

Swarming Geographic Event Profiling, Link Analysis, and Prediction

Sven A. Brueckner

Vector Research Center, a Division of TechTeam Government Solutions, Inc.
3250 Green Court, Suite 250, Ann Arbor, MI 48105
sven.brueckner@newvectors.net

Abstract

Geographically embedded processes with hidden origins are often observable in events they generate. It is common practice in criminological forensics to reverse simple equation-based models of trajectories that link the origin with its events to derive a probability estimate of a common origin location. This approach requires that linked events are manually extracted from larger event data sets. Also, as this approach is equation-based, it is generally not possible to take into account any specific geographic characteristics that may affect the trajectory.

In this paper, we present a swarming model of simple geographic agents who reason “backward” from large sets of events to origin location probability distributions, use the overlap of these distributions to identify clusters of events that may share a common origin, and then reason “forward” from the clusters’ origin distributions to predict the risk of future events. We apply this model to the domain of Improvised Explosive Devices (IEDs).

Keywords: geographic profiling, prediction, application, self-organization, pheromones

1 Introduction

In various domains there are processes embedded in a geographic context whose patterns of observable events are indicative of their likely origin but strongly influenced by the respective local geographic context. For example, the pattern of discovery of mineral samples in field geology is shaped by the location of the source deposit and the local characteristics of the erosion process (examples in [4]). In criminal investigations we find that the locations of crimes by a serial offender follow a simple walk-to-crime pattern from the offender’s home ([2], [3]). In the first case, the geographic context is the local topography of the area around the source and, possibly, weather-related influences. In the second case, the context is typically the availability of transportation or land-use patterns.

Using this insight, we propose to reason “backwards” from the events to the likely origin of the respective process, taking into account the geographic characteristics of the domain and their impact on the event-generating process. Furthermore, we seek to identify clusters of events from the same process with a likely common origin in a larger event data set. Finally, we attempt to esti-

mate the spatial probability (“risk”) of future events generated by the process by reasoning “forward” from the estimated origin, again taking into account the geographic context. These three tasks, profiling locations for likelihood of origin, clustering past events according to common origin, and finally predicting the spatial likelihood of future events from the same process, require extensive probabilistic reasoning and the seamless integration of disparate geographic context and the process’ probabilistic response to this context.

Geographic profiling is predominantly applied in criminological investigations. Current, off-the-shelf tools, such as CrimeStat [3] or Rigel [2] apply sophisticated statistical methods to the locations of past crime events to estimate their spatial origin probability distribution. But, because of their equation-based statistical approach, these tools cannot integrate the geographic context (often referred to as the backcloth and required to be close to uniform) that affects the process that generated the events. These tools also require a manual extraction (link analysis) of related events before they can apply their methods. We propose to apply a swarming agent system to this challenge as it allows us to model competing contextual process drivers and filter automatically large-scale and dynamically changing event data sets.

This paper reports on the first six-month phase of our Office of Naval Research (ONR) funded GP3 project that seeks to expand our existing *swarming prediction* framework for the future risk of emplacement of Improvised Explosive Devices (IEDs) with *reasoning about the origin* (e.g., weapons cache, safe houses, emplacer home) of events observed in the past. Our research hypothesis for this project is that modeling the characteristics of the underlying *emplacement process* from a surmised origin to likely event locations will yield better IED risk predictions than more simplistic models that project past *emplacement patterns* into the future.

The remainder of this paper is structured as follows. In the background section 2, we provide a brief overview of past efforts in geographic profiling and introduce our existing IED-risk prediction platform. In section 3, we discuss the overall system architecture and the population-level information flows in the GP3 system. Section 4 then reviews individual agent behavior in detail. Section 5 reports on metrics, preliminary evaluation experiments, and our experimental plan. Section 6 presents future re-

search opportunities for the GP3 project and beyond. We conclude in section 7.

2 Background

GP3 seeks to address a combined problem (prediction based on profiling) that, thus far, has only been solved separately in its two main components. We believe that by combining profiling from past events back to their origin and prediction from origin to future events, we can achieve better prediction accuracy. In this section, we first review the current predominant approach to profiling. Then we discuss briefly our swarming approach to prediction of IED risk, which we will extend in the subsequent sections.

2.1 Profiling in Crime Investigations

Geographic profiling is an investigative support technique for serial violent crime investigations. The process analyzes the locations connected to a series of crimes to determine the most probable area of offender home. It should be regarded as an information management system designed to help focus an investigation, prioritize tips and suspects, and suggest new strategies to complement traditional methods.

Geographic profiling uses analysis of crime sites to determine the likely area of offender residence. The methodology is based on a model that describes offenders' hunting behavior. Geographic profiling tools, such as Rigel [2] or CrimeStat [3], produce "jeopardy surfaces", which are presented to the expert user as 3-D probability surfaces that indicate the most likely area of offender home.

Current approaches to geographic profiling are based on a closed-form equation that combines for any given location its distance to those crime events that the investigator considers related to the same offender. In Rigel, for instance, it takes the following form:

$$P_{ij} = k \sum_{c=1}^T \left[\frac{\phi}{(|x_i - x_{cl}| + |y_i - y_{cl}|)^f} + \frac{(1-\phi)(B^g - f)}{(2B - |x_i - x_{cl}| - |y_i - y_{cl}|)^g} \right]$$

For a single event, the origin probability is a radial-symmetric distribution as a function of distance to the event. This distribution peaks at a distance from the event, which is derived from the median "walk-to-crime" distance of crimes of that kind observed in the past. The lengths of the tails of this distribution (towards and away from the event) are also characteristic of the crime type.

