Active Surveying: A Probabilistic Approach for Identifying Key Opinion Leaders

Hossam Sharara
Computer Science Department
University of Maryland
College Park, MD

Lise Getoor
Computer Science Department
University of Maryland
College Park, MD

Myra Norton
Community Analytics
Baltimore, MD

Abstract

Opinion leaders play an important role in influencing people's beliefs, actions and behaviors. Although a number of methods have been proposed for identifying influentials using secondary sources of information, the use of primary sources, such as surveys, is still favored in many domains. In this work we present a new surveying method which combines secondary data with partial knowledge from primary sources to guide the information gathering process. We apply our proposed active surveying method to the problem of identifying key opinion leaders in the medical field, and show how we are able to accurately identify the opinion leaders while minimizing the amount of primary data required, which results in significant cost reduction in data acquisition without sacrificing its integrity.

1 Introduction

Studying influence in social networks is an important topic that has attracted the attention of a variety of researchers in different domains [Raven, 1965; Kempe *et al.*, 2003]. People often seek the opinion and advice of their peers regarding various decisions, whether it is to try a new restaurant, buy a certain product or even to support a particular politician [Keller and Berry, 2003]. This behavior gives rise to a certain set of individuals in the social network, referred to as *influentials* or *opinion leaders*, who have a huge impact on other people's opinions, actions and behavior.

In the commercial space, the question of how to identify opinion leaders within a given population of purchasers or decision makers is of great importance [Myers and Robertson, 1972; Krackhardt, 1996]. Identifying these individuals properly leads to more effective and efficient sales and marketing initiatives [Valente and Davis, 1999]. This is true in multiple industries; here we begin our exploration in the medical domain, studying the influence networks of local physicians relative to the treatment of specific disease states. Key opinion leader identification has been the focus of multiple studies in the health care literature [Soumerai *et al.*, 1998; Doumit *et al.*, 2007].

Secondary data describing suggested influence, is often easy to obtain; whereas primary data, representing surveys that measure trust and advice-seeking, is harder and much more expensive to acquire. For instance, citations are often used as an indirect indicator of influence in an academic settings, where influential authors' publications tend to receive higher citations than average. Obtaining a citation network between a set of authors in a certain field (e.g. infectious disease) can be easily constructed by looking at the publication record of each author. However, measuring the influence of each author directly requires more work, and often involves a labor-intensive process of interviewing subjects and extracting their "network of influence", e.g., who they turn to for advice and recommendations.

Methods for identifying opinion leaders can be classified into two categories according to the type of data they use for drawing their conclusions. Primary methods rely on manually collecting information about peer-influence in a given population from the individuals themselves. One of the most commonly used primary methods is surveys, where the respondents are asked to report their opinion about who they perceive as opinion leaders. Although primary methods are considered to be the most informative about actual peer-influence, their main drawback is the high associated costs due to the time-intensive nature of the process: in many cases surveys are obtained through one-on-one interviews with the respondents, sometimes over the phone, but often in person.

On the other hand, secondary methods rely mainly on using an underlying interaction network as a "*proxy*" for influence, thus avoiding the manual aspect of primary methods. One of the most widely used techniques in this setting is relying on network centrality measures of these secondary networks (e.g., citation, co-authorship, etc.) to identify the opinion leaders. However, the major drawback of these methods is the fact that the correlation between peer-influence in the actual social network and the interactions occurring in the proxy networks cannot be verified. In a recent study on public opinion formation [Watts and Dodds, 2007], the authors showed through a series of experiments that the customers who are critical in accelerating the speed of diffusion need not be the most connected in their corresponding social network.

In this work, we show how to combine the use of primary and secondary methods for leadership identification in the medical domain. We use primary data describing a physician nomination network in which physicians are surveyed to nominate other physicians whom they turn to for professional

advice. We augment this network with secondary data describing publication history (citation and co-authorship), as well as hospital affiliation information. We use ideas from the active learning literature to build a model that can use partial knowledge of primary data, together with secondary data, to guide the survey process. By targeting the most informative physicians for additional primary data collection, we minimize the amount of primary data needed for accurate leadership identification. As this type of primary data collection requires significant investment, this technique empowers organizations to tackle the task of accurate leadership identification in a much more cost effective and efficient manner.

