

Viewing the minimum dominating set and maximum coverage problems motivated by “word of mouth marketing” in a problem decomposition context

Anand Narasimhamurthy, Pádraig Cunningham, and Derek Greene

School of Computer Science and Informatics, University College Dublin
{anand.narasimhamurthy,padraig.cunningham,derek.greene}@ucd.ie

University College Dublin
Technical Report UCD-CSI-2009-5¹

Abstract. Modelling and analyzing the flow of influence is a key challenge in social network analysis. In scenarios where the network is too large to analyze in detail for computational reasons graph partitioning is a useful aid to decompose the large graph into manageable subgraphs. The question that arises in such a situation is *how to partition a given graph such that the solution obtained by combining the solutions from the individual subgraphs is as close as possible to the optimal solution obtained from the full graph (with respect to a particular objective)*. While graph cuts such as the *min cut*, *ratio cut* and *normalised cut* are a useful aid in breaking down the large problem into tractable subproblems, they may not yield the optimal graph partitioning with respect to a given objective. A natural question that arises in this scenario is *“How close is the solution given by the graph cut to that of the optimal partitioning?”* or in other words *Are the above graph cuts good heuristics?* In this report we pose the above questions with respect to two graph theoretic problems namely the minimum dominating set and maximum coverage. We partition the graphs using the normalised cut and present results that suggest that the normalised cut provides a “good partitioning” with respect to the defined objective.

1 Introduction

Social network analysis is of relevance in marketing due to the potential to model the flow of influence in the network. A useful objective is to identify “influential” nodes in the graph in order to selectively target them so as to “cover” as large a portion of the network as possible. Problems of this nature and their variants have been studied in various domains such as “word of mouth” promotion of new

¹ This work was supported by Science Foundation Ireland Grant No. 05/IN.1/I24 and Enterprise Ireland Grant No. PC/2007/010

products [1, 2], diffusion of technological innovations [3, 4], spread of communicable diseases [5, 6], cascading failures in power systems [7] and so on. Different types of models have been employed for modeling the dynamics of interaction between individuals in different domains.

An important problem in these models is the choice of an initial set of target nodes. This often leads to one of the well known NP-hard problems in graph theory. The Minimum Dominating Set (MDS) (and the set cover problem of which the MDS may be regarded as a special case) and Maximum Coverage problems arise in many applications. We motivate these problems as arising from applying a “first approximation model” in a word-of-mouth marketing scenario. The minimum dominating set can be shown to arise as a result of a special case of Independent Cascade Model, (in fact Kempe *et al.* [8] show that the discrete optimization problem for the Independent Cascade Model is NP-hard by showing that the Set Cover problem is a special case). These aspects are discussed in subsequent sections.

Given that many real world networks consist of millions of nodes, it may often not be possible to work on the entire graph. For instance it may be more feasible to run computationally intensive analysis techniques on subgraphs obtained via graph partitioning rather than on the entire graph. While there exists a significant amount of literature related to the dominating set and maximum coverage problems, the main contribution of this report is examining the above graph covering problems in a problem decomposition context.

Although most of the analysis in this report is general (i.e. not specific to any application), the analysis is motivated from a marketing scenario. While many such problems are amenable to be formulated as facility location problems, we make the following broad distinction. While fairness is an important consideration in facility location problems (i.e. it is important that every node is covered adequately by its nearest facility), this is usually not the case in a marketing setting. The aim is to *cover* as many nodes as possible, possibly within a given budget allocation.

This report is organised as follows. We present a snapshot of relevant previous work in section 2. In section 3 we provide the background and briefly review some of the results in the literature most relevant to this report. We motivate the problem in section 4. In section 5 we examine the graph coverage problems in a problem decomposition context. We present experimental results on synthetic and real data in section 6. The report finishes with conclusions and future work in section 7.

2 Previous work

The problem of identifying influential nodes in graphs arises in many applications. In marketing the aim is to identify influential nodes in the graph in order to target them with promotions in order to maximise the viral impact of a marketing campaign. Which nodes are influential and how best to target them depends on the actual application. In models for the spread of diseases,

the main concerns are to devise optimum disease surveillance strategies and/or identify high risk individuals for vaccination; the overall goal being to contain an outbreak of a disease. On the other hand in devising marketing strategies, the aim is to target individuals with the overall goal of spreading the innovation within the population as much as possible. For example given a call graph of mobile phone subscribers, the object is a targeted placement of offers to selected mobile phone subscribers with the goal of spreading the word out to as many subscribers as possible.

A number of researchers have looked at the propagation of viruses, whether biological or computer viruses, through a social network, in the context of epidemic modelling. In many such approaches individuals in the network are assumed to be in one of a number of possible states. In this context, state I corresponds to *infected* individuals, state S corresponds to *susceptible* individuals who may become infected through contact with other infected people and state R corresponds to *recovered* individuals. Commonly used models include the SIS and SIR. In a SIR model, recovered individuals cannot be reinfected whereas, in an SIS model, infected individuals return to the susceptible state after recovery.

Various models have been employed for modeling the dynamics of interaction between individuals in different domains. Many mathematical models for the spread of communicable diseases use differential equations based on uniform mixing assumptions [9] or other ad hoc models for the contact process. Granovetter proposed a model incorporating node specific thresholds to model collective behaviour [10]. Domingos and Richardson [1] consider a marketing context, they model the social network as a Markov random field where each customer's probability of buying is a function of both the intrinsic desirability of buying the product as well as the influence of other customers (network effects). They pose the question of which individuals to select for targeted marketing of new innovations or products with a goal of triggering a large cascade of further adoptions and suggest heuristics for identifying customers who have a large overall effect on the network.

