Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media

Network Weirdness: Exploring the Origins of Network Paradoxes

Farshad Kooti
USC Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA 90292
kooti@usc.edu

Nathan O. Hodas
USC Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA 90292
nhodas@isi.edu

Kristina Lerman
USC Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA 90292
lerman@isi.edu

Abstract

Social networks have many counter-intuitive properties, including the "friendship paradox" that states, on average, your friends have more friends than you do. Recently, a variety of other paradoxes were demonstrated in online social networks. This paper explores the origins of these network paradoxes. Specifically, we ask whether they arise from mathematical properties of the networks or whether they have a behavioral origin. We show that sampling from heavy-tailed distributions always gives rise to a paradox in the mean, but not the median. We propose a strong form of network paradoxes, based on utilizing the median, and validate it empirically using data from two online social networks. Specifically, we show that for any user the majority of user's friends and followers have more friends, followers, etc. than the user, and that this cannot be explained by statistical properties of sampling. Next, we explore the behavioral origins of the paradoxes by using the shuffle test to remove correlations between node degrees and attributes. We find that paradoxes for the mean persist in the shuffled network, but not for the median. We demonstrate that strong paradoxes arise due to the assortativity of user attributes, including degree, and correlation between degree and attribute.

Introduction

The interplay between individual's attributes and her choice of who to link to in a social network produces much of the network's observed complexity and leads to counter-intuitive phenomena, such as the "friendship paradox". This paradox states that regardless of which individual you pick in a social network, on average, her friends have more friends than she does (Feld 1991). The paradox has been observed both in online (Ugander et al. 2011; Hodas, Kooti, and Lerman 2013) and offline social networks (Feld 1991; Zuckerman and Jost 2001; Eom and Jo 2014). Paradoxes for attributes other than the number of friends have also been observed. For example, a Twitter user posts fewer messages and receives less-viral content than her friends do on average (Hodas, Kooti, and Lerman 2013). Similarly, a scientist's co-authors are more productive and better cited on average (Eom and Jo 2014). Additionally, paradoxes can lead to systematic biases in how individuals perceive their world (Zuckerman and Jost 2001; Sgourev 2006; Wolfson 2000; Yoganarasimhan 2012; Kanai et al. 2012) and in how peers affect an individual's behavior. For instance, studies have shown that teenagers over-estimate alcohol and drug use of their peers, because the correlation between connectivity and drug/alcohol use allows popular drinkers and smokers to skew perceptions of their many friends (Tucker et al. 2011; Wolfson 2000). On a positive note, the friendship paradox has been used as a basis for early detection of flu outbreaks on a college campus (Christakis and Fowler 2010) and of trending topics on social media (Garcia-Herranz et al. 2012).

We examine the origins of these network paradoxes. Specifically, we ask whether the paradoxes are simply a consequence of the mathematical properties of networks and populations of users comprising them or whether they have a behavioral origin. First, we show that the conventional measure of paradox, which compares a user's attribute, e.g., number of neighbors, to the *mean* of the attributes of her network neighbors, will almost always produce a paradox, even if the network is completely random. This can be explained by the statistical properties of sampling from a heavy-tailed distribution, and many social attributes, such as degree, have such heavy-tailed distributions. Hence, it is not surprising to observe this sort of 'traditional paradox' in the mean on any social network for almost any attribute. We show that using the median, instead of the mean, is a more robust measure of "paradoxical behavior".

Next, we measure network paradoxes on social media sites Digg and Twitter. We confirm traditional paradoxes for several user attributes, including number of neighbors, activity, and also diversity and virality of content user sees. In addition, we find that paradoxes persist when users' attributes are compared to the *median*, rather than the mean, of the attributes of their neighbors. As a consequence, network paradoxes can be restated in a stronger form: For a variety of attributes, a majority of a user's neighbors have a value that exceeds the user's own value, i.e., not only do user's friends have more friends on average, but *a majority of friends have more friends than the user*. Because the strong paradox does not trivially arise from the statistical properties of distributions, it must have a behavioral origin.

To better understand behavioral causes of network paradoxes, we perform shuffle tests to destroy correlations between degree and user attributes. There exist two types of correlations in social networks. First, there is a correlation between attribute of a node, such as activity, and its degree (within-node correlation). In addition, there is assortativity, or correlation between the attributes of a node and the attributes of its neighbors (between-node correlation). Each of these correlations might explain the paradoxes. To understand the exact origin, we separate the effect of different correlations by destroying one of the correlations while keeping the other one, which is done using a controlled shuffle test. We find that both within-node and between-node correlations are each responsible for the paradoxes.

Existence of the strong paradox for an attribute implies that for most users, a randomly selected friend is likely to exceed the user in that attribute. Because the same behavioral factors that create between-node correlations in attributes and within-node correlations are often related to desirability of that attribute (extraversion, wealth, etc.), people end up dynamically positioning themselves in the network to remain subject to the strong paradox. The users will perceive themselves as inferior to friends, even if comparison is done on a one-to-one basis. This may explain why self-assessments are negatively correlated with exposure to online social media (Kross et al. 2013; Chou and Edge 2012).

Our specific contributions are as follows:

1. We characterize the differences between the traditional ('weak') paradox, which is based on the mean, and our strong paradox, based on the median. We empirically demonstrate that previously observed paradoxes still hold as strong paradoxes.
2. We show that upon shuffling the network, weak paradoxes persist but the strong paradoxes are diminished or destroyed, demonstrating that the magnitude of strong paradox can be used as a measure of behavioral factors in the creation and maintenance of the social network.
3. We also show that the observation of a weak paradox does not imply any within-node or between-node correlation between attributes and connectivity; they can arise in completely random networks devoid of correlation.

Our work suggests that network structure and node attributes in social media are intimately connected. Accounting for these often non-intuitive relationships is necessary for understanding, modeling, and predicting individual and social behavior in networks.

