Social Network Analysis

Stefan Dimitrov
School of Computer Science
McGill University
Montreal, Quebec, Canada
stefan.dimitrov@mail.mcgill.ca

***Abstract*—This paper surveys a number of existing works on the topic of Social Network Analysis and presents some of the most essential concepts in the field. We focus on issues such as the definition of a Social Network and Social Network Analysis, important theories from sociology and psychology applicable to social networks in the digital age, the concept of influence and its importance for the analysis of social networks, and the buzzword technique known as Viral Marketing.**

***Index Terms*—Social, network, influence, viral, marketing**

I. Introduction

With the growing popularity of mobile devices and the omnipresence of high speed Internet connections, online social networks such as Twitter and Facebook have emerged to connect individuals and organizations. Thus, it is of great interest to study the underlying theories which explain the formation, structure and functioning of social networks. Sociologists and psychologysts have made great progress in modeling human behaviour and relationships, the balance of power and influence in groups and the ideas powering viral information propagation in society. This paper aims to investigate the applicability of these existing theories to the recently emerged online social network services. Gaining better understanding of the science behind these services will facilitate future work to solve important problems in social networks such as the identification of malicious or automated accounts publishing undesired commercial messages (SPAM) in social networks.

The following section defines the most important concepts for the study of social networks, such as actors, relations and centrality. Then we focus on influence in social networks and viral marketing. Finally, we conclude with a summary of our study.

II. Definition

Social networks have been studied in society long before the existance of Internet. For example, Wikipedia defines a Social Network as: ”a social structure made up of a set of social actors (such as individuals or organizations) and a set of the dyadic ties between these actors,”[2] and this definition can also explain an online social network like Twitter or Facebook (fig. 1).

Graph Theory is often applied to model social networks. In this case we define actors (vertices), see fig. 2, as individuals or organizations that participate in relationships. Again, this definition is closely related to the one given by Wasserman et al. in 1984 with respect to social and behavioural sciences: ”Actors are discrete individual, corporate, or collective social units”[4].
We can say that online social networks closely mirror the structure of the existing relationships in society. Thus, a relationship between two nodes can be modelled as a an edge in a graph. It is also known as a tie, or the link between a pair of actors [4]. Depending on the social network, the link can be directed (eg. a follower on Twitter) or undirected (eg. a friend on Facebook). It can also be signed or unsigned; mutual or not.

Two nodes examined together with the relationship between them constitute a dyad (fig. 3).

When we consider a third node in addition to the existing two, we get a group of three actors which is known as a

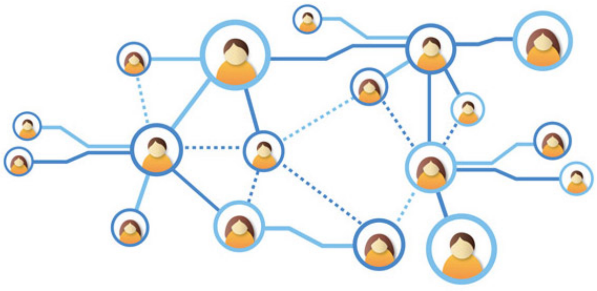

Fig. 1. A graph of a social network. The thikness of lines represents the strength of relations between actors. [1]

”triad”. The concept of triad is important because it allows us to study the balance between the links in a signed network (Balance Theory) as well as the transitivity and network closure well known in the fields of sociology and psychology (fig. 4).
These theories have application to online social networks for the purpose of discovering links which are not explicit (eg. suggesting friends on Facebook or LinkedIn) and for recommending content based on the tastes of friends we ”like”.

A. *Centrality*

Centrality is an important concept in graph theory as it identifies the most important (or influential) vertices in a graph (fig. 5). With respect to social networks analysis, measuring centrality allows us to study the structure of the network and measures of influence that actors have.

