End-to-End Influence Maximization in the Field

Bryan Wilder, Laura Onasch-Vera, Juliana Hudson, Jose Luna, Nicole Wilson, Robin Petering,
Darlene Woo, Milind Tambe, Eric Rice
Center for Artificial Intelligence in Society, University of Southern California
{bwilder,onaschve,jnhudson,joseluna,wilsonnj,petering,darlene,tambe,ericr}@usc.edu

ABSTRACT

This work is aims to overcome the challenges in deploying influence maximization to support community driven interventions. Influence maximization is a crucial technique used in preventative health interventions, such as HIV prevention amongst homeless youth. Drop-in centers for homeless youth train a subset of youth as peer leaders who will disseminate information about HIV through their social networks. The challenge is to find a small set of peer leaders who will have the greatest possible influence. While many algorithms have been proposed for influence maximization, none can be feasibly deployed by a service provider: existing algorithms require costly surveys of the entire social network of the youth to provide input data, and high performance computing resources to run the algorithm itself. Both requirements are crucial bottlenecks to widespread use of influence maximization in real world interventions.

To address the above challenges, this innovative applications paper introduces the CHANGE agent for influence maximization. CHANGE handles the end-to-end process of influence maximization, from data collection to peer leader selection. Crucially, CHANGE only surveys a fraction of the youth to gather network data and minimizes computational cost while providing comparable performance to previously proposed algorithms. We carried out a pilot study of CHANGE in collaboration with a drop-in center serving homeless youth in a major U.S. city. CHANGE surveyed only 18% of the youth to construct its social network. However, the peer leaders it selected reached just as many youth as previously field-tested algorithms which surveyed the *entire* network. This is the first real-world study of a network sampling algorithm for influence maximization. Simulation results on real-world networks also support our claims.

KEYWORDS

Innovative applications; influence maximization; pilot study

1 INTRODUCTION

This innovative applications paper presents and field-tests a novel, practical agent for influence maximization, the challenge of selecting a small set of seed nodes in a social network who will diffuse information to many others. Such techniques have important applications ranging from preventative health [1, 23] to international development [2]. It is inherently a multiagent problem because nodes (agents) make decisions in response to those around them [17, 29].

In this paper, we are particularly motivated by the challenge of preventing HIV spread among homeless youth [19, 20, 28] (although our contributions would also assist other public health interventions). Here, influence maximization is used to select homeless youth who will serve as *peer leaders* and spread messages about HIV prevention through their social network. Pilot studies in this domain have shown that algorithmic approaches have great promise, substantially outperforming status-quo heuristics [27]. However, current algorithms have a high barrier to entry: they require a great deal of time to gather the complete social network of the youth, expertise to select appropriate parameters, and computational power to run the algorithms. None of these are likely available to resource-strained service providers who will ultimately be the ones to deploy influence maximization.

Gathering network data is particularly onerous because it requires individually surveying upwards of a hundred youth. Further, network collection is more time intensive than simple survey methods, requiring days of time for a dedicated team of social work researchers. It is not feasible for service providers with many other responsibilities.

The other barriers are also serious impediments to wide-scale adoption of influence maximization. Service providers will not have access to the high-performance computing resources required by previous algorithms, where high computational cost is often incurred to find solutions robust to unknown parameters. For instance, DOSIM, the state of the art algorithm for robust influence maximization [25], takes hours of runtime on a high-performance computing system. In reality, a deployed system would need to run in minutes on a laptop.

This paper addresses the challenge of deployed influence maximization. We present CHANGE (CompreHensive Adaptive Network samplinG for social influencE), a novel, end-to-end agent for influence maximization which addresses the above barriers via a set of algorithmic contributions. CHANGE is easy to deploy, but this simplicity is crucially enabled by a series of insights into the social structure of homeless youth (which may be useful for other vulnerable populations). We conducted a pilot test of CHANGE's performance in a real deployment by a drop-in center serving homeless youth in a major U.S. city. CHANGE was used to plan a series of interventions designed to spread HIV awareness among the youth. *CHANGE obtained comparable influence spread to state of the art algorithms while surveying only 18% of nodes for network data*, a finding which is backed by additional simulation results.

Overall, CHANGE offers a practical, field-tested vehicle for deployed influence maximization which drastically lowers the barrier to entry. *To our knowledge, this is the first real-world pilot study of a network sampling algorithm for influence maximization and only the second ever field test of* any *influence maximization algorithm.*

Overview of algorithmic contributions: We now summarize how CHANGE handles the challenges above. We discuss related work in Section 2; however, none addresses these challenges.

First, to address the *data gathering* challenge, we present an easily deployable sampling protocol which randomly selects a small set of youth to interview. For each of these youth, a randomly chosen

Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018), M. Dastani, G. Sukthankar, E. Andre, S. Koenig (eds.), July 2018, Stockholm, Sweden

https://doi.org/doi

neighbor is also interviewed. We show that this procedure gathers enough of the network to enable high-quality influence maximization even though it surveys only a small number of nodes directly.

Second, to address *computational power* challenge (which in turn stems from unknown parameters), we present a heuristic for selecting influence maximization solutions which are robust to uncertainty in the probability p that influence will spread. We show that this heuristic finds solutions which obtain approximately 90% of the maximum possible influence spread under *any* value for p. Importantly, this heuristic runs in minutes on a laptop, while DOSIM (the previously proposed algorithm for this problem) requires hours or even days of time on a high performance computing cluster.

Third, we integrate these components with an *adaptive greedy* algorithm for planning interventions and prove the first theoretical guarantee for influence maximization under execution errors. The challenge is that some youth selected as peer leaders may not attend the intervention [25, 27]. Our algorithm selects its action with such uncertainties in mind, observes which youth do attend, and then plans the next round using this observation. We prove that it obtains a constant-factor approximation to the *optimal* adaptive policy.

Overview of field deployment contributions: We conducted two pilot studies of CHANGE, each addressing distinct questions.

