Trust and Distrust Across Coalitions: Shapley Value Based Centrality Measures for Signed Networks (Student Abstract Version)

Varun Gangal[1], **Abhishek Narwekar**[1], **Balaraman Ravindran**[1], **Ramasuri Narayanam**[2]
[1] Indian Institute of Technology Madras, Chennai, Tamil Nadu, 600036, India, +91-44-22574350
vgtomahawk@gmail.com,abhishekknkar@gmail.com,ravi@cse.iitm.ac.in
[2] IBM Research India, Bangalore, India
ramasurn@in.ibm.com

Abstract

We propose Shapley Value based centrality measures for signed social networks. We also demonstrate that they lead to improved precision for the troll detection task.

Introduction

Signed social networks (SSNs) are social networks comprising of trust/distrust or friend-foe relationships betweeen users. These relationships may be explicit, such as in the Slashdot network (2009) , or inferred from interactions such as elections and conversations, such as on Wikipedia . Mathematically, a SSN can be specified as (V, E^+, E^-), where V is the set of vertices, with E^+ and E^- being sets of directed edges of the form (a, b), denoting a *trusts* b or a *distrusts* b, for E^+ and E^- respectively.

In a social network, a centrality measure assigns each node a value, which denotes its importance within the network. The notion of importance may vary based on the application, leading to a wide variety of such measures, based on position, betweenness and local importance. A centrality measure for a SSN needs to incorporate two sources of information and the interplay between them - trust and distrust edges as compared to just one type of edge in unsigned networks. This, in addition to the imbalance in real-world SSNs between positive and negative edges, makes defining SSN centrality measures a non-trivial task. A simple centrality measure for SSNs, first proposed in (2009), is the simple net positive in- degree, also called Fans Minus Freaks (FMF) centrality measure . Other measures include generalizations of eigenvector centrality and PageRank. A disadvantage of some of these centrality measures is that they consider every node in isolation when computing the centrality. This ignores the synergy between nodes, where a node is important by virtue of its combination with other groups of nodes. Moreover, ignoring the synergy can also make some of these centrality measures vulnerable to attacks wherein groups of nodes work together, as noted in (Kumar, Spezzano, and Subrahmanian 2014), to boost their individual centralities or reduce other nodes's centralities. One approach to incorporate this synergy has been to define a cooperative game, which assigns a value to every possible subset of nodes $C \subseteq V$ given by the characteristic function $\nu(C)$ of the game. The value assigned to a node is a weighted sum of the marginal contributions it makes to the values of all possible subsets, also known as the Shapley Value (SV). SV based centrality provides an intuitive way of capturing a node's centrality in combination with different groups of other nodes in the network.

Earlier works on SV based centrality measures, such as Suri and Narahari (2008) used MC sampling to compute centrality. Aadithya et al. (2010) introduced the possibility of deriving a closed form expression by defining $\nu(C)$ appropriately. To the best of our knowledge, ours is the first work to define cooperative game theoretic centrality measures for SSNs. Moreover, we are also the first to evaluate such measures for a centrality-based ranking task, in traditional or signed networks. Earlier works have evaluated these measures for other tasks such as influence maximization (Suri and Narahari 2008). Here we consider the task of ranking users to detect trolls in a SSN. Trolls or malicious users, are users with a highly negative reputation amongst the users of the network. In the availability of ground truth, one can evaluate a centrality measure by considering how low these "trolls" rank in a ranklist of users according to the measure.

Definitions

We define four different games, deriving the closed form expression for the SV for each of them. We only present one of the expressions here, including the remaining expressions & proofs in the appendix [1]. We denote the positive and negative in-degrees, out-degrees and neighbor sets by $d^+_{in}(V)$, $d^-_{in}(V)$, $d^+_{out}(V)$, $d^-_{out}(V)$, $N^+_{in}(V)$, $N^-_{in}(V)$, $N^+_{out}(V)$ and $N^-_{out}(V)$.

Net Positive Fringe (NPF)

Aadithya et al. (2010) defined the fringe (a generalization of degree) of a coalition, as the set of all nodes either in the coalition or having a positive neighbor in the coalition. In a similar spirit, we define the positive fringe $\nu^+(C)$ as the set of all nodes which are in the coalition [2] or have a positive out-neighbor in the coalition. The negative fringe

[1] http://tinyurl.com/o2g9a53

[2] We assume a node always trusts itself.

$\nu^-(C)$ is the set of all nodes with a negative out neighbor in the coalition. The characterisitic function $\nu(C)$ is given by $(|\nu^+(C)| - |\nu^-(C)|)$. For a node v_i, the SV of **NPF** gives the closed form expression

$$SV(v_i) = \sum_{v_j \in N_{in}^+(v_i) \cup v_i} \frac{1}{1 + d_{out}^+(v_j)} - \sum_{v_j \in N_{in}^-(v_i)} \frac{1}{d_{out}^-(v_j)}$$

In some sense, NPF can be thought of as generalizing FMF to a set. However, it ignores the interplay between positive and negative edges, since it is just the difference of their respective fringes.

