

# SINAS: Suspect Investigation Using Offenders’ Activity Space

Mohammad A. Tayebi<sup>1</sup>, Uwe Glässer<sup>1</sup>,  
Patricia L. Brantingham<sup>2</sup>, and Hamed Yaghoubi Shahir<sup>1</sup>

<sup>1</sup>School of Computing Science, Simon Fraser University

<sup>2</sup>School of Criminology, Simon Fraser University

{tayebi, glaesser, pbrantin, syaghoub}@sfu.ca

**Abstract.** Suspect investigation as a critical function of policing determines the truth about how a crime occurred, as far as it can be found. Understanding of the environmental elements in the causes of a crime incidence inevitably improves the suspect investigation process. Crime pattern theory concludes that offenders, rather than venture into unknown territories, frequently commit opportunistic and serial violent crimes by taking advantage of opportunities they encounter in places they are most familiar with as part of their activity space. In this paper, we present a suspect investigation method, called **SINAS**, which learns the activity space of offenders using an extended version of the random walk method based on crime pattern theory, and then recommends the top- $K$  potential suspects for a committed crime. Our experiments on a large real-world crime dataset show that **SINAS** outperforms the baseline suspect investigation methods we used for the experimental evaluation.

## 1 Introduction

Crime is a purposive deviant behavior that is an integrated result of different social, economical and environmental factors. Crime imposes substantial costs on society at individual, community, and national levels.

An important policing task is investigating committed or reported crimes—known as criminal or *suspect investigation*. Spatial studies of crime, and more specifically environmental criminology, play an essential role in criminal intelligence [1, 6, 8, 5, 7].

Modeling spatial aspects of criminal behavior can be seen as an intractable case of the human mobility problem [4, 2]. This is mainly because available information about the spatial life of offenders is usually limited to police arrest data on their home and crime locations. Further, spatial displacement of crime is a common phenomenon, meaning offenders shift their crime locations.

In this paper, we propose an approach to Suspect INvestigation using offenders’ Activity Space, called **SINAS**. It first learns activity space of offenders based on crime pattern theory, using existing crime records. Then, given the location of a newly occurred crime, **SINAS** ranks known offenders based on their activity space that influences offenders’ criminal activity, and finally it recommends the top- $K$  suspects of that crime with the highest probability. Our experiments on a large real-world crime dataset show that **SINAS** outperforms the baseline suspect investigation methods significantly.

Section 2 explores related work, and Section 3 presents the fundamental concepts. Next, Section 4 introduces the proposed model, and Section 5 discusses the experimental evaluation and the results. Section 6 concludes the paper.

## 2 Background and Related Work

People do not move randomly across urban landscapes. For the most part, they commute between a handful of routinely visited places such as home, work, and their favorite places. With each and every trip, they get more familiar with, and gain new knowledge about, these places and everything along the way. A person will eventually be at ease with a place and it becomes part of their activity space (see Figure 1). Nodes and Paths are two main components of an activity space. The (activity) nodes are the locations that the person frequents (e.g., workplace, residence, recreation).

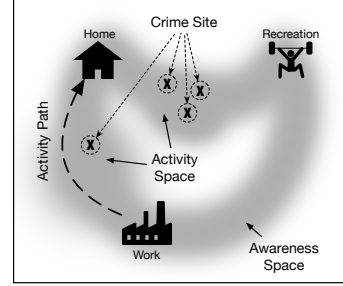


Fig. 1: Activity Space

These are the endpoints of a journey. The (activity) paths connect the nodes and represent the person's path of travel between nodes. Activity space of offenders is explored in several studies. The geographic profiling method by Rossmo [8] is widely recognized for inferring the activity space of an offender to determine the likely home location based on their crime locations. This method assumes that offenders select targets and commit crimes near their homes. Frank [3] proposes an approach to infer the activity paths of all offenders in a region based on their crime and home locations. Assuming the home location as the center of an offender's movements, the orientation of activity paths of each individual offender is calculated so as to determine the major directions, relative to their home location, into which they tend to move to commit crimes.

Based on criminological theories, several studies propose mathematical models for spatial and temporal characteristics of crime to predict future crimes. For instance, in [7], the authors use a point-pattern-based transition density model for crime space-event prediction. This model computes the likelihood of a criminal incident occurring at a specified location based on previous incidents. In [9], the authors model the emergence and dynamics of crime hotspots by using a two-dimensional lattice model for residential burglary, where each location is assigned a dynamic attractiveness value, and the behavior of each offender is modeled with a random walk process.

