Link-Sign Prediction In Social Networks
CS401 - Applied Data Analysis

Romain Gros [*], Balz Marty[†], Jimmy Wilde[‡]
EPFL - Lausanne, Switzerland
Email: [*]romain.gros@epfl.ch, [†]balz.marty@epfl.ch, [‡]jimmy.wilde@epfl.ch

Abstract—**User interactions on social media - like all human communication – range from affection to antagonism. Online social networks can thus provide convenient datasets for studying signed networks, a subfield of social psychology. Predictions from psychological theories on signed networks have been tested against online social networks before. In the present study, we extend this idea to predicting edge signs between user pairs, based on the local graph context in which they are embedded. A predictive model allows us to disentangle the effects of different phenomena foretold by signed network theory.**

I. Introduction

Social psychology has developed theories that make (sometimes conflicting) claims about the structure of social networks. Certain theories can be formalized by viewing networks as graphs with positive or negative edges connecting pairs of people depending on the quality of their relation. On some online social networks, connections between users are naturally positive or negative, for instance, when voting to give or not give a fellow user admin status. In ”Signed Networks in Social Media”[2], Leskovec et al. investigated if concepts from social psychology - *structural balance, status, reciprocation, embeddedness* - can provide plausible explanations for such networks observed on online platforms. In this report, we go beyond validating these concepts independently and present a logistic regression model that predicts edge signs. This enables us to estimate the magnitudes of their respective effects.

II. Dataset description

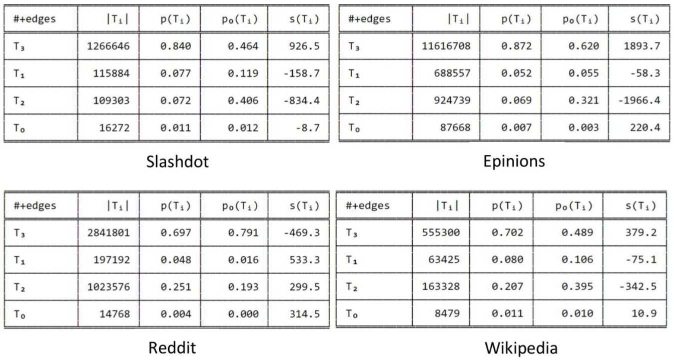
Four large signed networks from online social media platforms were analyzed. In addition to the same three datasets used by Leskovec et al. from Wikipedia, Slashdot, and Epinions, we studied a similar network from the social news aggregation website Reddit. It contains links between subreddits. A subreddit is a community on the social media platform Reddit. In all these datasets, links between users are represented by a weight from a source, i.e. the user that judges another user, to a target, who is being judged. The binary weight of such an edge can either be positive (+) or negative (-).

III. Analyzing networks properties

As a first step, we assessed whether the additional network from Reddit is similar to the other three networks in terms of graph-theoretical properties. To answer this question, we replicated *Table 3* of Leskovec et al. (see Figure 1), which shows the number of balanced and unbalanced undirected triads in each graph. The difference between the observed and expected proportions of the triads is represented in column $s(T_i)$. It is computed using the following formula:

$$s(T_i) = \frac{|T_i| - E[|T_i|]}{\sqrt{\Delta p_0(T_i)(1 - p_0(T_i))}}$$

where $|T_i|$ represents the number of observed triad of a particular type, $E[|T_i|]$ is the expected number of triads, and $p_0(T_i)$ is the a priori probability of obtaining a triad of that type. We found that the Reddit dataset does not follow the same pattern as the other datasets when it comes to under- and overrepresentation of certain triad types. In the Reddit dataset, triads with all-positive edges are underrepresented and all other types are overrepresented. This goes against both structural balance theory and its weaker variant (see the original paper by Leskovec et al. for context). One could argue that a subreddit is not a person and thus not subjected to the same rules of social psychology. But it would have been quite plausible that the principle ”the friend of my friend is my friend” applies also to groups of people. We don’t see any particular reason why two groups that are friendly with a third group should not be friendly with one another. Further research is needed to explain this observation.



#+edges	$\lvert T_i \rvert$	$p(T_i)$	$p_0(T_i)$	$s(T_i)$
T_3	1266646	0.840	0.464	926.5
T_1	115884	0.077	0.119	-158.7
T_2	109303	0.072	0.406	-834.4
T_0	16272	0.011	0.012	-8.7

Slashdot

#+edges	$\lvert T_i \rvert$	$p(T_i)$	$p_0(T_i)$	$s(T_i)$
T_3	11616708	0.872	0.620	1893.7
T_1	688557	0.052	0.055	-58.3
T_2	924739	0.069	0.321	-1966.4
T_0	87668	0.007	0.003	220.4

Epinions

#+edges	$\lvert T_i \rvert$	$p(T_i)$	$p_0(T_i)$	$s(T_i)$
T_3	2841801	0.697	0.791	-469.3
T_1	197192	0.048	0.016	533.3
T_2	1023576	0.251	0.193	299.5
T_0	14768	0.004	0.000	314.5

Reddit

#+edges	$\lvert T_i \rvert$	$p(T_i)$	$p_0(T_i)$	$s(T_i)$
T_3	555300	0.702	0.489	379.2
T_1	63425	0.080	0.106	-75.1
T_2	163328	0.207	0.395	-342.5
T_0	8479	0.011	0.010	10.9

Wikipedia

Fig. 1: *graph-theoretical properties. Top left: Slashdot, top right: Epinions, bottom left: Reddit, bottom right: Wikipedia*

We then estimated the clustering coefficient and the average shortest path length for each graph to get a sense of their efficiencies (Table I).

