Reducing Gang Violence Through Network Influence Based Targeting of Social Programs

Paulo Shakarian
Arizona State University
Tempe, AZ 85281
pshakari@asu.edu

Joseph Salmento and
William Pulleyblank
Network Science Center
U.S. Military Academy
West Point, NY 10996
joseph.salmento@usma.edu
william.pulleyblank@usma.edu

John Bertetto
Chicago Police Dept.
Chicago, IL 60653
john.bertetto@chicagopolice.org

ABSTRACT

In this paper, we study a variant of the social network maximum influence problem and its application to intelligently approaching individual gang members with incentives to leave a gang. The goal is to identify individuals who when influenced to leave gangs will propagate this action. We study this emerging application by exploring specific facets of the problem that must be addressed when modeling this particular situation. We formulate a new influence maximization variant - the "social incentive influence" (SII) problem and study it both formally and in the context of the law-enforcement domain. Using new techniques from unconstrained submodular maximization, we develop an approximation algorithm for SII and present a suite of experimental results - including tests on real-world police data from Chicago.

Categories and Subject Descriptors

Applied Computing [**Law, social and behavioral sciences**]: Sociology

General Terms

Algorithms, Experimentation

Keywords

complex networks, network diffusion, propagation in networks

1. INTRODUCTION

Violent street gangs are a major cause of criminal activity in the United States [2, 5]. A recent trend has been toward development of "smart policing" tactics to reduce the effectiveness of these gangs. Typically, these strategies have focused on the allocation of law enforcement resources to increase arrests and deter criminal behavior. In this paper we focus on a use of "smart policing" in a different manner: we wish to intelligently target individual gang members with incentives to leave a gang. While "pulling levers" or encouraging dis-enrollment from a gang, is already a tactic employed in cities such as Boston and Chicago, the selection of which specific gang members to focus on is still largely unanswered - and hence currently based on ad-hoc methods. In this paper, we study this emerging application as a variant of a social network influence maximization problem [14] that we refer to as the "social incentive influence" (SII) problem. We study this problem both formally and in the context of law enforcement. Then, using new techniques from unconstrained submodular maximization [7], we develop an approximation algorithm for SII and present a suite of experimental results - including tests on real-world police data.

The paper is organized as follows. In Section 2 we discuss some current methods in law-enforcement used for social program targeting. In Section 3 we introduce the SII problem an associated technical preliminaries. This is followed by a discussion of our algorithmic approach in Section 4 and associated experimental results in Section 5. Finally, we discuss related work in the literature in Section 6.

2. BACKGROUND

Recent successes with so called "pulling levers" (gang membership dis-enrollment) approaches to deterring violence include the Boston Gun Project and Operation CeaseFire as well as Project Safe Neighborhoods in Chicago, and continue with the on-going Group Violence Reduction Strategy in Chicago. Using this approach, law enforcement partners work with social service providers and victims advocacy groups to attempt to abate gang violence by 'pulling' whatever 'levers' need to be applied to street gangs. The types of levers applied, and the degree to which they are employed, depend upon the particular gang. Adjustments are made so that the application is both customized to the specific situation and, hopefully, more successful.

To facilitate these interactions between law enforcement, social service providers, victim's advocacy groups, and street gang members, two approaches are commonly employed. In the case of the first, law enforcement engages known street gang members on the street as part of regular patrol activities. While performed under the auspices of focused de-

KDD'14, August 24–27, 2014, New York, NY, USA.
Copyright 2014 ACM 978-1-4503-2956-9/14/08 ...$15.00.
http://dx.doi.org/10.1145/2623330.2623331.

terrence, this type of interaction is likely to have the least impact. Time spent with the gang member may be limited, the remaining levers (social service providers, victims groups, etc.) are absent so that the message conveyed to the street gang member is incomplete or biased toward criminal enforcement aspect. Moreover, the selection of target gang members is often random – the result of opportunity as the gang member is seen on the street.

In the second approach, gang members are subjected to a "call-in." The call-in sessions are prearranged meetings organized by law enforcement, social service providers, and victims groups during which messages of non-violence are conveyed. The call-in is a full-spectrum effort: law enforcement makes clear to attendees that further violence will be met with relentless police operations and enforcement efforts, social service providers offer information on how members can exit the street gang and obtain educational and vocational training, and victims groups tell stories of loss in an effort to make an emotional plea for violence cessation.

