Information Network or Social Network? The Structure of the Twitter Follow Graph

Seth A. Myers, Aneesh Sharma, Pankaj Gupta, and Jimmy Lin
Twitter, Inc.
@seth_a_myers @aneeshs @pankaj @lintool

ABSTRACT

In this paper, we provide a characterization of the topological features of the Twitter follow graph, analyzing properties such as degree distributions, connected components, shortest path lengths, clustering coefficients, and degree assortativity. For each of these properties, we compare and contrast with available data from other social networks. These analyses provide a set of authoritative statistics that the community can reference. In addition, we use these data to investigate an often-posed question: Is Twitter a social network or an information network? The "follow" relationship in Twitter is primarily about information consumption, yet many follows are built on social ties. Not surprisingly, we find that the Twitter follow graph exhibits structural characteristics of both an information network and a social network. Going beyond descriptive characterizations, we hypothesize that from an individual user's perspective, Twitter starts off more like an information network, but evolves to behave more like a social network. We provide preliminary evidence that may serve as a formal model of how a hybrid network like Twitter evolves.

Categories and Subject Descriptors: H.3.5 [Online Information Services]: Web-based services

Keywords: graph analysis; social media

1. INTRODUCTION

We provide a characterization of the topological properties of a snapshot of the Twitter follow graph with two goals: First, we present a set of authoritative descriptive statistics that the community can reference for comparison purposes. Second, we use these characterizations to offer new insight into a question that many have asked: Is Twitter a social network or an information network?

To answer this question in a meaningful way, we must first define a social network. Unfortunately, there is no universally-accepted definition by researchers in network science. However, one can point to a prototype such as Facebook, which researchers would all agree is a social network. Our operational definition of a social network is simply one that exhibits characteristics we observe in other social networks. These include high degree assortativity, small shortest path lengths, large connected components, high clustering coefficients, and a high degree of reciprocity. An information network, on the other hand, is a structure where the dominant interaction is the dissemination of information along edges: these are characterized by large vertex degrees, a lack of reciprocity, and large two-hop neighborhoods.

Intuitively, Twitter appears to be *both*. On the one hand, the follow relationship seems to be primarily about information consumption. Users follow a news outlet not because of any meaningful social relationship, but to receive news. Except in a few special cases, these edges are not reciprocated—this is Twitter being used as an information network. On the other hand, it is undeniable that many follow relationships are built on social ties, e.g., following one's colleagues, family members, and friends. In these cases, Twitter behaves like a social network. Not surprisingly, our analyses show that the Twitter graph displays characteristics of both an information network and a social network.

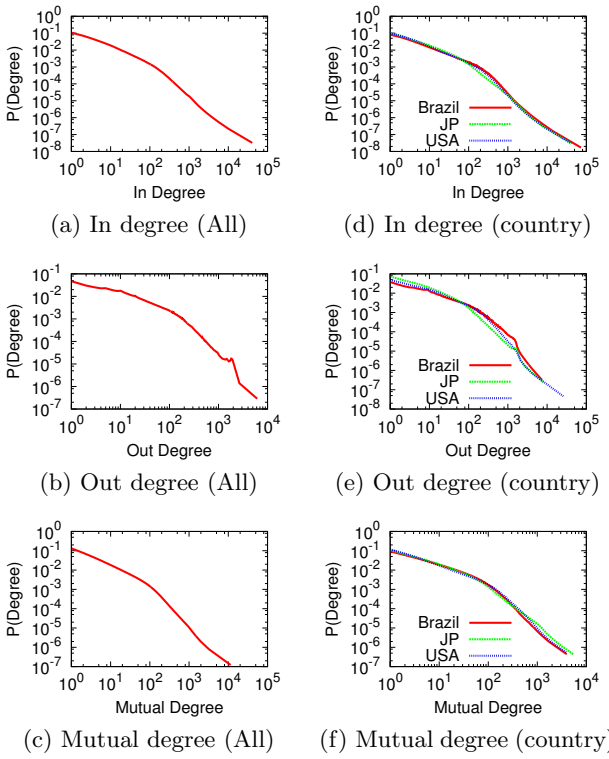
Why is this question important? From an intellectual perspective, we believe that Twitter exemplifies a hybrid system and it is important to understand how such networks arise and evolve. From a practical perspective, a better understanding of user and graph behavior helps us build products that better serve users.