Our proposed method, **SINAS**, addresses the problem of recommending most likely suspects based on historical spatial information, which is a problem that none of the existing methods is able to address. Thus, there is a challenge in evaluating our experimental results. The model in [7] predicts only the time and location of crimes at an aggregate level. The method proposed in [8] discovers offender home locations based on their crime locations. Finally, the method in [3] finds locations which are centers of interest for committing crime.

### 3 Crime Data Characteristics

We evaluate the efficacy of our approach on a crime dataset representing five years (2001-2006) of police arrest-data for the Province of British Columbia, comprising several million data records, each refers to a reported offence<sup>1</sup>. Our experiments consider all subjects in four main categories: *charged*, *chargeable*, *charge recommended*, and *suspect*. Being in one of these categories means that the police is serious enough about the subject's involvement in a crime as to warrant calling them 'offender'. Here, we concentrate on crimes in Metro Vancouver (population: 2.46 M), with different regions connected through a road network composed of road segments having an average length of 0.2 km (see Table 1).

Table 1: Statistical properties of the used dataset

Property	Value	Property	Value
#crimes	125,927	#offenders	189,675
#offenders with more then 1 crime	25,162	#co-offending links	68,577
#co-offenders in co-offending network	17,181	avg. node degree in co-offending network	4
#road-segments	64,108	avg. crime per road segment	2

Figures 2a and 2b illustrate the distribution function of crime incidents per offender and per road segment respectively. Both distributions have heavy-tailed pattern. 83% of the offenders committed only one crime, while less than 1% of the offenders committed 10 or more crimes. Further, 38% of the road segments are linked to at least one crime and 9% of the road segments are linked to 10 or more crimes. Half of all the crimes occurred in only 1% of all road segments, and a total of 25% of all the crimes occurred in only 100 road segments. The average home to crime location distance of 80%, 63% and 40% of all offenders is less than 10 km, 5 km and 2 km, respectively. The average crime location distance of 73%, 52% and 26% of all offenders is less than 10 km, 5 km and 2 km, respectively. One can assume that frequent offenders are generally mobile and have several home locations identified in their records. 41% of the offenders who committed more than one crime have more than one home location.

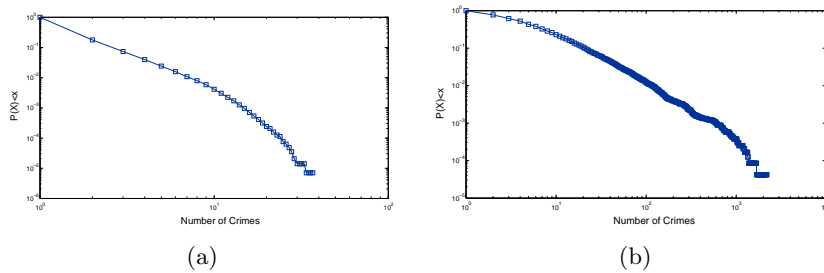


Fig. 2: Distribution of: (a) crimes per offender; (b) crimes per road segment

<sup>1</sup> This data was made available for research purpose by Royal Canadian Mounted Police (RCMP) and retrieved from the Police Information Retrieval System (PIRS).

### 3.1 Fundamental Concepts and Definitions

This section introduces the fundamental concepts, definitions, and notations.

**Offender.** Let  $V$  be a set of offenders and  $C$  be a set of crimes. Each crime event  $e \in C$  involves a non-empty subset of criminal offenders  $U \subseteq V$ .  $C_i$  is the set of crimes committed by offender  $u_i$ . With each crime incident we associate a type of crime, a time when the crime occurred as well as longitude and latitude coordinates of the crime location and home location of involved offenders.

**Co-offending Network.** A *co-offending network* is an undirected graph  $G(V, E)$ . Each node represents a known offender  $u_i \in V$ . Offenders  $u$  and  $v$  are connected,  $u_i, u_k \in V$  and  $(u_i, u_k) \in E$ , if they are known to have committed one or more offences together, and are not connected otherwise.  $\Gamma_i$  denotes the set of neighbors of offender  $u_i$  in the co-offending network.

