Report 2: Analysing Epinions Signed Network Data

Robert Hoff

March 6, 2012

1 Introduction

I'm considering a signed dataset collected from the website Epinions[1]. The analysis is inspired by the paper 'Signed Networks in Social Media' by Leskovec *et al*[2]. The dataset used here is not identical, but similar, it consists of about 90,000 nodes and 640,000 signed edges that are either -1 or 1 (see Table 5), and carry a Unix timestamp giving the time of creation.

The most interesting part of the paper is a description of a new model that indicates some predictive ability on link creation. The model is called 'status', and works by assuming that every user has an intrinsic value that will attract positive edges from, and negative edges to, lower-ranking individuals. The 'status', or degree, of a given user is inferred implicitly from existing connections. Status indicates that reciprocal edges between two nodes will tend to be of opposite signs. In the case of triads, for example if the links are cyclic we'd expect the appearance of exactly either one positive or one negative edge. Status theory is a contrast to the theory of balance[1], where nodes are conceptually considered to belong to dividing factions, that share reciprocal relationships of the same signs.

Considering the analysis in the paper, we can see that neither balance nor status explains the observed data very well. But, using the theories in particular contexts agree when predicting edge formation. For example, when predicting reciprocal edges it is overwhelmingly likely that they will be reciprocated of equal sign if positive, and significantly more likely if the sign is negative - but this case is not as conclusive. Equal sign reciprocation is in agreement with balance theory, status theory suggests otherwise. The paper[2] suggests that status theory in some cases holds in the case of triads, not pairs. I'm exploring in this report why this is the case, and to consider the user's conceptual models in more detail, how this may be taken into account and to look for agreements in different aspects of the data. I will also explore why status theory doesn't universally hold for all possible triads, and suggest a different way to approach the link prediction is by considering how existing edges define the characteristics of the users. This seems to suggest that it is the user's characteristics and conceptual model of the network dynamics that determine probabilities of edge-formations.

nodes	edges	+ive edges	$P(+)$	-ive edges	$P(-)$
88,951	639,078	529,974	0.829	109,104	0.171

Table 1: Epinions graph characteristics

2 Analysis

In criticism of the paper, the dynamics of the Epinions community hasn't seemed to been taken carefully into account. By considering the characteristics of the site, and user behaviours, we should be able to formulate a better

[1] http://www.epinions.com/

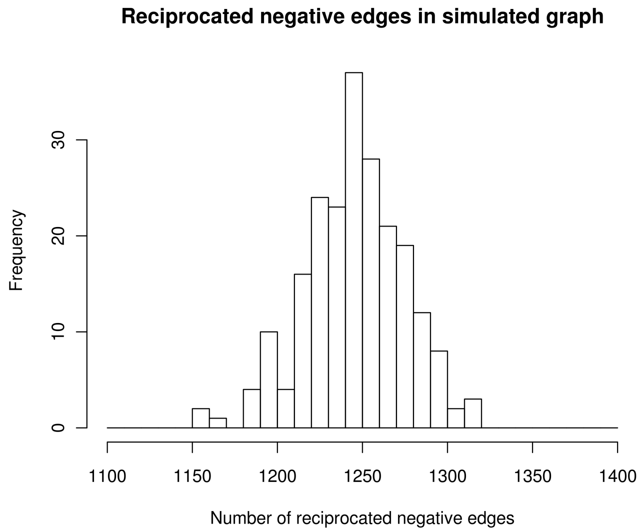

Figure 1: Simulation shows that random modelling doesn't exactly explain the number of observed reciprocated edges. Here the mean $\mu = 1246.7$ and standard deviation $\sigma^2 = 29.1$.

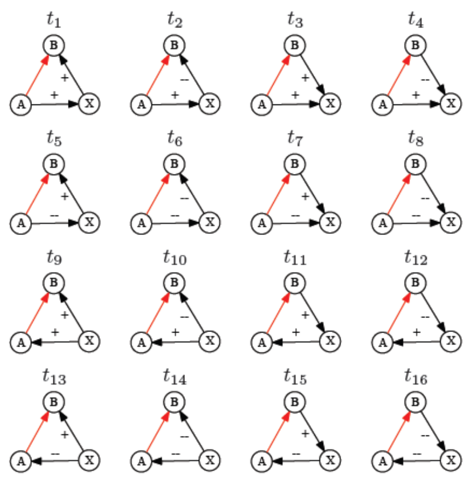

Figure 2: Triad configurations

model for the edge predictions. Results considered from the paper are also based three different websites, with this amalgamation important subtleties based on individual sites are lost. Pragmatically, by trying out Epinions and a comparison site Slashdot, edge objectives are conceptually different. In Epinions there is a community understanding that the edges are related to 'quality of reviews', in Slashdot however, the positive edges have important social networking functions that are missing on Epinions such information sharing and reduced barriers for getting in touch. With this in mind it is logically incompatible to compare the datasets directly.

