Predicting Reciprocity in Social Networks

Anonymous
Anonymous Department
Anonymous Institution
Anonymous Location
anon@whoare.you

Anonymous
Anonymous Department
Anonymous Institution
Anonymous Location
anon@whoare.you

Anonymous
Anonymous Department
Anonymous Institution
Anonymous Location
anon@whoare.you

***Abstract*—In this paper we investigate methods of predicting reciprocity in social networks, and use machine learning and regression to determine good indicators of reciprocity. Using the Twitter @ message graph, we discover that the ratio of a node's outdegree to its indegree is the best predictor of reciprocity, and that heuristics used in link prediction do not necessarily perform well in reciprocity prediction. In fact, using only simple properties that relate to the degree of a node or the number of messages that it sends or receives is sufficient in obtaining maximum accuracy in this prediction ¡¡rephrase?¿¿.**

I. Introduction

Reciprocity prediction and link prediction are inherently different problems - while link prediction is about predicting the occurence of rare events, reciprocity prediction predicts the "balance", or directions of an edge.

A. Related work

Tyler and Tang showed that reciprocity http://www.hpl.hp.com/research/idl/papers/rhythms/ECSCWFinal.pdf To be written.

B. Twitter as a domain to analyze

Twitter is a good domain to explore the superposition of the reciprocated and unreciprocated networks. The reciprocated network consists of mainly mutual interactions between friends or people in the same social circle, while the unreciprocated network consists of interactions between individuals in different social circles. We can also relate these types of interactions to the concept of status - where people with similar status participate in reciprocal interactions (e.g. messages between friends), while those with dissimilar status participate in unreciprocal interactions (e.g. messages from fans to celebrities).

C. Problem definition

The input to our prediction problem is a graph $G = (V, E)$ and a node pair $\{v, w\}$, where $v, w \in V$ but all edges between v and w removed. Our task is to predict the direction of edges between v and w.

The problem of predicting reciprocation can be defined in two ways. In the first, we decide whether both (v, w) and (w, v) exist (bidirected/symmetric), or only one of (v, w), (w, v) exists (asymmetry), given that at least one edge exists between v and w. In the second, we decide whether (w, v) exists given an edge (v, w). It would appear that predicting symmetry in the relation between nodes is more difficult than predicting reciprocation in a specific direction, and we show that this is indeed the case.

D. Notation

We consider the subgraphs of the form $G_n = (V_n, E_n)$, where $V_n = \{v \mid v \in V, v \text{ sent } \geq n \text{ messages}\}$ and $E_n = \{e = (v, w) \mid e \in E,\ v \text{ and } w \in V_n\}$.

We also define $v \xrightarrow{k} w$, which means v sent w k messages. From this definition we can formalize reciprocity in terms of k. We define an edge (v, w) to be reciprocated if $v \xrightarrow{k} w$ and $w \xrightarrow{k} v$, and unreciprocated if $v \xrightarrow{k} w$ and $w \xrightarrow{0} v$.

Let the set of reciprocated edges be $E^r_k = \{(v, w) : v \xrightarrow{k} w \text{ and } w \xrightarrow{k} v\}$, and the set of unreciprocated edges be $E^u_k = \{(v, w) : v \xrightarrow{k} w\}$.

Let $\deg^-(v)$ and $\deg^+(v)$ be the indegree and outdegree of node v respectively, $\text{msg}^-(v)$ and $\text{msg}^+(v)$ be the messages received and sent by a node v, and $\Gamma^-(v) = \{w | (w, v) \in E\}$, or the set of people who send messages to v.

II. Dataset description

The sample dataset consisted of the directed @ message graph $G = (V, E)$ of the Twitter network from (TIME) to (TIME). 12,795,683 unique users ($|V|$) sent a total of 819,305,776 messages, with 156,868,257 unique directed interactions ($|E|$) taking place between users during this time.

We focus our analysis on a subgraph of this, the graph in which each person sends at least 1000 messages each (G_{1000}). In G_{1000}, $|E^r_{10}| = 797,342$, $|E^u_{10}| = 349,258$.