**Road Network.** Intuitively, a road network can be decomposed into *road segments*, each of which starts and ends at an intersection. We use the *dual* representation where the role of roads and intersections is reversed. All physical locations along the same road segment are mapped to the same node. Formally, a road network is an undirected graph  $R(L, Q)$ , where  $L$  is a set of nodes, each representing a single road segment. Road segments  $l_j$  and  $l_k$  are connected,  $\{l_j, l_k\} \in Q$ , if they have an adjacent intersection in common. Crime locations within a studied geographic boundary are mapped to the closest road segment. Henceforth, the term “road” is used to refer to a road segment.

**Road Features.** A vector  $\bar{y}_j$  denotes the features of the road  $l_j$  including road length  $d_j$ , and *road attractiveness* features vector  $\bar{a}_j$ . Further,  $\bar{a}_j$  is a vector of size  $m$  where the value of the  $k^{th}$  entry of  $\bar{a}_j$  corresponds to the total number of crimes of type  $k$  committed previously at  $l_j$ .  $\Pi_j$  denotes the set of neighbors of road  $l_j$  in the road network.  $\Delta \subset L$  denotes a set of roads with the highest crime rate, called crime *hotspots*.  $D_{l_j, l_k}$  is the shortest path distance of road  $l_j$  from road  $l_k$ , and  $f_j$  denotes the total number of crimes at road  $l_j$ .

**Anchor Locations.**  $L_i$  is the set of roads at which offender  $u_i$  has been observed, including all of his known home and crime locations.  $f_{i,j}$  and  $t_{i,j}$  respectively denote the frequency and the last time  $u_i$  was at anchor location  $l_j$ . *Offender trend* is given by a vector  $\bar{x}_i$  of size  $m$  which indicates the crime trend of  $u_i$  as extracted from his criminal history. That is, the value of the  $k^{th}$  entry of  $\bar{x}_i$  corresponds to the number of crimes of type  $k$  committed by offender  $u_i$ .

### 3.2 Problem Scope

Crime analysis captures a broad spectrum of facets pertaining to different needs and using different analytical methods. For instance, intelligence analysis aims at recognizing relationships between criminal network actors to identify and arrest offenders. It typically starts with a known crime problem or identified co-offending network, then uses these resources to collect, analyze and compile information about a predetermined target. An important problem is to identify most likely perpetrators of previous crimes. Criminal profiling approaches contribute to criminal intelligence using offenders’ characteristics. Also, methods

like geographic profiling build on environmental criminology theories and use information related to the environment of offenders and crimes.

**Problem Definition—Suspect Investigation:** In the following, we formally address the problem of suspect investigation. Assume there is a collection of crime records,  $C$ , from past crime events where each element in  $C$  uniquely identifies a single crime incident. When a new crime incident  $e$  occurs, police investigates suspects who potentially committed  $e$  based on the existing information, that is, anchor locations (home and crime locations) of every offender in  $C$  mapped on a road network, the type of crimes they committed, and also the (known) co-offending network  $G$  extracted from  $C$ . The problem definition in abstract formal terms is as follows:

*Given a crime dataset  $C$  and new crime incident  $e$  at location  $l_e$ , the goal is to recommend the top- $K$  suspects for  $e$  with the highest probability.*

Geographic profiling addresses a similar problem of detecting home locations of suspects of a crime incident, given a series of past crimes. However, the novelty of our approach is two-fold: 1) it directly targets the identity of offenders rather than their home locations; and 2) the input of our approach is a single crime incident, while the input of geographic profiling is a series of crimes.

## 4 SINAS Method

### 4.1 Learning Activity Space

A *random walk* over a graph is a stochastic process in which the initial state is known and the next state is decided using a transition probability matrix that identifies the probability of moving from a node to another node of the graph. Under certain conditions, the random walk process converges to a stationary distribution assigning an importance value to each node of the graph. The random walk method can be modified to satisfy the locality aspect of crimes, which states that offenders do not attempt to move far from their anchor locations. For instance, the random walk method works locally if the likelihood of terminating the walk increases with the distance from the anchor locations.