Attendees for call-in sessions are typically chosen in two ways: compulsory attendance and invitation. Gang members currently on probation or parole are compelled to attend. The remaining attendees are invitees, and it is here where the selection criteria may become a bit vague. Invitees may be selected via some form of social network analysis (as in [5]) however such concerted efforts are not entirely relied upon. Often, a large amount of discretion is afforded local law enforcement in selecting invitees. This allows for local gang experts and command staff members to identify those gang members whom they know or suspect to be influential or key members of the gang and invite them to the call-in. This can be a very successful process if the agency has access to these experts or knowledge of the gang's organizational structure. However, there is no guarantee that those persons invited are, in fact, genuinely influential in the gang or are just the "most well known to law enforcement" members of the gang. In law enforcement, where financial, personnel, and time resources are increasingly constrained, turning a more objective and analytical eye toward invitee selection grows more important.

A key aim of gang dis-enrollment programs is to enable law enforcement to invite to call-ins those gang members who, should the efforts to dis-enroll them be successful, are most likely to pull additional gang members out with them. However, there is a key challenge: are influential members also easy to encourage to leave the gang? A recent empirical study exploring non-criminal online social networks suggests that highly influential individuals are typically not susceptible [1]. However, we argue that taking both influence and susceptibility into account are necessary; identifying individuals (or groups of individuals) that possess both qualities is needed for the behavior to spread.

Though, to the best of our knowledge, influence maximization techniques have not been applied to law enforcement before, there is some anecdotal evidence that such approaches could bear fruit. For instance, there have been cases where gang members thought to be "influential" successfully convinced others to dis-enroll from the gang. In one case, in Chicago (in the summer of 20130, the district personnel (local plainclothes gang officers) knew this person to be an influential and as such targeted him for intervention. When he was contacted he indicated that he had already gotten a job, but knew several fellow gang members who could use the offered social services and dis-enrollment opportunities. He personally contacted 20 fellow gang members, of which 7 walked into the local social services facility.

In this paper we frame the problem of "pulling levers" formally as a variant of the social network maximum influence problem [14] in what we call the "social incentive influence" (SII) problem. However, there are some key nuances of the "pulling levers" strategy that we integrated into our framework that are not inherently considered in the maximum influence problem. We list these items here and address our technical approach to each in the next section.

1. **Duration of the diffusion process.** One key difference SII has from other maximum influence formulations is the length of time it takes for the diffusion process to occur. The reason for this is that gang dis-enrollment is a major life decision for an individual, hence the spread of this idea will likely take time. Further complicating the matter is that there may be changes to social network structure while the diffusion process occurs - based on events such as arrests, homicides, gang conflict and cooperation, etc.

2. **Interaction with the population during the diffusion process.** Not only does the diffusion process take time to occur in this domain, but also the law enforcement personnel will often make multiple attempts to "pull levers" as the diffusion occurs.

3. **Geographic locations and strength of connections.** Often, law enforcement data has an inherent geospatial component. In this work, we leverage this information to inform the strength of connections in the social network - as the street gangs are also inherently territorial.

4. **Notion of cost.** Cost also becomes an important factor in SII as the law enforcement personnel are attempting to encourage a major change in the life of the gang members. Conducting a call-in session with certain members costs time, money, and other resources. We also note that not all gang members will be equally susceptible to this type of intervention - some may require more or less effort to dis-enroll. Further, there are real costs associated with encouraging dis-enrollment. For example, in North Carolina personalized letters are created for the gang members that show the individual how his association with others involved in violence puts him or her at risk. A similar tactic is used in Chicago, where the letters are often delivered to the homes of the gang members. This utilizes police manpower and resources hence further increasing the cost.

5. **Profit maximization.** As we consider cost, we also model "benefit" in SII - which is the value of each expected infectee to the diffusion process. This allows us to adopt a profit-maximizing model (similar to the ProMax problem of [15]) where we look to maximize the expected benefit minus the cost.

3. TECHNICAL PRELIMINARIES AND ANALYSIS

Throughout this paper we assume the existence of a *social network* $G = (V, E)$ where V is a set of vertices and E is a set of directed edges. We let n and m denote the cardinality of V and E respectively. For any node $v \in V$, the set of incoming neighbors is η_v^{in}, and the set of outgoing neighbors is η_v^{out}. The cardinalities of these sets (and hence the in- and out-degrees of node v) are denoted by k_v^{in}, k_v^{out} respectively. For each node v, we assume a cost of marketing to that node denoted by $c_v \in \Re^+$. We let C denote the vector of costs indexed by V. We let $\langle c \rangle$ denote the average cost $\sum_v c_v/n$. We also assume a benefit value, $b \in \Re^+$, which is the associated benefit for having marketed to a given node.

We assume that each node in G has an associated geolocation and there exists a distance function $d : V \times V \to \Re^+$ that meets the normal axioms: $d(u, u) = 0$, $d(u, v) = d(v, u)$, and $d(u, w) \leq d(u, v) + d(v, w)$.