2.2 Swarming IED Risk Prediction

The choice of emplacement location for an Improvised Explosive Device (IED) by insurgents is driven by many factors. In [1], we present a swarming agent architecture that uses fine-grained agents to emulate the emplacement decision by acting out the presumed responses to these drivers in the sub-symbolic behavioral model of the agents. For instance, in the probabilistic walk of an agent, we guide the agent towards roads (where possible

IED targets are), towards locations of successful past IED detonations (where there may be a hole in the ground to emplace the new IED), away from heavily patrolled areas (where the emplacer might be discovered), and towards any known high-value targets (that the insurgents might want to hit). The advantage of using many agents repeating such a probabilistic movement model is that we can expand the list of behavioral drivers at any time and adapt the level of response of the agent to each driver automatically, using artificial evolution.

As the agents execute their movement model concurrently, they mark up the geographic map in each cycle at their respective location. Thus, areas that have a high probability of being visited under the current scenario and motivational model will receive a higher mark-up. It is important to point out that the *agents do not emulate an actual traversal of the geographic domain*, but perform a "random" walk that is probabilistically shaped by the geographic backcloth to assess the relative utility of visited locations for IED emplacement. To identify areas that are under threat for future IED attacks, we apply a proportional threshold on the emerging mark-up.

This approach to IED risk prediction is open to changing data (new events occurring, suppression or target pattern changing) and changing insurgent tactics. But it does not take into account the origin of insurgents' IED attacks (e.g., weapons cache, safe house, emplacer home), which may dramatically affect the distribution of risk areas. As the GP3 project seeks to address this shortfall by providing an origin estimate derived from past IED events, we will use the "origin free" prediction model as the baseline in our evaluation of performance gains.

3 System-Level Architecture

Swarming systems achieve complex application-level behavior through the *self-organization* of many simple agents with the desired functionality of the system *emerging* from repeated interactions of these agents in a shared and often dynamical environment. These processes are hard to understand and often non-intuitive due to the complex feedback loops that are established among the agents depending on the current scenario. Therefore, it is helpful to describe the principal information flows among major agent populations first, before discussing the individual agent behavior.

Our GP3 application comprises three different classes of agents that interact through a shared digital pheromone environment. These agents are the *profiler* agents, the (link) *clustering* agents, and the *prediction* agents. Each such agent population responds to pheromone fields as well as objects in their immediate local environment. Their responses are movement in the topology of the environment (geographic maps in GP3) as well as manipulation of the pheromone fields through local deposits. In the following, we discuss the structure and dynamics of

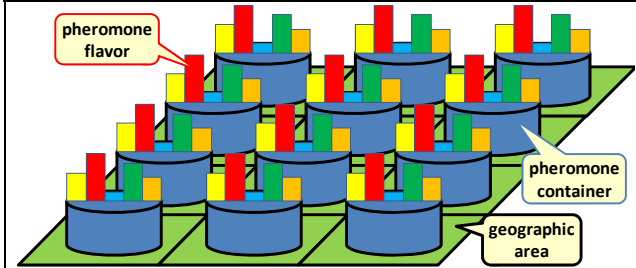


Figure 1. A pheromone container carries local concentrations of pheromone flavors for a small geographic area.

the environment first and then present the population-level behavior and information exchanges among the agents next.

3.1 Pheromone Infrastructure

The pheromone infrastructure is an architecture component that we have applied in essentially all our previous swarming agent systems in domains as varied as manufacturing control [5], adversarial prediction in urban combat [6], swarming robotics [7], or automotive vehicle design [8]. It provides the computational equivalent of chemical markers in the physical environment used by social insects to self-organize in the achievement of complex tasks.

The basic component of the pheromone infrastructure is the pheromone container. Each pheromone container holds one numerical “strength” value associated with a particular flavor, and optionally, with a dynamically assigned tag within a flavor. In the GP3 system, we have a variety of flavors to facilitate the information flows that we will discuss in the following sections.

The strength of a flavor (and tag) in the pheromone container may be read by the agents and they may increase (or decrease) the strength through deposits of arbitrary values. The deposits by the agents may be constant in each cycle or they may be modulated by their current internal state. These deposits are labeled with the particular flavor identifier and an optional tagging identifier and multiple deposits from the agents aggregate in the strength value for that flavor and tag.

Just as chemical pheromones evaporate over time, digital pheromone concentrations at any given location (x,y) are reduced in strength by the pheromone infrastructure by the repeated multiplication with a flavor-specific evaporation factor E_{Flavor} between zero and one ($strength_{Flavor}(x,y,t+1) = strength_{Flavor}(x,y,t) * E_{Flavor}$). If the strength of a pheromone falls below a flavor-specific threshold value, it will be set to zero. In other application we also use the equivalent of Brownian-motion dispersion of chemical markers to nearby locations modeled through an exponentially decaying propagation process, but in GP3 we do not need the digital pheromones to disperse.

While the pheromone infrastructure may be applied to any number of topologies in which agents coordinate

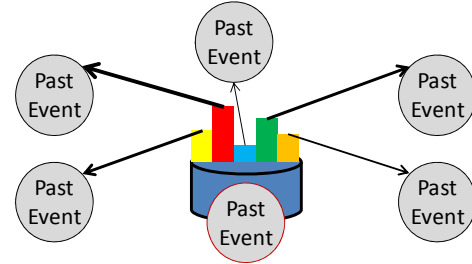


Figure 2. Other pheromone containers (associated with individual past events) carry flavors tagged with other past events.

their activities, in GP3, we primarily associate pheromone containers with small geographic areas in the region of interest (“playbox”, Figure 1). These areas form a non-overlapping square grid, where any coordinate within the playbox is associated with exactly one pheromone container. Thus, for any pheromone flavor and tag, GP3 supports a discretized spatial field of strength values of the flavor and tag across the playbox. In absence of pheromone propagation processes and assuming fixed-strength repeated deposits by the agents that contribute to the field, it is permissible to normalize the strength field into spatially distributed probability values.