	Clustering coeff.	ASPL
Epinions	0.128	4.156
Slashdot	0.059	4.101
Wikipedia	0.141	3.220
Reddit	0.217	3.668

TABLE I: *Efficiency metrics for the four datasets. ASPL (Average Short Path Length).*

In Signed Netwoks in Social Media Leskovec *et al.* discuss all-positive and all-negative subnetworks. The following interpretation of their clustering was proposed:

> The giant components of real social networks are believed to consist of densely connected clusters linked by less embedded ties. The all-positive and all-negative networks in the real (rather than randomized) datasets are each biased toward one side of this balance: the all-positive networks have dense clusters without the bridging provided by less embedded ties, while the all-negative networks lack a sufficient abundance of dense clusters to sustain a large component.

We explored this hypothesis by performing spectral clustering and calculating the positive and negative edge sign fraction of inter- and intracluster links. According to the interpretation from above, the intracluster edge signs are expected to be more positive and the intercluster edge signs are expected to be more negative than the corresponding values of the full graph.

The complexity of the spectral clustering algorithm is $O(n^3)$ in the number of nodes. It is therefore not suitable for larger networks. Instead of sampling the larger graphs (from Epinions, Slashdot, and Reddit), which might alter the result, we ran it only on the Wikipedia dataset. Because the real number of clusters is unknown, we vary this parameter. Note also, that the networks are not fully connected. Thus, we first identify the largest connected component in the Wikipedia network before performing the clustering analysis.

Figure 2 shows that the fraction of positive and negative edges is very similar within and between clusters, no matter how many clusters the graph is cut into. Because there are fewer inter- than intracluster connections, the former undergo larger fluctuations as the number of clusters is varied. Overall, there seem to be slightly more positive edges in the largest connected component of the Wikipedia graph. These findings are not consistent with the hypothesis that negative edges are mostly formed between large clusters of mostly positive connections. This idea was not further pursued when predicting edge signs because the effect is very small at best.

IV. Sign prediction

Next, we build and trained a logistic regression model to predict the sign of a given edge. As features we used local properties of the graph, which are related to notions of structural balance, reciprocation, and embeddedness:

- Four features reflecting the proportion of positive edges from and to nodes i and j, which form an edge. The edges that we predict are removed from the fraction. These features were added primarily to predict the sign of edges that were not contained in any triad.
- Features created from the different combinations of directed signed triads, which are detailed in the following paragraph, and also here [1].

Let A_+ and A_- be the matrices of positive and negative edges such that $A = A_+ + A_-$. With these matrices, we count the number of the 16 possible configurations of signed directed triads given an edge $i \to j$. Indeed, if k is the third node of the triad, there are 4 possible configurations of edges between i and k which are

$$i \xrightarrow{+} k, \quad i \xrightarrow{-} k, \quad i \xleftarrow{+} k, \quad i \xleftarrow{-} k$$

Symmetrically, there are 4 possible configurations between j and k, hence 16 possible configurations, which will be the features representing the weight of embeddedness in our model. To count the number of each of the sixteen different triads for an edge $i \to j$ we compute the sixteen multiplications possible between A_+, A_-, A_+^T, and A_-^T and look at the entry (i, j) of each matrix. The regression model was optimized using a standard grid search procedure and 10-fold cross-validation for each dataset (and the combined data). We chose to train a model for each dataset to improve the results and spot differences between networks. The parameters which were optimized and their range of possible values are shown in Table II.

Parameters	Possible values
Penalty	l1, l2
C (regularization)	logspace(-6, 2, 15)
Solver	liblinear, lbfgs

TABLE II: *Optimization parameters*

The hyper-parameter combination yielding the highest AUC value for each dataset was selected. The final metrics were obtained with these best combinations (see Table III).

	Penalty	C (regularization)	Solver
Epinions	l2	0.0014	lbfgs
Slashdot	l2	0.0048	lbfgs
Wikipedia	l2	2976.3	liblinear
Reddit	l2	3.36e-06	lbfgs
All	l2	0.00043	liblinear

TABLE III: *Hyper-parameter optimization results.*

The percentage of positive edges is shown as a reference value for the accuracy because simply assuming all the edges to be positive would already yield a high accuracy (see Table IV). Our model does slightly better than this simple classifier, except for the Reddit dataset, where positive and neutral connections are confounded. There, the edge sign is determined by sentiment analysis, and both neutral and positive sentiments are counted as positive links. The precision follows the same pattern as the accuracy, and the recall is close to one for all networks. Training the model on all networks combined decreases its predictive power for all but the network from Reddit. This simple model exploiting the notions of balance, embeddedness, and reciprocation already yield better than chance predictions.

