Efficient Respondents Selection for Biased Survey using Online Social Networks *

Donghyun Kim[1], Jiaofei Zhong[2], Minhyuk Lee[1], Deying Li[3], and Alade O. Tokuta[1]

[1] Dept. of Math. and Physics, North Carolina Central University, Durham, NC 27707, USA. [donghyun.kim,atokuta]@nccu.edu, mlee28@eagles.nccu.edu

[2] Dept. of Math. and Computer Science, University of Central Missouri, Warrensburg, MO 64093, USA. zhong@ucmo.edu

[3] School of Information, Renmin University of China, Beijing 100872, China. deyingli@ruc.edu.cn

Abstract. Online social networks are getting lots of attentions from the research communities since they are rich sources of data to learn about the members of our society as well as the relationship among them. With the advances of Internet related technologies, online surveys are established as an essential tool for a wide range of applications. One significant issue of online survey is how to select a good respondent group so that the survey result is reliable. This paper investigates the use of online social network to form a biased survey respondent group which is useful for certain applications. We formally introduce a new optimization problem called the *minimum inverse k-core dominating set problem (MIkCDSP)* for this purpose, show its NP-hardness, and finally and mostly importantly introduce a greedy approximation algorithm for it.

1 Introduction

Recently, online social networks are receiving lots of attentions from the research communities due to the growing popularity of social networking web sites such

* This work was supported in part by US National Science Foundation (NSF) CREST No. HRD-1345219. This research was jointly supported by National Natural Science Foundation of China under grant 91124001, the Fundamental Research Funds for the Central Universities, and the Research Funds of Renmin University of China 10XNJ032.

as Facebook, Twitter, Google+, etc. It is widely recognized that the online social networks are rich sources of data to learn about the interest of each user as well as the relationship among them. Due to the reason, online social networks are investigated for a wide range of applications such as shared interest discovery among users [4], information propagation [6, 1], online advertising [5], efficient information propagation [7], community clustering [2], and so on.

These days, online surveys are established as an essential tool for a wide range of applications such as marketing and political decision making. It is known that in 2006, around 20% of global data-collection expenditure was spent for online survey research [8]. In 2012, US spent more than $1.8 billion for all survey research spending [9]. There are a number of reasons, not to mention its low cost (than the traditional methods), that online survey becomes so popular [10]. In online survey researches, how to find a right sample group of respondents is a long lasting conundrum since this is directly related to the reliability of the survey. Frequently, a biased respondent group is considered to be lack of its reliability. This is because the result from the surveys are mainly used to obtain a statistical information about the general public by consulting with a sample group from the public, the survey result from a sample group which lost its representative is not reliable for this purpose. Due to the reason, many efforts are made to find a representative and unbiased respondent group [9].

Interestingly enough, however, we observe the bias in the survey is not always something to avoid. Consider a product quality manager of a new smartphone, e.g. iPhone 5c, who wants to collect the feedback via an online survey from users so that he/she can improve the quality of the product. Also, suppose most of the customers using the product are happy with it. Then, while the manager is only interested in hearing complaints from the users, it is likely that the online survey result from the respondents selected by the methods whose common goal is to make the result representative and unbiased, would be mostly about their satisfaction about the new product. As a result, such survey is quite wasteful in practice to the manager who is only interested in complaints. Therefore, it would be helpful to form a biased respondent group so that it includes more unsatisfied users.

In this paper, we investigate the use of online social networks to compute a biased but representative respondent group such that the rate of the minority opinion group (e.g. those who are not satisfied with the product) in the respon-

dent group can be magnified. To the best of our knowledge, this is the first effort in the literature which exploits online social network to enhance to the quality of online survey. The rest of this paper is organized as follows: In Section 2, we introduce several important notations and definitions. Especially, we introduce the formal definition of our problem of interest, the *minimum inverse k-core dominating set problem (MIkCDSP)*, corresponding justification, and its NP-hardness result. Section 3 proposes a new greedy approximation algorithm for MIkCDSP. Finally, we conclude this paper in Section 4.