In GP3, we also associate additional non-geographic pheromone containers with the objects that represent individual past events (Figure 2). As we will discuss in the following sections, here pheromones serve as accumulation points for knowledge about the relation among these events (link analysis).

3.2 Profiler Agent Population

The profiler agents implement a swarming variety of the traditional geographic profiling process that is used, for instance, in criminological forensics or in geology. These agents effectively reverse our model of the geographic process that led to the occurrence of a particular known past event, providing a probabilistic hypothesis about the likely origin location of the event. In criminological terms, the profiler agents reverse the walk-to-crime, where the past event is object is at the location of the crime.

Traditional geographic profiling (see section 2.1) assumes an origin probability for any given location that is simply a function of the distance of that location from the location of the event. That function peaks at a certain distance (median walk-to-crime distance), forming a ring around the event location.

Our swarming profiling process refines this process, taking into account characteristics of the environment within which the events are embedded. Events in GP3 are sites of deployed Improvised Explosive Devices (IED) and therefore the profiler agents reverse the deployment process back to the location of the weapons cache or the home of the placer. Hence, environmental characteris-

tics that our profiler agents take into account reflect spatial characteristics of the origin of the deployed IED, such as the absence of roads or the presence of buildings. Additional characteristics, such as the presence of a supportive (insurgent-friendly) population or the absence of suppressing forces could be included too if such data was available.

We represent environmental characteristics as (static or dynamic) pheromone fields of various flavors. Our current implementation includes a field that peaks at the location of roads (“road” pheromone flavor) and another that peaks in areas that are built up (“building” pheromone flavor). These pheromones are maintained by the GP3 “world” model.

The movement of profiler agents is driven by local road and building pheromone concentrations and by their position relative to their respective past event. With each model cycle, as they move across the geographic topology of the pheromone infrastructure, profiler agents deposit a constant amount of “profiler” pheromone flavor at their current location. That deposit is further tagged with the identifier of the respective past event whose origin a particular profiler is trying to estimate. Thus, as shown in Figure 3, profiler agents consume road and building pheromones and produce tagged profiler pheromones that represent the spatial origin probability distribution for each past event.

In addition to reading the local concentrations of road and building pheromones to determine their movement in geographic space, profiler agents also read profiler pheromones. These concentrations do not affect the agents directly, but serve to build up the current link-probability estimate for their respective past event as they are deposited with the event identifier tag intact in the pheromone container of the profilers’ past event object. Thus, if a profiler from event A encounters higher concentrations of profiler pheromones tagged with event B than C, B-tagged pheromones in the profiler’s past event will be stronger than C-tagged ones. Effectively, the profilers estimate the pair-wise probability that their past event shares a common origin with any of the other past events. Figure 3 also shows this information flow from the geographic space to the collection of past events.

3.3 Clustering Agent Population

The profiler agent population constructs probabilistic spatial origin estimates for the current set of past events. In addition, they use the overlap of these probability fields to estimate the probability that any two given events share a common origin. This so called link-probability is computed by normalizing the tagged profiler pheromones in the containers of the past events for a given event across all tags.

From this emergent and dynamically adjusted link analysis, we derive discrete clusters of linked events us-

ing the cluster agent population. Cluster agents are also embedded in geographic space, but their current absolute location only has a meaningful relative interpretation in respect to each other. Cluster agents execute a self-organizing, iterative, and emergent clustering process. There is one cluster agent for each past event and their movement in space is driven by artificial forces that attracts them to the location of their event (increasing with distance) and to other cluster agents (decreasing with distance). Their attraction to other cluster agents in their current neighborhood is modulated by the current link-probability estimate between them and the other agents. Therefore, cluster agents whose events occurred close to each other or who share strongly overlapping origin probability fields tend to be close to each other.

We define an arbitrary and configurable distance threshold below which two cluster agents are considered part of the same cluster of linked events. Using this threshold, we turn the real-valued distance relationships among cluster agents into a binary (linked yes or no) statement that allows us to enumerate all non-overlapping clusters of linked events (transitive hull of binary link relations) in the current arrangement of clustering agents. Figure 4 shows the information flow from profiler pheromones on past event objects to enumerated cluster objects (model artifacts) through the clustering agents.

Any cluster that emerges from the interaction of the clustering agents has a population of at least one past event as members. Somewhat arbitrarily, we define the spatial location of a cluster as the center of gravity of the location of all its member events. To reinforce the clusters which are primarily based on the overlapping origin probability fields of its members, we use the center of gravity of the cluster to which an

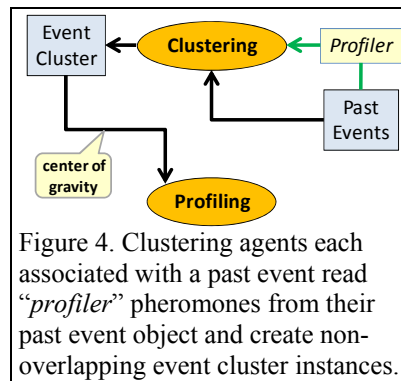


Figure 4. Clustering agents each associated with a past event read “profiler” pheromones from their past event object and create non-overlapping event cluster instances.

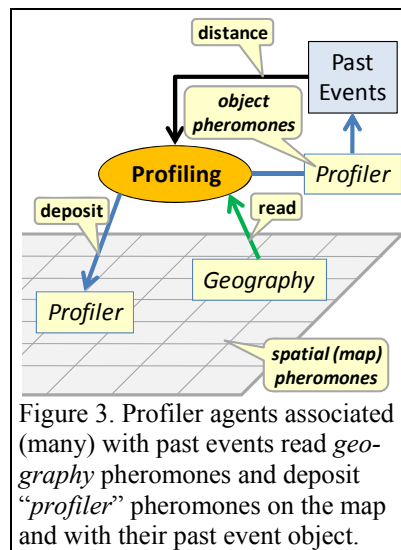


Figure 3. Profiler agents associated (many) with past events read *geography* pheromones and deposit “profiler” pheromones on the map and with their past event object.

