Exploration Study of Retweet Propensity

Ryan Compton
rcompton@ucsc.edu

Shachi H Kumar
shulluma@ucsc.edu

Diego Rodriguez
dirodrig@soe.ucsc.edu

Abstract

Twitter, a social media platform, can be used to spread information across social networks. A critical aspect of this spread of information is user's engagment within such environments, specifically the level that a user will 'retweet'. This paper describes three experiments on observational data collected on Twitter. First, this paper will explore the possible causal structure that can be found from particular social metrics of users on Twitter along with their likelihood to retweet. A structure is found suggesting that many interactions are taking place. The second experiment will explore a subset of these interactions to find the magnitude of the effect certain social metrics have on a user's likelihood to retweet. It is found that a user's number of Followers and Statuses have a relationship to their *Retweet Propensity*, while Friends Count has no relation. Lastly, an experiment is conducted to discover if within this Twitter network, Homophily and Social Contagion effects influence retweeting. Such factors are found to very slightly predict retweeting.

1 Introduction

Social media services, such as Twitter, can be used by organizations to spread messages through online word-of-mouth communications. A critical part of such communication efforts is engagement, the sum of actions performed by the organizational followers after receiving a Tweet. Engagement is the sum of "retweets", likes and mentions received by the message sender (Zhang et. al, 2011). Engagement is important because it measures how effective the word-of-mouth communication was with the senders' followers, a gate to a much broader audience.

What is lacking is an understanding of the causal entities for engagement. While there are certain factors that have been explored correlationally (Tsugawa et. al, 2015) , (Zhang et. al, 2011) (Luo et. al, 2015), the key elements causing engagement have yet to be found. In this project we will focus on analyzing potential causes for one aspect of engagment, retweets, on the social network platform Twitter, where a retweet is a message that a user receives from another source and shares this message with their network.

To accomplish this, this project has three sections.

1. Finding any potential causal structures. The dataset used is observational which creates many challenges for discovering causal effects. It is not possible to segregate a particular potential cause during data collection, as can be done during an experimental setting. To find potential strucutres, an algorithmic procedure, the PC algorithm (Spirtes et. al, 2000), will be used.

2. Once a possible causal structure has been found, the magnitude or quantifiable influence that variables have on each other will need to be analyzed. To accomplish this, conditioning on potential effect sizes as referenced within (Rubin et. al, 2011) can be used to account for possible confounding factors.

3. Based on the previous findings of this project, possible social network effects may be occurring, this section will attempt to model

those social effects and their power to predict propensity to retweet.

The main contribution of this project is that it gives a peak into the causal structure that influences engagment behavior in social media networks.

2 Related Work

Previous work has focused on observational studies, partially due to the difficult practicalities in setting up an experiment that matches the true environment of social media. This section will describe some of these studies.

(Zhang et. al, 2011) explored the influence of business engagement on the level of consumer engagement on Twitter. The authors explored the effect of the number of business posting and the number of individuals the business follows on consumer engagement. Consumer engagement is defined by the authors as the number of posting related to the business by individuals, and the number of followers the organization has. The analysis done through a variation of the Structural Equations Modeling method, called Path Analysis, showed that business engagement is positively correlated with user word-of-mouth engagement (Zhang et. al, 2011). The project described in this paper will expand on this work by not examining volume of organizational engagement, but also bring in social metrics and network effects.

Another study that has been conducted purely on examining the content of the tweet is that of Tsugawa and Ohsaki, 2015. The authors examined tweets sentiment level in relation to the"virality" of a tweet . Virality of a tweet was measured by the number of messages that were retweeted and the time elapsed from the original posting. They found that negative tweets, text that was classified as having a negative sentiment, had a more rapid and frequent retweet than positive or netural tweets. Negative messages were found to be retweeted by a factor of 1.2-1.6 times more and would be retweet quicker at a rate of 1.25 faster (Tsugawa et. al, 2015). This work suggests, rather expectedly, that the content of a tweet will effect the rate and amount of retweets of that message. While acknowledging that content is an important factor in retweeting, this paper will examine sentiment within the context of causal structure discovery, but will also examine possible social effects for causal modeling.

Work by (Luo et. al, 2015) which is closest to our work, focuses on how messages propagate as a retweet chain, and specifically on detecting burst of retweets and predicting them. They develop a predictive model to predict bursts, with features based on their analysis of the burst patterns. They use models like Logistic Regression, SVM, Naive Bayes, and others, and analyse the effect of the features on prediction using the logistic regression coefficients. They observe that the average number of retweets of the user is most predictive for burst prediction. While their work focused on retweet burst prediction, in this work we focus on a more precise measure of the user's propensity to retweet, based on social factors, rather than bursts of retweeting.

