Role of Conformity in Opinion Dynamics in Social Networks

Abhimanyu Das
Microsoft Research
Mountain View, CA
abhidas@microsoft.com

Sreenivas Gollapudi
Microsoft Research
Mountain View, CA
sreenig@microsoft.com

Arindam Khan
School of Computer Science
Georgia Tech, Atlanta, GA
akhan67@gatech.edu

Renato Paes Leme
Google Research
New York, NY
renatoppl@google.com

ABSTRACT

Social networks serve as important platforms for users to express, exchange and form opinions on various topics. Several opinion dynamics models have been proposed to characterize how a user iteratively updates her expressed opinion based on her innate opinion and the opinion of her neighbors. The extent to how much a user is influenced by her neighboring opinions, as opposed to her own innate opinion, is governed by a measure of her "conformity' parameter. Characterizing this degree of conformity for users of a social network is critical for several applications such as debiasing online surveys and finding social influencers. In this paper, we address the problem of estimating these conformity values for users, using only the expressed opinions and the social graph. We pose this problem in a constrained optimization framework and design efficient algorithms, which we validate on both synthetic and real-world Twitter data. Using these estimated conformity values, we then address the problem of identifying the smallest subset of users in a social graph that, when seeded initially with some non-neutral opinions, can accurately explain the current opinion values of users in the entire social graph. We call this problem seed recovery. Using ideas from compressed sensing, we analyze and design algorithms for both conformity estimation and seed recovery, and validate them on real and synthetic data.

1. INTRODUCTION

The widespread use of online social networks has a very direct bearing on how users form and express opinions on various issues such as politics, technology, consumer products, healthcare etc. Though users are increasingly spending more time on these networks, not all of them adopt and propagate ideas and opinions in a homogeneous way. Understanding how users in an online social network shape and influence each other's opinions is important in the context of viral marketing, behavioral targeting and information dissemination of users in the network.

There has been a plethora of work on modeling opinion dynamics in social networks [7, 10, 29, 21] in the sociology and economics literature. Specifically, these works model how an individual user updates her opinion in the context of information learned from her neighbors in her social network, and then use the model to characterize the evolution of opinions in the network in terms of equilibria, convergence time, and the emergence of a consensus or polarization.

More recent models [11, 1, 20] explicitly incorporate the notion of an *innate opinion* – i.e., an endogenous opinion of each user about a certain topic – as opposed to her *expressed opinion*, which is a result of a social process. These models describe how the opinion of each user in a social network evolves (i.e. how she changes her opinion over time) based on the following: (i) her innate opinion, which is an immutable property of the user, (ii) the expressed opinion of users in her social circle and (iii) internal parameters of the model. In the case of the Friedkin and Johnsen [11] model, the model parameters correspond to a real quantity for each user which measures her *conformity*, i.e., the likelihood of this user adopting the opinions of her neighbors in the network.

While the expressed opinion of users are readily observable in the social network (e.g. from a user's tweets or Facebook posts), both the innate opinion and conformity parameter of users are hidden and can only be inferred by reverse engineering the opinion dynamics process. In particular, identifying these endogenous conformity values of users are a critical component toward enabling several interesting applications that leverage opinion dynamics processes. Salient examples include efficient sampling of innate user opinions [4], seeding opinions to maximize opinion adoption [12], and identifying candidates for viral marketing or targeting.

Inferring the conformity parameter for each user by simply analyzing the history of opinions emitted by the user is not practical largely due to lack of availability of a reasonable amount of historical opinion formation data for each user. Therefore, we ask the following question, that we call ConformityExtraction : *Can we reliably estimate conformity parameters of users in a social network from a single snapshot of the stationary state of the opinion dynamic?*

Mathematically this is an under-determined problem since it involves n constraints (the fact that the dynamic is in a stationary state for each user) and $2n$ variables (the innate opinion and conformity parameter for each user). We overcome this problem by relying on two natural assumptions, *viz.*,

- *homophily*: we assume that if two users are close in the social network, then their innate opinions are likely to be also close.
- *access to a coarse conformity distribution*: we assume that we have access to a coarse-grained estimate of the distribution of conformity parameters in the entire network.

We next define a novel problem, which we call SeedRecovery , that is fundamentally different from the well-studied problem of influence maximization [23, 28]. While the motivating application for both the problems is finding influencers, the influencers from SeedRecovery are a small set of users whose early adoption (in the past) of an opinion has been critical in shaping the current snapshot of opinions in the network. On the other hand, the users computed via influence maximization are individual who, when currently seeded with the opinion, will be critical in shaping (or maximizing) the future spread of the opinion in the network, agnostic to any historical opinion dynamics.

We note that the study of SeedRecovery is enabled by our estimates of the conformity parameters. More formally, we are given the current state of expressed opinions in the social network and we want to identify if there is a sparse set of innate opinions in the network that can explain the current expressed opinions. This enables us to distinguish opinions that might have been seeded, i.e., are the result of a small number of planted opinions (say by means of a viral market campaign) from opinions that arise naturally. It also helps capture a measure of the "heterogeneity" or "richness" of the opinion dynamics process in a social network. In fact, we identify this problem as a special case of the sparse recovery problem that is commonly studied in the compressed sensing and signal processing literature [2, 8, 32] and for which greedy strategies are known to work well[1]. We apply a similar Greedy algorithm for the SeedRecovery problem, and validate it for a special case analytically and more general, using experimental analysis on synthetic and Twitter data.

[1]We note here that unlike the influence maximization problem, the sparse recovery objective is not submodular

Contributions of this study

Our contributions in the study are three-fold:

- We address the ConformityExtraction problem of recovering the conformity values of users in a social network, using only the stationary snapshot of the opinion dynamics, along with assumptions on homophily and a coarse-grained empirical distribution of user conformity.
- We formulate the problem of SeedRecovery in a social network, show that it is related to the well-studied sparse recovery problem, and propose a Greedy algorithm for recovering the seed set of users for the opinion dynamics.
- We perform extensive experiments on both synthetic graphs and a large set of real-world Twitter data to validate the performance of our algorithms for both ConformityExtraction and SeedRecovery .

2. RELATED WORK

The broad research area of opinion formation is quite classical, and we refer the interested reader to [21] for a survey. The earliest work in this domain comes from the sociology and statistics literature [29, 7, 10].

Several models for opinion formation and consensus have been studied in the sociology community. One notable example is the work by DeGroot [7] which studies how consensus is formed and reached when individual opinions are updated using the average of the neighborhood opinions in a network. The work of Friedkin and Johnsen (FJ) [11], is perhaps the first study to extend the DeGroot model to include both disagreement and consensus, by associating with each node an innate opinion in addition to her expressed opinion. In their model, a user adheres to her initial opinion with a certain weight α_i, while she is socially influenced by others in her network with a weight $1 - \alpha_i$.

On the subject of conformity, recent work [31] focused on computing conformity parameters under three different notions of *individual*, *peer*, and *group* conformities. We differ from this line of work in that our focus is on leveraging the underlying opinion dynamics in a social network to estimate user specific conformity parameters. Further, estimating conformity values in the context of opinion dynamics also allows us to identify sparse seeded opinions (SeedRecovery).

There has been a large body of work on modeling the adoption or spread of ideas, rumors or content among online users. Well known models in this domain include

Threshold [18], Cascade models [14], and conformity-aware cascade models [25] that specify how a node adopts a particular idea or product based on the adoption pattern prevalent in its neighborhood. Subsequently, several papers studied, both theoretically [23] and empirically [13, 16, 27], the phenomenon of diffusion of ideas or content in a social network and the related problem of identifying influential nodes to seed, in order to maximize adoption rates. Several of these papers are mainly concerned with binary-valued propagation of an idea or products where a user decides to either adopt or not adopt the idea, instead of a more continuous opinion dynamics model where a user opinion is influenced by her neighboring opinions to varying degrees. However a recent paper by Terzi et al. [12] considers influence maximization in the context of opinion adoption based on the Friedkin-Johnsen model. They pose the problem of selecting a small set of nodes and seeding them with a single positive opinion to maximize the adoption of the overall positive opinion in the network. They show that the resulting problem is submodular and can hence be maximized efficiently. However, they assume knowledge of the user conformity values in the opinion dynamics model and do not address the problem of how to estimate these parameters.