Statistical Origins of Paradoxes

The friendship paradox is thought to be rooted in the heterogeneous distribution of node attributes, such as node degree (Feld 1991). Such distributions are characterized by a "heavy tail", where extremely large values, e.g., high degree nodes, appear much more frequently than expected compared to a normal or exponential distribution. These large values skew the "average", giving rise to a large difference between the mean and the median. In this section, we show that randomly sampling from a heavy-tailed distribution can produce a paradox when using the mean, but not the median. Therefore, if the friendship paradox also exists in the median, it cannot be purely statistical in nature, and must have a behavioral origin. Conversely, a paradox using the mean may arise simply from statistics of the attribute distribution, without any behavioral component or correlation between users.

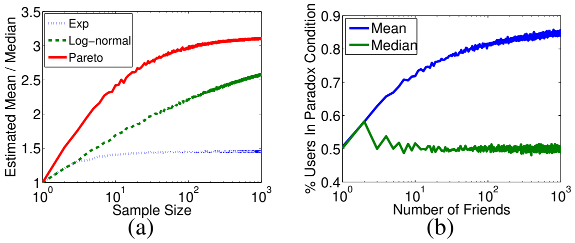


Figure 1: (a) Estimated mean grows with sample size. Three distributions, exponential (Exp), log-normal, and Pareto, each result in different biased estimates of the mean. The larger the difference between median and population mean, the larger the discrepancy. Thus, a user will always observe a paradox when calculating the mean of its neighbors when the population mean is greater than the median. (b) Effect of using mean vs. median on fraction of users with given number of friends estimated to be in paradox condition in a random network with no correlations. Users' attributes are drawn independently from $\sim x^{-1.2}$.

To be more precise, consider the definition of mean vs median for continuous, non-negative distributions[1]. The mean is defined as $\mu \equiv \int_0^\infty xP(x)\,dx$. The median, m, is defined by the solution to $\frac{1}{2} = \int_0^m P(x)\,dx$. Given a sample consisting of a single random instance drawn from $P(x)$, there is an equal chance that it will be larger or smaller than m. What value would you estimate that minimizes the mean absolute error with the mean of a sample with size = 1? This is the median of the distribution, m (Lee 1995). At the other extreme, the value that minimizes the absolute error with the mean of an infinitely large sample is μ, by definition. Social behavior is often characterized by unimodal distributions with heavy tails, so $m \leq \mu$. Thus, as the sample size from a distribution increases, the observed mean of the sample increases monotonically from m to μ.

Figure 1(a) provides an illustration of this behavior. We randomly sample values from three example distributions, exponential ($\sim e^{-2x}$), log-normal ($\mu = -0.3, \sigma = 1.5$), and Pareto ($\sim x^{-1.2}$), and calculate the mean and median of the samples of varying sizes. The true mean and median for these distributions are ($\mu = \frac{1}{2}, m = \frac{1}{2}\log 2$), ($\mu = 2.28, m = 0.74$), and ($\mu = 6, m = 1.78$), respectively. As explained above, the estimated sample mean changes monotonically from m to μ as the sample size increases from 1, shown in Fig. 1(a). However, the sample

[1] The conclusions in the present work hold for any distribution where the mean is greater than the median. This is almost always the case for heavy-tailed distributions, but may be violated for small-support discrete distributions. When median > mean, the present conclusions are simply reversed; we would expect 'anti-paradoxes' when considering the mean.

median does not vary with sample size, because half of the numbers in the sample are below the population median and half above. Thus, if you consider the fraction of users in paradox condition, shown in Figure 1(b), when users attributes are drawn iid from the previous Pareto distribution, the more friends a user has, the more likely the mean of their friends' attributes exceeds their own, but when using the median, no paradox is observed.

Thus, consider the following explanation of how network paradoxes arise and how they depend on what you mean by "average." Assume we are measuring some empirical quantity — user attribute x — in a purely random network where each user's x is an independent, identically distributed (iid) variable (and x is not node degree). The x for a user is compared to the "average" x for the user's network neighbors. The best estimate (with respect to mean absolute error) of x for the user is m_x, because it is a sample of size 1. On the other hand, the number of users' neighbors is at least one, meaning that the best estimate for the mean of the neighbors' x is $\geq m_x$. Therefore, even in a purely random network, as long as the mean of x is greater than the median of x, one would be led to the conclusion that "your x is smaller than the mean of your neighbors' x." But, if you consider the *median*, both you and your neighbors will have the same median, and no paradox will be observed. One may show that in a fully-connected network, where attributes are iid, the fraction of users in the strong paradox condition is no larger than $0.5 + 1/N$, but the weak paradox condition may hold for as many as $N - 1$ of the users.

Observation of paradoxes utilizing the median provides a test of the origin of such paradoxes: do they arise simply from heavy-tailed distributions or are they due to humans positioning themselves in the network according to some nontrivial behavioral mechanism?

Measuring Network Paradoxes

In the previous section, we showed that a network paradox could always exist when considering the mean of some user attribute but not necessarily the median. In this section, we analyze online social networks of Twitter and Digg users and show that network paradoxes exist for several user attributes when considering both the mean and median. Therefore, they cannot arise simply due to properties of heavy-tailed distribution of user attributes.

Data

We use the Twitter dataset gathered by (Yang and Leskovec 2011). The dataset includes 476M tweets, which are 20-30% of all tweets posted between June and December 2009. We complement this dataset with the social network of Twitter as of August 2009 (Kwak et al. 2010). To guarantee that we have information about all user's connections and activity data, we only use the first two months of tweets, because we do not have the follow links that were established later. We also only consider the links that are between users tweeted at least once during the observation period. The remaining network includes 5.8M users with 193.9M follow links. This network was used to measure friendship paradoxes.

