STANFORD UNIVERSITY
Stanford Center for Professional Development

HOMEWORK ROUTE FORM

Stanford Center for Professional Development Student Information

Course No. CS224W

Faculty / Instructor Name Jure Leskovec

Date 12/10/2012

Student Name Olga Kapralova, Valentina Kroshilina, and Philip Scuderi

Phone 413-262-0693

Company

Email pscuderi@gmail.com

City. Boston

State MA

Country U.S.

Check One: ○ Homework #: ○ Midterm ● Other Project Report

The email address provided on this form will be used to return homework, exams, and other documents and correspondence that require routing.

Total number of pages faxed including cover sheet 10

	For Stanford Use Only	
Date Received by the Stanford Center for Professional Development	Date Instructor returned graded project Score/Grade: (to be completed by instructor or by teaching assistant)	Date the Stanford Center for Professional Development returned graded project:

STANFORD USE ONLY

Please attach this route form to ALL MATERIALS and submit ALL to:

Stanford Center for Professional Development
496 Lomita Mall, Durand Building, Rm 410, Stanford, CA 94305-4036
Office 650.725.3015 | **Fax** 650.736.1266 or 650.725.4138
For homework confirmation, email scpd-distribution@lists.stanford.edu
http://scpd.stanford.edu

Last modified October 27, 2008

CS224w:
Social and Information Network Analysis

Assignment number: Project Report

Submission time: 11:50 AM PST **and date:** 12/10/2012

Fill in and include this cover sheet with each of your assignments. It is an honor code violation to write down the wrong time. Assignments are due at 9:30 am, either handed in at the beginning of class or left in the submission box on the 1st floor of the Gates building, near the east entrance.
Each student will have a total of *two* free late days. *One late day expires at the start of each class.* (Homeworks are usually due on Thursdays, which means the first late day expires on the following Tuesday at 9:30am.) Once these late days are exhausted, any assignments turned in late will be penalized 50% per late day. However, no assignment will be accepted more than *one* late day after its due date.

Your name: Group 12
Email: pscuderi@gmail.com **SUID:** 05855811

Collaborators: Olga Kapralova, Valentina Kroshilina, and Philip Scuderi

I acknowledge and accept the Honor Code.

(Signed) /Philip Scuderi/

(For CS224w staff only)

Late days: 1 2

Section	Score
1	
2	
3	
4	
5	
6	
Total	

Comments:

DETECTION OF OPINION POLARIZATION

OLGA KAPRALOVA, VALENTINA KROSHILINA, PHILIP SCUDERI

Group #12

1. Introduction

The value of having a greater understanding of the opinion formation process is hard to ignore. Such an understanding could improve decision makers understanding and predictions of many important variables such as customer behavior, tax rates, regulatory changes, and political risk. With such societal and institutional benefits in mind, we focused our research efforts on improving our understanding of the opinion formation process.

Our research efforts were focused on a particular opinion formation model that was recently proposed by Dandekar et al in [2]. While Dandekar's model gives a set of metrics that allows to predict the dynamics of the discussion, however no guidance is provided on setting up parameters for using the model on the real-life data. We tested Dandekar's model for two discussions in Twitter: health care reform (Obamacare) and death of Michael Jackson. To validate our results we compared opinions produced by Dandekar's model with user opinions based on the result of a sentiment analysis. Our results show that knowing biases of only two users is sufficient to predict the opinions and polarization index coefficient for the specified topics. We would also like to note that the developed method can be easily generalized to analyze broader set of topics.

2. Prior Work

2.1. **Model.** Let us first describe the model of opinion formation process [2] that we would like to validate in our project. In [2] the society is considered as weighted undirected graph, where nodes represent its members, and edges represent relations between them. Weight of the edge (u, v) represents influence that nodes u and v have on each other. In general, since the model uses an undirected graph, weights on the edges of the graph represent the relative value of the neighbors' opinions. The model also allows loops that represent how much a person values his own opinion.