Other models include dynamic cascade models drawing from work on interacting particle systems. The **Linear Threshold** and **Independent Cascade** models are examples of two widely used diffusion models, both are briefly outlined below. Although many variants of the Linear Threshold model exist, a characteristic of such models is that the tendency of a node to become active increases as more of its neighbours become active. Specifically, a node v is influenced by each neighbour w according to a weight $b_{v,w}$ such that $\sum_w b_{v,w} \leq 1$. Also each node v chooses a threshold θ_v which is the fraction of neighbouring nodes that are activated. In the case of the independent cascade model, there is an initial set of active nodes A_0 and the process unfolds in discrete time steps as follows: When a node i becomes active say at time step t , it attempts to activate its inactive neighbours. Let the probability that node i is successful in activating node j ($j \in N(i)$) be $P(i, j)$. If i succeeds in activating j , then j becomes active at time step $t + 1$. However, node i is not allowed any further attempts to activate any of its neighbours later on.

Goldenberg *et al.* [2] investigated the independent cascade model in the context of word of mouth marketing. Kempe *et al.* [8] study the problem of maximizing the spread of influence through a social network in a marketing context. They consider the issue of choosing influential individuals as a discrete optimization problem. They show that the influence maximization problem is NP-hard for linear threshold and independent cascade models and obtain provable bounds for a natural greedy strategy using an analysis framework based on sub modular functions.

3 Preliminaries : Minimum dominating set, maximum coverage problems and submodular functions

We briefly outline two results from the literature which are used in the rest of the paper. The first concerns the **minimum dominating set** problem which is an instance of the well known **set cover problem** [11]. The second is a result from Nemhauser *et al.* [12] utilising the theory of **submodular functions**.

3.1 Minimum Dominating Set

The minimum dominating set (MDS) problem may be stated formally as, “Given a graph $G = (V, E)$, a dominating set is a subset $V' \subseteq V$ such that for all $u \in V$ there is a $v \in V'$ for which $(u, v) \in E$.” The MDS problem entails finding a dominating set of the smallest size for a given graph $G = (V, E)$ i.e. finding the smallest set of nodes V' such that for any node in the graph either it or one of its neighbours belongs to V' .

This problem and its variants have been addressed by various researchers. We only aim to give a small sample of the recent literature in graph theory pertaining to the minimum dominating set. The MDS problem is NP-hard even for classes of graphs such as bipartite graphs and chordal graphs, however recently there has been interest in exact algorithms for minimum dominating set [13–15] (these are of course exponential time algorithms). In [13] Fomin *et al.* present algorithms with running times $O(1.4143^n)$, $O(1.7321^n)$ and $O(1.5144^n)$ respectively for split graphs, bipartite graphs and graphs of maximum degree three. They also present an $O(1.9379^n)$ algorithm for general graphs. Exact algorithms for MDS on general graphs have also been proposed by Randerath and Schiermeyer [15] ($O(1.8899^n)$) and by Grandoni [14] ($O(1.8026^n)$). Kuhn and Wattenhofer [16] develop a new fully distributed approximation algorithm based on LP relaxation techniques. For an arbitrary, possibly constant parameter k and maximum node degree Δ , their algorithm computes a dominating set of expected size $O(k\Delta^{2/k}\log(\Delta) | DSOPT |)$ in $O(k^2)$ rounds. They claim this to be the first algorithm which achieves a non-trivial approximation ratio in a constant number of rounds.

The MDS is a special instance of the *Set Cover problem* [11], stated formally as follows.

Let U be a finite set and F a family of subsets of U , such that every element of

U belongs to at least one subset in F . i.e. $U = \cup_{S \in F} S$. The problem is to find a minimum size subset $C \subseteq F$ whose members cover all of U i.e. each member of U belongs to at least one set in C . The set cover problem is NP-hard [17], however, a polynomial time greedy approximation algorithm may be employed; in each step of the greedy set cover algorithm, the set S covering the largest number of remaining uncovered elements is picked. This algorithm is reproduced below from Cormen et al. [11].

GREEDY-SET-COVER(X,F)

```

 $U \leftarrow X$ 
 $C \leftarrow \phi$ 
while  $U \neq \phi$ 
do select an  $S \in F$  that maximizes  $|S \cap U|$ 
 $U \leftarrow U - S$ 
 $C \leftarrow C \cup S$ 
return  $C$ 

```

It can be proved that the greedy set cover algorithm returns a set cover whose size is at most $H(d_{max})$ times the size of an optimal (the smallest) set cover where

- d_{max} is the cardinality of the largest set i.e. $d_{max} = \max |S| : S \in F$ and
- $H(d) = \sum_{i=1}^d 1/i$ is the d th harmonic number.

The reader is referred to [11] for a description and related analysis. The result in [11] expressing the size of the greedy set cover in terms of the optimal set cover can be written as

$$c^* \geq \frac{g}{H(d_{max})} \tag{1}$$

where

- g = the size of the greedy set cover i.e. the number of sets selected in the set cover by the greedy set cover algorithm.
- c^* = the size of an optimal (the smallest) set cover.
- d_{max} = the maximum degree among all the nodes and

$$H(d_{max}) = \sum_{i=1}^{d_{max}} 1/i$$

The size of the greedy cover is a $1 + \log(n)$ bound on the optimal since $H(d_{max}) \leq 1 + \log(n)$.

Another NP-hard problem that arises in many practical applications is the **maximum coverage problem** described below. *Given a universe U of n elements and a family F of subsets of U , and an integer $k \leq |F|$, find the k -element subset $C \in F$ (i.e. $|C| = k$) that maximises $\cup_{S \in C} S$. i.e. choose a subset of the family F such that the number of elements covered is maximised.* The same greedy algorithm as the set cover can be employed here as well, the only difference being that the algorithm terminates after k sets are chosen. Utilising the

theory of *sub modular functions* [12], Nemhauser *et al.* show that the greedy algorithm provides a value within $(1 - 1/e)$ of the optimal. We briefly review sub modular functions below.

