Casting a Web of Trust over Wikipedia: an Interaction-based Approach *

Extended Version

Silviu Maniu Talel Abdessalem Bogdan Cautis

Télécom ParisTech - CNRS LTCI, Paris, France

{firstname.lastname@telecom-paristech.fr}

ABSTRACT

We report in this short paper results on inferring a signed network (a "web of trust") from user interactions. On the Wikipedia network of contributors, from a collection of articles in the politics domain and their revision history, we investigate mechanisms by which relationships between contributors - in the form of signed directed links - can be inferred from their interactions. Our preliminary study provides valuable insight into principles underlying a signed network of Wikipedia contributors that is captured by social interaction. We look into whether this network (called hereafter WikiSigned) represents indeed a plausible configuration of link signs. We assess connections to social theories such as *structural balance* and *status*, which have already been considered in online communities. We also evaluate on this network the accuracy of a learned predictor for edge signs. Equipped with learning techniques that have been applied before on three explicit signed networks, we obtain good accuracy over the WikiSigned edges. Moreover, by cross training-testing we obtain strong evidence that our network does reveal an implicit signed configuration and that it has similar characteristics to the explicit ones, even though it is inferred from interactions. We also report on an application of the resulting signed network that impacts Wikipedia readers, namely the classification of Wikipedia articles by importance and quality.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data Mining*

General Terms

Algorithms, Experimentation

Keywords

Online communities, social applications, web of trust, signed networks, Wikipedia

1. INTRODUCTION

We are witnessing today the rapid emergence of large online communities that contribute and share content on the Web. These are essentially collaborative efforts oriented towards building repositories of quality user-generated content. Examples of applications include online encyclopedias (Wikipedia[1], Knol), photo sharing sites (Flickr) or rating sites (Epinions). An important trend in such platforms aims at exploiting user relationships, links between users (e.g., social links), in order to improve core functionalities in the system. For instance, search, recommendation or access control can benefit from socially-driven approaches. This is especially the case when links can be viewed as being *signed*, indicating a *positive* or *negative* attitude; possible meanings for positive links could be trust, friendship or similarity, while for negative links they could be distrust, opposition or antagonism. In settings where explicit relationships do not exist, are sparse or are inadequate indicators of one's attitude towards fellow members of the community, it becomes thus important to uncover *implicit* user inter-connections, positive or negative links, from relevant user activities and their interactions.

We present in this short paper a study of the Wikipedia network of contributors. For a collection of 320 articles from the politics[2] domain, starting from the revision history, we investigate mechanisms by which relationships between contributors - in the form of signed directed links - can be inferred from their interactions. We take into account *edits* over commonly-authored articles, activities such as *votes* for adminship, the *restoring* of an article to a previous version, or the assignment of *barnstars* (a prize, acknowledging valuable contributions).

Our model for user relationships is a local one: for a given ordered pair of members of the online community - called in the following the link *generator* and the link *recipient* - it will assign a positive or negative value, whenever such a value can be inferred. This could be interpreted as subjective trust / distrust in a contributor's ability to improve the Wikipedia, and we call the set of such values in the network the "web of trust". In short, our approach aims at converting interactions into indicators of user affinity or compatibility: to give a brief intuition, deleting one's text or reverting modifications (backtracking in the version thread) would support a negative link, while surface editing text or restoring a previous version would support a positive one.

Our preliminary study provides valuable insight into principles underlying a signed network of Wikipedia contributors that is captured by social interaction. We look into whether this network, denoted in the rest of the paper as WikiSigned, represents indeed a plausible configuration of link signs. First, we assess connections to social theories such as *structural balance* and *status*, which have already been considered in online communities [8]. Second,

*The networks used in this paper are available at http://www.infres.enst.fr/wp/maniu/www2011-datasets/. The Wikipedia datasets are available upon request.

[1] www.wikipedia.org.

[2] http://en.wikipedia.org/wiki/Wikipedia:WikiProject_Politics

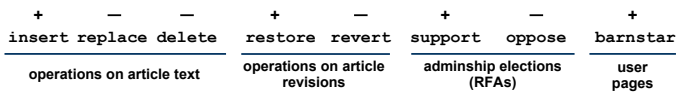

Figure 1: The interaction vector (from generator to recipient).

we evaluate on WikiSigned the accuracy of a learning approach for *edge sign prediction*. This amounts to exploiting existing links - in particular *link triads* - to infer new links and could be viewed as *propagation* of signed relationships. Equipped with the learning techniques that have been applied in [7] on three explicit signed networks, namely

- Slashdot (friend-foe tags),
- Epinions (trust-distrust tags),
- Wikipedia adminship votes (support-oppose tags)

we obtain good accuracy over the WikiSigned network (slightly better than the one achieved in [7] over a Wikipedia votes network). By cross training-testing we obtain strong evidence that our network does reveal an implicit signed configuration and that these networks have similar characteristics at the local level, even though WikiSigned is inferred from interactions while the other networks are explicitly declared.