Opinion formation process developed in [2] is modeled as an iterative process. At each time step t a node in the graph has opinion which is between 0 and 1, i.e. $x_i(t) \in [0, 1]$. Opinions 0 and 1 represent extreme opinions on some particular topic, while an opinion of 0.5 shows that the node is undecided. The weight of an edge is denoted by w_{ij} such that $w_{ij} \geq 0$. The weight $w_{ii} \geq 0$ on the loop (i, i) represents the value of the opinion of the individual represented by node i to itself. The parameter $b_i \geq 0$ represents the node's bias. At the next time moment $t + 1$ node i assimilates information from its neighbors and updates its own opinion to the weighted average of its own opinion and its neighbors' opinions, i.e.

$$x_i(t+1) = \frac{w_{ii}x_i(t) - (x_i(t))^{b_i}s_i(t)}{w_{ii} + (x_i(t))^{b_i}s_i(t) + (1 - x_i(t))^{b_i}(d_i - s_i(t))}, \tag{2.1}$$

where $s_i(t) = \sum_j w_{ij}x_j(t), d_i = \sum_j w_{ij}$ and in both cases the sum is over all j that are neighbors of i.

To measure polarization, the network disagreement index (NDI) is defined in [2] as

$$\eta(G, \mathbf{x}(t)) = \sum_{(i,j)\in E} w_{ij}(x_i(t) - x_j(t))^2. \tag{2.2}$$

The opinions in the society are defined to be polarizing if $\eta(G, \mathbf{x}(t + \tau)) > \eta(G, \mathbf{x}(t))$ for some $\tau > 0$, i.e. when the opinions are diverging.

Thus, to test the theoretical model proposed in [2] on a Twitter dataset, we need to convert the directed who-follows-whom network to an undirected graph, as well as provide edges' weights and nodes' biases. We now describe the most relevant work on this topic.

2.2. **Measuring biases and weights.** The most relevant work in terms of estimating biases and edge weights is the work of [1]. In it the authors are investigating problems that arise in real-time sentiment analysis. In particular, they show that in order to correctly predict the sentiment, one needs to analyze how biased the opinion holder is. To calculate the bias of the node, a notion of opinion agreement graph (OAG) is introduced in [1]. We describe this graph in more detail in section 3.1, as we use it in our method for the theoretical model [2] evaluation. To summarize, OAG is a weighted undirected graph that captures how individuals value each other opinion on a particular topic. As we are testing the theoretical model [2] on different topics, we decided to use the dynamic (i.e. topic dependent) OAG instead of a static who-follows-whom network structure (we elaborate on that in section 3.1 where we specify the graph construction procedure). Let us highlight that OAG has the same set of nodes as the initial graph, however its set of edges is different. While in the initial graph edges are static and show if two individuals 'know' each other, in OAG the edges represent that two individuals value each other opinion on a given topic. Thus, the set of OAG edges might differ from topic to topic.

Another reason for using OAG is that the model proposed in [2] operates with undirected graphs. A simple removing the directions on the edges of the who-follows-whom network seemed inadequate to us, as it obviously looses a lot of information. We initially considered the way of balancing this leak of information by creating weighted self-loops (for example, Obama's self-loop would have a very large weight w_{ii}, which would indicate that he probably values his opinion the most). However, by using a single parameter, i.e. the weight of the self-loop, it is difficult to capture the relationship between a user and each of his neighbors.

We will use the OAG as an underlying weighted undirected graph that model proposed in [2] takes as an input.

3. Methodology

In this section we present the algorithm we used to set initial parameters for the Dandekar's model. We discuss how Opinion Agreement Graph (OAG) was used to set initial weights and biases. We also describe the procedure used to set initial opinions for the users. Then we present a set of metrics that Dandekar's model provides to asses polarization of the discussion and how we validated the findings against the Twitter data.

3.1. **Construction of undirected weighted graph: Opinion Agreement Graph (OAG).** Recall that polarization model proposed in [2] operates with the undirected weighted graph. In this section we describe the procedure which takes as its input the directed graph constructed using Twitter data, and outputs the weighted undirected graph. This graph is called the Opinion Agreement Graph (OAG), and it was proposed in [1], where authors argue that OAG captures the global opinion of the whole network. We will elaborate on that later in this section.

