A Research of Directed Closure among Small Nodes on Twitter

Abstract

The relationships between users on Twitter form triadic closure, called directed closure. Directed closure plays an important role driving the evolvement of complex social networks. Former research has only confirmed that directed closure commonly exists among nodes with tens of thousands of followers on Twitter. This paper focuses on exploring how common the phenomenon of “directed closure” is among small user nodes (with less than 200 followers) on Twitter. By collecting and handling data of 30 small user nodes through APIs provided by Twitter and calculating “closure ratio”, I found that even users with only a dozen followers normally present the strong trend of directed closure among relationships between their followers. Though the distribution of closure ratio among small nodes has a lot of varieties and generally is not as high as that of big nodes, the phenomenon of “directed closure” still commonly exists. In 30 user samples, 27 exhibit strong influence by the drive of directed closure. The result is verified to be correct and precise, however, for more specific and accurate results, future work should concentrate on building better data collector which needs abundant fund and human resource.

Introduction

As the rapid development of network, nowadays so many online social networking service websites are getting popular among us, such as Facebook, Twitter, and Flickr. The online social network, or briefly called social network, has a lot of special properties which are worth researching. For instance, it is very important to analyze the structure of these social networks so that developers could build better applications, businessmen could modify their advertising strategies and so on. People get involved in various social network activities which gradually form the structure of a social network. Among the activities, the most important and fundamental is the directed connection between different users. Take Twitter as an example, for some reason, user (node) A connects to another user B to form a directed connection (edge). User B is called a “following” of user A. On the contrary, user A is called a “follower” of user B. On Twitter, here comes a quite interesting property: it seems that nodes prefer to form a triadic directed closure in social network. Romero, D. M., & Kleinberg, J. M. (2010, March) have defined this directed closure process as Figure 1.

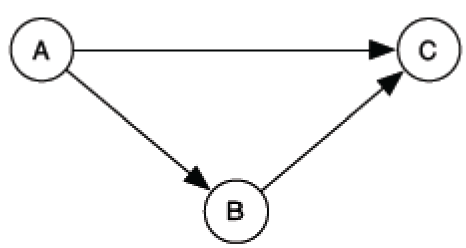

Figure 1. Directed “feed-forward” triangle. If A follows B and B follows C, A would also follow C with relatively high probability. Their relationships form a directed triadic closure.

Like the situation in real life that I introduce one of my friends to another of mine, the phenomenon of "directed closure" is actually the expression of a potential drive for users to follow others who share the same friends on social network. It's reason but not result. To form the triadic closure depicted in Figure 1, except "directed closure", there is always another drive: users always feel like following other users with a large amount of followers. To understand this drive, just think about famous singers and actors who have millions of fans all around the world. We all love chasing some famous people or they do have the quality and capability to become popular and be followed by many others. Romero, D. M., & Kleinberg, J. M. (2010, March) provided a feasible way to discern the same triadic closures caused by these two drives among nodes with large quantities of followers. They focused their efforts on big celebrities (users with many followers) because they thought these nodes played important roles in social network and they did have achievements.

However, former research of "directed closure" neglected to take the nodes with very low in-degree (number of followers) into consideration. Even the pioneer (D. M., & Kleinberg, J. M. 2010, March) of this topic only took 18 users on Twitter with between 10,000 and 50,000 followers as their sample. Though the phenomenon of "directed closure" does exist in these high in-degree nodes, the problem is that whether this phenomenon is still common in those nodes with only tens or hundreds of incoming edges which stand for a considerable part of all the nodes of the whole network. So the purpose of the paper is to take a look at the relationships between followers of users with less than 200 followers on Twitter, to assess how common the phenomenon of "directed closure" exists among the followers of these users and to explain the result in detail.

Method

I picked some of the users who had less than 200 followers on Twitter randomly as samples to analyze the phenomenon of directed closure. These users were chosen from different categories such as magazine, ordinary people, and computer technology and so on. The purpose is to see if the existence of directed closure is common or plays an important role among small nodes of online social network and to compare it to an assumed random network as what former researchers did.

