Team PAM : Project P4 - Report - Signed Networks paper

Sinan Ayhan, Axel Måneskiöld, Laurine Planat
École Polytechnique Fédérale de Lausanne - EPFL , Switzerland - 2020

***Abstract*—This report is a creative extension of the paper 'Signed Networks In Social Media' [2], where social status and balance theories are studied to explain relationships between persons in a signed network. We will study a new dataset from the Reddit website, which contains links between communities. We will see that the balance theory may not be consistent with a network structure and thus links predictions, questioning the robustness of the findings of the paper. Nevertheless, we will notice that balance and status theories explain well the sign for reciprocated edges. Then we will study how Time and weight (as sentiment score or number of votes per person) may influence network structure, and thus if they might be taken into account during its analysis-prediction. We will illustrate our finding using the famous situation of 3 people (ABC), two of whom are already linked(Ab, BC), trying to predict the remaining link (CA).**

I. Introduction

In the *strong balance theory* also called Heider's theory, the main idea is that triangles with 3 positive signs (3 mutual friends, +++, T3), or triangles with one positive signs (2 friends with a common enemy, +–, T1) are more plausible than other configurations. The weak version also called Davis' theory keeps only the great likelihood of T3 and the implausibility of T2. On the other hand, the status theory explains that A links positively to B means that A thinks B has a higher status. Therefore B should link negatively to A, as B is supposed to have a higher status.

Here we will identify the strength and weaknesses of these two theories in network structure analysis and try to identify new variables that may have an impact on it. To do so we will test if these 2 theories are suited for a new dataset, called *Reddit*, and identify new variables that may have an impact on the sign of a link between persons and network structure: *Time* and *weight* (sentiment score in *Reddit* and number of vote per person in *Wikipedia*). We will illustrate our conclusion using three person sketch situation.

II. Dataset Description

The new dataset is from the Reddit website, a "*network of communities based on people's interest*. It is a hyperlink network between subreddits (https://snap.stanford.edu/data/soc-RedditHyperlinks.html) [1]. A subreddit is a web forum of a particular topic where one can post links or create a self-post and discuss. A hyperlink is a word, phrase, or image that you can click on to jump to a new document" [1]. Here a hyperlink points to a subreddit page. This directed network is extracted from the posts that create hyperlinks from one subreddit to another. In this network, a node is a particular subreddit, and an edge from A to B means that a post from A contains a hyperlink pointing to B.

The hyperlink can be present in either the *title* of the post or in the content, called *body*. This dataset consists of 2 CSV files: one for the title, and one for the body. Its fields are :

- Name of the Subreddit source page
- Name of the Subreddit target page
- Id of the post
- Timestamp of the link with the format: y-m-d, h-min-sec
- Link sentiment (+1, positive ; or -1, negative)
- Properties (86 different features)

Parameters	Reddit title	Reddit body	Reddit Total
Nodes	35,776	54,075	67,180
Edges	286,561	571,927	858,488
+ edges	92.6%	89.3%	90.4%
- edges	7.4%	10.7%	9.6%

TABLE I
Decription of the reddit dataset

Table I summarizes the properties of the dataset. To assess status vs. balance theory validity we use the *Total* dataset, which contains links from both the body and the title networks (concatenation of separated networks) as it has more edges and nodes. It contains fewer nodes than the Slashdot dataset but has as many edges as the Epinion one. Thus it is more connected. Plus, in this dataset, we have around 90% of positive edges, which is 10% higher than for the 3 datasets studied in the paper [2]. In Time and Weight impact analysis, we will focus on the body part, because the sentiment score might illustrate better the idea of the post based on more text content.

III. Methods

A. Balance theory

We first check the consistency of balance theory on the total Reddit dataset following the analysis pipeline described in the paper. To do so we replicate Table 3 [2] for the Reddit dataset. The idea is to compare the observed proportion of triad $(p(T_i))$ with the one obtained after edge sign randomization using positive and negative network edges proportions $(p_0(T_i))$. The quantification of the difference is indicated by the *surprise*, $s(T_i)$, which is the standard error. If $p(T_i) > p_0(T_i)$ it means the triad type Ti is over-represented relative to chance, if $p(T_i) < p_0(T_i)$, it is under-represented in the graph.