In particular, for a given topic t let $G_t = (V_t, E_t)$ be the directed graph where V_t is the set of users who ever tweeted on topic t, and where an edge $(u, v) \in E_t$ shows that the user u has re-tweeted user v. Algorithm 1 describes the construction of the undirected OAG graph $G'_t = (V_t, E'_t)$ from the directed graph G_t (to simplify notation, we omit the subscript t in the algorithm description). Note, that graphs G_t and G'_t have same set of nodes, but different set of edges. After that, we estimate weights on the edges E'_t. Note, that whenever one or more users have re-tweeted the same user, they form a clique. Indeed, this captures that the views of these users on a given topic are correlated, and the amount of this correlation is specified by the edge weight. We will provide the formal definition of the edge weight later in this section, but informally weight captures how much the probability of an event that two users are in one clique deviates from the probability that the same users are in a clique if we consider them to be independent. Similarly, if some user re-tweeted a set of other users on a given topic, they also form a clique due to the same reason. Note that the resulting graph G' might have a larger set of edges (indeed this is what we observed in our experiments) than the original graph G, as it might connect nodes who share similar views but who do not know about each other.

Algorithm 1 Construction of the OAG: transforming G into G'

```
V' ← V
E' ← ∅
D_A ← ∅
for all v ∈ V do
  O(v) ← {u ∈ V : (v, u) ∈ E}
  for all u_1, u_2 ∈ O(v) such that u_1 ≠ u_2 do
    if E' does not contain (u_1, u_2) then
      E' ← E' ∪ {(u_1, u_2)}
    end if
    D_A ← D_A ∪ {(u_1, u_2)}
  end for
end for
Perform similar procedure for D_p. Use I(v) ← {u ∈ V : (u, v) ∈ E} instead of O(v)  1
```

Thus, we decided to use the OAG weighted graph instead of the initial who-follows-whom twitter network since its provides more information about how to users influence each other on a particular topic. Indeed, the who-follows-whom relationship does not change from topic to topic, i.e. it is static, while OAG is not static, but is topic dependent, since an edge in OAG shows how opinions of two users are similar on a given topic (when compared to the whole network).

In particular, sets D_A and D_P are called the passive and active sets. Indeed, set D_A corresponds to the set of users who re-tweeted other users, while D_P is the set of users who was re-tweeted. As suggested in [1], the weight of each edge in G' is calculated as $w(u,v) = \max(\alpha(u,v), \rho(u,v))$, where $\alpha(u,v)$ and $\rho(u,v)$ are lifts of the pair (u,v) in the sets D_A and D_P respectively. In particular,

$$\alpha(u,v) = \frac{P_{D_A}(u,v)}{P_{D_A}(u)P_{D_A}(v)} \quad \text{and} \quad \rho(u,v) = \frac{P_{D_P}(u,v)}{P_{D_P}(u)P_{D_P}(v)}.$$

Here $P_{D_A}(u,v)$ corresponds to the probability of seeing an edge (u,v) in the set of edges D_A. Similarly, $P_{D_A}(u)$ is the probability of seeing a node u in the set D_A (similar notation is used for the set D_P). Thus, lift shows how surprising a relationship between two nodes is, since it compares the frequency of their co-occurrence to the expected frequency of their co-occurrence, if they were independent.

3.2. **Estimation of opinions and biases.** We now describe the procedure we used for estimating opinions and biases of users. Denote the constructed OAG by $G_{OAG} = (V, E)$. The opinions and biases were defined as follows. For each topic two *attractors* $A_0 \subset V, A_1 \subseteq V$ were first chosen manually. These represented sets of individuals in the network that are known to have strong positive and negative opinions on the topic. As per results of [1] we used Pagerank personalized at the attractors to obtain opinion vectors. Let $\beta \in (0,1)$ denote the teleportation parameter. For each $x \in V$ and a starting set $S \subseteq V$ let $p_S(x)$ denote the pagerank of x personalized at S.

Opinions. We estimate opinions using the work of [1]. Unfortunately, this work does not contain an explicit algorithm for estimating opinions. The author only states that he used a random walk to estimate opinions. A natural interpretation of the work of [1] suggests using

$$\frac{p_{A_1}(x)}{p_{A_0}(x) + p_{A_1}(x)} \tag{3.1}$$

as the opinion vector of x. Note that this quantity is always between 0 and 1, with 0 corresponding to x leaning to A_0, 1 corresponding to A_1 and 1/2 corresponding to an unbiased individual. We found, however, that, this expression leads to high noise sensitivity, and instead used a regularized version:

$$\frac{p_{A_1}(x) + \delta}{(p_{A_0}(x) + \delta) + (p_{A_1}(x) + \delta)} \tag{3.2}$$

for a small $\delta > 0$. It was empirically observed that (3.2) is more robust to noise than (3.1), which may output essentially uniformly random values for nodes $x \in V$ for which $p_{A_0}(x)$ and $p_{A_1}(x)$ are small.