Fringe Of Absolute Trust (FAT)

The intuition for this measure is that every node contributing to the set's value should be such that it does not distrust any node in the set. Most SSNs have more positive edges than negative ones. For instance, Slashdot has only 23.9% of its edges marked as negative. Hence, the negative edges may be interpreted strongly as an explicit "vote of distrust". In this game, the value of a coalition $\nu(C)$ is given by the set of all nodes which are either in the coalition or have one positive out neighbor in the coalition, provided they do not have any negative out neighbors in the coalition. Note that distrusting even a single member in the coalition removes a node from the coalition's value. Using the fringe notation, $\nu(C)$ is given by $|\nu^+(C) - \nu^-(C)|$.

Negated Fringe Of Absolute Distrust (NFADT)

This is similar to the **NAT** game with the roles of trust and distrust reversed. The characteristic function $\nu(C)$ is given by $\nu(C) = -|\nu^-(C) - \nu^+(C)|$. This is more a measure of disrepute than of reputation, hence we add the negative sign for it to make sense as a centrality measure.

Net Trust Votes (NTV)

Here, $\nu(C)$ is defined as the number of positive edges into the coalition (from a node outside the coalition) minus the number of negative edges into the coalition [3]. The intuition behind this measure is that a collective importance of a group of nodes is the net number of "votes" or edges in its favour, by nodes outside the group.

k-Hop NPF

We can generalize the **NPF** measure to k hops, considering the sets of in-neighbours $N_{in}^+(v_i)$ and $N_{in}^-(v_i)$ to be the set of neighbors within a distance of k-hops using in-edges. The notion of a path being positive or negative is determined using the principle "The enemy of my enemy is my friend", motivated by balance theory (Leskovec, Huttenlocher, and Kleinberg 2010). The sign of the path is given by the product of its edge signs.

[3]Note that internal edges are excluded from the value, since we wish to quantify the external trust of the coalition as a whole.

Evaluation and Results

We consider the task of ranking users to detect trolls in the Slashdot SSN, with 96 users annotated as trolls. We evaluate a centrality measure by first ranking the nodes of the graph in ascending order according to them, and then evaluating these ranklists based on how high the ground truth trolls rank in them. For evaluation, we consider two metrics

1. The number of trolls in the top g elements of the ranklist, where g is the number of ground truth trolls
2. The average precision (AP) metric from IR, with the trolls corresponding to "relevant documents".

Besides computing these for the full graph, we also compute the mean of the AP (MAP) over 50 subgraphs formed by deleting 5 %, 10 % and 20% of the nodes. We observe that the **NPF**, **FAT** and **NFADT** measures perform considerably better than **FMF**, both on the full graph as well as on each of the random subsamples. Moreover, the **3-Hop NPF** gives the highest average precision amongst all the measures, while the performance of **k-Hop FMF** decreases when we go to 3 hops. In addition to this, we include evaluation the robustness of these measures to certain common attacks by trolls in the appendix.

Approach	In Top 96	AP	MAP-5	MAP-10	MAP-20
FMF	10	0.031	0.031	0.033	0.033
NTV	7	0.044	0.043	0.045	0.042
NPF	18	**0.104**	0.104	0.107	0.108
FAT	15	**0.092**	0.092	0.094	0.093
NFADT	**19**	**0.130**	0.131	0.134	0.133
2 Hop FMF	**25**	0.158	0.157	0.153	0.155
3 Hop FMF	7	0.023	0.025	0.025	0.025
2 Hop NPF	16	0.148	0.139	0.143	0.152
3 Hop NPF	**22**	**0.193**	0.193	0.193	0.192

Complexity

Note that computing **NTV** and **FMF** takes $O(V)$ time, while **NPF**, **FAT** and **NFADT** take roughly $O(V + E)$ time.

Conclusion

We propose here, for the first time, cooperative game theoretic centrality measures for a SSN, demonstrating that even a simple measure such as FMF can detect trolls more effectively by generalizing it using SV.

References

Aadithya, K. V.; Ravindran, B.; Michalak, T. P.; and Jennings, N. R. 2010. Efficient computation of the shapley value for centrality in networks. In *Internet and Network Economics*. Springer.

Kumar, S.; Spezzano, F.; and Subrahmanian, V. 2014. Accurately detecting trolls in Slashdot Zoo via decluttering. In *ASONAM, 2014*.

Kunegis, J.; Lommatzsch, A.; and Bauckhage, C. 2009. The slashdot zoo: mining a social network with negative edges. In *WWW*.

Leskovec, J.; Huttenlocher, D.; and Kleinberg, J. 2010. Signed networks in social media. In *SIGCHI*.

Suri, N. R., and Narahari, Y. 2008. Determining the top-k nodes in social networks using the Shapley value. In *AAMAS*.