As mentioned previously, the SeedRecovery problem that we introduce is fundamentally different from the above influence maximization problems, since the goal is not to seed users with products or opinions to maximize adoption, but rather to understand if the current state of expressed opinions in the social network can be explained (from the opinion dynamics process) using the opinions of a small set of seed nodes, and if so, to recover these seed nodes. Several results in the sparse recovery literature have shown [2, 8, 32, 5] that greedy and L_1-relaxation techniques can recover the linear combination efficiently as long as the matrix formed by taking the vectors x_i as columns is well-conditioned and k is sufficiently small. However, we are not aware of any prior application of sparse recovery techniques for opinion formation problems.

In other related work, Das et al [4] addresses the problem of sampling users in a social graph to estimate the average innate opinion of users, using only the expressed opinion of the sampled nodes. However, they too require knowledge of the per-user conformity values for their sampling algorithms.

For the cascade-based models for diffusion of an idea or product-adoption, the problem of estimating the adoption probabilities of a user has been studied in [17], [9] and [30]. Most of these papers use a probabilistic model, along with historical data of user adoption activity, to estimate adoption probabilities for each edge in the social graph. However, for the case of social opinion dynamics, to the best of our knowledge, we are not aware of any related work for estimating the user conformity parameters for applications that use opinion dynamics models.

3. OPINION MODEL

We consider a (possibly directed) social network graph $G = (V, E)$ with nodes $V = \{1, 2, \ldots, n\}$, corresponding to individuals and edges E corresponding to social interactions. We will say that $(i, j) \in E$ if i *is influenced by* j. We will denote by $N_i = \{j; (i, j) \in E\}$, the set of neighbors of node i and $d_i = |N_i|$, the out-degree of node i.

It will be convenient to express the graph in terms of its adjacency matrix $\mathbf{A}$, which corresponds to an $n \times n$ matrix such that $\mathbf{A}_{ij} = 1$ if $(i, j) \in E$ and $\mathbf{A}_{ij} = 0$ otherwise. We will also use $\mathbf{I}$ to represent the $n \times n$ identity matrix, $\mathbf{1}$ to represent the vector in $\mathbb{R}^n$ with all components 1 and given a vector $v \in \mathbb{R}^n$, we will represent by $\mathtt{dg}(v)$ the matrix with the components of v in the diagonal and zero elsewhere.

We are interested in studying opinion formation processes in social networks. We will distinguish between an agent's *innate opinion*, which reflects the agent's interval belief, and the agent's *expressed opinion*, which is the opinion an agent chooses to express in the network as a result of a social influence process. Here we encode the opinions as a single real quantity. Let $y_i^t \in \mathbb{R}$ be the opinion *expressed* by node i on time t. We express by z_i the innate opinion of agent i.

The classic model due to Friedkin and Johnsen [11] proposes a dynamic governing the opinion formation process. In their model, each agent is associated with conformity parameter α_i in the $[0, 1]$ range, which measures how strong her innate opinions are, and how likely will she be influenced by her neighborhood opinions. An α_i value close to 1 implies that the individual is highly opinionated, and her expressed opinion is similar to her innate opinion. While a value close to 0 implies that the individual has a very weak innate opinion and consequently her expressed opinion is largely governed by the opinions of neighbors around her. According to their model, in every timestep, each agent updates her expressed opinion to a convex combination between her innate opinion and the average of expressed opinions of her neighbors in the previous timestep. The weight of each term in the

$$y_i^{t+1} = \alpha_i \cdot z_i + (1 - \alpha_i) \cdot \frac{1}{d_i} \sum_{j \in N_i} y_j^t \tag{1}$$

In matrix form, we can re-write it as:

$$y^{t+1} = \mathtt{dg}(\alpha) \cdot z + (\mathbf{I} - \mathtt{dg}(\alpha)) \cdot \mathtt{dg}(d)^{-1} \mathbf{A} y^t \tag{2}$$

It has been shown in [4] that the above opinion formation dynamics converge to an equilibrium that depends only on the innate opinions and the structure of the network and not on the original opinions, as long as $\alpha_i > 0$ for all $i \in V$, i.e., each individual holds an innate opinion that has some impact on what they express.

The equilibrium can be obtained as the unique fixed point of equation (2), i.e.:

$$y = (\mathbf{I} - (\mathbf{I} - \mathtt{dg}(\alpha)) \cdot \mathtt{dg}(d)^{-1}\mathbf{A})^{-1}\mathtt{dg}(\alpha) \cdot z \qquad (3)$$

We will denote by $\mathbf{F}$ the matrix governing the Friedkin-Johnsen dynamic, i.e., $\mathbf{F} = (\mathbf{I} - (\mathbf{I} - \mathtt{dg}(\alpha)) \cdot \mathtt{dg}(d)^{-1}\mathbf{A})^{-1}\mathtt{dg}(\alpha)$. Also, given any matrix $\mathbf{M}$ of size $m \times n$ and a subset $S \subseteq [n]$, we denote by $\mathbf{M}_S$ the matrix of size $m \times |S|$ corresponding to the columns of $\mathbf{M}$ with indices in S. Similarly, given a vector $x \in \mathbb{R}^n$, we will denote by x_S the vector in $\mathbb{R}^S$ corresponding to the components of x with indices on S.

4. CONFORMITY EXTRACTION

Based on the Friedkin-Johnsen opinion dynamics model described in the previous section, we now address the problem of estimating the conformity parameter α_i of a user in a social network. Since we expect that these per-user conformity parameters depend on the topic that is being opined, in the remainder of this section we formulate the conformity extraction problem in the context of a particular topic. We then present a Linear Programming based approach for this problem.

We only assume knowledge of the current steady state expressed opinions of users in the social network. In practice, these opinions cam be obtained from opinion mining [26] of recent content posted by the user on the social network, for example, her recent tweets on Twitter or posts on Facebook. We are also given the directed social graph among these users, which could correspond to, say, the Twitter follow graph or Facebook friend graph.

Using these two pieces of information, we would like to estimate each user's α_i value. Clearly, as seen from Equation 1, if we knew the user's innate opinion z_i, we could directly calculate her α_i based on how far is her expressed opinion compared to her innate opinion and the mean of her neighboring expressed opinions. However, in practice, it is very hard to glean information about an individual user's innate opinions. Hence, the main technical challenge is that we have an underdetermined system of equations where we have two unknowns per user: α_i and z_i while the opinion dynamics model only gives us one equation per user.

To overcome the underdeterminacy of this system, we devise two score functions and pick the candidate (α, z) pair that optimizes a combination of such scores:

Homophily score: We assume that users that are 'close' in the social network are also likely to have innate opinions that are close. This leads to the following score function on the innate opinions:

$$H(z) = \sum_{i,j \in V} |z_i - z_j| \cdot (1 - D_{ij}/n)$$

where D_{ij} is the distance from node i to node j in the social graph.

Conformity distribution score: We bucket the range of α_i into three categories: low $B_1 = [0, 1/3]$, medium $B_2 = (1/3, 2/3]$ and high $B_3 = (2/3, 1]$ and assume that we have a coarse-grained estimate of the empirical distribution of the conformity values across buckets: let λ_j be the estimate of what fraction of the α_i values fall in bucket j. Then, given a vector α of conformity parameters, the *log-likelihood* that this vector was generated by a distribution with bucket estimates $(\lambda_1, \lambda_2, \lambda_3)$ is given by:

$$L(\alpha) = \beta_1(\alpha) \cdot \log \lambda_1 + \beta_2(\alpha) \cdot \log \lambda_2 + \beta_3(\alpha) \cdot \log \lambda_3$$

where $\beta_i(\alpha) = \sum_{j=1}^{n} \beta_{ij}$ and $\beta_{ij} = 1$ if $\alpha_i \in B_j$ and zero otherwise.

