User-User Relationship on Reddit: A Paper Replication Experiment

Felix Grezes
Speech Lab @ Queens College
CUNY Graduate Center
fgrezes@gradcenter.cuny.edu
May 8th - 2015

Abstract

This project paper explores the results and reproducibility of said results of the 2012 by Hassan, A., Abu-Jbara, A., and Radev, D. titled *Detecting Subgroups in Online Discussions by Modeling Positive and Negative Relations among Participants* (Hassan et al., 2012). While the paper claimed that subgroups of internally consistent opinions could be automatically detected on online discussion forums, they did not provide access to their dataset, nor their software. This is chronic issue in computational science papers. This project worked on a new dataset, the Reddit Dataset.

1 Introduction

The main novelty proposed in the original paper was to work at the level of relationship between individuals, as opposed to labeling the sentiment or polarity of words, sentences, or individuals. The work was done on text data extracted from online discussion forums, precisely the Wikipedia editor discussions and the PoliticalForum.com and CreateDebate.com forums. Four different tasks were performed on this data:

1. Automatically classify messages from one user to another as being targeted to this other user. Not all message from user to another carry any sentiment toward the relationship between these users.

2. Automatically classify relationship as having a positive or negative polarity. Users with a positive polarity should be in agreement over most topics.

3. Test the validity of the above polarity classification using structural balance theory.

4. Partition the discussion into subgroups of users with internally consistent relationships. Users within a subgroup should have mostly positive polarity relationships, and users from different subgroups should have mostly negative polarity relationships.

For these tasks, Hassan et al. used the popular Amazon Mechanical Turk to obtain a large volume of annotations on the data set.

1.1 Project Goals

Unfortunately, as previously noted the data is not publicly available. To test the reproducibility of the paper, we created a new data set, extracted from the popular Reddit.com website. Because we did not have the time to crowd-source the manual labeling of the data, the first two tasks of the paper are impossible to reproduce. However, the structure of discussions on Reddit is such that users choose to reply to specific comments left by other users. Because of this, we posited that the vast majority of text on Reddit contains directed sentiment, and can be used infer the polarity of user-user relationship. Similarly we cannot test the validity of polarity predictions against gold labels, but we are able to use the psychologically rooted structural balance theory to obtain a baseline validation.
Thus the goals of this project are:

1. Transform the raw XML-tree discussion data from Reddit into a network based architecture, with users as nodes and relationships as edges.

2. Apply OpinionFinder (Wilson et al., 2005) on messages between users and use the results to extrapolate the positive/negative polarity of the relationship.

3. Test if the labeled network satisfies the structural balance theory.

4. Implement the greedy approximation algorithm described in the original paper for subgroup detection in the labeled network.

2 Related Work

In addition to research related to the subject of the project, I have explored works that have cited the original paper, as well as works on the meta-goal of the project i.e. computational paper reproducibility.

2.1 Classical Sentiment Analysis

The original paper by Hassan et al. had the goal of mining the sentiment between two individuals, which make it related to all the work in the large and well established field of computational sentiment analysis. In 1997 McKeown et al (McKeown et al., 1997) worked on predicting the polarity of specific words. Since then many including Hatzivassiloglou et al. (Hatzivassiloglou et al, 2000) have worked on analysis the sentiment at the sentence level, in this case with the idea of analysis the subjectivity or objectivity of sentences.
This work differs from these approaches by generalizing the previous ideas to the level interaction between individuals, based on the multiple publicly available messages they exchange.

2.2 Stance Classification

Another goal of the paper is to identify subgroups and communities within the discussions forums studied. This is closely related to the field of stance classification. In single-topic discussions, Tan et al. (Tan et al., 2011) showed that a user's word polarity was often enough information to correctly classify him into the group of users sharing his views. However in more complex discussion involving many topics (healthcare, Obama, abortion, welfare, race, etc.) the simple word polarity is not sufficient, in fact more than 2 opposing groups may exist.

2.3 Social Network Extraction

While the idea of automatically extracting networks from data is well established, see the work of Matsuo et al. (Matsuo et al., 2007) in 2007; doing so exclusively from text is not common. In 2010 Elson et al. (Elson et al., 2010) worked on on texts literary fiction to extract the network of conversing characters. This work differs from these by using human conversation text data, and by assigning a positive/negative sign to the network connections.