Using this distance function, we shall assume a level of influence p_{uv} for each edge $(u, v) \in E$ that we define using an exponential distance-decay model [23, 16, 22] as follows:

$$p_{uv} = e^{-(d(u,v)/\gamma)^r}$$

where γ, r are parameters in the interval $(0, \infty)$ and e is the base of the natural logarithm. The parameter γ is used as a scaling parameter - and we shall set it to be the average distance between two nodes connected with an edge. The parameter r controls the shape of the distance decay curve, and we shall typically use $r = 2$. The use of the distance decay function as a proxy for influence is the primary way we address the geographic nature of the law-enforcement data in this application.

Diffusion Process. A key property that we utilize in our study is submodularity, which we review below:

Definition 3.1 (Submodularity). *Function $f : 2^U \to \Re^+$ is* ***submodular*** *if for every $A \subseteq B \subseteq U$ and $u \in U$: $f(A \cup \{u\}) - f(A) \geq f(B \cup \{u\}) - f(B)$.*

Intuitively, the idea of submodularity represents a notion of diminishing returns: adding an element u to a set B can provide no greater benefit than that gained by adding it to any proper subset of B.

Equivalently, function $f : 2^U \to \Re^+$ is submodular if for every $A, B \subseteq U$, $f(A) + f(B) \geq f(A \cup B) + f(A \cap B)$.

A set function is *supermodular* if its negation is submodular.

Next, we define a diffusion process function (*dpf*) which accepts an initial set of vertices (called the "seed set") and returns the expected number of infectees once the diffusion process completes. In this paper, we shall require this function to be sub-modular and normalized. We provide a formal definition below.

Definition 3.2 (Diffusion Process Function). *A diffusion process function, $dpf : 2^V \to [0, n]$, is any function such that: (1.) $dpf(\emptyset) = 0$ and (2.) $\forall V_1, V_2 \subseteq V$: $dpf(V_1) + dpf(V_2) \geq dpf(V_1 \cup V_2) + dpf(V_1 \cap V_2)$.*

We argue that, in general, these are reasonable restrictions. For instance, the σ function of the independent cascade and linear threshold models [14], the oracle of the MIA model [11], and the value function of the logic-programming framework of [20] are all valid diffusion process functions. We also note that these previous studies focused only on maximizing the number of expected infectees (subject to a cardinality constraint). In this work, we disregard the cardinality constraint and instead seek to maximize profit, which we formally define below.

Definition 3.3 (Profit). *Profit, $pft : 2^V \to \Re$ is defined by $pft(X) = b \times dpf(X) - \sum_{v \in X} c_v$*

The SII Problem. We now have all components necessary to formally define the social incentive influence (SII) problem:

Definition 3.4 (SII Problem). *We are given diffusion process function dpf, social network $G = (V, E)$, cost vector C and benefit value b. Find $SII(dpf, G, C, b) = V^* \subseteq V$ such that $pft(V^*) \geq pft(V')$ for all $V' \subseteq V$.*

Not surprisingly, the social incentive influence problem is NP-hard.

Theorem 3.1. *SII is NP-hard.*

Proof. We show NP-hardness by reducing SIMPLE MAX CUT [13] to SII. The SIMPLE MAX CUT problem takes as input a graph $G = (V, E)$ and returns sets $V_1, V_2 \subseteq V$ such that $|\{(u, v) \in E : u \in V_1, v \in V_2\}|$ is maximized. The following construction can be performed in polynomial time. Let $dpf(X) = \sum_{v \in X} f_v(X)$ where $f_v(X) = 1$ if $v \in X$ and $|\eta_v^{in} \cap X|$ otherwise. Note that $dpf(\emptyset) = 0$ and, because each f_v is submodular, *dpf* is submodular. For each v set $c_v = 1$ and set $b = 1$. Then $pft(X) = \sum_{v \in V \setminus X} : \eta_v^{in} \cap X| = |\{(u, v) \in E | u \in X, v \in V \setminus X\}|$. Hence *pft* becomes equivalent to the objective function for SIMPLE MAX CUT. □

However, note that our profit function *pft* is submodular.

Proposition 3.1. *pft is submodular.*

Proof. It is well known that subtracting a supermodular function from a submodular function yields a submodular function. Since *dpf* is submodular (and b is positive) and the sum of costs is supermodular, the proposition follows. □

One-Step Diffusion. As stated earlier, two key challenges in this domain are the duration of the diffusion process and the effect of the law-enforcement personnel interacting during the diffusion process. This has led us to model the diffusion process as a "one-step" influence model where we only consider the immediate effect of the diffusion process one time step in the future. Our envisioned use case is that the law-enforcement analysts will use the most current data available to make a decision as to which gang members to reach-out to based on this model. Attempts will be made to influence those individuals, after which changes to the social network (both resulting the outreach and other external factors) will be incorporated before repeating the cycle. Because we expect the time for diffusion to generally take longer, the repetition of the cycle will generally occur after about one time period. We formally define the following"one-step" influence model.