First, we conducted a *feasibility study* with two objectives. (i) We confirm that CHANGE's mechanism for sampling the network to gather edge data is implementable with a homeless youth population. This is nontrivial because homeless youth are often difficult to locate, making finding particular youth to query for network ties difficult. (ii) We validate that the data gathered is sufficiently accurate to enable influence maximization. Self-reported ties are subject to bias and forgetting [3], making it important to investigate whether they are accurate enough to find influential nodes. This point is of broader interest, since previous work on influence maximization has largely used self-reported ties [25, 27], but *no previous field study has validated their accuracy for influence maximization*. To address these questions we collected network data from 72 youth at a drop-in center via a range of methods: CHANGE's sampling mechanism, self-reports from the entire network, field observations by research staff, and interviews with staff members. Our results show that CHANGE's sampling mechanism is feasible, and that self-reported data is sufficient to enable high-quality influence maximization.

Second, we conduct an *intervention study* of the entire CHANGE agent with an additional set of 64 homeless youth. This includes network data collection, peer leader selection, and HIV awareness trainings for the selected peer leaders. We then conducted a follow-up survey to assess how many youth received information about HIV. While CHANGE only collected data from 18% of youth in the network, the peer leaders that it selects successfully reached 80% of the youth. This is comparable to previously tested algorithms HEALER and DOSIM which gather the *entire* network. This result provides evidence that CHANGE can obtain influence spread comparable to the highly sophisticated algorithms proposed by previous work, while eliminating crucial barriers to real world deployment.

Third, we give an analysis of the real network data to explain why CHANGE can succeed while gathering such a small portion of the network. Our explanation draws on *friendship paradox*, a phenomenon observed in social networks where a typical node's neighbors have more ties than the node itself. We demonstrate this phenomenon occurs across both of the networks that we gathered and show how CHANGE exploits it to produce sampled networks which are substantially more informative for influence maximization than a comparable number of uniformly random samples.

2 RELATED WORK

The influence maximization problem was introduced by Kempe et al. [13], and has been extensively studied since then [4, 5, 8, 9, 12, 15, 18, 21]. Most of this literature has focused on algorithms which are scalable to extremely large networks, primarily in the context of online viral marketing. Recently, HIV prevention (and preventative health more broadly) has emerged as a new application area for influence maximization which brings its own set of research challenges. Yadav et al. [26] proposed HEALER, a POMDP-based algorithm for selecting influential peer leaders. Subsequently, Wilder et al. [25] introduced the DOSIM algorithm which accounts for uncertainty about the true probability of influence propagation using a robust optimization approach. We note that our approach to parameter robustness is similar to techniques used in robust MDP planning [14], though the domains are entirely different.

Yadav et al. [27] conducted a real-world pilot study of HEALER and DOSIM, and found that both algorithms significantly outperformed the status-quo heuristic used by agencies (selecting high-degree nodes). However, neither algorithm addresses any of the challenges described above. Both assume that the entire social network is provided as input, which is unrealistic in practice due to the enormous effort required. Further, only DOSIM handles uncertainty about the probability of influence spread, and its method for doing so is extremely computationally intensive, far beyond the reach of an average service provider (see Section 4.3 for further discussion). Separate work by Wilder et al. [24] considered network data collection. They proposed the ARISEN algorithm which samples a portion of youth in the network to collect data from. While ARISEN can be theoretically analyzed for certain network structures, it is not practically suitable to deployment because it relies on querying a sequence of specific youth who may be difficult to locate (see Section 4.2). *Thus, ARISEN has never been tested in a field study.* Moreover, ARISEN does not consider either parameter uncertainty or execution errors (the possibility that some peer leaders will not attend), both of which we incorporate into CHANGE.

3 PROBLEM DESCRIPTION

Motivating domain: Our work is designed to overcome the challenges in deploying influence maximization techniques to support community-driven interventions. We are specifically motivated by the challenge of raising awareness about HIV among homeless youth. Typically, an HIV awareness intervention will be provided by a drop in center or other organization which serves homeless youth. Each intervention is a day-long class followed by weekly hour-long meetings. Hence (as is typical in many intervention domains), the service provider will almost never have enough resources to deliver the intervention to all of the youth that frequent the center; instead, the intervention is usually delivered to 15-20% of the population[1]. Further, limitations on space and personnel mean that the intervention

[1] Note that while CHANGE directly surveys ∼18% of youth, they name others as friends, resulting in a larger sampled graph.

can typically be delivered to only 4-6 youth at a given time, so the training is broken up over a series of small sessions. These youth are trained as *peer leaders* who communicate with other youth about HIV prevention. This amplifies the reach of the intervention through the social network of the homeless youth. The question is which youth will make the most effective peer leaders, able to reach the greatest number of their peers. This is an *influence maximization* problem, which we now formalize.

Influence: The youth have a social network represented as a graph $G = (V, E)$. Each youth is initially inactive, meaning that they have not received information about HIV prevention. Once nodes are activated by the intervention, they have a chance to influence their peers. We model this process through a variant on the classical independent cascade model (ICM) which has been used by previous work on HIV prevention and better reflects realistic time dynamics [25–27]. The process unfolds over discrete time steps $t = 1...T$, where T is a time horizon. There is a propagation probability p. When a node becomes active, it attempts to activate each of its neighbors. Each attempt succeeds independently with probability p. Activation attempts are made at each time step until either the neighbor is influenced or the time horizon is reached.

Note that the assumption that p is uniform across edges is without much loss. As noted by He and Kempe [10], a uniform p is equivalent to each edge drawing an individual propagation probability i.i.d. from a distribution with mean p. This is because the following processes are analytically equivalent: (1) propagate influence with probability p and (2) draw a propagation probability q from a distribution with $\mathbb{E}[q] = p$ and then propagate influence with probability q. Hence, our model actually subsumes *any* stochastic model where the propagation probabilities are drawn from a common prior.

Interventions: At each time step $t = 1...T$, the algorithm selects a seed set A_t containing up to K nodes. However, each seed node may or may not actually attend the intervention. This problem is particularly acute with homeless youth since a number of factors could prevent a given youth from attending (e.g., being arrested, running out of money for a bus ticket, etc.). Hence, we assume that each node v has a hidden state $x_v \in \{present, absent\}$. Each node’s state is drawn independently from some prior distribution D. For simplicity, we will take D to set each node to be present with probability q. However, all of our analysis applies to arbitrary distributions. For each $v \in A_t$, if $x_v = present$, then v is activated. Nothing occurs if $x_v = absent$. Note that an absent node can still become activated by others, since they may still be in contact with others in the social network. After the set A_t is chosen, the intervention occurs and the hidden state of each $v \in A_t$ is observed. We denote the set of all observations received at time t as O_t.

