

Scalable Influence Maximization for Multiple Products in Continuous-Time Diffusion Networks

Nan Du

DUNAN@GOOGLE.COM

Google Research, 1600 Amphitheatre Pkwy, Mountain View, CA 94043

Yingyu Liang

YINGYUL@CS.PRINCETON.EDU

Department of Computer Science, Princeton University, Princeton, NJ 08540

Maria-Florina Balcan

NINAMF@CS.CMU.EDU

School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213

Manuel Gomez-Rodriguez

MANUELGR@MPI-SWS.ORG

MPI for Software Systems, Kaiserslautern, Germany 67663

Hongyuan Zha

ZHA@CC.GATECH.EDU

Le Song

LSONG@CC.GATECH.EDU

College of Computing, Georgia Institute of Technology, Atlanta, GA 30332

Editor: Jeff Bilmes

Abstract

A typical viral marketing model identifies influential users in a social network to maximize a single product adoption assuming unlimited user attention, campaign budgets, and time. In reality, multiple products need campaigns, users have limited attention, convincing users incurs costs, and advertisers have limited budgets and expect the adoptions to be maximized soon. Facing these user, monetary, and timing constraints, we formulate the problem as a submodular maximization task in a continuous-time diffusion model under the intersection of one matroid and multiple knapsack constraints. We propose a randomized algorithm estimating the user influence¹ in a network ($|\mathcal{V}|$ nodes, $|\mathcal{E}|$ edges) to an accuracy of ϵ with $n = \mathcal{O}(1/\epsilon^2)$ randomizations and $\tilde{\mathcal{O}}(n|\mathcal{E}| + n|\mathcal{V}|)$ computations. By exploiting the influence estimation algorithm as a subroutine, we develop an adaptive threshold greedy algorithm achieving an approximation factor $k_a/(2 + 2k)$ of the optimal when k_a out of the k knapsack constraints are active. Extensive experiments on networks of millions of nodes demonstrate that the proposed algorithms achieve the state-of-the-art in terms of effectiveness and scalability.

Keywords: Influence Maximization, Influence Estimation, Continuous-time Diffusion Model, Matroid, Knapsack

1. Introduction

Online social networks play an important role in the promotion of new products, the spread of news, the success of political campaigns, and the diffusion of technological innovations. In these contexts, the influence maximization problem (or viral marketing problem) typically

1. Partial results in the paper on influence estimation have been published in a conference paper: Nan Du, Le Song, Manuel Gomez-Rodriguez, and Hongyuan Zha. Scalable influence estimation in continuous time diffusion networks. In *Advances in Neural Information Processing Systems* 26, 2013.

has the following flavor: identify a set of influential users in a social network, who, when convinced to adopt a product, shall influence other users in the network and trigger a large cascade of adoptions. This problem has been studied extensively in the literature from both the modeling and the algorithmic aspects (Richardson and Domingos, 2002; Kempe et al., 2003; Leskovec et al., 2007; Chen et al., 2009, 2010a,b, 2011, 2012; Ienco et al., 2010; Goyal et al., 2011a,b; Gomez-Rodriguez and Schölkopf, 2012), where it has been typically assumed that the host (*e.g.*, the owner of an online social platform) faces a single product, endless user attention, unlimited budgets and unbounded time. However, in reality, the host often encounters a much more constrained scenario:

- **Multiple-Item Constraints:** multiple products can spread simultaneously among the same set of social entities. These products may have different characteristics, such as their revenues and speed of spread.
- **Timing Constraints:** the advertisers expect the influence to occur within a certain time window, and different products may have different timing requirements.
- **User Constraints:** users of the social network, each of which can be a potential source, would like to be exposed to only a small number of ads. Furthermore, users may be grouped by their geographical locations, and advertisers may have a target population they want to reach.
- **Product Constraints:** seeking initial adopters entails a cost to the advertiser, who needs to pay to the host and often has a limited amount of money.

For example, Facebook (*i.e.*, the host) needs to allocate ads for various products with different characteristics, *e.g.*, clothes, books, or cosmetics. While some products, such as clothes, aim at influencing within a short time window, some others, such as books, may allow for longer periods. Moreover, Facebook limits the number of ads in each user’s sidebar (typically it shows less than five) and, as a consequence, it cannot assign all ads to a few highly influential users. Finally, each advertiser has a limited budget to pay for ads on Facebook and thus each ad can only be displayed to some subset of users. In our work, we incorporate these myriads of practical and important requirements into consideration in the influence maximization problem.

We account for the multi-product and timing constraints by applying product-specific continuous-time diffusion models. Here, we opt for continuous-time diffusion models instead of discrete-time models, which have been mostly used in previous work (Kempe et al., 2003; Chen et al., 2009, 2010a,b, 2011, 2012; Borgs et al., 2012). This is because artificially discretizing the time axis into bins introduces additional errors. One can adjust the additional tuning parameters, like the bin size, to balance the tradeoff between the error and the computational cost, but the parameters are not easy to choose optimally. Extensive experimental comparisons on both synthetic and real-world data have shown that discrete-time models provide less accurate influence estimation than their continuous-time counterparts (Gomez-Rodriguez et al., 2011; Gomez-Rodriguez and Schölkopf, 2012; Gomez-Rodriguez et al., 2013; Du et al., 2013a,b).

However, maximizing influence based on continuous-time diffusion models also entails additional challenges. First, evaluating the objective function of the influence maximization

problem (*i.e.*, the influence estimation problem) in this setting is a difficult graphical model inference problem, *i.e.*, computing the marginal density of continuous variables in loopy graphical models. The exact answer can be computed only for very special cases. For example, Gomez-Rodriguez and Schölkopf (2012) have shown that the problem can be solved exactly when the transmission functions are exponential densities, by using continuous time Markov processes theory. However, the computational complexity of such approach, in general, scales exponentially with the size and density of the network. Moreover, extending the approach to deal with arbitrary transmission functions would require additional nontrivial approximations which would increase even more the computational complexity. Second, it is unclear how to scale up influence estimation and maximization algorithms based on continuous-time diffusion models to millions of nodes. Especially in the maximization case, the influence estimation procedure needs to be called many times for different subsets of selected nodes. *Thus, our first goal is to design a scalable algorithm which can perform influence estimation in the regime of networks with millions of nodes.*

We account for the user and product constraints by restricting the feasible domain over which the maximization is performed. We first show that the overall influence function of multiple products is a submodular function and then realize that the user and product constraints correspond to constraints over the ground set of this submodular function. To the best of our knowledge, previous work has not considered both user and product constraints simultaneously over general unknown different diffusion networks with non-uniform costs. In particular, (Datta et al., 2010) first tried to model both the product and user constraints only with uniform costs and infinite time window, which essentially reduces to a special case of our formulations. Similarly, (Lu et al., 2013) considered the allocation problem of multiple products which may have competitions within the infinite time window. Besides, they all assume that multiple products spread within the same network. In contrast, our formulations generally allow products to have different diffusion networks, which can be unknown in practice. Soma et al. (2014) studied the influence maximization problem for one product subject to one knapsack constraint over a known bipartite graph between marketing channels and potential customers; Ienco et al. (2010) and Sun et al. (2011) considered user constraints but disregarded product constraints during the initial assignment; and, Narayanam and Nanavati (2012) studied the cross-sell phenomenon (the selling of the first product raises the chance of selling the second) and included monetary constraints for all the products. However, no user constraints were considered, and the cost of each user was still uniform for each product. *Thus, our second goal is to design an efficient submodular maximization algorithm which can take into account both user and product constraints simultaneously.*

Overall, this article includes the following major contributions:

- Unlike prior work that considers an a priori described simplistic discrete-time diffusion model, we first **learn** the diffusion networks from data by using continuous-time diffusion models. This allows us to address the timing constraints in a principled way.
- We provide a novel formulation of the influence estimation problem in the continuous-time diffusion model from the perspective of probabilistic graphical models, which allows heterogeneous diffusion dynamics over the edges.

- We propose an efficient randomized algorithm for continuous-time influence estimation, which can scale up to millions of nodes and estimate the influence of each node to an accuracy of ϵ using $n = \mathcal{O}(1/\epsilon^2)$ randomizations.
- We formulate the influence maximization problem with the aforementioned constraints as a submodular maximization under the intersection of matroid constraints and knapsack constraints. The submodular function we use is based on the actual diffusion model learned from the data for the time window constraint. This novel formulation provides us a firm theoretical foundation for designing greedy algorithms with theoretical guarantees.
- We develop an efficient adaptive-threshold greedy algorithm which is linear in the number of products and proportional to $\tilde{\mathcal{O}}(|\mathcal{V}| + |\mathcal{E}^*|)$, where $|\mathcal{V}|$ is the number of nodes (users) and $|\mathcal{E}^*|$ is the number of edges in the largest diffusion network. We then prove that this algorithm is guaranteed to find a solution with an overall influence of at least $\frac{k_a}{2+2k}$ of the optimal value, when k_a out of the k knapsack constraints are active. This improves over the best known approximation factor achieved by polynomial time algorithms in the combinatorial optimization literature. Moreover, whenever advertising each product to each user entails the same cost, the constraints reduce to an intersection of matroids, and we obtain an approximation factor of $1/3$, which is optimal for such optimization.
- We evaluate our algorithms over large synthetic and real-world datasets and show that our proposed methods significantly improve over previous state-of-the-arts in terms of both the accuracy of the estimated influence and the quality of the selected nodes in maximizing the influence over independently hold-out real testing data.

In the remainder of the paper, we will first tackle the influence estimation problem in section 2. We then formulate different realistic constraints for the influence maximization in section 3 and present the adaptive-thresholding greedy algorithm with its theoretical analysis in section 4; we investigate the performance of the proposed algorithms in both synthetic and real-world datasets in section 5; and finally we conclude in section 6.

2. Influence Estimation

We start by revisiting the continuous-time diffusion model by Gomez-Rodriguez et al. (2011) and then explicitly formulate the influence estimation problem from the perspective of probabilistic graphical models. Because the efficient inference of the influence value for each node is highly non-trivial, we further develop a scalable influence estimation algorithm which is able to handle networks of millions of nodes. The influence estimation procedure will be a key building block for our later influence maximization algorithm.

2.1 Continuous-Time Diffusion Networks

The continuous-time diffusion model associates each edge with a transmission function, that is, a density over the transmission time along the edge, in contrast to previous discrete-time models which associate each edge with a fixed infection probability (Kempe et al., 2003).

Moreover, it also differs from discrete-time models in the sense that events in a cascade are not generated iteratively in rounds, but event timings are sampled directly from the transmission function in the continuous-time model.

Continuous-Time Independent Cascade Model. Given a *directed* contact network, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, we use the independent cascade model for modeling a diffusion process (Kempe et al., 2003; Gomez-Rodriguez et al., 2011). The process begins with a set of infected source nodes, \mathcal{A} , initially adopting certain contagion (idea, meme or product) at time zero. The contagion is transmitted from the sources along their out-going edges to their direct neighbors. Each transmission through an edge entails *random* waiting times, τ , drawn from different independent pairwise waiting time distributions (one per edge). Then, the infected neighbors transmit the contagion to their respective neighbors, and the process continues. We assume that an infected node remains infected for the entire diffusion process. Thus, if a node i is infected by multiple neighbors, only the neighbor that first infects node i will be the *true* parent. As a result, although the contact network can be an arbitrary directed network, each diffusion process induces a Directed Acyclic Graph (DAG).

Heterogeneous Transmission Functions. Formally, the pairwise transmission function $f_{ji}(t_i|t_j)$ for a directed edge $j \rightarrow i$ is the conditional density of node i getting infected at time t_i given that node j was infected at time t_j . We assume it is shift invariant: $f_{ji}(t_i|t_j) = f_{ji}(\tau_{ji})$, where $\tau_{ji} := t_i - t_j$, and causal: $f_{ji}(\tau_{ji}) = 0$ if $\tau_{ji} < 0$. Both parametric transmission functions, such as the exponential and Rayleigh function (Gomez-Rodriguez et al., 2011), and nonparametric functions (Du et al., 2012) can be used and estimated from cascade data.

Shortest-Path Property. The independent cascade model has a useful property we will use later: given a sample of transmission times of all edges, the time t_i taken to infect a node i is the length of the shortest path in \mathcal{G} from the sources to node i , where the edge weights correspond to the associated transmission times.

2.2 Probabilistic Graphical Model for Continuous-Time Diffusion Networks

The continuous-time independent cascade model is essentially a directed graphical model for a set of *dependent* random variables, that is, the infection times t_i of the nodes, where the conditional independence structure is supported on the contact network \mathcal{G} . Although the original contact graph \mathcal{G} can contain directed loops, each diffusion process (or a cascade) induces a directed acyclic graph (DAG). For those cascades consistent with a particular DAG, we can model the joint density of t_i using a directed graphical model:

$$p(\{t_i\}_{i \in \mathcal{V}}) = \prod_{i \in \mathcal{V}} p(t_i | \{t_j\}_{j \in \pi_i}), \quad (1)$$

where each π_i denotes the collection of parents of node i in the induced DAG, and each term $p(t_i | \{t_j\}_{j \in \pi_i})$ corresponds to a conditional density of t_i given the infection times of the parents of node i . This is true because given the infection times of node i 's parents, t_i is independent of other infection times, satisfying the local Markov property of a directed graphical model. We note that the independent cascade model only specifies explicitly the pairwise transmission function of each directed edge, but does not directly define the conditional density $p(t_i | \{t_j\}_{j \in \pi_i})$.

However, these conditional densities can be derived from the pairwise transmission functions based on the Independent-Infection property:

$$p(t_i | \{t_j\}_{j \in \pi_i}) = \sum_{j \in \pi_i} f_{ji}(t_i | t_j) \prod_{l \in \pi_i, l \neq j} S(t_i | t_l), \quad (2)$$

which is the sum of the likelihoods that node i is infected by each parent node j . More precisely, each term in the summation can be interpreted as the likelihood $f_{ji}(t_i | t_j)$ of node i being infected at t_i by node j multiplied by the probability $S(t_i | t_l)$ that it has survived from the infection of each other parent node $l \neq j$ until time t_i .

Perhaps surprisingly, the factorization in Equation (1) is the same factorization that can be used for an arbitrary induced DAG consistent with the contact network \mathcal{G} . In this case, we only need to replace the definition of π_i (the parent of node i in the DAG) to the set of neighbors of node i with an edge pointing to node i in \mathcal{G} . This is not immediately obvious from Equation (1), since the contact network \mathcal{G} can contain directed loops which seems to be in conflict with the conditional independence semantics of directed graphical models. The reason why it is possible to do so is as follows: any fixed set of infection times, t_1, \dots, t_d , induces an ordering of the infection times. If $t_i \leq t_j$ for an edge $j \rightarrow i$ in \mathcal{G} , $h_{ji}(t_i | t_j) = 0$, and the corresponding term in Equation (2) is zeroed out, making the conditional density consistent with the semantics of directed graphical models.

Instead of directly modeling the infection times t_i , we can focus on the set of mutually *independent* random transmission times $\tau_{ji} = t_i - t_j$. Interestingly, by switching from a node-centric view to an edge-centric view, we obtain a fully factorized joint density of the set of transmission times

$$p(\{\tau_{ji}\}_{(j,i) \in \mathcal{E}}) = \prod_{(j,i) \in \mathcal{E}} f_{ji}(\tau_{ji}), \quad (3)$$

Based on the Shortest-Path property of the independent cascade model, each variable t_i can be viewed as a transformation from the collection of variables $\{\tau_{ji}\}_{(j,i) \in \mathcal{E}}$. More specifically, let \mathcal{Q}_i be the collection of directed paths in \mathcal{G} from the source nodes to node i , where each path $q \in \mathcal{Q}_i$ contains a sequence of directed edges (j, l) . Assuming all source nodes are infected at time zero, then we obtain variable t_i via

$$t_i = g_i(\{\tau_{ji}\}_{(j,i) \in \mathcal{E}} | \mathcal{A}) = \min_{q \in \mathcal{Q}_i} \sum_{(j,l) \in q} \tau_{jl}, \quad (4)$$

where the transformation $g_i(\cdot | \mathcal{A})$ is the value of the shortest-path minimization. As a special case, we can now compute the probability of node i infected before T using a set of independent variables:

$$\Pr \{t_i \leq T | \mathcal{A}\} = \Pr \{g_i(\{\tau_{ji}\}_{(j,i) \in \mathcal{E}} | \mathcal{A}) \leq T\}. \quad (5)$$

The significance of the relation is that it allows us to transform a problem involving a sequence of dependent variables $\{t_i\}_{i \in \mathcal{V}}$ to one with independent variables $\{\tau_{ji}\}_{(j,i) \in \mathcal{E}}$. Furthermore, the two perspectives are connected via the shortest path algorithm in weighted directed graph, a standard well-studied operation in graph analysis.

2.3 Influence Estimation Problem in Continuous-Time Diffusion Networks

Intuitively, given a time window, the wider the spread of infection, the more influential the set of sources. We adopt the definition of influence as the expected number of infected nodes given a set of source nodes and a time window, as in previous work (Gomez-Rodriguez and Schölkopf, 2012). More formally, consider a set of source nodes $\mathcal{A} \subseteq \mathcal{V}$, $|\mathcal{A}| \leq C$ which get infected at time zero. Then, given a time window T , a node i is infected within the time window if $t_i \leq T$. The expected number of infected nodes (or the influence) given the set of transmission functions $\{f_{ji}\}_{(j,i) \in \mathcal{E}}$ can be computed as

$$\sigma(\mathcal{A}, T) = \mathbb{E} \left[\sum_{i \in \mathcal{V}} \mathbb{I}\{t_i \leq T | \mathcal{A}\} \right] = \sum_{i \in \mathcal{V}} \Pr\{t_i \leq T | \mathcal{A}\}, \quad (6)$$

where $\mathbb{I}\{\cdot\}$ is the indicator function and the expectation is taken over the the set of *dependent* variables $\{t_i\}_{i \in \mathcal{V}}$. By construction, $\sigma(\mathcal{A}, T)$ is a non-negative, monotonic nondecreasing submodular function in the set of source nodes shown by Gomez-Rodriguez and Schölkopf (2012).