Definition 3.5. *The* ***one-step diffusion model****, $\sigma_1 : 2^V \to \Re^+$ is defined as follows:*

$$\sigma_1(V') = \sum_{u \in V} \big(1 - \prod_{u \in \eta_v^{in} \cap V'} (1 - p_{uv})\big)$$

Note that we also assume that a node v is infected by a node u independently of which others of its incoming neighbors were previously infected. An easy proof shows that σ_1 is a valid diffusion process function.

PROPOSITION 3.2. *σ_1 is a valid diffusion process.*

PROOF. Clearly, $\sigma_1(\emptyset) = 0$ by inspection. Next, we show that the quantity $\prod_{u\in\eta_v^{in}\cap V'}(1-p_{uv})$ is supermodular. Suppose, by way of contradiction, that it is not, then we have for V' and nodes $q, r \notin V'$ the following for each $v \in V$:

$$\prod_{u\in\eta_v^{in}\cap(V'\cup\{q,r\})}(1-p_{uv}) - \prod_{u\in\eta_v^{in}\cap(V'\cup\{r\})}(1-p_{uv}) < \prod_{u\in\eta_v^{in}\cap(V'\cup\{q\})}(1-p_{uv}) - \prod_{u\in\eta_v^{in}\cap V'}(1-p_{uv})$$

Let us assume that q, r are both neighbors of v (the other cases cause both sides to be equal). This gives us the following:

$$(1-p_{rv})\Big(\prod_{u\in\eta_v^{in}\cap(V'\cup\{q\})}(1-p_{uv}) - \prod_{u\in\eta_v^{in}\cap V'}(1-p_{uv})\Big) < \prod_{u\in\eta_v^{in}\cap(V'\cup\{q\})}(1-p_{uv}) - \prod_{u\in\eta_v^{in}\cap V'}(1-p_{uv})$$

As $\prod_{u\in\eta_v^{in}\cap(V'\cup\{q\})}(1-p_{uv}) \leq \prod_{u\in\eta_v^{in}\cap V'}(1-p_{uv})$, we have $1 < 1 - p_{rv}$ which is clearly a contradiction. Note that the supermodularity of this quantity implies the submodularity of $1-\prod_{u\in\eta_v^{in}\cap V'}(1-p_{uv})$. The rest of the statement follows from the fact that σ_1 is a positive linear combination of submodular functions. □

We note that we can cause nodes in the argument of this function to be assigned a probability of 1.0 by simply adding self-directed edges to each node in the network. We also note, that with many diffusion processes functions, the calculation of their outcome may yield an individual probability of activation for each node. Further, the one-step model also allows for the consideration of benefit as a vector - the probability for each node can obtained by inspecting the inner summation - this allows for a more customized setting of benefit on basis of each node (we are currently discussing this as a possibility with our law enforcement partners). In this case, we can identify a specific benefit for each node. The framework can be easily adapted for such a case.

4. APPROACH

While the submodularity of the *pft* function is encouraging, we note that because marketing to each node incurs an associated cost, it is possible to experience a loss by marketing to additional nodes. For instance, if we market to an additional individual who provides us no increase in the diffusion process, this reduces our profit and could lead to a loss. This is not considered in previous diffusion models designed to maximize the expected number of infectees. Hence, the greedy approximation of [17] no longer provides us an approximation guarantee. Our case can instead be viewed as an "unconstrained" submodualr function. Recently, a deterministic approximation algorithm was introduced in [7] that requires only a linear number of evaluations of the function. We recall their algorithm here (adapted for SII).

For positive, unconstrained submodular maximization, [7] proves that SII-Approx provides a result that is at least 1/3 of optimal. Note that *pft* can potentially provide a solution with negative value. However, we can leverage their results to provide the following approximation guarantee:

Algorithm 1 SII-Approx[7]

INPUT: Social network $G = (V, E)$, cost vector C, benefit b, distance function d.
OUTPUT: Approximation V' to SII.

```
$V' = \emptyset, V'' = V$
for $v \in V$ do
    $a = pft(V' \cup \{v\}) - pft(V')$
    $b = pft(V'' \setminus \{v\}) - pft(V'')$
    if $a \geq b$ then
        $V' = V' \cup \{v\}$
    else
        $V'' = V'' \setminus \{v\}$
    end if
end for
return $V'$.
```

COROLLARY 4.1. *Given V_{ALG} as returned by SII-Approx for an instance of SII and optimal soluition V_{OPT} we have the following relationship:*

$$\frac{pft(V_{OPT})}{3} - \frac{n}{3}(\langle c\rangle - b) \leq pft(V_{ALG})$$

PROOF. In the proof of Theorem I.1 of [7], the authors show that $pft(V_{OPT}) \leq 3pft(V_{ALG}) - pft(\emptyset) - pft(V)$. We note that, by definition, $pft(\emptyset) = 0$ and $pft(V) = bn - \langle c\rangle n = -n(c-b)$, which gives the result. □

Note that if $\langle c\rangle \leq b$ then we recover the 1/3 approximation of [7].