There are many opportunities that present to us for exploiting such a network at the application level, e.g., in the management tasks of contributors. But we chose to discuss in this paper one application that also impacts the readers, namely the classification of Wikipedia articles by importance and quality. The intuition here is that such article features depend on how contributors relate to one another.

A core contribution of this paper is a thesis: user interactions in online social applications can provide good indicators of implicit relationships and should be exploited as such.

Main related work. To the best of our knowledge, this is the first study on inferring a signed network (a "web of trust") directly from user interactions. The work that is closest in spirit to ours uses a semi-supervised approach and existing links to build a predictor of trust-distrust from interactions in Epinions [9]. Several papers deal with edge sign prediction using an existing network, among which [4, 7] (see also the references therein). These approaches use the explicit signed network, either for verifying the accuracy of the predictor or as a basis for the inference of new links. In [3], the authors deal with interactions between contributors of Wikipedia articles, using the concept of an "edit network" to measure the degree of polarization in articles. In [2, 1], a contributor reputation system and a measure of trustworthiness of text are derived based on their interactions over Wikipedia content. Another paper that experiments with reputation systems using the editing interactions between contributors is [6].

2. METHODOLOGY

The Wikipedia dataset. We extracted the full revision history of 320 articles, giving us 442297 revisions by 105177 contributors. A contributor to a revision can do one of the following: edit the text or revert to a previous revision. In the case of text modification, we track several metrics: the amount of text *inserted* near the text of another contributor, the amount of the text *replaced* and the amount of text *deleted*. We also track the *count* of these edit interactions on text, i.e. the number of revisions the respective contributors interacted upon.

For each revision, we establish ownership at word level based on the text difference between two consecutive revisions of an article. The interaction thus formed is between the author of the current revision and the owners of the text in the previous revision. For revisions that are restored (reverting to a previous revision), we track the author of the restored (previous) revision and the authors of the revisions that were discarded in the process. For a given pair of contributors, we then track the number of revisions that have been *restored* and the number of revisions that have been *reverted* (i.e., discarded).

For deriving the election votes, we retrieved the adminship elections and the votes (positive/negative) submitted by our contributors, thus deriving two measures: the number of *support votes* and the number of *oppose votes* between a pair of contributors.

Finally, *barnstars* are "prizes" that users can give to each other for perceived valuable contributions; barnstars are usually present on the prize receiver's page. We have thus retrieved the user profile pages of all our contributors and extracted this information.

In order to quantify the interactions between the authors, we had to establish a list of authorship of the text for each revision. This presents itself as a list of triples of the form $(auth, pos_{from}, pos_{to})$ for each revision R. This list is created from the differences between two subsequent revisions R^{t-1} and R^t between authors $auth^{t-1}$ and $auth^t$; using a text difference tool that outputs the numbers of words (and old/new positions of these characters) that were inserted, deleted, replaced or kept we will construct the author list as follows: for text inserted and replaced the new author is $auth^t$ and pos_{from} and pos_{to} are the new positions resulted from the diff tool; for deleted text, the author is removed for the words that were deleted in R^t.

Building the signed network. We use an *interaction vector* as the basis for inferring signed edges between users. This vector contains measures capturing the four types of interactions: text edits, reverts on article versions, adminship votes and barnstars. We describe next how these are further organized and then interpreted as positive or negative units.

Figure 1 shows the components of an interaction vector and the sign interpretation of each (positive or negative). For instance, for edits on text we interpret inserts as positive interactions while replacements and deletions of text as negative and we keep their respective counts. This component will be non-empty only if there were at least $k = 3$ text interactions.

Then, the restores of a revision are interpreted as positive interactions, while conversely the reverts of a revision are negative ones.

The votes cast in an election are interpreted accordingly as positive or negative interactions while barnstars are seen as positive interactions.

Note that these vectors denote *directed interactions*, from a generator to a recipient, hence the presence of interactions in one direction does not necessary imply that interactions in the other direction exist. We obtained in this way 800057 vectors, in which participate 42631 adminship votes and 2913 barnstars.

To infer a signed edge from these interactions, we adopt the following straightforward heuristic. We decide for each of the four dimensions of interactions whether it is overall a positive or a negative contribution by considering the sign that is more present (in the case of barnstars, it can be either positive or non-existent). Then, at vector level, the result (the sign of the edge) is given by a simple voting mechanism (i.e., the sign of the sum over the dimensions). Note that this may not result in a signed edge, when the various dimensions cancel out.