Essentially, the influence estimation problem in Equation (6) is an inference problem for graphical models, where the probability of event $t_i \leq T$ given sources in \mathcal{A} can be obtained by summing out the possible configuration of other variables $\{t_j\}_{j \neq i}$. That is

$$\Pr\{t_i \leq T | \mathcal{A}\} = \int_0^\infty \cdots \int_{t_i=0}^T \cdots \int_0^\infty \left(\prod_{j \in \mathcal{V}} p(t_j | \{t_l\}_{l \in \pi_j}) \right) \left(\prod_{j \in \mathcal{V}} dt_j \right), \quad (7)$$

which is, in general, a very challenging problem. First, the corresponding directed graphical models can contain nodes with high in-degree and high out-degree. For example, in Twitter, a user can follow dozens of other users, and another user can have hundreds of “followers”. The tree-width corresponding to this directed graphical model can be very high, and we need to perform integration for functions involving many continuous variables. Second, the integral in general can not be evaluated analytically for heterogeneous transmission functions, which means that we need to resort to numerical integration by discretizing the domain $[0, \infty)$. If we use N levels of discretization for each variable, we would need to enumerate $\mathcal{O}(N^{|\pi_i|})$ entries, exponential in the number of parents.

Only in very special cases, can one derive the closed-form equation for computing $\Pr\{t_i \leq T | \mathcal{A}\}$. For instance, Gomez-Rodriguez and Schölkopf (2012) proposed an approach for exponential transmission functions, where the special properties of exponential density are used to map the problem into a continuous time Markov process problem, and the computation can be carried out via a matrix exponential. However, without further heuristic approximation, the computational complexity of the algorithm is exponential in the size and density of the network. The intrinsic complexity of the problem entails the utilization of approximation algorithms, such as mean field algorithms or message passing algorithms. We will design an efficient randomized (or sampling) algorithm in the next section.

2.4 Efficient Influence Estimation in Continuous-Time Diffusion Networks

Our first key observation is that we can transform the influence estimation problem in Equation (6) into a problem with *independent* variables. With the relation in Equation (5),

we can derive the influence function as

$$\begin{aligned}\sigma(\mathcal{A}, T) &= \sum_{i \in \mathcal{V}} \Pr \{g_i(\{\tau_{ji}\}_{(j,i) \in \mathcal{E}} | \mathcal{A}) \leq T\} \\ &= \mathbb{E} \left[\sum_{i \in \mathcal{V}} \mathbb{I} \{g_i(\{\tau_{ji}\}_{(j,i) \in \mathcal{E}} | \mathcal{A}) \leq T\} \right],\end{aligned}\tag{8}$$

where the expectation is with respect to the set of independent variables $\{\tau_{ji}\}_{(j,i) \in \mathcal{E}}$. This equivalent formulation suggests a naive sampling (NS) algorithm for approximating $\sigma(\mathcal{A}, T)$: draw n samples of $\{\tau_{ji}\}_{(j,i) \in \mathcal{E}}$, run a shortest path algorithm for each sample, and finally average the results (see Appendix A for more details). However, this naive sampling approach has a computational complexity of $O(nC|\mathcal{V}||\mathcal{E}| + nC|\mathcal{V}|^2 \log |\mathcal{V}|)$ due to the repeated calling of the shortest path algorithm. This is quadratic to the network size, and hence not scalable to millions of nodes.

Our second key observation is that for each sample $\{\tau_{ji}\}_{(j,i) \in \mathcal{E}}$, we are only interested in the neighborhood size of the source nodes, *i.e.*, the summation $\sum_{i \in \mathcal{V}} \mathbb{I} \{\cdot\}$ in Equation (8), rather than in the individual shortest paths. Fortunately, the neighborhood size estimation problem has been studied in the theoretical computer science literature. Here, we adapt a very efficient randomized algorithm by Cohen (1997) to our influence estimation problem. This randomized algorithm has a computational complexity of $O(|\mathcal{E}| \log |\mathcal{V}| + |\mathcal{V}| \log^2 |\mathcal{V}|)$ and it estimates the neighborhood sizes for *all* possible single source node locations. Since it needs to run once for each sample of $\{\tau_{ji}\}_{(j,i) \in \mathcal{E}}$, we obtain an overall influence estimation algorithm with $O(n|\mathcal{E}| \log |\mathcal{V}| + n|\mathcal{V}| \log^2 |\mathcal{V}|)$ computation, nearly linear in network size. Next we will revisit Cohen's algorithm for neighborhood estimation.

2.4.1 RANDOMIZED ALGORITHM FOR SINGLE-SOURCE NEIGHBORHOOD-SIZE ESTIMATION

Given a fixed set of edge transmission times $\{\tau_{ji}\}_{(j,i) \in \mathcal{E}}$ and a source node s , infected at time zero, the neighborhood $\mathcal{N}(s, T)$ of a source node s given a time window T is the set of nodes within distance T from s , *i.e.*,

$$\mathcal{N}(s, T) = \{i \mid g_i(\{\tau_{ji}\}_{(j,i) \in \mathcal{E}}) \leq T, i \in \mathcal{V}\}.\tag{9}$$

Instead of estimating $\mathcal{N}(s, T)$ directly, the algorithm will assign an exponentially distributed random label r_i to each network node i . Then, it makes use of the fact that the minimum of a set of exponential random variables $\{r_i\}_{i \in \mathcal{N}(s, T)}$ is still an exponential random variable, but with its parameter being equal to the total number of variables, that is, if each $r_i \sim \exp(-r_i)$, then the smallest label within distance T from source s , $r_* := \min_{i \in \mathcal{N}(s, T)} r_i$, will distribute as $r_* \sim \exp\{-|\mathcal{N}(s, T)|r_*\}$. Suppose we randomize over the labeling m times and obtain m such least labels, $\{r_*^u\}_{u=1}^m$. Then the neighborhood size can be estimated as

$$|\mathcal{N}(s, T)| \approx \frac{m-1}{\sum_{u=1}^m r_*^u}.\tag{10}$$

which is shown by Cohen (1997) to be an unbiased estimator of $|\mathcal{N}(s, T)|$. This is an elegant relation since it allows us to transform the counting problem in (9) to a problem of finding the minimum random label r_* . The key question is whether we can compute the least label r_* efficiently, given random labels $\{r_i\}_{i \in \mathcal{V}}$ and any source node s .

Cohen (1997) designed a modified Dijkstra’s algorithm (Algorithm 3) to construct a data structure $r_*(s)$, called least label list, for each node s to support such query. Essentially, the algorithm starts with the node i with the smallest label r_i , and then it traverses in breadth-first search fashion along the reverse direction of the graph edges to find all reachable nodes. For each reachable node s , the distance d_* between i and s , and r_i are added to the end of $r_*(s)$. Then the algorithm moves to the node i' with the second smallest label $r_{i'}$, and similarly find all reachable nodes. For each reachable node s , the algorithm will compare the current distance d_* between i' and s with the last recorded distance in $r_*(s)$. If the current distance is smaller, then the current d_* and $r_{i'}$ are added to the end of $r_*(s)$. Then the algorithm move to the node with the third smallest label and so on. The algorithm is summarized in Algorithm 3 in Appendix B.

Algorithm 3 returns a list $r_*(s)$ per node $s \in \mathcal{V}$, which contains information about distance to the smallest reachable labels from s . In particular, each list contains pairs of distance and random labels, (d, r) , and these pairs are ordered as

$$\infty > d_{(1)} > d_{(2)} > \dots > d_{(|r_*(s)|)} = 0 \quad (11)$$

$$r_{(1)} < r_{(2)} < \dots < r_{(|r_*(s)|)}, \quad (12)$$

where $\{\cdot\}_{(l)}$ denotes the l -th element in the list. (see Appendix B for an example).

If we want to query the smallest reachable random label r_* for a given source s and a time T , we only need to perform a binary search on the list for node s :

$$r_* = r_{(l)}, \text{ where } d_{(l-1)} > T \geq d_{(l)}. \quad (13)$$

Finally, to estimate $|\mathcal{N}(s, T)|$, we generate m *i.i.d.* collections of random labels, run Algorithm 3 on each collection, and obtain m values $\{r_*^u\}_{u=1}^m$, which we use in Equation (10) to estimate $|\mathcal{N}(i, T)|$.

The computational complexity of Algorithm 3 is $O(|\mathcal{E}| \log |\mathcal{V}| + |\mathcal{V}| \log^2 |\mathcal{V}|)$, with expected size of each $r_*(s)$ being $O(\log |\mathcal{V}|)$. Then the expected time for querying r_* is $O(\log \log |\mathcal{V}|)$ using binary search. Since we need to generate m set of random labels and run Algorithm 3 m times, the overall computational complexity for estimating the single-source neighborhood size for all $s \in \mathcal{V}$ is $O(m|\mathcal{E}| \log |\mathcal{V}| + m|\mathcal{V}| \log^2 |\mathcal{V}| + m|\mathcal{V}| \log \log |\mathcal{V}|)$. For large-scale network, and when $m \ll \min\{|\mathcal{V}|, |\mathcal{E}|\}$, this randomized algorithm can be much more efficient than approaches based on directly calculating the shortest paths.

2.4.2 CONSTRUCTING ESTIMATION FOR MULTIPLE-SOURCE NEIGHBORHOOD SIZE

When we have a set of sources, \mathcal{A} , its neighborhood is the union of the neighborhoods of its constituent sources

$$\mathcal{N}(\mathcal{A}, T) = \bigcup_{i \in \mathcal{A}} \mathcal{N}(i, T). \quad (14)$$

This is true because each source independently infects its downstream nodes. Furthermore, to calculate the least label list r_* corresponding to $\mathcal{N}(\mathcal{A}, T)$, we can simply reuse the least label list $r_*(i)$ of each individual source $i \in \mathcal{A}$. More formally,

$$r_* = \min_{i \in \mathcal{A}} \min_{j \in \mathcal{N}(i, T)} r_j, \quad (15)$$

where the inner minimization can be carried out by querying $r_*(i)$. Similarly, after we obtain m samples of r_* , we can estimate $|\mathcal{N}(\mathcal{A}, T)|$ using Equation (10). Importantly, very little additional work is needed when we want to calculate r_* for a set of sources \mathcal{A} , and we can reuse work done for a single source. This is very different from a naive sampling approach where the sampling process needs to be done completely anew if we increase the source set. In contrast, using the randomized algorithm, only an additional constant-time minimization over $|\mathcal{A}|$ numbers is needed.

2.4.3 OVERALL ALGORITHM

So far, we have achieved efficient neighborhood size estimation of $|\mathcal{N}(\mathcal{A}, T)|$ with respect to a given set of transmission times $\{\tau_{ji}\}_{(j,i) \in \mathcal{E}}$. Next, we will estimate the influence by averaging over multiple sets of samples for $\{\tau_{ji}\}_{(j,i) \in \mathcal{E}}$. More specifically, the relation from (8)

$$\sigma(\mathcal{A}, T) = \mathbb{E}_{\{\tau_{ji}\}_{(j,i) \in \mathcal{E}}} [|\mathcal{N}(\mathcal{A}, T)|] = \mathbb{E}_{\{\tau_{ji}\}} \mathbb{E}_{\{r^1, \dots, r^m\} | \{\tau_{ji}\}} \left[\frac{m-1}{\sum_{u=1}^m r_*^u} \right], \quad (16)$$

suggests the following overall algorithm :

Continuous-Time Influence Estimation (CONTINEST):

1. Sample n sets of random transmission times $\{\tau_{ij}^l\}_{(j,i) \in \mathcal{E}} \sim \prod_{(j,i) \in \mathcal{E}} f_{ji}(\tau_{ji})$.
2. Given a set of $\{\tau_{ij}^l\}_{(j,i) \in \mathcal{E}}$, sample m sets of random labels $\{r_i^u\}_{i \in \mathcal{V}} \sim \prod_{i \in \mathcal{V}} \exp(-r_i)$.
3. Estimate $\sigma(\mathcal{A}, T)$ by sample averages $\sigma(\mathcal{A}, T) \approx \frac{1}{n} \sum_{l=1}^n \left((m-1) / \sum_{u=1}^m r_*^{u_l} \right)$.

What is even more important is that the number of random labels, m , does not need to be very large. Since the estimator for $|\mathcal{N}(\mathcal{A}, T)|$ is unbiased (Cohen, 1997), essentially the outer-loop of averaging over n samples of random transmission times further reduces the variance of the estimator in a rate of $\mathcal{O}(1/n)$. In practice, we can use a very small m (e.g., 5 or 10) and still achieve good results, which is also confirmed by our later experiments. Compared to (Chen et al., 2009), the novel application of Cohen’s algorithm arises for estimating influence for multiple sources, which drastically reduces the computation by cleverly using the least-label list from single source. Moreover, we have the following theoretical guarantee (see Appendix C for the proof).

Theorem 1 *Draw the following number of samples for the set of random transmission times*

$$n \geq \frac{C\Lambda}{\epsilon^2} \log \left(\frac{2|\mathcal{V}|}{\alpha} \right) \quad (17)$$

where $\Lambda := \max_{\mathcal{A}: |\mathcal{A}| \leq C} 2\sigma(\mathcal{A}, T)^2 / (m-2) + 2\text{Var}(|\mathcal{N}(\mathcal{A}, T)|)(m-1)/(m-2) + 2a\epsilon/3$ and $|\mathcal{N}(\mathcal{A}, T)| \leq |\mathcal{V}|$, and for each set of random transmission times, draw m sets of random labels. Then $|\hat{\sigma}(\mathcal{A}, T) - \sigma(\mathcal{A}, T)| \leq \epsilon$ uniformly for all \mathcal{A} with $|\mathcal{A}| \leq C$, with probability at least $1 - \alpha$.

The theorem indicates that the minimum number of samples, n , needed to achieve certain accuracy is related to the actual size of the influence $\sigma(\mathcal{A}, T)$, and the variance of the neighborhood size $|\mathcal{N}(\mathcal{A}, T)|$ over the random draw of samples. The number of random labels, m , drawn in the inner loop of the algorithm will monotonically decrease the dependency of n on $\sigma(\mathcal{A}, T)$. It suffices to draw a small number of random labels, as long as the value of $\sigma(\mathcal{A}, T)^2/(m-2)$ matches that of $\text{Var}(|\mathcal{N}(\mathcal{A}, T)|)$. Another implication is that influence at larger time window T is harder to estimate, since $\sigma(\mathcal{A}, T)$ will generally be larger and hence require more random samples.

3. Constraints of Practical Importance

By treating our proposed influence estimation algorithm CONTINEST as a building block, we can now tackle the influence maximization problem under various constraints of practical importance. Here, since CONTINEST can estimate the influence value of any source set with respect to any given time window T , the **Timing Constraints** can thus be naturally satisfied. Therefore, in the following sections, we mainly focus on modeling the **Multiple-Item Constraints**, the **User Constraints** and the **Product Constraints**.

3.1 Multiple-Item Constraints

Multiple products can spread simultaneously across the same set of social entities over different diffusion channels. Since these products may have different characteristics, such as the revenue and the speed of spread, and thus may follow different diffusion dynamics, we will use multiple diffusion networks for different types of products.

Suppose we have a set of products \mathcal{L} that propagate on the same set of nodes \mathcal{V} . The diffusion network for product i is denoted as $\mathcal{G}_i = (\mathcal{V}, \mathcal{E}_i)$. For each product $i \in \mathcal{L}$, we search for a set of source nodes $\mathcal{R}_i \subseteq \mathcal{V}$ to which we can assign the product i to start its campaign. We can represent the selection of \mathcal{R}_i 's using an assignment matrix $A \in \{0, 1\}^{|\mathcal{L}| \times |\mathcal{V}|}$ as follows: $A_{ij} = 1$ if $j \in \mathcal{R}_i$ and $A_{ij} = 0$ otherwise. Based on this representation, we define a new ground set $\mathcal{Z} = \mathcal{L} \times \mathcal{V}$ of size $N = |\mathcal{L}| \times |\mathcal{V}|$. Each element of \mathcal{Z} corresponds to the index (i, j) of an entry in the assignment matrix A , and selecting element $z = (i, j)$ means assigning product i to user j (see Figure 1 for an illustration). We also denote $\mathcal{Z}_{*j} = \mathcal{L} \times \{j\}$ and $\mathcal{Z}_{i*} = \{i\} \times \mathcal{V}$ as the j -th column and i -th row of matrix A , respectively. Then, under the above mentioned additional requirements, we would like to find a set of assignments $S \subseteq \mathcal{Z}$ so as to maximize the following **overall influence**

$$f(S) = \sum_{i \in \mathcal{L}} a_i \sigma_i(\mathcal{R}_i, T_i), \quad (18)$$

where $\sigma_i(\mathcal{R}_i, T_i)$ denote the influence of product i for a given time T_i , $\{a_i > 0\}$ is a set of weights reflecting the different benefits of the products and $\mathcal{R}_i = \{j \in \mathcal{V} : (i, j) \in S\}$. We now show that the overall influence function $f(S)$ in Equation (18) is submodular over the ground set \mathcal{Z} .

Lemma 2 *Under the continuous-time independent cascade model, the overall influence $f(S)$ is a normalized monotone submodular function of S .*

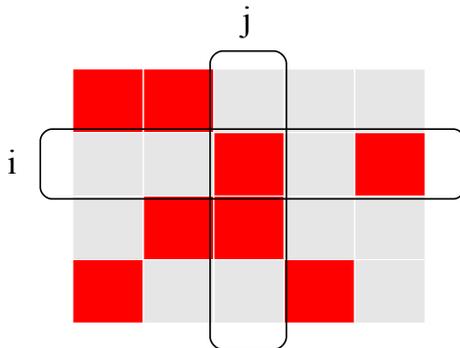


Figure 1: Illustration of the assignment matrix A associated with partition matroid \mathcal{M}_1 and group knapsack constraints. If product i is assigned to user j , then $A_{ij} = 1$ (colored in red). The ground set \mathcal{Z} is the set of indices of the entries in A , and selecting an element $(i, j) \in \mathcal{Z}$ means assigning product i to user j . The user constraint means that there are at most u_j elements selected in the j -th column; the product constraint means that the total cost of the elements selected in the i -th row is at most B_i .

Proof By definition, $f(\emptyset) = 0$ and $f(S)$ is monotone. By Theorem 4 in Gomez-Rodriguez and Schölkopf (2012), the component influence function $\sigma_i(\mathcal{R}_i, T_i)$ for product i is submodular in $\mathcal{R}_i \subseteq \mathcal{V}$. Since non-negative linear combinations of submodular functions are still submodular, $f_i(S) := a_i \sigma_i(\mathcal{R}_i, T_i)$ is also submodular in $S \subseteq \mathcal{Z} = \mathcal{L} \times \mathcal{V}$, and $f(S) = \sum_{i \in \mathcal{L}} f_i(S)$ is submodular. ■

3.2 User Constraints

Each social network user can be a potential source and would like to be exposed only to a small number of ads. Furthermore, users may be grouped according to their geographical locations, and advertisers may have a target population they want to reach. Here, we will incorporate these constraints using the matroids which are combinatorial structures that generalize the notion of linear independence in matrices (Schrijver, 2003; Fujishige, 2005). Formulating our constrained influence maximization task with matroids allows us to design a greedy algorithm with provable guarantees.