One type of centrality is degree centrality. It is defined as the number of ties a node has; the simples way to measure centrality. In directed networks such as Twitter there are two types (fig. 6): indegree, measuring the number of ties to the node (followers), and outdegree - the number of ties from the node (following). Indegree in particular is an important concept for social networks because it expresses the popularity of a given node.
A second type of centrality measure is betweenness centrality. It can be expressed as g(v):

$$g(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

where the numerator represents the total number of shortest paths through a vertice v, while the denominator is the total number of shortest paths. This measure is particularly useful for networks where information is known to always take the shortest path. In such networks high betweenness centrality indicates high influence on the transfer of information through the network. It allows us to indetify the number of times a node acts as a bridge between two nodes.
Another type of centrality is closeness centrality, defined as the inverse of the sum of the distances from a node to all other nodes [10]. It can be useful for studies of information propagation within social networks.
A very important measure is eigenvector centrality. We can look at eigenvector centrality as a generalization of PageRank [11]. Just like PageRank measures the importance of a webpage on the Internet relative to other webpages, eigenvector centrality measures the influence of a node in a social network. It assigns assigns relative scores to nodes, with nodes of higher scores contributing more than a large number of nodes of lower scores.
Eigenvector centrality can be augmented to account for ”external influence” which is often present on social networks. For instance, a famous person on Twitter may be followed by other famous people which makes her influential as measured by eigenvector centrality, but she may also be influential because of her real life popularity. The alpha centrality of a node x can be expressed as:

$$x_i = \alpha A^T_{i,j} x_j + e_i$$

Where alpha is a tradeoff constant between links and external influence. When alpha is equal to 0 only external influence determines alpha centrality. On the other hand, as alpha approaches infinity alpha centrality becomes equal to eigenvector centrality.

B. *Structural holes*

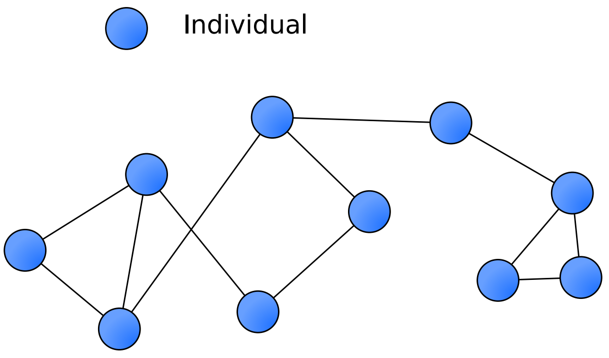

Fig. 2. An actor (vertice) can be an individual or organization. [3]

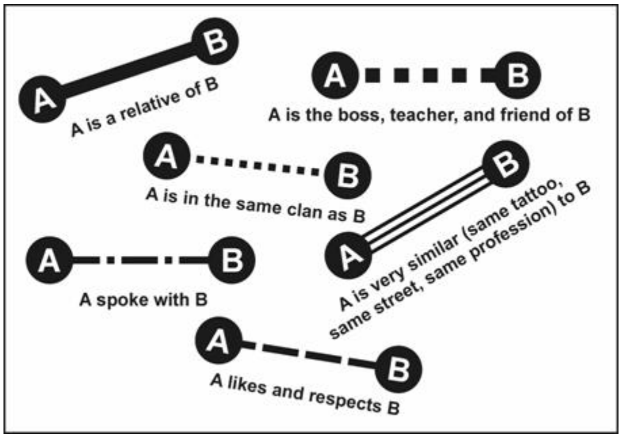
A is a relative of B
A is the boss, teacher, and friend of B
A is in the same clan as B
A is very similar (same tattoo, same street, same profession) to B
A spoke with B
A likes and respects B

Fig. 3. A dyad: two nodes and the edge between them. [5]

In networks with homogeneous vertices (e.g. students who finished the same university) we observe the formation of clusters. Clusters are characterized by strong ties between the vertices within them and fewer (or weaker) ties to vertices outside of the cluster. When two clusters contain non-overlapping information, there is a structural hole between them. Nodes ”bridging” the structural holes are called ”borkers” and can leverage social capital (fig. 7). For example, a structural hole can exist between the amlumni of an academia and the employees of a company on LinkedIn. An alumna who graduated from the academia and is currently employed by the company can act as a broker in recruiting talent for her employer and in helping her fellow alumni secure employment.