B. Reciprocated edges

Then we consider the reciprocated edges, which consist of mutual edges: let A have a link towards B, if B links to A, it means B link reciprocally to A. We describe both edges as *reciprocated edges*. As balance and status theories doesn't give the same prediction for reciprocated edges, their study gives a good interpretation of theories' consistency. To do so we enumerate the different conditional situations and compute the number of associated observations in the graph $[P(+|+), P(+|-), P(-|+), P(-|-)]$, as in Table 4 of [2]. To assess our method, we first replicate Table 4 for the Wikipedia dataset and quantify the error according to expected numbers. Thus we were able to validate the quality of the function implemented and therefore pursue the analysis on the Reddit dataset. It also shows good replication power for this part of the paper. In the method, we first find all couples A-B that have reciprocated edges. Then given the sign of the link A-B, we compute the number of couples for which the link B-A is positive or negative.

[1] from https://techterms.com/definition/hyperlink

C. Time variable analysis

In this part, we study how the link creation time can impact the sign of the link and the network structure. First, we identify the overall trend of edge distribution according to the variables by plotting the ratio of negative vs positive edges over the time index variables; year, month, week day, and hour. Thus we identify relevant subgroups and create associated sub-graphs, and compute network structure parameters. We will especially investigate the centrality, clustering, and behavior network characteristics.

To analyze the **centrality** of the sub-graphs, we compute the *Average Node Degree (AND)*, which corresponds to the average number of adjacent nodes for a node. We also calculate the *Average Closeness (AC)*, as a quantification of the node's relationship with all nodes in the network. High closeness means a short average distance to all the other nodes and thus an important sensitivity for edge changes even on another part of the network. We will also compute the average *Local Clustering Coefficient (LCC)*, it indicates to what extent are neighboring nodes autonomous from a common node (What impact on their ability to influence each other if removed?). Finally, we will compute the *Average Centrality Degree (ACD)* as it represents the total number of connection a node might have. Then, to analyze the **clustering** and the **behavior** of the sub-graph we will compute the *Global Clustering Coefficient (GCC)* and the *Assortativity Degree (AD)* which represent the preference of a node to link with a similar one.

D. Taking weights into account

Finally, we assign weights to the graphs. The idea is to analyze the structure of sub-graphs with different weights to see the possible impact of this variable on network organization. Here the weight is understood as a property of the edge.

For the Reddit dataset, we use the sentiment score as the weight of the post. By plotting the number of posts according to this variable we define three subgraphs with low, middle, and high scores. Then we compute the same network structure variable as for time analysis, using the same implementation. The Wikipedia dataset will be divided into significantly different parts in terms of weight. It will also be taken to account that the subgraph has to represent a sufficient amount of the data set, to do the analysis. In total three different subgraphs will be created.

In a general manner, during the analysis, we will replicate some tables (1, 3, 4, and 6) of the paper[2]. Here we will focus on the analysis of the results, in any details needed please refer to the original paper and the *jupyter notebook* with the details of our implementation. We are conscious that replication isn't a creative extension, however, our goal is to use the original data analysis pipeline on another dataset to identify its strength and limits and thus modulates the paper's conclusion. Thus it adds something to the original study.

IV. Results

A. Balance theory

According to Table II, balance theory doesn't fully explain signs in a triangle. Indeed, We have that T_1 (2 friends with a common enemy) are overrepresented, which follows the theory. But we also see that triads T_3 (3 mutual friends) are underrepresented compared to chance, which shouldn't be the case. Another result at odd with the theory is that triads T_0 (3 mutual enemies) and triads T_2 (2 enemies with the same friend) are both over-represented compared to chance.

Triad Ti	$\|Ti\|$	$p(Ti)$	$p_0(T_i)$	$s(Ti)$
T_3 | + + +	1,644,323	0.863	0.895	-143.6
T_1 | + - -	41,431	0.022	0.009	166.3
T_2 | + + -	209,148	0.109	0.093	76.4
T_0 | - - -	9,225	0.004	0.001	164.8

TABLE II
Number of balanced and unbalanced undirected triads for Reddit dataset

Probability	Count	Fraction
$P(+\|+)$	27,002	0.954
$P(-\|+)$	1287	0.045
$P(+\|-)$	1482	0.878
$P(-\|-)$	205	0.121

TABLE III
Edge reciprocation. Given that the first edge was of sign X $P(Y|X)$ give the probability that reciprocated edge is Y

For *Epinions*, *Slashdot* and *Wikipedia* datasets, the results were that Davis's weaker notion of balance - T_2 should be heavily under-represented and there is no reason to have T_1 over-represented and T_0 under-represented - explained well the results. But it's not the case for the Reddit dataset.