Biases. The biases were calculated according to the formula

$$\frac{1}{F} \max \left\{ \frac{p_{A_1}(x)}{p_{A_0}(x)}, \frac{p_{A_0}(x)}{p_{A_1}(x)} \right\}, \tag{3.3}$$

where F is a scaling factor which we choose to be 2 and 8 in the two experiments performed (see simulation results section below). Note that equation (3.3) outputs number between $1/F$ and $+\infty$. The formula for the bias was chosen based on the main intuition behind the proofs in [2]. In particular, in the model of [2] the network becomes sufficiently biased if all bias parameters are above 1. Thus, the division by F in (3.3) was chosen to ensure that some unpolarized behavior is possible in the model of [2].

Estimation of parameters. Personalized PageRank was estimated using random sampling. For each of the two attractors 200000 random walks were performed to approximate the resulting distribution. Thus, by the Chernoff bound one can expect faithful approximations of the elements of the distribution that are above, say, 1/20000. The teleportation parameter was chosen as $\beta = 1/10$. Empirically, choosing substantially smaller β led to convergence to the stationary distribution from each of the attractors and resulted in the opinions being equal to 1/2 for most nodes. The value of the regularization parameter δ was chosen as 1/10000 so as to minimize the influence of estimation error, i.e. ensure that nodes x with both $p_{A_0}(x)$ and $p_{A_1}(x)$ below our precision threshold of 1/20000 only contribute to the set of nodes with neutral opinions.

3.3. **Measuring Polarization.** We measure divergence of opinions by calculating Network Disagreement Index (NDI) that is introduced in [2] as given by (2.2). Discussion is defined to be polarizing if NDI is a strictly increasing function of time. Opinion of a user is a function of time and initial opinion $x_i(0)$, that according to [2] can be defined as in equation (2.1). Note that according to our weights construction process $w_{ii} = 0$, thus, $x_i(t+1)$ that we used for our calculations is:

$$x_i(t+1) = \frac{(x_i(t))^{b_i} s_i(t)}{(x_i(t))^{b_i} s_i(t) + (1 - x_i(t))^{b_i}(d_i - s_i(t))}. \tag{3.4}$$

Another metric that was introduced in Dandekar's model is polarization threshold for the individual defined below:

$$\hat{x}(s, b) = \frac{s^{1/(1-b)}}{s^{1/(1-b)} + (1-s)^{1/(1-b)}},$$

where $s = s_i/d_i$ and s_i, d_i are as defined in section 2.

If discussion follows biased opinion formation process described in [2] (equation (3.4)), then $\hat{x}$, $x_t(0)$ and user bias b_i are 3 parameters sufficient to predict opinion of a user as a limit when $t \to \infty$:

- If $b > 1$
  - if $x(t) > \hat{x}$, then $x(t+1) > x(t)$, and $x(t) \to 1$ as $t \to \infty$
  - if $x(t) < \hat{x}$, then $x(t+1) < x(t)$, and $x(t) \to 0$ as $t \to \infty$
  - if $x(t) = \hat{x}$, then for all $t' > t$, and $x(t') = \hat{x}$.
- If $b < 1$
  - if $x(t) > \hat{x}$, then $x(t+1) < x(t)$
  - if $x(t) < \hat{x}$, then $x(t+1) > x(t)$
  - $x(t) \to \hat{x}$ as $t \to \infty$.

For each node at $t = 0$ we calculated $\hat{x}$ to predict networks opinion distribution when $t \to \infty$.

3.4. **Validation of results.** While the model's premise was that environment stays static, meaning that weights, biases and the network structure remains unchanged during the discussion, common sense tells us that users opinion, as well as external sources, can result in a significant shift in characteristics of the network. Thus, our goal was not to measure the error in the opinion prediction for a particular user, but to validate how accurate our NDI

measure was for the network as a whole. In particular, we compared theoretical opinion distribution for a fixed time t with the results based on sentiment analysis, as well as with the analysis of the network based on OAG.