For measuring other network paradoxes, we further restrict the network to users who joined Twitter before the beginning of the dataset. These are the users whose activity is recorded over the entire observation period. This leaves us with 29.4M tweets, with 2.2M users and 113M links.

In addition, we also use the Digg dataset presented by (Sharara, Rand, and Getoor 2011). Digg is distinct from Twitter because it is primarily a social news site. Despite this, it is similar to Twitter in many ways. Digg users submit links to news stories they find online, which is similar to posting a tweet. Other users vote for, or digg, these stories, which is similar to retweeting. Moreover, Digg allowed users to follow submissions and diggs of other users, creating a friend-follower network similar to Twitter. However, unlike Twitter, Digg promoted popular news stories to the front page, where they could be seen by all Digg users. To reduce potentially confounding factors, we focus only on pre-promotion votes, when the stories were mainly visible through the social network (Hodas and Lerman 2013).

The Digg dataset includes all diggs on news stories of 11.9K users submitted to Digg over a six months period of July to December 2010. There are more than 1.9M diggs on 29K stories in the dataset, with 1.3M follow links.

For convenience, we refer to the people followed by a user U as U's *friends* and those who follow her as U's *followers*. Thus, U receives content from her friends in the form of messages on Twitter or recommendations for news stories on Digg, and sends content to her followers.

Distribution of Attributes

Social networks share some common structural properties, such as a heavy-tailed degree distributions. Most of the nodes in such networks have few connections, or small degree, while few nodes have a large number of connections, or high degree. Besides degree, many other attributes of Twitter (Digg) users have a heavy-tailed distribution. In this paper we focus on the following attributes:

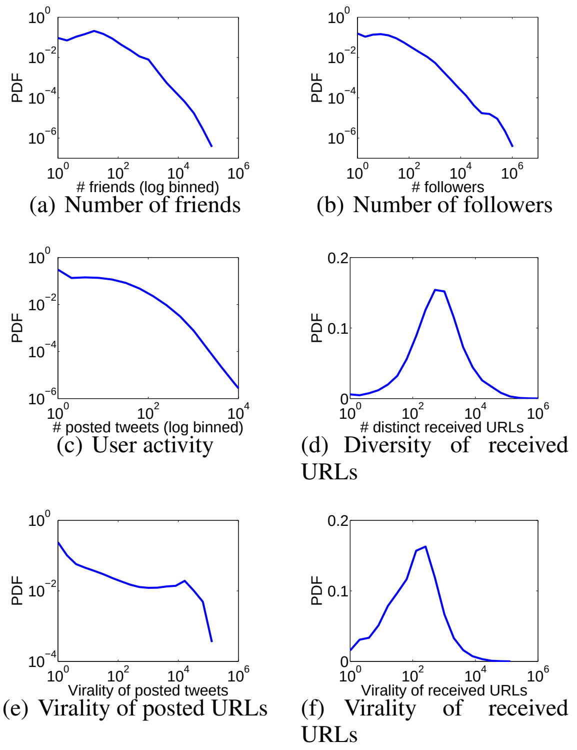
Degree: number of friends and followers of a user

Activity: number of tweets (diggs) made during the observation period by the user

Diversity: number of distinct URLs (or news stories on Digg) received by user from friends

Virality: popularity of content posted or received by the user, as measured by the number of retweets (or pre-promotion votes on Digg) it receives

Figure 2 shows the probability distribution of the observed values of each attribute, i.e., fraction of users in our sample who have that attribute value on Twitter. We logarithmically bin the values to reduce sparseness of extreme values. The distributions are characteristically heavy-tailed, regardless of whether the attribute depends on the decisions made by the user (number of friends, activity) or the decisions of others (number of followers, diversity). Except for diversity (Fig. 2(d)) and virality of received content (Fig. 2(f)), which resemble log-normals, the user attribute distributions have a power law-like shape. The results for Digg are similar,



(a) Number of friends (b) Number of followers

(c) User activity (d) Diversity of received URLs

(e) Virality of posted URLs (f) Virality of received URLs

Figure 2: Distribution of user attributes on Twitter.

except for the virality of Digg posts, which is a not heavy-tailed, due to the fact that posts accumulate only a limited number of diggs before the promotion.

Besides heavy-tailed attribute distributions, networks have other important statistical regularities. For example, many social networks are assortative, meaning that nodes tend to connect to other nodes having a similar degree (Newman 2002). In addition to degree *assortativity*, other correlations may exist between attributes of connected nodes. This phenomenon, known as *homophily*, is a generic property of social networks and results in connected users being similar (McPherson, Smith-Lovin, and Cook 2001) and becoming more similar over time (Kossinets and Watts 2009). In addition to between-node correlations, *within-node* correlations are important in social networks. The most important of these is the correlation of user's attributes, such as income, activity, or productivity, with her degree (Hodas, Kooti, and Lerman 2013; Eom and Jo 2014). Later in the paper we study the role of these correlations in explaining network paradoxes.

Network Paradoxes

When measuring the paradox for some attribute x, we consider a node to be in the paradox regime if the average of the neighbors' values of x is larger than the node's own value. We state that a paradox exists for that attribute if most of the nodes are in the paradox regime.