III. Methods for Reciprocity Prediction

Intuitively, features can measure whether v and w have similar status or a similar social circle, and both are potentially useful in predicting reciprocation. This section presents a survey of various methods that can be used in predicting reciprocity in networks. Each method assigns a value $\text{val}(v, w)$ to a node pair (v, w), or a value $\text{val}(v)$ to a single node v. Given values corresponding to all node pairs (or nodes) in question, we can then choose threshold values or ranges where we predict reciprocity, and none otherwise.

For each property, we picked a single value val_{OPT} for which we predict every pair with value lower than val_{OPT} as unreciprocated and reciprocated otherwise, or vice versa, to maximize prediction accuracy. Intuitively, we expect that larger values of each property correspond to a stronger indication of reciprocity. For example, a high mutual neighbor count for the nodes v and w could strongly indicate the existence of a reciprocated link between them.

Now, given the two ways that we can formulate the prediction problem, we present four different mechanisms that we can use to predict reciprocity. The first attempts to answer the question of symmetry, while the other three answer the problem of reciprocity, with different limits on the amount of information about the nodes in question used.

1) SYM (predicting symmetry), where we predict whether an edge is bidirectional or asymmetric after removing all information about the edge in question but using existing information about v and w,
2) REV (predicting a reverse edge), where we predict whether a reverse edge exists given that the forward edge (v, w) exists using information about v and w, and finally
3) REV-w (predicting a reverse edge using only w), where we predict whether a reverse edge exists given that (v, w) exists, but only using information about w in making that prediction.
4) REV-v (predicting a reverse edge using only v), where we predict whether a reverse edge exists given that (v, w) exists, but only using information about v in making that prediction.

With this framework, we can now choose specific features of the @ message graph that we can plug in and compare their relative performance, and also their combined performance.

A. Degree/message features

It seems intuitive that the relative indegree or outdegree of nodes would indicate whether a pair of nodes are in a one-sided or two-sided relationship. If both have a similar indegree, this might indicate that they are at a similar social status in the network. In contrast, a disproportionate indegree would indicate that one might be a celebrity and the other an average Joe, making it unlikely that their relationship is reciprocal.

For these features, we also looked at their absolute "counterparts" (ex. the indegree of v or the indegree of w of the edge (v, w)).

Indegree and outdegree ratio both measure the ratio of outdegrees or indegrees of two nodes, and $\mathrm{val}(v, w) = \deg^-(v)/\deg^-(w)$ or $\deg^+(v)/\deg^+(w)$ respectively.

Incoming message and outgoing message ratio are similar, but instead uses the total number of messages that a node receives or sends, rather than the unique nodes that a node sends messages to or receives messages from.

Incoming message/indegree ratio and outgoing message/outdegree ratio compares the ratio of two nodes' incoming message to indegree ratio or outgoing message to outdegree ratio. A high incoming message to indegree ratio might characterize people who have a small group of friends with which they exchange lots of messages, while a low incoming message to indegree ratio might characterize highly connected (and thus high-status) people in a network (as the messages they receive is "spread" over many more users).

Outdegree/indegree ratio is a heuristic that attempts to characterize the messaging activity of a single node - a celebrity might have high outdegree/indegree ratio because she receive many messages from many followers but herself sends relatively few messages. We then compute the ratio of the outdegree/indegree ratio of two nodes, or $\mathrm{val}(v, w) = \frac{\deg^+(v)}{\deg^-(v)} / \frac{\deg^+(w)}{\deg^-(w)}$.

B. Link prediction features

It is not intuitive whether methods that work well for link prediction would work well in reciprocity; while link prediction asks whether an edge could exist between two nodes, reciprocity prediction asks whether a known edge is bidirectional.

Mutual neighbors calculates the number of common people whom v and w both send messages to ($|\Gamma^+(v) \cap \Gamma^+(w)|$), or the number of people who send messages to both v and w ($|\Gamma^-(v) \cap \Gamma^-(w)|$).

Jaccard's coefficient, also based on the concept of mutual neighbors, calculates the similarity between two sets by taking the ratio of the cardinality of their intersection and their union. $\mathrm{val}(v, w) = \frac{|\Gamma^-(v) \cap \Gamma^-(w)|}{|\Gamma^-(v) \cup \Gamma^-(w)|}$.