Note that the assumption about knowledge of the coarse-grained conformity distribution across the three buckets is much weaker than an assumption about the exact distribution of the α_i values. Furthermore, in practice this is not an unrealistic assumption, since one can manually go through a small set of users who have tweeted or created a post about the topic, peruse their past postings and use human judgment to infer what fraction of them are likely to be highly conforming, moderately conforming, or stubborn.

4.1 Mathematical Program for Identifying Conformity

Based on the score functions identified, we propose identifying conformity via the following Mathematical Program:

$$\begin{aligned}
&\textsc{Minimize}_{\alpha,z} \quad H(z) - c \cdot L(\alpha) \\
&\text{s.t} \\
&\qquad y = \mathtt{dg}(\alpha) \cdot z + (\mathbf{I} - \mathtt{dg}(\alpha))\mathtt{dg}(d)^{-1}\mathbf{A}y \\
&\qquad \alpha_i \in (0,1), \qquad\qquad \forall i \in V
\end{aligned}$$

where c is a regularization constant that adjust the trade-off between the score functions, and that is set in our experiments using cross validation. By rearranging the update rule of the Friedkin-Johnsen model and denoting the average opinion of neighbors of node i by $m_i = \frac{1}{d_i} \sum_{j \in N_i} y_i$, we can write:

$$\text{MINIMIZE}_{\alpha,z} \quad H(z) - c \cdot L(\alpha)$$

S.T

$$\alpha_i(z_i - m_i) = y_i - m_i, \quad \forall i \in V$$
$$\alpha_i \in (0,1), \quad \forall i \in V$$

Observe that this is not a Linear Program since both terms α_i and z_i are variables and the product $\alpha_i \cdot z_i$ appears in the constraints. Now we describe how to change the program to get rid of the product. We note that the objective function doesn't depend on α_i directly, but rather on β_{ij} which are indicators of the event $\alpha_i \in B_j$. Therefore, we propose to substitute the constraint $\alpha_i(z_i - m_i) = y_i - m_i$ by

$$\beta_{ij} = 1 \Rightarrow \frac{y_i - m_i}{z_i - m_i} \in B_j$$

and enforce integrality constraints on β_{ij} as well as $\sum_{j=1}^{3} \beta_{ij} = 1$. We note that this doesn't affect the value of the mathematical program, since for any solution of the program on (z, β), it is possible to recover a solution (z, α) with the same objective simply by taking $\alpha_i = (y_i - m_i)/(z_i - m_i)$.

Now, we re-write the newly introduced constraints in a more amenable form. First, we enforce the constraint that the ratio $(y_i - m_i)/(z_i - m_i)$ is non-negative. Given this constraint, we can rephrase $\beta_{ij} = 1 \Rightarrow \underline{b}_j \leq \frac{y_i - m_i}{z_i - m_i} \leq \overline{b}_j$ where $B_j = [\underline{b}_j, \overline{b}_j]$ as follows:

$$\overline{b}_j \cdot \texttt{sgn}(y_i - m_i) \cdot (z_i - m_i) \geq \beta_{ij}|y_i - m_i|$$

$$\underline{b}_j \cdot \texttt{sgn}(y_i - m_i) \cdot (z_i - m_i) \leq \beta_{ij}|y_i - m_i| + K \cdot (1 - \beta_{ij})$$

where $\texttt{sgn}(x) = 1$ if $x \geq 0$ and $\texttt{sgn}(x) = -1$ otherwise and $K = c_1(\texttt{sgn}(y_i - m_i) \cdot (z_i - m_i))$ where c_1 is a constant to be chosen later. To see that those are equivalent, notice that if $\beta_{ij} = 1$, this gives exactly $\underline{b}_j \leq \frac{y_i - m_i}{z_i - m_i} \leq \overline{b}_j$. If $\beta_{ij} = 0$, then both constraints are trivially satisfied for $c_1 \geq 2/3$.

This leads to the following Integer Program:

$$\text{MINIMIZE} \quad H(z) - c \cdot \textstyle\sum_{j=1}^{3} (\sum_i \beta_{ij}) \log \lambda_i$$

S.T

$$(y_i - m_i)(z_i - y_i) \geq 0, \forall i$$
$$\overline{b}_j \cdot \texttt{sgn}(y_i - m_i) \cdot (z_i - m_i) \geq \beta_{ij}|y_i - m_i|, \forall i, j$$
$$\underline{b}_j \cdot \texttt{sgn}(y_i - m_i) \cdot (z_i - m_i) \leq \beta_{ij}|y_i - m_i| + K \cdot (1 - \beta_{ij}), \forall i, j$$
$$\textstyle\sum_{j=1}^{k} \beta_{ij} = 1, \forall i \in V$$
$$\beta_{ij} \in \{0,1\}, \forall i, j$$

Relaxation and rounding. We refer to the solution of this integer program as the IPRECOVERY . However, since solving an integer program is not feasible for large instances, we begin with an LP relaxation, LPRECOVERY , and offer a simple rounding technique for the solution of LPRECOVERY to obtain the final solution. Our approach consists of relaxing the integrality constraints to $\beta_{ij} \in [0,1]$, solving the resulting Linear Program, obtaining z and β_{ij} and then recovering the values of α_i by setting:

$$\alpha_i = \frac{y_i - m_i}{z_i - m_i}$$

Note that in the LP objective, there are no α_i's. We assign node i to bucket B_j based on its α_i value. This in a way, is a rounding of β_{ij} based on only α_i values. In the analysis, we show that original β_{ij} values from the LP and the rounded β_{ij}'s from α_i's are quite close. Therefore, the objective value resulting from the rounding is also close to the objective value of IPRECOVERY since the solution to LPRECOVERY is a lower bound to the objective function. Further, we also show empirically (see Section 6) our approach performs comparably to IPRECOVERY .

Analysis. In the above linear program, we show that the parameters β_{ij} and the α_i derived using the equation $\alpha_i = (y_i - m_i)/(z_i - m_i)$, are closely related.

LEMMA 1. $\alpha_i \in [0,1]$.

PROOF. From the constraint $(y_i - m_i)(z_i - y_i) \geq 0$, $\texttt{sgn}(y_i - m_i) = \texttt{sgn}(z_i - y_i) = \texttt{sgn}(z_i - m_i)$.
Thus y_i lies between z_i and m_i. Hence, $|y_i - m_i| \leq |z_i - m_i|$. Thus $\alpha_i = (y_i - m_i)/(z_i - m_i) \leq 1$.
On the other hand, from the same constraint we get,

$$\begin{aligned} \frac{(y_i - m_i)(z_i - y_i)}{(z_i - m_i)^2} &\geq 0 \\ \Rightarrow (y_i - m_i)/(z_i - m_i) &\geq 0 \\ \Rightarrow |y_i - m_i|/|z_i - m_i| &\geq 0 \\ \Rightarrow \alpha_i &\geq 0 \end{aligned}$$

□

LEMMA 2. *If $\beta_{i1} > c_2$ then $\alpha_i < \frac{1}{3c_2}$.*

PROOF. From $\overline{b}_j \cdot \texttt{sgn}(y_i - m_i) \cdot (z_i - m_i) \geq \beta_{ij}|y_i - m_i|$,
$\frac{1}{3}|z_i - m_i| \geq \beta_{i_1}|y_i - m_i| > c_2|y_i - m_i|$
$\Rightarrow \frac{1}{3c_2} > \frac{|y_i - m_i|}{|z_i - m_i|} = \alpha_i$. □