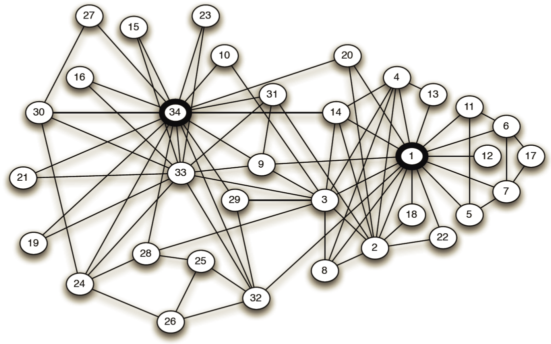
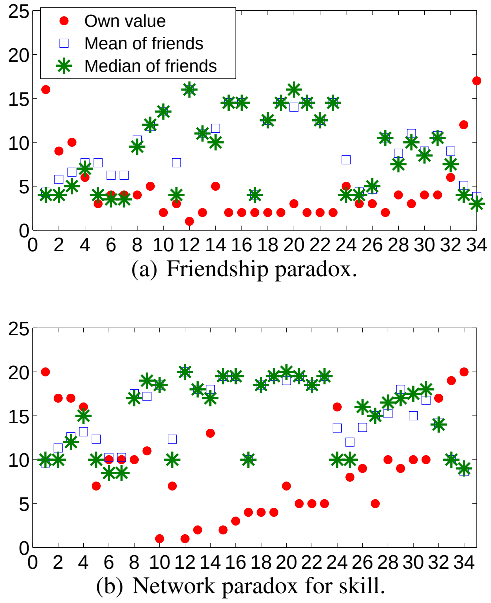


Figure 3: Friendship network of Karate Club



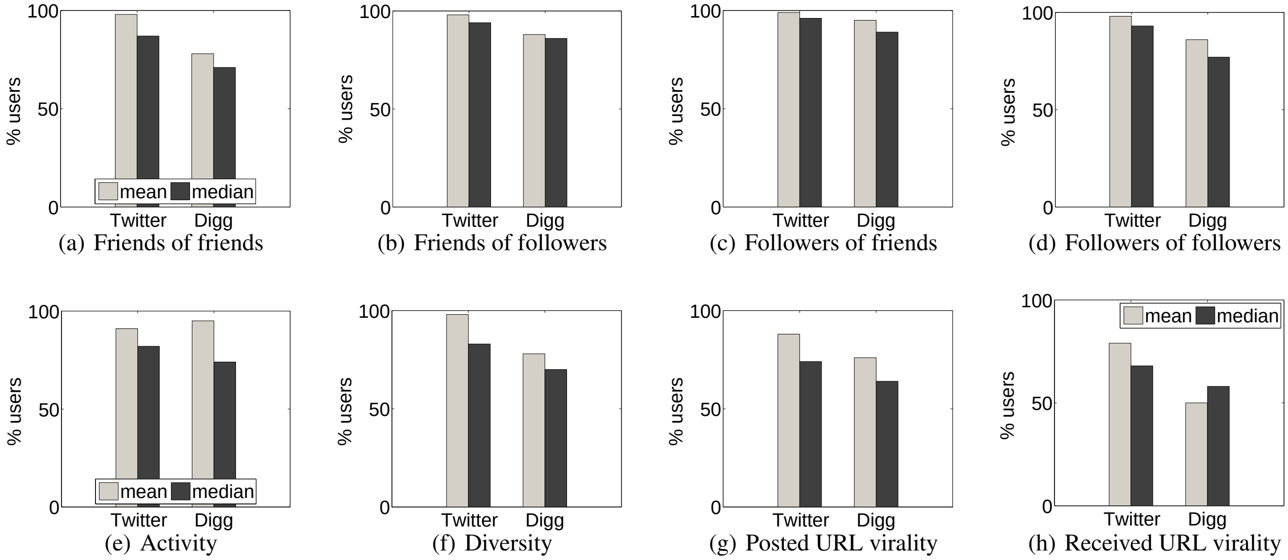
(a) Friendship paradox.

(b) Network paradox for skill.

Figure 4: Network paradoxes in the Karate Club network. For most individuals, the mean (or median) of the friends' (a) degree and (b) skill level is larger than the individual's own value.

We demonstrate the network paradoxes in the Karate Club, a small benchmark social network (Zachary 1977). The network, shown in Figure 3, includes 34 individuals and the friendship links between them. There are two nodes, 1 and 34, with a high degree, whereas most of the other nodes have just a few connections. We calculate the degree of each node along with mean and median degree of the friends, to demonstrate the friendship paradox. We observe that for 29 out 34 nodes (85%), the friends have a higher mean degree, and 26 of the nodes (76%) friends also have a higher median degree (Figure 4(a)). Therefore, both weak and strong friendship paradoxes hold for most nodes in this network.

We also consider network paradoxes for node attributes. In the Karate Club data, there are no specific attributes assigned to the individuals, but to demonstrate the network paradoxes, we assign a hypothetical skill level to each indi-



(a) Friends of friends (b) Friends of followers (c) Followers of friends (d) Followers of followers

(e) Activity (f) Diversity (g) Posted URL virality (h) Received URL virality

Figure 5: Demonstration of network paradoxes on Twitter and Digg. Each figure shows the percentage of users in the paradox regime, i.e., when the friends's mean (median) attribute value is larger than the user's own attribute value.

vidual. We assume that the skill level is correlated with individual's degree. We create skill levels by generating random numbers from a uniform distribution with range of 1-20, and assign the largest number to the node with the largest degree and so on. Similar to the friendship paradox, an attribute paradox exists in the network: for most individuals, their friends have a higher mean (and median) skill level than the individual herself (Figure 4(b)).

Several studies have confirmed that friendship and other network paradoxes exist in a variety of social networks, including Twitter (Garcia-Herranz et al. 2012; Hodas, Kooti, and Lerman 2013) and scientific collaboration networks (Eom and Jo 2014). These paradoxes were measured for the *mean* value of the friends' attributes. Surprisingly, these paradoxes also exist for the *median*.