3 Dataset

The dataset consists of a collection of tweets from the time period May 2012 to December 2014. We focus on the political organizations of Latin America and identify a set of 63 organizations that include individual politicians and political groups, for example, anarchists. The politicians are mainly from countries like Venezuela, Columbia, Mexico, and so on. The tweets were collected using the twitter API, based on these 63 organizations of interest. The tweets included :

- Tweets by the organizations
- Tweets that mention these organizations

The data has the follwing information:

1. Time of the tweet
2. Retweet or not ?
3. User details: location, counts of status, followers, friends, klout score
4. Mentions: list of screennames and corresponding ids of users mentioned in the tweet
5. Retweet count: number of times the tweet has been retweeted
6. GeoLocation Enrichment
   - Latitude, longitude
   - Location/country
7. Basis Enrichment : The Twitter feed is enriched through the Basis pipeline, that adds

- Parts-of-Speech tags
- Named Entity Recognition : Identifies if an expression in the tweet is an organization or an individual or a URL.
- Noun-phrases

8. Sentiment score : computed by a third party[1]. These scores range from -24 to 24 and are mapped to values between [0,1], as detailed in the experiments section.

Approximately 6.9 million tweets were collected, that had about 10,400 tweets by the organizations, and the rest being the tweets mentioning these organizations. This project will focus mainly on the features of User Follower Counts, User Friends Counts, User Status Counts, User Klout Score, Mentions, and Sentiment in relation to if the tweet was retweeted or not. Each section will use a different subset of these features in order to address the question at hand.

4 Discovering Causal Structure

Our first goal is to identify the factors that cause or influence retweet propensity. The variables that we consider in this work are User Friends Count, User Status Count, User Followers Count, User Klout score and Tweet Sentiment. As a first step, we need to identify if there is any causal structure involving these variables and retweets. We use the PC algorithm (Spirtes et. al, 2000) described in section 4.2.

4.1 Data

For computational feasibility, we consider a subset of the dataset described in section 3. We randomly sample about 1 million tweets that are used for determining the causal graph structure. As mentioned above, we only look at factors involving the user social metrics and tweet sentiment.

4.2 The PC Algorithm

Given data over a set of observed random variables, and a conditional independence test, the PC (Peter Spirtes, Clark Glymour) algorithm builds a causal graph over these set of variables. The PC algorithm is based on two assumptions: *Causal Markov Property* and *Causal Faithfulness*. In our work, we ensure causal sufficiency by assuming that all the factors the variables involved are observed, and there are no hidden factors influencing retweets. The PC algorithm builds the causal graph structure in two main steps. In the first step, from the data, it learns a skeleton graph, i.e., a graph with only undirected edges. As a second step, it orients the undirected edges to form a markov equivalence class of DAGs. Consider a graph consisting of variables X, Y and a set of variables Z. The PC algorithm is based on the fact that, if there is no edge between variables X and Y, then there is a set of vertices Z either connected to X or Y such that X is independent of Y, conditioned on Z, or Z d-separates X and Y.

We use the R package *pcalg*, that contains the implementation for PC algorithm for estimating the causal structure. We use a gaussian test for conditional independence, also built into pcalg.

4.3 Findings

Figure 1 shows the output of the PC algorithm. Interstingly, we see that all the factors, i.e., Sentiment, Followers Count, Status Count, Friends Count and Retweets, influence the Klout Score, and this could be explained by the fact that Klout Score is calculated based on the popularity of the user(followers, friends, status). The graph also finds that followers count influences retweets, friends count and status counts. Sentiment of a tweet also seems to influence retweet. A bi-directional edge indicates that no direction of influence was found and this was the case for Status Counts and Retweets. Overall, from the graph, we find an interesting interaction between Followers Count, Status Count, Friends Count and Retweets. We explore this in more detail and attempt to find the magnitude of the influence of these factors on retweets in the next section.

5 Social Metrics Influence on Retweet Propensity

After identifying the possible influence directions of each metric, the magnitude or quantified effect of each variable is yet to be discovered . A key aspect of causal inference is to find the causal estimates associated with the cause of interest (Rubin et. al, 2011). The results of section 4 gave a structure to test, where in this section the causal estimates will be explored.