COROLLARY 1. *If $\beta_{i1} > 1/2$ then $\alpha_i < \frac{2}{3}$.*

LEMMA 3. *If $\beta_{i3} > c_3$ then $\alpha_i > c_1 - \frac{3c_1 - 2}{3c_3}$.*

PROOF. From $\underline{b}_j \cdot \texttt{sgn}(y_i - m_i) \cdot (z_i - m_i) \leq \beta_{ij}|y_i - m_i| + K \cdot (1 - \beta_{ij})$,
$\Rightarrow \frac{2}{3} \leq \beta_{i3} \cdot \alpha_i + c_1(1 - \beta_{i3})$.
$\Rightarrow (c_1 - \frac{2}{3}) \geq \beta_{i3}(c_1 - \alpha_i) > c_3(c_1 - \alpha_i)$
$\Rightarrow \alpha_i > c_1 - \frac{3c_1 - 2}{3c_3}$. □

COROLLARY 2. *If $\beta_{i3} > 1/2$ and $c_1 = 1$ then $\alpha_i > \frac{1}{3}$.*

THEOREM 1. *If $\beta_{ij} > 1/2$ for some $j \in \{1, 2, 3\}$ and derived $\alpha \in B_k$, then $|j - k| \leq 1$.*

PROOF. For $j = 2$, the theorem is trivially true. For $j = 1, 3$ the proof follows from Corollary 1 and Corollary 2. □

In words, Theorem 1 says that if β has reasonable weight (¿1/2) on some bucket B_j, then the derived bucket B_k from α is very close to B_j.

5. SEED RECOVERY

In the previous section, we discussed how to infer the conformity parameters α_i for users in a social network. The knowledge of those parameters enables various interesting applications. In this section, we discuss one such application called the SEEDRECOVERY problem.

Consider the scenario where y_i represents the opinion about a certain product. By means of a marketing campaign, a company might try to influence the general opinions on this product by planting few nodes with very high innate opinions about such products. This could be done, for example, by paying celebrities to tweet about certain products or events. The problem of how to choose a few nodes to seed an opinion on a network has been extensively studied [23, 12]. In the context of SEEDRECOVERY , we ask the following question: given the expressed opinions in a network, how likely it is that those were seeded by a small number of nodes ?

We assume that opinions are normalized in such a way that $z_i = 0$ represents a default neutral opinion, i.e., the node hasn't heard about that particular product or has neither positive nor negative innate opinions about it. Now, given a certain observed expressed opinions y, is there a vector z with a small number of innate non-neutral opinions that could have produced y as an outcome of the Friedkin-Johnsen dynamic ? In other words, for a given k, can we estimate:

$$\texttt{error}_k = \min_z \|y - \mathbf{F}z\|_2^2 \text{ s.t. } \|z\|_0 = k$$

This can be cast as an instance of the *sparse recovery problem* from compressed sensing. The two main approaches to solve this problem are convex relaxation [2, 8] and greedy algorithms [32, 5]. We take the latter route and apply the well-known *Forward Regression algorithm* that was analyzed in [5]. Davis et al. [6] showed that the problem of minimizing $\texttt{error}_k$ admits no multiplicative approximation, by showing that it is NP-hard to check for a given instance if $\texttt{error}_k = 0$. As a result, approximation guarantees for this problem are usually given in terms of the squared multiple correlation or R^2 objective ($\mathrm{R}^2 = 1 - \texttt{error}_k/\|y\|_2^2$), which is a well known measure of goodness-of-fit in statistics [22].

We note that since $0 \leq \texttt{error}_k \leq \|y\|_2^2$, the R^2-objective is in the $[0, 1]$ range, where $\mathrm{R}^2 = 1$ corresponds to a solution of $\texttt{error}_k = 0$. We will say that a solution z is an α-approximation if $\mathrm{R}^2(z) \geq \alpha \cdot \mathrm{R}^2(z')$ for any $z' \in \mathbb{R}^n$ with $\|z'\|_0 = k$.

Now we describe the Forward Regression algorithm and derive its approximation guarantees for the SEEDRECOVERY problem. First, we note that the problem can be re-written as:

$$\texttt{error}_k = \min_{S:|S|=k} \left[\min_{z_S} \|y - \mathbf{F}_S z_S\|_2^2 \right]$$

For a fixed $S \subseteq [n]$, it is a classic result in Linear Algebra (see [15] for example) that $\|y - \mathbf{F}_S z_S\|_2^2$ is minimized by the vector $z_S = (\mathbf{F}_S^T \mathbf{F}_S)^{-1} \mathbf{F}_S^T y$, therefore the error can be written as:

$$\begin{aligned} \texttt{error}_k &= \min_{S:|S|=k} \|y - \mathbf{F}_S (\mathbf{F}_S^T \mathbf{F}_S)^{-1} \mathbf{F}_S^T y\|_2^2 \\ &= \min_{S:|S|=k} \|y\|_2^2 - y^T \mathbf{F}_S (\mathbf{F}_S^T \mathbf{F}_S)^{-1} \mathbf{F}_S^T y \end{aligned}$$

since $\mathbf{F}_S (\mathbf{F}_S^T \mathbf{F}_S)^{-1} \mathbf{F}_S^T y$ and $y - \mathbf{F}_S (\mathbf{F}_S^T \mathbf{F}_S)^{-1} \mathbf{F}_S^T y$ are perpendicular vectors. This transformation allows us to write the problem in terms of the R^2 objective as:

$$\mathrm{R}^2 = \max_{S:|S|=k} f(S) \text{ where } f(S) = \hat{y}^T \mathbf{F}_S (\mathbf{F}_S^T \mathbf{F}_S)^{-1} \mathbf{F}_S^T \hat{y}$$

and $\hat{y} = y/\|y\|_2$. The Forward Regression algorithm builds a family of sets incrementally by adding the element that provides the maximum increase in value for f. The algorithm is initialized with $S_0 = \emptyset$ and for each $k = 1 \ldots, n$, we define $S_{k+1} = S_k \cup \{i\}$ for some $i \in \mathrm{argmax}_{j \notin S_k} f(S_k \cup \{j\})$.

We will provide a two-fold validation for the Forward Regression algorithm for the Seed Recovery problem: the first is theoretical. We will use a result due to Das and Kempe [5] to give a theoretical approximation guarantee for this problem for a special case. The Forward Regression algorithm notoriously performs much better in practice than its theoretical bounds [5], however the theoretical guarantee is useful to highlight the dependency of the algorithm on parameters of the instance. In particular, we will show that the approximation guarantee improves for higher values of α_i. In Section 7, we also perform experimental validation of this algorithm: we construct synthetic instances of the problem for which the innate opinions form a sparse vector and evaluate the outcome of the Forward Regression algorithm against the ground truth.