Formally, suppose that each user j can be assigned to at most u_j products. A matroid can be defined as follows:

Definition 3 A matroid is a pair, $\mathcal{M} = (\mathcal{Z}, \mathcal{I})$, defined over a finite set (the ground set) \mathcal{Z} and a family of sets (the independent sets) \mathcal{I} , that satisfies three axioms:

1. *Non-emptiness:* The empty set $\emptyset \in \mathcal{I}$.
2. *Heredity:* If $Y \in \mathcal{I}$ and $X \subseteq Y$, then $X \in \mathcal{I}$.
3. *Exchange:* If $X \in \mathcal{I}, Y \in \mathcal{I}$ and $|Y| > |X|$, then there exists $z \in Y \setminus X$ such that $X \cup \{z\} \in \mathcal{I}$.

An important type of matroid is the partition matroid where the ground set \mathcal{Z} is partitioned into disjoint subsets $\mathcal{Z}_1, \mathcal{Z}_2, \dots, \mathcal{Z}_t$ for some t and

$$\mathcal{I} = \{S \mid S \subseteq \mathcal{Z} \text{ and } |S \cap \mathcal{Z}_i| \leq u_i, \forall i = 1, \dots, t\}$$

for some given parameters u_1, \dots, u_t . The user constraints can then be formulated as

Partition matroid \mathcal{M}_1 : partition the ground set \mathcal{Z} into $\mathcal{Z}_{*j} = \mathcal{L} \times \{j\}$ each of which corresponds to a column of A . Then $\mathcal{M}_1 = \{\mathcal{Z}, \mathcal{I}_1\}$ is

$$\mathcal{I}_1 = \{S \mid S \subseteq \mathcal{Z} \text{ and } |S \cap \mathcal{Z}_{*j}| \leq u_j, \forall j\}.$$

3.3 Product Constraints

Seeking initial adopters entails a cost to the advertiser, which needs to be paid to the host, while the advertisers of each product have a limited amount of money. Here, we will incorporate these requirements using knapsack constraints which we describe below.

Formally, suppose that each product i has a budget B_i , and assigning item i to user j costs $c_{ij} > 0$. Next, we introduce the following notation to describe product constraints over the ground set \mathcal{Z} . For an element $z = (i, j) \in \mathcal{Z}$, define its cost as $c(z) := c_{ij}$. Abusing the notation slightly, we denote the cost of a subset $S \subseteq \mathcal{Z}$ as $c(S) := \sum_{z \in S} c(z)$. Then, in a feasible solution $S \subseteq \mathcal{Z}$, the cost of assigning product i , $c(S \cap \mathcal{Z}_{i*})$, should not be larger than its budget B_i .

Now, without loss of generality, we can assume $B_i = 1$ (by normalizing c_{ij} with B_i), and also $c_{ij} \in (0, 1]$ (by throwing away any element (i, j) with $c_{ij} > 1$), and define

Group-knapsack: partition the ground set into $\mathcal{Z}_{i*} = \{i\} \times \mathcal{V}$ each of which corresponds to one row of A . Then a feasible solution $S \subseteq \mathcal{Z}$ satisfies

$$c(S \cap \mathcal{Z}_{i*}) \leq 1, \forall i.$$

Importantly, these knapsack constraints have very specific structure: they are on different groups of a partition $\{\mathcal{Z}_{i*}\}$ of the ground set and the submodular function $f(S) = \sum_i a_i \sigma_i(\mathcal{R}_i, T_i)$ is defined over the partition. In consequence, such structures allow us to design an efficient algorithm with improved guarantees over the known results.

3.4 Overall Problem Formulation

Based on the above discussion of various constraints in viral marketing and our design choices for tackling them, we can think of the influence maximization problem as a special case of the following constrained submodular maximization problem with $P = 1$ matroid constraints and $k = |\mathcal{L}|$ knapsack constraints,

$$\begin{aligned} & \max_{S \subseteq \mathcal{Z}} f(S) & (19) \\ & \text{subject to } c(S \cap \mathcal{Z}_{i*}) \leq 1, \quad 1 \leq i \leq k, \\ & S \in \bigcap_{p=1}^P \mathcal{I}_p, \end{aligned}$$

where, for simplicity, we will denote all the feasible solutions $S \subseteq \mathcal{Z}$ as \mathcal{F} . This formulation in general includes the following cases of practical importance :

Uniform User-Cost. An important case of influence maximization, which we denote as the Uniform Cost, is that for each product i , all users have the same cost c_{i*} , *i.e.*, $c_{ij} = c_{i*}$. Equivalently, each product i can be assigned to at most $b_i := \lfloor B_i/c_{i*} \rfloor$ users. Then the product constraints are simplified to

Partition matroid \mathcal{M}_2 : for the product constraints with uniform cost, define a matroid $\mathcal{M}_2 = \{\mathcal{Z}, \mathcal{I}_2\}$ where

$$\mathcal{I}_2 = \{S | S \subseteq \mathcal{Z} \text{ and } |S \cap \mathcal{Z}_{i*}| \leq b_i, \forall i\}.$$

In this case, the influence maximization problem defined by Equation (19) becomes the problem with $P = 2$ matroid constraints and no knapsack constraints ($k = 0$). In addition, if we assume only one product needs campaign, the formulation of Equation (19) further reduces to the classic influence maximization problem with the simple cardinality constraint.

User Group Constraint. Our formulation in Equation (19) essentially allows for general matroids which can model more sophisticated real-world constraints, and the proposed formulation, algorithms, and analysis can still hold. For instance, suppose there is a hierarchical community structure on the users, *i.e.*, a tree \mathcal{T} where leaves are the users and the internal nodes are communities consisting of all users underneath, such as customers in different countries around the world. In consequence of marketing strategies, on each community $C \in \mathcal{T}$, there are at most u_C slots for assigning the products. Such constraints are readily modeled by the Laminar Matroid, which generalizes the partition matroid by allowing the set $\{\mathcal{Z}_i\}$ to be a laminar family (*i.e.*, for any $\mathcal{Z}_i \neq \mathcal{Z}_j$, either $\mathcal{Z}_i \subseteq \mathcal{Z}_j$, or $\mathcal{Z}_j \subseteq \mathcal{Z}_i$, or $\mathcal{Z}_i \cap \mathcal{Z}_j = \emptyset$). It can be shown that the community constraints can be captured by the matroid $\mathcal{M} = (\mathcal{Z}, \mathcal{I})$ where $\mathcal{I} = \{S \subseteq \mathcal{Z} : |S \cap C| \leq u_C, \forall C \in \mathcal{T}\}$. In the next section, we first present our algorithm, then provide the analysis for the uniform cost case and finally leverage such analysis for the general case.

4. Influence Maximization

In this section, we first develop a simple, practical and intuitive adaptive-thresholding greedy algorithm to solve the continuous-time influence maximization problem with the aforementioned constraints. Then, we provide a detailed theoretical analysis of its performance.

4.1 Overall Algorithm

There exist algorithms for submodular maximization under multiple knapsack constraints achieving a $1 - \frac{1}{e}$ approximation factor by (Sviridenko, 2004). Thus, one may be tempted to convert the matroid constraint in the problem defined by Equation (19) to $|\mathcal{V}|$ knapsack constraints, so that the problem becomes a submodular maximization problem under $|\mathcal{L}| + |\mathcal{V}|$ knapsack constraints. However, this naive approach is not practical for large-scale scenarios because the running time of such algorithms is exponential in the number of knapsack constraints. Instead, if we opt for algorithms for submodular maximization under k knapsack constraints and P matroids constraints, the best approximation factor achieved by polynomial time algorithms is $\frac{1}{P+2k+1}$ (Badanidiyuru and Vondrák, 2014). However,

Algorithm 1: Density Threshold Enumeration

Input: parameter δ ; objective f or its approximation \widehat{f} ; assignment cost $c(z), z \in \mathcal{Z}$
1 Set $d = \max \{f(\{z\}) : z \in \mathcal{Z}\}$;
2 **for** $\rho \in \left\{ \frac{2d}{P+2k+1}, (1+\delta)\frac{2d}{P+2k+1}, \dots, \frac{2|\mathcal{Z}|d}{P+2k+1} \right\}$ **do**
3 Call Algorithm 2 to get S_ρ ;
Output: $\operatorname{argmax}_{S_\rho} f(S_\rho)$

this is not good enough yet, since in our problem $k = |\mathcal{L}|$ can be large, though $P = 1$ is small.

Here, we will design an algorithm that achieves a better approximation factor by exploiting the following key observation about the structure of the problem defined by Equation (19): the knapsack constraints are over different groups \mathcal{Z}_{i^*} of the whole ground set, and the objective function is a sum of submodular functions over these different groups.

The details of the algorithm, called BUDGETMAX, are described in Algorithm 1. BUDGETMAX enumerates different values of a so-called density threshold ρ , runs a subroutine to find a solution for each ρ , which quantifies the cost-effectiveness of assigning a particular product to a specific user, and finally outputs the solution with the maximum objective value. Intuitively, the algorithm restricts the search space to be the set of most cost-effective allocations. The details of the subroutine to find a solution for a fixed density threshold ρ are described in Algorithm 2. Inspired by the lazy evaluation heuristic (Leskovec et al., 2007), the algorithm maintains a working set G and a marginal gain threshold w_t , which geometrically decreases by a factor of $1 + \delta$ until it is sufficiently small to be set to zero. At each w_t , the subroutine selects each new element z that satisfies the following properties:

1. It is feasible and the density ratio (the ratio between the marginal gain and the cost) is above the current density threshold;
2. Its marginal gain

$$f(z|G) := f(G \cup \{z\}) - f(G)$$

is above the current marginal gain threshold.

The term “density” comes from the knapsack problem, where the marginal gain is the mass and the cost is the volume. A large density means gaining a lot without paying much. In short, the algorithm considers only high-quality assignments and repeatedly selects feasible ones with marginal gain ranging from large to small.

Remark 1. The traditional lazy evaluation heuristic also keeps a threshold, however, it only uses the threshold to speed up selecting the element with maximum marginal gain. Instead, Algorithm 2 can add multiple elements z from the ground set at each threshold according to the descending order of their marginal gains, and thus reduces the number of rounds from the size of the solution to the number of thresholds $\mathcal{O}(\frac{1}{\delta} \log \frac{N}{\delta})$. This allows us to trade off between the runtime and the approximation ratio (refer to our theoretical guarantees in section 4.2).

Algorithm 2: Adaptive Threshold Greedy for Fixed Density

Input: parameters ρ, δ ; objective f or its approximation \hat{f} ; assignment cost $c(z), z \in \mathcal{Z}$; set of feasible solutions \mathcal{F} ; and d from Algorithm 1.

- 1 Set $d_\rho = \max \{f(\{z\}) : z \in \mathcal{Z}, f(\{z\}) \geq c(z)\rho\}$;
- 2 Set $w_t = \frac{d_\rho}{(1+\delta)^t}$ for $t = 0, \dots, L = \operatorname{argmin}_i [w_i \leq \frac{\delta d}{N}]$ and $w_{L+1} = 0$;
- 3 Set $G = \emptyset$;
- 4 **for** $t = 0, 1, \dots, L, L + 1$ **do**
- 5 **for** $z \notin G$ with $G \cup \{z\} \in \mathcal{F}$ and $f(z|G) \geq c(z)\rho$ **do**
- 6 **if** $f(z|G) \geq w_t$ **then**
- 7 Set $G \leftarrow G \cup \{z\}$;

Output: $S_\rho = G$

Remark 2. Evaluating the influence of the assigned products f is expensive. Therefore, we will use the randomized algorithm in Section 2.4.3 to compute an estimation $\hat{f}(\cdot)$ of the quantity $f(\cdot)$.

4.2 Theoretical Guarantees

Although our algorithm is quite intuitive, it is highly non-trivial to obtain the theoretical guarantees. For clarity, we first analyze the simpler case with uniform cost, which then provides the base for analyzing the general case.

4.2.1 UNIFORM COST

As shown at the end of Section 3.4, the influence maximization, in this case, corresponds to the problem defined by Equation (19) with $P = 2$ and no knapsack constraints. Thus, we can simply run Algorithm 2 with $\rho = 0$ to obtain a solution G , which is roughly a $\frac{1}{P+1}$ -approximation.

Intuition. The algorithm greedily selects feasible elements with sufficiently large marginal gain. However, it is unclear whether our algorithm will find *good* solutions and whether it will be *robust* to noise. Regarding the former, one might wonder whether the algorithm will select just a few elements while many elements in the optimal solution O will become infeasible and will not be selected, in which case the greedy solution G is a poor approximation. Regarding the latter, we only use the estimation \hat{f} of the influence f (i.e., $|\hat{f}(S) - f(S)| \leq \epsilon$ for any $S \subseteq \mathcal{Z}$), which introduces additional error to the function value. A crucial question, which has not been addressed before (Badanidiyuru and Vondrák, 2014), is whether the adaptive threshold greedy algorithm is robust to such perturbations.

Fortunately, it turns out that the algorithm will provably select sufficiently many elements of high quality. First, the elements selected in the optimal solution O but not selected in G can be partitioned into $|G|$ groups, each of which is associated with an element in G . Thus, the number of elements in the groups associated with the first t elements in G , by the property of the intersection of matroids, are bounded by Pt . See Figure 2 for an illustration. Second, the marginal gain of each element in G is at least as large as that of any

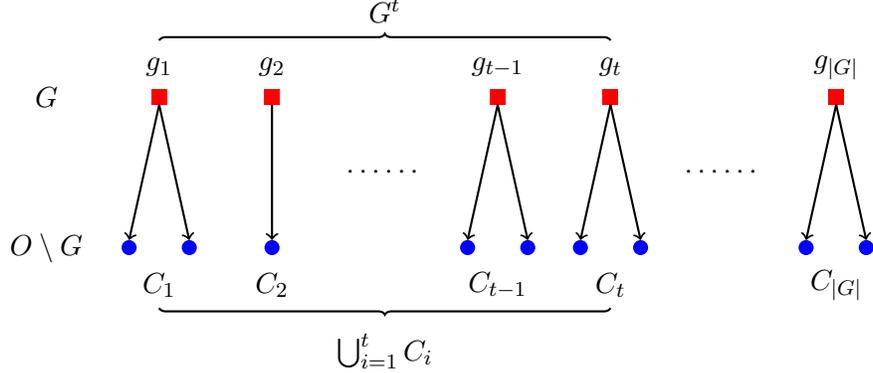


Figure 2: Notation for analyzing Algorithm 2. The elements in the greedy solution G are arranged according to the order in which Algorithm 2 selects them in Step 3. The elements in the optimal solution O but not in the greedy solution G are partitioned into groups $\{C_t\}_{1 \leq t \leq |G|}$, where C_t are those elements in $O \setminus G$ that are still feasible before selecting g_t but are infeasible after selecting g_t .

element in the group associated with it (up to some small error). This means that even if the submodular function evaluation is inexact, the quality of the elements in the greedy solution is still good. The two claims together show that the marginal gain of $O \setminus G$ is not much larger than the gain of G , and thus G is a good approximation for the problem.

Formally, suppose we use an inexact evaluation of the influence f such that $|\hat{f}(S) - f(S)| \leq \epsilon$ for any $S \subseteq \mathcal{Z}$, let product $i \in \mathcal{L}$ spread according to a diffusion network $\mathcal{G}_i = (\mathcal{V}, \mathcal{E}_i)$, and $i^* = \operatorname{argmax}_{i \in \mathcal{L}} |\mathcal{E}_i|$. Then, we have:

Theorem 4 *Suppose \hat{f} is evaluated up to error $\epsilon = \delta/16$ with CONTINEST. For influence maximization with uniform cost, Algorithm 2 (with $\rho = 0$) outputs a solution G with $f(G) \geq \frac{1-2\delta}{3} f(O)$ in expected time $\tilde{O}\left(\frac{|\mathcal{E}_{i^*}| + |\mathcal{V}|}{\delta^2} + \frac{|\mathcal{L}||\mathcal{V}|}{\delta^3}\right)$.*

The parameter δ introduces a tradeoff between the approximation guarantee and the runtime: larger δ decreases the approximation ratio but results in fewer influence evaluations. Moreover, the runtime has a linear dependence on the network size and the number of products to propagate (ignoring some small logarithmic terms) and, as a consequence, the algorithm is scalable to large networks.

Analysis. Suppose $G = \{g_1, \dots, g_{|G|}\}$ in the order of selection, and let $G^t = \{g_1, \dots, g_t\}$. Let C_t denote all those elements in $O \setminus G$ that satisfy the following: they are still feasible before selecting the t -th element g_t but are infeasible after selecting g_t . Equivalently, C_t are all those elements $j \in O \setminus G$ such that (1) $j \cup G^{t-1}$ does not violate the matroid constraints but (2) $j \cup G^t$ violates the matroid constraints. In other words, we can think of C_t as the optimal elements “blocked” by g_t . Then, we proceed as follows.

By the property of the intersection of matroids, the size of the prefix $\bigcup_{i=1}^t C_i$ is bounded by Pt . As a consequence of this property, for any $Q \subseteq \mathcal{Z}$, the sizes of any two maximal independent subsets T_1 and T_2 of Q can only differ by a multiplicative factor at most P . This can be realized with the following argument. First, note that any element $z \in T_1 \setminus T_2$,

$\{z\} \cup T_2$ violates at least one of the matroid constraints since T_2 is maximal. Then, let $\{V_i\}_{1 \leq i \leq P}$ denote all elements in $T_1 \setminus T_2$ that violate the i -th matroid, and partition $T_1 \cap T_2$ arbitrarily among these V_i 's so that they cover T_1 . In this construction, the size of each V_i must be at most $|T_2|$, since otherwise by the Exchange axiom, there would exist $z \in V_i \setminus T_2$ that can be added to T_2 , without violating the i -th matroid, leading to a contradiction. Therefore, $|T_1|$ is at most P times $|T_2|$.

Next, we apply the above property as follows. Let Q be the union of G^t and $\bigcup_{i=1}^t C_t$. On one hand, G^t is a maximal independent subset of Q , since no element in $\bigcup_{i=1}^t C_t$ can be added to G^t without violating the matroid constraints. On the other hand, $\bigcup_{i=1}^t C_t$ is an independent subset of Q , since it is part of the optimal solution. Therefore, $\bigcup_{i=1}^t C_t$ has size at most P times $|G^t|$, which is Pt . Note that the properties of matroids are crucial for this analysis, which justifies our formulation using matroids. In summary, we have

Claim 1 $\sum_{i=1}^t |C_i| \leq Pt$, for $t = 1, \dots, |G|$.