Adamic and Adar [1], defined the similarity between web sites v, w to be $\sum_{\{x | v,w \text{ share feature } x\}} \frac{1}{\log \text{frequency}(x)}$, and we similarly define $\mathrm{val}(v, w)$ to be

$$\sum_{\{x | x \in \Gamma^-(v) \cap \Gamma^-(w)\}} \frac{1}{\log \deg^-(x)}.$$

Preferential attachment is another popular heuristic in modeling network growth, where the probability that an edge forms with a specific node is proportional to its existing indegree. Here, $\mathrm{val}(v, w) = \deg^-(v) \cdot \deg^+(w)$, or $\deg^+(v) \cdot \deg^-(w)$. Notice that taking the ratio of these two values is equivalent to the outdegree/indegree ratio between two nodes.

2 step paths (ratio) is a simplification of what Katz [2] developed as a measure of status by calculating the number of paths between two nodes. In this study, we only consider paths of length 2, and $\mathrm{val}(v, w) = \mathrm{paths}^2(v, w)|$, where $\mathrm{paths}^2(v, w)$ is the set of paths from v to w of length 2. The 2 step paths ratio is simply the ratio of number of two step paths from v to w to that from w to v.

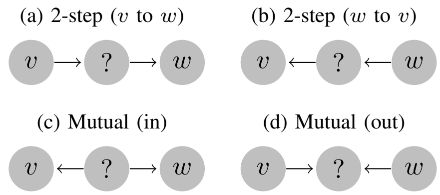
TODO I can also calculate that in the case where you predict a reverse edge, you use the path from v to w in your calculation, thus adding a single on-step path to the calculation. But this ends up being worse, so probably it's not important to include.

C. Different sets of features

For convenience, we further break down these into four sets of features:

1) Absolute degree/message features - degree, messages, message-degrees, outdegree-indegrees

Fig. 1: Two-step hops



2) Relative degree/message features - degree ratios, message ratios, message-degree ratios, and outdegree-indegree ratios
3) Two-step hop features - mutual neighbors (in and out), and two step paths (v to w and w to v)
4) Link prediction features - all other link prediction features not mentioned

D. Two-step hops

The importance of "friends of friends" or people 2 hops away from a given node lends itself to exploring features that directly arise out of the directed @-message graph. There are essentially four types of two-step hops, as shown in figure 1, corresponding to either the number of common in-neighbors or out-neighbors (mutual neighbors), or the number of directed paths from v to w or vice versa (two-step paths).

If both v and w send messages to many common people, it is likely that they are in the same social circle, or it could be that v and w simply message the same celebrities. If v and w receive many messages from the same group of people, it could be that both v and w are in the same community, or that both v and w are celebrities (who may or may not talk to each other) and there is an overlapping fanbase.

As the number of paths from v to w increases, there are two conflicting forces of w being popular and thus it unlikely for w to reciprocate an edge from v to w. The reverse case is simpler – as the number of paths from w to v increases, intuitively the likelihood that w knows v is a lot higher.

IV. Results and Discussion

A. Individual properties

To calculate the accuracy of the individual heuristics, we calculated val for each feature on the subset $E_{10}^r \cup E_{10}^u$ of the graph G_{1000}, where equal numbers of edges were taken from the two sets of reciprocated and unreciprocated edges. This would give a baseline accuracy is 0.500, and you would achieve this by predicting that all edges were of one type. We applied the SYM and REV mechanisms to feature sets 2-4, and REV-v and REV-w to feature set 1.

We then picked a threshold value val_{OPT} to optimize prediction accuracy - we predicted reciprocity above the threshold, and non-reciprocity below (or vice versa depending on which performed better). Tables III and IV summarize the performance of each heuristic on the subgraph $G_{1000}, k = 10$, while table II summarizes the different mechanisms of prediction for a single heuristic.