2 Notations, Definitions, and Problem Statement

2.1 Notations and Definitions

In this paper, $G = (V, E)$ represents an online social network graph with a node set $V = V(G)$ and an edge set $E = E(G)$. We assume the relationship between the members are symmetric and thus the edges in E are bidirectional. Also, we use n to denote the number of nodes in V, i.e. $n = |V|$. For any subset $D \subseteq V$, $G[D]$ is a subgraph of G induced by D. For each node $v \in V$, $N_{v,V}(G)$ is the set of nodes in V neighboring to v in G. Now, we introduce some important definitions.

Definition 1 (DS). *Given a graph G, a subset $D \subseteq V$ is a* dominating set (DS) *of G if for each node $u \in V \setminus D$, $\exists v \in D$ such that $(v, u) \in E$.*

Definition 2 (MDSP). *Given a graph G, the goal of the minimum dominating set problem (MDSP) is to find a minimum size DS of G.*

Definition 3 (Inverse k-core). *Given a graph G, a subset $D \subseteq V$, and a positive integer k such that $0 \leq k \leq \Delta$, where Δ is the degree of G, D is an* inverse k-core *in G if for each $v \in D$, $|N_{v,D}(G)| \leq k$.*

2.2 Problem Statement

In this paper, we study an online survey sample (respondents) selection problem such that the rate of people with minority opinion in the sample can be higher than their rate in the overall group.

We claim that people who share similar opinions have better chance to be a friend with each other in the online social networking, which is frequently true

in the professional social networks. Suppose for a survey, there exists an online social network relevant to the survey topic. For instance, for the survey on a new smartphone, we assume the existence of some online social network among the technicians. Then, based on our assumption, even though we do not know the opinion of each user in the social network regarding the smartphone, we can assume that two neighboring users in the social network have a smaller chance to have two drastically opposite opinion on the product.

A randomly selected DS of the whole group might be one approach to compute a good representative group since any node in the whole group either is a member of the DS or has a close friend who share similar opinion in the DS. However, note that while the DS has a representativeness, it is quite hard to tell if the DS is biased or not, and if biased, how much it is biased.

Based on our previous discussion, we claim the bias of the DS can be observed by checking its cohesiveness. That is, if there exists a clear majority opinion group in the overall group and the DS is completely randomly selected, then it is likely that the rate of the majority opinion group in the overall group is similar to the rate of the majority group in the DS. Furthermore, they will be appeared as a well-connected subgraph with relatively larger size in the social network graph induced by the DS. Meanwhile, there can be one or more well-connected subgraph with relatively smaller size in the graph, each of which represents a unique minority opinion group in the DS.

This observation implies that when we select a DS for the respondents if the degree of the induced graph by the DS is limited, then, the DS will include more amount of non-majority opinion group members. Formally, such a DS can be defined as the inverse k-core dominating set (IkCDS) showed below, where k is the degree of bias (with higher k, the DS is less biased, and with k equivalent to the degree of the social network, the DS is completely unbiased).

Definition 4 (IkCDS). *Given a graph G, a subset $D \subseteq V$, and a positive integer k, D is an* inverse k-core dominating set (IkCDS) *of G if (a) D is a DS of G and (b) for each $v \in D$, $|N_{v,D}(G)| \leq k$.*

It is noteworthy that there are a number of ways to compute IkCDS of a social network. Apparently, it is more desirable to reduce the size of IkCDS since it will cost less for the actual survey. As a result, the problem of computing a biased online survey respondent group can be formulated as MIkCDSP shown below.