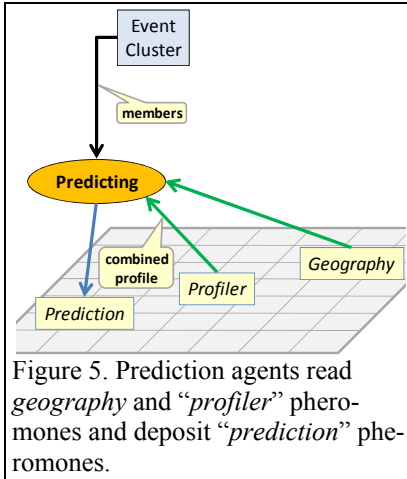


Figure 5. Prediction agents read *geography* and “*profiler*” pheromones and deposit “*prediction*” pheromones.

the edge of the area populated by the members of the cluster. Figure 4 illustrates this self-organizing adjustment in the individual origin probability fields through the cluster-recruitment feedback loop between the profiler agents and the clustering agents.

3.4 Prediction Agent Population

The interaction between the profiler agents and the clustering agents for a given set of past events yields an estimate for the spatial origin probability of each event and a set of non-overlapping clusters of linked events. We use these products to inform the population of prediction agents as they identify areas that have a high likelihood of the occurrence of future events. This input assumes that the same process that resulted in the past events at least in part contributes to the future events.

Each cluster of linked events constitutes our current hypothesis of the existence of a single geographically embedded process that produced these events from a common origin. In geological applications, this could be an open mineral deposit (origin) where erosion led to the dispersion of samples that were found (events) by geologists. In criminological forensics, the events would be crimes perpetrated by the same suspect and the origin is the suspect’s home. In GP3, a cluster of linked past IED events is assumed to be delivered from the same origin, which could be a weapons cache, safe house, or emplacer’s home.

In the baseline IED prediction process (see section 2.2) uses geospatial motivational drivers such as the presence of roads or past IED events to estimate the risk of future events for a particular area. These forecasts do not take into account the origin of past events nor the walk-to-crime process that constrains the emplacement relative to the origin. In GP3, we seek to demonstrate that using this additional knowledge will improve our prediction accuracy compared to this baseline.

Profiler agents reason “backwards” from past events to event-specific origin distributions. In contrast, predic-

tion agents reason “forward” from assumed origins to locations of high risk for future events. This forward reasoning (movement in and markup of space), includes the motivational drivers from the baseline prediction process conveyed through driver-specific pheromone fields (e.g., attraction to roads, attraction to past event locations, repulsion from suppressing forces, attraction to high-value targets). In addition, in GP3 it is informed by the combined origin probability of the linked-event clusters, which probabilistically determines the initial location of the prediction agents.

Prediction agents sample the origin probability estimate for linked event clusters enumerated from the arrangement of clustering agents. This cluster-origin probability field is the aggregate of the origin probability distributions of all past events that make up a cluster. The sample determines the initial locations for the prediction agents from where they start their walk-to-crime.

As the prediction agents walk away from their initial location, they follow their motivational drivers listed above. They also deposit a fixed amount of “prediction” pheromones to mark up the geographic probability field of the probability of future events.

We identify future event risk areas through an arbitrary and configurable threshold on the normalized prediction pheromones (currently at 15%). Figure 5 shows that the profiler agent population consumes merged profiler pheromones of a particular linked events cluster and produces prediction pheromones which identify risk areas. In a later phase of the project (see section 6), we plan to feed back the recall accuracy, that is the prediction agents’ ability to capture past events within the risk areas, to the clustering and profiling agents. But at this point, the prediction agents only consume information from the profiler and clustering agents without pushing any knowledge back to them.

3.5 Summary

Figure 6 shows the combination of the information flows in the GP3 profiling, link analysis, and prediction model. **(Profiling)** Starting with the current set of past events and integrating geographic characteristics, profiler agents reverse the walk-to-crime that led to a specific event and produce an origin probability distribution. **(Link Analysis)** In addition, they estimate for each past event the probability that it has a common origin with any of the other events. **(Link Cluster Identification)** The clustering agents process the implicit graph of link probabilities among past events and produce a set of non-overlapping event clusters that also take into account the distance among the event locations. These clusters further refine the profiling process. **(Prediction)** Finally, the prediction agents use the combined origin probability distribution of individual clusters and geographic characteristics that may affect their walk-to-crime to estimate the

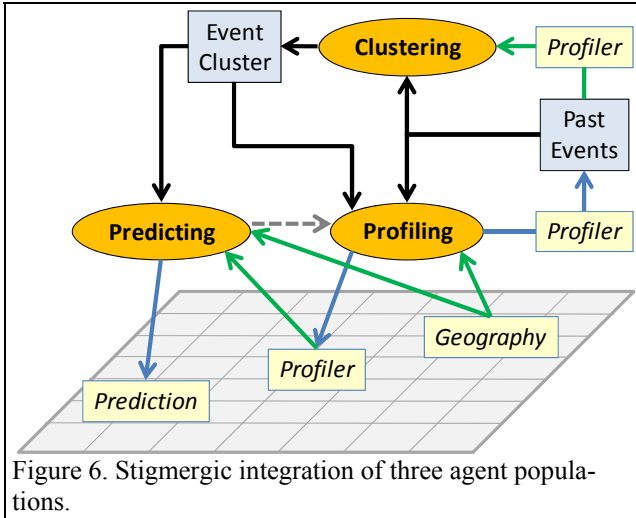


Figure 6. Stigmergic integration of three agent populations.

level of risk of future events for locations within the geographic area. Future research (dashed line in Figure 6) will provide feedback of the prediction agents' ability to recall the past events in the risk areas that they identify to the profiling and link clustering processes.