5. EXPERIMENTAL RESULTS

All experiments were run on a computer equipped with an Intel X5677 Xeon Processor operating at 3.46 GHz with a 12 MB Cache and 288 GB of physical memory under the Red Hat Enterprise Linux version 6.1 operating system. Only one core was used for experiments. Our implementation of SII-Approx was written in Python 2.7 using the NetworkX library[1].

Police Dataset. We used a dataset consisting of arrest records of individuals from March 2010 - March 2013 in a single police district in Chicago. This data set included arrest location and relationships among the individuals. From this data, we were able to construct a social network ("arrest network") consisting of 1836 nodes and 2531 edges. Two individuals in the arrest network are connected if they were arrested together. We note that this is likely an incomplete picture of the full network, but as we move to deployment of this approach by integrating it with our GANG/ORCA analysis software [18], law enforcement personnel can easily supplement or replace an arrest network with information from additional intelligence sources, obersvations by police patrolmen, and data from correctional facilities.

Additionally, for some experiments we also generated simulated networks to supplement our analysis.

[1] http://networkx.github.io/

Dataset	Num. Sams.	Avg. Size	Min. Appx.	Max. Appx.	Avg. Appx.	Std. Dev.
Police	20	13.55	0.70	1.00	0.92	0.09
Compl.	2	22.5	1.00	1.00	1.00	0.00
E-R	9	20	0.84	1.00	0.93	0.05
SF	27	20	0.80	1.00	0.98	0.06
FF	1	15	1.00	1.00	1.00	0.00

Table 1: Empirically Determined Approximation Ratios for SII-Approx (when compared to the optimal solution)

Comparison with Optimal Solution. Our first test was to evaluate SII-Approx compared to an optimal solution found by enumeration. We did this by sampling the police dataset and by generating simulated networks. We generated 20 connected samples from the overall police network ranging in size from 11 to 20 nodes. We defined cost $c_v = 1$ for all $v \in V$ and we set benefit $b = 1$. The worst approximation ratio obtained in these tests was 0.70 - more than double the theoretical bound of 1/3 in this case. The average-case bound was better still at 0.92. Additionally, we also studied the behavior of SII-Approx on several standard generated graph types including complete graphs of size 20 and 25, Erdos-Reyni (E-R) random graphs, preferential-attachment generated scale-free graphs (SF), and the "Florentine Families" (FF) network [6]. In all of these tests, we never achieved an approximation ratio lower than 0.80. The results are shown in Table 1.

Runtime Evaluation. We evaluated runtime in two ways: (1) we compared the runtime of SII-Approx with an exact enumeration based computation and (2) we studied how SII-Approx scaled with network size. Both results are depicted in Figures 1 and 2. For the comparison with the exact computation, we studied the effect of runtime on our 20 samples from the police dataset. We studied at the speedup provided by SII-Approx (defined as the runtime for the exact approach divided by the runtime for SII-Approx on the same input) as a function of network size. We found a significant speedup in all cases and that the speedup increased exponentially with network size ($R^2 = 0.96$) - which is clearly due to the exponential runtime of the enumeration approach.

Runtime also scaled monotonically with the size of the network (quadratic fit, $R^2 = 0.97$). Hence, for the size of the datasets used by the Chicago police department (order 10^3 nodes), this is a viable approach with the current implementation. However, we think further improvement in runtime for the heuristic is possible with further practical modifications.

Cost Model Evaluation. One of the more useful characteristics of our framework is the ability to consider node costs. We studied two variants. First, we studied the case in which all nodes have the same cost, considering several different values. Second, we set the cost to be proportional to a network centrality measure. In both cases, we also varied the value of the benefit. We used the entire police dataset in these trials. The results for both sets of trials are shown in Figures 3 and 4.

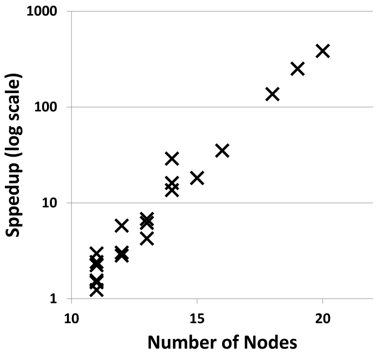

Figure 1: Network size vs. speedup obtained by SII-Approx over exact approach.