The theoretical guarantee on the approximation of the Forward Regression algorithm can be obtained from spectral properties of the matrix $\mathbf{F}$:

THEOREM 2 (DAS AND KEMPE [5]). *For each $k =$*

$1, 2, \ldots, n$,

$$f(S_k) \geq \left[1 - \exp\left(\frac{-\lambda^k_{\min}(\mathbf{F}^T\mathbf{F})}{\max_i \|\mathbf{F}^i\|_2^2}\right)\right] \cdot \max_{S:|S|=k} f(S)$$

where F^i is the i-th column of matrix $\mathbf{F}$ and $\lambda^k_{\min}(\mathbf{F}^T\mathbf{F})$ is the smallest k-sparse eigenvalue of the matrix $\mathbf{F}^T\mathbf{F}$, i.e., $\lambda_{k,\min}(\mathbf{F}^T\mathbf{F}) = \min_{x\in\mathbb{R}^n\setminus 0, \|x\|_0=k} \|\mathbf{F}x\|_2^2/\|x\|_2^2$

In what follows, we provide a lower bound on the exponent $\lambda^k_{\min}(\mathbf{F}^T\mathbf{F})/\max_i \|\mathbf{F}^i\|_2^2$ for the special case of regular undirected graphs and uniform α_i values, i.e., we will assume that α_i is the same for all i and that the graph is undirected and all nodes have the same degree. Since α_i is the same and d_i is the same for all nodes, we drop the subscript i for the rest of the section. The Friedkin-Johnsen matrix $\mathbf{F}$ in this case is symmetric and can be written as:

$$\mathbf{F} = \alpha \cdot \left(\mathbf{I} - \frac{1-\alpha}{d} \cdot \mathbf{A}\right)^{-1} = \alpha \sum_{k=0}^{\infty} (1-\alpha)^k \left(\frac{1}{d}\mathbf{A}\right)^k$$

using the matrix identity $(\mathbf{I} - \mathbf{M})^{-1} = \sum_{k=0}^{\infty} \mathbf{M}^k$. From this we can observe that the entries of $\mathbf{F}$ are non-negative. We use this fact to show the following result:

LEMMA 4. *If $\mathbf{F}$ is the Friedkin-Johnsen matrix associated with an undirected regular graph with degree d and uniform values $\alpha_i = \alpha (> 0)$, then all columns of the matrix $\mathbf{F}$ have their norm bounded by 1.*

PROOF. Let $\mathbf{1}$ be the vector with all components equal to 1. Then if $\mathbf{A}$ is the matrix associated with the regular graph, then clearly $\frac{1}{d}\mathbf{A}\mathbf{1} = \mathbf{1}$. Therefore $\mathbf{F}\mathbf{1} = \alpha \sum_{k=0}^{\infty} (1-\alpha)^k (\frac{1}{d}\mathbf{A})^k \mathbf{1} = \mathbf{1}\alpha \sum_{k=0}^{\infty} (1-\alpha)^k = \mathbf{1}$. Therefore, for all rows i, $\sum_j \mathbf{F}_{ij} = 1$. Since the entries are non-negative, it means that all entries are in the $[0, 1]$ range. Finally, we use the symmetry of $\mathbf{F}$ to see that:

$$\|\mathbf{F}^i\|_2^2 = \sum_j \mathbf{F}_{ij}^2 \leq \sum_j \mathbf{F}_{ij} = 1$$

□

The remaining term in Theorem 2 is $\lambda^k_{\min}(\mathbf{F}^T\mathbf{F})$, which we bound by the smallest eigenvalue of $\mathbf{F}^T\mathbf{F}$:

$$\lambda^k_{\min}(\mathbf{F}^T\mathbf{F}) \geq \lambda_{\min}(\mathbf{F}^T\mathbf{F}) = \min_{x\neq 0} \frac{\|\mathbf{F}x\|_2^2}{\|x\|_2^2}$$

In the subsequent proof, we use the concept of the operator norm of a matrix. Given a square matrix $\mathbf{M}$, we define its operator norm $\|\mathbf{M}\|_2 = \max_{x\in\mathbb{R}^n\setminus\{0\}} \|\mathbf{M}x\|_2/\|x\|_2$ and use the following matrix inequalities:

$$\|\mathbf{M}_1 + \mathbf{M}_2\|_2 \leq \|\mathbf{M}_1\|_2 + \|\mathbf{M}_2\|_2 \quad (4)$$

$$\|\mathbf{M}_1 \cdot \mathbf{M}_2\|_2 \leq \|\mathbf{M}_1\|_2 \cdot \|\mathbf{M}_2\|_2 \quad (5)$$

We refer to [15] for an extensive exposition on matrix norms and spectral properties of matrices.

Also, given the adjacency matrix $\mathbf{A}$ of a regular graph of degree d. We will use the fact [3] that the operator norm of the adjacency matrix is at most the maximum degree of its vertices, hence,

$$\|\frac{1}{d}\mathbf{A}\|_2 \leq 1. \quad (6)$$

LEMMA 5. *If $\mathbf{F}$ is the Friedkin-Johnsen matrix associated with an undirected regular graph d and uniform values $\alpha_i = \alpha(> 0)$, then*

$$\lambda_{\min}(\mathbf{F}^T\mathbf{F}) \geq \frac{\alpha^2}{4}$$

PROOF. Since $\alpha > 0$ then $\mathbf{F}$ is invertible, therefore:

$$\lambda_{\min}(\mathbf{F}^T\mathbf{F}) = \min_{x\neq 0} \frac{\|\mathbf{F}x\|_2^2}{\|x\|_2^2} = \min_{y\neq 0} \frac{\|y\|_2^2}{\|\mathbf{F}^{-1}y\|_2^2}$$
$$= \left[\max_{y\neq 0} \frac{\|\mathbf{F}^{-1}y\|_2}{\|y\|_2}\right]^{-2} = \|\mathbf{F}^{-1}\|_2^{-2}$$

By the definition of $\mathbf{F}$ we have that:

$$\mathbf{F}^{-1} = \alpha^{-1} \cdot (\mathbf{I} - (1-\alpha) \cdot \frac{1}{d}\mathbf{A})$$

Therefore:

$$\|\mathbf{F}^{-1}\|_2 \leq \alpha^{-1}(\|\mathbf{I}\|_2 + (1-\alpha)\|\tfrac{1}{d}\mathbf{A}\|_2) \leq \alpha^{-1}(1 + 1 - \alpha) \leq 2\alpha^{-1}$$

Here the first inequality follows from the matrix inequalities 4 and 5 and the second inequality follows from inequality 6. Thus, $\lambda_{\min}(\mathbf{F}^T\mathbf{F}) \geq \frac{\alpha^2}{4}$ □

Thus from Theorem 2, Lemma 4 and 5, we get the following theorem:

THEOREM 3. *For each $k = 1, 2, \ldots, n$,*

$$f(S_k) \geq \left[1 - \exp\left(-\alpha^2/4\right)\right] \cdot \max_{S:|S|=k} f(S).$$

6. CONFORMITY EXTRACTION EXPERIMENTS

In this section, we present the results for our experiments for the CONFORMITYEXTRACTION problem on both real world Twitter data and synthetic data.

6.1 Conformity in Synthetic Data

We first conduct synthetic experiments (where we have complete access to ground truth and can therefore obtain fine-grained validation) to show that our algorithms described in Section 4 can extract conformity values with high accuracy. Toward this end, we generated graphs having regular, random and power law degree distributions. The number of nodes in the graph was varied from 100 to 1000. The degree in the regular graph case was set to 20 while the maximum degree in the power-law and random graph was set to 100.

Every node was assigned one of 10 innate opinions in $\{0, 1, 2, \ldots, 9\}$. To capture a homophily effect on the innate opinions, we imposed a Lipschitz constraint on the assignment of innate opinions to nodes, such that for 85% of the graph edges, the difference in opinions between the nodes forming the edge is at most 1.

For assigning α values to nodes, we used three different distributions: α values distributed uniformly in $[0, 1]$, α values taken from a power-law distribution with most of the values close to 0, and a bimodal Gaussian distribution with peaks close to 0 and 0.5. Finally, we note that all results are averaged over 10 runs.

Using the α distribution and innate opinions of nodes in the graph, we ran 5000 rounds of the Friedkin-Johnsen opinion dynamics, and considered the final opinions of the nodes as their expressed opinions.

Based on these expressed opinions and the graph structure, we then estimated the α values of all nodes in the graph, using our LPRecovery and IPRecovery algorithms for ConformityExtraction described in Section 4. For all our experiments, we used a commercial optimization solver [19] to run LPRecovery and IPRecovery . To measure the effectiveness of our algorithms, we compared the estimated α values with the ground-truth α parameters. We first categorized the nodes into three buckets based on their ground truth α values: low, medium, and high corresponding to the ranges $[0, 1/3]$, $(1/3, 2/3]$, and $(2/3, 1]$ respectively. Then, we plotted the performance of our algorithms (separately for each bucket and also for for the overall set of nodes) using two metrics: 1) a fine-grained measure corresponding to the absolute error between the ground truth α and the estimated α; 2) a coarse-grained measure corresponding to the accuracy or percentage of nodes for which we estimated their buckets correctly.