4 Detailed Agent Behavior

The previous section discussed the operation of the GP3 agents at the population level and identified information flows among them. In the following, we highlight relevant aspects of the simple individual behavior of these agents that lead to the emergence of the complex system functions of profiling, link clustering, and risk prediction.

The three agent types share a common behavioral framework. Our model maintains a complete collection of all agents in the system and activates them all (in randomized sequence) in each model cycle. After the agents complete their simple decision processes, the model then activates the pheromone infrastructure to execute evaporation on all pheromone containers, collects any data for experimental metrics, and finally updates the graphical user interface if it is turned on.

When an agent is activated, it considers its local environment (e.g., samples pheromone concentrations, identifies nearby objects), and computes and executes a movement vector. The agent may also deposit pheromones if it finds itself within a geographic area associated with a pheromone container or if it contributes to pheromone fields maintained for specific object types (e.g., past events and their links to other events). The agents' movement vector calculation is a length-limited weighted vector sum of component movement vectors, each associated with a behavioral response to either a particular pheromone flavor or to the presence of objects in the agents' environment. Our discussion of specific agent behavior focuses on these sub-symbolic calculations.

4.1 Event Profiler Agents

The movement vector of a profiler agent associated with a particular past event is the weighted vector sum of five behavioral components: "buffer" step, "road" step, "building" step, "cluster" step, and "random" step. The weight associated with each component vector is configured globally and reflect the impact each behavioral component has on the overall profiler movement. Figure 7 illustrates the various component steps (building and random step omitted for clarity) that comprise the profiler movement model.

The combination of the buffer and random steps implements the traditional geographic profiling process. A global buffer radius parameter defines the median distance between origin and event location characteristic for the particular type of event. In the buffer step, the profiler measures its current distance to its event's location and then computes a movement vector either towards or away from its event depending on whether the distance is smaller (away from) or larger (towards) the buffer radius parameter. The length of the buffer step vector increases exponentially with the distance of the profiler from the desired buffer-radius distance to the event. In other words, the buffer step is an attractive force towards the ring around the event as applied in energy-minimizing spring models for graph embedding. The combination with a random noise step component ensures that the profilers visit locations outside of the actual buffer radius but with decreasing likelihood as the distance to the ring increases. It also ensures that the entire ring around the event is visited with equal likelihood.

The road and building steps modify the distribution of profiler agents on the ring around their event to take into account geographic characteristics. In GP3, the events are emplaced IEDs and thus we expect their origins (i.e., weapons caches, safe houses, emplacer homes) to be away from roads and inside built-up areas. Therefore, the road step seeks to avoid areas with high road pheromone concentrations while the building step is attracted to high building pheromone concentrations.

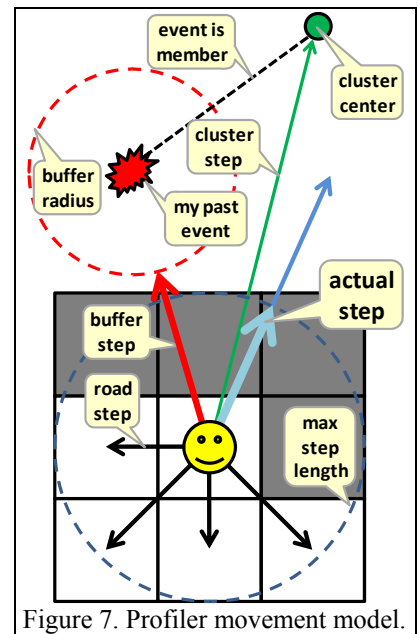


Figure 7. Profiler movement model.

While the buf-

fer step is calculated using a spring model based on actual geographic locations independent of the square gridding of the topology of the pheromone infrastructure, the road and building steps base their component vectors on the arrangement of the pheromone containers. In both steps are calculated as weighted vector sums to the geographic center of each pheromone container in its immediate neighborhood. In the case of the road step, the weights are inverse proportional to the concentration of road pheromone in these containers (repulsion), while, they are proportional to the building pheromones in the building step. In other words, the profiler “climbs” down the road pheromone gradient and up the building pheromone gradient.

Finally, the cluster step attracts the profiler to the center of gravity of the cluster to which its event currently belongs. As with the buffer step, this movement component vector is a spring force that increases in length with the distance of the profiler to that center of gravity.

The *direction* of the movement vector of the profiler agent in each model cycle is the weighted vector sum of its component steps. It is proper in the design of self-organizing algorithms to avoid large-scale changes in the step-by-step evolution of the agent system. Therefore we limit the *length* of the step that a profiler is permitted to take by a global parameter that is small (1%) relative to the overall size of the geographic area associated with the model. If the weighted component vector sum is larger than the maximum step length, then we scale the movement vector down to that maximum length. Shorter steps are permitted of course, effectively reducing the speed with which the agent moves in space.

After completing its movement by applying the length-limited weighted component vector sum, the profiler deposits a unit amount of profiler pheromone tagged with the identifier of its associated event into the pheromone container

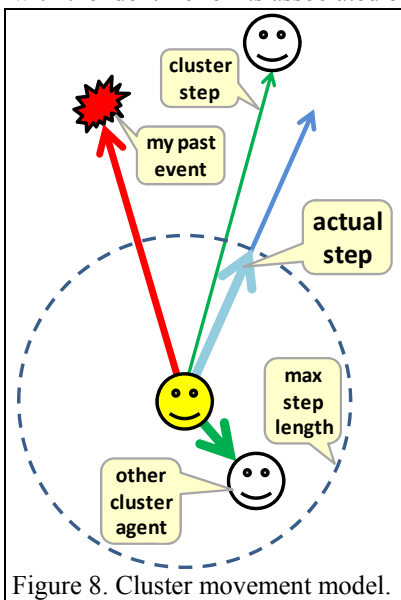


Figure 8. Cluster movement model.

that is associated with its new geographic location. From this container it also reads all profiler pheromone concentrations for tags other than its own and deposits the same amounts into the pheromone container of its past event object. The fixed deposits on the geographic map by a fixed population of pro-

filer agents allows us to normalize the profiler pheromones across the geographic map into spatial origin probability fields and across past event objects into link probabilities.