Since we are testing the causal structure that

[1] http://datasift.com

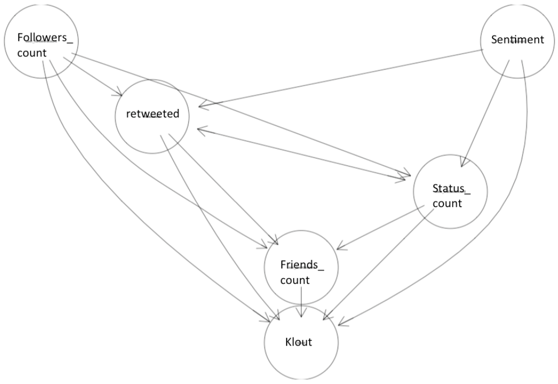


Figure 1: Output from the PC algorithm

was outputted from the PC algorithm, the main causes of interest are going to be the following three variables: User Friends Count, User Followers Count, and User Status Count. These three variables will be examined for their influence on the propensity to retweet. Propensity to retweet is measured as the total amount of retweets made divided by the total amount of tweets.

$$Propensity\ to\ retweet = \frac{Total\ \#\ of\ Retweets}{Total\ \#\ of\ Tweets}$$

These variables were choosen due to the interesting triangle structure they form around retweeting within the PC output. Each variable has some effect over each other while also having a relationship with retweeting. Finding the causal estimates also serves as a testing method for the PC algorithm, checking if the relationship given is actually observable. There are two assumptions being made that differ from the PC output, (1) the interaction between Friends Count and Retweeted is inverse than what is outputed from PC and (2) Status Count is directly influencing Retweeted instead of unknown. These assumptions allow for simplier testing methods and provide answer to the magnitude of influence across all four of these variables.

5.1 Data

The dataset for this subsection needed to be a reduced set of the data described in section 3 in order for the procedure to be computational tractable. A subsample was made thorugh a random sample of the main dataset, taking 15% of the main data set, thus giving around 1,041,000 tweets within this subsample. Furthermore, tweets that were missing data on Friends Count, Followers Count, and Status Count were removed. To further simplify the analysis in order to lower to complexity of the analysis, each variable of interest was reduced to two bins, High or Low. This binning was completed through comparing the user measure to the median within the overall distribution of each metric. For example: If user A has a Friends Count less than the median Friends Count, they would be classified as having a Low Friends Count.

Individual Effects	**High**	**Low**
Friends Count	0.67	0.72
Followers Count	0.71	0.77
Status Count	0.72	0.79

Table 1: Individual effect sizes of the variables of interest

5.2 Method

Effect sizes can be compared across multiple bins of social measures and allow for tractable conditioning across each variable. Conditioning is needed to discover the true effect for each variable of interest. As stated in previous work(Rubin et. al, 2011) , confounding variables can bias causal results due to the effect of the confounder being passed through to the variable of interest. As shown within the PC output, the possible confounders for these three metrics can be each other, therefore allowing us to condition on said variables to discover any causal effects.

5.3 Results

First to be examined is individual effect sizes. Table 1 shows that for each metric, there is a higher level of retweeting when the metric is lower. Thus saying that twitter users that have a lower number of friends, a lower number of followers, and a lower number of statuses, will then retweet more. Followers and Status Count have higher effect sizes of 0.06 and 0.07 increases in retweeting propensity compared to Friends Count having only a 0.05 increase.

Next, is to condition each effect on a level of the other metrics. Since there are three metrics of interest, each metric will need to be conditioned two times. Table 2 shows the results of such conditioning. When conditioning on the other metrics, the effect previously found from Friends Count is no longer existing. Friends when conditioned on either Followers or Status, produces no effect on Retweet Propensity. However, when Followers

One Level Conditioning	High		Low	
Effect Level	High	Low	High	Low
Friends — Followers	0.66	0.65	0.71	0.73
Friends — Status	0.65	0.67	0.72	0.73
Followers — Friends	0.66	0.71	0.65	0.73
Followers — Status	0.68	0.77	0.77	0.76
Status — Friends	0.65	0.72	0.67	0.73
Status — Followers	0.68	0.77	0.77	0.76

Table 2: Effect sizes when conditioning on one possible confounder

Two Level Conditioning	High	Low
Friends — Followers, Status	0.702	0.701
Followers — Friends, Status	0.683	0.72
Status — Friends, Followers	0.677	0.726

Table 3: Effect sizes when conditioning on two possible confounders

and Status are both conditioned on Friends Count, the effects still remain. This table does show an interesting interaction existing between Followers and Status. When Followers are conditioned on Status, Followers Count was found to effect the level of Retweeting by a factor of 0.09 when Status is High, but when Status is Low, that effect is no longer present. This is represented as the inverse when Status is conditioned on Followers.