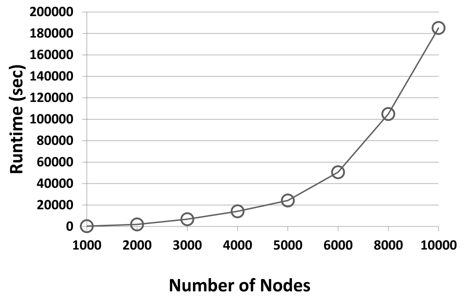

Figure 2: Network size vs. runtime for the heuristic algorithm.

We examined a fixed/uniform cost model where all nodes we assigned the same value. We examined cost values from 0.25 to 2.0 in intervals of 0.25 and compared them to uniform benefit values in the same range. In general, there was a linear relationship between the benefit and profit for all fixed cost models examined (R^2 values ranged from 0.97 to 0.99). Further, as expected, decreased cost led to increased profit.

For our centrality-based cost trials, we studied degree centrality (number of adjacent edges), closeness centrality (see [24]), eigenvector centrality (see [4]), shell number (based on shell-decomposition, see [19]), and clustering coefficient (see [24]). Cost was set to be proportional to these values for each node. We also normalized the cost so that the average would be 1.0 in each case. Just as with the fixed-cost trials, we compared the profit for various benefit values in the range $[0.25, 2.0]$ in intervals of 0.25.

For centrality-based cost models, we also observed a linear relationship between benefit and profit (R^2 values approaching 1.0). In examining the difference among centrality measures, we found the most expensive centralities were degree and shell-number followed by closeness. As these can be considered radial measures of centrality, meaning they measure centrality in terms of the number of paths that originate from a given node, then this result should be expected. Clustering coefficient was less expensive than these measures, which again was as expected as this measure is less dependent upon the number of adjacent nodes and more dependent upon the neighborhood. Perhaps most interesting was that eigenvector centrality was the "least expensive" cost model. We believe that this is due to the wide distribution of values assigned by this measure which ranged from 1.66×10^{-43} to 166.45 (compared to degree, which ranged from 0.36 to 6.17).

We note that the idea of a cost model is an important feature in our model as it has previously been shown that influential nodes are often not those who most susceptible [1]. This may imply that an individuals who may be influential in the network from a topological perspective may also be of high cost. This is why we considered cost models in our experiment where more central nodes were given a higher cost.

Heuristic for Improved Solution Quality. SII-Approx, as presented in this paper, does not take into account the order in which the vertices are selected. We found that if vertices are examined in descending order by their ClusterRank [8] then the algorithm provided a higher-profit solution when compared to our random baseline (average over 10 runs) for the case of uniform cost ($\forall v \in V, c_v = 1$) and various settings for benefit. The results are depicted in Figure 5. In [8] nodes of high ClusterRank were shown to encourage diffusion under the SIR model - which is related to the one-step process of this paper. The ClusterRank of node v is defined as follows:

$$\mathsf{cr}_v = 10^{-\mathcal{C}_v} \sum_{u \in \eta_v^{out}} (1 + k_u^{out})$$

Where $\mathcal{C}_v$ is the clustering coefficient for node v. This is particularly helpful as the computation of these measures relies only on local information and can be calculated quickly. Additionally, we examined ordering by degree, clustering coefficient, closeness centrality, shell number, and weighted degree centrality (for each $v \in V$ the quantity $\sum_{u \in V} p_{vu}$). While all of these measures showed some improvement over the random baseline, they were out performed by ClusterRank for all benefit values above 0.5. We are currently examining the performance of centrality-based ordering heuristics on a variety of inputs for the algorithm.

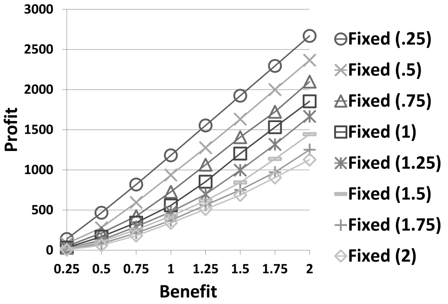

Figure 3: Profit obtained from SII-Approx for the police dataset for fixed/uniform cost models with various benefit settings.

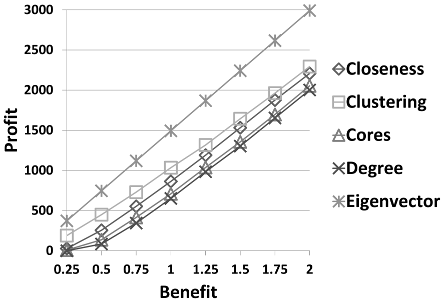

Figure 4: Profit obtained from SII-Approx for the police dataset for centrality-based cost models with various benefit settings.