First, we consider the friendship paradox. Since both Twitter and Digg are directed networks, the friendship paradox is manifested in four different ways (Garcia-Herranz et al. 2012; Hodas, Kooti, and Lerman 2013): (*i*) On average, your friends have more friends than you do. (*ii*) On average your friends have more followers than you do. (*iii*) On average, your followers have more friends than you do. Finally, (*iv*) on average, your followers have more followers than you do. Figures 5(a)–5(d) show the percentage of users in the paradox regime in each case. The paradox holds for almost all users when considering both the mean and the median, although the fraction of users in the paradox regime is slightly lower for the median. Thus, on both Digg and Twitter, a stronger statement of the friendship paradox holds.

Strong friendship paradox: *The majority of your friends (and followers) have more friends (and followers) than you do.*

Next, we consider network paradoxes for other user attributes besides degree. The activity paradox (Hodas, Kooti, and Lerman 2013) compares the number of tweets posted by a user on Twitter (or votes on Digg) with the average number of tweets (votes) made by her friends. Again, the paradox holds for a large fraction of users, both for the mean and the median (Figure 5(e)). We also observe network paradoxes for the virality of posted (Fig. 5(g)) and received (Fig. 5(h)) content and the diversity (Fig. 5(f)) of the received content on Twitter and Digg. The paradoxes exist regardless of whether the mean or the median is used to measure friends' values. This leads us to restate the paradoxes in their stronger form.

Strong activity paradox: *The majority of your friends are more active than you are.*

Strong diversity paradox: *The majority of your friends receive more diverse content than you do.*

Strong virality paradox: *The majority of your friends send and receive more viral content than you do.*

Remarkably, the paradoxes exist for attributes that are beyond user's direct control. Specifically, both diversity and virality of received content depend on the decisions friends make about the content they post. Somehow users position themselves in the network in a way that leads to a paradox: for most of the users in the network a randomly picked friend of the user is highly likely to be better connected, more active, and receive better content. This suggests that if your goal is exposure to interesting and novel content, a promising strategy for identifying new people to follow on Twitter would be to pick a random friend of a user you meant to follow, rather than following that user.

As explained earlier in this paper, it is not surprising to observe a paradox in the mean for an attribute with a heavy-tailed distribution. Any attribute with such a distribution will manifest the paradox, even in the absence of underlying behavioral factors. As an example, consider the fraction of posted URLs that point to YouTube videos. We do not expect this attribute to be "paradoxical", since it is unlikely that users selectively link to others who post a higher fraction of YouTube videos. However, 98% of users appear to be in the paradox regime when the mean is used. Using the median, on the other hand, puts only 43% of users in the paradox regime; hence, there is no true paradox – no surprising behavior. The apparent "paradox" only exists because the underlying distribution has a mean greater than the median.

Network paradoxes are not a manifestation of Simpson's paradox. Simpson's paradox refers to a phenomenon when a specific trend is observed within different sub-groups of data, but the trend does not appear if these groups are combined. For example, the mean within sub-groups could be above a threshold, but the aggregate mean over all groups is below the threshold. Simpson's paradox arises from mixing of heterogeneous populations, but in our case we are not grouping users based on any attribute. We simply check each individual to see whether she is in a paradox regime and report the fraction of individuals in the paradox regime.

Behavioral Origins of Paradoxes

We test two potential sources of the behavioral mechanisms. The first source is the correlation between a user's degree and her own attributes. We call this "within-node correlation". We use Pearson's Correlation Coefficient to measure the within-node correlation between its number of friends and its attribute, as defined in (Eom and Jo 2014). We use the number of friends as the degree and not the number of followers, because only the former characteristic is under user control. The second potential source of the paradoxes is the correlation between an attribute of the node and the attributes of its neighbors. This correlation is at the link level, and we call it "between-node correlation". We use assortativity to measure this correlation. Table 1 reports the empirical values (*Emp.*) of assortativity and correlation for a variety of attributes in the Twitter and Digg networks. Note that the follower graphs of Twitter and Digg have a slight degree disassortativity, as found by a previous study of Twitter (Kwak et al. 2010). Other attributes, on the other hand, are somewhat assortative. Within-node correlations are higher, as observed also in co-authorship networks (Eom and Jo 2014).

We use the shuffle test to probe the behavioral explanation of the paradoxes. The shuffle test randomizes node attribute values, destroying the within-node and/or between-node correlations. We then measure network paradox for the attribute in the shuffled network. If the paradox disappears, because most of the users are no longer in the paradox regime, we conclude that the correlation is the root cause of the paradox. First, we start by destroying both correlations and observe that the strong form of the paradoxes disappear, so the paradoxes are caused by these correlations. Then, we do a controlled shuffle to differentiate between within-node and between-node correlations.

Assortativity of the attribute						
Attribute	**Twitter**			**Digg**		
	Emp.	*Contr.*	*Shuffle*	*Emp.*	*Contr.*	*Shuffle*
num. friends	0.015	—	0.000	-0.040	—	0.001
num. followers	-0.047	—	0.000	-0.157	—	0.001
activity	0.037	0.016	0.000	0.152	0.005	0.000
diversity	0.055	0.012	0.000	-0.041	0.022	-0.001
posted virality	0.030	0.000	0.000	0.061	0.000	0.000
received virality	0.191	0.001	0.000	0.105	0.010	0.000

Correlation of node's attribute with num. friends						
Attribute	**Twitter**			**Digg**		
	Emp.	*Contr.*	*Shuffle*	*Emp.*	*Contr.*	*Shuffle*
activity	0.191	0.138	-0.001	0.097	0.108	-0.002
diversity	0.895	0.867	0.001	0.999	0.690	0.005
posted virality	-0.001	-0.001	0.000	-0.019	0.040	0.003
received virality	0.000	0.000	0.000	0.287	0.281	0.001