Definition 5 (MIkCDSP). *Given a graph G and a positive integer k, the goal of the* minimum inverse k-core dominating set problem (MIkCDSP) *is to find a minimum size IkCDS of G.*

Remark 1. It is noteworthy that as k decreases, the DS will be more biased and the rate of minority opinion in the survey will increase. At the same time, the size of the IkCDS will decrease. This means that with very small k value, the survey respondent set can be very small and less practical given that the usual degree of social networks is not small. On the other hand, with very high k value, the survey respondent group can be negligibly biased, which also may not be desirable for our application. While selecting proper k value is very significant, it is also application dependent. Since this question is the out of this paper, we assume that k value is given as a part of the inputs of the problem.

The below theorem shows our problem is NP-hard.

Theorem 1. *MIkCDSP is NP-hard.*

Proof. A special case of MIkCDSP with $k = n$ is equivalent to the minimum dominating set problem, the problem of computing a minimum size dominating set of G, which is proven to be NP-hard [3]. As a result, MIkCDSP is NP-hard.

Remark 2. Given any graph G and a non-negative integer k, there exists a feasible solution of MIkCDSP in G. This claim is true since (a) a feasible solution of MIkCDSP with $k = 0$ is clearly a feasible solution of MIkCDSP with any $k \geq 1$, and (b) the following coloring strategy can be used to compute an independent set of G, the subset of nodes in G which are pairwise disjoint with each other, which is a feasible solution of MIkCDSP with $k = 0$: (i) initially color all nodes white, (ii) pick each white node black and its neighbors in gray until there is not white node left, and (iii) return the set of black nodes. Clearly, the set of black nodes is a dominating set and each pair of black nodes are not neighboring from each other.

3 Greedy-MIkCDSA: A Simple Greedy Approximation for MIkCDSP

In this section, we introduce Greedy-MIkCDSA, a simple greedy strategy for MIkCDSP and show that its performance ratio is $(1 + \Delta)$, where Δ is the de-

Algorithm 1 Greedy-MIkCDSA ($G = (V, E), k$)

1: Prepare an empty set D, i.e. $D \leftarrow \emptyset$.
2: For each $v_i \in V$, prepare a counter n_i which is initialized to 0, i.e. $n_i \leftarrow 0$.
3: Suppose $X_j = \{v_i | v_i \in V \text{ and } n_i = j\}$.
4: **while** $X_0 \neq \emptyset$ **do**
5: Find $v_i \in V \setminus \left((\bigcup_{j \geq k} X_j) \bigcup D \bigcup Q\right)$ so that $|N_{v_i, X_0}(G)|$ is maximized, where $Q = \{w_1, \cdots, w_q\}$ such that $w_l \in Q$ has at least one neighbor in $(\bigcup_{j \geq k} X_j)$ and $w_l \in D$ is true. A tie can be broken arbitrarily.
6: Set $D \leftarrow D \cup \{v_i\}$.
7: **for** each node $v_j \in N_{v_i, V}(G)$ **do**
8: $n_j \leftarrow n_j + 1$.
9: **end for**
10: **end while**
11: Output D.

gree of the input online social network graph. The formal description of Greedy-MIkCDSA is Algorithm 1. Given an MIkCDSP instance $\langle G, k \rangle$, Greedy-MIkCDSA first prepares an empty set D (Line 1), which will eventually include the output, an inverse k-core dominating set (IkCDS) of G. For each node $v_i \in V$, we create a counter n_i which is initialized to 0 (Line 2). The counter will be used to track the number of neighbors of v_i in D. Depending on the counter, we create a partition of the nodes in V, $X_0, X_1, \cdots$, where X_j is the subset of nodes in V whose counter is j (Line 3). This means that initially X_0 is equal to V and each of the rest is empty. Clearly, the number of the subsets is bounded by n. From Lines 4-10, we iteratively pick a node v_i from $\left((\bigcup_{j \geq k} X_j) \bigcup D \bigcup Q\right)$, i.e. v_i is a node which is

- **Condition 1**: with a counter n_i whose value is less than k (i.e. has less than k neighbors in DS),
- **Condition 2**: not selected as a DS node yet, and
- **Condition 3**: without any neighboring node w_l which is in D and, at the same time, in X_j for some $j \geq k$,

such that the number of neighbors of v_i in X_0 is the maximum. Any tie can be broken arbitrarily. This loop is repeated until all nodes in V is either in D or dominated by some node in D while maintaining $G[D]$ as an inverse k-core.

