MINIMUM INFORMATION DOMINATING SET FOR CRITICAL SAMPLING OVER GRAPHS

Jianhang Gao[†], *Qing Zhao*[†], *Ananthram Swami*[§]

[†]Electrical and Computer Engineering, UC Davis, Davis CA 95616, jhgao, qzhao@ucdavis.edu

[§]Army Research Laboratory, Adelphi MD 20783, a.swami@ieee.org

ABSTRACT

We consider the problem of sampling a node-weighted graph. The objective is to infer the values of all nodes from that of a minimum subset of nodes by exploiting correlations in node values. We first introduce the concept of *information dominating set* (IDS). A subset of nodes in a given graph is an IDS if the value of these nodes is sufficient to infer the information state of the entire graph. We focus on two fundamental algorithmic problems: (i) how to determine whether a given subset of vertices is an IDS; (ii) how to construct a minimum IDS. Assuming binary node values and the local majority rule, we show that the first problem is co-NP-complete and the second problem is NP-hard in a general network. We then show that in acyclic graphs, both problems admit linear-complexity solutions by establishing a connection between the IDS problems and the vertex cover problem. For general graphs, we develop algorithms for solving both problems based on the concept of *essential differential set*. These results find applications in opinion sampling such as political polling and market survey in social-economic networks, and inferring epidemics and cascading failures in communication and infrastructure networks.

Index Terms— sampling; information dominating set; NP-complete; opinion polling; social networks.

1 Introduction

In this paper, we introduce and study the problem of critical sampling in node-weighted graphs. The objective is to infer the values of all nodes in a given graph from that of a minimum subset of nodes by exploiting correlations in node values.

This problem is motivated by opinion sampling in social or economic networks for applications such as political polling and market survey. Specifically, in social and information networks, it is often necessary to gauge the general opinion of a large population on a certain issue. Since polling often incurs a cost (either monetary or in terms of delay), an important question is how to infer the opinion of the entire network through a strategic sampling of a minimum subset of nodes by exploiting correlations in node opinions.

Besides the applications in social-economic networks, the problem of critical sampling over graphs and the results obtained in this paper also bear significance in identifying critical nodes in information networks. Identifying such critical nodes has important applications in learning and inference under resource constraints as well as security considerations in terms of protecting critical information hubs. When the value of a node indicates whether the node is infected, our results also apply to inferring, tracking, and controlling epidemics, worms and virus in communication networks, and cascading failures in infrastructure networks. This problem may also be applicable to data compression, given that the identified subset of nodes completely represents the information of the entire network.

[0]This work was supported by the Army Research Laboratory Network Science CTA under Cooperative Agreement W911NF-09-2-0053.

1.1 Information Dominating Set

We presents an algorithmic study of critical sampling over graphs. We first introduce the concept of information dominating set (IDS). A subset of nodes in a given node-weighted graph is an IDS if knowing the values of nodes in this subset is sufficient to infer the values of all nodes in the graph. We focus on two fundamental questions: (i) given a subset of nodes, how to determine whether it is an IDS; (ii) how to construct an IDS with a minimum number of nodes for a given graph. The former is referred to as the IDS checker (IDSC) problem, and the latter the minimum IDS (MIDS) problem.

While the concept of IDS applies to general information and information correlation models, in this paper, we focus on binary node values and adopt the local majority rule to model node correlation. Specifically, each node in the given graph has a binary value that is consistent with the majority opinion of its neighbors. Binary node values are sufficient to model yes/no opinions in social-economic networks and to indicate whether a node is infected in the study of epidemics and cascading failures. Local majority rule is also commonly used in studying opinion dynamics in social networks (see, for example, [1, 2]).

