Proceedings of the First AAAI Conference on Human Computation and Crowdsourcing

Scalable Preference Aggregation in Social Networks

Swapnil Dhamal and **Y. Narahari**
{swapnil.dhamal, hari}@csa.iisc.ernet.in
Department of Computer Science and Automation
Indian Institute of Science
Bangalore 560012, India

Abstract

In social choice theory, preference aggregation refers to computing an aggregate preference over a set of alternatives given individual preferences of all the agents. In real-world scenarios, it may not be feasible to gather preferences from all the agents. Moreover, determining the aggregate preference is computationally intensive. In this paper, we show that the aggregate preference of the agents in a social network can be computed efficiently and with sufficient accuracy using preferences elicited from a small subset of critical nodes in the network. Our methodology uses a model developed based on real-world data obtained using a survey on human subjects, and exploits network structure and homophily of relationships. Our approach guarantees good performance for aggregation rules that satisfy a property which we call expected weak insensitivity. We demonstrate empirically that many practically relevant aggregation rules satisfy this property. We also show that two natural objective functions in this context satisfy certain properties, which makes our methodology attractive for scalable preference aggregation over large scale social networks. We conclude that our approach is superior to random polling while aggregating preferences related to individualistic metrics, whereas random polling is acceptable in the case of social metrics.

Introduction

Social networks have been harnessed for a variety of purposes, ranging from viral marketing to controlling spreading of virus, from determining the most powerful personalities in a society to determining the behaviors of people. Social networks explain several phenomena which cannot be explained otherwise, primarily because such phenomena are a result of the social networks themselves. Many of these phenomena can be explained with an important feature of social networks - *homophily* (Easley and Kleinberg 2010). Homophily refers to a bias in friendships towards similar individuals - individuals with similar interests, behaviors, opinions, etc. The tendency of individuals to form friendships with others who are like them is termed *selection*. On the other hand, similarities may also be a result of friendships; people may change their behaviors to align themselves more closely with the behaviors of their friends. This process is termed *social influence*. Hence selection and social influence can be viewed as complements of each other. It is evident that social networks and homophily are inseparable.

How individuals settle at some steady-state behaviors or interests they hold is an interesting question. Due to lack of empirical evidence, it may not be possible to claim that individuals indeed aggregate the behaviors or preferences of their neighbors in a social network based on some private aggregation rule in order to occupy advantageous positions amongst their neighbors. It would be interesting to understand this inherent ability of individuals, which perhaps works towards aggregating the preferences of their neighbors. All these and more revolve around homophily of social networks. We exploit this important feature of homophily in addressing the problem of *preference aggregation*. In particular, we explain how human computation helps not only in simplifying the task of aggregating preferences of a large population, but also in deducing data that is not available.

In this paper, we use the terms voters, individuals, agents, and nodes interchangeably, so also neighbors and friends.

Preliminaries

Given a set of alternatives, individuals have certain preferences over them. These alternatives can be any entity, ranging from political candidates to food cuisines. We assume that an individual's preference can be represented as a complete ranked list of alternatives. We refer to a ranked list of alternatives as *preference* and the multiset consisting of the preferences of the individuals as *preference profile*. For example, if the set of alternatives is $\{X, Y, Z\}$ and individual i prefers Y the most and X the least, then i's preference can be written as $(Y, Z, X)_i$. Suppose individual j's preference is $(X, Y, Z)_j$, then the preference profile of the population $\{i, j\}$ is $\{(Y, Z, X), (X, Y, Z)\}$.

Preference aggregation is a well-studied topic in social choice theory. An *aggregation rule* takes a preference profile as input and outputs the *aggregate preference(s)*, which in some sense reflect(s) the collective opinion of all the individuals. A widely used measure of dissimilarity between two preferences is *Kendall-Tau distance*. It counts the number of pairwise inversions with respect to the alternatives. In this paper, given that the number of alternatives is r, we normalize Kendall-Tau distance to be in $[0, 1]$, by dividing actual distance by $\binom{r}{2}$, the maximum distance between any

two preferences on r alternatives. For example, the Kendall-Tau distance between preferences (X, Y, Z) and (Y, Z, X) is 2, because two pairs $\{X, Y\}$ and $\{X, Z\}$ are inverted between them; the normalized Kendall-Tau distance is $2/\binom{3}{2}$.

We consider **10** voting rules for our study, namely, Bucklin, Smith set, Borda, Veto, Minmax (pairwise opposition), Dictatorship, Schulze, Plurality, Kemeny, and Copeland. Of these rules, only Kemeny outputs the entire aggregate preference; others either determine a winning alternative or give each alternative a score. For consistency, for all rules except Kemeny, we use a well accepted approach - rank a winning alternative as first, then vote over the remaining alternatives and rank a winning alternative therein as second, and so on until all alternatives have been ranked (Brandt, Conitzer, and Endriss 2013). As we are indifferent among alternatives, we assume no tie-breaking rules while determining a winner. So an aggregation rule may not output a unique preference.

Motivation

In real-world scenarios, it may not be feasible to gather the preferences from all the voters owing to factors like time and interest of the voters. One such scenario is preference aggregation in a large online social network. We consider the structure of the underlying social network, wherein owing to homophily, most (if not all) friendship relations imply similar preferences. In order to estimate the aggregate preference of the entire population, an attractive approach would be to select a subset of individuals based on the above phenomenon and incentivize them to report their preferences.

Most aggregation rules are computationally intensive, some of them being hard. Hence, it is important to have efficient workarounds. In almost all aggregation rules (apart from those similar to dictatorship), as the number of voters decreases, computation of the aggregate preference becomes faster. So the problem we consider is potentially one such workaround where we use a subset of preferences to arrive at an acceptable result. As we will see, the way we aggregate the preferences does not reduce the number of voters, but it certainly reduces the number of distinct preferences. This approach works particularly well if this reduced number is significantly less than the number of voters as well as $r!$ (number of possible distinct preferences) because repeated calculations for duplicate preferences can be avoided.