Now, we consider the marginal gain of each element in C_t associated with g_t . First, suppose g_t is selected at the threshold $\tau_t > 0$. Then, any $j \in C_t$ has marginal gain bounded by $(1 + \delta)\tau_t + 2\epsilon$, since otherwise j would have been selected at a larger threshold before τ_t by the greedy criterion. Second, suppose g_t is selected at the threshold $w_{L+1} = 0$. Then, any $j \in C_t$ has marginal gain approximately bounded by $\frac{\delta}{N}d$. Since the greedy algorithm must pick g_1 with $\widehat{f}(g_1) = d$ and $d \leq f(g_1) + \epsilon$, any $j \in C_t$ has marginal gain bounded by $\frac{\delta}{N}f(G) + \mathcal{O}(\epsilon)$. Putting everything together we have:

Claim 2 Suppose g_t is selected at the threshold τ_t . Then $f(j|G^{t-1}) \leq (1 + \delta)\tau_t + 4\epsilon + \frac{\delta}{N}f(G)$ for any $j \in C_t$.

Since the evaluation of the marginal gain of g_t should be at least τ_t , this claims essentially indicates that the marginal gain of j is approximately bounded by that of g_t .

Since there are not many elements in C_t (Claim 1) and the marginal gain of each of its elements is not much larger than that of g_t (Claim 2), we can conclude that the marginal gain of $O \setminus G = \bigcup_{i=1}^{|G|} C_t$ is not much larger than that of G , which is just $f(G)$.

Claim 3 The marginal gain of $O \setminus G$ satisfies

$$\sum_{j \in O \setminus G} f(j|G) \leq [(1 + \delta)P + \delta]f(G) + (6 + 2\delta)\epsilon P|G|.$$

Finally, since by submodularity, $f(O) \leq f(O \cup G) \leq f(G) + \sum_{j \in O \setminus G} f(j|G)$, Claim 3 shows that $f(G)$ is close to $f(O)$ up to a multiplicative factor roughly $(1 + P)$ and additive factor $\mathcal{O}(\epsilon P|G|)$. Given that $f(G) > |G|$, it leads to roughly a 1/3-approximation for our influence maximization problem by setting $\epsilon = \delta/16$ when evaluating \widehat{f} with CONTINEST. Combining the above analysis and the runtime of the influence estimation algorithm, we have our final guarantee in Theorem 4. Appendix D.1 presents the complete proofs.

4.2.2 GENERAL CASE

In this section, we consider the general case, in which users may have different associated costs. Recall that this case corresponds to the problem defined by Equation (19) with $P = 1$ matroid constraints and $k = |\mathcal{L}|$ group-knapsack constraints. Here, we will show that there is a step in Algorithm 1 which outputs a solution S_ρ that is a good approximation.

Intuition. The key idea behind Algorithm 1 and Algorithm 2 is simple: spend the budgets *efficiently* and spend them *as much as possible*. By spending them efficiently, we mean to only select those elements whose density ratio between the marginal gain and the cost is above the threshold ρ . That is, we assign product i to user j only if the assignment leads to large marginal gain without paying too much. By spending the budgets as much as possible, we mean to stop assigning product i only if its budget is almost exhausted or no more assignments are possible without violating the matroid constraints. Here we make use of the special structure of the knapsack constraints on the budgets: each constraint is only related to the assignment of the corresponding product and its budget, so that when the budget of one product is exhausted, it does not affect the assignment of the other products. In the language of submodular optimization, the knapsack constraints are on a partition \mathcal{Z}_{i^*} of the ground set and the objective function is a sum of submodular functions over the partition.

However, there seems to be a hidden contradiction between spending the budgets efficiently and spending them as much as possible. On one hand, efficiency means the density ratio should be large, so the threshold ρ should be large; on the other hand, if ρ is large, there are just a few elements that can be considered, and thus the budget might not be exhausted. After all, if we set ρ to be even larger than the maximum possible value, then no element is considered and no gain is achieved. In the other extreme, if we set $\rho = 0$ and consider all the elements, then a few elements with large costs may be selected, exhausting all the budgets and leading to a poor solution.

Fortunately, there exists a suitable threshold ρ that achieves a good tradeoff between the two and leads to a good approximation. On one hand, the threshold is sufficiently small, so that the optimal elements we abandon (*i.e.*, those with low-density ratio) have a total gain at most a fraction of the optimum; on the other hand, it is also sufficiently large, so that the elements selected are of high quality (*i.e.*, of high-density ratio), and we achieve sufficient gain even if the budgets of some items are exhausted.

Theorem 5 *Suppose \hat{f} is evaluated up to error $\epsilon = \delta/16$ with CONTINEST. In Algorithm 1, there exists a ρ such that*

$$f(S_\rho) \geq \frac{\max\{k_a, 1\}}{(2|\mathcal{L}| + 2)(1 + 3\delta)} f(O)$$

where k_a is the number of active knapsack constraints:

$$k_a = |\{i : S_\rho \cup \{z\} \notin \mathcal{F}, \forall z \in \mathcal{Z}_{i^*}\}|.$$

The expected running time is $\tilde{O}\left(\frac{|\mathcal{E}_{i^*}| + |\mathcal{V}|}{\delta^2} + \frac{|\mathcal{L}||\mathcal{V}|}{\delta^4}\right)$.

Importantly, the approximation factor improves over the best known guarantee $\frac{1}{P+2k+1} = \frac{1}{2|\mathcal{L}|+2}$ for efficiently maximizing submodular functions over P matroids and k general knapsack constraints. Moreover, since the runtime has a linear dependence on the network size, the algorithm easily scales to large networks. As in the uniform cost case, the parameter δ introduces a tradeoff between the approximation and the runtime.

Analysis. The analysis follows the intuition. Pick $\rho = \frac{2f(O)}{P+2k+1}$, where O is the optimal solution, and define

$$\begin{aligned} O_- &:= \{z \in O \setminus S_\rho : f(z|S_\rho) < c(z)\rho + 2\epsilon\}, \\ O_+ &:= \{z \in O \setminus S_\rho : z \notin O_-\}. \end{aligned}$$

Note that, by submodularity, O_- is a superset of the elements in the optimal solution that we abandon due to the density threshold and, by construction, its marginal gain is small:

$$f(O_-|S_\rho) \leq \rho c(O_-) + \mathcal{O}(\epsilon|S_\rho|) \leq k\rho + \mathcal{O}(\epsilon|S_\rho|),$$

where the small additive term $\mathcal{O}(\epsilon|S_\rho|)$ is due to inexact function evaluations. Next, we proceed as follows.

First, if no knapsack constraints are active, then the algorithm runs as if there were no knapsack constraints (but only on elements with density ratio above ρ). Therefore, we can apply the same argument as in the case of uniform cost (refer to the analysis up to Claim 3 in Section 4.2.1); the only caveat is that we apply the argument to O_+ instead of $O \setminus S_\rho$. Formally, similar to Claim 3, the marginal gain of O_+ satisfies

$$f(O_+|S_\rho) \leq [(1 + \delta)P + \delta]f(S_\rho) + \mathcal{O}(\epsilon P|S_\rho|),$$

where the small additive term $\mathcal{O}(\epsilon P|S_\rho|)$ is due to inexact function evaluations. Using that $f(O) \leq f(S_\rho) + f(O_-|S_\rho) + f(O_+|S_\rho)$, we can conclude that S_ρ is roughly a $\frac{1}{P+2k+1}$ -approximation.

Second, suppose $k_a > 0$ knapsack constraints are active and the algorithm discovers that the budget of product i is exhausted when trying to add element z to the set $G_i = G \cap \mathcal{Z}_{i^*}$ of selected elements at that time. Since $c(G_i \cup \{z\}) > 1$ and each of these elements has density above ρ , the gain of $G_i \cup \{z\}$ is above ρ . However, only G_i is included in our final solution, so we need to show that the marginal gain of z is not large compared to that of G_i . To do so, we first realize that the algorithm greedily selects elements with marginal gain above a decreasing threshold w_t . Then, since z is the last element selected and G_i is nonempty (otherwise adding z will not exhaust the budget), the marginal gain of z must be bounded by roughly that of G_i , which is at least roughly $\frac{1}{2}\rho$. Since this holds for all active knapsack constraints, then the solution has value at least $\frac{k_a}{2}\rho$, which is an $\frac{k_a}{P+2k+1}$ -approximation.

Finally, combining both cases, and setting $k = |\mathcal{L}|$ and $P = 1$ as in our problem, we have our final guarantee in Theorem 5. Appendix D.2 presents the complete proofs.

5. Experiments on Synthetic and Real Data

In this section, we first evaluate the accuracy of the estimated influence given by CONTINEST and then investigate the performance of influence maximization on synthetic and real networks by incorporating CONTINEST into the framework of BUDGETMAX. We show that our approach significantly outperforms the state-of-the-art methods in terms of both speed and solution quality.

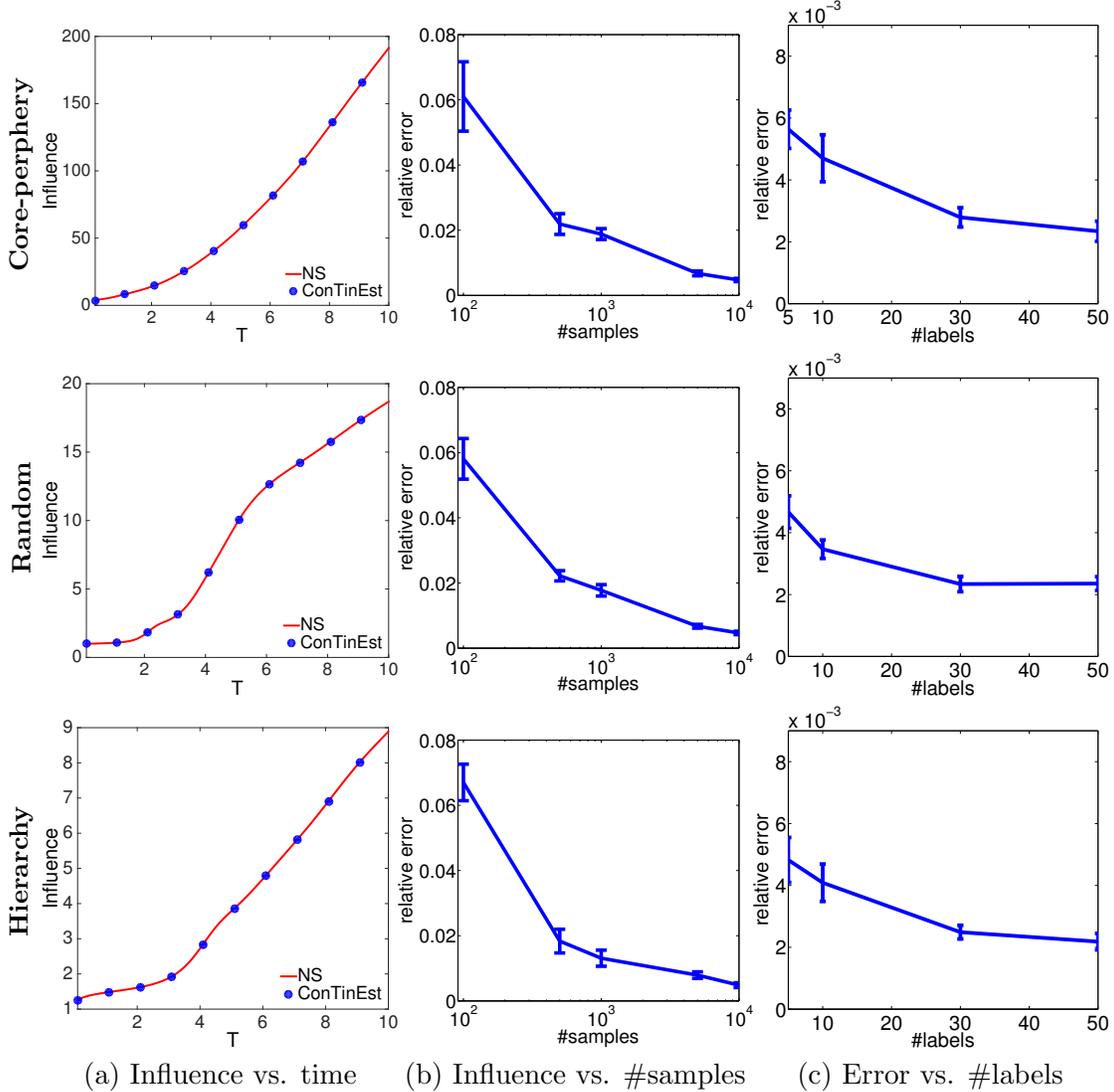


Figure 3: Influence estimation for core-periphery, random, and hierarchical networks with 1,024 nodes and 2,048 edges. Column (a) shows estimated influence by NS (near ground truth), and CONTINEST for increasing time window T ; Column (b) shows CONTINEST's relative error against number of samples with 5 random labels and $T = 10$; Column (c) reports CONTINEST's relative error against the number of random labels with 10,000 random samples and $T = 10$.

5.1 Experiments on Synthetic Data

We generate three types of Kronecker networks (Leskovec et al., 2010) which are synthetic networks generated by a recursive Kronecker product of a base 2-by-2 parameter matrix with itself to generate self-similar graphs. By tuning the base parameter matrix, we are able to generate the Kronecker networks which can mimic different structural properties of many real networks. In the following, we consider networks of three different types of structures: (i) core-periphery networks (parameter matrix: $[0.9 \ 0.5; 0.5 \ 0.3]$), which mimic the information diffusion traces in real-world networks (Gomez-Rodriguez et al., 2011), (ii) random networks ($[0.5 \ 0.5; 0.5 \ 0.5]$), typically used in physics and graph theory (Easley and Kleinberg, 2010) and (iii) hierarchical networks ($[0.9 \ 0.1; 0.1 \ 0.9]$) (Clauset et al., 2008). Next, we assign a pairwise transmission function for every directed edge in each type of network and set its parameters at random. In our experiments, we use the Weibull distribution from (Aalen et al., 2008),

$$f(t; \alpha, \beta) = \frac{\beta}{\alpha} \left(\frac{t}{\alpha} \right)^{\beta-1} e^{-(t/\alpha)^\beta}, t \geq 0, \quad (20)$$

where $\alpha > 0$ is a scale parameter and $\beta > 0$ is a shape parameter. The Weibull distribution (Wbl) has often been used to model lifetime events in survival analysis, providing more flexibility than an exponential distribution. We choose α and β from 0 to 10 uniformly at random for each edge in order to have heterogeneous temporal dynamics. Finally, for each type of Kronecker network, we generate 10 sample networks, each of which has different α and β chosen for every edge.

5.1.1 INFLUENCE ESTIMATION

To the best of our knowledge, there is no analytical solution to the influence estimation given Weibull transmission function. Therefore, we compare CONTINEST with the Naive Sampling (NS) approach by considering the highest degree node in a network as the source, and draw 1,000,000 samples for NS to obtain near ground truth. In Figure 3, Column (a) compares CONTINEST with the ground truth provided by NS at different time window T , from 0.1 to 10 in networks of different structures. For CONTINEST, we generate up to 10,000 random samples (or sets of random waiting times), and 5 random labels in the inner loop. In all three networks, estimation provided by CONTINEST fits the ground truth accurately, and the relative error decreases quickly as we increase the number of samples and labels (Column (b) and Column (c)). For 10,000 random samples with 5 random labels, the relative error is smaller than 0.01.

5.1.2 INFLUENCE MAXIMIZATION WITH UNIFORM COST

In this section, we first evaluate the effectiveness of CONTINEST to the classic influence maximization problem where we have only one product to assign with the simple cardinality constraint on the users. We compare to other influence maximization methods developed based on discrete-time diffusion models: traditional greedy by (Kempe et al., 2003), with discrete-time Linear Threshold Model (LT) and Independent Cascade Model (IC) diffusion models, and the heuristic methods SP1M, PMIA and MIA-M by (Chen et al., 2009, 2010a, 2012). For INFLUMAX, since it only supports exponential pairwise transmission functions,

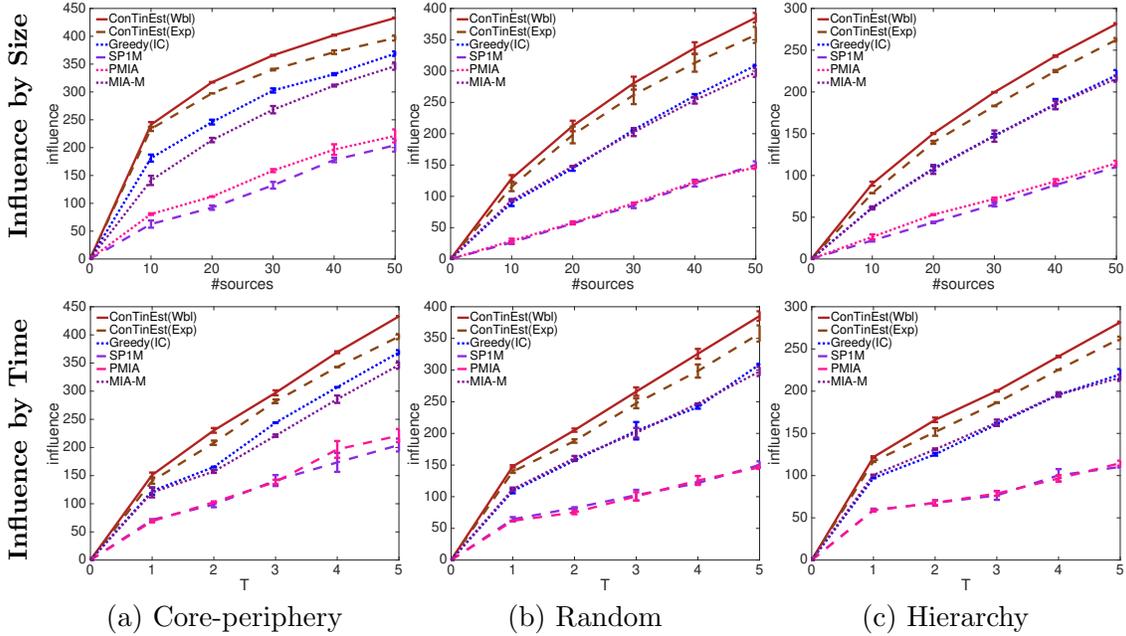


Figure 4: Influence $\sigma(\mathcal{A}, T)$ achieved by varying number of sources $|\mathcal{A}|$ and observation window T on the networks of different structures with 1,024 nodes, 2,048 edges and heterogeneous Weibull transmission functions. Top row: influence against $\#sources$ by $T = 5$; Bottom row: influence against the time window T using 50 sources.