For binary node values and under the local majority correlation model, we show that the IDSC problem is co-NP-complete and the MIDS problem is NP-hard in a general graph. We then focus on graphs with special structures, in particular, acyclic graphs. We show that in acyclic graphs, both IDSC and MIDS problems admit linear-complexity solutions by establishing a connection between the IDS problem and the vertex cover problem. Our technique for establishing the hardness of the IDS problems is based on a novel graph transformation that transforms the IDS problems in a general graph to that in an odd-degree graph. This graph transformation technique not only gives an approximation algorithm to the NP-hard problem, but also provides a useful tool for general studies related to the local majority rule. For general graphs, we develop an efficient algorithm based on the concept of essential differential set to solve both the IDSC and the MIDS problems. This approach applies to general node values and general correlation models.

1.2 Related Work

Statistical sampling is a classic problem pioneered by Neyman in 1934 [3]. Different from the deterministic model and the algorithmic approach taken in this paper, statistical sampling assumes that the value associated with each node is a random variable obeying a known probability distribution, and designing the sampling strategy amounts to choosing the probability with which each node will be sampled. More recent work on statistical sampling can be found in [4–7].

In recent years, the concept of uniqueness set in a graph was proposed and studied in [8, 9] for sampling Paley-Wiener functions on graphs. However, the uniqueness set is different from the information dominating set because the former uniquely determines a band width limited function while the information on IDS uniquely determines the information in the rest of the graph.

The minimum vertex cover (MVC) [10, 11] and the minimum dominating set (MDS) [12, 13] are related to the IDS problem. The MVC asks for a minimum subset of vertices such that each edge in the original graph is adjacent to at least one vertex in this subset. And the MDS asks for a minimum subset such that each vertex is either in this subset or adjacent to at least one vertex in this subset. The minimum IDS problem is inherently more complex than MVC and MDS. For instance, as shown in this paper, it is co-NP-complete to verify whether a given subset is an IDS, while MVC and MDS have trivial polynomial time checkers simply based on their definitions.

The local majority rule has been adopted in studying opinion dynamics in social networks (see, for example, [1, 2]). The focus there is on characterizing the evolution of network opinions when each node dynamically changes its opinion by following the majority opinion of its neighbors. But absent from that line of work is the inference problem, which is the main objective of this paper: we aim to infer the network opinions *after* the opinion of each node has reached an equilibrium value.

2 Problem Formulation

In this section, we introduce the concept of IDS and formulate the IDSC and the MIDS problems. While the concept of IDS applies to critical sampling of graphs with general node weights and node correlations, we present the basic concepts and main results in the context of opinion sampling where nodes are binary valued satisfying the local majority rule.

2.1 Information Dominating Set

Given a graph $G = (V, E)$ with $n = |V|$ vertices, a *binary opinion profile* μ on G is a binary vector $(\mu_{v_1}, \ldots, \mu_{v_n})$ indicating where $\mu_{v_i} \in \{0, 1\}$ represents the opinion of vertex v_i. For a given a binary opinion profile μ on G, the neighbors of a vertex v_i are partitioned into two groups: the same-minded and opposite-minded neighbors, depending on whether they share the same opinion with v_i. In Fig. 1, the same-minded neighbors of v_3 are v_1, v_2 while its opposite-minded neighbor is v_4.

A *valid opinion profile* μ under the *local majority rule* in G is a binary opinion profile such that for each vertex v_i, the number of its same-minded neighbors is greater than or equal to the number of its opposite-minded neighbors. In other words, the opinion of each vertex is consistent with the majority opinion among its neighbors. If there is no such majority opinion, this vertex may take either opinion. Fig 1 demonstrates a valid opinion profile u.

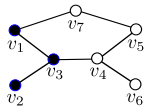

Fig. 1: The colors of vertices represents their opinions. In this example, the opinion profile is (1,1,1,0,0,0,0) and it is a valid opinion profile. Though the neighbors of both v_1 and v_7 are half black half white, they are still valid based on the definition.

The *valid opinion profile set* U of a given graph G is the set of all valid opinion profiles on G.

An *information dominating set* (IDS) in a given graph G is a subset of vertices $D \subseteq V$ such that under any opinion profile, the opinions of vertices in D is sufficient to infer the opinions of all the other vertices. Based on the definition, IDS has an important property as follows.