Clearly, Algorithm 1 produces a feasible solution of MIkCDSP since the algorithm repeatedly constructs D until X_0 becomes empty (which means D is a DS of G) and by Line 5, the degree of $G[D]$, the graph induced by D in G, will be bounded by k (which means D is an inverse k-core). One may wonder if there is a situation in which some node x, which has to be included in D to dominate some other node y, cannot be included in D since it has already k neighbors in D. However, this never becomes a problem since if x cannot be selected, then y itself will be included in D by our algorithm, which means that D is always a valid output.

Now, we show Algorithm 1 is a $(1+\Delta)$-approximation algorithm for MIkCDSP.

Lemma 1. *Given a graph $G = (V, E)$, let OPT_{MDSP} and OPT_{MIkCDS} be an optimal solution of MDSP and an optimal solution of MIkCDSP defined over $\langle G, k\rangle$ for some $k \geq 1$, respectively. Then,*

$$|OPT_{MDSP}| \leq |OPT_{MIkCDS}|.$$

Proof. By definitions, the goal of MDSP is to find a DS of G with minimum cardinality and the goal of MIkCDSP is to find a DS of G with minimum cardinality such that for each node in the DS, the node is allowed to be adjacent with at most k other nodes in the DS. Therefore, in any given G, an IkCDS of G is also a DS of G, but our choice of IkCDS is more limited than that of DS. As a result, this lemma is true.

Lemma 2. *Given a graph $G = (V, E)$, suppose we have an α-approximation algorithm of MDSP such that the output O of the algorithm is also a feasible solution of MIkCDSP. Then, we have $|O| \leq \alpha|OPT_{MIkCDS}|$.*

Proof. By the definition of an α-approximation algorithm of MDSP, we have $|O| \leq \alpha|OPT_{MDSP}|$. By combining this with Lemma 1, we have $|O| \leq \alpha|OPT_{MDSP}| \leq \alpha|OPT_{MIkCDS}|$, and thus this lemma is true.

Recall that Algorithm 1 produces a feasible solution of MIkCDSP defined over $\langle G, k\rangle$ which is also a feasible solution of MDSP defined over G. Therefore, by Lemma 2, we can obtain the performance ratio of Algorithm 1 for MIkCDSP by bounding the ratio between the size of an output of Algorithm 1 and the size of an optimal DS.

Theorem 2. *The performance ratio of Algorithm 1 for MIkCDSP is $1 + \ln \Delta$, where Δ is the maximum degree of G.*

Proof. Given $G = (V, E)$ and k, consider $OPT_{MDSP} = \{o_1, o_2, \cdots, o_l\}$ be a minimum DS of G. Then, for each $o_i \in OPT_{MDSP}$ in the increasing order of i, we compute

$$P_1 = \{o_1\} \bigcup N_{o_1, V \setminus OPT_{MDSP}}(G), \text{ and}$$
$$P_i = \left(\{o_i\} \bigcup N_{o_i, V \setminus OPT_{MDSP}}(G)\right) \setminus \left(\bigcup_{1 \le j \le i-1} P_j\right) \text{ for } i \neq 1.$$

Then, V is partitioned into $\mathcal{P} = \{P_1, P_2, \cdots, P_l\}$ such that each $P_i \in \mathcal{P}$ exactly includes one $o_i \in OPT_{DS}$.