TABLE I: Reciprocity Prediction Features

Feature	$\mathbf{val}(v)$ or $\mathbf{val}(v, w)$
Absolute degree/message features	
Indegree or outdegree	$\deg^-(v)$ or $\deg^+(v)$
Incoming or outgoing messages	$\mathrm{msg}^-(v)$ or $\mathrm{msg}^+(v)$
Message-degree (in or out)	$\frac{\mathrm{msg}^-(v)}{\deg^-(v)}$ or $\frac{\mathrm{msg}^+(v)}{\deg^+(v)}$
Outdegree-indegree	$\frac{\deg^+(v)}{\deg^-(v)}$
Relative degree/message features	
Indegree ratio	$\deg^-(v)/\deg^-(w)$
Outdegree ratio	$\deg^+(v)/\deg^+(w)$
Incoming message ratio	$\mathrm{msg}^-(v)/\mathrm{msg}^-(w)$
Outgoing message ratio	$\mathrm{msg}^+(v)/\mathrm{msg}^+(w)$
Message-degree ratio (in)	$\frac{\mathrm{msg}^-(v)}{\deg^-(v)} / \frac{\mathrm{msg}^-(w)}{\deg^-(w)}$
Message-degree ratio (out)	$\frac{\mathrm{msg}^+(v)}{\deg^+(v)} / \frac{\mathrm{msg}^+(w)}{\deg^+(w)}$
Outdegree-indegree ratio	$\frac{\deg^+(v)}{\deg^-(v)} / \frac{\deg^+(w)}{\deg^-(w)}$
Link prediction features	
Mutual neighbors (in)	$\lvert\Gamma^-(v) \cap \Gamma^-(w)\rvert$
Mutual neighbors (out)	$\lvert\Gamma^+(v) \cap \Gamma^+(w)\rvert$
Jaccard's coefficient (in)	$\frac{\lvert\Gamma^-(v)\cap\Gamma^-(w)\rvert}{\lvert\Gamma^-(v)\cup\Gamma^-(w)\rvert}$
Jaccard's coefficient (out)	$\frac{\lvert\Gamma^+(v)\cap\Gamma^+(w)\rvert}{\lvert\Gamma^+(v)\cup\Gamma^+(w)\rvert}$
Adamic/Adar	$\sum_{\{x \mid x\in\Gamma^-(v)\cap\Gamma^-(w)\}} \frac{1}{\log \deg^-(x)}$
Preferential attachment (v to w)	$\deg^+(v) \cdot \deg^+(w)$
Preferential Attachment (w to v)	$\deg^-(v) \cdot \deg^-(w)$
Two-step paths (v to w)	$\lvert \mathrm{paths}^2(v, w)\rvert$
Two-step paths (w to v)	$\lvert \mathrm{paths}^2(w, v)\rvert$
Two-step paths ratio	$\frac{\lvert \mathrm{paths}^2(v,w)\rvert}{\lvert \mathrm{paths}^2(w,v)\rvert}$

In tables III and IV, a star ($*$) indicates that reciprocity was predicted when val was below the threshold, and a lack thereof indicates reciprocity was predicted when val was above the threshold.

In table II, SYM^+ refers to the prediction mechanism where we aim to predict symmetry and predict all edges with values *above* val_{OPT} to be reciprocated, and REV^- refers to the mechanism where we aim to predict whether a reverse edge (w, v) exists given (v, w) and predict all edges with values *below* val_{OPT} to be reciprocated.

1) Comparison of prediction mechanisms: We observe higher accuracy for the REV task than SYM, as REV is "easier" than SYM since you know more information about the edge (v, w).

Comparing REV-v to REV-w, we see REV-w obtains higher accuracy compared to REV-v, and this is because we're trying to predict the existence of the edge from w to v given (v, w), and knowing about w is a lot more valuable than knowing about v.

Note that SYM^-, REV^-, $\mathrm{REV}{-}w^+$ and REV-v^- are such poor predictors that simply predicting that everything was reciprocated (or unreciprocated) would have been better.

2) Comparison of methods of prediction:

a) Trends: On the whole, outdegree-indegree ratio and the two-step paths ratio are the best indicators of reciprocity. In

TABLE II: Indegree performance - different methods

Mechanism	**val**$_{OPT}$ (Percentile)	**Accuracy**
	Indegree ratio	
SYM$^+$	0.256 (40)	0.702
SYM$^-$	-	-
REV$^+$	0.414 (46)	0.759
REV$^-$	-	-
	Indegree of v or w	
REV-w^+	-	-
REV-w^-	74 (61)	0.731
REV-v^+	61 (60)	0.582
REV-v^-	-	-

fact, outdegree-indegree ratio alone already achieves accuracy to within $\pm 5\%$ of a decision tree using every feature.

b) Sending and receiving: When we look at features using one of the four mechanisms, the larger the value, the more likely it is for reciprocation to occur, and this is the case for a majority of features. For example, the *larger* the indegree ratio, the more in-links v has in relation to w, increasing the probability that w will link to v, given that we know that v already links to w.

