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 aimed at overcoming 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 among 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 are crucial bottlenecks to widespread use of influence maximization in real-world interventions.

To address the above challenges, this 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

ACM Reference Format:

1 INTRODUCTION
This 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].

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, expertise to select appropriate parameters, and computational power to run the algorithms. None of these are likely available to the resource-strained service providers ultimately responsible for deployment.

Gathering network data is particularly onerous because it requires individually surveying over a hundred youth. Network collection is more time intensive than simple survey methods, requiring days of time for a dedicated team of social work researchers. It is infeasible 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, a state of the art algorithm for robust influence maximization [25], requires hours on a high-performance computing system. A deployed system would need to run in minutes on a laptop.

This paper presents 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 neighbor is also interviewed. We show that this procedure gathers enough of the network to enable 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 to days of time on a high performance 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 influence maximization work 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 for 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

Influence maximization was introduced by Kempe et al. [13], and has been extensively studied since then [4, 5, 8, 9, 12, 15, 18, 21]. Most work 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 uses robust optimization to account for uncertainty about the true probability of influence propagation. Our approach to parameter robustness is similar to techniques 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 (see Section 4.3). 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). 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\footnote{Note that while CHANGE directly surveys ~18% of youth, they name others as friends, resulting in a larger sampled graph.}. Further, limitations on space and personnel mean that the
intervention 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 $E[q] = p$ and then propagate influence with probability $q$. Hence, our model subsumes any stochastic model where the 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 $x_v \in \{\text{present}, \text{absent}\}$. 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 $x_v = \text{present}$, then $v$ is activated. Nothing occurs if $x_v = \text{absent}$. 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 $A = \{S \subseteq V : |S| \leq K\}$ 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 $\Psi$ be the set of all possible histories. A policy is a mapping $\pi : \Psi \rightarrow A$. Let $A(\psi_t) = (A_1...A_t)$ be the sequence of actions taken and $O(\psi_t) = (O_1...O_t)$ 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 $f(\pi) = E_{\psi \sim \pi}[f(A(\psi)|O(\psi))]$ be the expected reward from running policy $\pi$, where the expectation ranges over the hidden state $x$ (which determines $\pi$’s actions) as well as the influence process. We seek a policy maximizing $f(\pi)$.

**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 $p$. Algorithms 2 and 3 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 $2^{|\Psi|}$ possible hidden states and $\binom{n}{K}$ possible actions. While it is possible to formulate the problem as a POMDP, these exponentially large state
We show that our problem can be reformulated as maximizing an Adaptive greedy $G$ The adaptive greedy algorithm assumes that the graph

4.2 Network collection

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 $\Delta$. However, $p$ is never known precisely in practice; each intervention takes months to deploy so we are unlikely to observe the many repeated cascades needed to for learning-based approaches. Previous work has attempted to resolve
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.

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 algorithms. 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.

We also tried comparing to the DOSIM agent [25] and obtained 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 devise several seeds to.

### 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. We also tried comparing to the DOSIM agent [25] and obtained

<table>
<thead>
<tr>
<th>$p$</th>
<th>0.01</th>
<th>0.2</th>
<th>0.5</th>
<th>0.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p$</td>
<td>100</td>
<td>86.8</td>
<td>88.3</td>
<td>89.7</td>
</tr>
<tr>
<td>0.2</td>
<td>88.7</td>
<td>100</td>
<td>95.6</td>
<td>92.9</td>
</tr>
<tr>
<td>0.5</td>
<td>85.5</td>
<td>99.3</td>
<td>97.1</td>
<td>100</td>
</tr>
<tr>
<td>0.8</td>
<td>84.9</td>
<td>93.1</td>
<td>97.8</td>
<td>100</td>
</tr>
</tbody>
</table>

**Table 1:** Percentage of optimum obtained by planning based on parameter on row, when true parameter is given by column.
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, 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 others 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 number 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, 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, 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. Figure 4 shows the three networks, along with the composite graph obtained by combining edges from all three data sources. We see 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.

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.

<table>
<thead>
<tr>
<th>Youth recruited</th>
<th>CHANGE</th>
<th>HEALER</th>
<th>DOSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Query for links</td>
<td>64</td>
<td>62</td>
<td>56</td>
</tr>
<tr>
<td>PL trained</td>
<td>18.75%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>Retained</td>
<td>54.7%</td>
<td>73%</td>
<td>73%</td>
</tr>
</tbody>
</table>
Table 3: Number of edges gathered by each method and the percentage overlap with edges gathered via self-report.

<table>
<thead>
<tr>
<th>Number edges</th>
<th>Self-report</th>
<th>Observed</th>
<th>Staff</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>51</td>
<td>23</td>
<td>46</td>
<td>112</td>
<td></td>
</tr>
</tbody>
</table>

Overlap with self-reports: 100% 8.7% 13% 40%

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].

While many ties may be absent in self-reported data, 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 possible under full information. The values are consistently high, with very low standard deviation. Even when all of the unreported edges are added, 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.

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

<table>
<thead>
<tr>
<th></th>
<th>CHANGE</th>
<th>HEALER</th>
<th>DOSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diameter</td>
<td>12</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>Density</td>
<td>0.043</td>
<td>0.079</td>
<td>0.059</td>
</tr>
<tr>
<td>Avg. path length</td>
<td>4.88</td>
<td>3.38</td>
<td>3.15</td>
</tr>
<tr>
<td>Avg. clustering coeff.</td>
<td>0.221</td>
<td>0.397</td>
<td>0.195</td>
</tr>
<tr>
<td>Modularity</td>
<td>0.654</td>
<td>0.568</td>
<td>0.568</td>
</tr>
</tbody>
</table>

youth in the follow-up group who were not peer leaders (and hence 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.

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 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.