Figure 1 illustrates the results for random graphs using the LPRecovery algorithm. We varied the number of nodes in $\{100, 250, 500, 750, 1000\}$ and ran the experiments with the three different types of α distributions. In almost all the cases, the absolute error in estimating the conformity parameters was less then 0.12. In general, we observe that the accuracy of our algorithm is slightly better for the sparse α distribution where most of the nodes have low α values, than with other distributions. We also note that the estimation error for the α values is lower for large graphs than for small graphs. This is likely due to the fact that as we increase the number of nodes in the random graph, the expected degree of a node increases, which then strengthens the homophily assumptions used by the linear program to prune its feasible solution space.

The results for the accuracy measure are qualitatively similar. As seen from the figure, we recover at least 80% of the values in each bucket. Again, in the natural case of sparse α distribution, this number increases to 90%.

We observe that we obtained qualitatively similar results for the case of regular and power-law graphs and for the case where we used the IPRecovery algorithm instead of LPRecovery . The results are omitted due to lack of space.

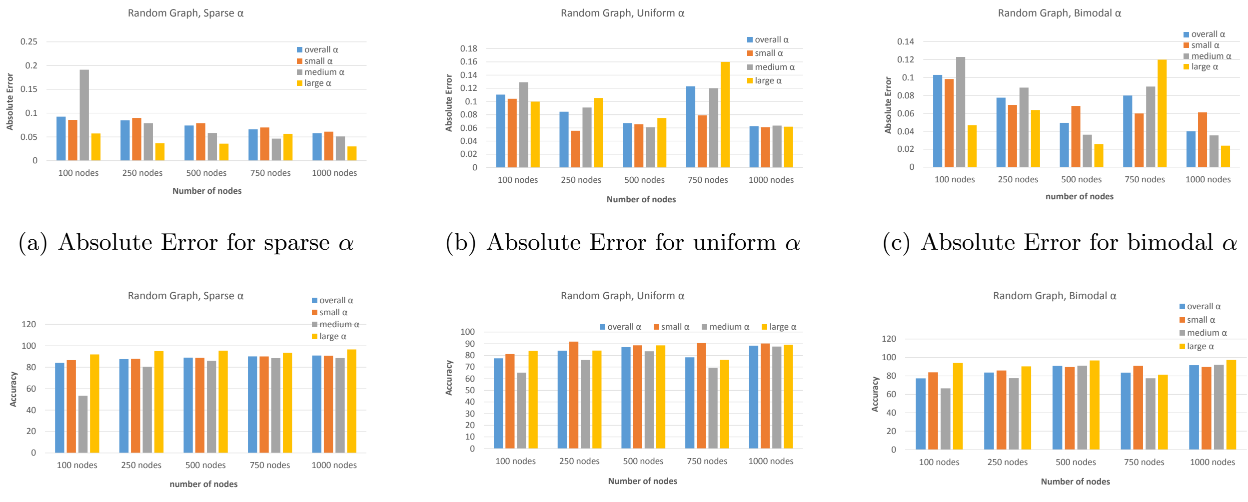
6.2 Conformity in Twitter Data

Next, we ran our algorithms for ConformityExtraction on real world social network data from Twitter. Our data set comprised of user opinions extracted from a large set of tweets corresponding to one of three topics, namely: *organic food*, *weight loss* and *electric cars*. We considered all tweets related to these topics (using simple keyword-based classifiers) within a 6-month timeline from 12/1/2012 to 5/31/2013. The total number of tweets in each topic varied between 100000 and 2000000. We then ran each tweet through a commercial sentiment analyzer [24] to obtain sentiment values ranging from -1 (corresponding to negative sentiment) to 1 (corresponding to positive sentiment), which was then smoothed into one of 10 opinion buckets. We treated the median of a user's last three opinion values as her expressed opinion. Using the Twitter follow graph, we obtained the induced subgraph over the nodes (across all topics) with around 1 million nodes and 100 million edges. As before, our goal is to extract α values of the users using the expressed opinions and graph structure.

One significant difference in this experiment is that, unlike the synthetic experiments, there is no explicit ground truth α values available for the users. To overcome this problem, we identified, using tweet histories, a small set of ground truth consisting of users with high α and low α. To obtain this ground truth for each topic, we first extract users with at least 5 tweets on the topic, and have at least 5 neighbors in the Twitter graph who have also tweeted about the topic. We categorize a user's tweet into a positive, negative or neutral opinion based on the sentiment value. We then define a stubborn user (high α bucket) as one whose opinions differs from her majority neighboring opinion for ϵ fraction of her tweets. Similarly a user is conforming (low α bucket) if her final opinion is different from her initial opinion and whose set of opinions is the same as the final majority opinion in her neighborhood for γ fraction of her tweets. We ignored the nodes that do not satisfy these conditions in our analysis. For our experiments we set ϵ to be 0.7 and γ to be 0.3.

Note that the above method to compute the ground truth cannot be used as a general algorithm to extract conformity values for *all* users since only a small set of users satisfy the aforementioned criteria to reliably measure their conformity. This necessitates the design of algorithms such as the ones proposed in this paper.

For each of the topics, we ran our IPRecovery and LPRecovery algorithms using the expressed opinions and the induced graph structure on all Twitter users



(a) Absolute Error for sparse α (b) Absolute Error for uniform α (c) Absolute Error for bimodal α

(d) Accuracy for sparse α (e) Accuracy for uniform α (f) Accuracy for bimodal α

Figure 1: Effectiveness of LPRecovery on random graphs

Topic	Algorithm	Percentage of stubborn nodes recovered	Percentage of conforming nodes recovered
Electric Car	LPRecovery	75	65
Electric Car	IPRecovery	77	82.5
Organic Food	LPRecovery	87.5	66.7
Organic Food	IPRecovery	89	59
Weight Loss	LPRecovery	55	65
Weight Loss	IPRecovery	55	61

Table 1: Conformity Extraction on Twitter

who have tweeted about the topic at least 5 times and have at least 5 neighbors, and estimated their conformity parameter α. This yielded around 2000 to 15000 users for each topic. From among these users, we then extracted a smaller ground truth set of stubborn users and conforming users. We considered the estimated α value for each user in the ground truth set, and used a threshold of 0.66 and 0.33 on their estimated α to predict whether these users are stubborn, conforming, or neither. We then measure the recall with respect to the stubborn user set (i.e., what fraction of ground truth stubborn users were correctly recovered by our algorithm) and similarly with respect to the conforming user set. Table 1 summarizes the results.

As the table shows, our algorithms perform well in recovering the stubborn users from the ground truth set for many instances. In particular, for the topic of organic food, our algorithms recover almost 90% of the stubborn users, but recover a lesser (60%) fraction of the conforming user. On the other hand, for weight loss, the performance of the algorithms in recovering the conforming users (65%) is better than that for the stubborn users (55%). This correlated with the observed skew toward stubborn users for *organic food* and conforming users for *weight loss* (see Figure 3).

The table above also validates empirically that the LPRecovery algorithm is a good approximation to the (much slower) IPRecovery algorithm, since the gap in performance between the two algorithms for most cases is less than 10%.

6.3 Validation of homophily in Twitter

The assumption of homophily in the innate opinions of users in a social networks is crucial in our LPRecovery and IPRecovery algorithms, since it helps us solve an under-determined system of equations. To validate this assumption, we set out to observe the difference between the innate values of neighbors in the Twitter follow graph for the three topics. For the homophily experiments, we define a user's innate opinion on a particular topic to be the average of the sentiment values associated with the first three tweets posted by the user in the 6 month time period. Recall that the user's opinion values as extracted from our sentiment analyser are bucketed into $\{0, 1, 2, \ldots, 9\}$. Table 2 reports the average difference in opinions between every pair of users connected by an edge in the Twitter graph, as well as the total fraction of edges in the graph for which this difference in opinions is less than 1. As seen in the table for our topics of interest, we observe that for more than 64% of the users, the difference in opinions across an edge is less than 1. Furthermore the average difference between a pair of neighboring users is less than 1 for all the topics.