Property 1. *A subset of vertices D of a graph G is an IDS iff. for any pair of different valid opinion profiles μ, ν, there exists a vertex $v \in D$ such that $\mu_v \neq \nu_v$.*

The significance of Property 1 is that it provides a way to determine whether a subset of vertices is an IDS or not without considering any specific inference method. It is used repeatedly in this paper. Fig. 2 demonstrates the valid opinion profile set U and an IDS.

We focus on two problems on IDS. The IDS checker (IDSC) problem, seeks to determine whether a given set is an IDS. The second problem we consider is the main objective of this paper, which is to find the minimum IDS (MIDS). In hardness analysis, the corresponding decision problem is: given a graph G and a parameter k, does there exists an IDS D in G with size at most k.

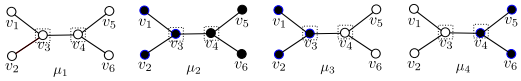

Fig. 2: There are only four valid opinion profiles on this graph. By Property 1, subset $\{v_3, v_4\}$ is an IDS.

2.2 Odd-degree Graph Transformation

We propose a graph transformation that allows us to study both the IDSC and the MIDS problems by considering odd-degreed graphs only without losing generality of the results. This transformation plays an important role in the hardness analysis and algorithm development given in subsequent sections.

Given an arbitrary graph $G = (V, E)$, we first copy every vertex and edge to G'. Then, for every even degree vertex v_i in G', we attach an auxiliary neighbor u_i (see Fig. 3). We call G' the *odd-degree transformation* of G. Given any valid opinion profile μ in G, we construct its odd-degree transformation opinion profile μ' by $\mu'_{v_i} = \mu_{v_i}$ and $\mu'_{u_i} = \mu_{v_i}$. In other words, those vertices derived from the original graph take the original opinions, and every auxiliary vertex

take the opinion of the vertex to which it is attached. Fig. 3 demonstrates an example of the odd-degree transformation from G to G' and a valid opinion profile μ to μ'.

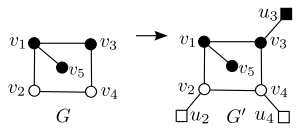

Fig. 3: An example of the odd-degree transformation from G to G'. The round vertices in G' are derived from G and the square vertices are the auxiliary vertices. It also shows the odd-degree transformation from μ to μ'.

The following theorem establishes a reduction from both IDSC and MIDS in G to the corresponding problems in G'. All results in this paper are stated without proof due to the space limitation.

Theorem 1. *There exists an IDS D in G if and only if there exists an IDS D' in G' such that for any vertex $v_i \in D$, either $v_i \in D'$ or its auxiliary vertex $u_i \in D'$.*

Based on Theorem 1, for both the IDSC and MIDS problems, it suffices to consider only odd-degree graphs. Unless otherwise noted, the graphs considered in the remaining part of this paper are all odd-degree graphs.

3 Hardness Analysis

In this section, we study the computational hardness of IDSC and MIDS. The following theorem establishes the co-NP-completeness of the IDSC problem.

Theorem 2. *Given a graph G and a subset of vertices D, it is co-NP-complete to determine whether D is an IDS of G.*

Since the checker problem is co-NP-complete, the minimum IDS problem may not belong to NP space. The following theorem establishes the NP-hardness of the MIDS problem.

Theorem 3. *Given a graph G, it is NP-hard to find the minimum IDS.*

4 IDS in Acyclic Graphs

In this section, we consider both IDSC and MIDS problem in acyclic graphs. An acyclic graph is a forest (i.e., a collection of trees). Since each connected component of the graph can be considered separately when studying the IDS problems, it suffices to focus on trees. We show, in Lemma 1, that an IDS without any leaf node is a vertex cover in an odd-degree tree. Since both an IDS or a vertex cover with leaf vertex can be transformed into a same size IDS or a vertex cover without any leaf vertex, respectively, we can solve IDSC and MIDS by solving the vertex cover problem.