The WikiSigned network obtained in this way has 71770 nodes and 463312 edges, of which 85.93% are positive (a link proportion

	nodes	edges	positive	negative
Wikipedia k=2	80,133	720,353	86.40%	13.60%
Wikipedia k=3	71,770	463,312	85.93%	14.07%
Wikipedia k=4	67,603	351,623	85.03%	14.97%

Table 1: Web of trust statistics.

triad	count	P(+)	lrn	bal	stat
t_1	1,818,176	0.97	0.1571		
t_2	164,846	0.90	-0.2020		
t_3	1,888,239	0.94	0.0183		
t_4	99,776	0.92	0.0866	X	
t_5	136,277	0.53	-0.3393		
t_6	25,308	0.40	-0.3349	X	
t_7	83,154	0.44	-0.4121		
t_8	24,373	0.54	-0.1448	X	
t_9	2,866,505	0.93	0.0148		
t_{10}	98,883	0.76	-0.2137		
t_{11}	283,149	0.83	-0.0177	X	
t_{12}	43,036	0.78	-0.1400		
t_{13}	171,551	0.91	0.0306	X	
t_{14}	99,135	0.80	-0.0750	X	
t_{15}	62,732	0.83	0.0389	X	
t_{16}	16,223	0.58	-0.0822	X	X

Table 2: Triad statistics. 'X' marks contradiction with theory.

that is very similar to the ones of the existing signed networks). One observation we can make: our mined election network did not skew the results, seeing that we have extracted 42631 votes which represent less that 10% of the links of WikiSigned. The link proportion of this network is consistent with the other studied signed networks. In Table 1 we give for comparison the resulting networks for other possible values of k (the minimum required number of text interactions so the text interactions part of the interaction vector is taken into account).

3. VALIDATION AND RESULTS

Global properties of WikiSigned. We first analyze the global properties of WikiSigned, checking whether it represents indeed a plausible configuration of link signs. For that, we study the role of "edge triads" in our signed network (similar to [8, 7]).

A triad represents a composition of a link from A to B and the possible links between them and third party nodes C. Depending on the direction and sign of the link connecting A, respectively B, with C, there are sixteen such types of triads. As in [8], they are encoded as follows: for the $A - C$ link we encode by 8 if the link is backward and by 4 if the link is negative; for the $C - B$ link we encode by if the link is backward and by 1 if the link is negative. The sum of these elements plus one gives us the triad number (ranging from 1 to 16). For example, t_6 is an encoding of "the enemy of my enemy is my friend", t_1 an encoding of "trust transitivity", t_9 is a triad in which C points positively to both A and B.

We looked at the distribution of link triads and at the proportion of positive $A - B$ links in each type of triad. We found that both measures calculated on WikiSigned are very similar with the results in [8] (see Table 2 columns *count* and *P(+)*).

Next, we study the configuration of our network in comparison with two social theories, *status* and *balance*. These theories define the formation of links between individuals. *Structural balance theory* posits that triads which are "balanced" (three or one positive sign, disregarding the direction of the links), are more prevalent in real-world networks that the other types of triads [5]. *Status* posits that a directed negative link between A and B means that A regards B as being of lower "status", while a positive link signifies higher "status" [4]. For instance, for the network to be "in line" with balance theory, in triads t_1, t_3, t_6, t_8, t_9, t_{11}, t_{14} and t_{16} the $A - B$ link should be positive and negative in the rest. For status, triads t_1, t_4, t_{13}, t_{16} should have a positive $A - B$ link and triads t_6, t_7, t_{10}, t_{11} should have negative link. For the rest of the triads, status does not predict a link sign.

In order to understand how the WikiSigned network relates to the theories of status and balanced, we used regression learning in order to compare the signs of the learned coefficients for each type of triad with the prediction of status and balance. The feature vector for each link in the network consists, for each type of triad, of the number of the respective triads that involve the link. We perform this comparison on WikiSigned, counting the contradictions i.e. the differences between the sign of the learned coefficients for each triads and the prediction of the two social theories (the same methodology has been used in [7]). We find that at a global level our interaction-based network is strongly consistent with the theory of status (only one contradiction with status in t_{16}), in line with the study on the Wikipedia election network (see Table 2, columns *lrn, bal* and *stat*; X marks a contradiction with a social theory).

Local properties of WikiSigned. Using the same methodology as [8], we consider then the problem of predicting link signs. More precisely, we run logistic regression learning with 10-fold cross-validation based on a feature set consisting of the number of triads of each type in which the link participates. For that, we use a randomly selected (via reservoir sampling) balanced dataset of links involved in at least 10 triads.

The predictive accuracy we obtained for the WikiSigned network is of 0.822, with an AUC of 0.899.

