Reciprocity Prediction 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

(TODO) What are @ messages (@ mentions, @ replies)? What are our motivations? Why is this important? What's the related work?

A. 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 (symmetry), or only one of (v, w), (w, v) exists (asymmetry). In the second, given an edge (v, w), decide whether (w, v) exists.

B. 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.

III. Methods for Reciprocity Prediction

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 in all others.

A. Degree-based prediction methods

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 perhaps one was 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 compares the ratio of two nodes' incoming message to indegree ratio. (INSERT RATIONALE?)

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 methods

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.

TABLE I: Reciprocity Prediction Methods

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	$\frac{\mathrm{msg}^-(v)}{\deg^-(v)} / \frac{\mathrm{msg}^-(w)}{\deg^-(w)}$
Outdegree-indegree ratio	$\frac{\deg^+(v)}{\deg^-(v)} / \frac{\deg^+(w)}{\deg^-(w)}$
Jaccard's coefficient	$\frac{\lvert\Gamma^-(v)\cap\Gamma^-(w)\rvert}{\lvert\Gamma^-(v)\cup\Gamma^-(w)}$
Adamic/Adar	$\sum_{\{x\mid x\in\Gamma^-(v)\cap\Gamma^-(w)\}} \frac{1}{\log\deg^-(x)}$
Preferential Attachment	$\deg^-(v)\cdot\deg^-(w)$
Katz	$\sum_{i=1}^{2}\beta^i\lvert\mathrm{paths}^i(v,w)\rvert$

TABLE II: Reciprocity Prediction Method Performance: Individual

Method	$\mathbf{val}_{OPT}$ (Percentile)	Accuracy
Indegree ratio		
Outdegree ratio		
Incoming message ratio		
Outgoing message ratio		
Incoming message-indegree ratio		
Outdegree-indegree ratio	0.498 (45)	0.768
Jaccard's coefficient		
Adamic/Adar	0.48 (??)	0.574
Preferential Attachment	- (-)	0.500
Katz	?? (52)	0.760

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

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

Katz [2] also developed a measure of status by calculating the number of paths between two nodes. In this study, we only consider paths of up to 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.

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 $E^r_{10} \cup E^u_{10}$, 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. Table II summarizes the performance of each heuristic on the subgraph $G_{1000}, k = 10$.

(INCLUDE a graph for one/all of the properties)

Surprisingly, preferential attachment does badly, even though...

B. Decision tree analysis

(TODO) We were also interested in how well these methods fare in comparison, and also wanted to know if we could achieve higher accuracy in reciprocity prediction if we combined these heuristics. Our results indicate that the Katz measure was the most important, followed by outdegree-indegree ratio.

(Old) Just using degree-based heuristics, get accuracy of 0.825. With all heuristics (with skewed Katz), 0.878

V. Twitter as a superposition

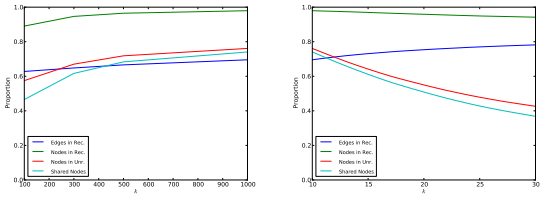
We also analyzed how various network properties of the subgraphs G_n, as well as the edge sets E^r_k and E^u_k varied as we adjusted n and k. Again, G_n is the subgraph where all nodes have sent at least n messages each, and E^r_k and E^u_k are the subsets of edges of some G_n that are either reciprocated or unreciprocated, where reciprocity is defined by the threshold 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 1. While reciprocated communication is the dominant form of interaction, we see a significant number of interactions that are "unreciprocated", indicating that a significant number of relationships on Twitter are unbalanced, or when a less known person (of lower status) tries to get the attention of a more influential person (of higher status) by messaging him or her (e.g. when fans message 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" users do not exist might be because active users of Twitter use it because of these reciprocated relationships.

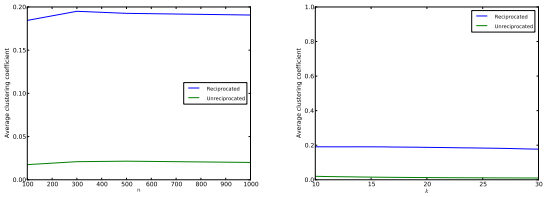
c) Clustering coefficient remains relatively stable as n, k vary.:

d) The graph corresponding to E^r_k and E^u_k are connected.:



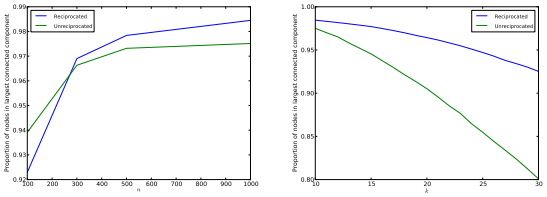
(a) Varying n (b) Varying k

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

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.

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