4.2 Link Clustering Agents

The movement vector of the clustering agent associated with a particular past event is the weighted vector sum of three behavioral components: “event” step, “cluster” step, and “random” step (Figure 8). The clustering agents do not use pheromone fields for their coordination. Therefore, both the event and the cluster step are attractive artificial forces directed towards specific locations.

The cluster agents seek to find an energy-minimizing balance among their dynamically changing event step and cluster step forces. This balance integrates the two optimization criteria of the emergent clustering mechanism: group events that a) have a strong overlap in their origin probability fields, and that b) occurred near to each other.

The event step computes a spring force from a clustering agent’s current location to the location of its event. The length of this component vector increases with increasing distance between agent and event. This behavioral component supports the second optimization objective.

The cluster step supports the first objective. It is the weighted sum of vectors towards other clustering agents where the weight is inversely proportional to the distance from the agent’s current location to the current location of the other agent. This attraction model acts not like a spring but like gravity or electro-magnetic forces, but instead of a quadratic force decline with distance, we chose an exponential decline. The shaping parameter of this decline determines the effective radius around the clustering agent beyond which forces to other clustering agents are negligible. This limitation of the clustering agents’ reach is in line with the proper design of self-organizing algorithms where an agent’s reasoning and action should be local in some topology.

Finally, the random step is a unit-length vector with a randomly selected heading. Adding random noise to the clustering process allows the system to escape from local force-balancing minima. At this point, we keep the random contribution constant rather than reducing the “temperature” over time as in simulated annealing.

The clustering agent maintains an internal table of its current distance to other agents (computed for the cluster step). Based on this table, the agent applies a global cut-off parameter to enumerate all other agents that are currently within the given radius. The GP3 model infrastructure recursively traverses these binary link relations among clustering agents to enumerate all unique non-overlapping clusters in a given model cycle. We took this centralized implementation shortcut, because the number of past events in our application is relatively small. To

scale up, we could apply off-the-shelf self-organizing hierarchy-formation mechanisms that would decentralize this processing step.

4.3 Event Risk Prediction Agents

The initial location of the walk-to-crime of a prediction agent is sampled from the combined origin probability distribution of a linked events cluster. The agent determines this location in a two-step process. First, it probabilistically selects a cluster from the current enumeration, where the selection probability is proportional to the number of past events that are included in a cluster. Thus, larger clusters attract more prediction agents that emulate their geographic process. The second step in the initialization requires the probabilistic selection of a geographic location based on the selected cluster. Here the prediction agent considers the centers of all areas associated with pheromone containers. The probability of selecting any one of these locations is proportional to the combined origin probability assigned to these locations by the individual events in the cluster (renormalized sum of all cluster members' origin probability distributions). The profiler agent will repeat this initialization process at regular intervals (currently configured to 200 model cycles) to provide a dense sampling of the origin distributions and to adapt to changing origin and clustering hypotheses.

Just as in the baseline IED prediction process, the movement vector of the prediction agent is the weighted vector sum of three behavioral components: "event" step, "road" step, and "random" step (Figure 9). The event step component vector is the weighted sum of vectors to all past events where the length of these vectors is inversely (exponential) proportional to the distance to the event (see cluster step in clustering agent section 4.2). The road step is the weighted sum of vectors to nearby pheromone

containers where the length of these vectors is proportional to the strength of the road pheromone in the containers. Thus, while a profiler agent climbs down the road-pheromone gradient in its reversed walk-to-crime, the prediction agent climbs up that gradient. The unit-length random noise step ensures that the prediction agent is

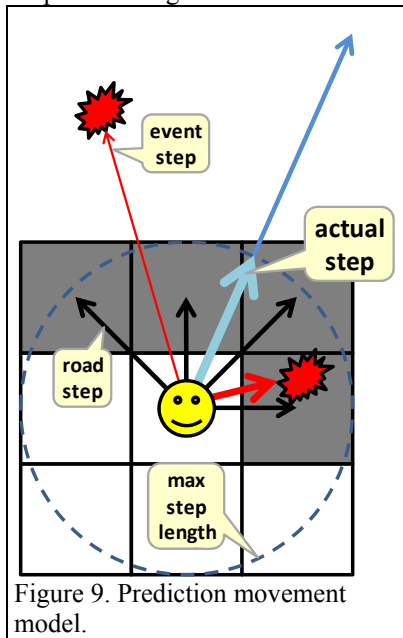


Figure 9. Prediction movement model.

not trapped in local minima but converges on areas that truly stand out as attractive future event candidates.

The prediction agent deposits a fixed amount of prediction pheromone at its current map location in each model cycle. Thus, higher prediction pheromone concentrations directly indicate more frequent visitations of these locations by prediction agents.

The model infrastructure designates risk areas for future events based on the current pattern of prediction pheromones in the square array across the geographic area of interest. An area associated with a single pheromone container is designated at risk for a future event if its prediction pheromone concentration is above a certain "risk-area threshold" (currently configured to 15%), relative to the globally highest prediction pheromone concentration.