The algorithm may use this information to plan the next intervention. In other words, the problem is *adaptive*. To model adaptivity, we introduce the notion of a *policy*. A policy maps from past actions and observations to the action that should be taken next. Let $\mathcal{A} = \{S \subseteq V : |S| \leq K\}$ be the set of all possible actions. A *history* is the current sequence of actions chosen and observations received, denoted by $\psi_t = ((A_1, O_1), (A_2, O_2), ...(A_t, O_t))$. Let Ψ be the set of all possible histories. A policy is a mapping $\pi : \Psi \rightarrow \mathcal{A}$. Let $A(\psi_t) = (A_1...A_t)$ be the sequence of actions taken and $O(\psi) = (O_1...O_t)$ be the corresponding observations (whether each peer leader was present or absent). Recall that youth are trained in groups of 4-6; the policy selects a group of youth to invite given who was trained previously. We denote the objective as $f(A(\psi)|O(\psi))$. f is the expected number of nodes influenced by the seed nodes in $A(\psi)$ conditioned on the observations in $O(\psi)$. We overload notation and let $f(\pi) = \mathbb{E}_{\psi\sim\pi}[f(A(\psi)|O(\psi))]$ be the expected reward from running policy π, where the expectation ranges over the hidden state $\boldsymbol{x}$ (which determines π’s actions) as well as the influence process. Our goal is to find a policy maximizing $f(\pi)$.

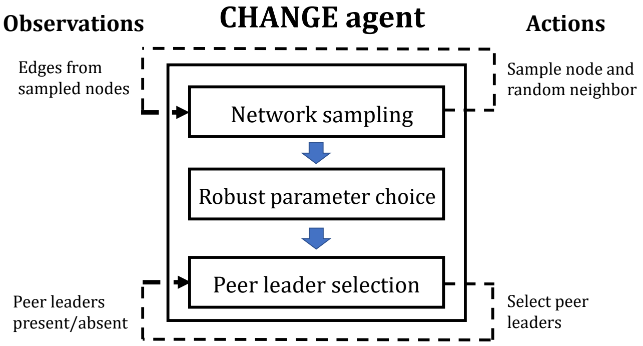

Figure 1: Illustration of the CHANGE agent.

Uncertainty about network structure and parameters: We consider extensions to the core adaptive influence maximization problem which account for the lack of information endemic in field deployments. First, we consider the case where the structure of the network (the edges E) are unknown. To address this challenge, we give our agent a budget of M queries to run before conducting the intervention. Each query may target either a uniformly random node, or the neighbor of a node already queried. When a node is queried, it reveals all of its edges. The goal is to use the M queries to uncover a set of edges which suffice to identify influential nodes.

We then consider an unknown propagation probability. Here, we take a robust optimization approach and look for a policy which performs well across a range of possible values for p. More detail on this part of the problem can be found in Section 4.3.

4 CHANGE: A NEW AGENT FOR INFLUENCE MAXIMIZATION IN THE FIELD

We now introduce the CHANGE agent for end-to-end influence maximization. Figure 1 illustrates the three components of the agent. We start with the last component, peer leader selection, since the other components exist to provide the data that the peer leader selection algorithm requires. Peer leader selection is performed by an adaptive greedy algorithm (Algorithm 1), which handles the chance that some peer leaders may not attend the intervention and plans solutions using the observations obtained so far. Algorithm 1 requires as input a (sample of) social network and a propagation probability p. We subsequently introduce Algorithms 2 and 3 to provide these inputs.

4.1 Adaptive greedy planning

Given as input the graph G and propagation probability p, finding the optimal policy is a difficult planning problem. There are 2^n possible hidden states and $\binom{n}{K}$ possible actions. While it is possible to formulate the problem as a POMDP, these exponentially large

Algorithm 1 Adaptive greedy

```
1: for $t = 1...T$ do
2:     $A_t = \emptyset$
3:     for $k = 1...K$ do //greedily select seeds for action $t$
4:         $v = \arg\max_{v \in V} \Delta(A_t \cup \{v\}|\psi_{t-1}) - \Delta(A_t|\psi_{t-1})$
5:         $A_t = A_t \cup \{v\}$
6:     execute $A_t$ and observe $O_t$
7:     $\psi_t = \psi_{t-1} + (A_t, O_t)$ //add action/observation to history
```

Algorithm 2 Network sampling

```
1: input: vertex set $V$, budget $M$
2: $E = \emptyset$ //set of edges observed
3: $S = \emptyset$ //set of nodes surveyed
4: for $i = 1...\frac{M}{2}$ do
5:     Sample $v$ uniformly at random from $V \setminus S$
6:     $S = S \cup \{v\}$
7:     $E = E \cup \{(v, u) : u \in N(v)\}$
8:     Sample $u$ uniformly at random from $N(v) \setminus S$
9:     $E = E \cup \{(u, w) : w \in N(u)\}$
10:    $S = S \cup \{u\}$
11: return $E$
```

state and action spaces place even small instances beyond the reach of off-the-self solvers. Hence, we exploit the structure of the problem to formulate a scalable greedy algorithm which obtains (provably) near-optimal solutions.

Pseudocode for adaptive greedy, our online planning algorithm, can be found in Algorithm 1. Algorithm 1 selects the action at each step which maximizes the expected gain in influence spread, conditioned on the observations received so far. Then, it waits until this action has been executed, observes which peer leaders attended the intervention, and greedily plans the next step. Formally, let $\Delta(A_t|\psi_{t-1}) = f(A(\psi_{t-1}) \cup A_t|O(\psi_{t-1})) - f(A(\psi_{t-1})|O(\psi_{t-1}))$ denote the expected marginal gain to selecting A_t at time t. The greedy policy is to select $A_t = \arg\max_{|A| \leq K} \Delta(A|\psi_{t-1})$ (the outer loop of Algorithm 1). However, computing the maximizing action is itself computationally intractable (as there are $\binom{n}{K}$ possible choices). Hence, Algorithm 1 uses an additional greedy inner loop which greedily selects the elements of A_t one at a time (lines 3-5). Note that Δ can be computed by averaging over random simulations over both the hidden state (which nodes are present/absent) as well as how influence spreads via the ICM.