The rest of the paper is organized as follows. Section 2 provides a brief overview of the related work and background for both opinion leader identification and active learning. In Section 3, we give a detailed description of the problem and an outline of the proposed method. Section 4 describes the details of the active surveying algorithm. Section 5 discusses the experimental settings, the dataset and the results of using our proposed method compared to different baselines. Finally, Section 6 concludes our work and proposes future directions.

2 Background

2.1 Opinion Leader Identification

In the diffusion of innovation literature, there are two main methods for identifying opinion leaders from primary sources: self-designation and surveys [Rogers and Cartano, 1962]. In the self-designation method, respondents are asked to report to what extent they perceive themselves to be influential. However, as can be expected, such methods are usually biased, and often reflect self-confidence rather than actual influence. On the other hand, surveys are based on having selected individuals, referred to as respondents, report who they perceive as opinion leaders in a given domain [Dorfman and Maynor, 2006]. Peer-identified opinion leaders are believed to be better sources of true influence compared to self-identified ones.

Due to the high costs associated with primary methods for leadership identification, there has also been a great deal of attention to methods that make use of secondary data sources. These methods rely mainly on using different structural measures for determining the importance of nodes in a proxy interaction network. In the sociology literature, various centrality measures [Wasserman and Faust, 1995] have been used to determine the most important individuals in a given social network. Among the most commonly used measures are degree centrality, indicating the most connected individuals in the network, and betweenness centrality, distinguishing the "brokers" in the network.

2.2 Active Learning

In this work, we build on ideas from the field of active learning, where the learner is able to acquire labels of additional examples to construct an accurate classifier or ranker while minimizing the number of labeled examples acquired. This is achieved by providing an intelligent, adaptive querying technique for obtaining new labels to attain a certain level of accuracy with minimal training instances. A generic algorithm for active learning is described in [Saar-Tsechansky and Provost, 2004], where a learner is applied to an initial sample L of labeled examples, then each example in the remaining unlabeled pool is assigned an "*effectiveness score,*" based on which the subsequent set of examples to be labeled is chosen until some predefined condition is met. The main difference between various active learning methods is how the effectiveness score of each example is computed; the score usually corresponds to the expected utility that the newly acquired example can add to the learning process.

One widely used method for active learning is uncertainty sampling [Lewis and Gale, 1994], where the learner chooses the most uncertain data point to query, given the current model and parameters. Measuring the uncertainty depends on the underlying model used, but it usually translates to how close the data point is to the decision boundary. For instance, if a probabilistic classifier is used, the posterior probability can be used directly to guide the selection process. By acquiring the labels for the data points closer to the decision boundary, the model can be improved by better defining the existing margin. A variety of active learning methods have been proposed [Settles, 2009], with various ways to reduce the generalization error of the underlying model during learning. Active learning has proved to be useful in settings where acquiring labeled data is expensive. It has been applied successfully in numerous domains, such as image processing [Tong and Chang, 2001], speech recognition [Tür *et al.*, 2005], and information extraction [Thompson *et al.*, 1999].

3 Problem Description

Our problem can be formulated as determining the minimal set of respondents needed to correctly identify at least $k\%$ of the set of opinion leaders present in a given population. In order to achieve this goal, we need a method that can guide the surveying process for selecting the next respondent, such that the expected set of identified opinion leaders is maximized at each step. We apply a simple threshold model on the survey responses to identify opinion leaders; if a candidate receives more than α nominations, she is considered an opinion leader.

A key difference between this problem setting and the traditional active learning setting is that the acquisition of a survey response is more complex than that of a single label. A survey response is a structured object that includes a *set* of nominations $\{nominate(v, u) : u \in population\}$ made by a given respondent v; all of which should be accounted for in both the learning and inference phases. In some cases there may be weights associated with each nomination; although here we are assuming uniform weights, it is straightforward to extend the model to cases where weights vary.

We propose an active surveying approach that combines partial knowledge from primary sources along with secondary information to provide a dynamic framework for intelligently gathering additional primary data for opinion leader identification. In our approach, the next survey respondent is chosen to maximize the likelihood of identifying new opinion leaders. After the proposed respondent is surveyed, the survey results are incorporated back into the model to update future predictions.

