Abstract

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 re-tweets, likes and mentions received by the message sender. 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. In this project we will focus on analyzing potential causes for re-tweets.

To accomplish this, this project has three separate stages.

1. To find any potential causal structures from observational data. The dataset used is observational which create many challenges for discovering causal effects since the data collection did not take that into account like an experimental setting would allow. To find potential strucutres, an algorithmic procedure (cite PC paper) will be used.

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

3. Based on the previous findings of this project, possible social effects may be occurring, this section will attempt to model those social effects in their power to predict propensity to retweet.

2 Related Work

- homophily paper

- COSN conf paper While meta data metrics are typically focused on, there are previous studies that have been conducted purely on examining the content of the tweet. Tsugawa and Ohsaki examined tweets sentiment level in relation to how "viral" a tweet was. Virality of a tweet was measured by the number of meesages 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 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. 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 only examine possible social effects for causal modeling.

- Truthy - papers

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 tweets posted, followers, friends, klout score

4. Mentions : list of screennames and corresponding ids of users mentioned in the tweet (When a tweet by a user, includes the username of another user, it is called a 'mention'.)

5. Retweet count : Number of times the tweet has been retweeted

6. GeoLocation Enrichment
   - Latitude, longitude
   - Location/country
7. Basis Enrichment
   - The tweets are tokenized and tagged with Parts-of-Speech tags
   - Entities identified if any, i.e., if an expression in the tweet is an organization or an individual or a URL.
   - Noun-phrases identified.
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.

4 Social Metrics Influence on Retweet Propensity

4.1 Motivation

After identifying the possible influence directions of each metric, it has yet to be discovered the magnitude or quantified effect of each variable. A key aspect of causal inference is to find the causal estimates associated with the cause of interest [Rubin 2005]. The previous section gives a structure to test, where in this section the causal estimates will be explored.

Since we are testing the causal structure that 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. $Propensity to retweet = \frac{Total Number of Retweets}{Total Number of Tweets}$

These variables were choosen due to the interesting triangle structure they form around retweeting within the PC output. Each variable has some affect 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.

[1] http://datasift.com

Table 1: My caption

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

4.2 Data

The dataset for this subsection needed to be a reduced set of the data described above in order for the procedure to be computational tractable. A subsample was made thorugh a random sample of the main dataset, taking 15

4.3 Method

Effect sizes can name 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, 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.

4.4 Results

First to be examined is individual effect sizes. Table XX 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.

Discussion points: Social metrics indicate that lack of creating one's own content may be an indicator of being less socially involved. People don't want to follow people who only repeat information.

Next, is to condition each effect on a level of the other metrics. Since there are three metrics of interest, this means each metric will need to be conditioned two times. Table XX shows the

Table 2: My caption

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 3: My caption

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

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 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 used 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. So 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. Table XX shows these results. 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.

Discussion Point: Followers and Status have an interaction indicating they are related. This is shown to be true within the PC output. This may be due to content creation from users (measured as Status Count) can be influencial in the number of Followers a user may receive.

5 Discovering Causal Structure

5.1 Motivation

Our goal is to identify the factors that cause or influence retweet propensity. The variables that we consider in this work are friends count, status count, followers count, klout score and 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 described in the next section.

5.2 Dataset

For computational feasibility, we consider a subset of the dataset described in the previous section. We randomly sample about 1 million tweets are used for determining the causal graph structure. As mentioned above, we only look at factors friends count, status count, followers count and klout score, associated with users, and sentiment, associated with the tweet itself.

5.3 The PC Algorithm

Given a dataset 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.

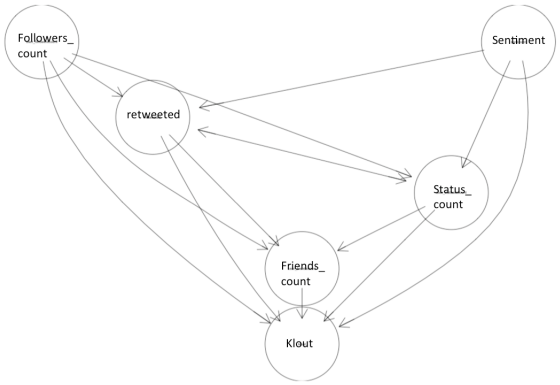

Figure 1: Output from the PC algorithm

5.4 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, though we will not be exploring this further in this work. The PC algorithm does not find the direction of influence between 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.

6 Diego Section

6.1 Motivation

In the third experiment we attempted to see if we could observe network effects in the dataset. It is well known that network effects can have a significant impact on the behavior of its members[Easley]. In particular we searched for evidence of Latent Homophily and Social Contagion in the Tweeter message network of a very well know 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]. 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].