We prove the following theorem, which shows that greedy planning is sufficient to obtain a guaranteed approximation ratio:

THEOREM 4.1. *Let π_G be Algorithm 1's greedy policy and π_* be an optimal policy. It holds that $f(\pi_G) \geq \left(\frac{e-1}{2e-1}\right) f(\pi_*)$.*

A proof may be found in the supplemental material[2]. We use the *adaptive submodularity* framework of Golovin and Krause, which generalizes the classical notion of a submodular set function to adaptive policies. Their framework does not straightforwardly apply to our problem since our algorithm selects a *sequence* of actions, not a set. The order in which actions are selected matters since peer leaders who are selected earlier will have more time to influence others. We show that our problem can be reformulated as maximizing an adaptive submodular set function subject to a more complex set of constraints (a partition matroid). *This is the first approximation guarantee for adaptive influence maximization under execution errors, which is a well-known challenge in domains such as ours [25, 27].*

4.2 Network collection

The adaptive greedy algorithm assumes that the graph G is fully specified. However, in order for an intervention to deployed in practice, the social network needs to be laboriously gathered by interviewing the entire population of homeless youth (potentially hundreds of youth in total). This is not practical for a service provider to carry out on their own. We present an approach (Algorithm 2) which randomly samples a small number of youth to survey. Our procedure is easy for a service provider to implement in the field without much computational assistance. This simplicity is enabled by underlying insights about the structure of homeless youth social networks, which may assist with intervention design in other vulnerable populations.

We assume that the service provider has the ability to survey up to M youth. Each youth, when surveyed, reveals all of their edges. Algorithm 2 chooses $\frac{M}{2}$ nodes uniformly at random from the population to survey (line 5). For each surveyed node, it choses a uniformly random neighbor to survey as well (line 8). Lastly, it returns the graph consisting of the reported edges. The intuition for why this procedure succeeds is that it leverages the *friendship paradox*: a phenomena where a random node's neighbor has more friends, on average, than the node itself. Essentially, high-degree nodes are overrepresented when we sample a random neighbor instead of a uniformly random node. Thus, Algorithm 2 is disproportionately likely to find central nodes in the network who will reveal many edges and may be good potential seeds. We elaborate using empirical data from our pilot studies in Section 6.4.

We contrast here our sampling procedure with the previously proposed algorithm for influence maximization with an unknown network, ARISEN [24]. ARISEN simulates a random walk by starting at a random node, moving to a random neighbor of the first node, then to a random neighbor of the second and so on. ARISEN's motivation is very different. It exploits community structure, where nodes form densely connected subgraphs which are only loosely connected to the rest of the network. ARISEN uses each random walk to estimate the size of the community that it lies in and attempts to seed the largest communities. By contrast, Algorithm 2 leverages a fundamentally different structural property (the friendship paradox). This shift is motivated by the practical realities of deployment. In our feasibility study, we found that only 53% of contacts listed by youth could be located at the center. Hence, it is relatively easy to find at least one contact, as prescribed by Algorithm 2, but much harder to reach a chain of 5-10 youth in succession as in ARISEN.

4.3 Parameter robustness

A further complication is that the adaptive greedy algorithm assumes that the propagation probability p is known, in order to calculate the marginal gain Δ. However, p is never known precisely in practice; each intervention takes months to deploy so we are unlikely to

[2] https://www.dropbox.com/s/9q2yieups401792/supplement_deployment.pdf?dl=0

Algorithm 3 Robust parameter selection

```
input: parameter values $p_1 ... p_L$
for $i = 1...L$ do
    for $j = 1...L$ do
        $g(p_i, p_j)$ = value obtained by Algorithm 1 using $p_i$ evaluated under $p_j$
return $\arg\max_{i=1...L} \min_{j=1...L} \frac{g(p_i,p_j)}{g(p_j,p_j)}$
```

observe the many repeated cascades needed to for learning-based approaches. Previous work has attempted to resolve this dilemma via *robust influence maximization* [4, 10, 16, 25] which finds a seed set which performs well in the worst case over an uncertainty set of possible parameters. However, the only previous work which addresses robust influence maximization in an adaptive domain is the DOSIM algorithm. DOSIM requires hours or even days of runtime on a high-performance computing cluster because it needs to brute force over a grid of possible parameter settings. Such computational expense is far beyond the capabilities of the average service provider, motivating the development of lightweight but effective heuristics for robust influence maximization.

Algorithm 3 gives the heuristic used by CHANGE. It searches for a good nominal value of the parameter p, which (when given to Algorithm 1) will result in high performance no matter what the true value of p actually is. We first discretize the interval $[0, 1]$ into L points $p_1 ... p_L$. Let $g(p_i, p_j)$ denote the expected influence obtained when we run adaptive greedy planning based on propagation probability p_i, but the true parameter is p_j. We then find $p^* = \arg\max_{i=1...L} \min j = 1...L \frac{g(p_i,p_j)}{g(p_j,p_j)}$. Here, $\frac{g(p_i,p_j)}{g(p_j,p_j)}$, is the ratio of the value based on planning with parameter p_i to the value that could have been obtained if we new the true parameter p_j. p^* is the parameter which maximizes the worst-case value of this ratio. Notably, this procedure requires only L^2 runs of adaptive greedy; we take $L = 10$ in practice. By contrast, DOSIM requires $O\left(\frac{n}{\epsilon}\right)^3$ runs of a greedy algorithm to achieve approximation error ϵ. This quickly reaches thousands (or tens of thousands) of runs even for moderately sized networks and requires high-performance computing resources.