Similar to the method by Romero, D. M., & Kleinberg, J. M. (2010, March), I considered a sample user as a node which represented a micro celebrity $\mu - celebrity$ as depicted in Figure 1. By traversing the $\mu - celebrity$ node's follower list, each follower was checked to determine if the follower was followed by another follower. If there was one (or more than one) different follower following a follower of the $\mu - celebrity$ node, then the follower was treated as having directed closure. In the end, I calculated the fraction of $\mu - celebrity$ node's followers having directed closure as closure ratio. That was the work for real-world online social network.

Next, for an assumed random network, the probabilities to form an edge between any two nodes are the same as p ($p \in [0, 1]$). So for a $\mu - celebrity$ node with N incoming edges, the probability for an edge having directed closure is $1 - (1 - p)^{N-1}$ ($p \in [0, 1]$). It is also the value of the node's closure ratio. As I investigated the MAG model's experiment by Kim, M., & Leskovec, J. (2010), I defined the value of p as 0.0009 (calculation omitted). So for a random network, it has a probability (estimated closure ratio) of $p_{CR}(N) = 1 - 0.9991^{N-1}$ for a node with Nfollowers to have directed closure. Thus the node should have taken part in $CR(N) = N(1 - 0.9991^{N-1})$ directed closures.

Through the comparison between the real closure ratio of sample nodes and estimated closure ratio of random network, I finally got my conclusion. If the sample node with N followers had a closure ratio bigger than $p_{CR}(N)$, the phenomenon of "directed closure" existed. Thus the corresponding $p_{CR}(N)$ acted as a deciding threshold.

There was a little difference from the original method. As I mentioned in introduction section, there were only two competing drives at work: one was "directed closure" (the topic of the paper), another was "the tendency to link first to celebrity" (a node always tended to connect to another node with more incoming edges first instead of another node with a few incoming edges). The two forces could both cause the closure shown in Figure 1. But the later one was in this order: A→C, B→C, and then A→B. In original method, this formation order was eliminated from the result. However, I didn't take the order of formation of the directed edges into consideration because these sample nodes were so small (less than 200 followers) that the tendency of link first to celebrity didn't exist. The order of edge formation was mostly in the order: A→B, B→C, and then A→C. So it could only be the result of "directed closure" if the closure ratio was obviously bigger than that of a random network.

The data collection was based on Twitter API 1.1 and Twitter4J library which was quite limited by the rank of developer. As a free user and developer, I could only call the API 150 times per hour. So to analyze a single user even with 200 followers, it might take a day. To develop a better data collection tool and raise fund to apply for high rank authority for gathering user information is a significant future work.

Results

I picked 30 users on Twitter completely randomly as the sample of my experiment. These users are individuals, couples, organizations and companies. Their numbers of followers vary from a dozen to two hundred. Through APIs provided by Twitter, I gathered all of their follower lists, and analyzed all of the followers' relationships to count the directed closures existed among these users' friends. The result is shown as Table 1.

Column "id" stands for the sampled user's account id on Twitter. "screenName" is the unique user name of a user on Twitter. "closureCount" shows how many followers of the user attend into a directed closure. "followers" is the number of the user's followers. "Closure Ratio" is the fraction of followers who have attended to form a directed closure. "Threshold" is the estimated closure ratio for a random network's node with this amount of followers. "Conclusion" exhibits that whether the phenomenon of "directed closure" really exists among the user's friends. If "closure ratio" is bigger than "Threshold", then the conclusion is "TRUE".

According to the analysis of data gathered, only 3 in 30 of the sample users' friendships on Twitter aren't strongly affected by the drive of "directed closure". In the end of Table 1, I provided an average calculation of all the samples. The average closure ratio of these randomly picked 30 sample users is 0.485456369 which is more than twice the value of the biggest threshold 0.164045124 calculated for a random network when the number of a node's followers is 200. What's more, there's only 1 of the 30 sample users whose followers don't have any sign of directed closure.