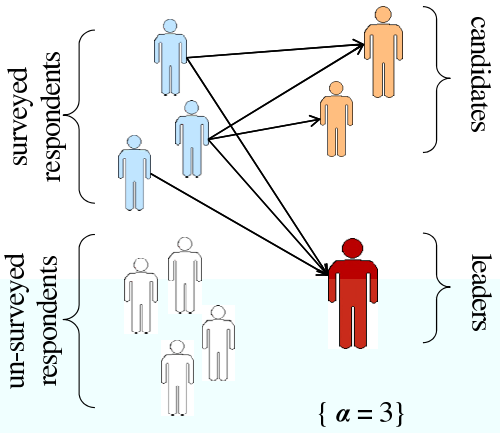

Figure 1: Example *candidates* and *leaders* sets

First, we need to define the conditions upon which the next respondent should be selected in order to maximize the set of identified opinion leaders. Suppose we are given an initial set of survey responses, and a threshold α that determines the minimum number of nominations an individual should obtain to be declared an opinion leader. Let the set of nominations received by a given nominee u be denoted as $nominations(u) = \{v : nominate(v, u) \wedge v \in respondents\}$. From the initial set of responses, we can generate the following two sets of individuals:

$$\begin{aligned} leaders = &\quad \{l : \quad |nominations(l)| \geq \alpha\} \\ candidates = &\quad \{c : \quad 0 < |nominations(c)| < \alpha\} \end{aligned}$$

where the *leaders* set represents the individuals who have received at least α nominations and are already identified as opinion leaders, while *candidates* is the set of individuals who have been nominated by at least one person, but have not yet received enough nominations to be declared opinion leaders. Figure 1 shows a toy example of how the *candidates* and *leaders* sets are generated.

Ideally, the best respondent to survey should be more likely to nominate new leaders, either from the ones already in the *candidates* set or introduce new individuals to expand it. In survey settings, there's typically a bound on the number of opinion leaders each respondent can nominate. Thus, we add a requirement that the respondent is also less likely to nominate individuals in the already identified *leaders* set, in order to minimize the "non-informative" nominations to already identified opinion leaders. In order to estimate the likely nominations of a given respondent, we model the expected survey responses based on existing secondary sources, along with primary information from the current available surveys. By using this model to predict the nominations of the yet-to-be-surveyed respondents, we can then follow a greedy approach based on the above criterion to pick the respondent who is likely to expand the set of identified opinion leaders at each step.

The set of possible nominations in a given population can be viewed as a directed graph $G(V, E)$, where each node $v \in V$ in the network corresponds to an individual in the population, and a directed edge $e(u, v) \in E$ indicates that v is a possible nominee for respondent u. Generally, the set of potential edges in the network can be as large as $|V| \times |V|$, yielding a fully connected graph. However, in real scenarios, the number of potential edges can often be limited by using appropriate filters on the incident nodes, such as evidence from secondary sources, local proximity, similarity, or any other constraint imposed by the problem. We refer to the subset of potential edges that correspond to actual respondent nominations as "positive" edges, and the ones that are not realized through the survey as "negative" edges. We refer to the set of edges corresponding to the initial set of surveys as the "observed" edges.

The secondary sources of information are represented in our model as: a) a set of features $\mathcal{F}_v$ associated with the nodes V in G, and b) a set of secondary networks $G^{(1)}(V^{(1)}, E^{(1)}) \dots G^{(n)}(V^{(n)}, E^{(n)})$ representing other types of interactions between the set of individuals in the target population (e.g. communication, co-authorship, co-affiliation, etc.). As these secondary networks may not necessarily align with the main graph G, we only consider the sub-networks comprising the nodes that overlap with our network of concern. Another set of edge features $\mathcal{F}_e$ is generated for the set of edges E in G, each representing a vector of the corresponding edge weights in each of the associated secondary networks. During this step, the set of node features $\mathcal{F}_v$ are also enriched by additional features from the secondary networks.