Suppose Algorithm 1 is applied to $\langle G, k \rangle$ and outputs D. Then, each P_i can include some nodes in D. During the rest of this proof, we will try to find the upper bound of the size (i.e. the number of nodes) of $P_i \cap D$. If we can bound this size by α, we have

$$|D| \le \max_{1 \le i \le l} |P_i \cap D| \cdot |OPT_{MDSP}| = \alpha \cdot |OPT_{MDSP}|.$$

Remember that D is also an IkCDS. Therefore, by Lemma 2, we have

$$|D| \le \alpha \cdot |OPT_{MDSP}| \le \alpha \cdot |OPT_{MIkCDS}|,$$

which will complete this proof.

To obtain the upper bound of $|P_i \cap D|$, we consider the following strategy: whenever a node $v \in P_i$ is selected as a member of D by Algorithm 1, we assume each neighbor $u \in (P_i \bigcap X_0)$ of v immediately (before updating its counter) receives an additional weight $w(u)$, which is equivalent to one divided by the number of neighbors of v in $(P_i \bigcap X_0)$, i.e.

$$w(u) \leftarrow w(u) + \frac{1}{N_{v,(P_i \bigcap X_0)}(G)}.$$

Clearly, $\sum_{v \in P_i} w(v) = |P_i \cap D|$.

Next, we show that $\sum_{v \in P_i} w(v) = 1 + \ln \Delta$. If $P_i \cap D = \emptyset$, then this proof is trivial, and thus we assume $P_i \cap D \neq \emptyset$. Let $P_i \cap D = \{z_1, z_2, \cdots, z_p\}$. Also, let X_0 be the set of nodes in 'P_i' whose counter is 0, i.e. has no neighbor in D, yet. Note that each time, a node is selected by Algorithm 1 using the greedy strategy and added to D, there will be less number of nodes left in X_0. Let us use $X_0^{(0)}, X_0^{(1)}, \cdots, X_0^{(p)}$, where $X_0^{(i)}$ is the remaining nodes in X_0 after ith iteration of while loop (Line 4-9 in Algorithm 1). Then, we have

$$|X_0^{(0)}| \ge |X_0^{(1)}| \ge \cdots \ge |X_0^{(p)}|. \tag{1}$$

Note that for any j, $|X_0^{(j-1)}|-|X_0^{(j)}|$ is the number of nodes removed from $X_0^{(j-1)}$ after jth iteration. In other word, $|X_0^{(j-1)}| - |X_0^{(j)}|$ is the number of nodes in $X_0^{(j-1)}$, which are not adjacent to any node in $\{z_1, z_2, \cdots, z_{j-1}\}$ yet, and at the moment that z_j is selected, they are adjacent to z_j.

Suppose the initial iteration is executed and z_1 is selected and added to D. Then, the weight added to each neighbor of z_1 in $P_i \bigcap X_0^{(0)}$ is $1/N_{z_1,(P_i \bigcap X_0^{(0)})}$ and the number of such nodes is $|N_{z_1,(P_i \bigcap X_0^{(0)})}|$. In general, after jth iteration, the weight added to each neighbor of v_j in $P_i \bigcap X_0^{(j-1)}$ is $1/N_{v_j,(P_i \bigcap X_0^{(j-1)})}$ and the number of such nodes is $|N_{v_j,(P_i \bigcap X_0^{(j-1)})}|$. Since we are using a greedy strategy, z_j is always neighboring more nodes in $P_i \bigcap X_0^{(j-1)}$ than $o_i \in OPT_{DS}$. Therefore, we have

$$|N_{z_j,(P_i \bigcap X_0^{(j-1)})}| \geq |N_{o_i,(P_i \bigcap X_0^{(j-1)})}|,$$

which implies

$$\frac{1}{|N_{z_j,(P_i \bigcap X_0^{(j-1)})}|} \leq \frac{1}{|N_{o_i,(P_i \bigcap X_0^{(j-1)})}|}.$$