Table 1
Closure ratio distribution of 30 sample users on Twitter

id	screenName	closureCount	followers	Closure Ratio	Threshold	Conclusion
140685387	TheAceMag	165	180	0.916666667	0.148854797	TRUE
818002789	YPinDC	115	146	0.787671233	0.122395077	TRUE
375563276	YM_all	9	12	0.75	0.00985557	TRUE
154255539	withlove_logan	103	152	0.677631579	0.127123493	TRUE
43200636	Bermudamark	7	11	0.636363636	0.008963637	TRUE
1451742258	Fado_Justina	66	126	0.523809524	0.106447917	TRUE
332818072	lilypaddc	94	182	0.516483516	0.150386169	TRUE
371361071	codydot619	7	15	0.466666667	0.012526555	TRUE
1622544864	Halverpolo	2	10	0.2	0.008070901	TRUE
1931296171	COTM_DC	8	47	0.170212766	0.04057261	TRUE
1866110604	ThingsDC	17	111	0.153153153	0.094297666	TRUE
1676794099	wilmatilinda	1	11	0.090909091	0.008963637	TRUE
365471008	gfish818	10	115	0.086956522	0.097553795	FALSE
34250043	malexjones	0	19	0	0.016076663	FALSE
426828567	EBPembrokePines	7	134	0.052238806	0.112861263	FALSE
940263678	lissa_randall	6	12	0.5	0.00985557	TRUE
251902551	sugeepie	20	39	0.512820513	0.033636672	TRUE
931255368	vickytadomingue	5	16	0.3125	0.013415281	TRUE
1683568550	GabrieAcosta	4	15	0.266666667	0.012526555	TRUE
254083485	luissierraochoa	2	35	0.057142857	0.030149922	TRUE
297085399	jperezco	6	17	0.352941176	0.014303207	TRUE
45285078	wigevi	8	53	0.150943396	0.045741875	TRUE
1468724030	JeNaShoes	6	62	0.096774194	0.053443598	TRUE
1170891884	karina28ro	24	54	0.444444444	0.046600707	TRUE
336241543	WilliamRueda2	46	58	0.793103448	0.050028314	TRUE
1918294825	ProfeKleyman	38	55	0.690909091	0.047458766	TRUE
155261261	SUAITA2011	39	85	0.458823529	0.072844554	TRUE
45749807	AQuil24	7	22	0.318181818	0.018730866	TRUE
552026064	Verdictdeposit	146	200	0.73	0.164045124	TRUE
	[Average]	968	1994	0.485456369		

For the need of analysis, the distribution of closure ratio and threshold is show in Figure 2. Any point above the threshold line is treated as a sample with the phenomenon of "directed closure". The x-axis of Figure 2 stands for the number of followers. With simple division, it's easy to analyze the closure ratio distribution in 3 parts: users with followers from 10 to 30, 40 to 100, and 100 to 200. It can be observed that half of the sample users' closure ratios are higher than 0.4, and more than two thirds of the sample users have higher closure ratio than the biggest threshold. 27 in 30 of the sample users have higher closure ratio than their corresponding threshold. Only one user has closure ratio higher than 0.9.

Also, there's commonness and varieties among all these small user nodes. When having less than 200 followers, users with different number of followers all tend to have different closure ratios between 0.1 and 0.9 instead of having similar closure ratio in a smaller range. However, for users with very few followers like a dozen or two, they tend to have random closure ratio between 0 and 0.75. For users with 40 to 100 followers, the distribution of closure ratio goes to two ends. Almost half of those sample users have closure ratio higher than 0.4, and the other half have closure ratio lower than 0.2. Similar situation happens when it comes to users with more than 100 followers. But more of those users have closure ratio higher than 0.5 while only a few of those users have closure ratio lower than 0.2.