We investigate the performance of this heuristic on two real homeless youth social networks, Network A and Network B [25, 26]. Both were gathered from youth at a different drop-in center and contain approximately 150 nodes. Table 1 shows $\frac{g(p_i,p_j)}{g(p_j,p_j)}$, the percentage of optimality, for several combinations of p_i and p_j. For instance, the entry for Network A in the row corresponding to 0.2 and the column corresponding to 0.01 indicates that when adaptive greedy plans on $p = 0.2$, but the true parameter is actually $p = 0.01$, it obtains 88.7% of the optimal value possible. In both networks, Algorithm 3 selects $p^* = 0.2$ as the optimal choice: it has value at least 88.7% of the optimum under all parameter combinations in Network A and value at least 92.9% of the optimum on Network B. While this still improves on a naive choice which ignores robustness, we observe that all of the values in the table are relatively high. This indicates that influence maximization in this domain may not be highly sensitive to the exact choice of parameter.

To explain this phenomenon, Figure 2 shows the seed set chosen for Network A under different values of p. We observe a clear trend: with low p, the seeds are clustered more tightly together in the core of the network, and as p grows an increasing fraction of the seeds move to the periphery of the network. Intuitively, when p is high, a few seed nodes suffice to influence the core of the network. Thus, the greedy algorithm extracts higher marginal return by using seed nodes to cover outlying regions which are less likely to have been reached from the core. $p = 0.2$ represents a "goldilocks" solution where the core of the network is heavily covered without being oversaturated, and hence performs well across many values of p. However, other parameter choices can still do well because the majority of the possible value is located in the core of the network, which all seed sets devote several seeds to.

Table 1: Percentage of optimum obtained by planning based on parameter on row, when true parameter is given by column.

	Network A				Network B			
p	0.01	0.2	0.5	0.8	0.01	0.2	0.5	0.8
0.01	100	81.0	83.4	88.1	100	86.8	88.3	89.7
0.2	88.7	100	97.0	96.8	93.2	100	95.6	92.9
0.5	85.5	95.7	100	98.8	88.6	96.9	100	97.1
0.8	84.9	93.1	97.8	100	89.1	92.0	99.3	100

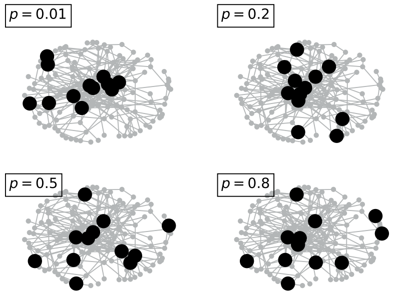

Figure 2: Seeds chosen under different values of p.

4.4 Simulation experiments

We now examine the performance of the CHANGE agent in a series of experiments using real-world data collected from homeless youth populations at different drop-in centers. We use networks collected from our own and previous pilot studies. The first network is the one we collected from the youth enrolled for CHANGE's intervention study. The other two networks were gathered by Yadav et al., also from real homeless youth, for their pilot studies of the HEALER and DOSIM algorithm. The main question is whether CHANGE is able to find influential seed nodes while only surveying a small fraction of the network. We ran CHANGE in simulation on each of the real-world networks, querying $M = 12$ nodes to obtain a sampled graph. This is 15-20% of the number of nodes in each network. Then, CHANGE selected $K = 4$ seed nodes in each of $T = 3$ rounds (reflecting the setup used in the intervention study). We conducted 30 independent trials for each network.

Figure 3 compares the number of non-peer leaders reached by CHANGE compared to the number reached by adaptive greedy (Algorithm 1) when it was given the entire network in advance.

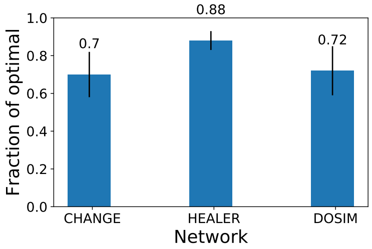

Figure 3: Simulated influence of CHANGE compared to adaptive greedy run on the full network. The x axis denotes which pilot study the network is taken from.

Table 2: Number of youth recruited, trained, and retained for follow-up in each study. CHANGE refers to the study conducted in this work to test the CHANGE agent. The other columns are taken from Yadav et al. [27], who conducted pilot tests of HEALER and DOSIM.

	CHANGE	HEALER	DOSIM
Youth recruited	64	62	56
Queried for links	18.75%	100%	100%
PL trained	15.6%	17.7%	17.85%
Retained	54.7%	73%	73%

We also tried comparing to the DOSIM agent [25] and obtained near-identical results. We see that CHANGE obtains 70-88% of the influence spread which is achievable if we knew the entire network in advance (comparable to previous work on network sampling [24]). However, CHANGE surveyed only 15-20% of the nodes in the network. This simulation, conducted on networks gathered from real homeless youth populations, provides evidence that CHANGE can find influential peer leaders using only a small amount of data.

5 PILOT STUDY PROCEDURE

The major contribution of this work is carrying out a pilot study which tests the CHANGE agent in a field deployment at a real drop in center serving homeless youth in a major U.S. city. Here, we outline the procedure followed for the pilot study. There were two studies, the feasibility study and the intervention study. In the feasibility study, we just tested the first component of CHANGE (network data collection) to validate that it works in practice to gather high-quality data. In the intervention study, we carried out actual interventions with homeless youth at the center. This step used all three steps of the CHANGE agent: we gathered the network, found a robust set of parameters, and then carried out interventions.

For each of the studies, we enrolled (respectively) 72 and 64 youth. Each youth was paid $20 to enroll in the study (all monetary incentives were the same as prior studies [27]). We ran CHANGE's data collection mechanism, randomly sampling a subset of youth to query for ties. Each youth who enrolled was also asked to complete a baseline survey. As part of this survey, we also gathered the *full* network consisting of ties from all of the youth. *We emphasize that this data was collected just for analysis. We did not use the full network to plan interventions, and we would not expect an agency to conduct this step in a regular deployment.* In the feasibility study, we also gathered edges via field observations and interviews with agency staff in order to validate our data collection via comparison to alternate mechanisms (see Section 6.1).