2. DATA AND METHODS

Our analysis is based on a snapshot of the Twitter follow graph from the second half of 2012 with 175 million active users and approximately twenty billion edges. In addition to the complete follow graph, we also examined the subgraphs associated with three different countries: Brazil (BR), Japan (JP), and the United States (US). Each subgraph contains only vertices corresponding to active users logging in from that particular country. Unlike previous work (e.g., [7]), our study is based on a complete graph snapshot, and thus our findings are free from artifacts that can be attributed to methodological issues around crawling.

Because the follow relationship is asymmetric, the Twitter follow graph is *directed*. To facilitate comparisons, our analyses make reference to the undirected *mutual graph*, which is the graph with just the edges that are reciprocated. An edge between two users in the mutual graph implies that both users follow each other. In all, 42% of edges in the follow graph are reciprocated, so there are a total of around four billion undirected edges in the mutual graph.

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2576939.



(a) In degree (All) (d) In degree (country)

(b) Out degree (All) (e) Out degree (country)

(c) Mutual degree (All) (f) Mutual degree (country)

Figure 1: Degree distributions in the follow graph.

Our findings are contrasted with studies of other social networks: Facebook [1, 14] (721m vertices, 68.7b undirected edges) and the network from users of MSN Messenger [8, 13] (180m vertices, 1.3b undirected edges). Properties of these two social networks provide a point of reference.

When considering the size of the Twitter graph, computing exact values of different statistical quantities is challenging. In many cases, we performed approximations, noted in each section. All analyses in this work were conducted on Twitter's Hadoop analytics stack using Pig. More details about analytics infrastructure at Twitter can be found elsewhere [10].

3. GRAPH CHARACTERISTICS

3.1 Degree Distributions

Since the Twitter follow graph is directed, vertices have both an *inbound* degree, or in-degree (the number of users who follow them) and an *outbound* degree, or out-degree (the number of users who they follow). Figure 1(a) shows the in-degree distribution across all users. Not surprisingly, we see a heavy tail resembling a power-law distribution. The out-degree distribution in Figure 1(b) also exhibits a heavy tail, although not to the same extent as the in-degree distribution. This is interesting, as one might expect that users' limited capacity to consume information would set a relatively low upper bound on the number of people they can follow. Instead, some users follow hundreds of thousands of accounts. These are often celebrities who choose to reciprocate the follows of some of their fans (in some cases, automatically). For example, in Summer 2012, the pop singer Lady Gaga was the most-followed user on Twitter, and she followed more than 130k other users; Barack Obama had 21m followers and followed more than 600k people. In another common case, businesses will reciprocate follows to better connect with customers (for example, the grocery store chain Whole Foods followed more than 500k people). More commonly, journalists have been found to follow many thousands of people.

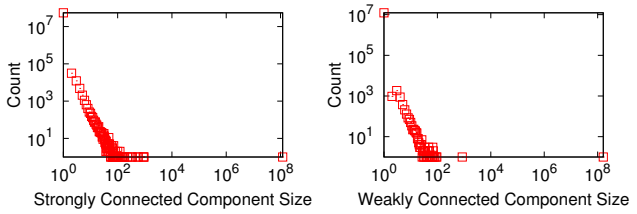
Network	25%	50%	75%	95%	Max	α	μ	σ^2
In-All	4	16	65	339	14.7m	1.35	2.83	3.36
Out-All	11	39	121	470	757k	1.28	3.56	2.87
Mut-All	3	13	50	223	563k	1.39	2.59	3.03
In-BR	6	32	127	514	3.0m	1.30	3.34	3.76
Out-BR	16	69	209	894	140k	1.25	4.03	3.28
Mut-BR	5	19	57	204	115k	1.35	2.83	2.63
In-JP	4	17	60	347	1.2m	1.35	2.84	3.21
Out-JP	6	23	71	360	297k	1.32	3.08	2.93
Mut-JP	4	15	50	253	276k	1.37	2.72	2.90
In-US	4	20	89	402	5.1m	1.33	3.01	3.59
Out-US	11	43	138	509	325k	1.28	3.62	3.05
Mut-US	4	16	64	257	235k	1.36	2.76	3.14