We found an example of a dataset where the balance theory doesn't explain the links. Therefore the use of this new dataset brings doubts to the conclusion that the authors make. The balance theory may not be able to explain links in a triad for an undirected configuration.

B. Reciprocated edges

Table III presents the results for the Reddit dataset. It seems they are consistent with the conclusion authors make: if A-B is positive, we have in 95.4% of cases that B-A is positive as well, which is higher than the 90% of positive links in the data. So balance theory explains well the proportion of $P(+|+)$, as a reciprocation of a positive link should be positive as well.

Plus, if A-B is negative, the status theory explains the high proportion of reciprocated positive links, but there is a small deviation in the direction of balance-based interpretation as the proportion is lower than the 90% of positive edges. If A links negatively to B (A thinks B has a lower status), then B should link positively to A (B thinks A has higher status).

C. Time variable analysis

In Table IV are presented the computed parameters for the sub-graphs drawn from the year's, month's, and hour's trend observation. In Reddit, the year 2014 corresponds to a high ratio of negative/positive edges while the year 2016 corresponds to a very small ratio. In Wikipedia, the ratio difference is less pronounced. The year 2005 corresponds to small, 2008 to middle, and 2007 to a high ratio, with a range of value smaller than for Reddit. First, we can see that the year variable seems to have an impact on the Wikipedia sub-graph network. The year 2008 appears to have a lower *AND* than 2005-7. Plus, *ACD* of the year 2007 is lowered by 2 compared to 2005-8. This may reveal a dissymmetry in incoming-outcoming edges proportion and perhaps reciprocated edges and could be an indicator of nodes activity: few incoming edges but an important total number of connections means a lot of outcoming edges and thus important activity. Also, its higher *GCC*, indicates the high connectivity of the network. Thus one could say the year 2008 is more tightly connected. Thus if we consider the three-person situation in 2008, the last edge

A-C is more likely to be negative than in 2005-7, also as nodes in 2008 appear to have important activity, they may greatly influence each other, thus it would be a bit restrained to only consider a single common neighbor B for AC edge prediction.

	Reddit		**Wikipedia**		
Year	**2014**	**2016**	**2005**	**2007**	**2008**
AND	5.768	5.857	20.407	21.854	4.292
AC	0.077	0.067	0.067	0.058	0.010
LCC	0.104	0.106	0.079	0.074	0.063
ACD	0.0004	0.0003	0.0160	0.006	0.015
GCC	0.052	0.052	0.079	0.061	0.091
AD	-0.074	-0.068	-0.188	-0.154	-0.188

	Reddit			**Wikipedia**		
Month	**february**	**march**	**june**	**november**	**june**	**july**
AND	4.060	4.211	3.915	10.052	10.283	9.593
AC	0.042	0.044	0.041	0.017	0.019	0.022
LCC	0.071	0.066	0.058	0.059	0.079	0.086
ACD	4e-4	4e-4	5e-4	0.005	6e-4	0.005
GCC	0.111	0.073	0.039	0.061	0.094	0.098
AD	-0.04	-0.04	-0.04	-0.08	-0.06	-0.1

	Reddit			**Wikipedia**	
Hour	**7h**	**12h**	**18h**	**6h**	**10h**
AND	3.213	3.260	3.081	3.428	3.058
AC	0.025	0.028	0.023	0.0050	0.0030
LCC	0.036	0.041	0.037	0.0047	0.0057
ACD	0.0005	0.0004	0.0004	0.0016	0.0016
GCC	0.064	0.036	0.029	0.0030	0.0021
AD	-0.015	-0.024	-0.008	-0.088	-0.099

TABLE IV

CENTRALITY, CLUSTERING AND BEHAVIOR PARAMETERS COMPUTED FOR SUB-GRAPHS OF REDDIT AND WIKIPEDIA FROM YEAR, MONTH AND HOUR'S TREND ANALYSIS

Furthermore also on the Wikipedia dataset, one can see Table IV that the *LCC* is raising according to the proportion of negative edges for the month analysis. With the *GCC* being higher in summer, it shows that summer months tend to be more connected and with nodes more independent from each other (the removal of one may have less effect on the ability of communication of its neighbors). Thus network may be less sensitive to node removal during summer than winter. On the other hand, one can see that the year doesn't seem to impact the Reddit year's and month's sub-graphs.