3.4.1. *Sentiment analysis opinions.* Our model called for the use of sentiment analysis to determine the opinion polarity of the Tweets in our dataset. A company called Semantria provides access to Lexalytics Salience Engine through an online API. Let us note here that Lexalytics Salience Engine is considered to be industry-standard. We contacted Oleg Rogynskyy, CEO and Founder of Semantria, who was gracious enough to provide us with enough API calls for academic use to complete this project.

The Semantria API service returns a sentiment score for each document (in our case, each Tweet) and claims to provide 65-70% precision out of the box (without any model training). Upon inspection of the sentiment scores and the corresponding Tweets, this is surely a believable number. We are not surprised that Lexalytics Salience Engine powers 4 out of the top 5 largest social media monitoring systems.

For each time period, our model called for the determination of a sentiment score for each user on a scale of 0.0 (for very negative) to 1.0 (for very positive) where 0.5 represents a neutral sentiment. We first normalized the sentiment scores returned by the Semantria API to fit to our scale. Next, we computed the average of the sentiment scores returned by the Semantria API across all Tweets by a particular user during each time period.

3.4.2. *OAG based opinions.* We constructed OAG for different timestamps to validate if predicted NDI reflected the shift in the opinions during the discussion. Our process of constructing the OAG and the following process of assigning individual opinions is described in section 3.1.

4. Results

4.1. **Topics.** We used methodology described in section 3 to predict dynamics for the discussions about two events that took place in 2009:
(1) President Obama's speech to Congress in Sept 2009, where he offered the outline for the health care reform;
(2) death of Michael Jackson on June 25th, which spawned an extremely emotional response across all social media platforms.

For both events we were able to find representatives of polar opinions about the event. For the first topic we found people supporting and rejecting the reform. In the second case two contrary sides of the discussion were fans vs. people who resented the canonization of the singer by his fans.

4.2. **Data.** The analysis was performed on historical Twitter data available at snap.stanford.edu. Available Twitter data set was filtered to collect posts regarding topic of interest. Below is the overview for the data sets used for this project:

Topic	Time	Period	Number of tweets related to topic	Number of re-tweets	Unique users
Health care	$t = 0$	Sept-Oct 2009	327847	74975	31679
Health care	$t = 1$	Nov-Dec 2009	167497	42719	28379
Michael Jackson	$t = 0$	Jul 2009	2262255	31090	34233
Michael Jackson	$t = 1$	Aug 2009	239247	29722	26777

Table 1. Project data statistics

We collected Twitter data for two time intervals for each topic. Each interval represented timestamp $t = 0$ and $t = 1$ accordingly. The data in the interval $t = 1$ was used to compare model opinion distribution with empirical values based on OAG as well as sentiment analysis. Sentiment analysis data in the interval $t = 0$ was used to validate initial setting for the opinion x_i for $t = 0$.

4.3. **Parameters of the constructed OAG.** We now give the parameters of the OAG graphs that we got for the above described topics. Let n and m be the number of nodes and edges in the original Twitter endorsement data, and let m' be the number of nodes and edges in the corresponding OAG (note that the number of nodes in both graphs coincide). Let us note that we can only calculate opinions and biases for the nodes that are in the same components as the attractors (see section 3 for details). Our experiments show that for our data both attractors always belong to the largest connected component (CC). Table 2 shows the OAG statistics as well as the size of the largest CC . We note that the largest CC contains about half the nodes but most of the edges.

Dataset	time	n	m	m'	num. nodes in the largest CC	num. edges in the largest CC
Health care	$t = 0$	31679	74975	1486574	18948	1483189
Health care	$t = 1$	28379	42719	1019818	18019	1014343
Michael Jackson	$t = 0$	34233	31090	977101	13113	967262
Michael Jackson	$t = 1$	26777	29722	885410	11358	875468

Table 2. OAG statistics

5. Simulation results

We now describe results of our simulations. At a high level the simulations suggest two conclusions. First, we observed a significant dependence of the NDI behavior on the choice of the bias vector. We believe that a promising direction toward stabilizing these results is to develop an algorithm for calculating bias that would allow tighter control of the dynamic range of bias values. We give more details on this in section 5.1 below.