Table 1: Statistics for the degree distributions (inbound, outbound, and mutual) for the four graphs we examined. The parameter α assumes $P(x) \sim x^{-\alpha}$ for degree x (power law). The μ and σ^2 parameters assume $P(x) \sim \frac{1}{x} \exp\left[\frac{(\ln x - \mu)^2}{2\sigma^2}\right]$ (log-normal).

The presence of users with thousands of followings is indicative of "non-social" behavior. It has been well-established that individuals are only able to maintain around 150 stable social relationships at a time [3]. Furthermore, it has been established through studying reciprocated direct messaging on Twitter that the number of social relationships a user can effectively maintain is limited by this constraint as well [5]. Also of note in the out-degree distribution is the apparent spike at 2,000 followings. Spambots have been observed in the past to arbitrarily follow a large number of people. To curtail this, Twitter does not allow users to follow more than 2,000 accounts unless they themselves have more than 2,200 followers. This is not to imply that all users who fall in this spike are spambots, but only more well-known users can "break through" this limit.

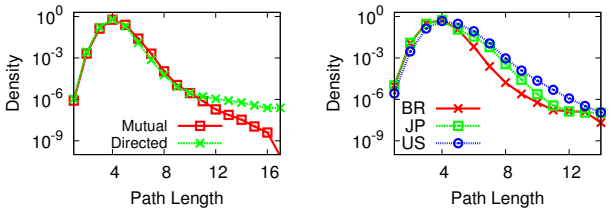
The degree distribution of the mutual graph is shown in Figure 1(c). Here, we still observe relatively large degrees, although smaller than both the in-degrees and out-degrees. Figures 1(d), 1(e), and 1(f) show the various degree distributions for each of the three country subgraphs. Surprisingly, there is very little variation between them.

Finally, Table 1 shows the statistics of the various degree distributions. In addition to the percentiles of each distribution, we also tried fitting each to both a power law and a log-normal distribution. Interestingly, both the in-degree distribution and the mutual degree distributions were best fit by a power law, while the out-degree distribution was best fit by a log-normal. Furthermore, each of the percentiles reported are *higher* for the out-degree distributions compared to the in-degree or mutual degree, even though the maximum out-degree is much smaller than the maximum in-degree. This means that the typical Twitter user follows more people than she has followers, but this does not hold for a small population of "celebrity" users who have very large in-degrees (i.e., many followers).



(a) SCC (b) WCC

Figure 2: The connected component size distributions of the follow graph.



(a) Entire Graph (b) Selected Countries

Figure 3: The distribution of path lengths in the mutual and follow graphs.

Conclusion: The degree distributions of the Twitter graph is inconsistent with that of a social network. It is highly unlikely that an individual can maintain as many social relationships as the out-degrees of the vertices suggest.

3.2 Connected Components

In directed graphs, the distinction is often made between *weakly* and *strongly* connected components. In a weakly connected component, determination of connectivity ignores edge direction, whereas in a strongly connected component, a pair of vertices must be reachable through a directed path.

Figure 2 shows the size distribution of the two types of components in the Twitter follow graph. One can see that in both cases, there is a single large component that dwarfs the other components in size. The largest weakly connected component contains 92.9% of all active users. Of the remaining vertices not in this largest component, the majority are completely disconnected because they contain no edges at all—they default to component sizes of one. Of all vertices with at least one edge (inbound or outbound), the largest weakly connected component contains 99.94% of all vertices.