As such, Epinions isn't predominantly a social networking site. The users are there to create product reviews, to discover products reviewed by other members, and indicate trust or distrust towards users (the signed edges on our graph). The conceptual model should be such that approval or distrust towards other users is related to perceived quality of their reviews. Balance theory certainly break down on Epinion because it assumes knowledge of other users' network connection, but there are generally hidden (only positive edges are shown and these are

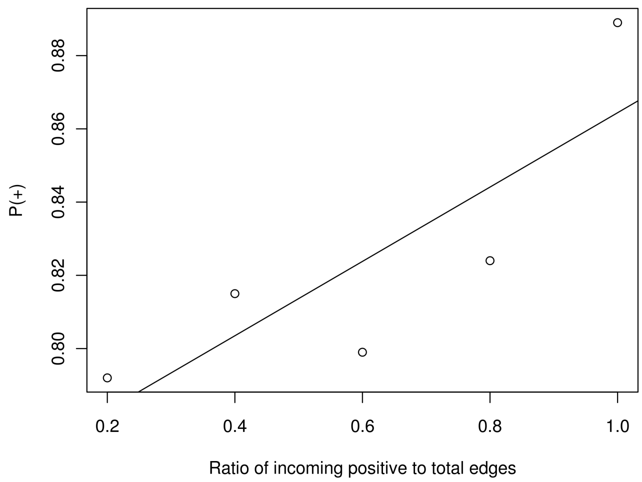

Figure 3: There is a weak correlation between the ratio of incoming positive edges, and the ratio of outgoing edges, which is equal to the probability of a given node to form new positive edges.

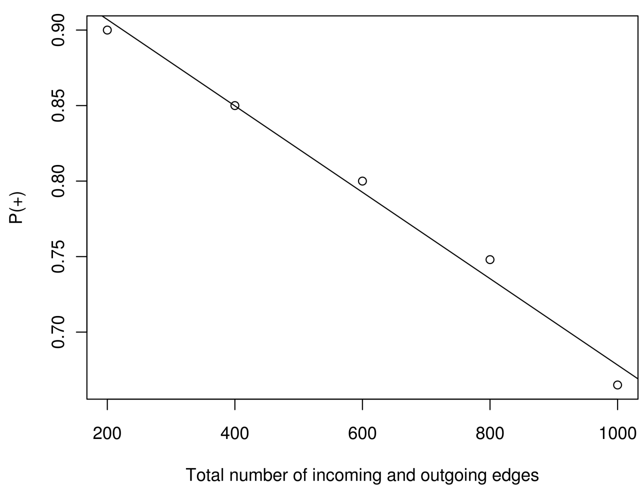

Figure 4: There is a strong linear correlation between the number of total edges for given node and the ratio of positive to total outgoing edges (which can be taken as the probability for a given node to form new positive edges).

not emphasised). Status theory is not so dependent on this knowledge, but it implies that user's are rating other's relative to oneself, and there is no clear evidence why this is the case.

We can verify and define a property of a given user as 'quality', instead of 'status'. A person of high quality will gain approximately the same ratio of positive links throughout his lifetime independently of the total number of neighbours. This hypothesis can be verified considering 72 users with more than 1000 neighbours. The absolute mean difference of the ratio of incoming positive to total number of edges between the time they had 100 neighbours and currently, is 0.05.

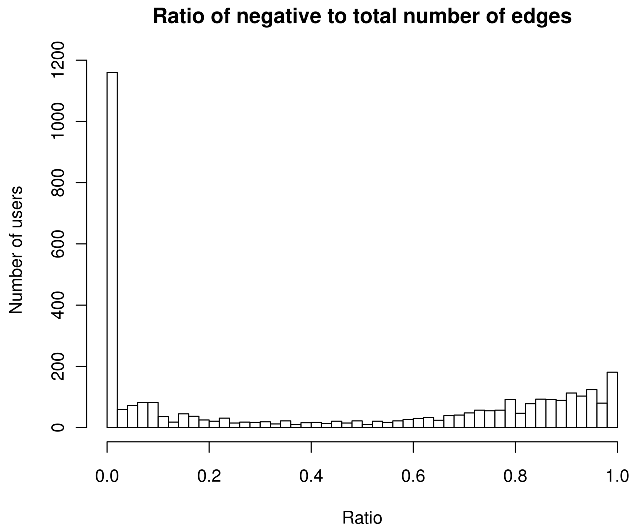

Figure 5: The graph shows active users that have at least 10 incoming and 10 outgoing edges. Most users receive only positive edges, which quickly drops. Towards the right of the chart we see a large group of users almost entirely receive *and* create negative edges.