In our proposed model, starting from an anchor location the offender explores the city through the underlying road network. At each road, he decides whether to proceed to a neighboring road or return to one of his anchor locations. The random walk process continues until it converges to a steady state which reflects the probability of visiting a road by the offender. This probability can be relevant to the offender's exposure to a crime opportunity.

For learning the activity space of an offender, we need to understand his daily life and routines. However, in the crime dataset, generally we miss the *paths* completely and the *nodes* partially. Thus, we improve our incomplete knowledge about offenders with available information in the dataset by defining two different sets of anchor locations: 1) *main anchor locations*, denoted by  $\mathcal{L}_i$  for offender  $u_i$ , is an extension of the offender anchor locations by adding his co-offenders' anchor locations with the assumption that friends in the co-offending network are

likely to share the same locations; and 2) *intermediate anchor locations*, denoted by  $\mathcal{I}_i$  for offender  $u_i$ , is the roads closest to the set of his main anchor locations, using a Gaussian model (see Section 4.1–Starting Probabilities for details).

An offender starts his random walk either from a main or intermediate anchor location. Given that the actual trajectories in an offender’s journey to crime are unknown, SINAS guides offender movements in directions with a higher chance of committing a crime. This is done by taking into account different aspects that influence the offender movement directionality in computing the transition probabilities in a random walk.

**Random Walk Process:** For each single offender  $u_i$ , we perform a series of random walks on the road network  $R(L, Q)$ . The random walk process starts from one of the anchor locations of  $u_i$  with predefined probabilities (see Section 4.1–Starting Probabilities) and traverses the road network to locate a criminal opportunity. At each step  $k$  of the random walk, the offender is at a certain road  $l_j$  and makes one of two possible decisions: 1) With probability  $\alpha$ , he decides to return to an anchor location and not look for a criminal opportunity this time, choosing an anchor location as follows: a) with probability  $\beta$ , he decides to return to a main anchor location  $l \in \mathcal{L}_i$ , and b) with probability  $1 - \beta$ , he returns to an intermediate anchor location  $l \in \mathcal{I}_i$ ; and 2) With probability  $1 - \alpha$  he continues looking for a crime opportunity. If he continues his random walk then he has two options in each step of the walk: a) with probability  $\theta(u_i, l_j, k)$ , he stops the random walk, which means the offender commits a crime at road  $l_j$ , and b) with probability  $1 - \theta(u_i, l_j, k)$  he continues the random walk, moving to another road which is a direct neighbor of  $l_j$ .

To continue the random walk at road  $l_j$ , we select a direct neighbor road from  $\Pi_{l_j}$ . Function  $\phi$  computes the transition probability from a road segment to one of its neighbor road segments (see Section 4.1–Movement Directionality for details). The probability of selecting road segment  $l_r$  in the next step is:

$$P(l_j \rightarrow l_r) = \frac{\phi(l_r)}{\sum_{l_p \in \Pi_{l_j}} \phi(l_p)} \quad (1)$$

The probability of being at road  $l_r$  at step  $k + 1$  given that the offender was at road  $l_j$  at step  $k$  is shown in Equation 2, where  $X_{u_i, k}$  is the random variable for  $u_i$  being at road  $l_r$  in step  $k$ . We terminate the random walks when  $\|F^{m+1}\| - \|F^m\| \leq \epsilon$ , where  $F^m = (F(u_i, l_1) \dots F(u_i, l_{|L|}))^T$  is the results for  $u_i$  after  $m$  random walks. For some offenders the random walks do not converge, in which case we terminate the overall process at  $m > 10000$ .

$$P(X_{u_i, k+1} = l_r | X_{u_i, k} = l_j) = (1 - \alpha)(1 - \theta_{l_j, k}) \times P(l_j \rightarrow l_r) \quad (2)$$

**Starting Probabilities:** The model distinguishes two types of starting nodes. 1) *Main anchor locations* are all anchor locations of a single offender and his co-offenders:  $\mathcal{L}_i = L_i \cup \{l_j : l_j \in L_v, v \in \Gamma_u\}$ . The rationale is that offenders who have collaborated in the past likely may have shared information on anchor locations in their activity space, an aspect that possibly affects their choice of

future crime locations. In computing the starting probability of each anchor location, the two primary factors are the *frequency* and the *last time* an offender visited an anchor location. The probability that offender  $u_i$  starts his random walk from  $l_j$  is shown in Equation 3, where  $t$  is the current time, and  $\rho$  is the parameter controlling the effect of the timing.