Finally, the effects are examined using a two level conditioning. Here each metric is conditioned on both of the remaining metrics. The effects are combined averages of the variable of interests condition across all possible conditions for the conditioning variables. Table 3 shows these results. Within this table, the cell for when Friends is High is an averaged effect of when Friends is High across all possible conditions for both Followers and Status. As before, the effect from Friends Count is no longer present, while Followers and Status are showing significant differences in Retweet propensity. Even with the interaction present, different levels of Followers and Status counts effect the level of Retweeting.

These results, in comparison to the PC output, show there is a negtative effect due to Follower and Status counts from a user. In contrast, these results are supporting that there may be no relation between Friends Count and Retweeting, unlike that which is found within the PC output. Due to the nature of Follower counts influencing the propensity for a user to retweet, this implies that there may be social influences occurring.

6 Predicting the Propensity to Retweet from Network Effects

6.1 Motivation

In the third experiment we attempt to see if we can observe network effects in the data. It is well known that network effects can have a significant impact on the behavior of its members (Easley et. al, 2010). In particular we look for evidence of Latent Homophily and Social Contagion in the twitter network of a very well known political figure in Latin American. Latent Homophily is the tendency for the members of a social network to share similar characteristics because a latent characteristics lead the individual to join the social network in the first place (Shalizi et. al, 2011). Social Contagion is the tendency for the members of a social network to share similar characteristics because of direct influence between the member of the social network (Shalizi et. al, 2011) .

Although (Shalizi et. al, 2011) showed that separating the Latent Homophily and Social Contagion is very difficult, we hypothesized that we could observe evidence of either or both by measuring the number of mentions an individual makes. Mentions are direct references to an individual or group in Twitter. We hypothesize that the strength of membership on an individual belonging to a social group could be estimated by counting the number of mentions that individual makes about the organization. Similarly, we hypothesize that the level of influence on an individual by other members of the social group could be estimated by counting the number of times an individual is mentioned by others. We develop a Probabilistic Soft Logic (PSL) (Kimmig et. al, 2012) predictive model to attempt to predict the propensity of an individual to retweet based on the number of mentions that individual received and made.

6.2 Data

This experiment was run on a subset of the dataset described in section 3. In particular, we used all Tweets associated with one well known political figure. These include all tweets by this user, all tweets that mention the user, and all retweets of postings made by them. This user is a well known and very polemic individual in Latin American. The training set consisted of 179 tweets that originated from the user's account; 89,303 retweets of the original tweets; and 288,564 tweets that mentioned the user (excluding retweets). The test set contained 87 tweets that originated from the user's account; 13,281 retweets of the original tweets; and 97,901 tweets that mentioned this user (excluding retweets). The train and test sets do not overlap in time.

6.3 Methods

6.3.1 Probabilistic Soft Logic

Probabilistic Soft Logic(PSL)(Kimmig et. al, 2012) is a framework for collective, probabilistic reasoning in relational domains. PSL is a weighted first order logical templating language that specifies a class of continuous, conditional graphical models *hinge-loss Markov random fields (HL-MRFs)*(Bach et. al, 2013).

6.3.2 Model

The PSL model to incorporate Homophily and Contagion effects is described below and the rules are detailed in Table 4.

1. Estimating Homophily Strength: The higher the number of organization mentions by an individual A implies a higher propensity for individual A to retweet messages posted by the organization.

$$\begin{aligned} \textsc{postedbyindividual}(U, M) \wedge \\ \textsc{hasgroupmention}(M, G) \\ \rightarrow \textsc{retweetedgroup}(U) \end{aligned}$$

   where, U is the individual, G is the group and M is the tweet.

2. Estimating Contagion: The higher the number of mentions received by an individual A implies a higher propensity for individual A to retweet messages posted by the organization. Individual A has been identified (outside of PSL) as belonging to the social network associated with the organization. Versions of this rule include adjustments for the sentiment of the tweets.

$$\begin{aligned} \textsc{postedbyindividual}(U1, M) \wedge \textsc{mentions}(M, U2) \\ \rightarrow \textsc{retweetedgroup}(U2) \end{aligned}$$

   where, U1 and U2 are individuals, G is the group, and M is the tweet

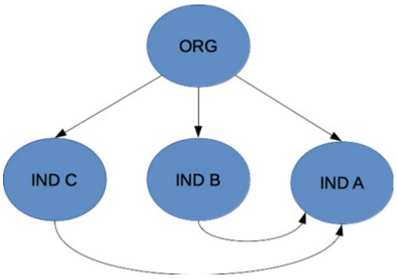


Figure 2: Graphical Interpretation of PSL Rules

A graphical interpretation of the rules can be observed in Figure 2 .