In Figure 2, the input graph G represents a partially observed author nomination network, where the shaded authors A_1 and A_7 are the ones who have already been surveyed. In this example, all of the potential nominations for author A_1 were realized (positive, denoted by solid edges), while for author A_7, although the nomination for A_2 was a potential edge, it was not realized (negative, denoted by a dashed outgoing edge). Each author in the primary nomination network G has a set of associated features, such as the geographical location, h-index, current academic position, etc. These features constitute the set of node features $\mathcal{F}_v$ in our model. In addition to the nomination network, we have two secondary sources of information in our example: a co-authorship network $G^{(1)}$ and a co-affiliation network $G^{(2)}$.

After aligning the secondary networks with the primary nomination network, the edge features we generate are indicators of the edge existence in the corresponding secondary network. For instance, the edge in G corresponding to author A_1 nominating author A_2 does not have corresponding coauthorship evidence in network $G^{(1)}$, but the two authors do share the same affiliation as indicated in network $G^{(2)}$. Thus, the resulting feature vector for edge $e(A_1, A_2)$ in this simple example would be $\mathcal{F}_{e(A_1,A_2)} = (0, 1)$, as shown on the resulting annotated input graph G_a in Figure 2. In addition to the generated edge features, extra node features are derived from these secondary networks, such as the number of publications from the co-authorship network, or the prestige of the affiliated organization from the co-affiliation network. These additional node features are then used to enrich the original set of author features obtained from the primary data.

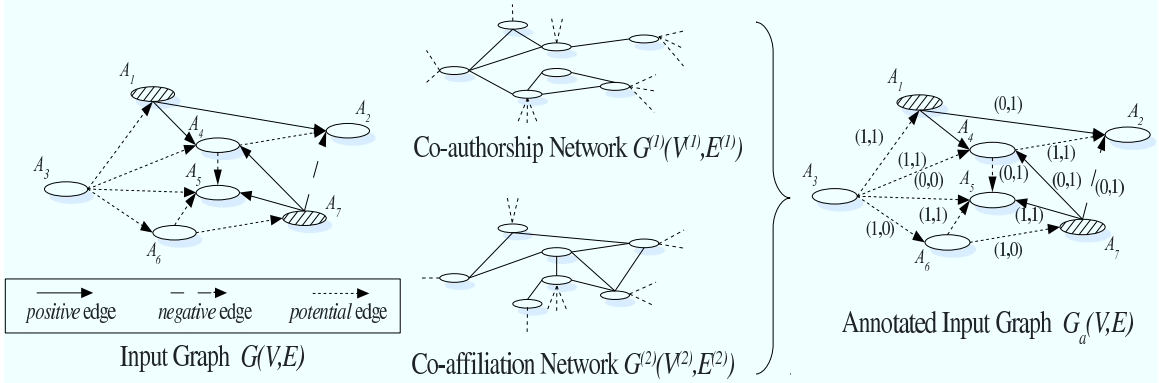

Figure 2: Feature generation for an example author network

4 Method

The proposed active surveying method uses a greedy probabilistic approach for solving the optimization problem. We use the current set of observed nominations as evidence for training a probabilistic classifier. The classifier is then used to infer how likely the potential nominations for each unsurveyed respondent are to be realized. Given the input graph G and the sets of node features $\mathcal{F}_v$ and edge features $\mathcal{F}_e$, a probabilistic classifier C is trained using the initial set of observed edges. For each un-observed, potential edge $e(u, v) \in E$, the classifier outputs the posterior probability of that edge being positive, denoted as $p(+|e(u, v))$, or negative, denoted as $p(-|e(u, v))$.

Given the output probabilities from the classifier along with the initial sets of *leaders* and *candidates* determined by the observed edges in G, we define a score function $S(v)$ for each node $v \in V$ as:

$$S(v) = \sum_{c \in candidates} p(+|e(v, c)) - \sum_{l \in leaders} p(+|e(v, l))$$

The score function $S(u)$ represents the difference between the expected number of nominated *candidates* and the expected number of nominated *leaders* for a given respondent u. Thus, following a greedy approach for finding the minimal set of respondents, the individual corresponding to node $v_S : \arg\max_v S(v)$ is then surveyed, and the resulting nominations are added to the training set and incorporated back into the model. Although there is an underlying independence assumption in predicting the respondents' choices of influential peers, we show in the experimental section that this approximation works well in practice.