Relevant Work

There have been studies in the literature that deal with the influence of social networks on voting in elections. The pioneering Columbia and Michigan political voting research is discussed in (Sheingold 1973) with an emphasis on importance of the underlying social network. It has been observed that the social network has more impact on one's political party choice than background attributes like class or ethnicity (Burstein 1976). On the other hand, it has been argued via maximum likelihood approach to political voting that, it is optimal to ignore the network structure (Conitzer 2012).

Results of certain behavioral experiments suggest that agents compromise their individual preferences to achieve unanimity in a situation where agents get some utility if and only if the entire population reaches a unanimous decision (Kearns et al. 2009). The scenario in a real group is similar, where members, who do not comply with group norms, either eventually compromise or leave the group to evade the tension between the preferences.

There have been efforts to detect the most critical nodes in social networks with respect to influence maximization, virus inoculation, etc. (Jackson 2008; Easley and Kleinberg 2010). There is extant literature on modeling individual preferences using *general random utility models* which consider the attributes of alternatives and agents, the most relevant being the problem of node selection by exploiting these attributes (Soufiani, Parkes, and Xia 2013) wherein, the underlying social network is not taken into consideration.

To the best of our knowledge, there do not exist any models that model preferences in social networks and furthermore, there do not exist any attempts to determine critical nodes that represent the social network in terms of preferences. Also, our work focuses on not only the winning alternative, but considers the entire aggregate preference.

Survey for Eliciting Preferences

A primary reason for conducting a survey was unavailability of data containing preferences of nodes as well as the underlying social network. The survey, seeking preferences of nodes for a variety of questions, was hosted on the Internet; it was triggered at the darkened seed nodes (Figure 1), who were known to the authors, and diffused along the actual social network. As the survey involved personal preferences, privacy was clearly a concern for most nodes. The underlying network was obtained from the reports of the seed nodes and nodes without privacy concerns (shaded nodes in Figure 1). A given node (say j) having privacy concerns shared a common URL (for completing the survey) with other nodes having privacy concerns and belonging to a cluster of which j is a member. A unique URL was given to each seed node and node having no privacy concerns. The survey was completed by the nodes using the allotted URLs. Below are the questions used in the survey (the first three relate to personal issues while the rest relate to social issues):

1. Where will you prefer to spend time with your friends?
2. Which of the recent movies did you like the most?
3. What will you order when you go to an all-cuisine restaurant with your friends?

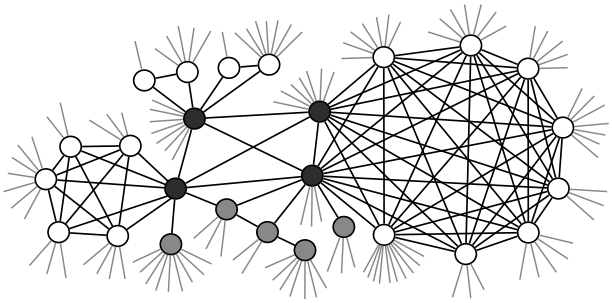


Figure 1: Survey Network with 26 nodes and 80 edges

4. Whom will you pick for the only vacant position of a cricket batsman in India's upcoming overseas tours?
5. Who deserves to be the next Prime Minister of India?
6. Who will be the next Prime Minister of India?
7. Which crime deserves the most extreme of punishments?
8. Which among the crimes that do not attract much public attention, deserves the most attention?

Each question had 5 alternatives to rank. Questions 5 and 6 had 'other' as one of their alternatives, the remaining four alternatives being the most relevant and popular ones. The participants were also asked to report the number of interactive friendships, apart from ones with the people taking the survey, to the best of their knowledge (faint links in Figure 1). For an actual detailed survey, the reader is referred to www.surveymonkey.com/s/F9Z8NZM .

The scale of our survey was modest with 26 nodes. Though the survey network is small, it is suitable for our study since homophily is present in any social network irrespective of its size. Furthermore, in the literature, experiments have been conducted with small scale participation (Golder and Yardi 2010).

The highlighted observations of the survey are as follows:

- Consistent with the theory of homophily, participants connected to each other displayed similar rankings. The rankings were significantly similar if these participants had several common friends who were interconnected.
- In the rankings of connected participants related to both social and personal issues, the first and last alternatives were observed to be exactly the same throughout most of the survey, with slightly jumbled ordering of intermediate alternatives. This can be explained from the fact that people generally praise or criticize an alternative, leaving alternatives that hold middle ground out of the discussion.
- Interestingly, a high level of similarity with respect to social issues was observed even for most unconnected participants, perhaps because of the impact of news and other channels. It may also be justified by a theory of political communication (Huckfeldt et al. 1995) which stresses the importance of citizen discussion beyond the boundaries of cohesive groups for the dissemination of public opinion.
- With regard to personal issues, preferences across different groups were widely different. The cluster size and network structure had a much greater influence on rankings related to personal issues than that to social ones.

Dataset Demographics and a Model

Certain demographics were obtained based on the information provided by the seed nodes and nodes without privacy concerns. A link between two nodes in the survey network implies that they are primarily either past/present college friends, labmates, or roommates. The relationships, in general, are more of an informal nature. 12 of the surveyed nodes belong to the age group 22-25 while the rest to 26-32, with 24 males and 2 females. The nodes have similar educational qualification with them holding either Bachelors or Masters degree in Engineering or Management. Of the 26 nodes, 7 work in industry, while 19 are a part of educational institutions out of which 13 have prior industrial experience. All the nodes hold the nationality of India, belong to the same economic class, and have similar exposure to and interest in political news. They have varying backgrounds, cultures, regions, and mother-tongues; however, these attributes are similar within either of the two large clusters - the leftmost and the rightmost in Figure 1. It will be useful to study a larger population which is more diverse with respect to various attributes so as to understand the effect of similarity in key attributes versus that of network structure.