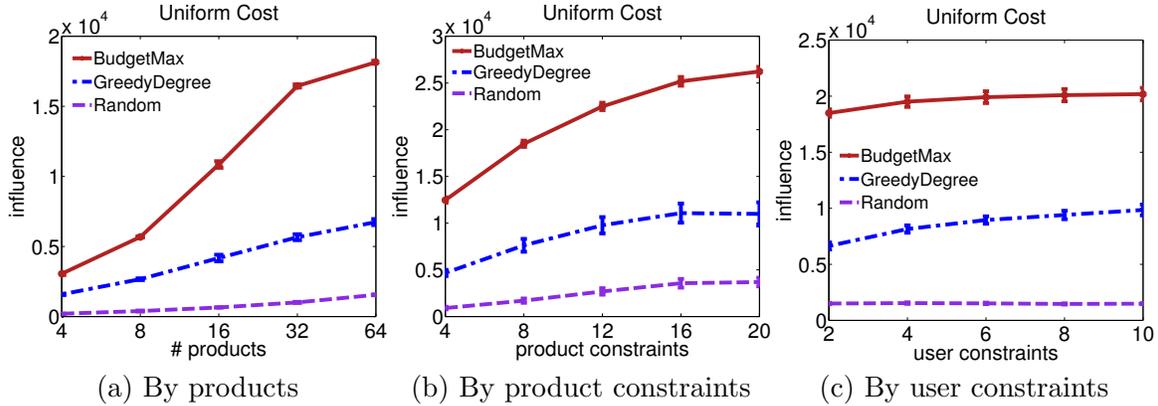


Figure 5: Over the 64 product-specific diffusion networks, each of which has 1,048,576 nodes, the estimated influence (a) for increasing the number of products by fixing the product-constraint at 8 and user-constraint at 2; (b) for increasing product-constraint by user-constraint at 2; and (c) for increasing user-constraint by fixing product-constraint at 8. For all experiments, we have $T = 5$ time window.

we fit an exponential distribution per edge by NETRATE (Gomez-Rodriguez et al., 2011). Furthermore, INFLUMAX is not scalable. When the average network density (defined as the average degree per node) of the synthetic networks is ~ 2.0 , the run time for INFLUMAX

is more than 24 hours. In consequence, we present the results of CONTINEST using fitted exponential distributions (Exp). For the discrete-time IC model, we learn the infection probability within time window T using Netrapalli’s method (Netrapalli and Sanghavi, 2012). The learned pairwise infection probabilities are also served for SP1M and PMIA, which approximately calculate the influence based on the IC model. For the discrete-time LT model, we set the weight of each incoming edge to a node u to the inverse of its in-degree, as in previous work (Kempe et al., 2003), and choose each node’s threshold uniformly at random. The top row of Figure 4 compares the expected number of infected nodes against the source set size for different methods. CONTINEST outperforms the rest, and the competitive advantage becomes more dramatic the larger the source set grows. The bottom row of Figure 4 shows the expected number of infected nodes against the time window for 50 selected sources. Again, CONTINEST performs the best for all three types of networks.

Next, using CONTINEST as a subroutine for influence estimation, we evaluate the performance of BUDGETMAX with the uniform-cost constraints on the users. In our experiments we consider up to 64 products, each of which diffuses over one of the above three different types of Kronecker networks with \sim one million nodes. Further, we randomly select a subset of 512 nodes $\mathcal{V}_S \subseteq \mathcal{V}$ as our candidate target users, who will receive the given products, and evaluate the potential influence of an allocation over the underlying one-million-node networks. For BUDGETMAX, we set the adaptive threshold δ to 0.01 and the cost per user and product to 1. For CONTINEST, we use 2,048 samples with 5 random labels on each of the product-specific diffusion networks. We repeat our experiments 10 times and report the average performance.

We compare BUDGETMAX with a nodes’ degree-based heuristic, which we refer to as GreedyDegree, where the degree is treated as a natural measure of influence, and a baseline method, which assigns the products to the target nodes randomly. We opt for the nodes’ degree-based heuristic since, in practice, large-degree nodes, such as users with millions of followers in Twitter, are often the targeted users who will receive a considerable payment if he (she) agrees to post the adoption of some products (or ads) from merchants. GreedyDegree proceeds as follows. It first sorts the list of all pairs of products i and nodes $j \in \mathcal{V}_S$ in descending order of node- j ’s degree in the diffusion network associated to product i . Then, starting from the beginning of the list, it considers each pair one by one: if the addition of the current pair to the existing solution does not violate the predefined matroid constraints, it is added to the solution, and otherwise, it is skipped. This process continues until the end of the list is reached. In other words, we greedily assign products to nodes with the largest degree. Due to the large size of the underlying diffusion networks, we do not apply other more expensive node centrality measures such as the clustering coefficient and betweenness.

Figure 5 summarizes the results. Panel (a) shows the achieved influence against number of products, fixing the budget per product to 8 and the budget per user to 2. As the number of products increases, on the one hand, more and more nodes become assigned, so the total influence will increase. Yet, on the other hand, the competition among products for a few existing *influential* nodes also increases. GreedyDegree achieves a modest performance, since high degree nodes may have many overlapping children. In contrast, BUDGETMAX, by taking the submodularity of the problem, the network structure and the diffusion dynamics of the edges into consideration, achieves a superior performance, especially as the number

of product (*i.e.*, the competition) increases. Panel (b) shows the achieved influence against the budget per product, considering 64 products and fixing the budget per user to 2. We find that, as the budget per product increases, the performance of GreedyDegree tends to flatten and the competitive advantage of BUDGETMAX becomes more dramatic. Finally, Panel (c) shows the achieved influence against the budget per user, considering 64 products and fixing the budget per product to 8. We find that, as the budget per user increases, the influence only increases slowly. This is due to the fixed budget per product, which prevents additional new nodes to be assigned. This meets our intuition: by making a fixed number of people watching more ads per day, we can hardly boost the popularity of the product. Additionally, even though the same node can be assigned to more products, there is hardly ever a node that is the *perfect* source from which all products can efficiently spread.

5.1.3 INFLUENCE MAXIMIZATION WITH NON-UNIFORM COST

In this section, we evaluate the performance of BUDGETMAX under non-uniform cost constraints, using again CONTINEST as a subroutine for influence estimation. Our designing of user-cost aim to mimic a real scenario, where advertisers pay much more money to celebrities with millions of social network followers than to normal citizens. To do so, we let $c_i \propto d_i^{1/n}$ where c_i is the cost, d_i is the degree, and $n \geq 1$ controls the increasing speed of cost with respect to the degree. In our experiments, we use $n = 3$ and normalize c_i to be within $(0, 1]$. Moreover, we set the product-budget to a base value from 1 to 10 and add a random adjustment drawn from a uniform distribution $U(0, 1)$.

We compare our method to two modified versions of the above mentioned nodes' degree-based heuristic GreedyDegree and to the same baseline method. In the first modified version of the heuristic, which we still refer to as GreedyDegree, takes both the degree and the corresponding cost into consideration. In particular, it sorts the list of all pairs of products i and nodes $j \in \mathcal{V}_S$ in descending order of degree-cost ratio d_j/c_j in the diffusion network associated to product i , instead of simply the node- j 's degree, and then proceeds similarly as before. In the second modified version of the heuristic, which we refer as GreedyLocalDegree, we use the same degree-cost ratio but allow the target users to be partitioned into distinct groups (or communities) and pick the most cost-effective pairs within each group locally instead. Figure 6 compares the performance of our method with the competing methods against four factors: (a) the number of products, (b) the budget per product, (c) the budget per user and (d) the time window T , while fixing the other factors. In all cases, BUDGETMAX significantly outperforms the other methods, and the achieved influence increases monotonically with respect to the factor value, as one may have expected. In addition, in Figure 6(e), we study the effect of the Laminar matroid combined with group knapsack constraints, which is the most general type of constraint we handle in this paper (refer to Section 3.4). The selected target users are further partitioned into K groups randomly, each of which has, $Q_i, i = 1 \dots K$, limit which constrains the maximum allocations allowed in each group. In practical scenarios, each group might correspond to a geographical community or organization. In our experiment, we divide the users into 8 equal-size groups and set $Q_i = 16, i = 1 \dots K$ to indicate that we want a balanced allocation in each group. Figure 6(e) shows the achieved influence against the budget per user for $K = 8$ equally-sized groups and $Q_i = 16, i = 1 \dots K$. In contrast to Figure 6(b), the total

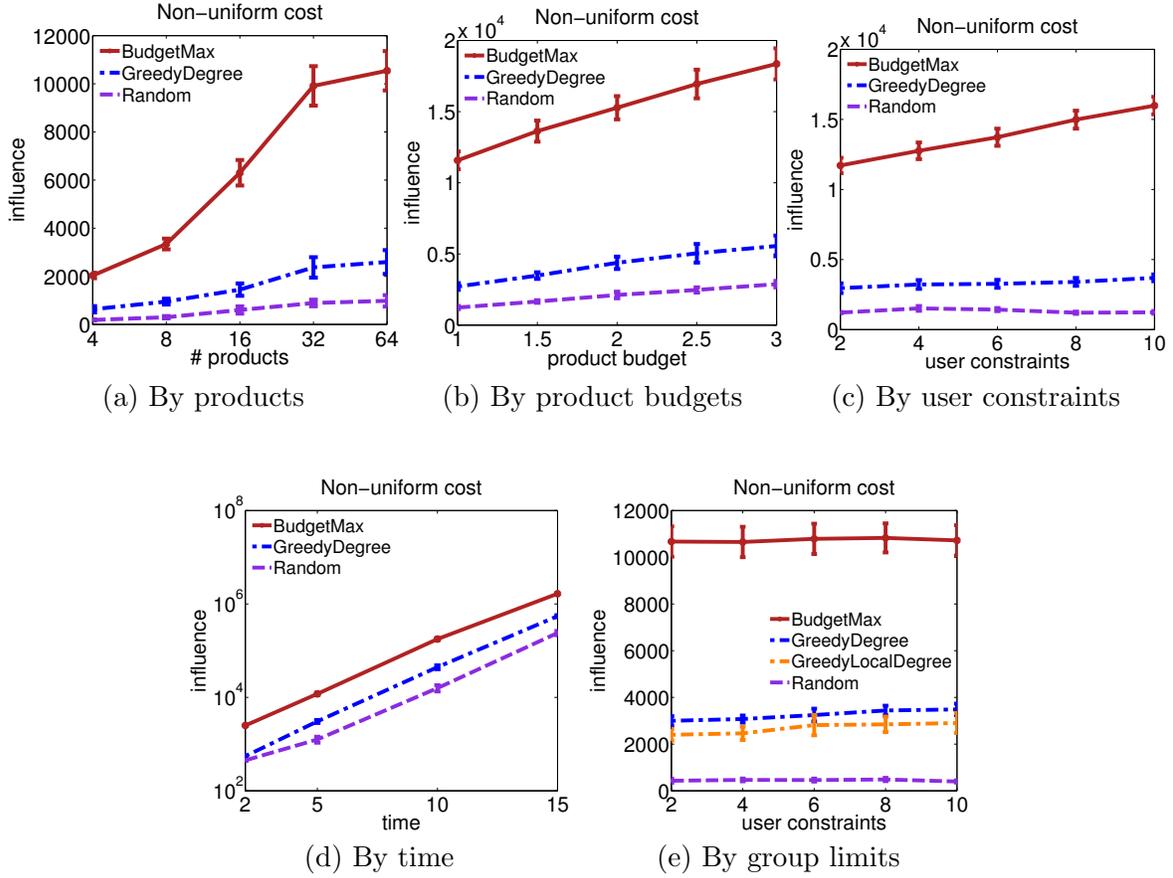


Figure 6: Over the 64 product-specific diffusion networks, each of which has a total 1,048,576 nodes, the estimated influence (a) for increasing the number of products by fixing the product-budget at 1.0 and user-constraint at 2; (b) for increasing product-budget by fixing user-constraint at 2; (c) for increasing user-constraint by fixing product-budget at 1.0; (d) for different time window T ; and (e) for increasing user-constraint with group-limit 16 by fixing product-budget at 1.0.

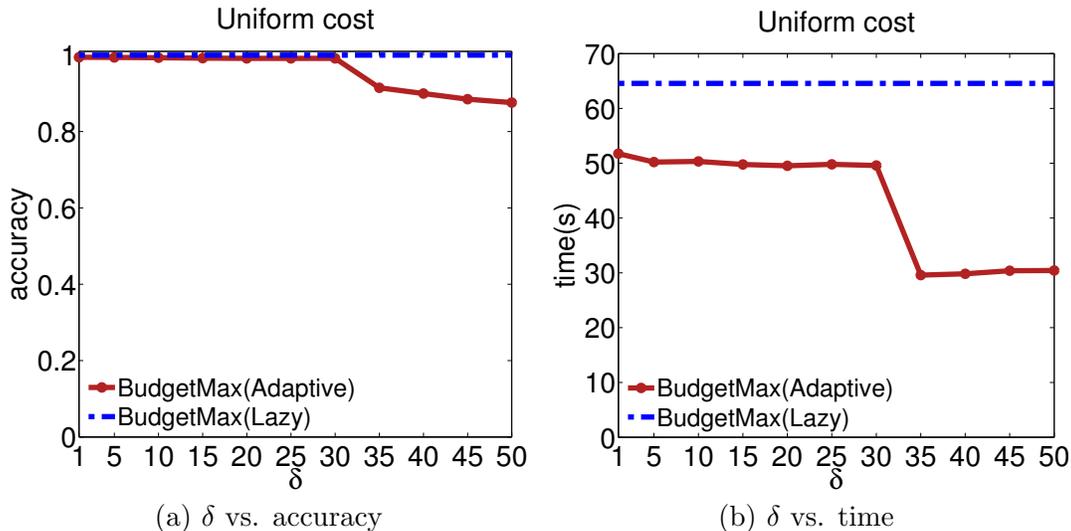


Figure 7: The relative accuracy and the run-time for different threshold parameter δ .

estimated influence does not increase significantly with respect to the budget (*i.e.*, number of slots) per user. This is due to the fixed budget per group, which prevents additional new nodes to be assigned, even though the number of available slots per user increases.

5.1.4 EFFECTS OF ADAPTIVE THRESHOLDING

In Figure 7, we investigate the impact that the threshold value δ has on the accuracy and runtime of our adaptive thresholding algorithm and compare it with the lazy evaluation method. Note that the performance and runtime of lazy evaluation do not change with respect to δ because it does not depend on it. Panel (a) shows the achieved influence against the threshold δ . As expected, the larger the δ value, the lower the accuracy. However, our method is relatively robust to the particular choice of δ since its performance is always over a 90-percent relative accuracy even for large δ . Panel (b) shows the runtime against the threshold δ . In this case, the larger the δ value, the lower the runtime. In other words, Figure 7 verifies the intuition that δ is able to trade off the solution quality of the allocation with the runtime time.

5.1.5 SCALABILITY

In this section, we start with evaluating the scalability of the proposed algorithms on the classic influence maximization problem where we only have one product with the cardinality constraint on the users.

We compare it to the state-of-the-art method INFLUMAX (Gomez-Rodriguez and Schölkopf, 2012) and the Naive Sampling (NS) method in terms of runtime for the continuous-time influence estimation and maximization. For CONTINEST, we draw 10,000 samples in the outer loop, each having 5 random labels in the inner loop. We plug CONTINEST as a subroutine into the classic greedy algorithm by (Nemhauser et al., 1978). For NS, we also draw 10,000 samples. The first two experiments are carried out in a single 2.4GHz processor.

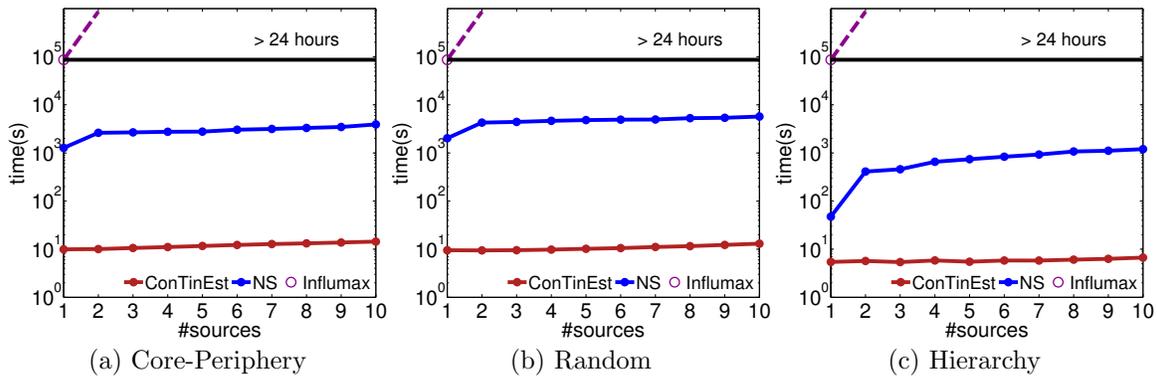


Figure 8: Runtime of selecting increasing number of sources on Kronecker networks of 128 nodes and 320 edges with $T = 10$.

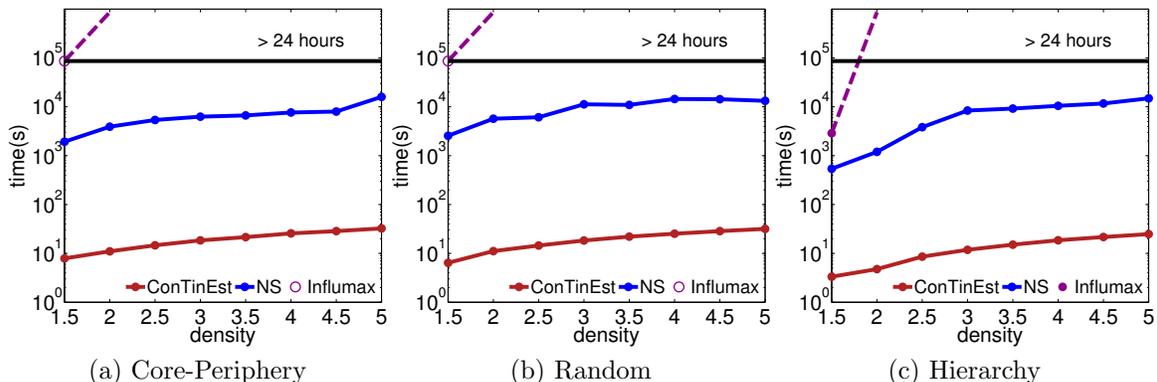


Figure 9: Runtime of selecting 10 sources in networks of 128 nodes with increasing density by $T = 10$.