In contrast, the largest strongly connected component only contains 68.7% of all active users. This is interesting in that the figure is much lower compared to other social networks—in both the Facebook graph [14] and the MSN messenger graph [8], the largest connected component contains more than 99% of all vertices. One possible explanation of this is that a reciprocated edge in Twitter is often an indication of a social connection, but perhaps Twitter is not entirely a social network to begin with. In fact, more than 30% of active users do not have a single mutual edge—these are instances where the user is exclusively using Twitter for either information dissemination or consumption.

Conclusion: Due to the abundance of unreciprocated edges, the Twitter graph is less well connected than one would expect if it were a social network.

3.3 Shortest Path Lengths

The path length between users is the number of traversals along edges required to reach one from another; the distribution of shortest path lengths quantifies how tightly users are connected. Here, we examine both symmetric paths over the mutual graph as well as directed paths across the follow graph. We see that the Twitter graph exhibits small path lengths between vertices, which previous work has argued to be a typical characteristic of social networks [2].

It is computationally infeasible to identify the shortest path length between every pair of vertices in a large graph. For the Twitter follow graph, there are about $N \times (N-1) = 2.6 \times 10^{20}$ different shortest paths (where N is the number of connected vertices), and the mutual graph has 7.3×10^{15}. Instead, we make use of the probabilistic shortest path length counter called the HyperANF algorithm [2]. This counter approximates the size of a user's neighborhood after a certain number of traversals using the HyperLogLog counter [4], which gives a probabilistic estimate of the number of unique items in a large stream. The number of shortest paths of length n through which a user is connected can be approximated as the change in her neighborhood size after the n^{th} jump. This technique was also used to analyze the Facebook graph [1]. For the HyperLogLog counter, we found that using registers of length 64 for each user (which leads to a relative standard deviation of accuracy of about 0.1325) to be sufficient for our purposes.

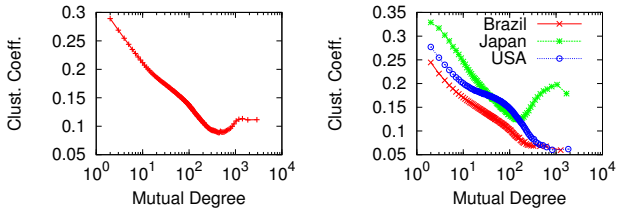
Figure 3 shows the distribution of both types of shortest path lengths in the Twitter follow graph. The average path length is 4.17 for the mutual graph and 4.05 for the directed graph. This is shorter than the social networks to which we compare (in both the directed and undirected cases): the average path length for Facebook is 4.74, and for MSN messenger, 6.6. The average degree of the MSN messenger graph is much lower than that of the mutual graph so perhaps this is less surprising, but the average degree of a vertex in the Facebook graph is higher than both the Twitter follow graph and mutual graph. Despite the fact that the Facebook graph has a higher branching factor (and a larger clustering coefficient—see below), users in the Twitter graph appear to be more closely connected. Note that although the Facebook graph contains more vertices, previous work suggests that in social networks the average path length should actually *decrease* with size [9]—this makes the short average path lengths in Twitter more surprising.

The *spid* (dispersion of the path length distribution) is another important graph metric. It is the ratio of the variance of the distribution to the mean of the distribution: social networks have a spid of less than one while web graphs have a spid greater than one. The spid of the mutual path length distribution is 0.115, and 0.108 for the directed path length distribution. These are well within the "social" range, but are slightly higher than that of the Facebook graph (0.09)—suggesting that the distribution for the Twitter graph is slightly "wider" than that of the Facebook graph.

The path length distribution of each country subgraph does not deviate much from that of the entire graph. The results are shown in Table 2. Brazil is the most tightly-connected subgraph with an average shortest path length of 3.78, and the US has the largest average shortest path length of 4.37. Note that these findings do not necessarily contradict the results of Leskovec et al. [9] (shrinking diam-

Graph	Avg. Path Length	spid
Twitter		
Follow Graph	4.05	0.12
Mutual Graph	4.17	0.11
Mutual BR	3.78	0.13
Mutual JP	3.89	0.16
Mutual US	4.37	0.18
Other Networks		
Facebook	4.74	0.09
MSN	6.6	-

Table 2: Summary of the average shortest path length distributions for the various graphs.