$$S(i, j) = \frac{f_{i,j} \times e^{\frac{-(t-t_{i,j})}{\rho}}}{\sum_{l_k \in \mathcal{L}_i} f_{i,k} \times e^{\frac{-(t-t_{i,k})}{\rho}}} \quad (3)$$

2) *Intermediate anchor locations* are the closest locations to main anchor locations. Human mobility models use Gaussian distributions to analyze human movement around a particular point such as home or work location [4]. We assume that offender movement around their main anchor locations follows a Gaussian distribution. Each main anchor location of offender  $u_i$  is used as the center, and the probability of  $u_i$  being located in a road is modeled with a Gaussian distribution. Given road  $l$ , the probability of  $u_i$  residing at  $l$  is:

$$S(i, l) = \sum_{l_j \in \mathcal{L}_i} \frac{f_{i,j}}{\sum_{l_k \in \mathcal{L}_i} f_{i,k}} \frac{\mathcal{N}(l|\mu_{l_j}, \Sigma_{l_j})}{\sum_{l_k \in \mathcal{L}_i} \mathcal{N}(l|\mu_{l_k}, \Sigma_{l_k})} \quad (4)$$

Here  $l$  is a road which does not belong to the set of main anchor locations.  $\mathcal{N}(l|\mu_{l_j}, \Sigma_{l_j})$  is a Gaussian distribution for visiting a road when  $u_i$  is at anchor location  $l_j$ , with  $\mu_{l_j}$  and  $\Sigma_{l_j}$  as mean and covariance. We consider the normalized activity frequency of  $u_i$  at  $l_j$ , meaning that a main anchor location with higher activity frequency has higher importance. For offender  $u_i$ , the roads with the highest probability of being an intermediate anchor location are added to the set  $\mathcal{L}_i$  as additional starting nodes besides the main anchor locations.

**Movement Directionality:** The creation of the main attractor nodes and paths are developed through normal mobility shaped by the urban backcloth or urban environment. Each individual has normal, routine pathways or commuting/mobility routes that are unique. However, the environment where we live influences our actions and movements. Highways, streets and road networks in general guide us to our destinations such as home, workplace, recreation center, and business establishments. In the aggregate, individuals routes overlap or intersect in time and space. These areas of overlap often have rush hours and congestion at intersections or mass transit stops associated with handling large numbers of people. These high activity locations can become crime attractors and crime generators when there are enough suitable targets in those locations. Crime attractors and generators affect directionality of offenders' movement.

One can conclude that starting from an anchor location the probability of offender movement toward crime attractors and generators is higher. To address this fact, in the random walk process, the transition probability is computed based on the proximity of a road to the crime hotspots and the importance of each crime hotspot, which is proportional to the number of crimes committed

there. Function  $\phi(l_j)$  is used in computing the transition probability (see Section 4.1–Random Walk Process) of moving offender  $u_i$  from  $l_k$  to  $l_j$ , where  $f_n$  is the number of crimes committed at  $l_n$ .  $D_{j,n}$  is the distance of road  $l_j$  from the hotspot  $l_n \in \Delta$ , which is equal to the length of shortest path between two roads.

$$\phi(l_j) = \sum_{n=1}^{|\Delta|} f_n \times \frac{1}{D_{j,n}} \quad (5)$$

**Stopping Criteria:** The probability of stopping the random walk for an offender at a given road corresponds to the probability of this offender committing a crime in that road segment. Two factors influence the stopping probability of offender  $u_i$  in the road  $l_j$ . The first one relates to the similarity of the crime trend of offender  $u_i$  and the criminal attractiveness of road  $l_j$ , where higher similarity means a higher chance of stopping  $u_i$  at  $l_j$ . The second factor is the distance of  $l_j$  from the starting point measured in the number of steps ( $k$ ) from the starting point. To satisfy the locality aspect of crimes, the probability of continuing the random walk decrease while getting farther from the starting point. Thus, the stopping probability (Equation 6) is inversely proportional to  $k$ . Also,  $\text{sim}(i, j)$  denotes the cosine similarity of crime trend of  $u_i$  and the attractiveness of  $l_j$ .