2.4 Structural Balance Theory

The structural balance theory is a psychological theory that tries to explain the dynamics of signed social interaction, first stated by Heider (Heider, 1946). Structural balance theory is rooted in the 'fiend of my friend' and 'enemy of my enemy is my friend' principle. In effect this states that relationship triangles in the network should contain and odd number of positive edges, any other configuration contains some kind of inconsistency ('the friend of my enemy is my friend'). This has been verified empirically by Leskovec (Leskovec et al., 2010).

2.5 Citing Papers

Since this project is an attempt at replicating the results of Hassan et al., I looked at the papers citing it. This gives us an idea how the work is currently being used. The most prolific author citing the original paper is Pr. Jing Jiang of the School of Information Systems at Singapore Management University. In her 2013 paper (Qiu et al., 2013) she examined methods called collaborative filtering and probabilistic matrix factorization that allowed her to predict the sing of the relations between users that never interact, which produce a much denser network. Her other works (Gottipati et al., 2013; Qui et al., 2013) citing Hassan et. al are most closely related to stance classification, in which the number and nature of the subgroups to be detected is known in advance; and with the goal of addressing the usual sparseness of the connectivity matrix that results from online discussion datasets.

2.6 Work on Replication

The development of computational research in the past 25 years has unfortunately not lead to a similar development of code and dataset sharing by the computational scientific community. My interest in reproducible research was sparked when attending Pr. Victoria Stodden's course at Columbia University. Her work with David Donoho (Donoho et al., 2009) aims to be an template for future reproducible computational research by offering open access to both the datasets used and the software created. Similar work include the RunMyCode.org (Stodden et al., 2012) and the IPOL.im (Image Processing On Line) (IPOL,) websites that allows authors to showcase the code tied to their papers' results, and lets peers rerun the computational experiments.

3 Reddit Data

3.1 Structure

The Reddit data is extracted in XML-tree form, based on messages replying to other messages, and contains a wealth of meta-data for each messages such as time, score (on Reddit users are incentivized to rate other people's messages), karma (how popular a user is). The data we extracted was from the the /r/Worldnews and the /r/Politics sub-forums. We chose these domains because the conversations should have a high negative or positive sentiment content, and not much neutral content. Currently the data consists of over 750K messages, spread over 25K different discussion threads, with 150K users, and this data continues to grow as the crawlers continue to run. For comparison, the original paper had 1.2M messages.

3.2 Data Exploration

To get a better understanding of the data, we performed a cursory exploration of the Reddit data. Figures 2 and 3 show a sample of the results. The inverse exponential pattern is common in social networks observed and modeled (Robins et al., 2007).

4 Project

4.1 Code & Data Distribution

All my code is available through my github repository found at

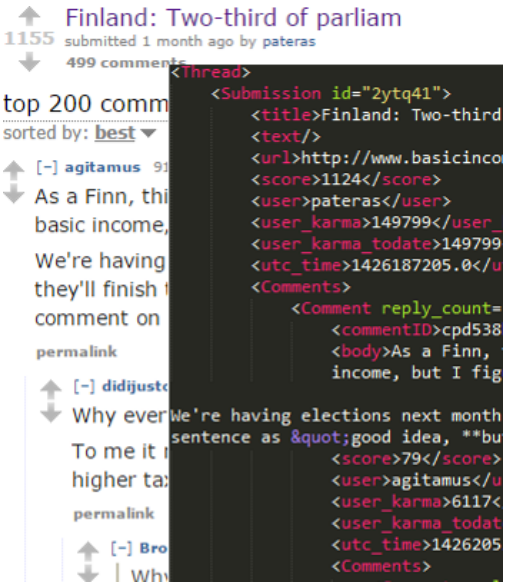

Figure 1: Visualization of the XML-tree structure of extracted from a Reddit Discussion

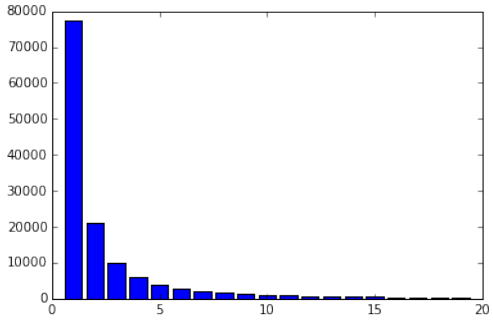

Figure 2: Distribution of the number of threads per user
The largest number of threads a user has posted in is 1782