Iterative Application. We envision real-world police use of SII-Approx to occur in an iterative manner. One way this could be done is as follows: we initially consider a uniform cost model and identify initial nodes to seed. Then, we calculate the diffusion process function based on that seed set in a manner that yields the probability p_v of each node v being activated. Then, for the next iteration, we remove from the network all previously seeded nodes (or whichever subset dis-enrolled from the gang) and set the cost for each node v to be $1 - p_v$. The intuition is that it will be less expensive to seed nodes that already obtained influence from other members departing the gang. The process of re-setting the cost function and social network prior to re-running SII-Approx is then performed continually.

We applied SII-Approx iteratively five times to the police dataset in the manner described above and studied the size of the set targeted as well as the resulting profit (see Figures 6-7). We observed, under the assumption that all previously seeded members left the gang, that the profit gained decreased monotonically with the number of iterations while the number of targeted vertices increased slightly in the second iteration, followed by a steep decrease and then converged to zero. The success of the second iteration indi-

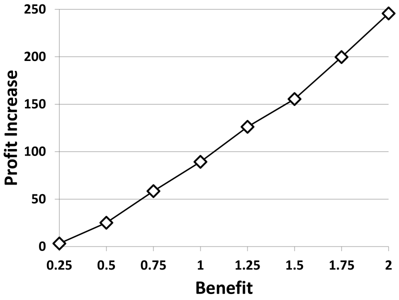

Figure 5: Improvement to profit as returned by SII-Approx when the vertices are ordered by Cluster-Rank.

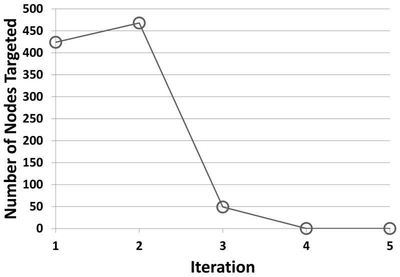

Figure 6: Iterative applications of SII-Approx – Number of nodes targetted at each iteration.

cated that viral marketing may be successful in encouraging neighboring individuals to dis-enroll. However, beyond the second step, there is limited profit to further marketing for dis-enrollment. We note that at this point, if successful, approximately half of the gang members are dis-enrolled, which would be a significant reduction. Further, we also note that topological changes to the network may become more significant after the second (and possibly even after the first) round of dis-enrollment.

Iterative application of the algorithm also opens up some new possibilities for future work. For instance, we can view our problem as a sequential decision making problem. The intuition in such an approach would be to not only to maximize the expected number of dis-enrolled gang members but also to position the law enforcement personnel to more easily influence the network in later iterations. Such an approach may also allow us to consider how the topology of the network will change over time.

6. RELATED WORK

The maximum influence problem was introduced in [14] and later studied in work such as [9, 11, 15, 21]. We refer the reader to the book [10] for a summary of recent work in this area. However, to our knowledge, no other work addresses all the challenges presented here for the SII problem simultaneously. For instance, [11] presents a model where the diffusion is restricted to shortest paths - which is a similar restriction to our one-step model, but does not consider the idea of profit. Likewise, [15] considers the idea of profit, but only applies it to the linear threshold model - which relies on the number of an individuals neighbors reaching a certain threshold. Another issue is that in most of these models, the diffusion process function is difficult to compute - for instance the *dpf* for the independent cascade model is shown to be $\#P$-hard in [11]. As a result, in most other pieces of work the diffusion process is approximated using simulation, which is as expensive operation. (Most law enforcement agencies we work with have limited computational power). One notable exception regarding this issue are deterministic models such as that described in [12, 21]. We note that in our previous work we have looked at utilizing this model in a law-enforcement setting [18, 3]. However the results of that work were primarily used to describe characteristics of the gangs and not to make operational decisions. This work did not study the operational issues associated with encouraging gang dis-enrollment as considered in this paper.

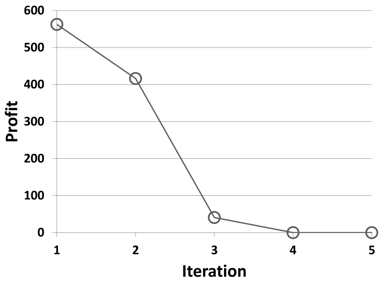

Figure 7: Iterative applications of SII-Approx – Profit obtained at each iteration.

7. CONCLUSION

In this paper we introduced the "social incentive influence" (SII) problem, a variant of the maximum influence problem, designed to help law-enforcement personnel identify members of street gangs that they can encourage to dis-enroll. We studied this problem both formally and experimentally in the context of the law-enforcement domain. Utilizing techniques from unconstrained submodular maximization, we developed a heuristic technique to help police better identify sets of influential individuals to target with dis-enrollment incentives. We implemented our approach and performed an experimental evaluation. We currently have our approach to the SII problem integrated into our GANG/ORCA analysis software [18] that is currently in use by the Chicago Police. Our next goal is to work with law enforcement personnel to better understand how SII is employed in practice - allowing us to identify components of this framework that can be adjusted for improved results in a real-world setting.