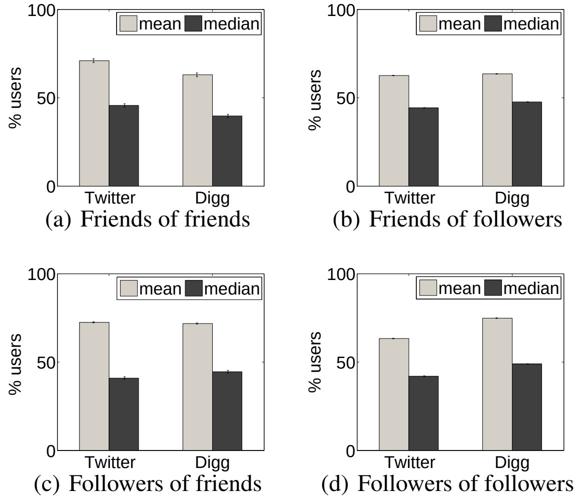
Table 1: Network properties. (*Top*) Assortativity of attributes of connected users and (*Bottom*) within-node correlations of the attribute with degree in the empirical data (*Emp.*) and in the shuffled networks after a controlled (*Contr.*) and full shuffle (*Shuffle*) of attributes.

Shuffle Test

We start by examining the number of friends attribute. As explained earlier, in directed networks there are four variants of the friendship paradox, which compare the number of friends or followers a user has with the average number of friends or followers of her (*i*) friends and (*ii*) followers. We shuffle the network to destroy the correlation between connected nodes as follows. We keep the links between users as is, preserving network structure, but assign a new number of friends to each user, which is randomly drawn from another network node. Note that "number of friends" is treated as an attribute of a node, unrelated to its degree. While this may be a non-conventional application of the shuffle test, we use it to probe how correlations affect the paradoxes. Shuffling the number of friends eliminates any correlation between the number of friends of connected users, but does not change its distribution. Table 1 (column *Shuffle*) shows that in all cases degree assortativity disappears in the shuffled network. Figures 6(a) and 6(b) show that the friendship paradox still holds in the shuffled network (though weaker) for the mean. However, the paradox no longer holds for the median, since fewer than 50% of users are in the paradox regime in the shuffled network.

Next, we consider the paradoxes involving the number of followers. We shuffle the number of followers by assigning to each user the number of followers from a randomly drawn user. The two paradoxes still hold for more than 60% of users for the mean, but only about 50% of users are in the paradox regime for the median on Twitter and Digg (Figs. 6(c)–6(d)).

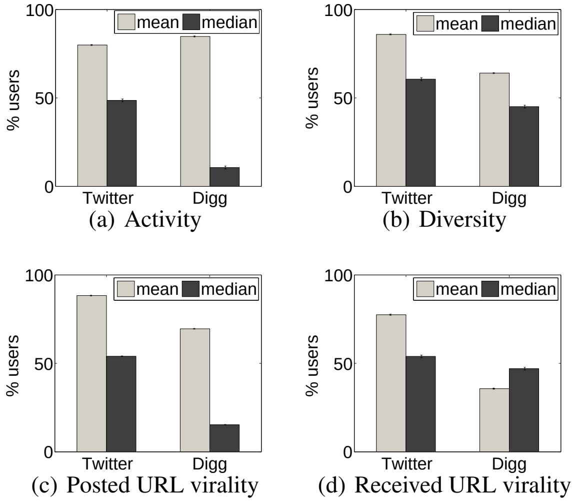
The empirically observed degree dissassortativity is an outcome of the mechanisms people use to select who to follow in online social networks. Disassortativity appears in a network where below-average users follow above-average users. This seems to be the case on Twitter (and Digg) where large fraction of follow links are from normal users to the celebrities (or top users on Digg) with orders of magnitude more followers. Friendship paradoxes in the online social networks of Digg and to some extent Twitter appear to be



(a) Friends of friends (b) Friends of followers

(c) Followers of friends (d) Followers of followers

Figure 6: Percentage of users in paradox regime on Twitter and Digg after shuffling the number of friends (top row) and the number of followers (bottom row). Error bars show the 0.95 confidence interval.



(a) Activity (b) Diversity

(c) Posted URL virality (d) Received URL virality

Figure 7: Percentage of users in paradox regime on Twitter and Digg after shuffling user attribute. Error bars show the 0.95 confidence interval.

related to degree disassortativity.

The remaining paradoxes are similar because each compares a user's attribute with the average value of this attribute among her friends. In each shuffle test, we shuffle the values of the attribute among all users. This eliminates both within-node and between-node correlations. Table 1 (column *Shuffle*) shows that none of the correlations exist in the shuffled network.

Figure 7 measures network paradoxes for the four attributes in the shuffled Twitter and Digg networks. In almost all cases, the paradoxes still hold for the mean. The only exception is the received virality paradox on Digg, which does not hold because virality of the stories does not have a heavy-tailed distribution on Digg, as mentioned earlier. When comparing user with friends using the median, the paradoxes mostly disappear. One ambiguous case is content diversity paradox on Twitter, which has 60% of the users in the paradox regime, a small statistically significant paradox.

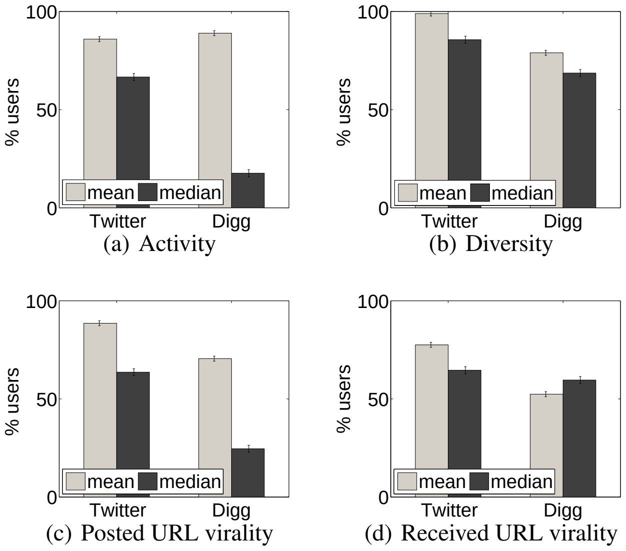
We conclude that the empirical observations of the paradox for the median cannot be explained purely by statistical sampling and imply a socio-behavioral dimension. The origin of these strong network paradoxes appears to be in the within- and between-node correlations.