One caveat with the above approach is the dependence between the quality of the decision of who to survey next with the accuracy of the underlying classifier. Therefore, a competing requirement is to choose respondents based on a criterion that will enhance the overall accuracy of the classifier. We rely on active learning to provide the necessary criterion for choosing the most informative respondents from the perspective of enhancing the overall accuracy of the underlying classifier.

In order to reduce the class probability estimation error, we use uncertainty sampling to select the respondents with the most uncertain responses, measured as the expected conditional classification error over their corresponding potential nominations. To choose the next respondent to survey, each nomination of a given respondent v is assigned a weight $w(e(v, u)) = (0.5 - |0.5 - p(+|e(v, u))|)$ indicating the distance of the class probability estimate from 0.5, which is used to quantify the amount of uncertainty in the class prediction. Then, the weight of each respondent $W(v)$ is computed as the average of all the weights on her outgoing nominations. The respondents' weights are then used to make a probabilistic choice of the next respondent. This weighted uncertainty sampling approach (WUS) has been shown to outperform traditional methods that pick the most uncertain sample [Saar-Tsechansky and Provost, 2004].

To provide a robust mechanism, we incorporate the two objectives of maximizing the likelihood to identify a new opinion leader and minimizing the expected classification error for choosing the next respondent. For that, we quantify the amount of uncertainty in the classifier output over all unobserved edges E_u as:

$$H_{avg} = \frac{1}{H_{max} \times |E_u|} \sum_{e(u,v) \in E_u} H(e(u, v))$$

where the entropy of the classifier output with respect to a given edge $e(u, v)$ is defined as:

$$H(e(u, v)) = - \sum_{l \in \{+,-\}} p(l|e(u, v)) log(p(l|e(u, v)))$$

and H_{max} is a normalization factor, representing the maximum entropy of the classifier output, so that H_{avg} is a valid probability value between $[0, 1]$

The next respondent to be surveyed v^* is then chosen via a probabilistic decision based on the current accuracy of the underlying classifier as follows:

$$v^* = \begin{cases} v \sim WUS & \text{with probability } p = H_{avg} \\ \arg\max_v S(v) & \text{with probability } p = (1 - H_{avg}) \end{cases}$$

Thus, the probability of choosing a respondent based on uncertainty sampling to enhance the classifier accuracy increases with higher uncertainty in the classifier output, while being more confident in the predictions yields a higher probability of choosing a respondent that optimizes the objective function $S(v)$. The full details are presented in Algorithm 1.

Algorithm 1 Active Survey Algorithm

repeat
  Train classifier C using observed nominations
  for each un-surveyed respondent v **do**
    Compute the objective function $S(v)$
    Compute the weight $W(v)$ using uncertainty sampling
  end for
  Normalize uncertainty sampling weights $W(v)$
  Set $v_S \leftarrow \arg\max_v S(v)$
  Set $v_{WUS} \sim W(v)$
  With probability $p = H_{avg}$, set $v^* \leftarrow v_{WUS}$, otherwise set $v^* \leftarrow v_S$
  Survey respondent v^*, update *leaders* and *candidates* sets according to the resulting nominations
  Remove v^* from the un-surveyed respondents and add her survey results to the set of observed nominations.
until required number of opinion leaders is obtained

5 Experimental Evaluation

To test our proposed method, we use a health care dataset generously provided by Community Analytics, a social marketing research organization which specializes in analyzing influence networks and identifying opinion leaders through conducting surveys of their clients' target audiences. The data represents survey information for nominating influential local physicians, provided by their peers.

The dataset consists of 2004 physicians, with 899 actual survey respondents generating 1598 nominations. As the surveys are based on identifying locally influential physicians, we limit the potential edges for each respondent to the physicians whose locations are within a 150 mile radius, yielding a set of 127,420 potential edges. By setting the nomination threshold (α = 2), we identified 260 opinion leaders.

By using the physicians' lists of publications from PubMed[1], we constructed both a citation and a co-authorship network among the physicians in the primary network. We also used the physicians' affiliation information to construct a co-affiliation network as a third supplementary source to leverage our data. Finally, using these three secondary networks, we generated a set of 20 edge features on the primary physician network and enriched the node features with additional attributes from these networks. A sample of the features included in the augmented network as the input to our model are illustrated in Table 1.

