core 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 offline real-world friendship? Our experiments suggest that these links are not sufficient. In Figure 7 we show the comparison on the distribution (among the 65 user evaluations) of precision, recall and F score between our algorithm CCD 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, let's take a closer examination at each user. We compute the difference of precision and recall between two solutions for each user.

In Figure 8, each point represents one user and the coordinate is defined as $(P_{CCD} - P_{naive}, R_{CCD} - R_{naive})$ where $P_{CCD}$ and $R_{CCD}$ is the precision and recall of our algorithm respectively, $P_{naive}$ and $R_{naive}$ is the precision and recall of the naive approach respectively. The result shows that for most users, our solution outperforms the naive solution for both precision and recall. In particular, in two cases, the difference is even close to 1. There is only one single case in which our algorithm is prevailed for both precision and recall.
Model Accuracy

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

Figure 5. AUC comparison for rankings with and without incorporating iteration information.

REFERENCES

1. R. Dunbar. The social brain hypothesis.


Parameters

- On # of Iterations
- On Robustness

- Count
- Measure value

- # of Iterations
- Perturbation on # of followers

- Precision
- Recall
- F-score
Future Work

1. Deeper understanding of the offline community
2. Social profiling
3. Spam, Zombie detection
4. Business competitive analysis
5. Relationship understanding.