Figure 8 compares the performance of increasingly selecting sources (from 1 to 10) on small Kronecker networks. When the number of selected sources is 1, different algorithms essentially spend time estimating the influence for each node. CONTINEST outperforms other methods by order of magnitude and for the number of sources larger than 1, it can efficiently reuse computations for estimating influence for individual nodes. Dashed lines mean that a method did not finish in 24 hours, and the estimated run time is plotted.

Next, we compare the run time for selecting 10 sources with increasing densities (or the number of edges) in Figure 9. Again, INFLUMAX and NS are order of magnitude slower due to their respective exponential and quadratic computational complexity in network density. In contrast, the run time of CONTINEST only increases slightly with the increasing density since its computational complexity is linear in the number of edges. We evaluate the speed on large core-periphery networks, ranging from 100 to 1,000,000 nodes with density 1.5 in Figure 10. We report the parallel run time only for CONTINEST and NS (both are implemented by MPI running on 192 cores of 2.4Ghz) since INFLUMAX is not scalable. In contrast to NS, the performance of CONTINEST increases linearly with the network size and can easily scale up to one million nodes.

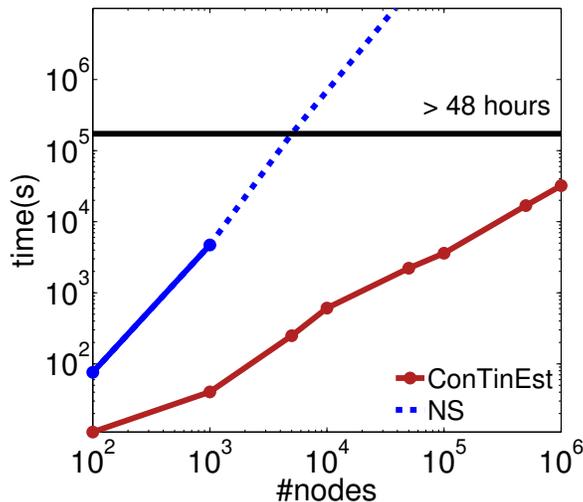


Figure 10: For core-periphery networks by $T = 10$, runtime of selecting 10 sources with increasing network size from 100 to 1,000,000 by fixing 1.5 network density.

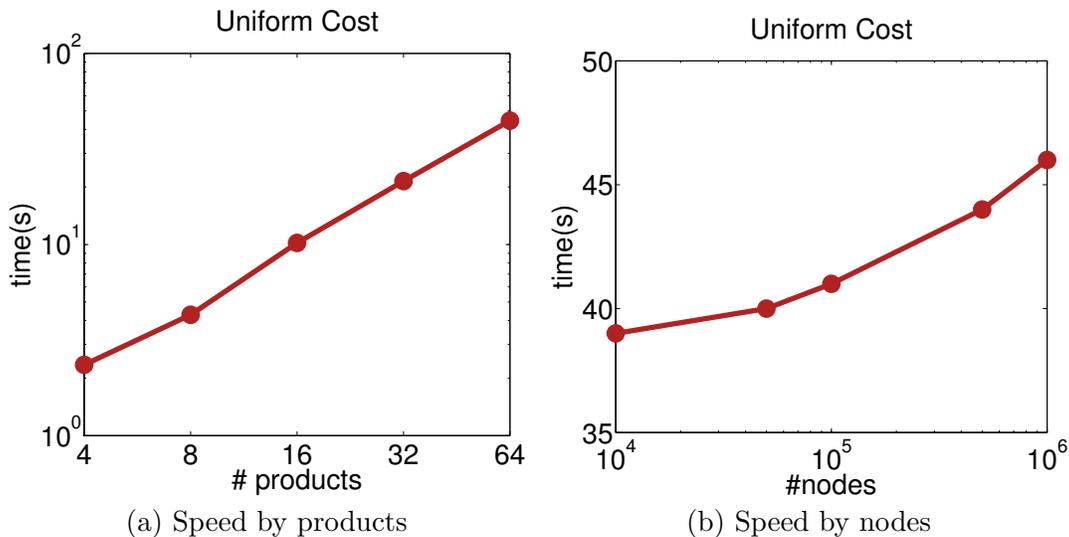


Figure 11: Over the 64 product-specific diffusion networks, each of which has 1,048,576 nodes, the runtime (a) of allocating increasing number of products and (b) of allocating 64 products to 512 users on networks of varying size. In all experiments, we have $T = 5$ time window and fix product-constraint at 8 and user-constraint at 2.

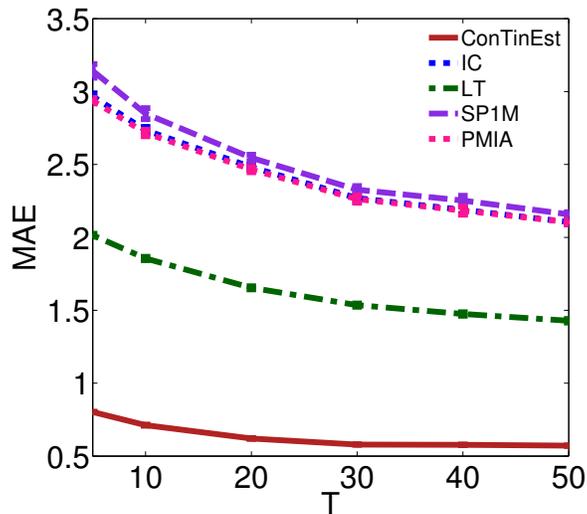


Figure 12: In MemeTracker dataset, comparison of the accuracy of the estimated influence in terms of mean absolute error.

Finally, we investigate the performance of BUDGETMAX in terms of runtime when using CONTINEST as subroutine to estimate the influence. We can precompute the data structures and store the samples needed to estimate the influence function in advance. Therefore, we focus only on the runtime for the constrained influence maximization algorithm. BUDGETMAX runs on 64 cores of 2.4Ghz by using OpenMP to accelerate the first round of the optimization. We report the allocation time for increasing number of products in Figure 11(a), which clearly shows a linear time complexity with respect to the size of the ground set. Figure 11(b) evaluates the runtime of allocation by varying the size of the network from 16,384 to 1,048,576 nodes.

5.2 Experiments on Real Data

In this section, we first quantify how well our proposed algorithm can estimate the true influence in the real-world dataset. Then, we evaluate the solution quality of the selected sources for influence maximization under different constraints. We have used the public MemeTracker datasets (Leskovec et al., 2009), which contains more than 172 million news articles and blog posts from 1 million mainstream media sites and blogs.

5.2.1 INFLUENCE ESTIMATION

We first trace the flow of information from one site to another by using the hyperlinks among articles and posts as in the work of Gomez-Rodriguez et al. (2011); Du et al. (2012). In detail, we extracted 10,967 hyperlink cascades among top 600 media sites. We then evaluate the accuracy of CONTINEST as follows. First, we repeatedly split all cascades into a 80% training set and a 20% test set at random for five times. On each training set, we learn one continuous-time model, which we use for CONTINEST, and a discrete-time model, which we use for the competitive methods: IC, SP1M, PMIA and MIAM-M. For the continuous-time

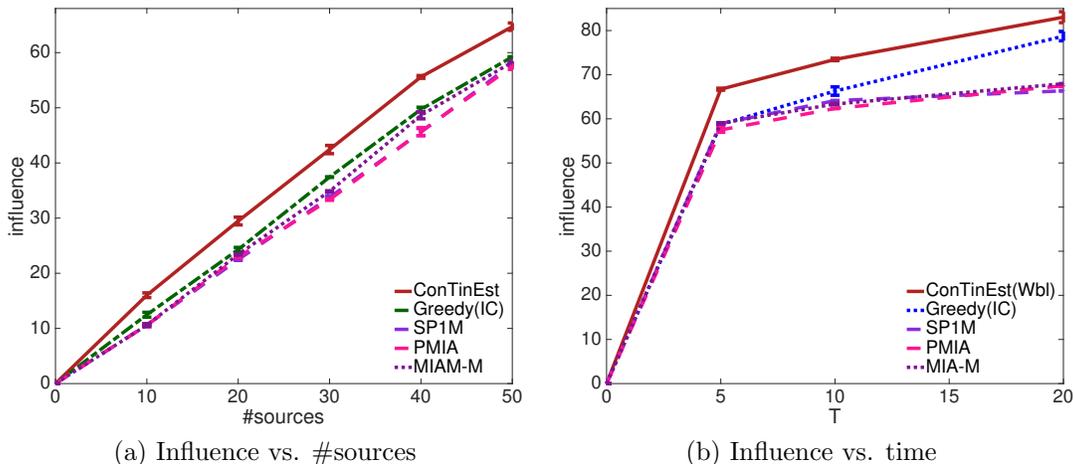


Figure 13: In MemeTracker dataset, (a) comparison of the influence of the selected nodes by fixing the observation window $T = 5$ and varying the number sources, and (b) comparison of the influence of the selected nodes by fixing the number of sources to 50 and varying the time window.

model, we opt for NETRATE (Gomez-Rodriguez et al., 2011) with exponential transmission functions (fixing the shape parameter of the Weibull family to be one) to learn the diffusion networks by maximizing the likelihood of the observed cascades. For the discrete-time model, we learn the infection probabilities using the method by Netrapalli and Sanghavi (2012).

Second, let $\mathcal{C}(u)$ be the set of all cascades where u was the source node. By counting the total number of distinct nodes infected before T in $\mathcal{C}(u)$, we can quantify the real influence of node u up to time T . Thus, we can evaluate the quality of the influence estimation by computing the average (across nodes) Mean Absolute Error (MAE) between the real and the estimated influence on the test set, which we show in Figure 12. Clearly, CONTINEST performs the best statistically. Since the length of real cascades empirically conforms to a power-law distribution, where most cascades are very short (2-4 nodes), the gap of the estimation error is not too large. However, we emphasize that such accuracy improvement is critical for maximizing long-term influence since the estimation error for individuals will accumulate along the spreading paths. Hence, any consistent improvement in influence estimation can lead to significant improvement to the overall influence estimation and maximization task, which is further confirmed in the following sections.

5.2.2 INFLUENCE MAXIMIZATION WITH UNIFORM COST

We first apply CONTINEST to the continuous-time influence maximization task with the simple cardinality constraint on the users. We evaluate the influence of the selected nodes in the same spirit as influence estimation: the true influence is calculated as the total number of distinct nodes infected before T based on $\mathcal{C}(u)$ of the selected nodes. Figure 13 shows that the selected sources given by CONTINEST achieve the best performance as we vary the number of selected sources and the observation time window.

Next, we evaluate the performance of BUDGETMAX on cascades from Memetracker traced from quotes which are short textual phrases spreading through the websites. Because all published documents containing a particular quote are time-stamped, a cascade induced by the same quote is a collection of times when the media site first mentioned it. In detail, we use the public dataset released by Gomez-Rodriguez et al. (2013), which splits the original Memetracker dataset into groups, each associated to a topic or real-world event. Each group consists of cascades built from quotes which were mentioned in posts containing particular keywords. We considered 64 groups, with at least 100,000 cascades, which play the role of products. Therein, we distinguish well-known topics, such as “Apple” and “Occupy Wall-Street”, or real-world events, such as the Fukushima nuclear disaster in 2013 and the marriage between Kate Middleton and Prince William in 2011.

We then evaluate the accuracy of BUDGETMAX in the following way. First, we evenly split each group into a training and a test set and then learn one continuous-time model and a discrete-time model per group using the training sets. As previously, for the continuous-time model, we opt for NETRATE (Gomez-Rodriguez et al., 2011) with exponential transmission functions, and for the discrete-time model, we learn the infection probabilities using the method by Netrapalli and Sanghavi (2012), where the step-length is set to one. Second, we run BUDGETMAX using both the continuous-time model and the discrete-time model. We refer to BUDGETMAX with the discrete-time model as the Greedy(discrete) method. Since we do not have no ground-truth information about cost of each node, we focus our experiments using a uniform cost. Third, once we have found an allocation over the learned networks, we evaluate the performance of the two methods using the cascades in the test set as follows: given a group-node pair (i, j) , let $\mathcal{C}(j)$ denote the set of cascades induced by group i that contains node j . Then, we take the average number of nodes coming after j for all the cascades in $\mathcal{C}(j)$ as a proxy of the average influence induced by assigning group i to node j . Finally, the influence of an allocation is just the sum of the average influence of each group-node pair in the solution. In our experiments, we randomly select 128 nodes as our target users.

Figure 14 summarizes the achieved influence against four factors: (a) the number of products, (b) the budget per product, (c) the budget per user and (d) the time window T , while fixing the other factors. In comparison with the Greedy(IC) and a random allocation, BUDGETMAX finds an allocation that indeed induces the largest diffusion in the test data, with an average 20-percent improvement overall.

In the end, Figure 15 investigates qualitatively the actual allocations of groups (topics or real-world events; in red) and sites (in black). Here, we find examples that intuitively one could expect: “japantoday.com” is assigned to Fukushima Nuclear disaster or “finance.yahoo.com” is assigned to “Occupy Wall-street”. Moreover, because we consider several topics and real-world events with different underlying diffusion networks, the selected nodes are not only very popular media sites such as nytimes.com or cnn.com but also several modest sites (Bakshy et al., 2011), often specialized or local, such as freep.com or localnews8.com.

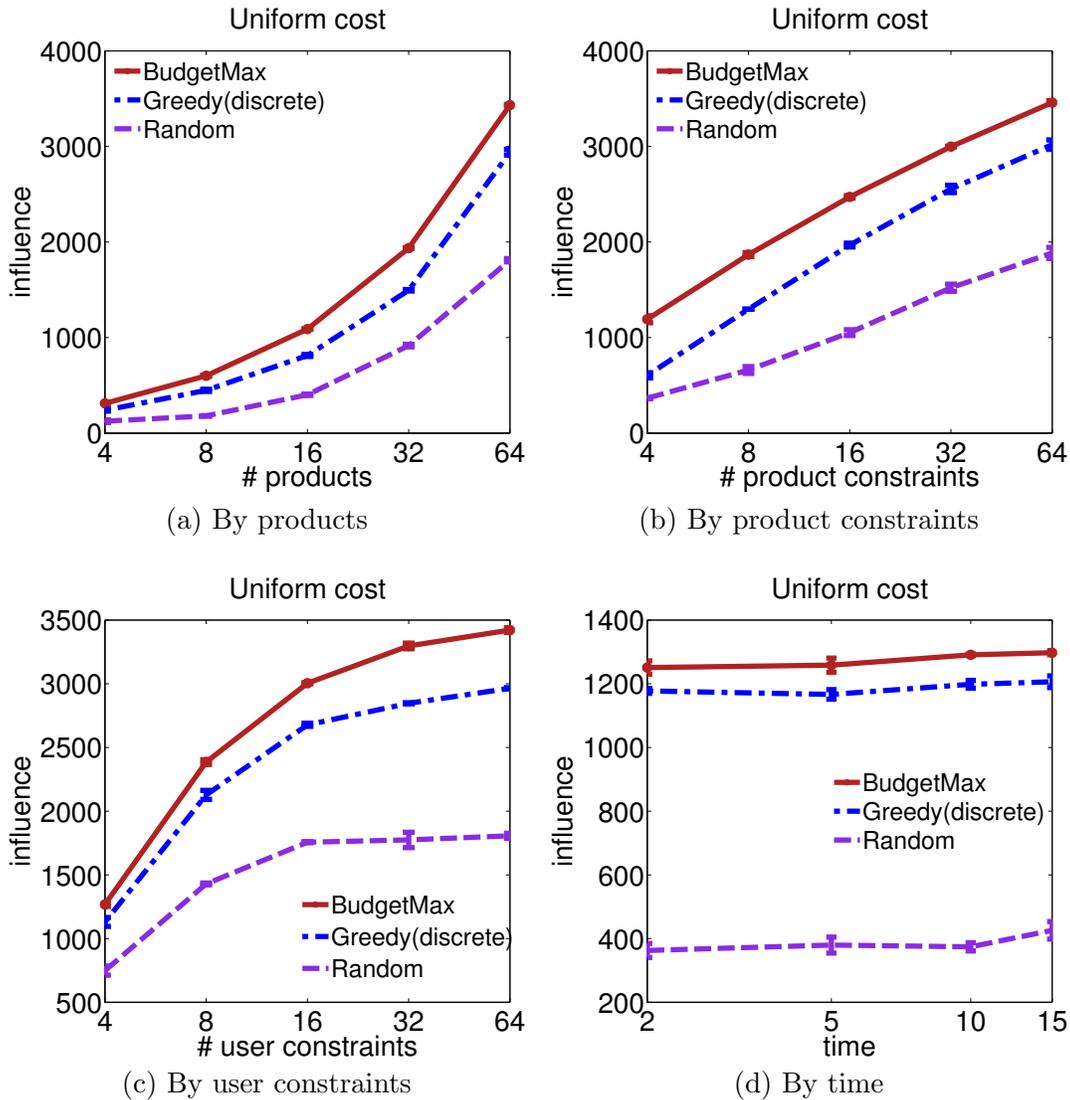


Figure 14: Over the inferred 64 product-specific diffusion networks, the true influence estimated from separated testing data (a) for increasing the number of products by fixing the product-constraint at 8 and user-constraint at 2; (b) for increasing product-constraint by fixing user-constraint at 2; (c) for increasing user-constraint by fixing product-constraint at 8; (d) for different time window T .