To conduct our experiments, we use a logistic regression classifier and vary the target percentage k of opinion leaders to be identified, showing the corresponding percentage of respondents required to reach this target using our proposed active survey method. We compare active surveying with a random baseline and a set of other baselines based on various centrality measures for determining the most informative physicians. We use three widely used centrality measures for the structural baselines: degree centrality, betweenness centrality, and page rank. In order to understand the relative contribution of the classifier versus active learning, we compare our proposed approach to a "*passive*" surveying method, which follows the same procedure of active surveying for optimizing the score function $S(v)$ based solely on the classifier's output, but does not incorporate uncertainty sampling. Finally, we show the performance of a method we refer to as perfect information (*PI*). *PI* uses the fully observed network and, at each step, greedily selects the survey respondent which identifies the maximum number of new opinion leaders. Note that the *PI* method represents a pseudo-optimal solution at each point, and hence the lower bound for the number of required respondents at each step.

[1] http://www.ncbi.nlm.nih.gov/pubmed

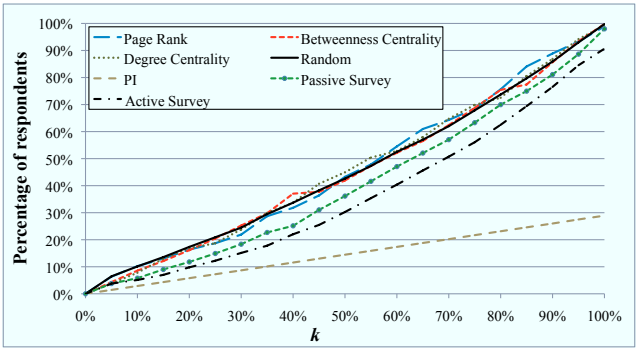

Figure 3: The percentage of respondents (y-axis) needed to identify k% of the opinion leaders (x-axis) at ($\alpha = 2$)

As can be seen in Figure 3, while the performance of the centrality-based methods is indistinguishable from the random baseline, both the passive and the active surveying methods perform significantly better than the baselines. Furthermore, our proposed active surveying method outperforms passive surveying, showing that our intelligent acquisition strategy helps to improve the quality of the learned classifier. Figure 4 shows the actual percentage of reduction in the size of the respondent set of both the active survey method and the perfect information method, with respect to the minimum set obtained by the best performing baselines at the corresponding value of k. As can be noted from the figure, our proposed approach yields a 30% average reduction in the number of respondents required, as compared to a 19% average reduction by the passive approach. The reduction attained by the active surveying method is reflected directly in surveying costs, thus helping survey conductors achieve their required goal at

Feature Name	Source Network
-Geographical Distance	$G_{nomination}$
-Respondent's current position (academic/non-academic)	$G_{nomination}$
-Nominee's current position	$G_{nomination}$
-Number of co-authored publications	$G_{co\text{-}authorship}$
-Nominee's publications count	$G_{co\text{-}authorship}$
-Number of respondent's citations of the nominee's publications	$G_{citation}$
-Nominee's h-index	$G_{citation}$
-Number of common affiliations	$G_{co\text{-}affiliation}$

Table 1: Sample features in the annotated physician nomination network

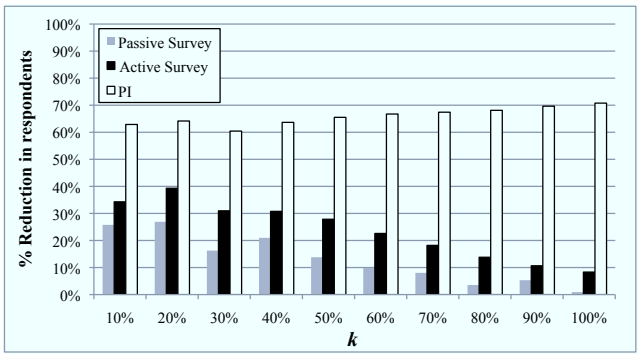

Figure 4: The percentage reduction in required respondents to identify $k\%$ of the opinion leaders at ($\alpha = 2$)

minimum cost. For instance, if a survey costs \$500 per person, then in order to identify 50% of the opinion leaders in the used physician network, the active survey method needs only 270 surveys rather than 375 surveys required by the best performing baseline; this leads to a savings of \$52,500.