7. SEED RECOVERY EXPERIMENTS

In the next set of experiments, we used the α values

Topics	Avg gap in neighboring opinions	Fraction of edges with gap < 1
Electric Car	0.75	64%
Weight Loss	0.79	69%
Organic Food	0.60	83%

Table 2: Homophily in Twitter data

from our previous ConformityExtraction experiments to address the SeedRecovery problem for both synthetic as well as Twitter graphs. Thus, we would like to compute a small set of nodes with a given innate opinion (and assuming neutral innate opinion on all other nodes) that can best explain the current expressed opinions in the network resulting from Friedkin-Johnsen dynamics. As described in Section 5, we measure the discrepancy in the predicted expressed opinions using this seed set versus the ground truth expressed opinion across all the nodes in the graph. As mentioned earlier, for measuring this discrepancy, we use the squared multiple correlation (R^2) metric, which lies in $[0, 1]$ and is essentially equal to $1 - L2Error$, where $L2Error$ is the normalized L_2 norm error between the predicted expressed opinion vector and the ground truth expressed opinion vector. Our goal is to recover a seed set of k nodes in the graph that can maximize the R^2 measure.

We use the Greedy algorithm (defined in Section 5) to recover the best seed sets of size k (we vary k from 1 to 20) for both synthetic and Twitter data. We also report the characteristics of the computed seed set as we vary its size k. We compare the performance of the Greedy algorithm against two natural baselines that have been used in similar problems ([23]): selecting nodes with the highest α-values and selecting nodes with the highest degrees in the graph.

7.1 Synthetic Data

Similar to Section 6, we generated synthetic graphs of 1000 nodes with regular, random and power law degree distributions, and assigned innate opinions in a similar manner as earlier. We also used the same three distributions of α values as earlier: uniform, bimodal and sparse. We only report the result of random graphs (the regular and power-law graphs had qualitatively similar results). In Figure 2(a), we plot the R^2 metric as a function of k for the various α distributions for the Greedy algorithm.

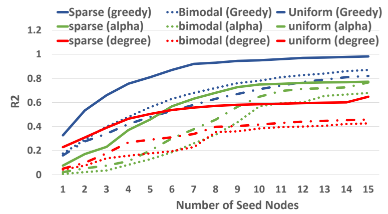
First, as expected, the R^2 increases as the size of the seed set increases. More interestingly, even for as low as 6 seed nodes, we get an R^2 value of close to 0.92 indicating a very good agreement between the original and predicted expressed opinions. For comparison, we also plot the corresponding R^2 values using the two baseline algorithms (selecting nodes according to the α values and degrees respectively) for each of the α distributions. Clearly, our Greedy algorithm performs significantly better than both the baselines. Secondly, in terms of the characteristics of the seed set selected by Greedy , we observe that the algorithm starts off by initially selecting high-degree, stubborn nodes, and then moves to nodes with lower α and degree values (Figures 2(b) and 2(d)) as the size of the seed set increases. We also measured the difference in the α values between a node in the seed set and its average neighboring α. Figure 2(c) indicates that our algorithm favors selecting seed nodes that have high α but whose neighbors have low α values. These observations agree with our intuitive expectation that the most likely seed nodes are ones that are stubborn and have a large number of conforming neighbors. Interestingly, similar behaviour has also been observed previously in [4] in the context of selecting nodes to best estimate the average innate opinion in the network.

In a more direct experiment (Figures 2(e) and 2(f)), we "planted" 10 seed nodes in a 1000 node random graph with $\alpha = 0.95$ for all the 10 nodes. Further, these 10 nodes were initialized with non-neutral innate opinions while those of the remaining nodes had neutral opinions. As before, the α values of the remaining nodes were drawn from three different distributions – sparse, bimodal, and uniform. The goal of this experiment was to validate if our algorithm can indeed "recover" the planted seed nodes, purely based on the expressed opinions and alpha values of all the nodes. Note that the graph also contained several (based on the specific α distribution) stubborn nodes that were not seeds, and hence it is not sufficient to simply pick nodes with large α values as the seeds. This is corroborated by Figure 2(e) which shows that the R^2 value of the seed set obtained by the Greedy algorithm outperforms the α-based algorithm. The α-based algorithm in turn outperforms the degree-based algorithm, due to the fact that the seed nodes in this case have high $alpha$ values.

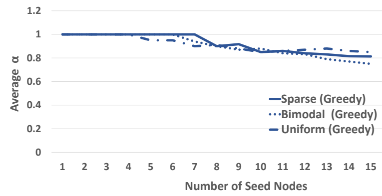
In particular, Figure 2(e) shows that the Greedy algorithm finds exactly the right set of 10 seed nodes. This is because for any size ≥ 10, the R^2 value is 1.0, implying that the selected seed set actually contains all the 10 seed nodes! This observation is further corroborated by Figure 2(f) where we see that the average value of α for the seed set of size 10 is precisely 0.95 which is indeed the α value for each of the planted stubborn node.

7.2 Twitter Data

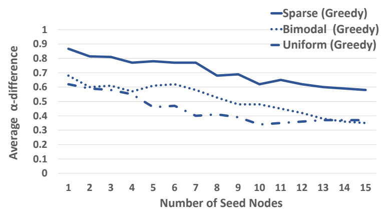
Next, we repeated the SeedRecovery experiments using Twitter data for various topics. The dataset and resulting social graph are exactly the same as in Section 6.2. As earlier, we used the α values generated



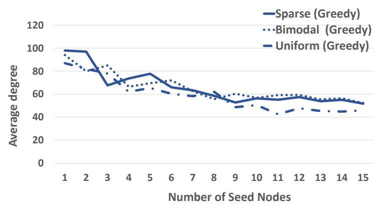
(a) Squared multiple correlation R^2



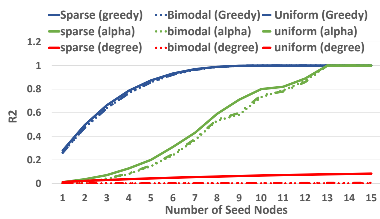
(b) Average α of the seed set



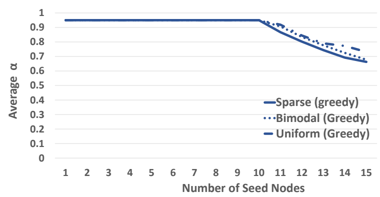
(c) Avg difference in α values of the seed set node and its neighborhood



(d) Average degree of seed set for a power-law graph



(e) Squared multiple correlation R^2 of the recovered seed set



(f) Average α of the recovered seed set

Figure 2: Characteristics of the seed set for synthetic random graphs, with uniform, bimodal, and power-law α-distributions. Figures 2(e) and 2(f) show results for the "planted" seed set experiment.

from the conformity extraction experiments described in Section 6.2. Figure 3 shows the distribution of these α values for different topics. The α distributions for all the topics resemble either sparse or bimodal distributions, and we remind the reader that we covered both of them in our simulations.