However, the *smaller* the outdegree-indegree ratio, the more likely reciprocation occurs. In other words, a small numerator and a large denominator in $\frac{\deg^+(v)}{\deg^-(v)} / \frac{\deg^+(w)}{\deg^-(w)} = \frac{\deg^+(v)\deg^-(w)}{\deg^-(v)\deg^+(w)}$ is a good indicator of reciprocation. And indeed, a large denominator indicates that v has many in-links and w has many out-links, both of which increase the probability that w links to v.

We can generalize this further: if we know that v ”receives” a lot of traffic and w ”sends” a lot of traffic, that naturally reciprocation between w and v is likely to occur.

Interestingly, separating the numerator and denominator from the outdegree-indegree ratio above, which corresponds to our two preferential attachment features, results in two very different results. While a small numerator does decently well (preferential attachment (v to w)), a large denominator does not (preferential attachment (w to v)) and this only performs marginally above chance. ¡¡TODO Why?¿¿

c) REV-v vs. REV-w: Not surprisingly, REV-w performs better than REV-v on almost all features, and where REV-v performs better, the difference is not as significant. Analagous to whether you would want to know more about the features of a marketer (v) vs. a subscriber (w), knowing about the subscriber tells us more about the relationship between both. ¡¡Expand¿¿

B. Decision tree analysis

We then combined subsets of features and evaluated their performance, by splitting the edges in $E_1^r 0 \cup E_1 0^u$ randomly into two sets, training on one set and evaluating on the other.

The following combined subsets, in addition to each individual subset, were considered:

1) **All** (sets 1-4) – every single feature was considered.

TABLE III: Reciprocity Prediction Method Performance: Individual (REV)

Method	**val**$_{OPT}$ (Percentile)	**Accuracy**
Indegree ratio	0.414 (46)	0.759
Outdegree ratio	0.667 (43)	0.628
Incoming message ratio	0.333 (48)	0.772
Outgoing message ratio	0.905 (46)	0.547
Incoming message-indegree ratio	0.650 (39)	0.569
Outgoing message-outdegree ratio	0.791 (33)	0.615*
Outdegree-indegree ratio	1.72 (53)	0.820*
Mutual neighbors (in)	10 (61)	0.552
Mutual neighbors (out)	8 (51)	0.580
Jaccard's coefficient (in)	0.0345 (48)	0.684
Jaccard's coefficient (out)	0.0637 (55)	0.660
Adamic/Adar	1.94 (55)	0.561
Two-step paths (v to w)	6 (59)	0.517*
Two-step paths (w to v)	5 (51)	0.657
Two-step paths ratio (directed)	0.556 (52)	0.760
Two-step paths ratio (undirected)	0.259 (34)	0.516
Preferential attachment (v to w)	10230 (58)	0.687*
Preferential attachment (w to v)	2610 (37)	0.534*

TABLE IV: Reciprocity Prediction Method Performance: Individual (REV-v,REV-w)

Method	**val**$_{OPT}$ (Percentile)	**Accuracy**
Indegree (v)	61 (60)	0.582
Indegree (w)	148 (61)	0.731*
Outdegree (v)	25 (14)	0.506*
Outdegree (w)	105 (60)	0.647*
Incoming messages (v)	619 (53)	0.637
Incoming messages (w)	1802 (54)	0.733*
Outgoing messages (v)	906 (51)	0.542
Outgoing messages (w)	506 (17)	0.524*
Incoming message-indegree (v)	9.4 (41)	0.596
Incoming message-indegree (w)	9.12 (30)	0.535
Outgoing message-outdegree (v)	13.2 (50)	0.523
Outgoing message-outdegree (w)	8.14 (36)	0.661
Outdegree-indegree (v)	1.28 (53)	0.679*
Outdegree-indegree (w)	0.747 (50)	0.777

2) **All ratio** (sets 2,3,4) – all features that used ratios were considered.
3) **All absolute** (sets 1,3,4) – we wanted to see how using ”absolute” features would affect our decision accuracy.