In the intervention study, trained social workers delivered the *Have You Heard* intervention, previously published in the public health literature [20]. The social workers conducted a day-long class with the selected youth, covering HIV awareness and prevention, and training the youth as peer leaders to communicate with other youth at the agency. Peer leaders were paid $60. Three sets of peer leaders were selected by CHANGE, with approximately 4 peer leaders in each set. This matches the size of trainings used in previous influence maximization pilot studies [27]. Table 2 reports specific values on the number of youth enrolled, queried for edges, and trained as peer leaders for our pilot test as well as pilot tests of previous algorithms. One month after the start of the study, we conducted a follow up survey with all of the youth who initially enrolled. Some youth were lost to follow up (see Table 2). We asked the youth whether they had received information about HIV prevention from a peer who was part of the study. Youth were paid $20 to respond to the follow up survey. We emphasize that all aspects of the intervention study (the training materials for peer leaders, survey instruments, etc.) are identical to Yadav et al. [27], so our results are directly comparable.

6 PILOT STUDY RESULTS

6.1 Feasibility study

We address two questions in the feasibility study. First, can Algorithm 2 (CHANGE's network sampling) be implemented with homeless youth? Second, is the resulting self-reported data accurate enough for influence maximization?

The challenge in the first question is that homeless youth can be difficult to locate. However, we were able to locate at least one neighbor for at least 80% of youth queried who were not isolates (i.e., named at least one neighbor). We conclude that Algorithm 2 is feasible for homeless youth populations. When no neighbor could be located, we drew a new random youth.

We now turn to the second question, which is of broader interest. Previous work on influence maximization in the field uses primarily self-reported network data [25–27]. Note that gathering ties from social media has proven unreliable for homeless youth populations both due to limited access to social media websites and mismatch between social media ties and true relationships. More broadly, self-reported network data is the best available to researchers in many field settings [3]. However, self reported ties are subject to their own limitations (forgetfulness, reticence on the part of the youth, etc. [3]). *To our knowledge, no previous work has validated whether self-reported ties suffice for influence maximization.* Our results show that self-reported data has important limitations (many edges discovered by other means were not self-reported), consistent with a large literature on network data collection methods [3]. However, using just the self-reported data sufficed to find near-optimal seed sets despite these limitations.

We gathered data via several methods: traditional self-reporting, field observations by the research staff, and interviews with staff members at the agency. This yielded three distinct sets of edges.

Table 3: Number of edges gathered by each method and the percentage overlap with edges gathered via self-report.

	Self-report	Observed	Staff	All
Number edges	51	23	46	112
Overlap with self-reports	100%	8.7%	13%	40%

Figure 4 shows the three networks, along with the composite graph obtained by combining edges from all three data sources. We see that self-reports give a fairly accurate global picture of the network. However, the other two data sources fill in many specific edges omitted in the self-reported data. Table 3 gives the number of edges gathered by each method, and the percentage of those edges which were contained in the self-reported data. We see that a high level of disagreement between the data collection methods on the status of individual edges: only 8.7% of ties from field observations and 13% of ties reported by staff members were reported by the youth themselves. In total, field observations and staff reports uncovered 69 edges, compared to 51 reported by the youth (with little overlap between the two). This is consistent with prior knowledge: a review of research on network data collection shows that anywhere from 10-80% of edges may be forgotten in self-reported data [3]. Another study comparing self-reported ties to observed interactions found that the two data sources were moderately correlated (median r = 0.51), but far from identical [7].

This establishes that many ties may be absent in self-reported data, but our ultimate objective is to find influential nodes (not reconstruct the network for its own sake). Hence, we now assess the robustness of influence maximization to missing edges. Given the propensity for forgetting in self-reported data, we conclude that all edges which *are* self-reported do exist [3], but many existing edges are not self-reported. Nevertheless, it is unlikely that all of the edges observed by field researchers or staff truly exist since reports by outside observers are typically less reliable than self-reports [7]. Thus, we conduct a simulation experiment in which a randomly selected portion of the non self-reported edges are added to the graph.

Figure 4 shows the performance of the greedy algorithm as the number of edges added increases. Each point on the x axis represents a fraction of edges which were observed by either field researchers or staff, but not reported by the youth themselves, to add to the graph. E.g., the point 0.25 indicates that a random set comprising 25% of edges which were not self reported are added to the self reported edges to obtain the final graph. Each point averages over 30 draws for this random set. The y axis shows the fraction of optimality obtained by running the adaptive greedy algorithm on just the self reported network. We approximate the optimum by running adaptive greedy on the full network, representing the best that we could have done had the entire true network been known. The values are consistently high, with very low standard deviation. Even when all of the unreported edges are added, so adaptive greedy does not know about the majority of edges in the graph, it still obtains at least 87% of the optimal value. In reality, not all of the unreported edges are real links, so we would expect even better performance in practice. We conclude that even though self-reported data may miss some edges, it still suffices to identify the influential nodes in the graph.

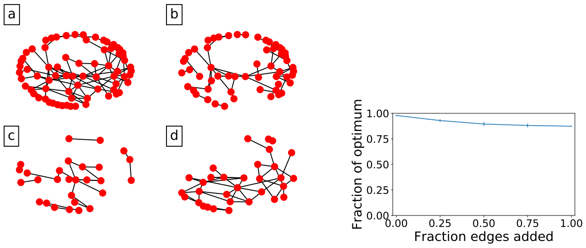

Figure 4: Left: Networks gathered using different methods. (a) All methods combined. (b) Self reported ties. (c) Field observations. (d) Staff observations. Right: Fraction of optimal value obtained using self-reported data as additional edges are added. Error bars show one standard deviation.

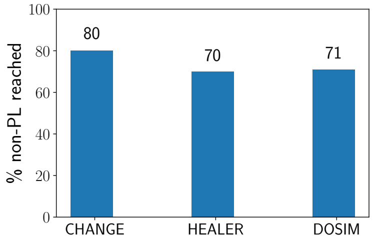

Figure 5: Percentage of youth who were not peer leaders reached by each algorithm in its respective real-world pilot test.