$$\theta(u_i, l_j, k) = \text{sim}(i, j) \times \frac{1}{1 + e^{\frac{-k}{2}}} \quad (6)$$

## 4.2 Suspect Recommendation

The crime location is neither even nor random, however, there is an underlying spatial pattern in it. Environmental criminology theories [1] suggest that crimes occur in predictable ways, at offenders' *awareness space* which includes their activity space. To recommend the most likely suspects of a new crime incident based on the learnt offenders' activity space, we rank offenders based on the proximity of the crime location to their activity spaces. An offender is considered as a 'potential suspect' if the crime location is close enough to the activity space of this offender. This approach is based on a crime pattern theory stating that future crime locations are within offenders' activity space which is dependent to their activity nodes and paths. To influence offenders' characteristics, we consider the history of the offenders including the *types* and *number* of their committed crimes. The probability of offender  $u_i$  commits a *new* crime  $e$  is computed in Equation 7, where  $\mathcal{T}(C_i, e)$  is a boolean function that returns *one* if in the crime records of offender  $u_i$ ,  $C_i$ , there is a crime event with the same type as crime  $e$ , and *zero* otherwise.  $\omega$  is a parameter that controls the influence of function  $\mathcal{T}$ .  $|C_i|$  is the number of crimes that offender  $u_i$  committed previously.  $F(u_i, l_k)$  is the probability of  $l_k$  being in activity space of offender  $u_i$ .  $D_{l_k, l_e}$  is the distance between roads  $l_k$  and  $l_e$ .

$$Z(u_i, e) = \omega \mathcal{T}(C_i, e) \times |C_i| \times \sum_{k=1}^n F(u_i, l_k) \times D_{l_k, l_e} \quad (7)$$



## 5 Experimental Evaluation

We divide the crime dataset chronologically into train and test data. The train data, used to learn the activity space of offenders, includes all crimes that happened in the first 54 months, and the test data includes the remaining six months. The crimes in the test data committed by known offenders are used for suspect investigation. SINAS recommends the top- $K$  suspects most likely to commit a new occurred crime.  $K$  is set to 50 in our experiments, but relative results for other values of  $K$  are also consistent. We use the recall measure (i.e., the percentage of crimes in which the offender who committed that crime appears in the list of top- $K$  recommended suspects) to evaluate the quality of methods. Before discussing the results, it is crucial to specify the experimental setting.

1) On the one hand, if a new crime occurs in a location where an offender has been observed previously (as his anchor location) then the probability that the same offender is involved in the new crime is higher, and this fact makes the investigation process easier. Formally speaking: assume a crime  $e$  committed by offender  $u_i$  at  $l_j$ . If  $l_j \in L_i$ , then we consider  $e$  as an *easy case*; and, if  $l_j \notin L_i$ , then we consider  $e$  as a *hard case*. We therefore define two scenarios: *easy scenario* which includes all (union of easy and hard) cases and *hard scenario* which only includes hard cases. We compare the performance of SINAS for both easy and hard scenarios.

2) On the other hand, repeat offenders are responsible for large percentage of committed crimes and there has long been an interest in the behavior of repeat offenders since controlling these groups of offenders can significantly reduce the overall crime level. Therefore, we distinguish two groups of offenders: repeat offenders with 10 or more crimes and non-repeat offenders with less than 10 crimes. We compare the performance of SINAS for repeat offenders with all (union of repeat and non-repeat) offenders.

### 5.1 Baseline Methods

As discussed in Section 2, there is no suspect investigation method using offenders' spatial information to the best of our knowledge; however, we evaluate the SINAS performance in comparison with some baseline methods. We also perform experiments on different settings of SINAS to learn the meaningfulness of its three principal elements: 1) the probabilistic aspect of offenders' activity space, 2) offenders' crime types, and 3) the frequency of crimes committed by offenders. Following is the list of comparison partners in our experiments.

**SINAS-PPN** takes the probabilistic aspect of offenders' activity space and their crime types into account while ignoring the frequency of committed crimes.

**SINAS-PNP** takes the probabilistic aspect of offenders' activity space and frequency of committed crimes into account while ignoring the type of crimes.

**SINAS-NPP** ignores the probabilistic aspect of offenders' activity space but considers type and frequency of crimes committed by them.

**SINAS** takes all available information including the probabilistic aspect of offenders' activity space, frequency and type of committed crimes into account.