The propensity to retweet in the training and test sets were calculated following the procedure below:

1. Take the average and standard deviation of the counts of retweets made by each individual.

2. Make the measure linear by taking 0 as no evidence observed and 1 for evidence equal or higher than the mean plus two standard deviations as determined in 1.

The training set was used to train (compute weights) the PSL model. The learned model was then used to predict the propensity of an individual to retweet in the test set.

We run two sub-experiments on the data. In the first one, we discretized the propensity to retweet by dividing the values into three groups (low, medium, and high propensity to retweet scores), each having the same number of members. The mid value in each range was used as the retweet propensity for the entire group. In the second sub-experiment, the propensity to retweet values without adjustments were used.

6.4 Results

Figure 3 and 5 shows the results of the observational study performed on the discretized and non-discretized propensity scores respectively. It displays the actual and predicted propensity to retweet calculated as described in the previous section. The effects are small as suggested by the slope of the best fit lines, but the effect are statistically significant with p values of 2.2e-16 in both cases. The low R-squared values indicate the model fits the data poorly as expected.

PSL rules for homophily and contagion, to predict retweet propensity
PostedByIndividual(U, M) ∧ HasGroupMention(M, G) → RetweetedGroup(U)
PostedByIndividual(U1, M) ∧ Mentions(M, U2) → RetweetedGroup(U2)
PostedByIndividual(U1, M) ∧ Mentions(M, U2) ∧ Positive(M) → RetweetedGroup(U2)
PostedByIndividual(U1, M) ∧ Mentions(M, U2) ∧ Negative(M) → RetweetedGroup(U2)

Table 4: Homophily, Contagion and Retweet Propensity

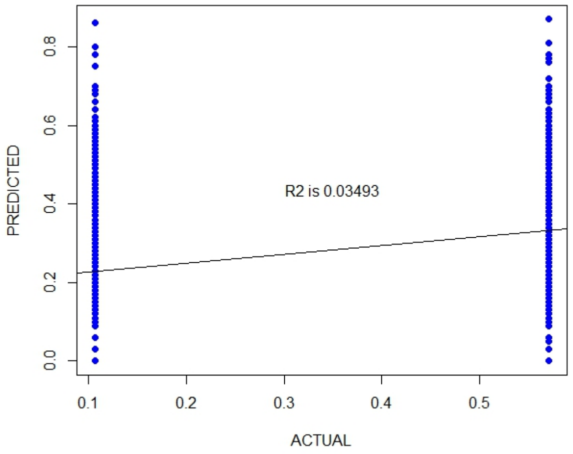
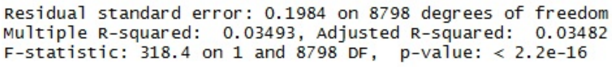


Figure 3: Predicted vs. Actual Propensity to Retweet by Individuals (discretized propensities)



```
Residual standard error: 0.1984 on 8798 degrees of freedom
Multiple R-squared:  0.03493,	Adjusted R-squared:  0.03482 
F-statistic: 318.4 on 1 and 8798 DF,  p-value: < 2.2e-16
```

Figure 4: R^2 results (discretized propensities)

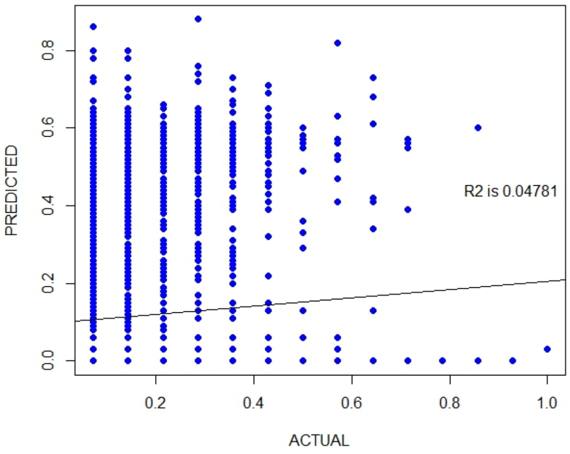
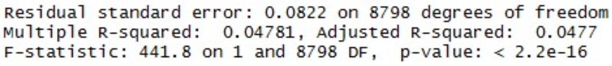