3.2 Submodular functions

Let $f(\cdot)$ be an arbitrary function that maps subsets of a finite ground set U to non-negative real numbers. $f(\cdot)$ is said to be submodular if it satisfies

$$f(S \cup v) - f(S) \geq f(T \cup v) - f(T) \quad (2)$$

for all elements v and all pairs of sets $S \subseteq T$. There are a number of equivalent definitions of submodular functions. Sub modular functions have a number of nice tractability properties. One such property is that of “diminishing returns” given by (2) above; the marginal gain from adding an element to a set S is at least as high as the marginal gain from adding the same element to a superset of S .

Let f be a non-negative, monotone submodular function, let $S_k^{(g)}$ be a set of size k obtained by selecting elements one at a time, each time choosing an element that provides the largest marginal increase in the function value. Let S_k^* be a set that maximizes value of f over all k -element sets. Nemhauser *et al.* [12] show that the greedy approach provides a value within $(1 - 1/e)$ of the optimal.

Switching back to the maximum coverage problem, let S be a set of nodes, and let $f(S)$ denote the number of nodes covered by this set. It can be seen easily that $f(S)$ is a submodular function. Let $S_k^{(g)}$ denote the k -element set obtained by a greedy selection of elements i.e. the set obtained by adding that element which yields the largest marginal increase in the value of the function f at each step. The result from Nemhauser *et al.* can be expressed as :

$$f(S_k^{(g)}) \geq (1 - 1/e)f(S_k^*) \quad (3)$$

4 Problem formulation

Given a social network graph we consider the problem of choosing an initial set of nodes so as to ensure with high probability, that as large a fraction of the whole network is covered. This can be approached in many ways. For instance, Domingos and Richardson [1] model the social network in a marketing context as a Markov random field. Kempe *et al.* [8] consider the issue of choosing influential sets of individuals as a discrete optimization problem. They consider the Linear Threshold and Independent Cascade models (and their generalizations) both of which involve an initial set of *active* nodes A_0 . In this section we formulate the problem as a discrete optimization problem similar in spirit to Kempe *et al.* [8]. We show that the Minimum Dominating Set and Maximum Coverage problems described in section 3 arise as special cases.

Let A_0 denote the set of initially active nodes. These active nodes try to activate other nodes which in turn attempt to activate others and so on. In

general the activation process is probabilistic, hence we define an **activation probability matrix** P where

$$P(i, j) = \text{Prob}(\text{Node } i \text{ succeeds in activating node } j \mid i \text{ is currently active})$$

This is equivalent to a graph $G = (V, E, P)$ where V is the set of vertices, E the set of edges and the edge weights are the corresponding activation probabilities given by P . The problem may be stated as :

Problem 1

Given an activation probability matrix P , choose a set A_0 consisting of k (a fixed number) initially active nodes so as to maximize the expected number of active nodes at the end of the process.

Given an initial set of active nodes $A_0 = \{u_1, u_2, \dots, u_k\}$ at time $t = 0$, the expected number of active nodes at $t = 1$ can be computed as follows:

Let S be a binary vector such that

$$S(i) = \begin{cases} 1 & \text{if node } i \text{ is active} \\ 0 & \text{otherwise.} \end{cases} \quad (4)$$

Let $S^{(0)}, S^{(1)}, \dots$ denote these vectors at time steps $t = 0, 1, \dots$ respectively. Let $E^{(1)}(A_0), E^{(2)}(A_0) \dots$ denote the expected number of active nodes at time steps $t = 1, 2, \dots$ respectively.

Let v be a node that has edges connecting one or more of nodes in A_0 . Assuming that the attempts of active neighbours of v to activate v are independent, the probability that v becomes active at time step $t = 1$ can be expressed as :

$$\text{Prob}(S^{(1)}(v) = 1 \mid A_0) = 1 - \prod_{u_i \in A_0} (1 - P(u_i, v)) \quad (5)$$

Compute the probabilities of activation for each node according to (5). The expected number of active nodes at time step $t = 1$ (excluding the set of initial active nodes) is then :

$$E^{(1)}(A_0) = \sum_{v \in V - A_0} \text{Prob}(S^{(1)}(v) = 1 \mid A_0) \quad (6)$$

Computing the expected number of nodes at the end of the process depends on the model employed for the propagation of information. The problem of choosing a set of k initially active nodes so as to maximize the expected number of nodes at the end is intractable in many cases since it usually entails computing the expected number of active nodes for each combination of k initially active nodes.

We simplify the problem as follows.

Problem 2

Choose a set of initially active nodes $A_0 = \{u_1, \dots, u_k\}$ so as to maximize the expected number of active nodes in the next time step.

However as we shall soon see, this problem can be shown to be NP-hard by considering a special case, namely that all the probabilities of activation are

either 1 or 0. In this case, every neighbour of any of the initially active nodes is activated. Thus for each of the nodes $i = 1, \dots, n$ we can associate a set of nodes covered by node i , namely $C_i = N(i)$ where $N(i)$ denotes the neighbours of i .

Given a fixed number of initially active nodes (say k) the problem reduces to finding a set of nodes $A_0 = \{u_1, u_2, \dots, u_k\}$ such that $|\cup_{j=1}^k C_{u_j}|$ is maximized. This is the **maximum coverage** problem described in section 3. The question of what is the minimum number of initially active nodes such that *all* nodes are active in the next time step, is the **minimum dominating set** problem.

For the rest of the report we consider the special case of deterministic activations and address the Minimum Dominating Set and Maximum Coverage problems. Let $G = (V, E)$ (V is the set of vertices and E the set of edges) be a binary *coverage graph* where

$$G(i, j) = \begin{cases} 1 & \text{if node } i \text{ "covers" node } j \\ 0 & \text{otherwise.} \end{cases} \quad (7)$$

In case of completely deterministic activations, $G(i, j) = 1$ implies that node i activates node j with probability 1, however this is unlikely to be the case in most real networks since there is uncertainty as to whether information is propagated from one node to another. In practice, G could be obtained from the original network, by applying criteria most suitable to the particular application. For instance the coverage matrix could be the binary matrix obtained by applying a threshold on the edge weights of the original network. If the edge weights represent probabilities of activation, then retaining only edges with weight above a threshold (say p_{thresh}) amounts to enforcing that any currently inactive node has a probability of activation of at least p_{thresh} if any of its neighbours are active (this can be seen easily from 5). In this case $G(i, j) = 1$ implies that i activates j with a probability $\geq p_{thresh}$.