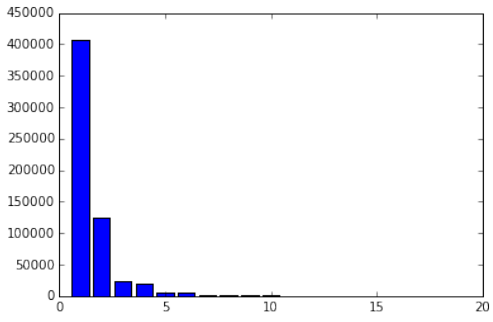

Figure 3: Distribution of the number of messages in a single thread per user
Maximum posts by a single user in a thread: 60

github.com/grezesf/Research/tree/master/reddit and can be free shared by the scientific community. That said the code is not universal in the sense that any researcher who wants to run it will need to adapt it slightly to be able to run it on their machine, installing proper dependencies and changing directory names.
The code is written in Python, and uses the notebook tool to display the results alongside the code. Notebook viewers, such as the online tool nbviewer.com, can convert the code-result hybrid to an HTML page viewable in a web browser.
The scripts included are described below:

- Data-Exploration This script transforms the data from XML-tree structure to a network of users connected by messages. It also performs simple statistical measures of the resulting network.
- Network-Graph-Visualization This script uses the NetworkX library to graphically visualize the above network.
- Network-Pickling This scripts saves the network in Pythong-pickled form, to allow others scripts to easily load them.
- Structural-Balance This scripts performs the two experiments of the project, automatically labeling the edges of the network, comparing the results to structural balance theory, and detecting subgroups within the network.
- Reddit-Tools This is a library of functions that the other scrips use.

Ideally, all computational papers would provide open access their data, and their code would be publicly available for the community to replicate the results. While this does add additional work on the authors, the effort has been shown to lead to more citations on such papers (Stodden, 2009), in addition to the purely scientific benefits.
The data is freely available at speech.cs.qc.cuny.edu/∼felix/Reddit_Data, in .tar compressed form. The File is approximatively 115M large, and contains the web crawler script that extracted the data from the web.

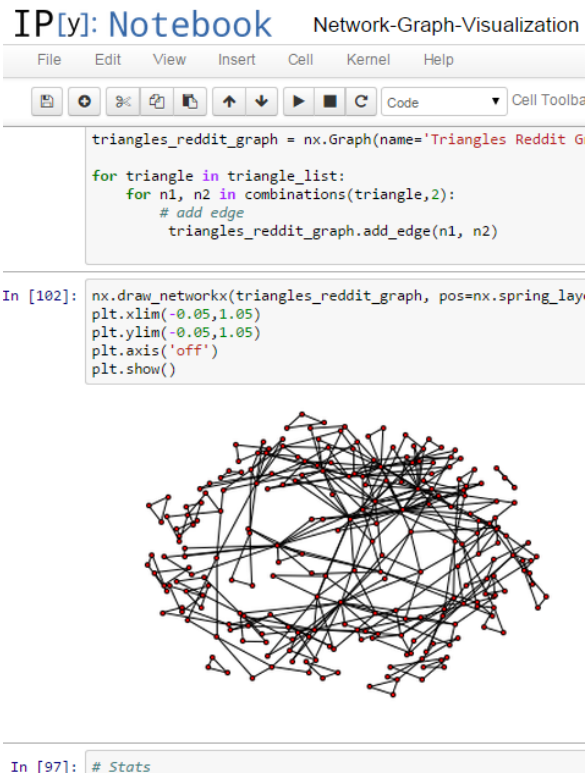

Figure 4: Example of python code and it's outcome displayed together using iPython Notebook

4.2 Results

The structural balance experiment results are shown in table 1, comparing to the results of Table 4 in the original paper. We observe mixed results, with the + - - balanced triangles appearing more often than chance, the + + - unbalanced triangle appearing less often which is in accordance with structural balance theory. However the two other triangles types have the opposite behavior. This could be caused by the fact that most edge polarities are negative, as we maybe should have expected from online discussions, which explains why the + + + is so much rarer in the labeled network. Overall this result cannot disprove the original paper's result, and differences in dataset are certainly the cause of any discrepancies.

Exctracted Network				Random Network			
+++	++ -	+ - -	- - -	+++	++ -	+ - -	- - -
1.0	11.7	51.1	36.2	11.4	36.5	38.6	13.5

Table 1: Percentage of different types of triangles in the extracted networks vs. the random networks.

The second experiment of this project is re-implementing the sub-group detection approximation algorithm described in the original paper. While the algorithm was adequately documented, the baseline comparisons were not, making it difficult to interpret the results of our implementation.
Upon manual inspection, the algorithm always suggested the largest number of subgroups as optimal, 5 subgroups in this case. This might be cause by the large number of relationships labeled as negative, leading to smaller subgroups with internal agreement. For the results to be interesting, one would need to find the first value for which the number of subgroups is too large, and agreeing users are forced into separate subgroups.