6. Conclusion

We have studied the influence estimation and maximization problems in the continuous-time diffusion model. We first propose a randomized influence estimation algorithm `CONTINEST`, which can scale up to networks of millions of nodes while significantly improves over previous state of the art methods in terms of the accuracy of the estimated influence. Once we have a subroutine for efficient influence estimation in large networks, we then tackle the problem of maximizing the influence of multiple types of products (or information) in

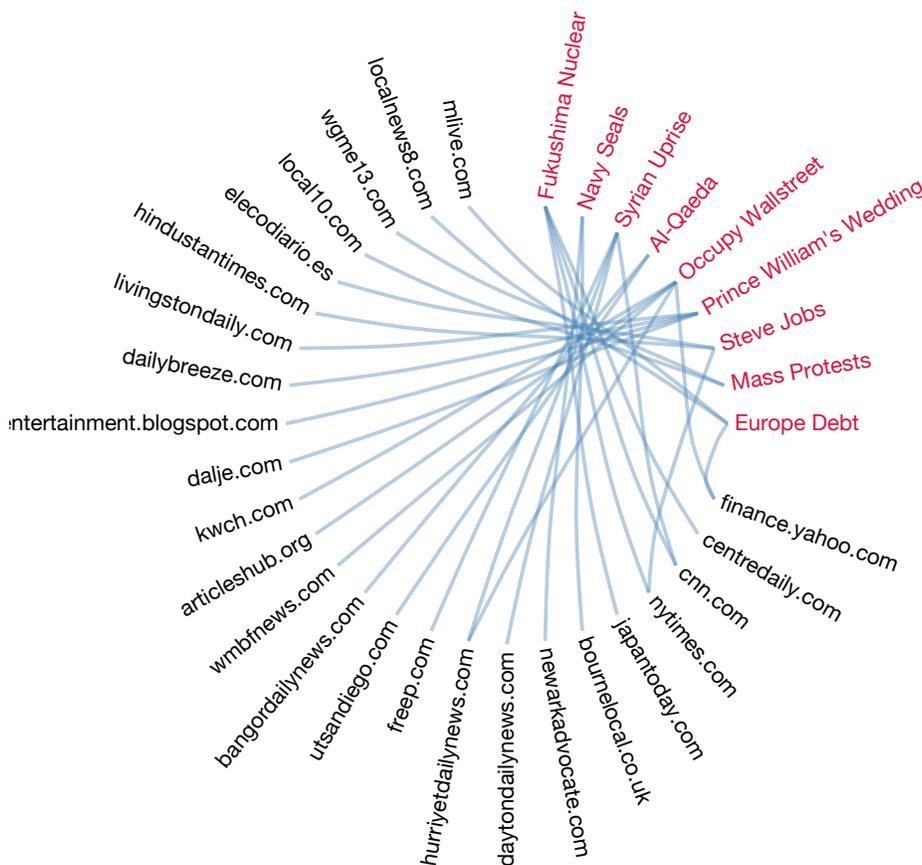


Figure 15: The allocation of memes to media sites.

realistic continuous-time diffusion networks subject to various practical constraints: different products can have different diffusion structures; only influence within a given time window is considered; each user can only be recommended a small number of products; and each product has a limited campaign budget, and assigning it to users incurs costs. We provide a novel formulation as a submodular maximization under an intersection of matroid constraints and group-knapsack constraints, and then design an efficient adaptive threshold greedy algorithm with provable approximation guarantees, which we call BUDGETMAX. Experimental results show that the proposed algorithm performs remarkably better than other scalable alternatives in both synthetic and real-world datasets. There are also a few interesting open problems. For example, when the influence is estimated using CONTINEST, its error is a random variable. How does this affect the submodularity of the influence function? Is there an influence maximization algorithm that has better tolerance to the random error? These questions are left for future work.

Acknowledgments

Nan Du is supported by the Facebook Graduate Fellowship 2014-2015. Maria-Florina Balcan and Yingyu Liang are supported in part by NSF grant CCF-1101283 and CCF-0953192, AFOSR grant FA9550-09-1-0538, ONR grant N00014-09-1-075, a Microsoft Faculty Fellowship, and a Raytheon Faculty Fellowship. Le Song is supported in part by NSF/NIH BIGDATA 1R01GM108341, ONR N00014-15-1-2340, NSF IIS-1218749, NSF CAREER IIS-1350983, Nvidia and Intel. Hongyuan Zha is supported in part by NSF/NIH BIGDATA 1R01GM108341, NSF DMS-1317424 and NSF CNS-1409635.

References

- Odd Aalen, Oernulf Borgan, and Håkon K Gjessing. *Survival and event history analysis: a process point of view*. Springer, 2008.
- Ashwinkumar Badanidiyuru and Jan Vondrák. Fast algorithms for maximizing submodular functions. In *SODA*. SIAM, 2014.
- Eytan Bakshy, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. Everyone’s an influencer: Quantifying influence on twitter. In *International Conference on Web Search and Data Mining (WSDM)*, pages 65–74, 2011.
- Christian Borgs, Michael Brautbar, Jennifer Chayes, and Brendan Lucier. Influence maximization in social networks: Towards an optimal algorithmic solution. *arXiv preprint arXiv:1212.0884*, 2012.
- Wei Chen, Yajun Wang, and Siyu Yang. Efficient influence maximization in social networks. In *Knowledge Discovery and Data Mining (SIGKDD)*, pages 199–208. ACM, 2009.
- Wei Chen, Chi Wang, and Yajun Wang. Scalable influence maximization for prevalent viral marketing in large-scale social networks. In *Knowledge Discovery and Data Mining (SIGKDD)*, pages 1029–1038. ACM, 2010a.
- Wei Chen, Yifei Yuan, and Li Zhang. Scalable influence maximization in social networks under the linear threshold model. In *International Conference on Data Mining series (ICDM)*, pages 88–97. IEEE, 2010b.
- Wei Chen, Alex Collins, Rachel Cummings, Te Ke, Zhenming Liu, David Rincon, Xiaorui Sun, Yajun Wang, Wei Wei, and Yifei Yuan. Influence maximization in social networks when negative opinions may emerge and propagate. In *SIAM International Conference on Data Mining (SDM)*, pages 379–390. SIAM, 2011.
- Wei Chen, Wei Lu, and Ning Zhang. Time-critical influence maximization in social networks with time-delayed diffusion process. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2012.
- Aaron Clauset, Cristopher Moore, and M.E.J. Newman. Hierarchical structure and the prediction of missing links in networks. *Nature*, 453(7191):98–101, 2008.

- Edith Cohen. Size-estimation framework with applications to transitive closure and reachability. *Journal of Computer and System Sciences*, 55(3):441–453, 1997.
- Samik Datta, Anirban Majumder, and Nisheeth Shrivastava. Viral marketing for multiple products. In *International Conference on Data Mining series (ICDM)*, pages 118–127, 2010.
- Nan Du, Le Song, Alex Smola, and Ming Yuan. Learning networks of heterogeneous influence. In *Neural Information Processing System (NIPS)*, pages 2789–2797, 2012.
- Nan Du, Le Song, Manuel Gomez-Rodriguez, and Hongyuan Zha. Scalable influence estimation in continuous time diffusion networks. In *Neural Information Processing System (NIPS)*, 2013a.
- Nan Du, Le Song, Hyenkyun Woo, and Hongyuan Zha. Uncover topic-sensitive information diffusion networks. In *Artificial Intelligence and Statistics (AISTATS)*, 2013b.
- David Easley and Jon Kleinberg. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, 2010.
- Satoru Fujishige. *Submodular functions and optimization*, volume 58. Elsevier Science Limited, 2005.
- Manuel Gomez-Rodriguez and Bernhard Schölkopf. Influence maximization in continuous time diffusion networks. In *International Conference on Machine Learning (ICML)*, pages 313–320, 2012.
- Manuel Gomez-Rodriguez, David Balduzzi, and Bernhard Schölkopf. Uncovering the temporal dynamics of diffusion networks. In *International Conference on Machine Learning (ICML)*, 2011.
- Manuel Gomez-Rodriguez, Jure Leskovec, and Bernhard Schölkopf. Structure and Dynamics of Information Pathways in On-line Media. In *International Conference on Web Search and Web Data Mining (WSDM)*, 2013.
- Amit Goyal, Wei Lu, and Laks V. S. Lakshmanan. Celf++: optimizing the greedy algorithm for influence maximization in social networks. In *World Wide Web Conference (WWW) (Companion Volume)*, pages 47–48, 2011a.
- Amit Goyal, Wei Lu, and Laks V. S. Lakshmanan. Simpath: An efficient algorithm for influence maximization under the linear threshold model. In *International Conference on Data Mining series (ICDM)*, pages 211–220, 2011b.
- Dino Ienco, Francesco Bonchi, and Carlos Castillo. The meme ranking problem: Maximizing microblogging virality. In *International Conference on Data Mining series (ICDM) Workshops*, 2010.
- David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 137–146. ACM, 2003.

- Jure Leskovec, Andreas Krause, Carlos Guestrin, Christos Faloutsos, Jeanne Van-Briesen, and Natalie Glance. Cost-effective outbreak detection in networks. In P. Berkhin, R. Caruana, and X. Wu, editors, *International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 420–429. ACM, 2007. URL <http://doi.acm.org/10.1145/1281192.1281239>.
- Jure Leskovec, Lars Backstrom, and Jon Kleinberg. Meme-tracking and the dynamics of the news cycle. In *International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 497–506. ACM, 2009.
- Jure Leskovec, Deepayan Chakrabarti, Jon Kleinberg, Christos Faloutsos, and Zoubin Ghahramani. Kronecker graphs: An approach to modeling networks. *Journal of Machine Learning Research (JMLR)*, 11(Feb):985–1042, 2010.
- Wei Lu, Francesco Bonchi, Goyal Amit, and Laks V. S. Lakshmanan. The bang for the buck: fair competitive viral marketing from the host perspective. In *International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 928–936, 2013.
- Ramasuri Narayanam and Amit A Nanavati. Viral marketing for product cross-sell through social networks. In *Machine Learning and Knowledge Discovery in Databases*. 2012.
- George Nemhauser, Laurence Wolsey, and Marshall Fisher. An analysis of the approximations for maximizing submodular set functions. *Mathematical Programming*, 14:265–294, 1978.
- Praneeth Netrapalli and Sujay Sanghavi. Learning the graph of epidemic cascades. In *SIGMETRICS/PERFORMANCE*, pages 211–222. ACM, 2012. ISBN 978-1-4503-1097-0.
- Matthew Richardson and Pedro Domingos. Mining knowledge-sharing sites for viral marketing. In *International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 61–70. ACM, 2002.
- Alexander Schrijver. *Combinatorial Optimization*, volume 24 of *Algorithms and Combinatorics*. Springer, 2003.
- Tasuku Soma, Naonori Kakimura, Kazuhiro Inaba, and Ken-ichi Kawarabayashi. Optimal budget allocation: Theoretical guarantee and efficient algorithm. In *International Conference on Machine Learning (ICML)*, pages 351–359, 2014.
- Tao Sun, Wei Chen, Zhenming Liu, Yajun Wang, Xiaorui Sun, Ming Zhang, and Chin-Yew Lin. Participation maximization based on social influence in online discussion forums. In *International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2011.
- Maxim Sviridenko. A note on maximizing a submodular set function subject to knapsack constraint. *Operations Research Letters*, 32:41–43, 2004.

Appendix A. Naive Sampling Algorithm

The graphical model perspective described in Section 2.2 suggests a naive sampling (NS) algorithm for approximating $\sigma(\mathcal{A}, T)$:

1. Draw n samples, $\left\{ \left\{ \tau_{ji}^l \right\}_{(j,i) \in \mathcal{E}} \right\}_{l=1}^n$, *i.i.d.* from the waiting time product distribution $\prod_{(j,i) \in \mathcal{E}} f_{ji}(\tau_{ji})$;
2. For each sample $\left\{ \tau_{ji}^l \right\}_{(j,i) \in \mathcal{E}}$ and for each node i , find the shortest path from source nodes to node i ; count the number of nodes with $g_i \left(\left\{ \tau_{ji}^l \right\}_{(j,i) \in \mathcal{E}} \right) \leq T$;
3. Average the counts across n samples.

Although the naive sampling algorithm can handle arbitrary transmission function, it is not scalable to networks with millions of nodes. We need to compute the shortest path for each node and each sample, which results in a computational complexity of $O(n|\mathcal{E}| + n|\mathcal{V}| \log |\mathcal{V}|)$ for a single source node. The problem is even more pressing in the influence maximization problem, where we need to estimate the influence of source nodes at different location and with increasing number of source nodes. To do this, the algorithm needs to be repeated, adding a multiplicative factor of $C|\mathcal{V}|$ to the computational complexity (C is the number of nodes to select). Then, the algorithm becomes quadratic in the network size. When the network size is in the order of thousands and millions, typical in modern social network analysis, the naive sampling algorithm become prohibitively expensive. Additionally, we may need to draw thousands of samples (n is large), further making the algorithm impractical for large-scale problems.

Appendix B. Least Label List

The notation “ $\text{argsort}((r_1, \dots, r_{|\mathcal{V}|}), \text{ascend})$ ” in line 2 of Algorithm 3 means that we sort the collection of random labels in ascending order and return the argument of the sort as an ordered list.

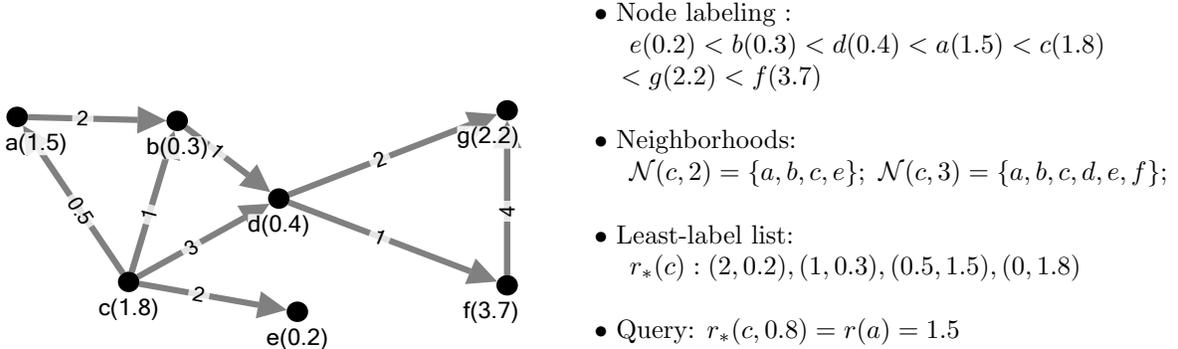


Figure 16: Graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, edge weights $\{\tau_{ji}\}_{(j,i) \in \mathcal{E}}$, and node labeling $\{r_i\}_{i \in \mathcal{V}}$ with the associated output from Algorithm 3.

Algorithm 3: Least Label List

Input: a reversed directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with edge weights $\{\tau_{ji}\}_{(j,i) \in \mathcal{E}}$, a node labeling $\{r_i\}_{i \in \mathcal{V}}$
Output: A list $r_*(s)$ for each $s \in \mathcal{V}$

```

1 for each  $s \in \mathcal{V}$  do  $d_s \leftarrow \infty, r_*(s) \leftarrow \emptyset$ ;
2 for  $i$  in  $\text{argsort}((r_1, \dots, r_{|\mathcal{V}|}), \text{ascend})$  do
3     empty heap  $H \leftarrow \emptyset$ ;
4     set all nodes except  $i$  as unvisited;
5     push  $(0, i)$  into heap  $H$ ;
6     while  $H \neq \emptyset$  do
7         pop  $(d_*, s)$  with the minimum  $d_*$  from  $H$ ;
8         add  $(d_*, r_i)$  to the end of list  $r_*(s)$ ;
9          $d_s \leftarrow d_*$ ;
10        for each unvisited out-neighbor  $j$  of  $s$  do
11            set  $j$  as visited;
12            if  $(d, j)$  in heap  $H$  then
13                Pop  $(d, j)$  from heap  $H$ ;
14                Push  $(\min\{d, d_* + \tau_{js}\}, j)$  into heap  $H$ ;
15            else if  $d_* + \tau_{js} < d_j$  then
16                Push  $(d_* + \tau_{js}, j)$  into heap  $H$ ;
```

Figure 16 shows an example of the Least-Label-List. The nodes from a to g are assigned to exponentially distributed labels with mean one shown in each parentheses. Given a query distance 0.8 for node c , we can binary-search its Least-label-list $r_*(c)$ to find that node a belongs to this range with the smallest label $r(a) = 1.5$.

Appendix C. Theorem 1

Theorem 1 *Sample the following number of sets of random transmission times*

$$n \geq \frac{C\Lambda}{\epsilon^2} \log \left(\frac{2|\mathcal{V}|}{\alpha} \right)$$

where $\Lambda := \max_{\mathcal{A}: |\mathcal{A}| \leq C} 2\sigma(\mathcal{A}, T)^2 / (m - 2) + 2\text{Var}(|\mathcal{N}(\mathcal{A}, T)|) / (m - 1) / (m - 2) + 2a\epsilon/3$, $|\mathcal{N}(\mathcal{A}, T)| \leq a$, and for each set of random transmission times, sample m set of random labels. Then we can guarantee that

$$|\hat{\sigma}(\mathcal{A}, T) - \sigma(\mathcal{A}, T)| \leq \epsilon$$

simultaneously for all \mathcal{A} with $|\mathcal{A}| \leq C$, with probability at least $1 - \alpha$.

Proof Let $S_\tau := |\mathcal{N}(\mathcal{A}, T)|$ for a fixed set of $\{\tau_{ji}\}$ and then $\sigma(\mathcal{A}, T) = \mathbb{E}_\tau[S_\tau]$. The randomized algorithm with m randomizations produces an unbiased estimator $\hat{S}_\tau = (m - 1) / (\sum_{u=1}^m r_*^u)$ for S_τ , i.e., $\mathbb{E}_{r|\tau}[\hat{S}_\tau] = S_\tau$, with variance $\mathbb{E}_{r|\tau}[(\hat{S}_\tau - S_\tau)^2] = S_\tau^2 / (m - 2)$ (Cohen, 1997).

Then \widehat{S}_τ is also an unbiased estimator for $\sigma(\mathcal{A}, T)$, since $\mathbb{E}_{\tau,r}[\widehat{S}_\tau] = \mathbb{E}_\tau \mathbb{E}_{r|\tau}[\widehat{S}_\tau] = \mathbb{E}_\tau[S_\tau] = \sigma(\mathcal{A}, T)$. Its variance is

$$\begin{aligned} \text{Var}(\widehat{S}_\tau) &:= \mathbb{E}_{\tau,r}[(\widehat{S}_\tau - \sigma(\mathcal{A}, T))^2] = \mathbb{E}_{\tau,r}[(\widehat{S}_\tau - S_\tau + S_\tau - \sigma(\mathcal{A}, T))^2] \\ &= \mathbb{E}_{\tau,r}[(\widehat{S}_\tau - S_\tau)^2] + 2 \mathbb{E}_{\tau,r}[(\widehat{S}_\tau - S_\tau)(S_\tau - \sigma(\mathcal{A}, T))] + \mathbb{E}_{\tau,r}[(S_\tau - \sigma(\mathcal{A}, T))^2] \\ &= \mathbb{E}_\tau[S_\tau^2/(m-2)] + 0 + \text{Var}(S_\tau) \\ &= \sigma(\mathcal{A}, T)^2/(m-2) + \text{Var}(S_\tau)(m-1)/(m-2) \end{aligned}$$

Then using Bernstein's inequality, we have, for our final estimator $\widehat{\sigma}(\mathcal{A}, T) = \frac{1}{n} \sum_{l=1}^n \widehat{S}_{\tau^l}$, that

$$\Pr \{ |\widehat{\sigma}(\mathcal{A}, T) - \sigma(\mathcal{A}, T)| \geq \epsilon \} \leq 2 \exp \left(- \frac{n\epsilon^2}{2\text{Var}(\widehat{S}_\tau) + 2a\epsilon/3} \right) \quad (21)$$

where $\widehat{S}_\tau < a \leq |\mathcal{V}|$.