5 Graph coverage problems in a problem decomposition context

5.1 Notation

We introduce some notation which will be used in the rest of the paper. Let G be the graph with nodes $1, 2, \dots, n$.

- Let $cover(G, S)$ where $S \subseteq V$ denote the set of nodes in graph $G = (V, E)$ **covered** by the set of nodes S
For an undirected, unweighted graph define $cover(G, S) = \cup_{x \in S} (N(x) \cup x)$ where $N(x)$ denotes the neighbours of x .
- Let $C^*(G, k)$ denote the maximum number of nodes in graph G covered by selecting k covering nodes i.e. $C(G, k) = \max_S |cover(G, S)|$ where $|S| = k$.
- Let $S_k^{(g)}$ denote the set of k nodes in $G = (V, E)$ ($S_k^{(g)} \subset V$) chosen according to the greedy algorithm.

- Let G be partitioned into $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ consisting of n_1 and n_2 nodes respectively. Let the maximum node degrees for the graphs G_1, G_2 and G be d_1, d_2 and d respectively (for a directed graph let these be the maximum out-degrees). Let g, g_1 and g_2 be the sizes of the greedy set covers (i.e. covers obtained by the greedy approach) of G, G_1 and G_2 respectively.

5.2 Processing subgraphs obtained by partitioning

We now consider the dominating set and maximum coverage problems in a problem decomposition context. We define the objective function $X^{(optimal)}$ as the ratio of the solution obtained by combining the optimal solutions from the individual subgraphs to the optimal solution obtained from the original graph. In the case of the MDS the aim is to *minimise* the number of elements in the dominating set whereas in the case of the maximum coverage the aim is to *maximise* the number of covered nodes. Accordingly we define a performance ratio ($\rho^{(optimal)}$) such that the maximum value of $\rho^{(optimal)}$ is 1 as

$$\rho^{(optimal)} = \min(X^{(optimal)}, 1/X^{(optimal)}) \quad (8)$$

Since computing the optimal solutions in the case of minimum dominating set and maximum coverage is NP-hard, $X^{(optimal)}$ and $\rho^{(optimal)}$ cannot be computed directly. We also define an empirical performance ratio $\rho^{(greedy)}$ as the ratio of corresponding empirical quantities computed using a greedy algorithm (described in section 3).

We now formulate the questions stated in general terms earlier, with respect to the MDS and maximum coverage problems as follows, focussing on the case of $p = 2$ partitions.

Minimum Dominating Set The objective in the case of $p = 2$ partitions is

$$X_{MDS}^{(optimal)} = \frac{c_1^* + c_2^*}{c^*}$$

where c_1^*, c_2^* and c^* be the sizes of the optimal (smallest) dominating sets of G_1 and G_2 and G respectively. The corresponding performance ratios are :

$$\text{Ratio of optimal quantities } \rho_{MDS}^{(optimal)} = \frac{c^*}{c_1^* + c_2^*} \quad (9)$$

$$\text{Ratio of empirical quantities } \rho_{MDS}^{(greedy)} = \frac{g}{g_1 + g_2} \quad (10)$$

where,

c_1^*, c_2^* and c^* are the sizes of the minimum dominating sets of G_1 and G_2, G respectively and g_1, g_2 and g are sizes of dominating sets obtained by the greedy algorithm. Note that while $\rho_{MDS}^{(optimal)}$ is always ≤ 1 , $\rho_{MDS}^{(greedy)}$ could exceed 1.

Maximum coverage We now consider the questions posed earlier with respect to the maximum coverage problem and derive the objective function as follows. As before let $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ be subgraphs obtained by partitioning $G = (V, E)$. For the maximum coverage, the given inputs are the graph G and a budget k . There is an additional criterion, namely the allocation of budgets i.e. the number of covering nodes allocated to G_1 and G_2 , let these be k_1 and k_2 respectively.

Let $S_1^* \subseteq V_1$ and $S_2^* \subseteq V_2$ be the sets of nodes which form the optimal k_1 -element and k_2 -element optimal covers in G_1 and G_2 respectively ($k_1 + k_2 = k$). Let S^* be the optimal k -element cover for G .

Analogous to an empirical performance ratio for the MDS ($\rho_{MDS}^{(greedy)}$) defined in (10) we define one for the maximum coverage problem as :

$$\rho_{MC}^{(greedy)} = \frac{|cover(G_1, S_{k_1}^{(g)})| + |cover(G_2, S_{k_2}^{(g)})|}{|cover(G, S_k^{(g)})|} \quad (11)$$

where $S_{k_1}^{(g)}$ ($S_{k_1}^{(g)} \subset V_1$) is the set of k_1 nodes in G_1 and $S_{k_2}^{(g)}$ is the set of $k_2 = k - k_1$ nodes in G_2 both chosen according to the greedy algorithm.

For both the MDS and maximum coverage problems, in the trivial case where the graph is not fully connected, it is obvious that the best possible bound of 1 is achieved if G is partitioned into G_1 and G_2 where G_1 and G_2 had no interconnections in G . However it is not obvious how to achieve a good (if not the optimal) cut with respect to the objective functions stated above in the general case.