Furthermore, we have also applied the same learning methodology over the three explicit networks considered in [7], asking the following question: how well a predictor learned on one network performs when applied on another network (see Table 3). First, one can notice that our results that use and apply to explicit networks are almost identical to the ones reported in [8]. Then, WikiSigned performs comparable (slightly better) to the election network, in that prediction on itself is worse than self-prediction over Epinions and Slashdot, while learning the predictor on WikiSigned and applying it on both Epinions and Slashdot yields good prediction rates, while the inverse performs slightly worse. All this indicates that these networks have quite similar characteristics at the local level, even though our WikiSigned network is inferred from interactions while the other three are explicitly declared by users.

Exploiting WikiSigned at the application level. We also investigate the usefulness of having the signed network in applications, by considering how link structure can be exploited in the classification of articles. There are two article features that are explicit in the Wikipedia politics portal: the quality and the priority. In our dataset, we have articles that span the top 3 article qualities (Featured Articles, A-class Articles, Great Articles) and all the importances (Top, High, Mid, Low).

For our experiments we have used separate datasets for predict-

	Epinions	Slashdot	Elections	WikiSigned
Epinions	0.926	0.905	0.787	0.727
Slashdot	0.929	0.806	0.792	0.732
Elections	0.922	0.895	0.814	0.733
WikiSigned	0.889	0.844	0.784	0.822

Table 3: Predictive accuracy in training on the row data and testing on the column data. The first three networks are the datasets used in [7].

type	Importance	Quality
contributors	0.683	0.566
contribs.+links	0.740	0.779
incoming pos+neg	0.683	0.500
outgoing pos+neg	0.740	0.574
inside pos+neg	0.700	0.676
all pos+neg	0.807	0.750

Table 4: Predictive rates for the article importance and quality.

ing the most populated classes of articles (FA and GA) and importances (Medium and Low). For predicting the importance, we have used 300 articles quasi-balanced between the two classes (133 mid-priority articles). For predicting the quality, we have 136 articles equally balanced between FA-class and GA-class articles.

We have used a set of 10 features for each article: the number of *authors* plus three features (total, positive and negative) for each of: *outgoing links* (links from the authors towards other contributors), *incoming links* (the links from other contributors towards the authors) and *inside links* (links from authors to authors).

We report the predictive accuracy we obtained via logistic regression in Table 4. Following the intuition that more important articles have a larger participation and thus more links, we tested the predictive power of these two values (*contributors* and *contribs.+links*). While using knowledge about positive or negative links in separation does not provide better accuracy, their combination yields significantly better results (*all pos+neg*). This suggests that the quality of an article is not defined solely by its authors, but also by the relationships between contributors and the rest define the importance of an article. One remark we can make is that, while predicting the quality of an article using the link structure performs worse than using only the number of contributors and their links, the results are comparable.

4. FUTURE WORK

These preliminary results encourage us to continue towards more refined heuristics and richer Wikipedia data. We believe that better heuristics for deciding the sign of a link can further improve the derived network. Having a dataset that includes all article qualities could improve the utility of WikiSigned for article quality classification. We plan to use the link prediction methods validated by our results to further enrich the WikiSigned network. At the application level, one goal is to establish and exploit a reputation system for contributors, for example based on exponential ranking on the derived links (while also taking into account the negative links). Another goal is to propose a text-trust system similar and comparable with [1].

5. ACKNOWLEDGEMENTS

This work was supported by the French ANR project ISICIL and the EU project ARCOMEM FP7-ICT-270239.

6. REFERENCES

[1] B. T. Adler, K. Chatterjee, L. de Alfaro, M. Faella, I. Pye, and V. Raman. Assigning trust to Wikipedia content. In *WikiSym*, 2008.

[2] B. T. Adler and L. de Alfaro. A content-driven reputation system for the Wikipedia. In *WWW*, 2007.

[3] U. Brandes, P. Kenis, J. Lerner, and D. van Raaij. Network analysis of collaboration structure in Wikipedia. In *WWW*, 2009.

[4] R. Guha, R. Kumar, P. Raghavan, and A. Tomkins. Propagation of trust and distrust. In *WWW*, 2004.

[5] F. Heider. Attitudes and cognitive organization. *Journal of Psychology*, 1946.

[6] S. Javanmardi, C. Lopes, and P. Baldi. Modeling user reputation in wikis. *Stat. Anal. Data Min.*, 2010.

[7] J. Leskovec, D. Huttenlocher, and J. Kleinberg. Predicting positive and negative links in online social networks. In *WWW*, 2010.

[8] J. Leskovec, D. Huttenlocher, and J. Kleinberg. Signed networks in social media. In *CHI*, 2010.

[9] H. Liu, E. Lim, H. W. Lauw, M. Le, A. Sun, J. Srivastava, and Y. Kim. Predicting trusts among users of online communities: an Epinions case study. In *EC*, 2008.