Lemma 1. *Given an odd-degree tree G, an IDS that does not contain any leaf is also a vertex cover in G.*

The following lemma extends this result to any IDS.

Lemma 2. *Given any IDS D, $\exists$ an IDS D' that contains no leaf nodes and has a size smaller than or equal to D.*

With Lemma 2, we can solve the IDSC on a tree by checking whether its non-leaf transformation is a vertex cover. Furthermore, the following theorem provide us a way to find the MIDS.

Theorem 4. *The non-leaf minimum vertex cover is a minimum IDS.*

Since the non-leaf minimum vertex cover can be solved in linear time by a greedy algorithm, we can solve the MIDS on trees in linear time.

5 IDS in General Graphs

In this section, we develop an efficient algorithm for solving both the IDSC and the MIDS problems in general graphs. Based on the definition of IDS and Property 1, a brute-force solution to these problems is to consider every pair of valid opinion profiles. However, given the exponential order of the number of valid opinion profiles, this approach requires $O(2^{2n})$ time complexity. To address this issue, we introduce a concept called the *essential differential set* (EDS) that is much smaller in number than the valid opinion profile pairs, but still contains all the information needed for solving both the IDSC and the MIDS problems. An efficient algorithm, referred to as the wall separation algorithm, is then developed to find the EDS.

5.1 Essential Differential Set

We define Essential Differential Set (EDS) and establish the connection between EDS and the IDS problems.

A *set representation* $S(\mu)$ of a opinion profile μ is the set of vertices with opinion 1 in μ, i.e., the set $S(\mu) = \{v \in V | \mu_v = 1\}$. A *differential set* $D(\mu, \nu)$ is the exclusive disjunction of the sets representing two valid opinion profiles, i.e., $D(\mu, \nu) = S(\mu) \oplus S(\nu)$. The *essential differential set* is the family of all differential sets such that no other differential set is a subset of any set in the EDS. Based on Property 1 a subset D is an IDS if each differential set contains at least one vertex from D, i.e., subset D is a hitting set of the family of differential sets. The following theorem formally establishes the connection between EDS and IDS.

Theorem 5. *A subset of vertices D in a graph G is and IDS if and only if D is an hitting set of the EDS of G.*

Based on Theorem 5, given the EDS E of a graph G, we can solve the IDSC problem by checking whether the given subset is a hitting set of D or not. And furthermore, the MIDS problem becomes the minimum hitting set problem. In Sec. 5.3, we demonstrate that the average size of EDS is much smaller than the average number of valid opinion profiles in all our simulation cases. Hence the concept of EDS significantly reduces the problem size. What remains is to find the EDS given a graph G. We propose a wall separation algorithm in the next subsection for this problem.

5.2 The Wall Separation Algorithm

Based on its definition, the EDS E of a given graph G can be found by the following steps: list all the valid opinion profiles by exhaustive search; list all differential sets by considering all pairs of valid opinion profiles; eliminate those differential sets that are proper super sets of other differential sets. However, this procedure requires all the valid opinion profiles. We propose a wall separation algorithm that utilizes a double layered "wall" to partition this problem to smaller sub-problems and increase the efficiency. Before that, let us first define some terminology used in the algorithm. An *opinion sub-profile* $\mu^{V'}$ is an opinion profile on a subset of vertices V'. The opinions of

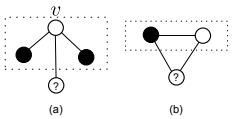

Fig. 4: The sub-profiles are on the vertices enclosed by the box. In (a), it is not a VOSP since vertex v has opinion 0 (denoted by white) but two of its three neighbors have opinion 1 (denoted by black). In (b), it is a VOSP even though in a complete valid opinion profile, all three vertices are either all black or all white.

the remaining vertices are undetermined. A *valid opinion sub-profile* (VOSP) $\mu^{V'}$ is a opinion sub-profile such that there is no known violation of the local majority rule for any vertex. Fig. 4 demonstrates two examples of opinion sub-profiles, one of which is valid, the other is not.