8. ACKNOWLEDGMENTS

The authors are supported by the Army Research Office (project 2GDATXR042) and the U.S. Air Force. The opinions in this paper are those of the authors and do not necessarily reflect the opinions of the funders, the U.S. Military Academy, the U.S. Army, the U.S. Air Force, or the U.S. Department of Defense.

9. REFERENCES

[1] S. Aral and D. Walker. Identifying influential and susceptible members of social networks. *Science*, 337(6092):337–341, 2012.

[2] J. Bertetto. Countering criminal street gangs: Lessons from the counterinsurgent battlespace. *Law Enforcement Executive Forum*, 12(3):43, 2012.

[3] J. Bertetto. Counter-gang strategy: Adapted coin in policing criminal street gangs. *Law Enforcement Executive Forum*, 13(3), 2013.

[4] P. Bonacich. Factoring and weighting approaches to status scores and clique identification. *The Journal of Mathematical Sociology*, 2(1):113–120, 1972.

[5] A. Braga, D. Hureau, and A. Papachristos. Deterring gang-involved gun violence: Measuring the impact of bostonâĂŹs operation ceasefire on street gang behavior. *Journal of Quantitative Criminology*, pages 1–27, 2013.

[6] R. L. Breiger and P. E. Pattison. Cumulated social roles: The duality of persons and their algebras. *Social Networks*, 8(3):215–256, Sept. 1986.

[7] N. Buchbinder, M. Feldman, J. S. Naor, and R. Schwartz. A tight linear time (1/2)-approximation for unconstrained submodular maximization. In *Proceedings of the 2012 IEEE 53rd Annual Symposium on Foundations of Computer Science*, FOCS '12, pages 649–658, Washington, DC, USA, 2012. IEEE Computer Society.

[8] D.-B. Chen, H. Gao, L. LÃij, and T. Zhou. Identifying influential nodes in large-scale directed networks: The role of clustering. *PLoS ONE*, 8(10):e77455, 10 2013.

[9] N. Chen. On the approximability of influence in social networks. *SIAM J. Discret. Math.*, 23:1400–1415, September 2009.

[10] W. Chen, L. V. Lakshmanan, and C. Castillo. *Information and Influence Propagation in Social Networks*. Morgan and Claypool Publishers, 2013.

[11] W. Chen, C. Wang, and Y. Wang. 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*, KDD '10, pages 1029–1038, New York, NY, USA, 2010. ACM.

[12] P. Dreyer and F. Roberts. Irreversible -threshold processes: Graph-theoretical threshold models of the spread of disease and of opinion. *Discrete Applied Mathematics*, 157(7):1615 – 1627, 2009.

[13] M. R. Garey and D. S. Johnson. *Computers and Intractability; A Guide to the Theory of NP-Completeness*. W. H. Freeman & Co., New York, NY, USA, 1979.

[14] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146, New York, NY, USA, 2003. ACM.

[15] W. Lu and L. Lakshmanan. Profit maximization over social networks. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*, pages 479–488, Dec 2012.

[16] J. C. Nekola and P. S. White. Special Paper: The Distance Decay of Similarity in Biogeography and Ecology. *Journal of Biogeography*, 26(4):867–878, 1999.

[17] G. L. Nemhauser, L. A. Wolsey, and M. Fisher. An analysis of approximations for maximizing submodular set functionsï£¡i. *Mathematical Programming*, 14(1):265–294, 1978.

[18] D. Paulo, B. Fischl, T. Markow, M. Martin, and P. Shakarian. Social network intelligence analysis to combat street gang violence. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '13, pages 1042–1049, New York, NY, USA, 2013. ACM.

[19] S. B. Seidman. Network structure and minimum degree. *Social Networks*, 5(3):269 – 287, 1983.

[20] P. Shakarian, M. Broecheler, V. S. Subrahmanian, and C. Molinaro. Using generalized annotated programs to solve social network diffusion optimization problems. *ACM Trans. Comput. Logic*, 14(2):10:1–10:40, June 2013.

[21] P. Shakarian and D. Paulo. Large social networks can be targeted for viral marketing with small seed sets. *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 1–8, 2012.

[22] H. Skov-Petersen. Estimation of distance-decay parameters: GIS-based indicators of recreational accessibility. In *ScanGIS*, pages 237–258, 2001.

[23] P. J. Taylor. Distance transformation and distance decay functions. *Geographical Analysis*, 3(3):221–238, 1971.

[24] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Number 8 in Structural analysis in the social sciences. Cambridge University Press, 1 edition, 1994.