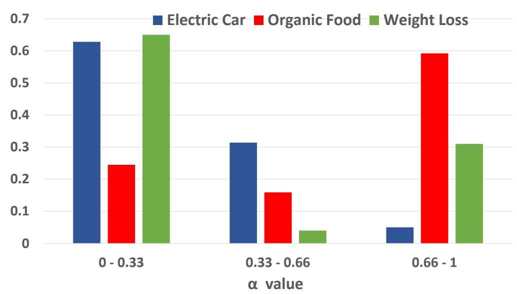


Figure 3: Distribution of α across the three buckets

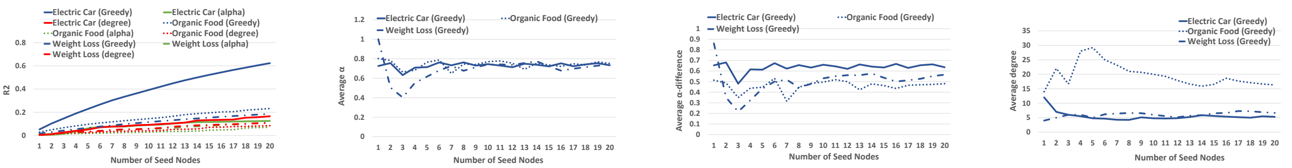
The results are summarized in Figure 4. Qualitatively, even for this data, we observe similar results to the simulations. The R^2 of the selected seed set is much larger than the α-based and degree-based baseline algorithms, for all the topics. Similarly, based on the plots showing the average α value, average degree, and average neighborhood difference in α values for the seed set, the Greedy algorithm shows a clear preference for selecting seed nodes that are moderately stubborn and have a large number of conforming neighbors. (Note that just selecting seed nodes based on high alpha values alone does not suffice, as shown by the corresponding baseline performance in 4(a)).

However, we do see interesting differences in the R^2 plots between various topics. For example, while the squared multiple correlation for *Electric cars* is around 0.6 for $k = 20$, it is only around 0.2 for the case of *Organic food* or *Weight loss*. This suggests that for the Twitter data, the equilibrium opinions for Electric cars is much more likely to have been generated by a small number of seed nodes, as compared to those for organic food or weight loss. We therefore surmise that the R^2 measure of the seed set for a topic might provide insights into a notion of how "heterogenous" or "diverse" is the opinion dynamics for that topic in a social network.

8. CONCLUSIONS

The notion of conformity plays a central role in shaping of users opinions in online social networks. In this study, we proposed algorithms for estimating conformity of users using only the expressed opinions of users resulting from the underlying opinion dynamics and the social graph. Under some natural conditions, we show using both simulations and Twitter data that our algorithms perform well on extracting the conformity values of the users.

Further, we propose efficient algorithms to recover the smallest set of source nodes in the graph that best explain the current distribution of opinions in the entire graph. We refer to this problem as *seed recovery* and we believe this and similar problems have many applications in running effective marketing campaigns, understanding information flow in social networks etc. As before, we validate our algorithms for this problem using both simulations and Twitter data. An interesting



(a) Squared multiple correlation R^2

(b) Average α of the seed set

(c) Average difference in α values of the seed set node and its neighborhood

(d) Average degree of seed set

Figure 4: Characteristics of seed set for the three different topics on Twitter

open question is to generalize the conformity extraction problem to other well-studied opinion dynamics models.

9. REFERENCES

[1] D. Bindel, J. M. Kleinberg, and S. Oren. How bad is forming your own opinion? In *FOCS*, pages 57–66, 2011.

[2] E. J. Candès, J. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59:1207–1223, 2005.

[3] F. Chung. *Spectral Graph Theory*. Number no. 92 in CBMS Regional Conference Series. American Mathematical Society, 1997.

[4] A. Das, S. Gollapudi, R. Panigrahy, and M. Salek. Debiasing social wisdom. In *Proceedings of the 19th ACM SIGKDD*, pages 500–508. ACM, 2013.

[5] A. Das and D. Kempe. Submodular meets spectral: Greedy algorithms for subset selection, sparse approximation and dictionary selection. In *ICML*, pages 1057–1064, 2011.

[6] G. Davis, S. Mallat, and M. Avellaneda. Adaptive greedy approximations. *Constructive Approximation*, 13(1):57–98, Mar. 1997.

[7] M. H. DeGroot. Reaching a consensus. *Journal of the American Statistical Association*, 69(345):118–121, 1974.

[8] D. Donoho. For most large underdetermined systems of linear equations, the minimal 11-norm near-solution approximates the sparsest near-solution. *Communications on Pure and Applied Mathematics*, 59:1207–1223, 2005.

[9] X. Fang, P. J.-H. Hu, Z. L. Li, and W. Tsai. Predicting adoption probabilities in social networks. *Information Systems Research*, 24(1):128–145, 2013.

[10] J. Fowler and N. Christakis. Cooperative behavior cascades in human social networks. *Proc. Nat. Acad. Sci.*, 107(12):5334–8, 2010.

[11] N. E. Friedkin and E. C. Johnsen. Social influence and opinions. *Journal of Mathematical Sociology*, 15(3-4):193–206, 1990.

[12] A. Gionis, E. Terzi, and P. Tsaparas. Opinion maximization in social networks. *arXiv preprint arXiv:1301.7455*, 2013.

[13] S. Goel, D. J. Watts, and D. G. Goldstein. The structure of online diffusion networks. In *Proc. of the 13th ACM EC*, pages 623–638. ACM, 2012.

[14] J. Goldenberg, B. Libai, and E. Muller. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing letters*, 12(3):211–223, 2001.

[15] G. H. Golub and C. F. Van Loan. *Matrix computations (3rd ed.)*. Johns Hopkins University Press, Baltimore, MD, USA, 1996.

[16] M. Gomez Rodriguez, J. Leskovec, and A. Krause. Inferring networks of diffusion and influence. In *Proc. of the 16th ACM SIGKDD*, pages 1019–1028. ACM, 2010.

[17] A. Goyal, F. Bonchi, and L. V. Lakshmanan. Learning influence probabilities in social networks. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 241–250. ACM, 2010.

[18] M. Granovetter. Threshold models of collective behavior. *American journal of sociology*, pages 1420–1443, 1978.

[19] Gurobi. Optimizer - http://www.gurobi.com/products/gurobi-optimizer/gurobi-overview.

[20] R. Hegselmann and U. Krause. Opinion dynamics and bounded confidence models, analysis, and simulation. *Journal of Artificial Societies and Social Simulation*, 5(3), 2002.

[21] M. O. Jackson. *Social and Economic Networks*. Princeton University Press, Princeton, NJ, USA, 2008.

[22] R. A. Johnson and D. W. Wichern. *Applied multivariate statistical analysis*, volume 5. Prentice hall Upper Saddle River, NJ, 2002.

[23] D. Kempe, J. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *Proc. of the ninth ACM SIGKDD*, pages 137–146. ACM, 2003.

[24] Lexalytics. Salience engine - http://www.lexalytics.com/technical-

info/salience-engine-for-text-analysis.
[25] H. Li, S. S. Bhowmick, and A. Sun. Cinema: conformity-aware greedy algorithm for influence maximization in online social networks. In *Proceedings of the 16th International Conference on Extending Database Technology*, pages 323–334. ACM, 2013.
[26] B. Liu. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1), 2012.
[27] S. A. Myers, C. Zhu, and J. Leskovec. Information diffusion and external influence in networks. In *Proc. of the 18th ACM SIGKDD*, pages 33–41. ACM, 2012.
[28] M. Richardson and P. Domingos. Mining knowledge-sharing sites for viral marketing. In *Proceedings of the eighth ACM SIGKDD*, pages 61–70. ACM, 2002.
[29] T. C. Schelling. Models of segregation. *American Economic Review*, 59(2):488–493, 1969.
[30] J. Tang, J. Sun, C. Wang, and Z. Yang. Social influence analysis in large-scale networks. In *Proceedings of the 15th ACM SIGKDD*, pages 807–816. ACM, 2009.
[31] J. Tang, S. Wu, and J. Sun. Confluence: Conformity influence in large social networks. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 347–355. ACM, 2013.
[32] T. Zhang. On the consistency of feature selection using greedy least squares regression. *Journal of Machine Learning Research*, 10:555–568, 2009.