5 Conclusion & Future Work

This project showed that the most interesting result from the original paper by Hassan et al., that automatic relationship polarity labeling based sentiment analysis of messages between users of online discussions, is compatible with the previously tested psychology theory of structural balance. However it also highlighted the difficulties in attempting to replicated full results from computational papers.

In the near future I plan to finish cleaning the code and properly documenting the functions, as well as provide a proper Readme.txt file for people that would be interested in working with the code and/or dataset.

Research-wise, there are many questions that this dataset can help answer, for example "What makes users have high-karma? Is it text based or network-connectivity related?". This type of question is interesting to those who wish to understand a facet of human-to-human online interactions.

Acknowledgments

I would like to thank Denys Katerenchuk of SpeechLab@QueensCollege for writing the scripts that crawled and downloaded the data from the Reddit website.

References

A. Hassan, A. Abu-Jbara, and D. Radev, "Detecting subgroups in online discussions by modeling positive and negative relations among participants," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 59–70, Association for Computational Linguistics, 2012.

V. Hatzivassiloglou and J. M. Wiebe, "Effects of adjective orientation and gradability on sentence subjectivity," in *Proceedings of the 18th conference on Computational linguistics-Volume 1*, pp. 299–305, Association for Computational Linguistics, 2000.

C. Tan, L. Lee, J. Tang, L. Jiang, M. Zhou, and P. Li, "User-level sentiment analysis incorporating social networks," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1397–1405, ACM, 2011.

Y. Matsuo, J. Mori, M. Hamasaki, T. Nishimura, H. Takeda, K. Hasida, and M. Ishizuka, "Polyphonet: an advanced social network extraction system from the web," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 5, no. 4, pp. 262–278, 2007.

D. K. Elson, N. Dames, and K. R. McKeown, "Extracting social networks from literary fiction," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 138–147, Association for Computational Linguistics, 2010.

M. Qiu, L. Yang, and J. Jiang, "Mining user relations from online discussions using sentiment analysis and probabilistic matrix factorization," 2013.

S. Gottipati, M. Qiu, L. Yang, F. Zhu, and J. Jiang, "Predicting users political party using ideological stances," in *Social Informatics*, pp. 177–191, Springer, 2013.

M. Qiu and J. Jiang, "A latent variable model for viewpoint discovery from threaded forum posts," 2013.

D. L. Donoho, A. Maleki, I. U. Rahman, M. Shahram, and V. Stodden, "Reproducible research in computational harmonic analysis," *Computing in Science & Engineering*, vol. 11, no. 1, pp. 8–18, 2009.

V. Stodden, C. Hurlin, and C. Pérignon, "Runmycode. org: a novel dissemination and collaboration platform for executing published computational results," in *E-Science (e-Science), 2012 IEEE 8th International Conference on*, pp. 1–8, IEEE, 2012.

"Image Processing On Line." `http://www.ipol.im`.

F. Pérez and B. E. Granger, "IPython: a system for interactive scientific computing," *Computing in Science and Engineering*, vol. 9, pp. 21–29, May 2007.

V. Hatzivassiloglou and K. R. McKeown, "Predicting the semantic orientation of adjectives," in *Proceedings of*

the 35th annual meeting of the association for computational linguistics and eighth conference of the european chapter of the association for computational linguistics, pp. 174–181, Association for Computational Linguistics, 1997.

Wilson, Theresa and Hoffmann, Paul and Somasundaran, Swapna and Kessler, Jason and Wiebe, Janyce and Choi, Yejin and Cardie, Claire and Riloff, Ellen and Patwardhan, Siddharth, *OpinionFinder: A system for subjectivity analysis*, Proceedings of hlt/emnlp on interactive demonstrations, 2005.

Heider, Fritz. "Attitudes and cognitive organization." The Journal of psychology 21.1 (1946): 107-112.

Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. 2010a. Predicting positive and negative links in online social networks. In Proceedings of the 19th international conference on World wide web, pages 641650, New York, NY, USA.

Robins, Garry, Pip Pattison, Yuval Kalish, and Dean Lusher. "An introduction to exponential random graph (p*) models for social networks." Social networks 29, no. 2 (2007): 173-191.

Stodden, Victoria. "Enabling reproducible research: Open licensing for scientific innovation." International Journal of Communications Law and Policy, Forthcoming (2009).