A VOSP μ^{V_1} *under* another VOSP ν^{V_2} such that $V_1 \cap V_2 = \emptyset$, denoted by $\mu^{V_1}|\nu^{V_2}$, is an opinion sub-profile such that the combination of both sub-profiles (an opinion profile such that the opinions on V_1 follows μ^{V_1}, the opinions of V_2 follows ν^{V_2} and the opinions of the remaining vertices are unknown) is still valid.

Now consider that graph G is partitioned into $k+1$ non-overlapping parts: a "double-layered wall" W and k other subgraphs $V_1, \ldots, V_k$ such that the distance between V_i and V_j is at least 3 hops. The following theorem states that under such a partition, given a particular sub-profile on the wall W, the sets of VOSPs on $V_1, \ldots, V_k$ are independent.

Theorem 6. *Given a graph G, consider an arbitrary partition of its vertices $\{W, V_1, \ldots, V_k\}$ such that the distance between V_i and V_j for arbitrary $i \neq j$ is at least 3. There exists a valid opinion profile μ in G if and only if there exists a VOSP ν^W and VOSPs $\nu^{V_1}|\nu^W, \ldots, \nu^{V_k}|\nu^W$ under ν^W such that all the VOSPs are consistent with μ.*

Based on Theorem 6, we proposed the wall separation algorithm (WSA) that contains two main steps. At the first step, WSA lists all the VOSPs on the wall and all the VOSPs on each V_i under every VOSP on the wall. At the second step, it builds the EDS based on the results in the first step. Additionally, there is a pre-processing algorithm that provides the partition of the graph. Based on Theorem 6, the partition of the graph does not affect the correctness of the algorithm. It only affect the time complexity of the two steps. We first describe the algorithm. And we then gives one realization of the pre-processing.

Given the graph G and a partition $\{W, V_1, \ldots, V_k\}$, the step 1 of WSA first list all VOSPs $\{\mu_1^W, \ldots, \mu_{m_l}^W\}$ on W by searching. Then for each $\mu_{m_l}^W$, we list all VOSPs $\{\mu_1^{V_i}|\mu_l^W, \ldots, \mu_{m_i'}^{V_i}|\mu_l^W\}$ for every V_i under μ_l^W.

In the second step, we construct the EDS by consecutively inserting candidate subsets to a family E of sets in a special way: if the candidate D is not a super set of any element in E, we add D in E and remove any element in E that is a super set of D. First, for any VOSP μ_l^W on the wall W, we insert $D(\mu_j^{V_i}|\mu_l^W, \mu_{j'}^{V_i}|\mu_l^W)$ into E for every i and every j, j'. Then, for every pair of VOSPs μ_l^W and $\mu_{l'}^W$, we compare the sets of VOSPs of each partition V^i under these two sub-profiles to check whether there exists $\mu_j^{V_i}|\mu_l^W$ and $\mu_{j'}^{V_i}|\mu_{l'}^W$ that are identical. Let $V^{i_1}, \ldots, V^{i_{k'}}$ be those partitions that does not have identical VOSPs under μ_l^W and $\mu_{l'}^W$. We insert $D(\mu_l^W, \mu_{l'}^W) \cup (\cup_{t=1}^{k'} D(\mu_{j_t}^{V_{i_t}}|\mu_l^W, \mu_{j_t'}^{V_{i_t}}|\mu_{l'}^W))$ into E for every j_t, j_t'. The following theorem establishes the correctness of the algorithm.

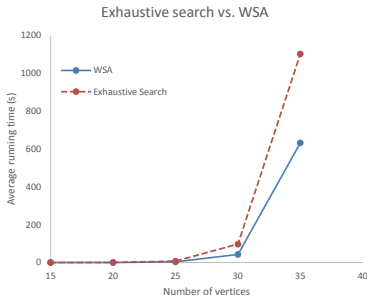

Fig. 5: The running time of exhaustive search algorithm and the WSA algorithm. Parameter $p = 0.2$ in the random graph model. The average is taken over 100 Monte Carlo runs.

