Classification in Social Networks based on Activity in Politics

Mihir Kulkarni - 12D020007 Aditya Kusupati - 130050054
CS 728 Project

Motivation:

Social networks are widespread today, and a lot of political activity is consumed and generated on these today. With the volumes of content on social media today, it seems natural to be able to say something about a person's political inclination by observing that of their friends, and the pages they follow and interact with. This information is useful in drawing inferences about what types of content they would want to consume, and about their biases.

While we have looked at the problem from a political lens, our methods could potentially extend to any sphere where there are clearly demarcated opinions.

Problem Statement:

Given a social network graph with nodes of the following types:

1. Individuals
2. Pages
3. Articles

with articles tagged either 0 or 1, standing for left or right bias respectively, quantify the bias of individuals using the biases of the pages they follow and the link structure of the graph.

Bias here is a number in [0,1].

Approach

We want to use this structure to learn biases of the users in the social network graph.

1. Compute biases of each page. This equals the average of the biases of each of the articles the page has shared.

2. Compute Page Rank of all nodes in subgraph P consisting of only pages.
3. Consider the subgraph N consisting only of nodes corresponding to individuals. For each node n in N, we assign the Page initial score of n, $initscore(n)$ as the weighted average of the biases of each of the pages it follows, weighted by the PageRank of each page in subgraph P

$$initScore(n) = \frac{\sum_{n\ follows\ x} PageRank(x) * Bias(x)}{\sum_{n\ follows\ x} PageRank(x)}$$

4. Now we compute the biases of each node n in the graph. For this, we hypothesize that the influence on each node n by nodes which share an edge and that from pages n follows are in the ratio $r : 1 - r$. Using this, we can write the equations for bias as:

$$bias(n) = r * initScore(n) + (1 - r) * influenceFromFriends(n)$$

where

$$influenceFromFriends(n) = \frac{\sum_{y\ is\ a\ friend\ of\ n} bias(y)}{\sum_{y\ is\ a\ friend\ of\ n} 1}$$

This gives us a linear system in $bias(n)$, which can be solved for obtaining the final answer.

$$r * A\bar{b} + (1 - r) * \bar{i} = \bar{b}$$

where

$\bar{b}$ is the bias vector

$\bar{i}$ is the vector of initial scores

A is a matrix with the j th row containing the number $\frac{1}{degree(j)}$ at all neighbours of j, and 0 elsewhere

For our experiments, we have used the value $r = 0.5$

Datasets:

We have used 2 major datasets and have tweaked around them as per our requirement.

Political Blogs Dataset[3]: AKA Adamic Dataset

As we were unable to come-up with novel and robust methods for classification of articles as per inclinations in various domains, we used a very simple blog dataset which is labelled with a binary classification among **Left** and **Right** in terms of political inclination. The label 0 => **Left** and label 1 => **Right**. As we don't have continuous regression based values for the above classification, which could be a great dataset for better analytics based on our approach. So 0 and 1 will be used as is in the future, the usage will be defined in the algorithm.

We assign certain articles to each Page node in the graph which in turn affects the inclination of the page. Our methodology for doing this was to associate each Page Node with an "activity", which corresponded to how many articles it was associated with. A combination of activity and an associated initial bias was used to associate each page with a subset of the available articles in the dataset. This set was used to compute its final score, as the average bias of all the articles it published.

Social circles: Facebook[1]

This dataset is actually meant for the ego centric analysis of social networks. There are many features of this dataset which we are using and a few which can be used but are not used due to anonymisation problems and complications into the existing model.

The dataset has 10 Ego nodes in the social network and 4029 normal nodes. To cater our problem we have changed definitions follows:

1) Ego Nodes are changed to **Pages**
    a) Pages can only be followed by normal nodes (ie., a page can't follow as person)
    b) Pages can be friends ie., if they are associated they form a undirected edge, this is in analogy to page affiliations, associations in facebook, which is internal but is evident by the behaviour. This closed graph on page nodes actually has a meaningful definition of **PageRank** score as the amount of Influence in that domain. Eg., Leaders in a Political Party, an influential leader is determined by his associations, thus it is equivalent to PageRank
2) Normal Nodes are changed to **People/Persons**
    a) People can follow Pages -> directed edge
    b) People can friend People -> undirected edge
3) Edges are parsed according to the above properties of both the nodes

Results:

On running the provided code, you shall get the final Bias of each Individual and also the Bias of each Page along with the influence of each Page. These are the factors which will be used for comparison. Given the final Biases and the Page Follow structure of each individual, we can deduce how much of the bias is actually translated from the page follow list and how

much is from the Friendship Graph as the society/friends in social network does affect the individuals in a way. Final value obtained is the Bias in the given Domain and this can be used for targeted advertising and e-campaigns.

As the given graph is not densely connected as far as the friendship graph is concerned, we find that most of biases are almost similar to the page influence on the person, but certain nodes show the effect of neighbours on them by showing a significant change from the initial bias obtained from the page follow list.

The results given out also contain a list of nodes which are influenced by friendship graph and thus are away from the absolute page influence. The count and the deviation can be found at the end of the results

For the given article allocation done in the code with the activity of the pages as initialised, we have the final Bias scores with the following insights:

1) The graph is not dense in terms of friendship
2) For a value of **r** = 0.5 => equal weightage for Pages and Friends and a threshold of 0.1 (the minimum change required to consider) (change is the difference of initial bias from pages and the final bias due to the graph, we found that 72 nodes out of 4029 have a significant impact due to Neighbourhood
3) When we change the values of **r** and the threshold (which are global variables that can be changed in the code), we find the results change accordingly

Future Work:

Given labelled data for the nodes we can actually learn the weights for the pages and the friendship graph. By having the learnt models that evaluate on the steady state scores of the graph nodes, we will be able to deduce generalised results to the specific domain over long periods of time and circumstances. This can be further extended by providing weights to each edge in the graph which denote the strength of the friendship, in turn representing the influence of that friend. Finally with a large amount of labelled data we will be able to learn higher order functions which work upon the steady state biases and help us understand the scenarios in a better way.

References:

[1] http://snap.stanford.edu/data/egonets-Facebook.html
[2] http://www-personal.umich.edu/~mejn/netdata/

[3] http://www-personal.umich.edu/~mejn/netdata/polblogs.zip
[4] http://i.stanford.edu/~julian/pdfs/nips2012.pdf
[5] https://networkx.readthedocs.io/en/stable/index.html