**CrimeFrequency** ranks offenders based on the number of crimes they have committed and includes top- $K$  offenders with the highest crime number in the recommendation list. The intuition behind this method is that *repeat offenders* are more probable to be involved in a new occurred crime.

**Proximity** uses a distance-decay function to compute the proximity of offenders' anchor locations from the location of a new crime. It considers the frequency of being an offender in each of his anchor locations as a factor of their importance. Proximity is comparable to the geographic profiling approach [8].

**Random** recommends suspects randomly from the pool of known offenders.

## 5.2 Experimental Results

Table 2 shows the performance of different variations of **SINAS** and the other baseline methods for both easy and hard scenarios. For both scenarios, **SINAS** outperforms the other baseline methods and significantly outperforms Proximity and CrimeFrequency. Interestingly, CrimeFrequency has a good performance in the easy scenario. In the easy scenario that crimes with known locations for offenders are included, Proximity gets the advantage of having those locations exactly in the offender's anchor locations, and therefore it is able to successfully recommend the suspects. CrimeFrequency has the weakest performance but still works much better than Random recommendation method.

Table 2: Recall (%) of different suspect investigation methods ( $K=50$ ) for hard and easy scenarios considering all offenders (repeat and non-repeat)

	SINAS-PPN	SINAS-PNP	SINAS-NPP	SINAS	CrimeFrequency	Proximity	Random
Hard Scenario	3.8	5.1	4.2	5.4	1.5	3.7	0.002
Easy Scenario	10.4	11.9	5.3	12.1	1.3	10.3	0.002

In our experimental setting, the number of potential suspects is about 25,000 and  $K = 50$ . **SINAS** recall is more than 5% and 12% respectively in the hard and easy scenarios. Contrary to geographic profiling which receives a series of crime locations as an input and criminal profiling which may have rich information about suspects to reduce the search space, **SINAS** only uses the location of a single crime, and we thus believe this result is a significant contribution to the difficult task of suspect investigation.

Looking at the experimental results of the **SINAS** variations, we notice all three elements of **SINAS** contribute to the method performance. Offenders' crime frequency has the most contribution and offenders' crime type has the least influence on the **SINAS** performance. As already discussed, a large percentage of crimes are committed by repeat offenders and taking this fact into account significantly improves the **SINAS** performance. As described in Section 3, in only half of the repeat offenders we observe strong patterns in their criminal trend. Recognizing complex and latent patterns in criminal activities to serve the suspect investigation task needs a more thorough study of offenders' trend which is beyond the scope of this paper.

Figure 3a shows the performance of SINAS for different values of  $K$ . As expected, the recall value is increased by increasing  $K$ , reaching to 9% and 16% in hard and easy scenarios for  $K = 100$ . Considering the major cost of investigation process for the law enforcement more specifically in serious crimes, using greater values of  $K$  to reduce the search space and optimize the spent cost and time is reasonable. Figure 3b shows the performance of SINAS for the repeat offenders with respect to different values of  $K$ . For  $K = 50$ , SINAS has the recall of 25% and 38% in the hard and easy scenarios. For the repeat offenders that we know more about their spatial activities, the SINAS performance is about two times greater than the method performance for all offenders.

For studying the SINAS performance for repeat and not-repeat offenders, we categorize offenders based on the number of crimes they have committed. Figure 3c shows the SINAS recall for each of these groups of offenders. As depicted, the SINAS performance increases linearly by increasing the number of crimes of the corresponding group, meaning that suspect investigation for a group of offenders who committed more crimes is generally more successful.

SINAS and criminal profiling approaches can be used as complementary tools for suspect investigation. Assume that for a new occurred crime, the police is able to guess the age of the offender based on evidence and witness interviews. Using this piece of information reduces the search space and increases the chance of success. In the following, we discuss the experimental results of applying SINAS on this subset of offenders instead of all offenders. Figure 3d shows the result for this suspect intelligence scenario ( $K = 50$ ), where the x-axis shows the exactness of our knowledge about the age (#years) of the offender. In other words, if the offender exact age is  $a$ , then the value  $b$  on x-axis means SINAS considers offenders with ages in the interval of  $[a - b, a + b]$ . As shown, having more precise information on the offender's age contributes more to the intelligence process.