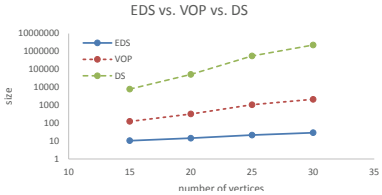

Fig. 6: The average size of EDS compared with the average number of valid opinion profiles and the number of differential sets. Parameter $p = 0.2$ in the random graph model. The average is taken over 100 Monte Carlo runs.

Theorem 7. *The family of sets, E, generate by WSA, is the EDS of the given graph G is the EDS of the given graph G.*

What remains is to find a partition $\{W, V_1, \ldots, V_k\}$. Note that only the efficiency but not the correctness of the algorithm depends on the partition. Since the algorithm is dominated by the first step, a good partition would ensure that all parts have the same size. In our simulation, we use a greedy algorithm to find the partition as follows. First, we calculate the shortest distances between all pairs of vertices. Then we start from a random vertex and sequentially select $k-1$ other vertices that on average are farthest away from all previous vertices. We use these k vertices as seed to grow to $V_1, \ldots, V_k$ by sequentially adding new adjacent vertices to V_i if the distance between any two sets is at least 3. The procedure stops either if there is no vertex to add in or the number of remaining vertices are smaller than that of largest among V_i.

5.3 Simulations

We compare the size of EDS to the size of all VOSPs and the size of all the differential sets under a random graph model where an edge occurs with probability p independently. As demonstrated in Fig. 6, the size of the EDS is significantly smaller than the number of valid opinion profiles and the number of all the differential sets. Fig. 6 shows the running time of WSA and exhaustive search algorithm to generate EDS under the same random graph model. The improvement in efficiency is clear.

6 References

[1] C. R. Plott, “A Notion of Equilibrium and Its Possibility Under Majority Rule”, In American Economic Review, volume 57, 1967.

[2] N. Mustafa, A. Pekec, “Majority Consensus and the Local Majority Rule”,Automata, Languages and Programming Lecture Notes in Computer Science Volume 2076, pp 530-542, 2001.

[3] J. Neyman, “On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection” Journal of the Royal Statistical, Society 97:558-625, 1934.

[4] F. Martin, L. Frankel,“Fifty Years of Survey Sampling in the United States”, Public Opinion Quarterly 51(Part 2), S127-38, 1987.

[5] D. Heckathorn, “Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations”, Social Problems, vol. 44, No. 2, pp. 174-199, 1997.

[6] S.K. Thompson, G.A.F. Seber, *Adaptive Sampling*, Wiley, 1996.

[7] P. Lavallée , *Indirect Sampling*, Springer, 2007.

[8] I. Pesenson, “Sampling in Paley-Wiener spaces on combinatorial graphs”, Transactions of the American Mathematical Society, vol. 360, no. 10, pp. 5603-5627, 2008.

[9] A. Anis, A. Gadde, A. Ortega, “Towards a sampling theorem for signals on arbitrary graphs”, Proc. IEEE ICASSP, pp. 3864-3868, 2014.

[10] R.M. Karp, “Reducibility Among Combinatorial Problems”, Complexity of Computer Computations, Plenum Press, pp. 85-103, 1972.

[11] G. Karakostas,“A better approximation ratio for the Vertex Cover problem”, Automata, Languages and Programming Lecture Notes in Computer Science Volume 3580, pp. 1043-1050, 2005.

[12] M.R. Garey, D.S. Johnson *Computers and Intractability: A guide to the theory of NP-completeness*, Freeman, San Francisco, 1978.

[13] U. Feige, “A Threshold of ln n for Approximating Set Cover”, Journal of the ACM, Vol. 45, No. 4, pp. 634-652, July 1998.