Abstract

1 Introduction

- About twitter - why causal study on twitter
- Aim of the project
- 3 stages in the study : describe , and why these stages
- causal discovery
- magnitude of causal effects
- homophily/contagion

2 Related Work

- homophily paper
- COSN conf paper
- 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.

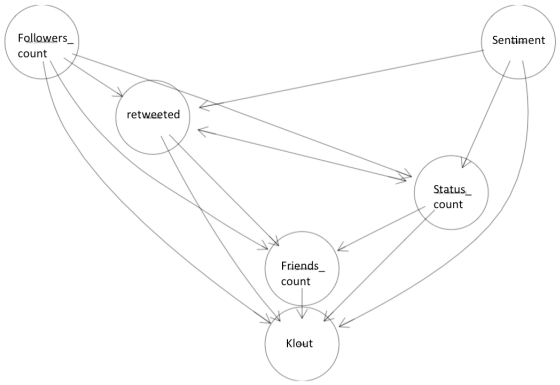
4 Ryan Section

- motivation
- data
- method
- findings
- motivation
- data
- method
- findings

5 Discovering Causal Structure

As discussed above, our goal is to identify the variables 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 try to identify if there is any causal structure involving these variables and retweets. We use the PC algorithm described below.

[1] http://datasift.com



5.1 The PC Algorithm

The PC algorithm is based on two assumptions: Causal Markov Property and Causal Faithfulness. Consider a graph consisting of variables {X,Y, Z, Z1,Z2...}. Given the data, the PC algorithm tries to come up with the causal graph structure. The PC algorithm has 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. PC algorithm is based on the fact that, if there is no edge between variables A and B, then there is a set of vertices C either connected to A or B such that A is independent of B, conditioned on C, or C d-separates A and B.
For each X and Y, see if X is independent of Y; if so, remove their edge. For each X and Y which are still connected, and each third variable Z, see if X is independent of Y given {Z}; if so, remove the edge between X and Y. For each X and Y which are still connected, and each third and fourth variables Z1 and Z2, see if X is independent of Y given {Z1,Z2}; if so, remove their edge.

5.2 Our findings

6 Diego Section

- motivation
- data
- method/models
- findings

7 Conclusion

overall conclusion/discussion - connect all sections