The normalized cut [18, 19] tends to produce balanced partitions. We suggest that the normalized cut provides a *good cut* with respect to minimising $\rho_{MDS}^{(optimal)}$ or $\rho_{MC}^{(optimal)}$. In the next section we present results on real graphs which suggest that the normalised cut may be a good heuristic towards this end. We refer the reader to [20] where Meila & Shi interpret the edge weights of a similarity matrix as edge flows in a Markov random walk and study the properties of the eigenvalues and eigenvectors of the resulting transition matrix. With this framework, they show the equivalence between spectral formulation of normalized cut algorithm and eigenvalues/eigenvectors of the stochastic matrix obtained by normalizing the similarity matrix. Armed with this they argue *the normalized cut results in partitions such that the random walk once in one of the parts tends to remain in it i.e. the probability of moving over to the other part is minimized.*

6 Experiments

6.1 Data description

We illustrate our results on synthetic and real datasets. For the synthetic data we generate power law graphs, for the real data we used graphs derived from two bibliographic datasets. The datasets are described in some detail next.

Synthetic data

It is often observed that for many real graphs the degree distribution follows a power law rather than a Poisson distribution [21]. Power law graphs are characterised by a significantly higher probability of nodes with large degrees as compared to a random graph. These nodes represent *hubs* which have a significant impact on overall connectivity of a graph, since a connected set of hubs can provide short paths between a large proportion of nodes in the network. Other characteristics often associated with power law graphs are high cluster coefficient and a scale-free property i.e. an observation of certain properties at different scales.

The degree sequence of a graph is the set of node degrees (d_1, d_2, \dots, d_n) ordered such that $d_1 \geq \dots \geq d_n$. To construct synthetic graphs that approximate real world power-law graphs, random graphs with a given fixed input degree sequence were generated and the degree sequence was chosen to follow an exact power law. For $k = 1, \dots, n$, node degrees d_k were chosen such that

$$d_k = \text{round}(ck^\alpha) \quad (12)$$

The constant of proportionality c was set by specifying the minimum degree $d_n = d_{min}$, this yields

$$d_k = \text{round}\left(\left(\frac{k}{n}\right)^\alpha d_{min}\right) \quad (13)$$

A reasonable value of α was chosen based on real-world graphs. For our experiments we used the values $\alpha = 10$ and $d_{min} = 8$ to generate the powerlaw graphs. Power law graphs consisting of different numbers of nodes ($n = 1000, 2000, 5000$ and 10000) were generated. For each value of n (number of nodes) 50 powerlaw graphs were generated and the experiments were repeated on each of the 50 graphs. The basic statistics of these graphs are shown in figure 1(a), the average values computed from 50 runs. The degree distribution of one of the power law graphs consisting of 1000 nodes is shown in figure 1(b).

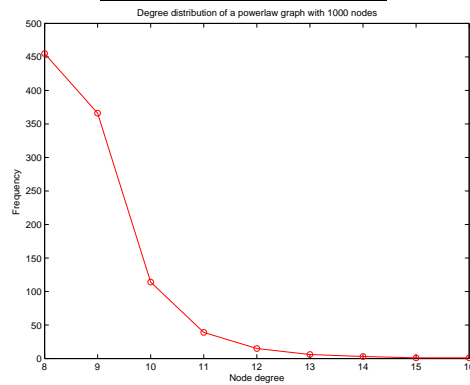
Real-world graphs do not exhibit the perfect power-law characteristics of these synthetic graphs, but rather tend to exhibit a power-law behaviour in certain ranges of node degrees. Nevertheless, this is a useful approximation that allows us to examine algorithms on large-scale graphs which exhibit some of the characteristics shown by real-world graphs.

Real data

It is often argued that co-authorship networks capture many key features of social networks in general(eg. [22]). Ideally, while it may be best to use graphs from the intended target application (eg. marketing data in our case), a practical issue in social network analysis is that many large datasets such as phone caller data are not publicly available due to privacy and/or proprietary reasons. On the other hand, bibliographic information tends to be more accessible publicly and hence graphs derived from bibliographic datasets tend to be used more often. We

# nodes	Average # edges	Average highest degree
1000	4417	16
2000	8837	17
5000	22095	19
10000	44195	20

(a)



(b)

Fig. 1. (a) Basic statistics of simulated power law graphs used in our experiments. These graphs were generated according to (13) with $d_{min} = 8$ and $\alpha = 10$ (b) Degree distribution of a power law graph consisting of 1000 nodes

suggest that the graphs derived from CORA dataset form a reasonable proxy for real-world data. For instance the communities are small despite the scale of the graph – typically containing 10 to 20 members, this characteristic is observed in real networks such as mobile subscriber networks. Moreover our analysis and results are quite general and hence not dependent on the specific nature of the graphs (although the actual numbers may vary significantly depending on the graph).

The CBR and CORA datasets are described next. The basic statistics of the resulting graphs are summarised in table 1. In most of our experiments the edge directions and weights were ignored and the largest weakly connected component was used (the numbers shown in table 1 correspond to the largest connected component). The experiments pertaining to the computation of empirical bounds from subgraphs obtained by partitioning the original graph are described in 6.2.

CBR dataset

Co-citation analysis has become the dominant method for the empirical study of the structures of scientific communication [23]. The Case Based Reasoning (CBR) co-citation dataset was collected from a bibliographic analysis of fifteen years of CBR conferences. The aim was to examine the themes that have evolved in CBR research as revealed by the implicit and explicit relationships between the conference papers. The interested reader is referred to [24] for more details. The CBR cocitation dataset consists of a total of 672 papers from the CBR

conference series, published by 828 individual authors. It is publicly available for research purposes². We refer to this dataset as **CBR Cocitation (Full)**. We also performed experiments on a subset of the CBR dataset this consisted of papers only within the CBR conferences, we refer to this as **CBR Cocitation (Small)**.

CORA dataset

The CORA bibliographic dataset [25] contains information and annotations such as authors, cited papers and topic for over 50,000 research papers³. We used the paper citation graph, author citation and coauthorship graphs derived from the CORA dataset for our experiments. Only the papers with available authors were selected. The number of individual papers and authors were 28400 and 24951 respectively. In each of the graphs the largest connected component was extracted and used for our experiments. The basic statistics of these graphs are shown in table 1, as can be observed the highest degree is quite large. Many real world networks are characterised by the presence of *hubs* which are well connected nodes, the presence of hubs tends to enhance overall connectivity.