6.2 Intervention study

We now turn to our second pilot study, which tested the entirety of the CHANGE agent. In this study, we recruited a separate population of 64 homeless youth from a drop-in center. Table 2 gives the total number of youth recruited for different activities, as well as the corresponding figures for previous pilot tests of the HEALER and DOSIM algorithms by Yadav et al. [27]. We gathered the full social network from all 64 youth, and in parallel ran Algorithm 2 with a budget of $M = 12$ youth to collect a sampled network (querying 18.75% of youth in total for links). *Only the sampled network was used to plan interventions; the full network was gathered only for analysis.* We then ran the CHANGE policy for three steps, training 10 total peer leaders (15.6% of the network). This percentage is comparable to previous studies (HEALER and DOSIM trained approximately 17% of the network each). However, HEALER and DOSIM used the entire network to plan their intervention, compared to the 18.75% of sampled youth used by CHANGE. At one month, we conducted a follow-up survey to assess whether youth received information about HIV prevention from the peer leaders. 54.7% of youth were retained in the follow-up survey, which is a somewhat lower percentage than in previous studies. Nevertheless, we obtain a population of 34 youth who provided follow-up data.

6.3 Influence spread results

We now present our core result: the number of youth who received a message about HIV prevention. We examine the percentage of youth in the follow-up group who were not peer leaders (and hence

Table 4: Aggregate network statistics for the complete network in each algorithm's pilot study. "Diameter" is the diameter of the largest connected component.

	CHANGE	HEALER	DOSIM
Diameter	12	8	8
Density	0.043	0.079	0.059
Avg. path length	4.88	3.38	3.15
Avg. clustering coeff.	0.221	0.397	0.195
Modularity	0.654	0.568	0.568

eligible to become influenced) who reported receiving information. Figure 5 shows this percentage for our pilot study of CHANGE as well as the percentages reported by Yadav et al. [27] in their pilot studies of the state of the art algorithms HEALER and DOSIM. CHANGE reached 80% of non-peer leaders compared to approximately 70% for each of HEALER and DOSIM. *Thus, CHANGE was able to reach just as many youth while gathering data from only 18.75% of the network.* The 10% difference between CHANGE and HEALER/DOSIM could be attributable to random variation; we do not claim that CHANGE is actually more effective than algorithms which gather the entire network. Nevertheless, this result provides empirical evidence that CHANGE can perform comparably to existing state of the art influence maximization agents while drastically reducing the amount of data required.

We now take steps to ensure that our results are not an artifact of a difference between the structures of the different networks from each pilot test or of random variation. First, we recall our simulation results in Figure 3, which indicate that CHANGE performs competitively with algorithms which are given the entire graph on three different real-world networks. Second, Table 4 shows a range of statistics for each network. CHANGE's networks is fairly similar to that of HEALER and DOSIM. However, it is somewhat sparser: its density (the fraction of possible edges which are present) is 0.043 compared to 0.079 for HEALER and 0.059 for DOSIM. This translates into somewhat longer average path lengths and larger diameter. However, sparser structure should only work *against* CHANGE since there are fewer edges along which influence can propagate. Hence, it is unlikely that CHANGE's strong performance is attributable to anomalous network structure.

6.4 Explaining CHANGE's success

In this section we attempt to explain why CHANGE can find seed sets which have near-optimal influence spread by surveying only a small fraction of youth. The intuitive explanation for this is a property that many social networks are known to possess: the friendship paradox [6, 11, 22]. Specifically, a randomly chosen neighbor of a given node is likely to have higher degree than the node itself. Our algorithm leverages the friendship paradox by surveying both a random node and a randomly chosen friend of that node.

Figure 6 plots two quantities for the networks collected in the feasibility and intervention studies. First, the degree distribution. Second, the distribution of the degree of a randomly chosen neighbor of a randomly chosen node. This is the degree distribution of the nodes that Algorithm 2 samples in its second step. We see that the neighbor degree distribution is skewed towards higher degrees. In the feasibility network, the mean degree is 3.11 while the mean friend's degree is 4.56. In the intervention network, the mean degree is 2.98 while the mean friend's degree is 4.04. This suggests that by querying a random neighbor of each node, our algorithm is able to preferentially locate nodes who are useful in two ways. First, high degree nodes provide more information about the network. Second, they are more likely to be influential peer leaders and may serve as a useful set of candidates which adaptive greedy can refine.

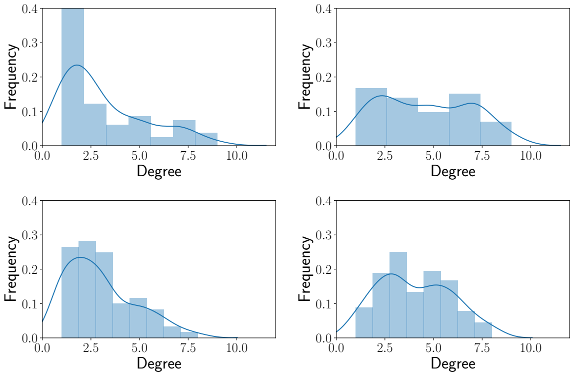

Figure 6: Degree distributions. The top row is from the feasibility study; the bottom from the intervention study. Left: standard degree distribution. Right: degree of a random neighbor.

7 DISCUSSION AND CONCLUSION

This paper presents the CHANGE agent for influence maximization, a multiagent problem with many applications in preventative health and other domains. CHANGE addresses major barriers to the deployment of influence maximization by service providers through a series of algorithmic contributions, backed by simulation results on real-world networks. We then conducted a real-world pilot study of CHANGE with a drop-in center serving homeless youth, the first such pilot study of sampling-based influence maximization and only the second study testing *any* influence maximization agent in the real world. CHANGE obtained comparable influence spread to previously field tested algorithms, but surveyed only 18% of youth to obtain network data. CHANGE has empirical promise in delivering high-quality influence maximization solutions in a manner which can be feasibly implemented by a service provider.

While the algorithms underlying CHANGE are easy to implement, they draw on a series of insights into the social behavior of homeless youth. One lesson learned from our study is that, to be successful in the field, algorithms must be designed with their target population and setting in mind. CHANGE both navigates challenges specific to homeless youth (e.g., the difficulty of locating youth to query for edges or serve as peer leaders) and leverages properties of their social network (the friendship paradox). Our experience shows that accounting for both challenges and opportunities in the target population is crucial to produce a practically deployable algorithm.

Acknowledgments: This research was supported by MURI Grant W911NF-11-1- 0332, the California HIV/AIDS Research Program, and a NSF Graduate Research Fellowship.