III. Influence

In preparing the dataset I first used the Google Finance feature of Google Spreadsheets to obtain end of day data for UNG since inception (2007-04-18) until 2014-09-17[1]. Then I processed the data to generate the Return, Day of Year, Season, Day of Week, Open, Overnight Change and Previous Return columns. To calculate mean temperature I downloaded all daily weather data for the New York Central Park weather station from 1763 until 2014. [2] Originally I intended to use the observed temperature at 12:00 AM, however this data was not available for recent years. Thus, to calculate the mean I averaged all minimum and maximum temperature measurements for a given day of the year. The observed temperature column I populated with the average of minimum and maximum temperature on the specific date.

$$I_i \leftarrow \sum_{j:(i,j)\in E} u_{ij} P_j$$

$$P_i \leftarrow \sum_{j:(j,i)\in E} v_{ji} I_j$$

Once I had the dataset, I attempted to implement logistic regression. Unfortunately, I was not successful. I encountered issues with float precision in Python and probably other implementation problems. I did implement a Naive Bayes classifier successfully, however it is very slow. In order to be able to present this report on time I resorted to developing a Scikit-learn based solution.[5] Thus, the rest of this section and the results discuss my findings with the aforementioned solution.

I used leave-one-out cross-validation to test the performance of my dataset with logistic regression, LDA and Naive Bayes classifiers, as well as to identify the best model dimension. I repeated the cross-validation experiment starting with only two dimensions (Day of Year and Season) and adding one dimension on every run until all 8 dimensions were included in the model, for every classifier. I also implemented a ”baseline” classifier which always chooses the predominant result from the test set.

I repeated all tests with half of the training data in order to evaluate how changing the data size affects the performance of the classifiers.

IV. Viral Marketing

A. Logistic Regression

I first performed logistic regression with half of the dataset (923 samples):

[1] https://support.google.com/docs/answer/3093281?hl=en

[2] http://www1.ncdc.noaa.gov/pub/data/ghcn/daily/by_year/

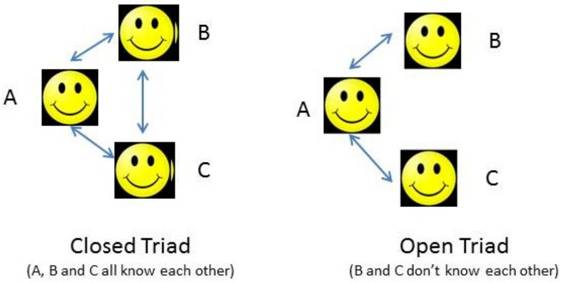
A
B
C
Closed Triad
(A, B and C all know each other)
A
B
C
Open Triad
(B and C don't know each other)

Fig. 4. A triad: three nodes and the edges between them. [6]

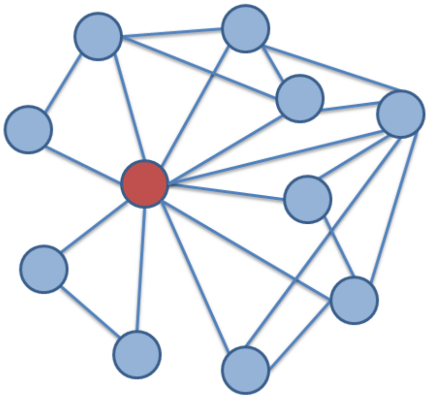

Fig. 5. An influential node with high degree centrality. [7]

As demonstrated in table 1, the model performs slightly better than the baseline (always choosing the most popular result). It appears that adding weather measurements (mean and observed temperature) helps reduce both the training error rate and the estimated true prediction error.

Next, I repeated the test with the full dataset (1847 samples), as can be seen on table 2. The results are very much consistent and show that the model of all 8 dimensions performs better than average and better than models of lower dimensions. However, my results reveal that increasing the amount of training data decreases the performance of the model, which is unexpected. It could be explained by the fact that the dataset was not randomized. Thus, there appear to be other factors which may help in predicting the direction of gas equities, and they could be an interesting topic for further research.

A slight overfitting can be observed when adding the overnight return variable to the larger dataset. What is interesting is that adding mean temperature alone does not improve performance, but adding the observed temperature does.

B. LDA

Table 3 summarizes the results of the LDA classifier. Similar results were obtained as with Logistic Regression. LDA performs even better on the smaller dataset, but performs worse than Logistic Regression on the full dataset.