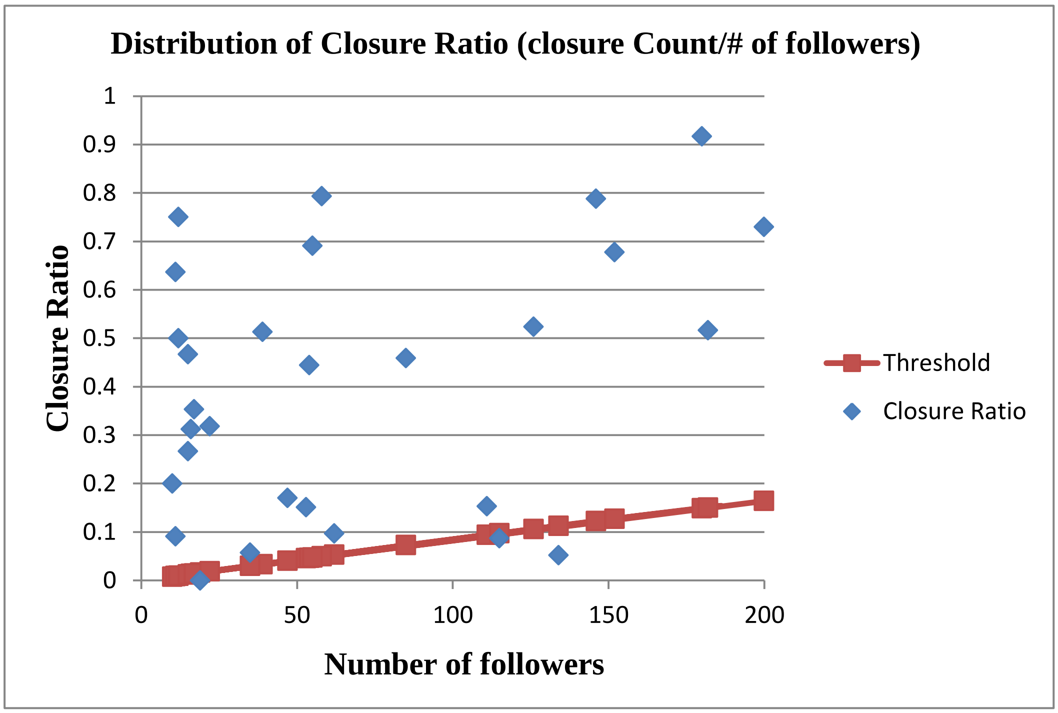

Figure 2. Distribution of closure ratio and threshold

Discussion

The purpose of the paper is to take a look at the relationships between followers of users with less than 200 followers on Twitter, to assess how common the phenomenon of "directed closure" exists among the followers of these users and to explain the results in detail.

I sampled 30 users with different numbers of followers from Twitter randomly representing different roles so that the sample would be representative and the result would be more convincing. Since the small-world theory proposed by Kleinberg, J. M. (2000), every user could connect to another user within small limited connections in social network. Also, as an example, the Twitter data collection of Romero, D. M., & Kleinberg, J. M. (2010, March) only concerned 18 users on Twitter. So it's enough to gather 30 randomly chosen users' followers' information for my research. Actually, the 30 users connect to about 2000 other users, which form a relatively large part of the whole social network of Twitter.

Generally, the result shows that almost all (27 in 30) of the users on Twitter with less than 200 followers do have the phenomenon of "directed closure" among their followers' relationships. Which indicates that users, who decide to follow another unfamiliar user X with less than 200 followers, always make the decision by seeing that if anyone of their friends has already followed the user X. So this explains a lot in social network. For instance, news diffused by friends is more convincing than that from stranger; Users prefer to make friends with strangers who share common friends with them; "ties are highly clustered" by Centola, D. (2010). Also, directly or indirectly, the result supports many of the properties observed from social network such as "small diameter" by Kim, M., & Leskovec, J. (2010).

However, the whole distribution is different from what was observed from users with 10000 to 40000 followers by Romero, D. M., & Kleinberg, J. M. (2010, March). In their research, those median user nodes (relatively smaller than users with hundreds of thousands or millions of followers) on Twitter always have closure ratio higher than 0.3. Most of them have closure ratio between 0.7 and 0.9. Differently, my research result shows that only two thirds of users on Twitter have closure ratio higher than 0.3. A few (3 in 30) of the users don't have big enough closure ratio for their size to reflect the drive of directed closure.