Type	Min.	Max.	Mean	Std. Dev.
Overall	0.16	0.55	0.34	0.073
Personal	0.07	0.77	0.40	0.130
Social	0.12	0.54	0.30	0.073

Table 1: Statistics about the expected normalized Kendall-Tau distances between all pairs of nodes in the survey dataset

Given a pair of nodes and a question asking for a ranked list of alternatives, the normalized Kendall-Tau distance between the pair can be obtained with respect to that question. As the questions are naturally grouped into either personal or social type, we obtain the distribution of distance and hence the expected distance between any pair, given the type of issue. Certain statistics about the expected normalized Kendall-Tau distances between all pairs of nodes in the data are given in Table 1. A completely unbiased sample is expected to have a mean of 0.5. But in real-world scenarios, the preferences on social issues are seldom unbiased. These preferences of the majority of the population are simultaneously affected by past events, news reports, and information spreading within the network itself. So it is expected to have a mean considerably less than 0.5. However, the mean should ideally be 0.5 in preferences related to personal issues. Our data has a mean of 0.40 in this regard which indicates some bias; this light bias can be attributed to the presence of a giant cluster (constituting the right half of Figure 1) and absence of any other cluster of equivalent size to counterbalance the group preference of the former. Another concern is the low standard deviations of preferences, particularly for personal issues, which again implies some bias. But as we will see, our formulation of the problem and the proposed solution are agnostic to these standard deviations.

The data consisting of distance between preferences of any pair of nodes in the network for personal or social type of issues was observed to follow a discrete distribution which was fit by a truncated Gaussian distribution with range $[0, 1]$ (though truncated Gaussian is a continuous distribution, it provided a best fit to the histogram). In the case of drawing data from such a distribution, we round off the drawn value to the nearest valid value of data. We call this discrete version of truncated Gaussian distribution as $\mathcal{D}$.

So for any pair $\{i, j\}$, the distance between their preferences for a type of issue was observed to follow distribution $\mathcal{D}$ with mean, say $d(i, j)$, and some standard deviation (our approach is agnostic to standard deviation). Note that the value of $d(i, j)$ is different for personal and social types

of issues. As our treatment for both types of issues is the same, we refer to the expected distance simply as $d(i,j)$; the actual value depends on the type of issue under consideration. Let *distance matrix* be a matrix whose cell (i,j) is $d(i,j)$ and *similarity matrix* be a matrix whose cell (i,j) is $c(i,j) = 1 - d(i,j)$. Henceforth, we will assume that the similarity matrix is known to us. We will later discuss a more practical setting where the matrix needs to be derived.

Problem Formulation

Given a network with a set of nodes N, the number of nodes to be selected k, and an aggregation rule f, our objective is to choose a set of nodes $M \subseteq N$ such that $|M| = k$, and aggregate their preferences to arrive at an aggregate preference that is 'close enough' in expectation to the aggregate preference of N using f. We now formalize this problem.

Let the expected distance between a set $S \subseteq N$ and node $i \in N$ be

$$d(S,i) = \min_{j \in S} d(j,i) \tag{1}$$

As $d(i,i) = 0 \ \ \forall i \in N$, we have $d(S,j) = 0 \ \ \forall j \in S$. Let

$$\Phi(S,i) \sim_{\mathcal{U}} \arg\min_{j \in S} d(j,i) \tag{2}$$

be a node chosen uniformly at random from the set of nodes in S that are closest to i in terms of preferences, in expectation. It can be said that $\Phi(S,i)$ represents node i in the set S. In other words, $\Phi(S,i)$ is the *representative* of i in S.

The problem under consideration can be viewed as a setting where given certain individuals representing a population, every individual of the population is asked to choose one among them as its representative; now the representatives vote on behalf of the individuals who chose them.

Aggregating Preferences of Critical Nodes

Recall that preference profile is a multiset containing preferences of the voters. Let the preference profile of the population N be P and that of the selected set M be Q. Suppose $M = \{i, j\}$ where j represents ten nodes, while i represents one. If the preferences are aggregated by inputting Q to aggregation rule f, the aggregate preference $f(Q)$ so obtained will not reflect the preference of the population, in general.

To capture this asymmetry in the importance of selected nodes, their preferences should be weighted. That is, the input for f should be a preference profile $R \neq Q$. In our approach, the weight given to the preference of a node in M is precisely the number of nodes that it represents.

Let Q' be the preference profile obtained by replacing every node's preference in P by its representative's preference. Clearly, $k = |M| = |Q| \leq |Q'| = |P| = |N|$. In our approach, the weight of a representative implies the number of times its preference appears in the new profile, that is, we use $R = Q'$. So in the above example, the new profile $R = Q'$ consists of ten preference of j and one of i. Thus we aggregate the preferences of selected nodes using $f(Q')$.

A Measure of 'Close Enough'

Now given k, our objective is to select a set of nodes M such that $|M| = k$, who report their preferences such that, in expectation, the distance between aggregate preference $f(P)$ obtained by aggregating the preferences of the individuals in N and $f(R)$ obtained by aggregating the preferences of the individuals in M (in an unweighted manner if $R = Q$ or in a weighted manner if $R = Q'$), is minimized. Recall that an aggregation rule f may not output a unique aggregate preference, that is, f is a correspondence.