5 Evaluation Experiments

In our project, we recently completed the first implementation of the model as it is reported in this paper. Therefore we are still early in our experimentation cycle and report only preliminary results. Furthermore, due to the sensitive application domain (IED risk prediction), we are precluded from publishing the data sets that we are experimenting with. But we can discuss aggregate performance measures.

In the following we first present metrics that measure the performance of our system. Then we discuss preliminary experimental results. Finally, we lay out our experimental plan.

5.1 Performance Metrics

Our agent populations produce a number of artifacts that warrant closer inspection. Primarily, there are three products: individual and combined origin probability distributions for past events, enumerated clusters of past events that hypothesize common origin for these events, and prediction distribution and risk-area designations for future events.

In GP3, our performance evaluation has to focus on the third product – the prediction of risk for future events. This limitation is due to the fact that in the IED domain we have some event data with their geographic and temporal context available, but ground-truth data on the insurgents that employed these IEDs and in particular the locations of their weapons caches, safe houses, or emplacements' homes is not available (in an unclassified context). Thus, our performance metrics need to focus on the accuracy with which our risk assessment recalls past events and predicts future events ("Recall Accuracy" and "Prediction Accuracy"). We achieve a high accuracy for a given set of past or future events if we designate only a small number of areas as risk areas but capture a large portion of the events in those risk areas.

For any given risk-area threshold between 0% and 100%, a given prediction pheromone field, and a set of

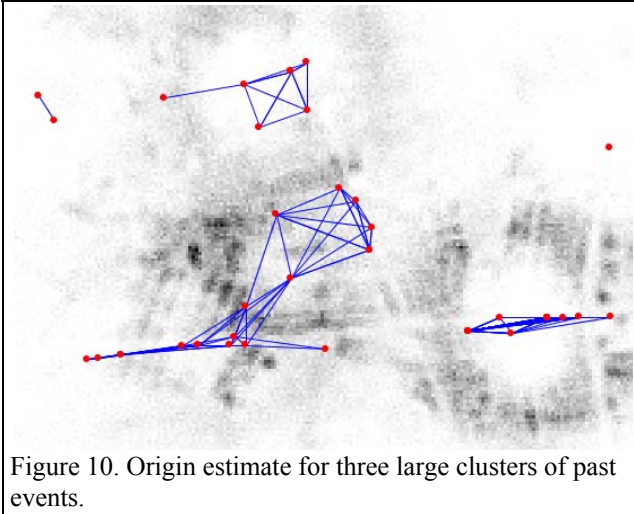


Figure 10. Origin estimate for three large clusters of past events.

past and future events, we compute the following three measures:

“**Effort**” – the portion of pheromone container areas with prediction pheromones above the relative risk area threshold (and for which a commander in the field may expend effort to protect it)

“**Recall**” – the portion of past events located within risk areas

“**Forecast**” – the portion of future events located within risk areas

The choice of a risk-area threshold for performance experiments is very arbitrary. Instead, we are interested in the general performance characteristics of our model independent of a specific threshold value. Thus, as in signal detection theory, we analyze the operating characteristics of our model through a “gains charts” that plot Recall or Forecast as a function of Effort. The diagonal of such a plot corresponds to the performance of a random predictor, while good performance is indicated by a steep rise of the curve above the random diagonal, approximating the 100% mark with small Effort values already.

5.2 Preliminary Results

We are currently experimenting with an unclassified event data set of approximately one month’s worth of IED events in a small geographic region (87 lat/lon and time stamped events). For our initial experiments, we arbitrarily split this data set by time stamp into 41 “past” and 46 “future” events.

During the implementation of our model, we manually tuned the various behavior weights, thresholds, and other model parameters to produce (visually) “reasonable” outcomes for the profiling, clustering, and prediction. As a result, the model now identifies three significant clusters with 13, 10, and 8 member events respectively (Figure 10). These clusters are associated (roughly) with three major roads that cross the area (not shown).

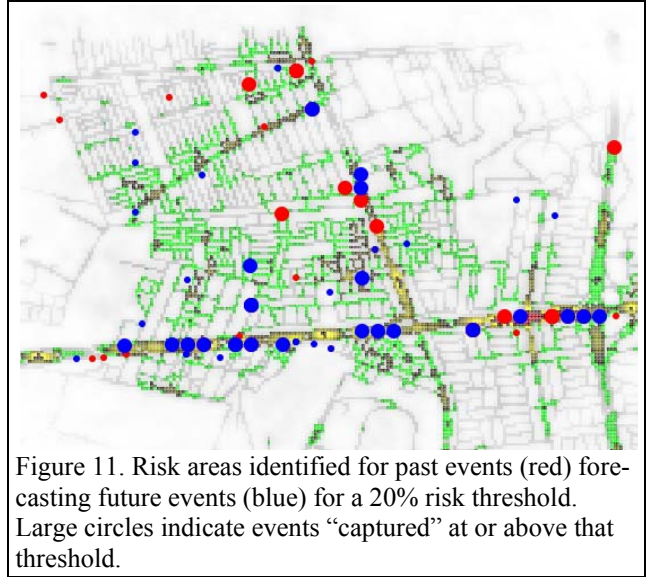


Figure 11. Risk areas identified for past events (red) forecasting future events (blue) for a 20% risk threshold. Large circles indicate events “captured” at or above that threshold.

An analysis of the combined origin probability distribution for each cluster reveals a distinct set of peaks at a safe distance from the main road inside residential areas, with fine-scale modulations of the residential road grid (small probability dips) and built-up blocks (small probability peaks). We conclude that the main peak of each cluster results from the buffer and cluster step components of the profiler movement model, while the fine-scale modulations stem from the road and building steps.

Figure 11 shows the pattern of risk areas emerging at a threshold of 20%. We clearly see the risk for further attacks focused on the main roads in the area as well as in the residential areas that the profiler agents designated as the likely origin of the past events.