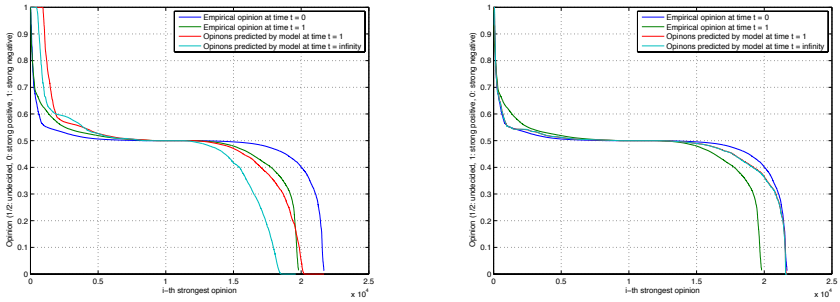
Our second conclusion is that, strikingly, opinion calculation using sentiment analysis gives very different results from our random walk based techniques (please see Fig. 4(a)). It would be very interesting to determine the cause of such behavior and find a more consistent set of methods for estimating opinions.

5.1. **Effect of bias vector on simulation results.** Our simulations indicate a significant dependence of the behavior of the model on the choice of the bias parameters. As follows from the work of [2], the evolution of a node's opinion depends on whether its bias is above or below 1. Thus, we simulated the model with different scaling factors in the expression for bias in (3.3). In some of the experiments we observed a remarkable match between the opinions predicted via our random walk based method described in section 3.2 (this happened for the Healthcare dataset with the scaling factor $F = 2$ in Fig. 1(a), the green and red curves) and the model predictions. On the other hand, experiments show that choosing the scaling factor $F = 8$ (with all other parameters equal) leads to no polarization. This suggests that further investigation of methods for evaluating user bias are needed in order to obtain a closer match between the model and real data. In would be particularly interesting to obtain an estimation scheme that would allow one to control the range of bias estimates. While we currently achieve this partially by using the damping factor F, tighter control seems necessary since the bias has exponential influence on the model of [2], resulting in high sensitivity. We now give the plots.

Healthcare dataset. The distribution of opinions for scaling factor equal to 2 and 8 is given in Fig. 1(a) and Fig. 1(b). Remarkably, our random walk based estimates for opinion in Fig. 1(a) match the model prediction quite closely. The model predicts polarization for $F = 2$ (Fig. 1(a)) and no polarization for $F = 8$ (Fig. 1(b)).

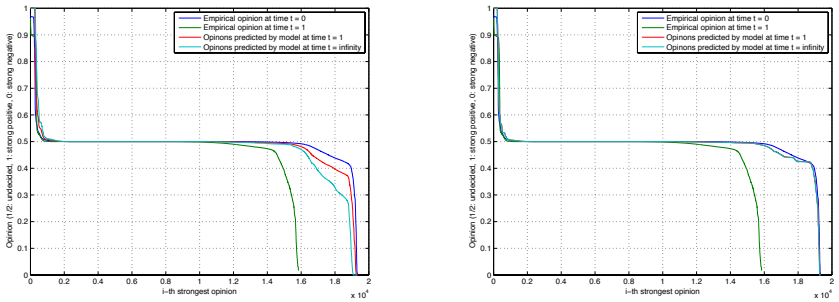
Michael Jackson dataset. The distribution of opinions for scaling factor equal to 2 and 8 is given in Fig. 2(a) and Fig. 2(b). We note that the model predicts polarization for $F = 2$ (Fig. 2(a)) and no polarization for $F = 8$ (Fig. 2(b)). Thus, simulations confirm that scaling the bias parameter is crucial for getting meaningful results.

5.2. **Comparison against sentiment analysis.** Another interesting conclusion from the simulation results is the following. As described in section 3.4, we ran the model using two sets of input opinion data: one given by our random walk based estimation and another given by sentiment analysis. In these experiments the OAG, edge weights and bias values were fixed, and only the input opinions were varied. The outcomes predicted by the model



(a) (b)

FIGURE 1. (a): Polarization for Health Care dataset with scaling factor $F = 2$
(b): No polarization for Health Care dataset with scaling factor $F = 8$