In Degree
Out Degree

Fig. 6. In degree and out degree in a directional social network. [8]

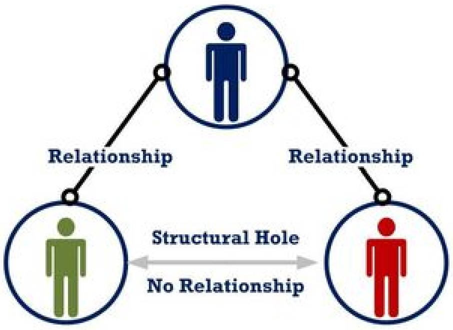

Fig. 7. A structural hole in a triad. [12]

C. Naive Bayes

Table 4 shows the results of running the scikit-learn Naive Bayes classifier. The Naive Bayes classifier performs the worst of all three evaluated in this paper. A possible reason would be that the dataset dimensions are not truly independent: season for example can be inferred from day of year.

I also implemented a Naive Bayes classifier, but it is much slower than the Scikit-learn version. Yet, the results are the same (apart for some very slight differences on the smaller dataset which I attribute to including the header line in the row count and rounding). The results of my classifier can be seen on Table 5.

While not shown, I did try various combinations of dimensions with all three classifiers and experienced similar or worse results. Therefore, based on the results above, the model with all 8 dimensions is able to predict the direction of daily returns for natural gas securities slightly better than the baseline, when used with Logistic Regression classifier (see fig.1) .

V. Conclusion

This is a first attempt at creating a dataset linking weather variations to intraday trading returns for a gas security. I chose the dimensions for this dataset based primarily on empirical observations, but also validated their relevance using the leave-one-out cross-validation method. Predicting intraday returns for any security, especially one as volatile as UNG, is a very challenging task. While the proposed dataset performs less than perfectly, it can be used as a good starting point for further research.

Gas supply is relatively fixed in the short term, because

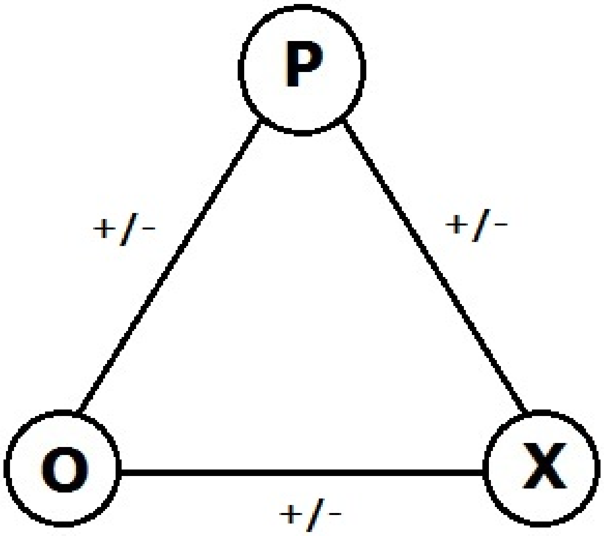

Fig. 8. Fritz Heider's P-O-X model. [14]

building wells and pipes takes time. On the other hand, temperature fluctuations affect gas demand as consumers and industries buy more gas for heating/cooling. Thus, the changes in demand with fixed supply lead to volatility in the equilibrium price. Traders of gas equities analyse supply and demand data to gain insight into the market direction. However, supply, storage and demand numbers are available only on the day when a report is released. Yet, traders need to make buy/sell decisions every day the market is open. Hence, weather appears to be the one input factor which is continuously available and affects prices.

Normally gas features are very volatile, still there are 36 days in the dataset for which the intraday return is 0.00% (open price is equal to close price). For these days I assign the intraday return to 0 (negative).

A possible future direction for research would be to consider temperature measurements throughout the day, as Chang et al. (2008) do for the overall market[4], but particularly with respect to gas equities. If enough DGAZ/UGAZ data is available, it may provide better insight into the psychology of intraday traders, due to the triple returns offered by these instruments. Other variables, such as the market demand/supply numbers, can help predict the trend for gas prices. At the same time, the day when these numbers are reported experiences unusual returns and volatility [3]. Augmenting the daily returns prediction model with the aforementioned extra dimensions could improve its performance, hence it would be a worthwhile direction of further research.