The author-author and coauthorship graphs were generated from the paper citation data by assigning edges between the respective authors of corresponding papers i.e. if paper A cited paper B, we assigned edges between each of the authors of papers A and B. While many possibilities have been suggested in co-citation analysis literature for assigning the weights of the edges of author co-citation and coauthorship graphs so constructed, we mostly used these graphs as a proxy for social networks; we ignored the edge weights in most cases. We also illustrate a few results on the graphs after thresholding the edge weights to exclude spurious and chance relationships, however this did not make any significant difference with respect to the overall results.

Graph description	# nodes	# edges	Highest degree
CBR Cocitation (Small)	247	8055	202
CBR Cocitation (Full)	657	24682	439
CORA paper citation	24542	100860	381
CORA author citation	21720	558143	1690
CORA coauthorship	20876	131936	428

Table 1. Basic statistics of real graphs (the largest connected component) derived from bibliographic datasets used in our experiments

² can be downloaded from <http://mlg.ucd.ie/cbr>

³ see <http://www.cs.umass.edu/mccallum/code-data.html>

6.2 Empirical bounds for partitioned graphs

In this section we describe experimental results pertaining to computation of empirical bounds from subgraphs obtained by partitioning the original graph. As mentioned earlier, while we would like to maximize the performance ratios $\rho_{MDS}^{(optimal)}$ and $\rho_{MC}^{(optimal)}$, these cannot be computed directly since the minimum dominating set and maximum coverage are NP-hard. However the empirical performance ratios $\rho_{MDS}^{(greedy)}$ and $\rho_{MC}^{(greedy)}$ defined as the ratio of corresponding empirical quantities in (10) and (11) respectively, can be directly computed. We argue that the normalised cut is a good candidate for partitioning the graph from the point of view of obtaining a good performance ratio. We used the software Graclus to perform the graph partitioning. This software implements the multilevel method proposed by Dhillon et al. [26]. One main advantage of this software is that it can compute the normalised and ratio cuts without a computationally intensive eigenvalue decomposition.

Table 2 shows the results for power law graphs described in section 6.1 for graphs consisting of 1000, 2000, 5000 and 10000 nodes. For each number of nodes (say 1000), the results shown are averaged over 50 runs. The numbers shown in the columns except the first and the last are the mean (standard deviation) of the respective quantities. The normalised cut provides good empirical performance ratios in most cases. These are shown in the last columns of tables 2 (power law graphs) and 3 (bibliographic datasets) respectively. For the power law graphs $\rho_{MDS}^{(greedy)}$ is computed as the ratio of the average sizes of the corresponding greedy dominating sets i.e. $\rho_{MDS}^{(greedy)} = \frac{g^{(avg)}}{g_1^{(avg)} + g_2^{(avg)}}$, where $g_1^{(avg)}$, $g_2^{(avg)}$ and $g^{(avg)}$ are the average values (computed over 50 runs) of the greedy dominating sets of the two subgraphs and the full graph respectively.

Full graph		Sub graph 1		Sub graph 2		$\rho_{MDS}^{(greedy)} = \frac{g^{(avg)}}{g_1^{(avg)} + g_2^{(avg)}}$
# nodes	Mean (std) $g^{(avg)}$	Mean (std) # nodes	Mean (std) $g_1^{(avg)}$	Mean (std) # nodes	Mean (std) $g_2^{(avg)}$	
1000	172 (2.2)	503 (37.9)	107 (7.1)	497 (37.9)	106 (6.3)	0.81
2000	332 (3.6)	1001 (48.7)	218 (8.4)	999 (48.7)	216 (8.4)	0.77
5000	861 (4.6)	2542 (168.3)	526 (32.2)	2548 (168.3)	543 (29.4)	0.80
10000	1712 (26.3)	5010 (194.3)	1050 (39.9)	4990 (194.3)	1045 (39.8)	0.82

Table 2. Power law graphs : Mean (std. dev) number of nodes, sizes of dominating sets obtained by greedy algorithm and empirical performance ratio. Columns 1-2 : Full graph, Columns 3-6 : Two subgraphs obtained by partitioning the graph using normalised cut. Column 7 : Empirical performance ratio $\rho_{MDS}^{(greedy)}$

The results pertaining to real graphs described in section 6.1 (and whose basic statistics are shown in table 1) are shown in table 3. The graphs were partitioned into $p = 2$ partitions and the corresponding empirical performance ratios were computed. For these bibliographic graphs too the performance ratio

of empirical quantities $\rho_{MDS}^{(greedy)}$ is quite high, even more pronounced for the Cora graphs.

Table 4 shows the situation for the same graphs for the maximum coverage problem with $k = 10, 100, 1000$ and number of partitions $p = 2$. For the maximum coverage problem, the number of covering nodes allocated to each subgraph was proportional to the size of the subgraph (number of nodes in the subgraph). Assume $G = (V, E)$ is partitioned into subgraphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ ($V_1 \cup V_2 = V$). Let k be the number of covering nodes for the full graph. The number of covering nodes allocated to subgraphs G_1 and G_2 were:

$$k_1 = \text{round} \left(k \frac{|V_1|}{|V|} \right)$$

$$k_2 = \text{round} \left(k \frac{|V_2|}{|V|} \right)$$

where $|V|, |V_1|, |V_2|$ are the number of vertices in graph G and subgraphs G_1 and G_2 respectively.