Although [Shalizi] 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 Tweeter. We hypothesized 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 hypothesized 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 developed a Probabilistic Soft Logic (PSL) 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 above. In particular, we used all Tweets associated with Nicolas Maduro, the president of Venezuela. These include all tweets by Nicolas Maduro, all tweets that mention Nicolas Maduro, and all retweets of postings made by him. Nicolas Maduro is a well known and very polemic individual in Latin American. The training set consisted of 179 tweets that originated from the Nicolas Maduro account; 89,303 retweets of the original tweets; and 288,564 tweets that mentioned Nicolas Maduro (excluding retweets). The test set contained 87 tweets that originated from the Nicolas Maduro account; 13,281 retweets of the original tweets; and 97,901 tweets that mentioned Nicolas Maduro (excluding retweets). The train and test sets do not overlap in time.

6.3 Methods

The PSL program described in the introduction contained the following rules:

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.

m.add rule : (PostedInd(U,M) & HasGroupMention(M,G)) ¿¿ RetweetedGroup(U), weight : 1

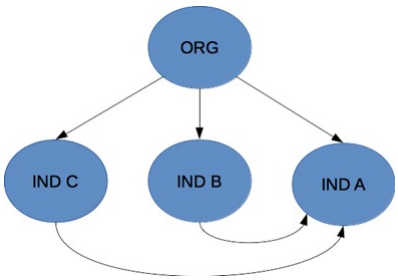

Figure 2: PSL Rules

Where: U is an individual G is the group 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.

m.add rule : (PostedInd(U1,M) & Mentions(M,U2)) ¿¿ RetweetedGroup(U2), weight : 1 m.add rule : (PostedInd(U1,M) & Mentions(M,U2) & Positive(M)) ¿¿ RetweetedGroup(U2), weight : 1 m.add rule : (PostedInd(U1,M) & Mentions(M,U2) & Negative(M)) ¿¿ RetweetedGroup(U2), weight : 1

Where: U1 and U2 are individuals G is the group M is the tweet

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

The propensity to retweet in the training set was 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.

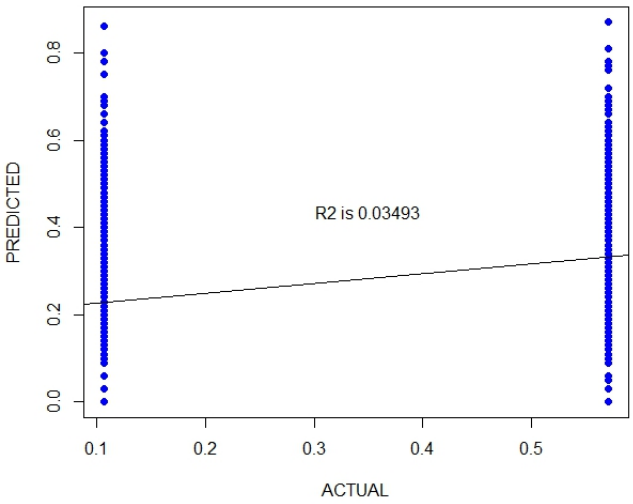

Figure 3: Predicted vs. Actual Propensity to Retweet by Individuals (discretized propensities)

6.4 Results

Figure 2 and 3 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 effect are small as suggested by the slop 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.

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. Factor related to the message are most likely very important, such us 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

overall conclusion/discussion - connect all sections

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

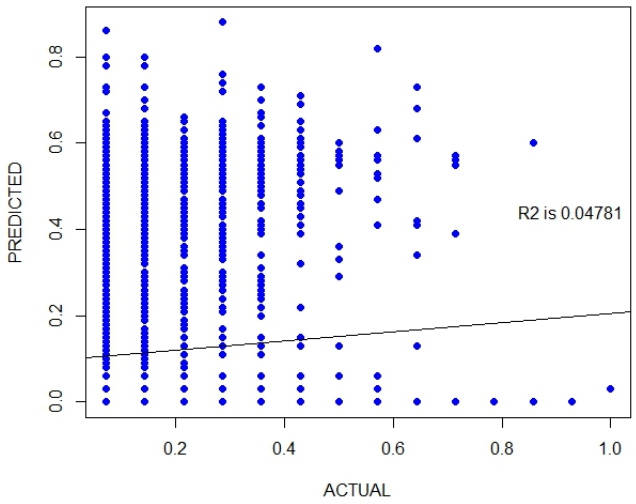

Figure 5: Predicted vs. Actual Propensity to Retweet by Individuals (discretized propensities)

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