(a) Entire Graph (b) Selected Countries

Figure 4: The average clustering coefficient of users as a function of their mutual degree.

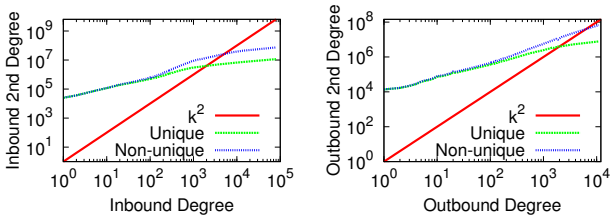
eters with increasing graph size) since there may be genuine connectivity differences between the country subgraphs.

Conclusion: Analysis of the shortest path lengths and *spid* shows that the Twitter follow graph exhibits properties that are consistent with a social network.

3.4 Clustering Coefficient

The clustering coefficient [15] in social networks measures the fraction of users whose friends are themselves friends—a high clustering coefficient is another property commonly attributed to social networks. Here, we examine the *local* clustering coefficient of vertices in the Twitter mutual graph. Figure 4(a) plots the average local clustering coefficient against the vertex degree. As expected, the local clustering coefficient decreases with increasing degree. While the clustering coefficient of the Twitter mutual graph is lower than that of the Facebook graph, it is still in the range that one would expect of a social network. Ugander et al. [14] found that the average clustering coefficient for degree = 5 is about 0.4 in the Facebook graph, compared to Twitter mutual graph's 0.23; for degree = 20, Facebook is 0.3 and the mutual graph is 0.19. Around degree = 100, the two graphs become very comparable with a coefficient around 0.14. Even though the clustering coefficient of the Twitter mutual graph is lower than that of Facebook's, it is still more tightly clustered than the MSN Messenger graph: for degree = 5 the mutual graph's clustering coefficient is more than 50% greater than that of the MSN graph, and for degree = 20 it is ∼90% greater.

Also of interest are the differences between the three countries shown in Figure 4(b). First, Japan has a higher clustering coefficient compared to the US or Brazil, and it is also higher overall compared to the entire globe. In Japan, the rate of reciprocity is much higher (i.e., if you follow someone, the chances that they follow you back is much higher)



(a) Inbound (b) Outbound

Figure 5: The size of users' inbound and outbound two-hop neighborhoods as a function of their degree.

and so the mutual graph has a higher edge to vertex ratio. What is even more interesting is that the clustering coefficient in the Japan subgraph begins to *increase* with degree at around a degree of 200 and peaks at a degree of 1,000. One possible explanation for this is the presence of massive cliques: members of these cliques would have high degrees as well as high clustering coefficients. The increase in clustering coefficients for large degrees in the overall graph, as in Figure 4(a), can be attributed to this idiosyncrasy in the Japan subgraph.

Conclusion: Analysis of clustering coefficients in the Twitter mutual graph suggests that Twitter exhibits characteristics that are consistent with a social network.

3.5 Two-Hop Neighborhoods

An important consideration in network analysis is a vertex's two-hop neighborhood, i.e., the set of vertices that are neighbors of a vertex's neighbors. Many algorithms, particularly edge prediction algorithms, use this set of vertices as the starting point for predicting the formation of new edges [6]. In the context of Twitter, the directed nature of the follow graph creates two such neighborhoods: the set of a vertex's followers' followers (inbound two-hop) and the set of a vertex's followings' followings (outbound two-hop).

Viewed from the perspective of an information network, the outbound two-hop neighborhood characterizes the "information gathering potential" of a particular user, in that any tweet or retweet from those vertices has the potential of reaching the user. Similarly, the inbound two-hop neighborhood characterizes the "information dissemination potential" of the user, in that any retweet of the user's tweet has the potential of reaching those vertices.