Graph description	Full graph		Sub graph 1		Sub graph 2		$\rho_{MDS}^{(greedy)} = \frac{g}{g_1 + g_2}$
	# nodes	g	# nodes	g_1	# nodes	g_2	
CBR Cocitation (Small)	247	11	102	7	145	6	0.84615
CBR Cocitation (Full)	657	27	288	13	369	20	0.8181
CORA paper citation	24542	4476	12689	2528	11853	2055	0.9766
CORA author citation	21720	1209	10546	545	11174	781	0.9118
CORA coauthors	20876	3107	13687	2099	7189	1183	0.9466

Table 3. Real graphs derived from bibliographic datasets : Number of nodes and sizes of corresponding dominating sets obtained by greedy algorithm and empirical performance ratio. Columns 2-3 : Full graph, Columns 4-7 : Two subgraphs obtained by partitioning the graph using normalised cut. Column 8 : Empirical performance ratio $\rho_{MDS}^{(greedy)}$

Next we varied the number of partitions and computed the empirical performance ratios for the minimum dominating set and maximum coverage problems. As mentioned above, for the maximum coverage problem, the number of covering nodes for the full graph was ($k = 100$) and the numbers allocated to the subgraphs (k_1 and k_2) were proportional to the sizes of the subgraphs. Figure 2 shows the empirical performance ratios $\rho_{MDS}^{(greedy)}$ and $\rho_{MC}^{(greedy)}$ (as defined by (10) and (11)) corresponding to the MDS and maximum coverage respectively, for the Cora graphs. The x-axis represents the number of partitions from 1 (no partitioning) to 10. Figures 2(a) and 2(b) correspond to the MDS and max coverage respectively, on the Cora graphs with no thresholds applied on edge weights. The author-author citation and coauthorship graphs were essentially weighted graphs. We applied a threshold (p_{thresh}) on the edge weights, retaining only

Graph	Full graph		Subgraph1		Subgraph2		$\rho_{MC}^{(greedy)} = \frac{ C(G_1, S_{k_1}^{(g)}) + C(G_2, S_{k_2}^{(g)}) }{ C(G, S_k^{(g)}) }$
	k	$ C(G, S_k^{(g)}) $	k_1	$ C(G_1, S_{k_1}^{(g)}) $	k_2	$ C(G_2, S_{k_2}^{(g)}) $	
CORA Paper citation	10	2403	5	1286	5	997	0.95
	100	8166	52	3831	48	4017	0.96
	1000	18550	517	9101	483	9197	0.99
CORA author	10	6411	5	3451	5	2421	0.92
	100	15377	49	7606	51	7080	0.95
	1000	21521	486	10491	514	10895	0.99
CORA coauthor	10	1942	7	1289	3	546	0.94
	100	6869	66	4323	34	2140	0.94
	1000	16931	656	10914	344	5620	0.98

Table 4. Maximum coverage on the entire graph and the subgraphs for different numbers of covering nodes k selected by greedy approach. $|C(G, S_k^{(g)})|$, $|C(G_1, S_{k_1}^{(g)})|$, $|C(G_2, S_{k_2}^{(g)})| = \#$ nodes covered by covering nodes selected by greedy approach for G, G_1, G_2 respectively.

those edges with weight $> p_{thresh}$ and repeated the experiments. The results for $p_{thresh} = 2$ are shown in figures 2(c) and 2(d) and those for $p_{thresh} = 5$ in 2(e) and 2(f). It is seen that the subgraphs obtained by using the normalised cut yield good empirical performance ratios for both the Minimum Dominating Set and Maximum coverage problems, for instance even for 10 partitions the drop in the empirical performance ratios are quite small.

7 Conclusions and Future Work

In this work we examine two graph coverage problems, namely the minimum dominating set and maximum coverage problems loosely motivated by a word-of-mouth marketing scenario, in a problem decomposition context. We consider the question of how to partition a given graph such that the solution obtained by combining the optimal solutions from the individual subgraphs is as close as possible to the optimal solution obtained from the original graph, focussing mostly on the case of $p = 2$ partitions. We define performance ratios in terms of the ratio of the combined optimal solutions from the individual subgraphs to the optimal solution obtained from the original graph as benchmarks for assessing the quality of a given partitioning. Considering that the above problems are NP-hard and hence the optimal performance ratio cannot be computed directly, we also define an empirical performance ratio expressed in terms of solutions obtained by applying a greedy algorithm. In our experimental results the normalised cut consistently provided a high empirical performance ratio, we suggest that it may also provide a good partitioning with respect to the optimal performance ratio, at the same time producing balanced partitions. Future work

involves exploration of this aspect, an explicit connection would provide a useful analysis framework.

References

1. Domingos, P., Richardson, M.: Mining the network value of customers. In: Seventh International Conference on Knowledge Discovery and Data Mining. (2001)
2. Goldenberg, J., Libai, B., Muller, E.: Using complex systems analysis to advance marketing theory development. *Academy of Marketing Science Review* (2001)
3. Valente, T.: Network models of the diffusion of innovations (1995)
4. Rogers, E.: Diffusion of innovations (1995)
5. Eubank, S., Guclu, H., Kumar, V.S.A., Marathe, M.V., Srinivasan, A., Toroczkai, Z., Wang, N.: Modelling disease outbreaks in realistic urban social networks. *Nature* **429** (2004) 180–184
6. Stanley, E.A.: Social networks and mathematical modelling. *Connections* **27** (2006) 43–49
7. Asavathiratham, C., Roy, S., Lesieutre, B., Verghese, G.: The influence model. *IEEE Control Systems* (2001)
8. Kempe, D., Kleinberg, J., Tardos, E.: Maximizing the spread of influence through a social network. In: SIGKDD'03, ACM (2003) 137–146
9. Kaplan, E., Craft, D., Wein, L.: Emergency response to a smallpox attack: the case for mass vaccination. *Proc. Natl Acad. Sci (PNAS)* **99** (2002) 10935–10940
10. Granovetter, M.: Threshold models of collective behaviour. *American Journal of Sociology* **83** (1978) 1420–1443
11. Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C.: Introduction to Algorithms. MIT Press, Cambridge, Massachusetts, USA (2001)
12. Nemhauser, G., Wolsey, L., Fisher, M.: An analysis of the approximations for maximizing submodular set functions. *Mathematical Programming* **14** (1978) 265–294
13. Fomin, F.V., Kratsch, D., Woeginger, G.J.: Exact (exponential) algorithms for the dominating set problem. In: LNCS. Volume 3353. (2004) 245–256
14. Grandoni, F.: A note on the complexity of minimum dominating set. *Journal of Discrete Algorithms* (to appear)
15. Randerath, B., Scheirmeyer, I.: Exact algorithms for minimum dominating set. (Technical Report zaik-469)
16. Kuhn, F., Wattenhofer, R.: Constant-time distributed dominating set approximation. *Distrib. Comput.* **17** (2005) 303–310
17. Karp, R.M.: Reducibility among combinatorial problems (1972)
18. Shi, J., Malik, J.: Normalized cuts and image segmentation. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR '97). (1997) 731–737
19. Shi, J., Malik, J.: Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22** (2000) 888–905
20. Meila, M., Shi, J.: (A random walks view of spectral segmentation)
21. Barabasi, A.L., Albert, R.: Emergence of scaling in random networks. *Science* **286** (1999) 509–512
22. Newman, M.: The structure of scientific collaboration networks. *Proc. Natl. Acad. Sci.* **98** (2001)