A strong social networking aspect of the Epinions site, is that a total of 22.7% (120,437) of all positive edges are in mutual reciprocal positive relationships. This is due to the fact that positive edges are published, it makes sense psychologically that people amicably tend to return such a favour. Conversely, only 1.7% of all negative edges (1879) are in mutual reciprocal negative relationships. Since negative edges are not published on Epinions it seems that retaliation in the case of negative edges is largely avoided. To determine if the negative reciprocation was a result of random formation, not one of retaliation or other factors related to the dynamics of the site. A simulation was run generating random networks with the same number of nodes and edges, but placing the edges randomly according to the outgoing and incoming degree distribution of each node. Results are shown in Figure 1. I believe the actual number of reciprocal edges are higher than simulated predictions because of dynamics of the site that are hard to take into account. Two users might exchange unfriendly emails, which leads to distrusting each other. And, the clustering of the graph structure isn't uniform, the reason is due to different topics (product categories) that appeal to different interest groups.

There are several correlations that may be discerned by logically considering the dynamics of the network. One such consideration arose from considering the triads shown in Figure 2, it was given that nodes receiving a positive edge (triads $t_9, t_{10}, t_{11}, t_{12}$) was shown to have a less then expected probability of generating positive edges themselves. Figure 3 shows however that users tend to receive a high fraction of positive edges also tend to create more. On the other hand, Figure 4 shows that the more neighbours a node has, the probability of creating positive edges significantly drops. Although it looks highly linear it is probably a coincidence, but the correlation is significant. It is probably explained that, when users become highly established, they don't care or notice so much about gaining friends or reputation. So possibly users with high degrees may be considered better indicators of objective quality.

Deviant behaviour was observed in two important groups. One of these groups is a set of users that almost entirely receive negative edges. Figure 5 shows towards the left of the chart that most people don't receive particularly many negative edges, and this quickly quickly drops in frequency. However, towards the right of the chart, some people attract many, which indicates that there are a large group of people that are producing reviews that are not appreciated by the community. The other group that deviant behaviour was observed are users with very few neighbours. For example in the case where they have one or two, ratios are often absolute. People with few links probably have different goals than established users, so I believe it would be better to most analysis with these nodes removed.

3 Results

By removing deviant users, and non-active users (defined to have less than either 10 incoming or 10 outgoing edges), it is observed that the link-predictions differ significantly, when re-running the analysis. In Table 2, dataset1 uses the entire graph, an analysis that is equivalent to the one in Lekovec[2]. The surprise values that are defined compared to the receptive and generative baselines over all the triads of a given type are sensitive to the composition of the network. Table 2 show the triads $t_{13}, t_{14}, t_{15}, t_{16}$ for their significant effect.

Considering Lekovec's results in more detail. I believe the link-prediction can be more readily be made by formulating different characteristics of the users based on the network structure. For example, existing edges indicate primarily two things, either quality of the receiving node, *or* affinity of the originating node as a certain sign. It seem that combinations of these characteristics can explain all the new link-formations. Consider for example t_8, in this case A is pointing a negative edge to X, which indicates that A is belligerent. However, since B is also pointing a negative edge to X, it increases instead the probability that X is a low quality node and reduced the likelihood of the first observation. Therefore the link creation of A to B in this case is overwhelmingly positive compared to all the other cases t_5, t_6, t_7 where the (B,X) edges are reversed or positive doesn't subtract from the likelihood that A has a greater than average affinity for assigning negative edges.

The main points demonstrated were that balance and status doesn't match well with conceptual models, that can be deduced by observing the behaviour and correlations of the network. They do apply partially in some cases, for example in reciprocation, and in the triads where quality is a determining factor. I've also shown that by removing certain groups of users that deviate in their behaviour from averages (very low quality, or inactive users) the measurements used by Leskovec are very sensitive. We would expect that a curated dataset applies more readily to the intended behaviour of the network.

t_i	$P(+)$	s_g	s_r
		dataset1	
t_{13}	0.78	22.7	26.0
t_{14}	0.69	-8.3	-20.2
t_{15}	0.73	5.0	19.0
t_{16}	0.79	12.0	17.9
		dataset2	
t_{13}	0.78	30.2	27.2
t_{14}	0.69	-10.2	-10.8
t_{15}	0.73	5.0	7.2
t_{16}	0.79	11.9	10.6

Table 2: After removing inactive, and deviant users (dataset2). The baseline generative and receptive surprises are seen to be sensitive.

References

[1] Heider F. Attitudes and cognitive organization. *The Journal of Psychology*, 21:107–112, 1943.

[2] Lekovec J., Huttenlocher D., and Kleinberg J. Signed networks in social media. *Proceedings of the 28th international conference on Human factors in computing systems*, 2010.