The aggregation rule f on the preference profile of the entire population outputs $f(P)$ which is a set of preferences. Suppose $f(R)$ also is a set of several preferences, the question arises: which of these to choose as the output? As $f(P)$ is generally not known and all preferences in $f(R)$ are equivalent to us, we choose a preference from $f(R)$ uniformly at random and see how far we are from the actual aggregate preference, in expectation. In order to claim that a chosen preference in $f(R)$ is a good approximation to $f(P)$, it suffices to show that it is close to at least one preference in $f(P)$. Also, as any preference y in $f(R)$ is chosen uniformly at random, we define the distance operator between the above mentioned sets $f(P)$ and $f(R)$ as

$$f(P) \ \Delta \ f(R) = \mathbb{E}_{y \sim_{\mathcal{U}} f(R)} \left[\min_{x \in f(P)} \tilde{\delta}(x,y) \right] \tag{3}$$

where $\tilde{\delta}(x,y)$ is the distance between preferences x and y in terms of the same distance measure as $d(\cdot,\cdot)$. Notice that in general, $f(P) \ \Delta \ f(R) \neq f(R) \ \Delta \ f(P)$. Also, Δ can be defined in several other ways depending on the application or the case we are interested in (worst, best, average, etc.). In this paper, for the reasons explained above, we use the definition of Δ as given in Equation (3).

Recall that for a particular type of issue, the distance between any pair of nodes is drawn from distribution $\mathcal{D}$, that is, the realized values for any particular question belonging to that type of issue are different in general. The value $f(P) \ \Delta \ f(R)$ can be obtained for every question and hence $\mathbb{E}[f(P) \ \Delta \ f(R)]$ for a type of issue can be computed by averaging the values for questions belonging to that type of issue. As our treatment for personal and social issues is the same, we refer to this expected distance as $\mathbb{E}[f(P) \ \Delta \ f(R)]$; the actual value depends on the type of issue under consideration. So now our objective is to find a set M such that $\mathbb{E}[f(P) \ \Delta \ f(R)]$ is minimized.

An Abstraction of the Problem

If the aggregation rule is known, an objective function can be $\phi(M) = 1 - \mathbb{E}[f(P) \ \Delta \ f(R)]$ with the objective of finding a set M that maximizes this value. However, even if the set M is given, computing $\phi(M)$ is computationally intensive for most aggregation rules and furthermore hard for rules like Kemeny. Consider a single question with two alternatives $\{X, Y\}$ and five voters $\{i, j, p, u, v\}$ with preferences $(X,Y)_i, (Y,X)_j, (Y,X)_p, (X,Y)_u, (X,Y)_v$. Using Plurality rule, we have $f(P) = (X,Y)$. For $R = Q$, it can be seen that $\phi(\{i\}) = 1$, while $\phi(\{i,j,p\}) = 0$. Similarly, $\phi(\{p\}) = 0$, while $\phi(\{p,u,v\}) = 1$. The non-monotonicity of $\phi(\cdot)$ can be checked for other non-dictatorial rules also. The non-monotonicity of $\phi(\cdot)$ for non-dictatorial rules for $R = Q'$ can be seen from the non-monotonic plots of the greedy hill-climbing approaches (Greedy-sum and Greedy-avg) in the average case plots of Figure 2 (in a run of greedy

hill-climbing, a set of certain cardinality is a superset of any set having a smaller cardinality). It can also be checked that $\phi(\cdot)$ is neither submodular nor supermodular for $R = Q, Q'$ for non-dictatorial rules. Owing to these properties of the objective function for non-dictatorial rules, it is hard to find a set M that maximizes $\phi(\cdot)$, even within any approximation factor. Moreover, the aggregation rule itself may not be known a priori or may be needed to be changed frequently in order to prevent strategic manipulation of preferences by the voters. This motivates us to propose an approach that finds set M agnostic to the aggregation rule being used.

To this end, we propose a property for preference aggregation rules, *weak insensitivity* which we define as follows.

Definition 1. *A preference aggregation rule satisfies weak insensitivity property under a distance measure and a Δ, if and only if a change of $\eta_i \leq \epsilon_d$ in the preferences of all i, results in a change of at most ϵ_d in the overall aggregate preference, for all ϵ_d. That is, $\forall \epsilon_d$,*

$$\eta_i \leq \epsilon_d \; \forall i \in N \implies f(P) \; \Delta \; f(P') \leq \epsilon_d$$

where P' is the preference profile of voters after deviations.

We call it 'weak' insensitivity property because it allows 'limited' change in the aggregate preference (strong insensitivity can be thought of as a property that allows no change). This is potentially an important property that an aggregation rule should satisfy as it is a measure of its robustness in some sense. It is clear that under normalized Kendall-Tau distance measure and Δ as defined in Equation (3), an aggregation rule that outputs a random preference does not satisfy weak insensitivity property as it fails the criterion for any $\epsilon_d < 1$, whereas dictatorship rule that outputs the preference of a single individual trivially satisfies the property. In fact, we observed using simulations that none of the aggregation rules under consideration, except dictatorship, satisfied this property. We use a weaker form of this property for our purpose, which we call *expected weak insensitivity*.

Definition 2. *A preference aggregation rule satisfies expected weak insensitivity property under a distribution, a distance measure, and a Δ, if and only if a change of η_i, where η_i is drawn from the distribution with mean $\delta_i \leq \mu_d$ and any permissible standard deviation σ_d, in the preferences of all i, results in a change with an expected value of at most μ_d in the overall aggregate preference, for all μ_d. That is, $\forall \mu_d$, $\forall$ permissible σ_d,*

$$\delta_i \leq \mu_d \; \forall i \in N \implies \mathbb{E}[f(P) \; \Delta \; f(P')] \leq \mu_d \quad (4)$$

where P' is the preference profile of voters after deviations.