Since o_i is adjacent to all nodes in P_i, $N_{o_i,(P_i \bigcap X_0^{(j-1)})} = X_0^{(j-1)}$. As a result, after the iteration is repeated for p times. we have

$$\sum_{v \in P_i} w(v) \leq \sum_{1 \leq j \leq p} \frac{|X_0^{(j-1)}| - |X_0^{(j)}|}{|X_0^{(j-1)}|}. \tag{2}$$

By Eq. (1), we have $|X_0^{(j-1)}| - |X_0^{(j)}| > 0$ for all j. Finally, p can be bounded by Δ since all nodes in P_i has to be adjacent to o_i. As a result, the second term of the right side of Eq. (2) can be bound by $H(\Delta)$, where H is a harmonic function. As a result, we have

$$\sum_{v \in P_i} w(v) \leq 1 + H(\Delta) \simeq 1 + \ln \Delta,$$

and this theorem is true.

4 Conclusion

In this paper, we introduce a new approach to use the information from an online social network to enhance the result of online survey. To perform this task efficiently, we introduce to solve a new NP-hard optimization problem,

propose a new greedy heuristic algorithm for it, and show the algorithm in fact has a theoretical performance guarantee. To the best of our knowledge, this is the first attempt to use online social network to improve the result of online survey. We plan to further investigate the use of social network to improve the reliability of online voting systems. In this paper, we assume the existence of a single social network for survey. However, in reality, there could be more than one social networks which can be used for this kind of computation. Also, it would be very interesting to consider a social network with weighted edges.

References

1. H. Zhang, T. N. Dinh, and M. T. Thai, "Maximizing the Spread of Positive Influence in Online Social Networks," *Proc. of the IEEE Int Conference on Distributed Computing Systems (ICDCS)*, 2013.
2. Donghyun Kim, Deying Li, Omid Asgari, Yingshu Li, and Alade O. Tokuta, "A Dominating Set Based Approach to Identify Effective Leader Group of Social Network," *Proc. of the 19th Annual International Computing and Combinatorics Conference (COCOON 2013)*, June 2013.
3. M.R. Garey and D.S. Johnson, "Computers and Intractability: A Guide to the Theory of NP-completeness," *Freeman*, San Francisco, 1978.
4. F. Wang, K. Xu, and H. Wang, "Discovering Shared Interests in Online Social Networks," *International Workshop on Hot Topics in Peer-to-peer computing and Online Social Networking*, July, 2012.
5. C. Kahl, S. Crane, M. Tschersich, and Kai Rannenberg, "Privacy Respecting Targeted Advertising for Social Networks," *Proc. of the 5th IFIP WG 11.2 International Conference on Information Security Theory and Practice: Security and Privacy of Mobile Devices in Wireless Communication (WISTP)*, pp. 361-370, 2011.
6. W. Zhang, W. Wu, F. Wang, and Kuai Xu, "Positive Influence Dominating Sets in Power-Law Graphs," *Social Network Analysis and Mining*, vol. 2, no. 1, pp. 31-37, Mar. 2012.
7. M. Cha, A. Mislove, and K. Gummadi, "A Measurement-driven Analysis of Information Propagation in the Flickr Social Network," *Proc. of the 18th International Conference on World Wide Web (WWW)*, pp. 721-730, 2009.
8. V. Vehovar and K.L. Manfreda, "Overview: Online Surveys," *The SAGE Handbook of Online Research Methods*, London: SAGE (edited by N.G. Fielding, R.M. Lee, and G. Blank), pp. 177-194, 2008.
9. G. Terhanian and J. Bremer, "A Smarter Way to Select Respondents for Surveys?," *International Journal of Market Research*, vol. 54, no. 6, pp. 751-780, 2012.

10. B. Duffy, K. Smith, G. Terhanian, and J. Bremer, J, "Comparing Data from Online and Face-to-face Surveys," *International Journal of Market Research*, vol. 47, no. 6, pp. 615-639, 2005.