23. Markus, G.: Co-citation analysis and the search for invisible colleges : A methodological evaluation. *Scientometrics* **57** (2003) 27–57
24. Greene, D., Freyne, J., Smyth, B., Cunningham, P.: An analysis of research themes in the cbr conference literature. In: Proc. 9th European Conference on Case-Based Reasoning (ECCBR 2008). (2008)
25. McCallum, A., Nigam, K., Rennie, J., Seymore, K.: Automating the construction of internet portals with machine learning. *Information Retrieval Journal* **3** (2000) 127–163 www.research.whizbang.com/data.
26. Dhillon, I.S., Guan, Y., Kulis, B.: Weighted graph cuts without eigenvectors a multilevel approach. *IEEE Trans. Pattern Anal. Mach. Intell.* **29** (2007) 1944–1957

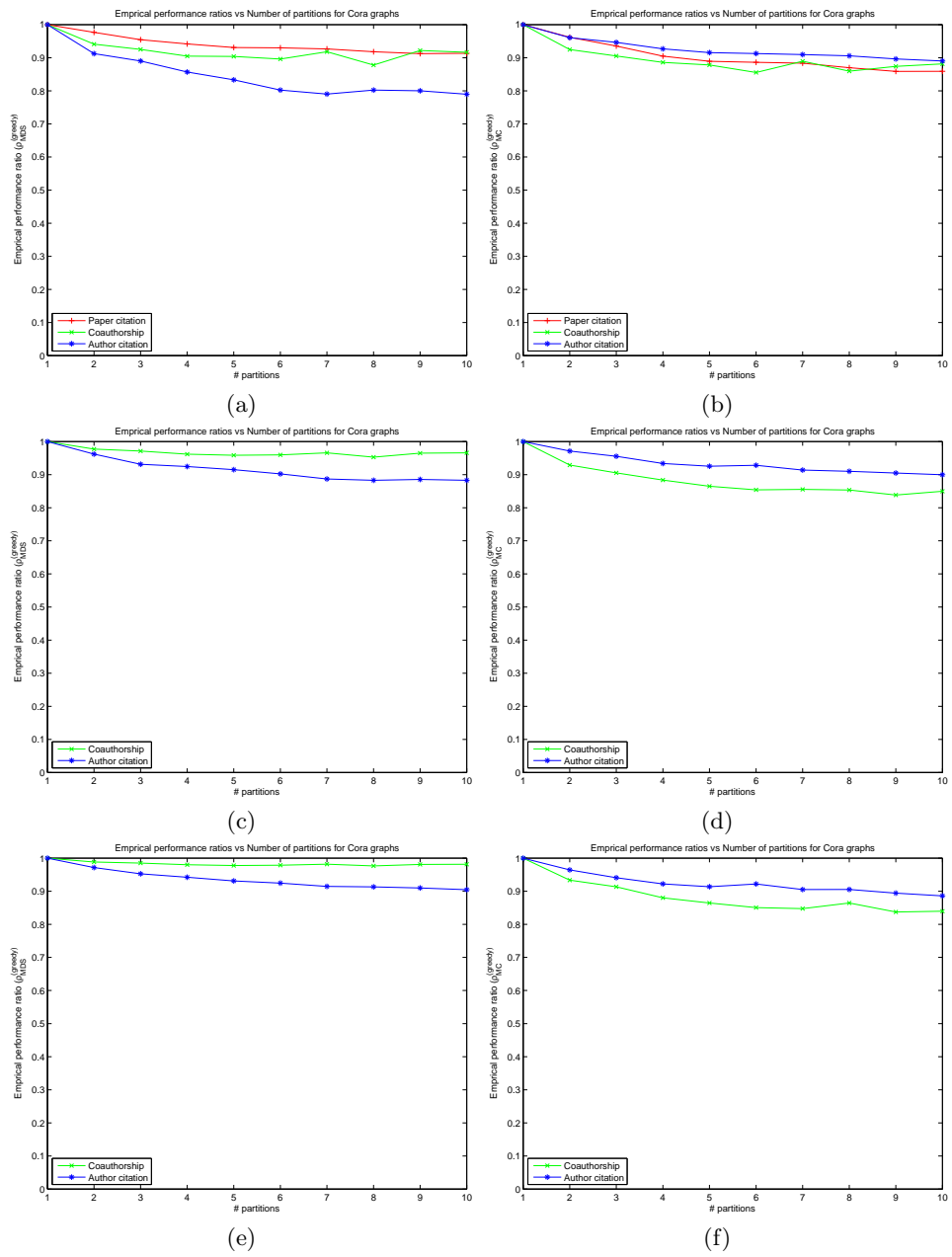


Fig. 2. Empirical performance ratios for Cora graphs for varying number of partitions. (a),(c),(e) : Minimum Dominating Set. (b),(d),(f) : Maximum Coverage (a) and (b) : Graphs with no thresholds on edge weights. (c) and (d) : $p_{thresh} = 2$ (e) and (f) $p_{thresh} = 5$