We notice that the accuracy when we only use degree/message features compared to that when we include all features is the same.

Whenever the outdegree-indegree value or ratio was included as an attribute, it became the most important.

C. Regression analysis

We used a logistic regression model on subsets of features as well, where $f(z) = \frac{e^z}{e^z+1}$ and $z = \beta_0 + \beta F$, where $f(z)$ is binary and takes the value 1 when an edge is reciprocated, and 0 otherwise. F is the vector of features.

TABLE V: Decision Tree Accuracy

Set	Accuracy	Top-level attribute
Degree/message (1)	0.828	Outdegree-indegree (w)
Degree/message ratio (2)	0.861	Outdegree-indegree ratio
Two step hops (3)	0.795	Two-step paths (w to v)
Link prediction (4)	0.742	Two-step paths ratio (directed)
Combined		
All ratio (2,3,4)	0.861	Outdegree-indegree ratio
All absolute (1,3,4)	0.828	Outdegree-indegree (w)
All (1-4)	0.861	Outdegree-indegree ratio

TABLE VI: Logistic regression – relative degree/message-based features

Feature	β	**p** value
Indegree ratio	0.0101903	$< 2 \times 10^{-16}$
Outdegree ratio	0.0005775	~~0.2545~~
Incoming messages ratio	0.0230161	$< 2 \times 10^{-16}$
Outgoing messages ratio	-0.0047152	$< 2 \times 10^{-16}$
Incoming messages-indegree ratio	-0.0005545	~~0.0798~~
Outgoing messages-outdegree ratio	-0.0049387	$< 2 \times 10^{-16}$
Outdegree-indegree ratio	**-0.0562983**	$< 2 \times 10^{-16}$

TABLE VII: Logistic regression – two-step hop features

Feature	β	**p** value
Mutual neighbors (in)	-0.0117269	$< 2 \times 10^{-16}$
Mutual neighbors (out)	0.0180579	$< 2 \times 10^{-16}$
Two-step paths (v to w)	-0.1193624	$< 2 \times 10^{-16}$
Two-step paths (w to v)	0.1296081	$< 2 \times 10^{-16}$

TABLE VIII: Logistic regression – All ratio

Feature	β	**p** value
Indegree ratio	0.0120256	$< 2 \times 10^{-16}$
Outdegree ratio	-0.0015554	0.005739
Incoming messages ratio	0.0145437	$< 2 \times 10^{-16}$
Outgoing messages ratio	-0.0043189	$< 2 \times 10^{-16}$
Incoming messages-indegree ratio	0.0048525	$< 2 \times 10^{-16}$
Outgoing messages-outdegree ratio	-0.0046674	$< 2 \times 10^{-16}$
Outdegree-indegree ratio	-0.0301592	$< 2 \times 10^{-16}$
Mutual Neighbors (in)	-0.0279290	$< 2 \times 10^{-16}$
Mutual Neighbors (out)	0.0147103	$< 2 \times 10^{-16}$
Two-step paths (v to w)	**-0.0530463**	$< 2 \times 10^{-16}$
Two-step paths (w to v)	0.0182572	$< 2 \times 10^{-16}$
Two-step paths (directed)	**0.0394657**	$< 2 \times 10^{-16}$
Jaccard (in)	-0.0238541	$< 2 \times 10^{-16}$
Jaccard (out)	**0.0572358**	$< 2 \times 10^{-16}$
Adamic-Adar	-0.0001424	~~0.881637~~
Preferential attachment (v to w)	0.0010837	0.000627
Preferential attachment (w to v)	-	-

V. Twitter as a superposition of networks

A. (Un)reciprocated subgraph analysis

We also analyzed how various properties of the subgraphs G_n, as well as the edge sets E_k^r and E_k^u varied as we adjusted n and k.