Note that in $\mathbb{E}[f(P) \; \Delta \; f(P')]$, the expectation is over the varying modified preferences of the agents (since η_i's vary across iterations and even if not, there are multiple preferences, in general, at a distance of η_i from any given preference). In this paper, we study expected weak insensitivity property under distribution $\mathcal{D}$, normalized Kendall-Tau distance, and Δ as defined in Equation (3). For distribution $\mathcal{D}$ with $\mu_d \in [0, 1]$, the permissible range of σ_d depends on μ_d. For instance, for most values of μ_d, the permissible range for $\sigma_d \leq \frac{1}{\sqrt{12}} \approx 0.28$ (value at which the truncated Gaussian becomes a Uniform distribution), while for $\mu_d \in \{0, 1\}$, the permissible $\sigma_d = 0$. Table 2 presents the results of extensive simulations investigating empirical satisfaction of this property by the considered aggregation rules.

Bucklin	✓	Smith	✓	Borda	✗	Veto	✗
Minmax	✓	Dictatorship	✓	Schulze	✗		
Plurality	✓	Kemeny	✗	Copeland	✗		

Table 2: Results of empirical satisfaction of expected weak insensitivity property by aggregation rules under distribution $\mathcal{D}$, normalized Kendall-Tau distance, and Δ as defined

Lemma 1. *Given a distance measure and a Δ, for a preference aggregation rule satisfying expected weak insensitivity property under distribution $\mathcal{D}$, the distance measure, and the Δ, $f(Q')$ is at an expected distance of at most ϵ_d from $f(P)$ if the expected distance between every individual and the set M is at most ϵ_d, for all ϵ_d. That is, $\forall \epsilon_d$,*

$$d(M, i) \leq \epsilon_d \; \forall i \in N \implies \mathbb{E}[f(P) \; \Delta \; f(Q')] \leq \epsilon_d$$

Proof. In the preference profile P of all voters, the preference of any node $i \in N$ is replaced by the preference of its representative node $p = \Phi(M, i)$ to obtain Q'. From Equations (1), (2), and the hypothesis, we have $d(p, i) \leq \epsilon_d$.

Since in P, preference of every i is replaced by that of the corresponding p to obtain a new profile Q', and distance between i and p is distributed according to distribution $\mathcal{D}$ with mean $d(p, i)$ and some standard deviation σ_d, the above is equivalent to node i deviating its preference by some value which is drawn from distribution $\mathcal{D}$ with mean $d(p, i) = d(M, i)$. So $\delta_i = d(M, i) \; \forall i$, $\mu_d = \epsilon_d$, and $P' = Q'$ in Equation (4). Also, recall that the expectation $\mathbb{E}[f(P) \; \Delta \; f(P')]$ is over the varying modified preferences of the agents. Here the expectation $\mathbb{E}[f(P) \; \Delta \; f(Q')]$ is over varying preferences of the agents' representatives in M with respect to different questions and preferences of the agents. These are equivalent given $P' = Q'$. As this argument is valid for any permissible σ_d, the result follows. □

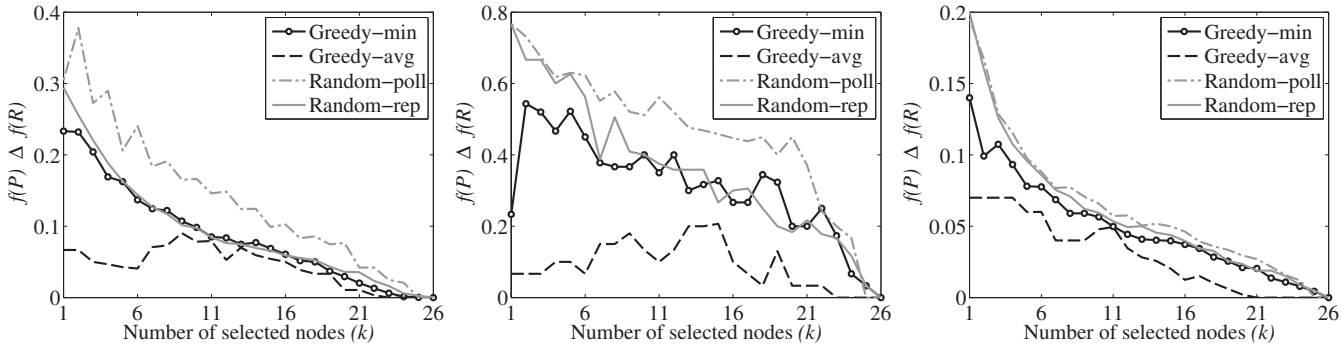
Under the proposed model, this lemma gives a theoretical guarantee on $\mathbb{E}[f(P) \; \Delta \; f(Q')]$ for aggregation rules that satisfy expected weak insensitivity property under distribution $\mathcal{D}$, and relevant distance measure and Δ.

Objective Functions in the Abstracted Problem

Recall that $c(\cdot, \cdot) = 1 - d(\cdot, \cdot)$. Our objective is now to find a set of critical nodes M that maximizes some objective function, with the hope of minimizing $\mathbb{E}[f(P) \; \Delta \; f(R)]$ where $R = Q'$ in our case. As the aggregation rule is anonymous, in order to ensure that the approach works well, even for rules such as random dictatorship, the worst-case objective function for the problem under consideration, representing least expected similarity, is

$$\rho(S) = \min_{i \in N} c(S, i) \quad (5)$$

It is clear that $\rho(S) = 1 - \epsilon_d$ in Lemma 1 and so this function provides a guarantee on $\mathbb{E}[f(P) \; \Delta \; f(Q')]$ for any aggregation rule satisfying expected weak insensitivity. However,



(a) Personal issue: average case (b) Personal issue: worst case (c) Social issue: average case

Figure 2: Performance of methods on the survey data with respect to different issues and cases

extreme aggregation rules like random dictatorship are seldom used in real-world scenarios; hence, an alternative objective function, representing average expected similarity, is

$$\psi(S) = \sum_{i \in N} c(S, i) \tag{6}$$

Proposition 1. *Given constants χ and ω, it is NP-hard to determine whether there exists a set M consisting of k nodes such that (a) $\rho(M) \geq \chi$, (b) $\psi(M) \geq \omega$.*

Proof. We reduce an NP-hard Dominating Set problem instance to the problem under consideration. Given a graph G of n vertices, the dominating set problem is to determine whether there exists a set D of k vertices such that every vertex not in D is adjacent to at least one vertex in D.