Controlled Shuffle Test

Eom and Jo (Eom and Jo 2014) argued that within-node correlation between attribute and degree results in the observed paradox in the mean for the attribute. Unfortunately, the shuffle test does not allow us to distinguish whether within-node or between-node correlation (assortativity) is responsible for the paradox. In this section, we disentangle these effects through a controlled shuffle test, which attempts to eliminate the between-node correlation while preserving within-node correlation. We achieve this by grouping together users with the same degree (number of friends) and shuffling the attribute values within each group. Thus, the reassigned attribute is still correlated with degree, because it's from another user with the same degree. We log-bin the data to deal with degree sparseness at high values. Table 1 shows that the within-node correlation has not changed significantly, but the between-node correlation is reduced in the shuffled network (column *Contr.*).

Figure 8 shows the result of the controlled shuffle test, which should be compared with empirical data in Fig. 5. No single type of correlation is responsible for all paradoxes. The activity paradox is greatly reduced by controlled shuffle of the Twitter network and disappears on Digg, suggesting that between-node correlation (here activity assortativity) is largely responsible for this paradox. This means that the paradox arises because active users preferentially link to other active users. The posted-URL virality paradoxes are similar in that it is largely reduced by controlled shuffling. We conclude that it mostly arises due to assortativity of this attribute – not within-node correlation. The diversity paradox, however, appears to be unaffected by controlled shuffling both on Digg and Twitter. This suggests that the diversity paradox is not caused by assortativity of diversity. Instead, it is due to within-node degree–diversity correlation. The received-URL virality paradox is similar to diversity in that it is largely unaffected by controlled shuffling. Hence, we conclude that degree–virality correlation plays a key role in creating the paradox, but it is not simply due to users selectively seeking out interesting users.

There are a few plausible behavioral mechanisms that could lead to these correlations. First, some of the correlations arise simply from the nature of the attribute. For example the within-node correlation of number of friends and diversity (as measured by number of distinct URLs) exists because as users add more friends, they will eventu-



(a) Activity

(b) Diversity

(c) Posted URL virality

(d) Received URL virality

Figure 8: Percentage of users in paradox regime for the shuffled attribute, but keeping attribute-connectivity correlation (controlled shuffling).

ally begin connecting with users outside of their immediate interests. Between-node correlations arise as users position themselves near people with certain characteristics. For example, active users are generally more engaged with the social networking site, consuming more information. To increase the amount of content they receive, they could seek out new friends (degree–activity correlation) or seek out more active friends (activity assortativity). Regardless of which correlation may be indicated as the source of the strong paradox for a particular attribute, the unifying theme is that they arise from decisions made by the users and not from statistical artifacts of averaging.

Related Work

The friendship paradox and other paradoxes have been shown to occur in a variety of contexts, both online and offline. As many have pointed out, starting with Feld (Feld 1991), these network paradoxes arise from users with high degree being overrepresented in the population of friends. Feld also claimed that the friendship paradox, that your friends have higher degree than you, holds with both the mean and the median. In this paper we build on existing works which have documented paradoxes using the mean, including (Garcia-Herranz et al. 2012; Christakis and Fowler 2010; Zuckerman and Jost 2001; Ugander et al. 2011; Eom and Jo 2014; Hodas, Kooti, and Lerman 2013).

Hodas et al. demonstrated a variety of paradoxes on Twitter beyond the friendship paradox (Hodas, Kooti, and Lerman 2013). They showed that users' friends and followers are more active and more highly connected. They claimed that any user attribute correlated with connectivity will ultimately result in a paradox. They don't establish if the activity or virality paradox results simply from the nature of heavy-tailed distributions or from how users choose to position themselves in the network, i.e. behavioral factors. On the other hand, Hodas et al. identify some key paradoxes that cannot be explained simply by correlations between degree and activity, such as how overloaded users receive more viral content than underloaded users.

Eom and Jo recently examined 'generalized friendship paradoxes' in coauthorship networks (Eom and Jo 2014). They showed that correlation between degree and any user characteristic may ultimately lead to a paradox in the mean between the average neighbor and the average user. Although they do not explore the differences between median and mean, and thus do not surprisingly find numerous new paradoxes, they identify some qualities of authors correlated with their degree, including citations per publication and number of publications. In short, they find that your coauthors are more prolific and highly cited than you are.

The present paper takes a closer look at the origin of various paradoxes on social networks and tests how they depend on the statistical methods employed. We reveal essential differences between utilizing the mean versus the median. For example, friendship paradoxes on shuffled networks disappear when using the median, revealing it is not simply due to statistical overrepresentation.

Conclusion

A network paradox exists when one expects the value of some user attribute (degree, activity, etc.) to be less than the average value of this attribute among her neighbors. Such paradoxes have been observed in many social networks for a variety of attributes. Although there exist explanations for the traditional friendship paradox in undirected networks, in this work we proposed causes of these paradoxes originating in correlations between user attributes the choices users make to correlate themselves with their neighbors. We showed that when the attribute has a heavy-tailed distribution, as is often the case for social attributes, the paradox always exists when mean is used to measure the average, regardless of the underlying system, as long as the mean of the attribute is larger than the median. However, utilizing the median does not inevitably lead to a paradox. Surprisingly, most of the paradoxes observed in online social networks still hold when sample median is used. This allows us to restate network paradoxes in their strong form: the *majority* of your friends are better connected, more active, and exposed to more viral and diverse content than you.