TABLE IX: Logistic regression - All

Feature	β	**p** value
Indegree ratio	0.0041791	6.09×10^{-8}
Outdegree ratio	0.0046914	1.28×10^{-13}
Incoming messages ratio	0.0029794	0.000125
Outgoing messages ratio	0.0033361	3.10×10^{-10}
Incoming messages-indegree ratio	0.0040884	1.11×10^{-14}
Outgoing messages-outdegree ratio	-0.0015075	0.006283
Outdegree-indegree ratio	-0.0057958	$< 2 \times 10^{-16}$
Indegree (v)	0.0050520	4.30×10^{-11}
Indegree (w)	-0.0089197	$< 2 \times 10^{-16}$
Outdegree (v)	-0.0063247	$< 2 \times 10^{-16}$
Outdegree (w)	0.0035881	3.58×10^{-7}
Incoming messages (v)	0.0063390	$< 2 \times 10^{-16}$
Incoming messages (w)	-0.0179975	$< 2 \times 10^{-16}$
Outgoing messages (v)	-0.0070869	$< 2 \times 10^{-16}$
Outgoing messages (w)	0.0095572	$< 2 \times 10^{-16}$
Incoming message-indegree (v)	-0.0023250	1.11×10^{-5}
Incoming message-indegree (w)	-0.0004044	~~0.42781~~
Outgoing message-outdegree (v)	0.0007430	~~0.175454~~
Outgoing message-outdegree (w)	0.0024155	2.54×10^{-5}
Outdegree-indegree (v)	-0.0110324	$< 2 \times 10^{-16}$
Outdegree-indegree (w)	0.0218874	$< 2 \times 10^{-16}$
Mutual Neighbors (in)	-0.0194635	$< 2 \times 10^{-16}$
Mutual Neighbors (out)	0.0050245	$< 2 \times 10^{-16}$
Two-step paths (v to w)	-0.0462950	$< 2 \times 10^{-16}$
Two-step paths (w to v)	0.0167156	$< 2 \times 10^{-16}$
Two-step paths (directed)	0.0440107	$< 2 \times 10^{-16}$
Jaccard (in)	-0.0398243	$< 2 \times 10^{-16}$
Jaccard (out)	0.0561815	$< 2 \times 10^{-16}$
Adamic-Adar	0.0111504	$< 2 \times 10^{-16}$
Preferential attachment (v to w)	0.0009537	0.003002
Preferential attachment (w to v)	-	-

d) Reciprocated and unreciprocated edges: we notice that the frequency of reciprocated edges is approximately 2 to 3 times that of unreciprocated edges, and the proportion of reciprocated edges increases as n and k increases (Fig. 2, 3). While reciprocated communication is the dominant form of interaction, we see a significant number of ”unreciprocated” interaction, indicating that a significant number of relationships on Twitter are unbalanced. This could occur when a user of lower status tries to get the attention of a more influential user (of higher status) by messaging him or her (e.g. when a fan messages a celebrity multiple times hoping to get a reply).

e) Reciprocated and unreciprocated nodes: a majority of nodes have reciprocated relationships, with a small proportion having only unreciprocated relationships. A significant proportion of nodes take part in both reciprocated and unreciprocated relationships - indicating that while there are two distinct types of relationships occuring on Twitter, this does not correspond to two distinct types of users. A reason that ”unreciprocated” Twitter users are not common might be that social, and hence reciprocated relationships are the driving factor of active,

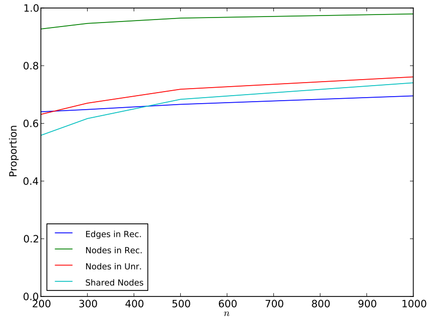


Fig. 2: Proportion of nodes or edges (varying n)

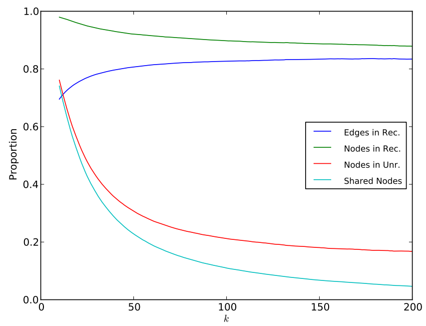


Fig. 3: Proportion of nodes or edges (varying k)

continued use of the platform.