As expected, followers of a user's followers are often not unique (it is likely that at least one user follows two people that follow the original user), so we consider both the unique and non-unique two-hop neighborhoods. The non-unique two-hop inbound neighborhood is simply the sum of the inbound degrees of a user's followers. To approximate the size of the unique second degree neighborhoods, we used the HyperANF algorithm described earlier.

Comparisons between the unique and non-unique two-hop neighborhoods are informative. If they are close in value, then the number of edges between users within those neighborhoods is low. This would also imply that, for instance, gaining a new follower dramatically increases the size of the two-hop neighborhood. This would also mean that the neighborhoods do not exhibit community structure, which provides evidence against Twitter as a social network.

Figures 5(a) and 5(b) show the average size of a vertex's two-hop neighborhoods as a function of its degree. For both

the inbound and outbound variants, the average number of unique and non-unique neighbors in the two-hop neighborhood is plotted against the degree, along with k^2, where k is the degree of the user. If a user links exclusively to other users with the same degree (high degree assortativity—see below), then these (non-unique) neighborhoods would be exactly of size k^2. From these figures, we see that most users have two-hop neighborhoods greater than k^2; it is not until a user has an inbound/outbound degree of greater than around 3,000 that the neighborhood sizes are less than k^2.

The fact that the two-hop neighborhoods are usually much larger than would be predicted by the branching factor (both inbound and outbound) suggests that there is a type of amplification effect. From the average Twitter user's perspective, the structure of the graph is very efficient for both information gathering (outbound) and information dissemination (inbound).

What is also striking about these plots is that for inbound/outbound degree of less than around 100, the number of unique and non-unique two-hop neighbors is practically the same. For example, the first 100 or so followers that a Twitter user receives adds almost *all* of the new user's followers as unique secondary followers. More specifically, for users with an inbound degree of less than 100, each new follower a user receives adds on average 4770 new secondary followers. Similarly, each new follow a user makes connects her to 3573 new secondary followings.

It is worth comparing these results to the Facebook graph. A Facebook user with 100 friends typically has 27,500 unique friends-of-friends. This more than an order of magnitude less than the 497,000 (unique) followers of followers of a typical Twitter user with 100 followers, or the 367,000 (unique) followings of followings of a user who follows 100 other users.

Conclusion: Analysis of two-hop neighborhoods suggests that Twitter behaves efficiently as an information network since the graph structure exhibits a pronounced amplification effect for information dissemination and reception.

3.6 Degree Assortativity

Degree assortativity, sometimes called assortative mixing, is the preference for a graph's vertices to attach to others that are similar (or dissimilar) in degree. Typically, this is quantified in terms of the correlation of the vertex degrees on either side of each edge [11]. Newman and Park [12] argue that degree assortativity is a fundamental characteristic that separates social networks from all other types of large-scale networks. In most social networks, an assortative measure between 0.1 and 0.4 is typical; for example, the Facebook global network measures 0.226.

Most previous analyses of degree assortativity have looked at undirected graphs. In the Twitter case, it is appropriate to look at both in-degree and out-degree correlations. This requires us to extend the standard formulation, illustrated in Figure 6. To be precise, in our analysis "source" refers to the follower, and "destination" refers to the person being followed. We consider four cases associated with each edge: source in-degree (SID), source out-degree (SOD), destination in-degree (DID), and destination out-degree (DOD).

In looking at the Pearson correlations across these degree measures, we find almost no correlation—most likely due to the heavy tails of the distributions. A more informative measure is the correlation of the logarithm of the degrees. Specifically, if x and y are two different types of degrees

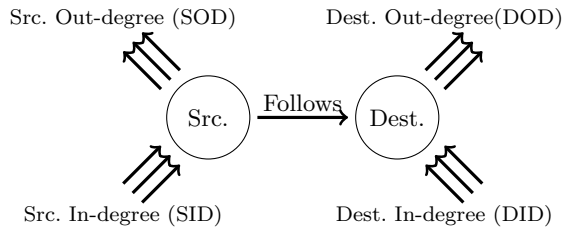


Figure 6: A diagram of the four different types of degrees between which we examine correlations.