	Accuracy	Precision	Recall	AUC	% + edges
Epinions	0.934	0.948	0.977	0.956	85.3
Slashdot	0.863	0.892	0.936	0.920	77.4
Wikipedia	0.880	0.902	0.950	0.927	78.8
Reddit	0.924	0.925	0.999	0.715	92.5
All	0.897	0.907	0.979	0.898	83.9

TABLE IV: *Results of the logistic regression model (10-fold cross-validation).*

V. Matrix factorization and visualization

In this last part, we used another technique for predicting edge signs to further explore the psychology of signed interactions on social media - matrix factorization on the networks' sparse adjacency matrices. These matrices contain the weight of the edge from node i to node j as entry (i, j). The output is, in principle, easily interpretable; latent variables can be seen as modeling the factors that contribute to the esteem for and from other types of users. Matrix factorization thus implicitly generates individual *preference profiles* for each user. In order to visualize these preference profiles, we used dimensionality reduction techniques, in this case, PCA (Principal Component Analysis). However, the results did not show clusters of similar users as we had hoped (see Notebook), whether we chose 10, 20, of 50 latent features. Therefore, we applied the same technique to the triad feature matrix we used for logistic regression. We know, from regression metrics, that it can be used to explain the data. Still, the results remained far from being interpretable. To obtain a suitable representation, we then tried to use t-SNE (t-distributed Stochastic Neighbor Embedding). We visualized the results in 2D (Figure 3) and 3D (see Notebook) density plots. One can, to some extent, recognize dense clusters and less dense regions connecting them. However, despite all our efforts, we could not obtain a meaningful representation for the data. Maybe the data we are working with is simply not suited for this kind of representation.

VI. Conclusion

We extended Leskovec et al.'s work on the explanatory capacity of theories from social psychology by :

- Studying a signed network from another online social media platform, which proved in many respects different from the previously characterized datasets.
- Estimating efficiency metrics of the graph, namely the global clustering coefficient and the average shortest path length between nodes.
- Showing that edges between clusters are not much more negative than edges within clusters.
- Using local network properties to predict edge signs. Features related to structural balance, reciprocation, and embeddedness sufficed to get better-than-random predictions. Note that no status-related features were necessary.
- Trying to generate preference profiles of user via matrix factorization without significant results.

Appendix

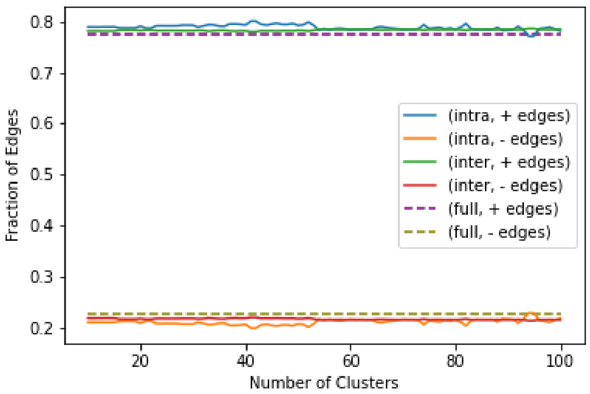


Fig. 2: Fraction of positive and negative intra- and intercluster links as a function of the number of clusters.

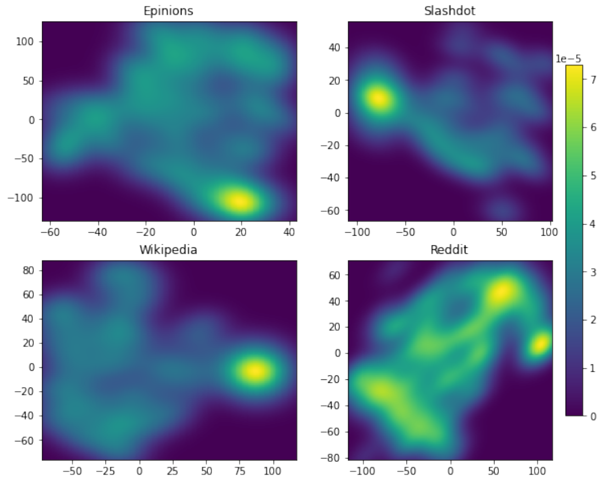


Fig. 3: 2D density plot of t-SNE vectors.

References

[1] Kai-Yang Chiang et al. "Prediction and Clustering in Signed Networks: A Local to Global Perspective". In: *Journal of Machine Learning Research* 15.34 (2014), pp. 1177–1213. URL: http://jmlr.org/papers/v15/chiang14a.html.

[2] Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. "Signed networks in social media". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '10. New York, NY, USA: Association for Computing Machinery, Apr. 2010, pp. 1361–1370. ISBN: 978-1-60558-929-9. DOI: 10.1145/1753326.1753532. URL: https://doi.org/10.1145/1753326.1753532 (visited on 10/24/2020).