Setting the right hand side of relation (21) to α , we have that, with probability $1 - \alpha$, sampling the following number sets of random transmission times

$$\begin{aligned} n &\geq \frac{2\text{Var}(\widehat{S}_\tau) + 2a\epsilon/3}{\epsilon^2} \log \left(\frac{2}{\alpha} \right) \\ &= \frac{2\sigma(\mathcal{A}, T)^2/(m-2) + 2\text{Var}(S_\tau)(m-1)/(m-2) + 2a\epsilon/3}{\epsilon^2} \log \left(\frac{2}{\alpha} \right) \end{aligned}$$

we can guarantee that our estimator to have error $|\widehat{\sigma}(\mathcal{A}, T) - \sigma(\mathcal{A}, T)| \leq \epsilon$.

If we want to insure that $|\widehat{\sigma}(\mathcal{A}, T) - \sigma(\mathcal{A}, T)| \leq \epsilon$ simultaneously hold for all \mathcal{A} such that $|\mathcal{A}| \leq C \ll |\mathcal{V}|$, we can first use union bound with relation (21). In this case, we have that, with probability $1 - \alpha$, sampling the following number sets of random transmission times

$$n \geq \frac{C\Lambda}{\epsilon^2} \log \left(\frac{2|\mathcal{V}|}{\alpha} \right)$$

we can guarantee that our estimator to have error $|\widehat{\sigma}(\mathcal{A}, T) - \sigma(\mathcal{A}, T)| \leq \epsilon$ for all \mathcal{A} with $|\mathcal{A}| \leq C$. Note that we have define the constant $\Lambda := \max_{\mathcal{A}: |\mathcal{A}| \leq C} 2\sigma(\mathcal{A}, T)^2/(m-2) + 2\text{Var}(S_\tau)(m-1)/(m-2) + 2a\epsilon/3$. \blacksquare

Appendix D. Complete Proofs for Section 4

D.1 Uniform Cost

In this section, we first prove a theorem for the general problem defined by Equation (19), considering a normalized monotonic submodular function $f(S)$ and general P (Theorem 7) and $k = 0$, and then obtain the guarantee for our influence maximization problem (Theorem 4).

Suppose $G = \{g_1, \dots, g_{|G|}\}$ in the order of selection, and let $G^t = \{g_1, \dots, g_t\}$. Let C_t denote all those elements in $O \setminus G$ that satisfy the following: they are still feasible before selecting the t -th element g_t but are infeasible after selecting g_t . Formally,

$$C_t = \{z \in O \setminus G : \{z\} \cup G^{t-1} \in \mathcal{F}, \{z\} \cup G^t \notin \mathcal{F}\}.$$

In the following, we will prove three claims and then use them to prove the Theorems 7 and 4. Recall that for any $i \in \mathcal{Z}$ and $S \subseteq \mathcal{Z}$, the marginal gain of z with respect to S is denoted as

$$f(z|S) := f(S \cup \{z\}) - f(S)$$

and its approximation is denoted by

$$\widehat{f}(z|S) = \widehat{f}(S \cup \{z\}) - \widehat{f}(S).$$

Also, when $|f(S) - \widehat{f}(S)| \leq \epsilon$ for any $S \subseteq \mathcal{Z}$, we have

$$|\widehat{f}(z|S) - f(z|S)| \leq 2\epsilon$$

for any $z \in \mathcal{Z}$ and $S \subseteq \mathcal{Z}$.

Claim 1. $\sum_{i=1}^t |C_i| \leq Pt$, for $t = 1, \dots, |G|$.

Proof We first show the following property about matroids: for any $Q \subseteq \mathcal{Z}$, the sizes of any two maximal independent subsets T_1 and T_2 of Q can only differ by a multiplicative factor at most P . Here, T is a maximal independent subset of Q if and only if:

- $T \subseteq Q$;
- $T \in \mathcal{F} = \bigcap_{i=1}^P \mathcal{I}_i$;
- $T \cup \{z\} \notin \mathcal{F}$ for any $z \in Q \setminus T$.

To prove the property, note that for any element $z \in T_1 \setminus T_2$, $\{z\} \cup T_2$ violates at least one of the matroid constraints since T_2 is maximal. Let $\{V_i\}_{1 \leq i \leq P}$ denote all elements in $T_1 \setminus T_2$ that violate the i -th matroid, and then partition $T_1 \cap T_2$ using these V_i 's so that they cover T_1 . Note that the size of each V_i must be at most that of T_2 , since otherwise by the Exchange axiom, there would exist $z \in V_i \setminus T_2$ that can be added to T_2 without violating the i -th matroid, leading to a contradiction. Therefore, $|T_1|$ is at most P times $|T_2|$.

Now we apply the property to prove the claim. Let Q be the union of G^t and $\bigcup_{i=1}^t C_i$. On one hand, G^t is a maximal independent subset of Q , since no element in $\bigcup_{i=1}^t C_i$ can be added to G^t without violating the matroid constraints. On the other hand, $\bigcup_{i=1}^t C_i$ is an independent subset of Q , since it is part of the optimal solution. Therefore, $\bigcup_{i=1}^t C_i$ has size at most P times $|G^t|$, which is Pt . \blacksquare

Claim 2. Suppose g_t is selected at the threshold τ_t . Then, $f(j|G^{t-1}) \leq (1 + \delta)\tau_t + 4\epsilon + \frac{\delta}{N}f(G)$, $\forall j \in C_t$.

Proof First, consider $\tau_t > w_{L+1} = 0$. Since g_t is selected at the threshold τ_t , we have that $\widehat{f}(g_t|G^{t-1}) \geq \tau_t$ and thus $f(g_t|G^{t-1}) \geq \tau_t - 2\epsilon$. Any $j \in C_t$ could have been selected at an

earlier stage, since adding j to G^{t-1} would not have violated the constraint. However, since $j \notin G^{t-1}$, that means that $\widehat{f}(j|G^{t-1}) \leq (1 + \delta)\tau_t$. Then,

$$f(j|G^{t-1}) \leq (1 + \delta)\tau_t + 2\epsilon.$$

Second, consider $\tau_t = w_{L+1} = 0$. For each $j \in C_t$, we have $\widehat{f}(j|G) < \frac{\delta}{N}d$. Since the greedy algorithm must pick g_1 with $\widehat{f}(g_1) = d$ and $d \leq f(g_1) + \epsilon$, then

$$f(j|G) < \frac{\delta}{N}f(G) + 4\epsilon.$$

The claim follows by combining the two cases. \blacksquare

Claim 3. *The marginal gain of $O \setminus G$ satisfies*

$$\sum_{j \in O \setminus G} f(j|G) \leq [(1 + \delta)P + \delta]f(G) + (6 + 2\delta)\epsilon P|G|.$$

Proof Combining Claim 1 and Claim 2, we have:

$$\begin{aligned} \sum_{j \in O \setminus G} f(j|G) &= \sum_{t=1}^{|G|} \sum_{j \in C_t} f(j|G) \leq (1 + \delta) \sum_{t=1}^{|G|} |C_t| \tau_t + \delta f(G) + 4\epsilon \sum_{t=1}^{|G|} |C_t| \\ &\leq (1 + \delta) \sum_{t=1}^{|G|} |C_t| \tau_t + \delta f(G) + 4\epsilon P|G|. \end{aligned}$$

Further, $\sum_{t=1}^{|G|} |C_t| \tau_t \leq P \sum_{t=1}^{|G|} \tau_t$ by Claim 1 and a technical lemma (Lemma 6). Finally, the claim follows from the fact that $f(G) = \sum_t f(g_t|G^{t-1}) \geq \sum_t (\tau_t - 2\epsilon)$. \blacksquare

Lemma 6 *If $\sum_{i=1}^t \sigma_{i-1} \leq t$ for $t = 1, \dots, K$ and $\rho_{i-1} \geq \rho_i$ for $i = 1, \dots, K - 1$ with $\rho_i, \sigma_i \geq 0$, then $\sum_{i=1}^K \rho_i \sigma_i \leq \sum_{i=1}^K \rho_{i-1}$.*

Proof Consider the linear program

$$\begin{aligned} V &= \max_{\sigma} \sum_{i=1}^K \rho_i \sigma_i \\ \text{s.t.} \quad &\sum_{i=1}^t \sigma_{i-1} \leq t, \quad t = 1, \dots, K, \\ &\sigma_i \geq 0, \quad i = 1, \dots, K - 1 \end{aligned}$$

with dual

$$\begin{aligned} W &= \min_u \sum_{i=1}^K t u_{t-1} \\ \text{s.t.} \quad &\sum_{t=i}^{K-1} u_t \geq \rho_i, \quad i = 0, \dots, K - 1, \\ &u_t \geq 0, \quad t = 0, \dots, K - 1. \end{aligned}$$

As $\rho_i \geq \rho_{i+1}$, the solution $u_i = \rho_i - \rho_{i+1}, i = 0, \dots, K - 1$ (where $\rho_K = 0$) is dual feasible with value $\sum_{t=1}^K t(\rho_{t-1} - \rho_t) = \sum_{i=1}^K \rho_{i-1}$. By weak linear programming duality, $\sum_{i=1}^K \rho_i \sigma_i \leq V \leq W \leq \sum_{i=1}^K \rho_{i-1}$. ■

Theorem 7 *Suppose we use Algorithm 2 to solve the problem defined by Equation (19) with $k = 0$, using $\rho = 0$ and \hat{f} to estimate the function f , where $|\hat{f}(S) - f(S)| \leq \epsilon$ for all $S \subseteq \mathcal{Z}$. It holds that the algorithm returns a greedy solution G with*

$$f(G) \geq \frac{1}{(1 + 2\delta)(P + 1)} f(O) - \frac{4P|G|}{P + c_f} \epsilon$$

where O is the optimal solution, using $\mathcal{O}(\frac{N}{\delta} \log \frac{N}{\delta})$ evaluations of \hat{f} .

Proof By submodularity and Claim 3, we have:

$$f(O) \leq f(O \cup G) \leq f(G) + \sum_{j \in O \setminus G} f(j|G) \leq (1 + \delta)(P + 1)f(G) + (6 + 2\delta)\epsilon P|G|,$$

which leads to the bound in the theorem.

Since there are $\mathcal{O}(\frac{1}{\delta} \log \frac{N}{\delta})$ thresholds, and there are $\mathcal{O}(N)$ evaluations at each threshold, the number of evaluations is bounded by $\mathcal{O}(\frac{N}{\delta} \log \frac{N}{\delta})$. ■

Theorem 7 essentially shows $f(G)$ is close to $f(O)$ up to a factor roughly $(1 + P)$, which then leads to the following guarantee for our influence maximization problem. Suppose product $i \in \mathcal{L}$ spreads according to diffusion network $\mathcal{G}_i = (\mathcal{V}, \mathcal{E}_i)$, and let $i^* = \operatorname{argmax}_{i \in \mathcal{L}} |\mathcal{E}_i|$.

Theorem 4. *In the influence maximization problem with uniform cost, Algorithm 2 (with $\rho = 0$) is able to output a solution G that satisfies $f(G) \geq \frac{1-2\delta}{3} f(O)$ in expected time $\tilde{\mathcal{O}}\left(\frac{|\mathcal{E}_{i^*}| + |\mathcal{V}|}{\delta^2} + \frac{|\mathcal{L}||\mathcal{V}|}{\delta^3}\right)$.*

Proof In the influence maximization problem, the number of matroids is $P = 2$. Also note that $|G| \leq f(G) \leq f(O)$, which leads to $4|G|\epsilon \leq 4\epsilon f(O)$. The approximation guarantee then follows from setting $\epsilon \leq \delta/16$ when using CONTINEST (Du et al., 2013a) to estimate the influence.

The runtime is bounded as follows. In Algorithm 2, we need to estimate the marginal gain of adding one more product to the current solution. In CONTINEST (Du et al., 2013a), building the initial data structure takes time

$$\mathcal{O}\left((|\mathcal{E}_{i^*}| \log |\mathcal{V}| + |\mathcal{V}| \log^2 |\mathcal{V}|) \frac{1}{\delta^2} \log \frac{|\mathcal{V}|}{\delta}\right)$$

and afterwards each function evaluation takes time

$$\mathcal{O}\left(\frac{1}{\delta^2} \log \frac{|\mathcal{V}|}{\delta} \log \log |\mathcal{V}|\right).$$

As there are $\mathcal{O}(\frac{N}{\delta} \log \frac{N}{\delta})$ evaluations where $N = |\mathcal{L}||\mathcal{V}|$, the runtime of our algorithm follows. ■

D.2 General case

As in the previous section, we first prove that a theorem for the problem defined by Equation (19) with general normalized monotonic submodular function $f(S)$ and general P (Theorem 8), and then obtain the guarantee for our influence maximization problem (Theorem 5).

Theorem 8 *Suppose Algorithm 1 uses \hat{f} to estimate the function f which satisfies $|\hat{f}(S) - f(S)| \leq \epsilon$ for all $S \subseteq \mathcal{Z}$. Then, there exists a ρ such that*

$$f(S_\rho) \geq \frac{\max\{1, |A_\rho|\}}{(P + 2k + 1)(1 + 2\delta)} f(O) - 8\epsilon |S_\rho|$$

where A_ρ is the set of active knapsack constraints:

$$A_\rho = \{i : S_\rho \cup \{z\} \notin \mathcal{F}, \forall z \in \mathcal{Z}_{i^*}\}.$$

Proof Consider the optimal solution O and set $\rho^* = \frac{2}{P+2k+1} f(O)$. By submodularity, we have $d \leq f(O) \leq |\mathcal{Z}|d$, so $\rho \in \left[\frac{2d}{P+2k+1}, \frac{2|\mathcal{Z}|d}{P+2k+1}\right]$, and there is a run of Algorithm 2 with ρ such that $\rho^* \in [\rho, (1 + \delta)\rho]$. In the following we consider this run.

Case 1 Suppose $|A_\rho| = 0$. The key observation in this case is that since no knapsack constraints are active, the algorithm runs as if there were only matroid constraints. Then, the argument for matroid constraints can be applied. More precisely, let

$$O_+ := \{z \in O \setminus S_\rho : f(z|S_\rho) \geq c(z)\rho + 2\epsilon\}$$

$$O_- := \{z \in O \setminus S_\rho : z \notin O_+\}.$$

Note that all elements in O_+ are feasible. Following the argument of Claim 3 in Theorem 7, we have:

$$f(O_+|S_\rho) \leq ((1 + \delta)P + \delta)f(S_\rho) + (4 + 2\delta)\epsilon P|S_\rho|. \quad (22)$$

Also, by definition, the marginal gain of O_- is:

$$f(O_-|S_\rho) \leq k\rho + 2\epsilon|O_-| \leq k\rho + 2\epsilon P|S_\rho|, \quad (23)$$

where the last inequality follows from the fact that S_ρ is a maximal independent subset, O_- is an independent subset of $O \cup S_\rho$, and the sizes of any two maximal independent subsets in the intersection of P matroids can differ by a factor of at most P . Plugging (22)(23) into $f(O) \leq f(O_+|S_\rho) + f(O_-|S_\rho) + f(S_\rho)$, we obtain the bound

$$f(S_\rho) \geq \frac{f(O)}{(P + 2k + 1)(1 + \delta)} - \frac{(6 + 2\delta)\epsilon P|S_\rho|}{(P + 1)(1 + \delta)}.$$

Case 2 Suppose $|A_\rho| > 0$. For any $i \in A_\rho$ (i.e., the i -th knapsack constraint is active), consider the step when i is added to A_ρ . Let $G_i = G \cap \mathcal{Z}_{i^*}$, and we have $c(G_i) + c(z) > 1$.

Since every element g we include in G_i satisfies $\widehat{f}(g|G) \geq c(g)\rho$ with respect to the solution G_i when g is added. Then $f(g|G) = f_i(g|G_i) \geq c(g)\rho - 2\epsilon$, and we have:

$$f_i(G_i \cup \{z\}) \geq \rho[c(G_i) + c(z)] - 2\epsilon(|G_i| + 1) > \rho - 2\epsilon(|G_i| + 1). \quad (24)$$

Note that G_i is non-empty since otherwise the knapsack constraint will not be active. Any element in G_i is selected before or at w_t , so $f_i(G_i) \geq w_t - 2\epsilon$. Also, note that z is not selected in previous thresholds before w_t , so $f_i(\{z\} | G_i) \leq (1 + \delta)w_t + 2\epsilon$ and thus,

$$f_i(\{z\} | G_i) \leq (1 + \delta)f_i(G_i) + 2\epsilon(2 + \delta). \quad (25)$$

Combining Eqs. 24 and 25 into $f_i(G_i \cup \{z\}) = f_i(G_i) + f_i(\{z\} | G_i)$ leads to

$$\begin{aligned} f_i(G_i) &\geq \frac{\rho}{(2 + \delta)} - \frac{2\epsilon(|G_i| + 3 + \delta)}{(2 + \delta)} \geq \frac{1}{2(1 + 2\delta)}\rho^* - \frac{2\epsilon(|G_i| + 3 + \delta)}{(2 + \delta)} \\ &\geq \frac{f(O)}{(P + 2k + 1)(1 + 2\delta)} - 5\epsilon|G_i|. \end{aligned}$$

Summing up over all $i \in A_\rho$ leads to the desired bound. ■

Suppose item $i \in \mathcal{L}$ spreads according to the diffusion network $\mathcal{G}_i = (\mathcal{V}, \mathcal{E}_i)$. Let $i^* = \operatorname{argmax}_{i \in \mathcal{L}} |\mathcal{E}_i|$. By setting $\epsilon = \delta/16$ in Theorem 8, we have:

Theorem 5. *In Algorithm 1, there exists a ρ such that*

$$f(S_\rho) \geq \frac{\max\{k_a, 1\}}{(2|\mathcal{L}| + 2)(1 + 3\delta)} f(O)$$

where k_a is the number of active knapsack constraints. The expected runtime to obtain the solution is $\tilde{O}\left(\frac{|\mathcal{E}_{i^*}| + |\mathcal{V}|}{\delta^2} + \frac{|\mathcal{L}||\mathcal{V}|}{\delta^4}\right)$.