Figure 5: Predicted vs. Actual Propensity to Retweet by Individuals



```
Residual standard error: 0.0822 on 8798 degrees of freedom
Multiple R-squared:  0.04781,	Adjusted R-squared:  0.0477 
F-statistic: 441.8 on 1 and 8798 DF,  p-value: < 2.2e-16
```

Figure 6: R^2 results

The probability of an individual to retweet a message from an organization is very difficult to predict because it is the result of a very complex process involving several factors, many of which are latent. The complete understanding of the process probably involves factors related to the Organization such as its political stance, popularity and the strength of its following. Factors related to the message are most likely very important, such as the topic (or how interesting the topic is to the intended audience), the language used (funny, motivating, informative). Factors related to the receiver such as its propensity to retweet, gender, age, culture and individual interests are all also likely very important.

7 Conclusion

Retweet propensity is a complex process involving many factors, many of which are latent. It is also a very important area of research as retweets are the first stage in message propagation in the Twitter social media. Much research remains to be done to fully explore the causal factors behind this process.

This paper shows that progress in this area will likely involve the usage of many causal inference techniques. In particular, it shows that the potential causal structure can be analyzed using techniques such as the PC Algorithm. This technique uncovered a direct effect between follower counts and retweets, and between tweet sentiment and retweets (an effect observed during previous research). Conditioning proved to be an effective tool to uncover the magnitude of the effects on retweet propensity, validating and enhancing the results obtained during causal structure extraction. We uncover an interesting relationship between number of followers and retweet propensity. People with less followers are more likely to retweet. It could well be that the more popular social media users are more likely to create original content (Tweets and Statuses), and therefore less likely to retweet. An interaction between the number of followers and status counts was observed by the PC algorithm and confirmed during conditioning. This hints that content creation (Statuses) and follower counts are prehaps infleuncing each other. It is reasonable to expect uses who creating more content will be more unqiue and more interesting to follow? Full exploration of this process is subject to future research.

A major limitation to this work was in complexity. Running these models on the causal graph outputed from PC would take an extremely long time, so the models needed to be reduced in order to accomplish results.

We showed that relational tools such as PSL can be used to explore network effects on social media. This is a relatively new area of research that promises to enhance the results obtained by other causal analysis techniques.

8 Acknowledgement

Special thanks to Rupinder Paul, PhD student at Virginia Tech, who has been of great help, for retrieving the tweets based on the organizations of our interest.

References

Easley, David and Kleinberg, Jon *Networks, crowds, and markets: Reasoning about a highly connected world* Cambridge University Press

Zhang, Mimi and Jansen, Bernard J and Chowdhury, Abdur *Business engagement on Twitter: a path analysis* Electronic Markets

The Klout Score `https://klout.com/corp/score`

Peter Spirtes, Clark Glymour, and Richard Scheines *Causation, Prediction, and Search* The MIT Press, 2nd edition, 2000.

Rubin, Donald B *Causal inference using potential outcomes* Journal of the American Statistical Association

Zhilin Luo , Yue Wang , Xintao Wu , Wandong Cai , and Ting Chen *On Burst Detection and Prediction in Retweeting Sequence* Proceedings of the 19th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)

Tsugawa, Sho and Ohsaki, Hiroyuki *Negative Messages Spread Rapidly and Widely on Social Media* Proceedings of the 2015 ACM on Conference on Online Social Networks

Cosma Rohilla Shalizi and Andrew C. Thomas *Homophily and Contagion Are Generically Confounded in Observational Social Network Studies* Sociological Methods and Research, vol. 40 (2011),

Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, Xiaoming Li *Comparing Twitter and Traditional Media Using Topic Models* In Advances in Information Retrieval, 2011.

Kimmig, Angelika and Bach, Stephen H. and Broecheler, Matthias and Huang, Bert and Getoor, Lise 2012. *A Short Introduction to Probabilistic Soft Logic* NIPS Workshop on Probabilistic Programming: Foundations and Applications, 2012.

Bach, Stephen H. and Huang, Bert and London, Ben and Getoor, Lise *Hinge-loss Markov Random Fields: Convex Inference for Structured Prediction* Uncertainty in Artificial Intelligence, 2013.

Emilio Ferrara, Zeyao Yang *Measuring Emotional Contagion in Social Media* Journal, PLOS ONE