Given a dominating set problem instance, we can construct a weighted undirected complete graph H consisting of the same set of vertices as G such that, the weight $c(i, j)$ of an edge (i, j) in H is some high value (say 0.9) if there is edge (i, j) in G, else it is some low value (say 0.6).

Now there exists a set D of k vertices in G such that the distance between any vertex in G and any vertex in D is at most one, if and only if there exists a set M of k vertices in H such that $\rho(M) \geq 0.9$ or $\psi(M) \geq k + 0.9(n - k)$. Here $\chi = 0.9$ and $\omega = k + 0.9(n - k)$. This shows that the NP-hard dominating set problem is a special case of the problem under consideration, hence the result. □

It can be shown that the marginal contribution of any node to the value of any set with respect to $\rho(\cdot), \psi(\cdot)$ is no less than that to its superset. Proposition 2 hence follows.

Proposition 2. *The objective functions $\rho(\cdot)$ and $\psi(\cdot)$ are non-negative, monotone, and submodular.*

For a non-negative, monotone, submodular function, selecting elements one at a time, each time choosing an element that provides the largest marginal increase in the function value (greedy hill-climbing), gives a $(1 - \frac{1}{e}) \approx 0.63$-approximation to the optimal solution (Nemhauser, Wolsey, and Fisher 1978). As the considered objective functions in Equations (5) and (6) satisfy these properties, we can use the greedy hill-climbing algorithm to obtain a good approximation to the optimal solution. Moreover, as desired, the functions are agnostic to the aggregation rule being used.

Experimental Results

Given k, our objective is to select a subset $M \subseteq N$ such that $|M| = k$ and $\mathbb{E}[f(P) \ \Delta \ f(R)]$ is minimized, where $R = Q, Q'$, etc. depending on the method. Note that $|M|$ is exactly k as opposed to the general trend in literature where it is at most k. This is because we select the k most critical nodes by solving the abstracted problem and so selecting at most k nodes instead of exactly k nodes does not guarantee that the solution obtained to the original problem by selecting at most k nodes is better than that obtained by selecting exactly k nodes. It is also to be noted that the selected set may be different for personal and social types of issues.

Recall that the profile of N is P, that of M is Q, and that obtained by replacing every node's preference in P by that of its representative in M, is Q'. We consider four methods for obtaining $f(R)$. In each method, we initialize M to $\{\}$.

Greedy-min (Greedy hill-climbing for maximizing $\rho(\cdot)$): Until $|M| = k$, choose a node $j \in N \backslash M$ that maximizes $\rho(M \cup \{j\}) - \rho(M)$. Then, obtain $f(R) = f(Q')$.

Greedy-avg (Greedy hill-climbing for maximizing $\psi(\cdot)$): Until $|M| = k$, choose a node $j \in N \backslash M$ that maximizes $\psi(M \cup \{j\}) - \psi(M)$. Then, obtain $f(R) = f(Q')$.

Random-poll (Random selection without representation): Until $|M| = k$, choose a node j uniformly at random such that $j \in N \backslash M$. Then, obtain $f(R) = f(Q)$.

Random-rep (Random selection with representation): Same as Random-poll, except that $f(R) = f(Q')$.

For Dictatorship, if the dictator $\mathbb{D} \notin M$, Random-poll outputs the preference of a node in M chosen uniformly at random, else it outputs $\mathbb{D}$'s preference itself; while other methods always output the preference of $\mathbb{D}$'s representative in M.

The values of $\mathbb{E}[f(P) \ \Delta \ f(R)]$ are computed using extensive simulations with the considered aggregation rules. The results for personal and social issues are obtained by averaging $f(P) \ \Delta \ f(R)$ for the questions related to those respective issues. Given k, the selected set M and also the uniquely chosen representatives of nodes vary in different runs and so, we observe average and worst case results for each method (the worst case plot for a type of issue is obtained by averaging the worst results for the questions that

are grouped in that type of issue). Apart from dictatorship, the plots for all aggregation rules are similar (albeit with different scaling) to the ones plotted in Figure 2 for average and worst cases with respect to personal issues and average case with respect to social ones. The plots of worst case for social issues are similar to the worst case for personal ones but with a scaling of approximately $\frac{2}{3}$ for Greedy-min, Greedy-avg, and Random-rep; here Random-poll performs almost at par with Random-rep. For dictatorship, the average case plots decrease linearly with k, while the worst case plots retain certain values after which they dip suddenly for some high values of k; here Greedy-avg performs the best, followed by Greedy-min, Random-rep, and Random-poll, in that order.

Our key observations are as follows:

- Greedy-avg consistently performs considerably better than other methods; but its plots display high entropy (lack of pattern) as compared to other methods.
- Greedy-min performs better than Random-rep for low values of k; this difference is not salient in average cases owing to low standard deviation of the data. Random-rep performs at par with Greedy-min for higher values of k.
- Random-poll performs quite poorly for personal issues, but reasonably well for social issues with moderate and high values of k. This justifies its use for social issues with a non-negligible sample size.
- Random methods do not perform very poorly in average cases owing to low standard deviation of the data. But in worst cases, they perform quite poorly for low values of k which precludes their use when the sample size is low.
- Random-rep consistently performs better than Random-poll, which justifies using $R = Q'$ instead of $R = Q$.
- The effect of satisfiability of expected weak insensitivity is not very prominent, because the property is not violated by an appreciable enough margin for any aggregation rule in case of only five alternatives. This makes our approach more attractive in practice as it performs well irrespective of the aggregation rule. Nonetheless, the property gives a guarantee for performance for an aggregation rule.