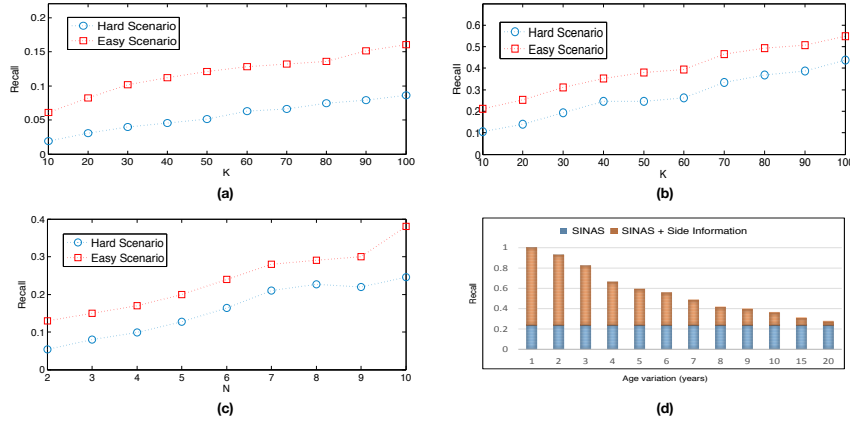


Fig. 3: SINAS performance in the easy and hard scenarios for: (a) different values of  $K$ , (b) repeat offenders respect to different values of  $K$ , and (c) group of offenders with greater than or equal  $N$  crimes ( $K=50$ ); (d) SINAS performance for repeat offenders in the hard scenario considering offender's age range ( $K=50$ )

With  $b = 1$ , SINAS is able to investigate all crimes successfully, and even  $b = 20$  improves the SINAS performance compared to the situation of having no side information. This result shows the importance of side information in the suspect investigation process.

## 6 Conclusions

This paper proposes the SINAS method for suspect investigation by analyzing the activity space of offenders. It utilizes an extended version of the random walk method and learns the activity space of offenders based on a widely accepted criminological theory, crime pattern theory. Our experimental results show: 1) learning the activity space of offenders from their spatial life contributes to high-quality suspect recommendation; 2) utilizing offenders' criminal trend improves suspect recommendation. Not only does SINAS significantly outperform baseline methods for both repeat and non-repeat offenders, but it also has more satisfying results for repeat offenders where there is more information available on their spatial activities; and 3) SINAS and criminal profiling approaches can be viewed as complementary tools for suspect investigation.

Data mining-based suspect investigation is a multi-step process that has significant operational challenges in practice. Three main steps of this process—question formulation, data preparation, and data mining—have been addressed in our proposed method. However, the ultimate steps, deployment and efficacy evaluation, are beyond the scope of this paper. Making a difference in real-world situations, calls for an iterative process where law enforcement and policymakers act on analytics inferred from data mining-based suspect investigation methods at the strategic, tactical and operational levels.

## References

1. P. J. Brantingham and P. L. Brantingham. *Environmental Criminology*. Sage Publications, 1981.
2. D. Brockmann, L. Hufnagel, and T. Geisel. The scaling laws of human travel. *Nature*, 439(7075):462–465, 2006.
3. R. Frank and B. Kinney. How many ways do offenders travel – evaluating the activity paths of offenders. In *Proceedings of the 2012 European Intelligence and Security Informatics Conference (EISIC'12)*, pages 99–106, 2012.
4. M. C. Gonzalez, C. A. Hidalgo, and A. Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.
5. W. Gorr and R. Harries. Introduction to crime forecasting. *International Journal of Forecasting*, 19(4):551–555, 2003.
6. K. Harries. Mapping crime principle and practice. *U.S. Department of Justice, Office of Justice Programs, National Institute of Justice*, 1999.
7. H. Liu and D. E. Brown. Criminal incident prediction using a point-pattern-based density model. *International Journal of Forecasting*, 19(4):603–622, 2003.
8. D. K. Rossmo. *Geographic Profiling*. CRC Press, 2000.
9. M. B. Short, M. R. D'orsogna, V. B. Pasour, G. E. Tita, P. J. Brantingham, A. L. Bertozzi, and L. B. Chayes. A statistical model of criminal behavior. *Mathematical Models and Methods in Applied Sciences*, 18(supp01):1249–1267, 2008.