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 find that XXX and XXX perform well, while XXX and XXX do not, despite the fact that they are good link prediction heuristics.**

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.

Reciprocity prediction on this graph can be defined in two ways. In the first, given that at least one edge exists between v and w, decide whether both (v, w) and (w, v) exist (bidirected/symmetric), or only one of (v, w), (w, v) exists (asymmetry). In the second, given an edge (v, w), decide whether (w, v) exists.

D. Notation

We subsequently 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 $E_n = \{e = (v, w) \mid e \in E,\ v \text{ and } w \in V_n\}$.

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 $E_k^r = \{(v, w) : v \xrightarrow{k} w \text{ and } w \xrightarrow{k} v\}$, and the set of unreciprocated edges be $E_k^u = \{(v, w) : v \xrightarrow{k} w\}$.

Let $\deg^-(v)$ and $\deg^+(v)$ be the indegree and outdegree of node v respectively, $\mathrm{msg}^-(v)$ and $\mathrm{msg}^+(v)$ 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.

In G_{1000}, $|E_{10}^r| = 797,342$, $|E_{10}^u| = 349,258$.

III. Methods for Reciprocity Prediction

Intuitively, features and measure whether v and w have similar status or a similar social circle, and each is 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 $\mathrm{val}(v, w)$ to a node pair $\{v, w\}$. Given values corresponding to all node pairs in question, we can then choose threshold values or ranges where we predict reciprocity, and non-reciprocity otherwise.

For each property, we picked a single value $v^\star$ for which we predict every edge with value lower than $v^\star$ is 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.

We considered 3 different mechanisms of prediction:

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) REVW (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.

A. Degree/message-based prediction 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, thus it would be unlikely that the relationship between them is reciprocated.

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.

Inmessage and outmessage ratio are similar, but instead also take into account 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. People with 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 those with 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 characterize 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.

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

TABLE I: Reciprocity Prediction Features

Method	$\mathbf{val}(v, w)$
Indegree ratio	$\deg^-(v)/\deg^-(w)$
Outdegree ratio	$\deg^+(v)/\deg^+(w)$
Incoming message ratio	$\mathrm{msg}^-(v)/\mathrm{msg}^-(w)$
Outgoing message ratio	$\mathrm{msg}^+(v)/\mathrm{msg}^+(w)$
Incoming message-indegree ratio	$\frac{\mathrm{msg}^-(v)}{\deg^-(v)} / \frac{\mathrm{msg}^-(w)}{\deg^-(w)}$
Outgoing message-indegree ratio	$\frac{\mathrm{msg}^+(v)}{\deg^+(v)} / \frac{\mathrm{msg}^+(w)}{\deg^+(w)}$
Outdegree-indegree ratio	$\frac{\deg^+(v)}{\deg^-(v)} / \frac{\deg^+(w)}{\deg^-(w)}$
Mutual neighbors	$\lvert\Gamma^-(v) \cap \Gamma^-(w)\rvert$ or $\lvert\Gamma^+(v) \cap \Gamma^+(w)\rvert$
Jaccard's coefficient	$\frac{\lvert\Gamma^-(v)\cap\Gamma^-(w)\rvert}{\lvert\Gamma^-(v)\cup\Gamma^-(w)} \text{ or } \frac{\lvert\Gamma^+(v)\cap\Gamma^+(w)\rvert}{\lvert\Gamma^+(v)\cup\Gamma^+(w)}$
Adamic/Adar	$\sum_{\{x\mid x\in\Gamma^-(v)\cap\Gamma^-(w)\}} \frac{1}{\log\deg^-(x)}$
Preferential Attachment	$\deg^-(v)\cdot\deg^-(w)$ or $\deg^+(v)\cdot\deg^+(w)$
2-step paths	$\lvert\mathrm{paths}^2(v, w)\rvert$
2-step paths ratio	$\frac{\lvert\mathrm{paths}^2(v,w)\rvert}{\lvert\mathrm{paths}^2(w,v)\rvert} \text{ or } \frac{\sum_{i=1}^2\beta^i\lvert\mathrm{paths}^i(v,w)\rvert}{\sum_{i=1}^2\beta^i\lvert\mathrm{paths}^i(w,v)\rvert}$

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

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

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, $\mathrm{val}(v, w) = \deg^-(v)\cdot\deg^-(w)$.

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 $\mathrm{val}(v, w) = \sum_{i=1}^2 \beta^i |\mathrm{paths}^i(v, w)|$, where β is an arbitary damping constant, and $\mathrm{paths}^i(v, w)$ is the set of paths from v to w of length i. 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.

IV. Results and Discussion

A. Individual properties

To calculate the accuracy of the individual heuristics, we calculated the values obtained for each method on a subset of $E_{10}^r \cup E_{10}^u$ of the graph G_{1000}, where equal numbers of edges were taken from the two sets of reciprocated and unreciprocated edges. The baseline accuracy is 0.500, since you would achieve this by simply predicting that all edges were of one type.

We then picked a threshold value val_{OPT} to optimize prediction accuracy, where we would predict reciprocity above the threshold, and non-reciprocity otherwise (or vice versa). Table III summarizes the performance of each heuristic on the subgraph $G_{1000}, k = 10$, while table IV summarizes the different mechanisms of prediction for a single heuristic.