associated with an edge, then we can measure the Pearson correlation coefficient between $\log(x+1)$ and $\log(y+1)$. This measures whether or not the degrees are correlated in *magnitude*. Below, we examine each of these correlations in turn (all of which are significant):

SOD vs. DOD has a positive correlation (0.272). This means that *the more people you follow, the more people that those people are likely to follow.* From a social network perspective, this makes sense: if we interpret "following" as social behavior, this correlation represents the type of assortativity we would observe in social networks—social users engage with other social users.

SID vs. DOD has a positive correlation (0.241). On Twitter, the number of followers is typically understood as a measure of popularity (or notoriety). This means that *the more popular you are, the people you follow will tend to follow more people.* This also appears to be consistent with social network theory: the more popular one becomes, the greater tendency one would engage with other socialable users (i.e., those who follow more people).

SOD vs. DID has a negative correlation (-0.118). This means that *the more people you follow, the less popular those people are likely to be.* This is highly unexpected, since the fact that the edge is present increases both the SOD and the DID by one, and this would suggest a positive correlation.

SID vs. DID has a negative correlation (-0.296). This means that *the more popular you are, the less popular the people you follow are.* In a social network, we would expect that popular people are friends with other popular people—in Twitter terms, we would expect popular users (i.e., with many followers) to follow other popular users. However, this is not the case, and stands in contrast with social network properties observed by others [11, 12].

Conclusion: Analysis of degree assortativity in the Twitter graph leads to conflicting and counter-intuitive results. In some ways, Twitter exhibits characteristics that are consistent with a social network; in other ways, the results are the opposite of what we observe in other social networks.

4. DISCUSSION

Quickly recapping the results: some analyses suggest that Twitter behaves more like an information network, but other analyses show that Twitter exhibits characteristics consistent with social networks. Beyond these descriptive characterizations, are there deeper insights that we can glean?

We are currently developing a model that explains these findings by taking into account the order in which edges in the Twitter follow graph are added. Our hypothesis is that from an individual user's perspective, Twitter starts

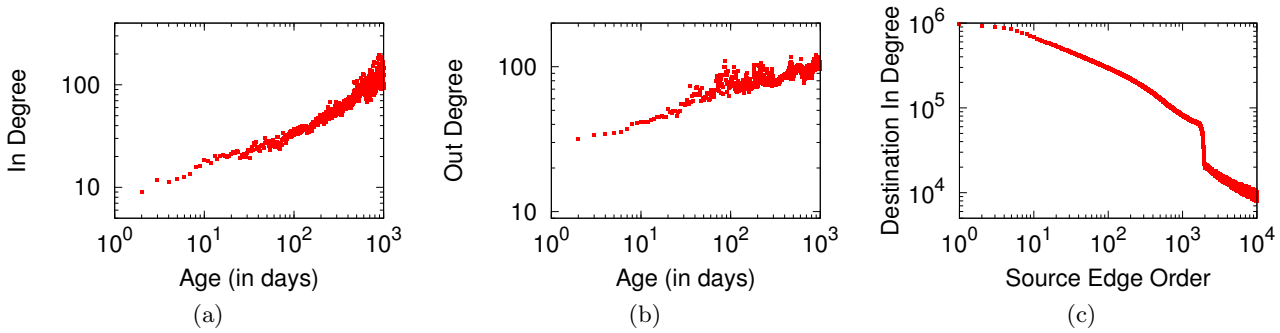


Figure 7: User age versus their in-degrees (a) and out-degrees (b). The source order of an edge versus the in-degree of the destination user (c).

off more like an information network, but evolves to behave more like a social network. Specifically, the first few accounts that a new user chooses to follow are likely high-profile, popular accounts with large inbound degrees (i.e., many followers), driven by preferential attachment—users with high in-degrees are more visible and are therefore more likely to receive new edges, further increasing their inbound degrees. However, as a user follows more people and becomes more "experienced" using Twitter, the preferential attachment effect diminishes. They become more selective and choose followings based on other criteria beyond popularity. The user typically discovers a community with which to engage—whether it be based on real-world social ties, common interests, or other factors—and Twitter starts behaving more like a social network.