Finally, we can note that the hour seems to have an impact on the Reddit network structure. The *AND, AC, ALC* and *ACD* are higher in the sub-graph from 12h which has the higher negative/positive edges ratio. Also, its *AD* is the smallest and 7h has the higher *GCC*. Also observations for *AND* and *AC* in Wikipedia 6h-10h dataset are the same, while for *LCC* and *ACD* it is opposite. From these observations we may draw two conclusions; first, the time might have an impact on network structure but second, this impact is specific to the dataset. As we were able to see the variable statistically insignificant were not the same for both datasets. Thus we cannot remove variable in a general manner, one should check for impact on the dataset of interest before withdrawing it from the study.

D. Taking weigths into account

The results for the Reddit dataset are presented in Table V. The *AND* between the graphs fluctuates in the range of 4 to 5. The AD does not seem linked to the level of sentiment score. For the AC, LCC, ACD, and GCC the result is fluctuating in the same way, where the middle sentiment score has a lower value in the properties of the graph. Why this occurs is to find a reason for, but overall that trend is not very significant.

	Reddit		
Sentiment score	$\mathbf{-1 \leq x \leq -0.875}$	$\mathbf{0 \leq x \leq 0.125}$	$\mathbf{0.875 \leq x \leq 1}$
AND	4.594	4.280	5.260
AC	0.045	0.044	0.069
LCC	0.074	0.054	0.112
ACD	6e-4	3e-4	2e-4
GCC	0.032	0.030	0.048
AD	-0.094	-0.096	-0.062

	Wikipedia		
Number of votes per user	$\mathbf{100 \leq x \leq 200}$	$\mathbf{400 \leq x \leq 500}$	$\mathbf{800 \leq x \leq 900}$
AND	19.0	3.488	1.997
AC	0.025	0.002	0.001
LCC	0.160	0.243	0.0
ACD	0.008	0.003	0.002
GCC	0.018	0.001	0.0
AD	-0.046	-0.070	NaN

TABLE V

CENTRALITY, CLUSTERING AND BEHAVIOR PARAMETERS COMPUTED FOR SUB-GRAPHS OF SENTIMENT SCORE (REDDIT) AND OF NUMBER OF VOTES PER USER (WIKIPEDIA

The result from the Wikipedia subgraphs in Table V show a clear tendency in increasing number of average degree and closeness between the users with lower number votes, thus it does not seem to exist a tighter link between the users with a similar number of votes. The number of votes can be seen as a measurement of trustworthiness. The properties in the subgraphs show that the more trustworthy a user is the less expected is the user to have a connection to similar users, in terms of several votes. Of course, this comes with the consideration that the number of users is considerably larger for a graph with a low number of votes, thus it is possible to have a total higher number of edges. At the same does the measurement of the subgraph, between the highest number of votes by a user, have the max degree by one user at 773 compared to 208 for the subgraph with the lowest votes. That shows that in fact, the subgraph with the lowest number of votes has the node with the highest degree.

V. CONCLUSION

To conclude, we found an example of a dataset where the balance theory can't explain the proportion of triads per type. Therefore, as opposed to the conclusion from the authors, we conclude that this theory is not well suited to describe the proportion of undirected triads. Then, we examined if balance and status theories could explain the sign of reciprocated edge for the Reddit dataset, and we conclude that using them, we can indeed explain it. Indeed, balance theory can explain the sign of the reciprocation of a positive link (the reciprocation has more chance to be positive), and status theory can explain the sign of the reciprocation of a negative link (the reciprocation has more chance to be positive).

Then, we studied the impact of time on the signs of the links, and we have seen that the global trend identified for each of the variables depends on the dataset. It seems logical as the two datasets are coming from two different websites and nodes-edges don't represent the same things. However, it is important to keep in mind that a prior social-historic analysis of the situation associated with the dataset or website might be very useful to understand the trend.

Finally, we saw how the strength of weights in a weighted network can impact the structure of the network, and we conclude that the strength of subgraphs of the network can be highly affected by how similar the nodes are. Furthermore, this can depend on how significant and extreme the similarity is.

References

[1] Srijan Kumar et al. “Community interaction and conflict on the web”. In: *Proceedings of the 2018 World Wide Web Conference on World Wide Web*. International World Wide Web Conferences Steering Committee. 2018, pp. 933–943.

[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. Atlanta, Georgia, USA: Association for Computing Machinery, 2010, pp. 1361–1370. ISBN: 9781605589299. DOI: 10.1145/1753326.1753532. URL: https://doi.org/10.1145/1753326.1753532.