6 Conclusion and Future Work

In this work, we presented a novel, dynamic framework for prioritizing the acquisition of survey information, for the purpose of leadership identification. The approach enables intelligent integration of both primary and secondary data to identify which respondents to survey, based on both the likelihood of them expanding the set of identified opinion leaders and also the utility of the information for improving future predictions. We then validated our results on a real-world dataset describing a physician nomination network.

Although our algorithm is focused on opinion leadership identification, we believe the idea of exploration vs. exploitation behind active surveying can generally be applied in different settings for guiding the survey process to reduce the associated costs. Future directions include introducing a weighting mechanism to account for varying acquisition costs, as well as dropping the independence assumption and utilizing a full joint approach for predicting nominations.

Acknowledgments

The authors would like to thank the anonymous reviewers for their valuable feedback. This work was supported by MIPS under Grant # 4409 and NSF under Grant # IIS-0746930.

References

[Dorfman and Maynor, 2006] S. Dorfman and J. Maynor. Under the influence. *Pharmaceutical Executive*, 26:148 – 150, 2006.

[Doumit *et al.*, 2007] Gaby Doumit, Melina Gattellari, Jeremy Grimshaw, and Mary Ann O'Brien. Local opinion leaders: effects on professional practice and health care outcomes. Cochrane Database of Systematic Revs., 2007.

[Keller and Berry, 2003] E. Keller and J. Berry. *One American in ten tells the other nine how to vote, where to eat, and what to buy. They are the influentials*. Free Press, 2003.

[Kempe *et al.*, 2003] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2003.

[Krackhardt, 1996] D. Krackhardt. Structural leverage in marketing. *Networks in Marketing*, pages 50 – 59, 1996.

[Lewis and Gale, 1994] D. Lewis and W. Gale. A sequential algorithm for training text classifiers. In *17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, 1994.

[Myers and Robertson, 1972] J. Myers and T. Robertson. Dimensions of opinion leadership. *Journal of Marketing Research*, 9:41 – 46, 1972.

[Raven, 1965] B. H. Raven. Social influence and power. *Current studies in social psychology*, pages 371 – 382, 1965.

[Rogers and Cartano, 1962] E. M. Rogers and D. G. Cartano. Methods of measuring opinion leadership. *Public Opinion Quarterly*, 26:435 – 441, 1962.

[Saar-Tsechansky and Provost, 2004] M. Saar-Tsechansky and F. Provost. Active sampling for class probability estimation and ranking. *Machine Learning*, 54(2):153 – 178, 2004.

[Settles, 2009] B. Settles. Active learning literature survey. Technical Report 1648, University of Wisconsin - Madison, 2009.

[Soumerai *et al.*, 1998] S. Soumerai, T. McLaughlin, J. Gurwitz, E. Guadagnoli, P. Hauptman, C. Borbas, N. Morris, B. McLaughlin, X. Gao, D. Willison, R. Asinger, and F. Gobel. Effect of local medical opinion leaders on quality of care for acute myocardial infarction: A randomized controlled trial. *The Journal of the American Medical Association*, pages 1358 – 1363, 1998.

[Thompson *et al.*, 1999] C.A. Thompson, M.E. Califf, and R.J. Mooney. Active learning for natural language parsing and information extraction. In *16th International Conference on Machine Learning*, 1999.

[Tong and Chang, 2001] S. Tong and E. Chang. Support vector machine active learning for image retrieval. In *ACM International Conference on Multimedia*, 2001.

[Tür *et al.*, 2005] G. Tür, D. Hakkani-Tür, and R.E. Schapire. Combining active and semisupervised learning for spoken language understanding. *Speech Communication*, 24(2):171 – 186, 2005.

[Valente and Davis, 1999] T. Valente and R. Davis. Accelerating the diffusion of innovations using opinion leaders. *The ANNALS of the American Academy of Political and Social Science*, 566(1):55 – 67, 1999.

[Wasserman and Faust, 1995] S. Wasserman and K. Faust. *Social network analysis: methods and applications*. Cambridge University Press, 1995.

[Watts and Dodds, 2007] D. Watts and P. Dodds. Influentials, networks, and public opinion formation. *Journal of Consumer Research*, 34(4):441 – 458, 2007.