One fact that sometimes gets lost in the analysis of a single graph snapshot is that the graph structure is constantly evolving. In particular, users accumulate followers and follow more people over time. This is shown in Figure 7(a) and (b), where in-degree and out-degrees are plotted against account age. In reality, our Twitter graph snapshot contains a mix of new users who recently just discovered Twitter and experienced users who have been active for a long time. Most revealing is Figure 7(c), where the x-axis shows the *source order* of the edge, or its ordering of when the source vertex added it, and the y-axis plots the average inbound degree of the destination vertex of that edge. For example, the 20^{th} account a user chooses to follow has an average in-degree of about 500,000, whereas the $1{,}000^{th}$ user has an average in-degree of about 70,000.

These graphs immediately explain many of the degree correlations we observe. SID and DOD are positively correlated because Twitter users tend to gain more followers and follow more accounts over time. The same explanation applies to SOD vs. DOD correlation as well. Figure 7(c) provides an explanation of why SOD and DID are inversely correlated—users who have lots of outbound edges are choosing followings who have fewer inbound edges. That is, over time, users' tendency to follow celebrities decreases (i.e., preferential attachment gives way to social ties).

5. FUTURE WORK AND CONCLUSIONS

In this paper, we present evidence that Twitter differs from previously-studied social networks in certain aspects, but it also demonstrates many social properties as well. Beyond descriptive characterizations that may be independently useful for the community, we have formulated a hypothesis that attempts to explain these findings and are developing a model to better formalize these ideas.

Ultimately, we believe that there are two major "modes" of behavior on Twitter: one that is based upon information consumption, and another that is based upon reciprocated social ties. The network structure we observe results from a mixture of the two, where the mix depends on the age of the account. However, to further analyze this mixture, we need to develop a more precise characterization of information networks and social networks. We provide only preliminary evidence for this hypothesis, but at an intuitive level, this hybrid structure seems to be plausible.

6. REFERENCES

[1] L. Backstrom, P. Boldi, M. Rosa, J. Ugander, and S. Vigna. Four degrees of separation. *WebSci 2012*.

[2] P. Boldi, M. Rosa, S. Vigna. HyperANF: approximating the neighborhood function of very large graphs on a budget. *WWW 2011*.

[3] R. Dunbar. Neocortex size as a constraint on group size in primates. *Journal of Human Evolution*, 1992.

[4] P. Flajolet, C. Fusy, O. Gandouet, and F. Meunier. HyperLogLog: the analysis of a near-optimal cardinality estimation algorithm *Analysis of Algorithms*, 2007.

[5] B. Goncalves, N. Perra, A. Vespignani. Modeling users' activity on Twitter networks: validation of Dunbar's number. *PLoS One*, 2011.

[6] P. Gupta, A. Goel, J. Lin, A. Sharma, D. Wang, and R. Zadeh. WTF: The Who to Follow service at Twitter. *WWW 2013*.

[7] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media? *WWW 2010*.

[8] J. Leskovec and Eric Horvitz. Planetary-scale views on a large instant-messaging network. *WWW 2008*.

[9] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. *KDD 2010*.

[10] J. Lin, D. Ryaboy. Scaling big data mining infrastructure: the Twitter experience. *SIGKDD Explorations*, 2012.

[11] M. Newman. Mixing patterns in networks. *Physical Review*, 2003.

[12] M. Newman and J. Park. Why social networks are different from other types of networks. *Physical Review*, 2003.

[13] X. Shi, L. A. Adamic, and M. Strauss. Networks of strong ties. *Physica A: Statistical Mechanics and its Applications*, 2007.

[14] J. Ugander, B. Karrer, L. Backstrom, and C. Marlow. The anatomy of the Facebook social graph. *arXiv*, 2011.

[15] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 1998.