We probed the behavioral origins of the paradoxes using the shuffle test. This test eliminates correlations between node degrees and attributes by shuffling attribute values of nodes. We show that the friendship paradox with the median disappears on Digg and Twitter after shuffling node degrees, suggesting the correlation of directed node degrees gives rise to the strong friendship paradox on directed networks. Shuffling other user attributes has a similar effect, leading us to conclude that the between-node correlation between attributes of connected nodes and the within-node correlation between user degree and attribute produce network paradoxes for different attributes. Furthermore, we conducted a controlled shuffle test to distinguish the effects of within-node and between-node correlations and we found that for activity and virality of posted content the between-node correlation has a major role, whereas for diversity and virality of received content within-node correlation is the key factor. More research remains to be done to clarify how attribute assortativity is related to within-node correlation, how these correlations affect behavior, and whether they are both caused by some other network factors.

Existence of the strong paradox for an attribute implies that, for most users, a randomly selected friend is likely to exceed the user in that attribute. Because the same behavioral factors that create between-node correlations in attributes and within-node correlations are often related to desirability of that attribute (extraversion, wealth, etc.), as people attempt to maximize that attribute, they end up dynamically positioning themselves in the network to remain subject to the strong paradox.

Our findings shed light on the causes of network paradoxes in social networks, and these paradoxes have implications in network sensing, early detection of outbreaks, and users' perceptions of their world. The present work suggests that models of network formation which cannot reproduce observed paradoxes are not successfully capturing behavioral factors that cause users to correlate themselves with their neighbors.

Acknowledgements

This material is based upon work supported in part by the Air Force Office of Scientific Research under Contract Nos. FA9550-10-1-0569, by the Air Force Research Laboratories under contract FA8750-12-2-0186, by the National Science Foundation under Grant No. CIF-1217605, and by DARPA under Contract No. W911NF-12-1-0034.

References

Chou, H.-T. G., and Edge, N. 2012. "they are happier and having better lives than i am": the impact of using facebook on perceptions of others' lives. *Cyberpsychology, Behavior, and Social Networking* 15(2):117–121.

Christakis, N. A., and Fowler, J. H. 2010. Social network sensors for early detection of contagious outbreaks. *PLoS ONE* 5(9):e12948+.

Eom, Y.-H., and Jo, H.-H. 2014. Generalized friendship paradox in complex networks.

Feld, S. L. 1991. Why Your Friends Have More Friends Than You Do. *American J. Sociology* 96(6):1464–1477.

Garcia-Herranz, M.; Egido, E.; Cebrian, M.; Christakis, N.; and Fowler, J. 2012. Using friends as sensors to detect global-scale contagious outbreaks. *arXiv preprint arXiv:1211.6512*.

Hodas, N. O., and Lerman, K. 2013. Attention and visibility in an information-rich world. In *Multimedia and Expo Workshops (ICMEW)*, 1–6. IEEE.

Hodas, N.; Kooti, F.; and Lerman, K. 2013. Friendship paradox redux: Your friends are more interesting than you. In *ICWSM*.

Kanai, R.; Bahrami, B.; Duchaine, B.; Janik, A.; Banissy, M.; and Rees, G. 2012. Brain structure links loneliness to social perception. *Current Biology*.

Kossinets, G., and Watts, D. J. 2009. Origins of homophily in an evolving social network. *The American Journal of Sociology* 115(2).

Kross, E.; Verduyn, P.; Demiralp, E.; Park, J.; Lee, D. S.; Lin, N.; Shablack, H.; Jonides, J.; and Ybarra, O. 2013. Facebook use predicts declines in subjective well-being in young adults. *PLoS ONE* 8(8):e69841.

Kwak, H.; Lee, C.; Park, H.; and Moon, S. 2010a. What is Twitter, a social network or a news media? In *WWW*, 591–600.

Lee, Y.-S. 1995. Graphical demonstration of an optimality property of the median. *The American Statistician* 49(4):369–372.

McPherson, M.; Smith-Lovin, L.; and Cook, J. M. 2001. Birds of a feather: Homophily in social networks. *Annual Review of Sociology* 27:415–444.

Newman, M. E. J. 2002. Assortative mixing in networks. *Phys. Rev. Lett.* 89:208701.

Sgourev, S. 2006. Lake wobegon upside down: the paradox of status-devaluation. *Social forces* 84(3):1497–1519.

Sharara, H.; Rand, W.; and Getoor, L. 2011. Differential adaptive diffusion: Understanding diversity and learning whom to trust in viral marketing. In *ICWSM*.

Tucker, J.; Green, H.; Zhou, A.; Miles, J.; Shih, R.; and Dï£¡Amico, E. 2011. Substance use among middle school students: Associations with self-rated and peer-nominated popularity. *J. adolescence* 34(3):513–519.

Ugander, J.; Karrer, B.; Backstrom, L.; and Marlow, C. 2011. The Anatomy of the Facebook Social Graph.

Wolfson, S. 2000. Students' estimates of the prevalence of drug use: Evidence for a false consensus effect. *Psychology of Addictive Behaviors* 14(3):295.

Yang, J., and Leskovec, J. 2011. Patterns of temporal variation in online media. In *WSDM*, 177–186.

Yoganarasimhan, H. 2012. Impact of social network structure on content propagation: A study using youtube data. *Quantitative Marketing and Economics* 10(1):111–150.

Zachary, W. 1977. An information flow modelfor conflict and fission in small groups1. *J. Anthropological Research* 33(4):452–473.

Zuckerman, E., and Jost, J. 2001. What makes you think you're so popular? self-evaluation maintenance and the subjective side of the" friendship paradox". *Social Psychology Quarterly* 207–223.