We can also see this in Fig. 8, a scatter plot of the number of users with each of 3 types of interaction - 1 reciprocated and 2 unreciprocated, as an unreciprocal interaction is by definition asymmetric. We differentiate between both ends in an unreciprocated edge ($v \xrightarrow{k} w$ and $w \xrightarrow{0} v$), where where a user could be v if she's not replied to, or w if she doesn't reply. The most common type of nodes are those which only have reciprocated edges, with a lot less having some unreciprocal interactions of some type.

f) Clustering coefficient remains relatively stable as n, k vary: this demonstrates that the network properties of these subgraphs do not change significantly even if we sample from a relatively smaller population of all users (Fig. 4,5).

g) Connected component remains stable as n varies, but decreases as k increases.: The graphs corresponding to E_k^r and E_k^u are connected for relatively low values of k and decreases as k increases. ¡¡TODO Something interesting about this?¿¿ (Fig. 6,7).

VI. Conclusion

To be written.

Acknowledgment

The authors would like to thank Twitter for providing our experimental dataset. This work was supported by (SOMEGRANT).

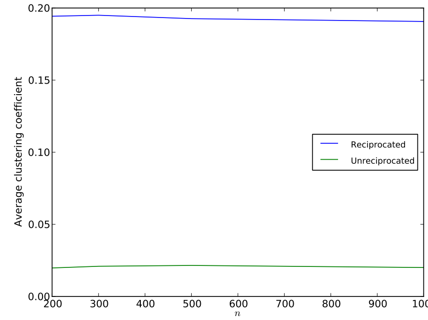


Fig. 4: Clustering coefficient (varying n)

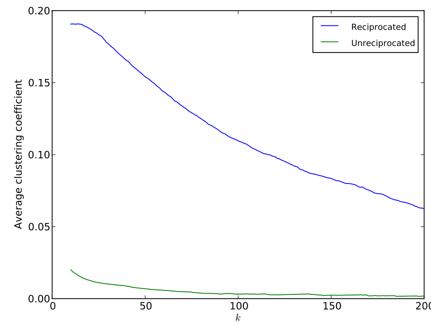


Fig. 5: Clustering coefficient (varying k)

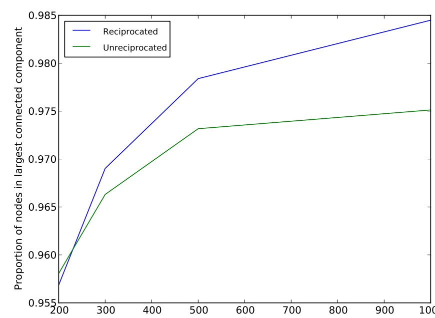


Fig. 6: Proportion in largest connected component (varying n)

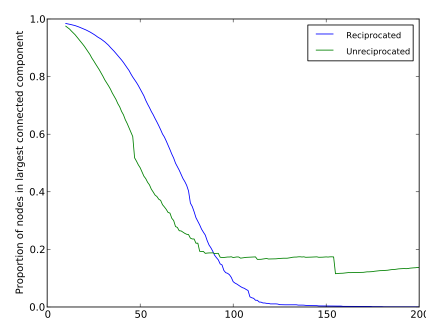


Fig. 7: Proportion in largest connected component (varying k)

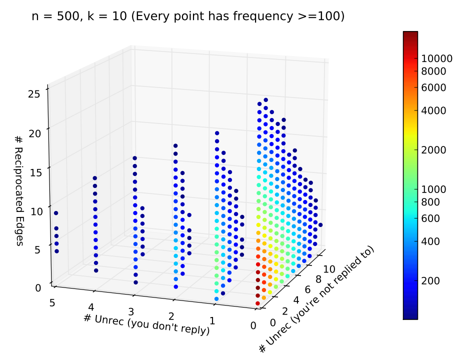


Fig. 8: Scatter plot of users' interaction types

References

[1] L. Adamic, "Friends and neighbors on the web," *Social Networks*, jan 2003.

[2] L. Katz, "A new status index derived from sociometric analysis," *Psychometrika*, 1953.