A Model for Deriving the Data

The above analysis was conducted assuming that the similarity matrix is known from some past data or it can be obtained from some parameters that reveal the expected similarities between nodes. But this data may be unavailable in practice. For instance, in online social networks, it is feasible to obtain expected distances between connected nodes by analyzing their interactions. However, the distances between unconnected pairs is generally unknown. In an effort to obtain expected distance between any two nodes in the network, we propose a model based on the survey.

Obtaining Distances between Unconnected Nodes

Recall that cell (i, j) of a *distance matrix* contains $d(i, j)$, the expected distance between preferences of nodes i and j. We initialize all values in this matrix to 0 for $i = j$ and to 1 (the upper bound on the value of the distance) for any unconnected pair $\{i, j\}$. In the case of a connected pair $\{i, j\}$,

d_x \ d_y	**0.00**	**0.10**	**0.20**	**0.30**	**0.40**	**0.50**
0.00	0.00					
0.10	0.10	0.17				
0.20	0.20	0.26	0.32			
0.30	0.30	0.33	0.37	0.40		
0.40	0.40	0.42	0.43	0.45	0.47	
0.50	0.50	0.50	0.50	0.50	0.50	0.50

Table 3: A partial view of the table T_5

the value $d(i, j)$ is initialized to the actual observed expected distance (this value is known). Following the initialization of the distance matrix, we now explain how to update it.

Consider nodes $\{p, i, j\}$ where we know the expected distances $d(p, i)$ and $d(p, j)$ and we are interested in finding $d(i, j)$ via node p. Given the preference of p and $d_x = d(p, i)$, let the preference of i be chosen uniformly at random from the set of preferences that are at a distance η from the preference of p, where η is drawn from distribution $\mathcal{D}$ with mean d_x and some standard deviation. Similarly, given $d_y = d(p, j)$, let the preference of j be obtained. Using this procedure, the distance between the obtained preferences of i and j via p over several iterations and various standard deviations, is observed to follow distribution $\mathcal{D}$. Let the corresponding expected distance constitute cell (d_x, d_y) of a table, say T_r, where r is number of alternatives. It is clear that this distance is independent of the actual preference of p.

We empirically observe that T_r is different from $T_{r'}$ for $r \neq r'$. Following are the general observed properties of T_r:

- $T_r(d_y, d_x) = T_r(d_x, d_y)$
- $T_r(1 - d_x, d_y) = T_r(d_x, 1 - d_y) = 1 - T_r(d_x, d_y)$
- $T_r(1 - d_x, 1 - d_y) = T_r(d_x, d_y)$

We define an operator $\oplus_r$ as follows:

$$d_x \oplus_r d_y = \begin{cases} T_r(d_x, d_y), & \text{if } d_x \leq 0.5 \text{ and } d_y \leq 0.5 \\ \max\{d_x, d_y\}, & \text{if } d_x > 0.5 \text{ or } d_y > 0.5 \end{cases}$$

The two different cases while defining $\oplus_r$ are based on the reasonable assumption that $d(i, j)$ via p should be assigned a value which is at least $\max\{d(p, i), d(p, j)\}$ (but T_r does not follow this rule when either $d(p, i)$ or $d(p, j)$ exceeds 0.5). As the questions of our survey has 5 alternatives, we obtain the table T_5 and hence $d_x \oplus_5 d_y$ for any pair $\{d_x, d_y\}$. Table 3 presents a partial view of T_5. Now the next question is to find $d(i, j)$ for any pair $\{i, j\}$. In order to provide a fit to the distances obtained from the survey, we initialize the distance matrix as explained in the beginning of this subsection and update it based on the *all pairs shortest path algorithm* (Cormen et al. 2009) with update rule:
if $d(p, i) \oplus_r d(p, j) < d(i, j)$ **then** $d(i, j) = d(p, i) \oplus_r d(p, j)$
where $r = 5$ for us. The corresponding similarity matrix is obtained by assigning value $c(i, j) = 1 - d(i, j)$.

As mentioned earlier, in online networks, it is feasible to obtain expected distances between connected nodes. But this is generally not the case for offline social networks. We propose a model in this direction based on the survey, quantifying homophily for connected nodes in social networks.

Obtaining Distances between Connected Nodes

In order to relate the graph structure with the observed distances between preferences of connected nodes, we propose a method that intuitively captures how similar a connected pair should be. Bigger the cluster(s) the pair is a part of (pair with more common friends which are well connected among themselves), the less distant their preferences should be, in expectation. However, it is possible that the pair is a part of multiple clusters and the individual clusters may not be exactly cliques. Hence we generalize the notion of clique size by proposing what we call *cliqueness coefficient* by defining it as an increasing function of the number of edges in the subgraph spanned by the pair and their common neighbors.

Definition 3. *Consider two connected nodes i and j. Let ξ be the number of edges in the subgraph spanned by i, j, and their common neighbors. The cliqueness coefficient between i and j is $\gamma \geq 1$ such that $\frac{\gamma(\gamma-1)}{2} = \xi$, or $\gamma = \frac{1+\sqrt{1+8\xi}}{2}$.*

If such a subgraph is a clique, then *cliqueness coefficient* is equal to the size of that clique.