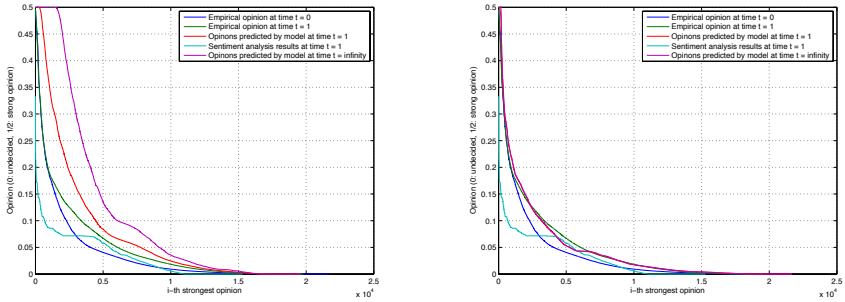


(a) (b)

FIGURE 2. (a): Polarization for Michael Jackson dataset with scaling factor $F = 2$
(b) No polarization for Michael Jackson dataset with scaling factor $F = 8$

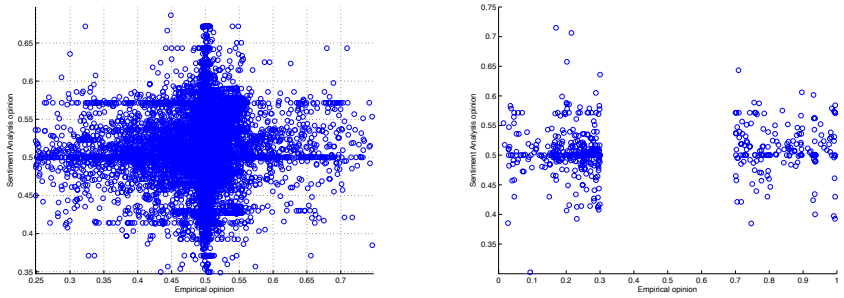
were drastically different, predicting polarization in the case of random walk based techniques and no polarization in sentiment analysis. This is despite the fact that the *distribution* of opinions derived from sentiment analysis matches the distribution of opinions derived using our random walk techniques quite closely. This is illustrated in Fig. 3(a), where we plot the folded opinions, i.e. instead of the opinion $x_i \in [0, 1]$ we plot the value $|1/2 - x_i|$. The reason for this is that because the sentiment analysis measures the strength of an individual's opinion, but does not provide information as to which of the two attractors the person is leaning towards. Fig. 3(a) also shows the opinion that each user converges to with time.

We did further investigation in the cause of such behavior and found out that the two methods of initialization give drastically different results. This is illustrated by the scatterplot in Fig. 4(a). Each point corresponds to a node x in the graph G_{OAG}, and the coordinates of x are the two opinions assigned to x by sentiment analysis and random walk based initialization respectively. Thus, if the initialization values were correlated, the points would cluster around the line $x = y$. However, we see that points are spaced quite uniformly and only cluster around lines $x = 1/2$ and $y = 1/2$. Such behavior is also true even for the range where the x values are far from $1/2$, i.e. when the random walk based opinions indicate a quite strong opinion.We thus conclude that these two methods of estimating opinions give drastically different results, and it is hence not natural to expect the outputs of the model



(a) (b)

FIGURE 3. (a): Opinion distribution for Health Care dataset ($F = 2$): folded plot $|1/2 - x_i|$ (b) Opinion distribution for Health Care dataset ($F = 8$): folded plot $|1/2 - x_i|$



(a) (b)

FIGURE 4. (a): Comparison of sentiment analysis based opinions and empirical opinions (see section 3.2) (b) Zoomed in at a confident interval of empirical opinions

to agree. It would be interesting to perform a more detailed investigation of this phenomenon and compare the two methods of initialization more carefully.

6. Conclusion

We implemented the Dandekar's model of opinion formation. We formalized the process of setting the initial parameters required for the model, and thus, our work can be generalized for the broader topics and various data sets. We showed that the model can be used to predict what network opinion converges to with time. We also provided insight into how initial bias affects the dynamics of the discussion. Being able to set optimal bias parameters as well as getting error margin for the calculated NDI is one of the possible directions of our work.

References

[1] P. H. C. Guerra, A. Veloso, W. Meira, Jr., and V.ílio Almeida. From bias to opinion: a transfer-learning approach to real-time sentiment analysis. KDD, 2011.

[2] Pranav Dandekar, Ashish Goel, and David Lee. Biased assimilation. homophily and the dynamics of polarization. WINE, 2012.