REFERENCES

[1] Cheryl Alexander, Marina Piazza, Debra Mekos, and Thomas Valente. 2001. Peers, schools, and adolescent cigarette smoking. *Journal of adolescent health* 29, 1 (2001), 22–30.

[2] Abhijit Banerjee, Arun G Chandrasekhar, Esther Duflo, and Matthew O Jackson. 2013. The diffusion of microfinance. *Science* 341, 6144 (2013), 1236498.

[3] Devon D Brewer. 2000. Forgetting in the recall-based elicitation of personal and social networks. *Social networks* 22, 1 (2000), 29–43.

[4] Wei Chen, Chi Wang, and Yajun Wang. 2010. Scalable influence maximization for prevalent viral marketing in large-scale social networks. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1029–1038.

[5] Edith Cohen, Daniel Delling, Thomas Pajor, and Renato F Werneck. 2014. Sketch-based influence maximization and computation: Scaling up with guarantees. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. ACM, 629–638.

[6] Scott L Feld. 1991. Why your friends have more friends than you do. *Amer. J. Sociology* 96, 6 (1991), 1464–1477.

[7] Scott D Gest, Thomas W Farmer, Beverley D Cairns, and Hongling Xie. 2003. Identifying children's peer social networks in school classrooms: Links between peer reports and observed interactions. *Social Development* 12, 4 (2003), 513–529.

[8] Amit Goyal, Wei Lu, and Laks VS Lakshmanan. 2011. Celf++: optimizing the greedy algorithm for influence maximization in social networks. In *Proceedings of the 20th international conference companion on World wide web*. ACM, 47–48.

[9] Amit Goyal, Wei Lu, and Laks VS Lakshmanan. 2011. Simpath: An efficient algorithm for influence maximization under the linear threshold model. In *Data Mining (ICDM), 2011 IEEE 11th International Conference on*. IEEE, 211–220.

[10] Xinran He and David Kempe. 2016. Robust Influence Maximization. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*. 885–894. https://doi.org/10.1145/2939672.2939760

[11] Nathan Oken Hodas, Farshad Kooti, and Kristina Lerman. 2013. Friendship Paradox Redux: Your Friends Are More Interesting Than You. *ICWSM* 13 (2013), 8–10.

[12] Kyomin Jung, Wooram Heo, and Wei Chen. 2012. Irie: Scalable and robust influence maximization in social networks. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*. IEEE, 918–923.

[13] David Kempe, Jon Kleinberg, and Éva Tardos. 2003. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 137–146.

[14] Jun-young Kwak, Pradeep Varakantham, Rajiv Maheswaran, Milind Tambe, Timothy Hayes, Wendy Wood, and Burcin Becerik-Gerber. 2012. Towards robust multi-objective optimization under model uncertainty for energy conservation. In *AAMAS workshop on agent technologies for energy systems (ATES)*.

[15] Weihua Li, Quan Bai, Tung Doan Nguyen, and Minjie Zhang. 2017. Agent-based Influence Maintenance in Social Networks. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 1592–1594.

[16] Meghna Lowalekar, Pradeep Varakantham, and Akshat Kumar. 2016. Robust influence maximization. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 1395–1396.

[17] Mahsa Maghami and Gita Sukthankar. 2012. Identifying influential agents for advertising in multi-agent markets. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*. International Foundation for Autonomous Agents and Multiagent Systems, 687–694.

[18] Mayur Mohite and Y Narahari. 2011. Incentive compatible influence maximization in social networks and application to viral marketing. In *The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 3*. International Foundation for Autonomous Agents and Multiagent Systems, 1081–1082.

[19] Eric Rice, Norweeta G Milburn, and Mary Jane Rotheram-Borus. 2007. Pro-social and problematic social network influences on HIV/AIDS risk behaviours among newly homeless youth in Los Angeles. *AIDS care* 19, 5 (2007), 697–704.

[20] Eric Rice, Eve Tulbert, Julie Cederbaum, Anamika Barman Adhikari, and Norweeta G Milburn. 2012. Mobilizing homeless youth for HIV prevention: a social network analysis of the acceptability of a face-to-face and online social networking intervention. *Health education research* 27, 2 (2012), 226–236.

[21] Youze Tang, Xiaokui Xiao, and Yanchen Shi. 2014. Influence maximization: Near-optimal time complexity meets practical efficiency. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*. ACM, 75–86.

[22] Johan Ugander, Brian Karrer, Lars Backstrom, and Cameron Marlow. 2011. The anatomy of the facebook social graph. *arXiv preprint arXiv:1111.4503* (2011).

[23] Thomas W Valente and Patchareeya Pumpuang. 2007. Identifying opinion leaders to promote behavior change. *Health Education & Behavior* (2007).

[24] Bryan Wilder, Nicole Immorlica, Eric Rice, and Milind Tambe. 2018. Maximizing influence in an unknown social network. In *AAAI Conference on Artificial Intelligence*.

[25] Bryan Wilder, Amulya Yadav, Nicole Immorlica, Eric Rice, and Milind Tambe. 2017. Uncharted but not Uninfluenced: Influence Maximization with an uncertain network. In *Proceedings of the 16th Conference on Autonomous Agents and Multi-Agent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 1305–1313.

[26] Amulya Yadav, Hau Chan, Albert Xin Jiang, Haifeng Xu, Eric Rice, and Milind Tambe. 2016. Using social networks to aid homeless shelters: Dynamic influence maximization under uncertainty. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 740–748.

[27] Amulya Yadav, Bryan Wilder, Eric Rice, Robin Petering, Jaih Craddock, Amanda Yoshioka-Maxwell, Mary Hemler, Laura Onasch-Vera, Milind Tambe, and Darlene Woo. 2017. Influence maximization in the field: The arduous journey from emerging to deployed application. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 150–158.

[28] Sean D Young and Eric Rice. 2011. Online social networking technologies, HIV knowledge, and sexual risk and testing behaviors among homeless youth. *AIDS and Behavior* 15, 2 (2011), 253–260.

[29] Haifeng Zhang, Ariel D Procaccia, and Yevgeniy Vorobeychik. 2015. Dynamic influence maximization under increasing returns to scale. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 949–957.