TABLE II: Indegree performance - different methods

Mechanism	$\mathbf{val}_{OPT}$ (Percentile)	Accuracy
SYM$^+$	0.256 (40)	0.702
SYM$^-$	-	-
REV$^+$	0.261 (40)	0.705
REV$^-$	-	-
REVW$^+$	-	-
REVW$^-$	74 (61)	0.731

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

Method	$\mathbf{val}_{OPT}$ (Percentile)	Accuracy
Indegree ratio	0.261 (40)	0.705
Outdegree ratio	0.398 (35)	0.579
Incoming message ratio	0.202 (42)	0.721
Outgoing message ratio	0.681 (61)	0.507
Incoming message-indegree ratio	0.462 (38)	0.551
Outgoing message-outdegree ratio	0.477 (33)	0.568
Outdegree-indegree ratio	0.496 (46)	0.777
Mutual neighbors (in)	10 (61)	0.552
Mutual neighbors (out)		
Jaccard's coefficient (in)	0.0345 (48)	0.684
Jaccard's coefficient (out)	0.0637 (55)	0.660
Adamic/Adar	XX 0.48 (??)	0.574
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)	?? (52)	0.760
Two-step paths ratio (undirected)		
Preferential attachment (in)	XX - (-)	0.500
Preferential attachment (out)		

1) Comparison of mechanisms of prediction: In table IV, 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.

We observe slightly higher accuracy for the REV task than SYM, as REV is "easier" than SYM since you know more information about the edge (v, w). Surprisingly, REVW performs even better TODO Why? We do notice that in all cases you want to predict that the top 40% of values as being reciprocated.

Note that SYM$^-$, REV$^-$ and REVW$^+$ did so poorly that simply predicting that everything was reciprocated (or unreciprocated) would do better.

2) Comparison of methods of prediction: On the whole, outdegree-indegree ratio and the two-step paths ratio are the best indicators of reciprocity.

B. Decision tree analysis

We can also focus on interesting subsets of features to see if they significantly improve the results obtained.

TABLE IV: Logistic regression on degree/message-based features

Feature	β	p value
Indegree ratio	0.00815	$< 2 \times 10^{-16}$
Outdegree ratio	-0.0002461	0.100
Incoming messages ratio	0.0139299	$< 2 \times 10^{-16}$
Outgoing messages ratio	-0.0040418	$< 2 \times 10^{-16}$
Incoming messages-indegree ratio	0.0003188	0.0183
Outgoing messages-outdegree ratio	0.0031140	$< 2 \times 10^{-16}$
Outdegree-indegree ratio	0.0404860	$< 2 \times 10^{-16}$

When we only use degree/message-based features, we obtain a prediction accuracy of 0.816. The most important factor was the outdegree-indegree ratio, and the least important factor the incoming message/indegree and outgoing message/outdegree ratios.

(OLD RESULT) With all heuristics (with skewed Katz), 0.878

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.

V. Twitter as a superposition of networks

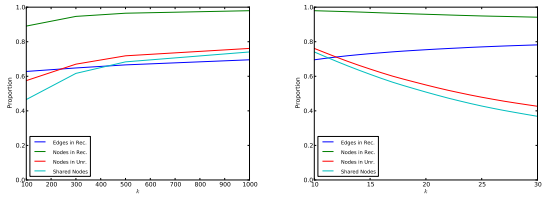
A. (Un)reciprocated subgraph analysis

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.

a) 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. 1). 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).

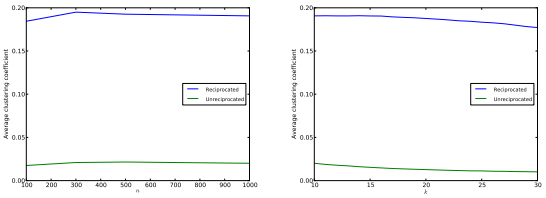
b) Reciprocated and unreciprocated nodes: a majority of nodes have reciprocated relationships, with a small proportion having only unreciprocated relationships. A significant proportion of nodes take part in both reciprocated and unreciprocated relationships - indicating that while there are two distinct 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, continued use of the platform.

We can also see this in Fig. 4, 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



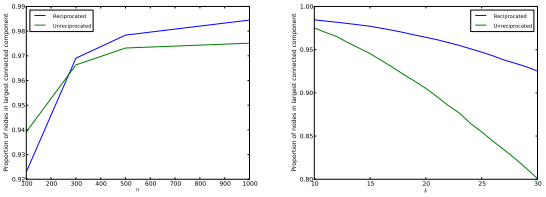
(a) Varying n (b) Varying k

Fig. 1: Proportion of nodes or edges



(a) Varying n (b) Varying k

Fig. 2: Clustering coefficient



(a) Varying n (b) Varying k

Fig. 3: Proportion in largest connected component

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.

c) Clustering coefficient remains relatively stable as n, k vary, and the graphs corresponding to E_k^r and E_k^u are connected: 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. 2,3).

VI. Conclusion

To be written.

Acknowledgment

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

References

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

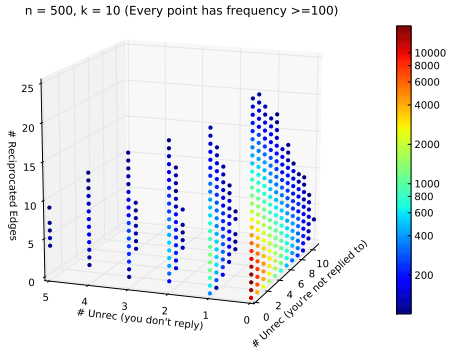


Fig. 4: Scatter plot of users' interaction types

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