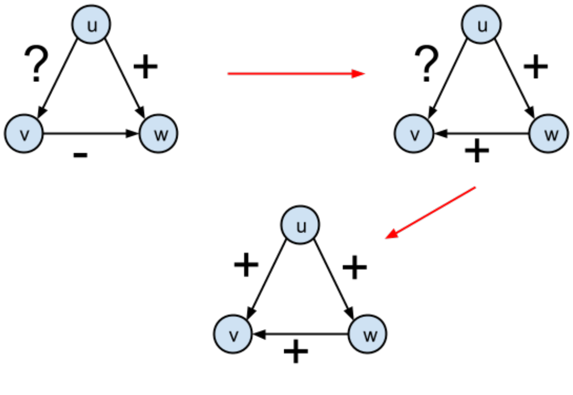

Fig. 9. Flipping the direction of an edge according to Status Theory.

Features	Epinions	Slashdot	Wikipedia
Positive edges	0.5612	0.5579	0.6983
Positive and negative edges	0.5911	0.5953	0.7114

Fig. 10. Accuracy of predicting positive and negative edges for three social networks. [13]

REFERENCES

[1] How to Use Crowdsourcing in the Classroom. Available: http://www.hollyclark.org/2013/11/03/how-to-use-crowdsourcing-in-the-classroom/

[2] Social Network. Available: http://en.wikipedia.org/wiki/Social_network

[3] An example of a social network diagram. Available: http://commons.wikimedia.org/wiki/File:Social-network.png

[4] Wasserman, Stanley; Faust, Katherine (1994). ”Social Network Analysis in the Social and Behavioral Sciences”. Social Network Analysis: Methods and Applications. Cambridge University Press. pp. 127. ISBN 9780521387071

[5] Examples of dyads. United States Army Training Support Center. Available: https://courseware.e-education.psu.edu/courses/bootcamp/lo09/08.html

[6] Tufekci, Zeynep; ”Is the Social Web Less Surprising? The Internet of People and Social Flneurism,” Available: http://technosociology.org/?p=693

[7] Weingart, Scott; ”Networks Demystified 2: Degree,” Available: http://www.scottbot.net/HIAL/?p=6526

[8] Keung, Tim; ”Social / Organizational Network Analysis: Common Terminology (Part 3),” Available: http://hr.toolbox.com/blogs/organizational-network-mapping/social-organizational-network-analysis-common-terminology-part-3-46734

[9] Betweenness centrality, Available: http://en.wikipedia.org/wiki/Betweenness_centrality

[10] Bavelas, Alex; ”Communication patterns in task-oriented groups,” J. Acoust. Soc. Am, 22

[11] Austin, David; ”How Google Finds Your Needle in the Web's Haystack,” Available: http://www.ams.org/samplings/feature-column/fcarc-pagerank

[12] Brenegar, Ed; ”Connect, Communicate & Contribute,” Available: http://edbrenegar.typepad.com/leading_questions/say-thanks-every-day/

[13] Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. 2010. Predicting positive and negative links in online social networks. In Proceedings of the 19th international conference on World wide web (WWW '10). ACM, New York, NY, USA, 641-650. DOI=10.1145/1772690.1772756 http://doi.acm.org/10.1145/1772690.1772756

[14] D Khanafiah , H Situngkir. Social Balance Theory. Revisiting Heiders Balance Theory for many agents, 2004

[15] W Galuba, S Asur, BA Huberman.Influence and passivity in social media, in Machine learning and knowledge discovery in databases, 2011, Springer Berlin Heidelberg.

[16] BA Huberman, DM Romero, F Wu. Social networks that matter: Twitter under the microscope, inarXiv preprint arXiv:0812.1045, 2008

[17] Louis Yu, SitaramAsur and Bernardo A. Huberman. ArtificialInflation: The Real Story of Trends in SinaWeibo, in arXiv preprint arXiv:1202.0327, 2011

[18] Jure Leskovec, Lada A. Adamic, and Bernardo A. Huberman. 2007. The dynamics of viral marketing. ACM Trans. Web 1, 1, Article 5 (May 2007). DOI=10.1145/1232722.1232727 http://doi.acm.org/10.1145/1232722.1232727

[19] Zarella, Dan; ”Informational Cascades Prove Tipping Points Exist,” Available: http://danzarrella.com/informational-cascades.html