It can be seen from the survey network as depicted in Figure 1 that several pairs of nodes have the same values of cliqueness coefficient. It was observed from the survey data that given a cliqueness coefficient and a type of issue, the expected normalized Kendall-Tau distance between the pairs of nodes having that particular cliqueness coefficient between them, followed distribution $\mathcal{D}$. Let the cliqueness coefficient between any two connected nodes i and j be γ_{ij} and let the distance between the two nodes be drawn from a distribution $\mathcal{D}$ with mean $d(i,j)$. In order to fit the survey data, $d(i,j)$ is set to be an inverse exponential function of γ_{ij}; specifically, there exist some $\alpha \in [0,1]$ and $\beta \geq 1$ such that $d(i,j) = \alpha\beta^{-\gamma_{ij}}$. We deduce the values of α and β from the data by minimizing weighted L_1 norm. Consider vectors A and B of size equal to the number of distinct values of γ in the network. Exactly one element of A is $\alpha\beta^{-\gamma}$ corresponding to a unique γ. The corresponding element in B is the actual mean distance between all pairs with that particular γ. By minimizing weighted L_1 norm of $A - B$ where the weight of any of its elements is equal to the number of pairs with the corresponding γ, the deduced values were: $\alpha_{personal} = 0.50, \beta_{personal} = 1.06$, $\alpha_{social} = 0.31, \beta_{social} = 1.01$. But α, β may not be known a priori. Two of the ways in which they can be inferred are:

- Directly use the parameters deduced from some previous data and update the expected distances based on the new γ's (owing to the change in network structure since then).
- Take a few samples of pairs of nodes who report their preferences, such that the number of edges in subgraph spanned by them and their common neighbors is known.

No concrete conclusions regarding the standard deviation of the distribution of distances between connected pairs, could be drawn from the data. However, it could be observed that the more mutually exclusive neighbors a pair $\{i,j\}$ had, the more the distance between them deviated from $d(i,j)$.

Discussion

Our main finding is that social network structure can be exploited for aggregating preferences related to personal issues; but for social issues, random polling is acceptable with a non-negligible sample size. Though the survey network was small, it was suitable for our experiments since homophily is present in any social network irrespective of its size. The nodes had very similar preferences regarding personal issues on average; also the standard deviation was on a lower side. These, actually, were some of the reasons why the random approaches did not fare very poorly for personal issues. Nonetheless, the experiments, conducted with this data, validated our theoretical results well. In fact, for any network with homophily property, the results guarantee excellent performance of our approach, particularly for aggregation rules satisfying expected weak insensitivity property.

If the similarity matrix is known, the time complexity for obtaining M and hence R using greedy hill-climbing is $O(k|N|^2)$, while that using random selection with representation is $O(k|N|)$. The time complexity for arriving at the aggregate preference(s) $f(R)$, however, depends on the aggregation rule. If the similarity matrix is unknown, the time complexity for deriving it is largely decided by the all pairs shortest path algorithm, which is $O(|N|^2 \log|N| + |N||E|)$ by Johnson's algorithm where $|E|$ the number of edges is generally small owing to sparsity of social networks.

Future Work

A primary objective of this paper was to select k nodes so as to minimize $\mathbb{E}[f(P) \ \Delta \ f(R)]$. This work can be extended to select minimum number of nodes such that this value is bounded. The expected weak insensitivity property may be of prime importance in social choice theory and so, it will be interesting to analytically determine the rules that satisfy it. It may be of practical interest to study its generalization where the aggregate preference changes by at most $\theta\mu_d$ instead of μ_d (see Equation (4)), in expectation, for some constant θ. We used a particular form of modified profile $R = Q'$. It will be interesting to study the 'best' form of R.

The time complexity of the greedy algorithm is large owing to the global nature of the abstracted optimization problem. It may be useful to consider localized algorithms like degree centrality heuristic. As the scale of our survey was modest, it is essential to have a survey on a larger scale to verify and refine the proposed model. It will also be useful to study models and approaches which take standard deviations of the data into consideration. We assumed that the voters are not strategic and so report their preferences truthfully. From a game theoretic viewpoint, it would be interesting to look at the strategic aspect of the problem. General random utility models are complementary to our model, exploiting attributes of nodes and alternatives instead of the underlying social network. It will be interesting to consider attributes as well as the underlying social network for node selection.

Acknowledgments

The authors thank Palash Dey, Prabuchandran K. J., and the anonymous reviewers for their useful comments.

References

Brandt, F.; Conitzer, V.; and Endriss, U. 2013. Computational social choice. www.cs.duke.edu/~conitzer/comsocchapter.pdf.

Burstein, P. 1976. Social networks and voting: Some Israeli data. *Social Forces* 54(4):833–847.

Conitzer, V. 2012. Should social network structure be taken into account in elections? *Mathematical Social Sciences* 64(1):100–102.

Cormen, T. H.; Leiserson, C. E.; Rivest, R. L.; and Stein, C. 2009. *Introduction to Algorithms*. MIT press.

Easley, D., and Kleinberg, J. 2010. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press.

Golder, S. A., and Yardi, S. 2010. Structural predictors of tie formation in twitter: Transitivity and mutuality. In *Social Computing, 2010 IEEE Second International Conference on*, 88–95.

Huckfeldt, R.; Beck, P.; Dalton, R.; and Levine, J. 1995. Political environments, cohesive social groups, and the communication of public opinion. *American Journal of Political Science* 1025–1054.

Jackson, M. 2008. *Social and Economic Networks*. Princeton Univ Press.

Kearns, M.; Judd, S.; Tan, J.; and Wortman, J. 2009. Behavioral experiments on biased voting in networks. *Proceedings of the National Academy of Sciences* 106(5):1347–1352.

Nemhauser, G. L.; Wolsey, L. A.; and Fisher, M. L. 1978. An analysis of approximations for maximizing submodular set functions-I. *Mathematical Programming* 14(1):265–294.

Sheingold, C. A. 1973. Social networks and voting: the resurrection of a research agenda. *American Sociological Review* 712–720.

Soufiani, H. A.; Parkes, D. C.; and Xia, L. 2013. Preference elicitation for general random utility models. In *The Twenty-Ninth Conference on Uncertainty In Artificial Intelligence*, 596–605.