Moreover, even in the small range for number of followers between 10 and 200, the distribution of closure ratio has big variety among these users. For users with 10 to 30 followers, their followers seem to depend on their friends to find new friends randomly. Some of them totally make friends with those have common friends. Some totally make friends with definite strangers. Some don't care if the one they are going to follow is an indirect friend or not. The distribution of closure ratio is scattered from 0 to 0.8 evenly. These users who have very few followers normally are not active in their social network. So it's not surprised to see how unpredictable their closure ratio is.

While for users with more followers, their followers tend to develop into two totally different groups. Members of one group pick their new friends carefully by discovering if they have common friends. Members of another group never pick new friends through anyone they've already known. This might be the result of users in online social network choose different living styles online. Some users would like to share their life with their real life friends or families more conveniently online. Their behaviors make them end up into the first group. Others want to make a different life online to do what they would never do and talk with people they would never know in real life. They are more likely to be categorized into the second group.

In the main, the phenomenon of "directed closure" is common even in small user nodes with less than 200 followers. One of the few differences from former research is that the influence of driving by "directed closure" is lighter than that of users with tens of thousands of followers. Also, the distribution of closure ratio is various from former research.

Conclusions

My current work extends the former research about directed closure to a wider range of users on Twitter. I investigated the distribution of closure ratio among 30 users with less than 200 followers on Twitter through former researchers' methods with some modification. I picked the sample carefully so that they would be representative and the result would be convincing.

27 in 30 sample users' data presented strong sign of "directed closure" with higher closure ratio than threshold. Half of the users explored had closure ratio higher than 0.4, and two thirds of the users had closure ratio higher than the biggest threshold. The results showed that it was also common to observe the consequence driven by "directed closure" among small user

nodes and their followers. Though the distribution of closure ratio was various from former research, the results could be completely explained.

The results could help explain some properties presented by social network structure such as “small-world”. It can also be treated as support material for former research of directed closure. Through the understanding of why the results happened in last section, people could change their behavior in social network to make their social life better.

However, there’s still further work need to be done. For example, there’s limited research on huge user nodes in social network. Also, to explore those huge nodes, more efficient data collector and higher rank API access are needed.

References

Barabási, A. L., Albert, R., & Jeong, H. (1999). Mean-field theory for scale-free random networks. Physica A: Statistical Mechanics and its Applications, 272(1), 173-187.

Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. science, 286(5439), 509-512.

Centola, D. (2010). The spread of behavior in an online social network experiment. science, 329(5996), 1194-1197.

Cioffi‐Revilla, C. (2010). Computational social science. Wiley Interdisciplinary Reviews: Computational Statistics, 2(3), 259-271.

Funk, S., & Jansen, V. A. (2010). Interacting epidemics on overlay networks. Physical Review E, 81(3), 036118.

Girvan, M., & Newman, M. E. (2002). Community structure in social and biological networks. Proceedings of the National Academy of Sciences, 99(12), 7821-7826.

Kim, M., & Leskovec, J. (2010). Multiplicative attribute graph model of real-world networks. In Algorithms and Models for the Web-Graph (pp. 62-73). Springer Berlin Heidelberg.

Leskovec, J., Kleinberg, J., & Faloutsos, C. (2005, August). Graphs over time: densification laws, shrinking diameters and possible explanations. In Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining (pp. 177-187). ACM.

Romero, D. M., & Kleinberg, J. M. (2010, March). The Directed Closure Process in Hybrid Social-Information Networks, with an Analysis of Link Formation on Twitter. In ICWSM.

Tang, J., Lou, T., & Kleinberg, J. (2012, February). Inferring social ties across heterogenous networks. In Proceedings of the fifth ACM international conference on Web search and data mining (pp. 743-752). ACM.

Zheleva, E., Sharara, H., & Getoor, L. (2009, June). Co-evolution of social and affiliation networks. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1007-1016). ACM.