Figure 12 plots the gains chart for the baseline prediction model (random initial locations for the prediction agents) versus the model configuration where we use the clusters’ origin probability distribution to initialize the walk-to-crime for the prediction agents. The figure shows that the performance of the baseline configuration (dashed line in Figure 12) in our scenario is already very good.

The scenario is dominated by a large number of future events that occur close to locations of past events and are thus captured by the good recall capability of the baseline model (initial sharp rise to 50% forecast accuracy in the plot). Our GP3 model that takes the estimated origin of past events into account also replicates the past event pattern but then rises above the performance of the baseline configuration as it quickly captures events that are further away from past events, reaching 100% forecast accuracy with only 50% effort. A more detailed analysis that specifically looks at the forecasting accuracy for outlier events will be included in a future publication.

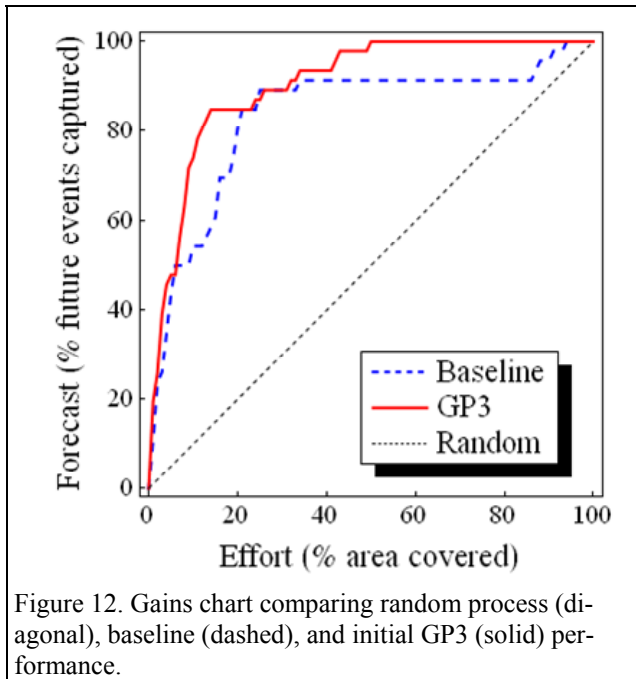


Figure 12. Gains chart comparing random process (diagonal), baseline (dashed), and initial GP3 (solid) performance.

5.3 Experimental Plan

As with many self-organizing systems with emergent properties, systematic experimentation is the only avenue to gain relevant insights into the model’s performance and to tune its parameters. This is due to the fact that the complex non-linear dynamics of these systems preclude us from a formal representation and analysis. Therefore, our GP3 project now enters the phase of systematic exploration of our model’s parameter space through large-scale simulation exercises.

To support systematic experimentations and data gathering, we integrated the model in a parameter sweep infrastructure that allows us to specify configuration parameter ranges for which individual experiments should be executed and selected metrics should be applied. Due to the probabilistic nature of our model, we also have to execute more than one experiment per configuration with varying random seeds. To manage this large volume of experiments, we use the Map/Reduce implementation by GridGain.com, which allows us to spawn off experiments on a dynamic grid of computers within our network.

We plan to execute two kinds of sweep experiments. First, we explore the impact of various model parameters, such as the behavioral weights of the agents’ movement models, on our performance metrics. As we vary these parameters, we will see different origin probability distributions and different cluster patterns that result in varying operating characteristics (gains charts) of the model. These experiments will help us select optimal parameter settings and identify candidate information flows for on-line model tuning based on the event recall accuracy.

In a second round of experiments, we plan to vary the data that we provide the model. Here we intend to change the ratio of past and future events from our static data set (currently 41/46), and the availability of road or building data. These experiments begin to test the robustness of the model in changing problem settings.

6 Future Research

The next phases of our GP3 project have us investigate the potential for further performance improvements of the IED risk prediction model along several avenues. First, we will use insights into the model dynamics gained from the systematic parameter sweep experiments to implement additional feedback loops among the agent populations to **adapt model parameters** according to the current scenario. In particular, we intend to close the loop between the prediction agents and the profiler and clustering agents as indicated in Figure 6.

Next, we intend to integrate **temporal reasoning** into our model, where the clustering includes also temporal proximity of events rather than just spatial. The temporal “center of gravity” (age) of a cluster affects the profiler agents’ behavior as well as the choice of the prediction agents as they determine their next initial location.

Finally, we may expand the range of **geographic characteristics** that are taken into account by the profiling and prediction processes (currently roads and buildings) and any **event characteristics** (e.g., type of explosive, nature of target) that may further refine our emerging hypotheses of linked event clusters. These refinements depend heavily on the availability of scenario data though.

7 Conclusion

In various domains there are processes embedded in a geographic context whose patterns of events are indicative of their likely origin but strongly influenced by the respective local geographic context. For example, the pattern of discovery of mineral samples in field geology is shaped by the location of the source deposit and the local characteristics of the erosion process. In criminal investigations we find that the locations of crimes by a serial perpetrator follow a simple walk-to-crime pattern from the suspect’s home. Using this insight, we are not only able to reason “backwards” from the events to the likely origin of the respective process, but also to identify clusters of events with a likely common origin in a larger event data set and to estimate the spatial probability (“risk”) of future events generated by the same process.

In this paper we presented a swarming agent model that implements the three major functions of profiling (estimating event origin), clustering (identifying events with likely common origin), and prediction (assessing spatial probability of future events) in an emergent fashion using self-organizing mechanisms. Given that be-

havior of swarming systems is hard to understand from the specification of individual agent behavior, we reviewed first the information flows among the three agent populations (profiler, clustering, and prediction agents). Our subsequent discussion of the simple agent logic for each type revealed a generic movement model that combines component movement vectors based on distinct motivational drivers. These vectors are computed relative to other entities in the model or gradients of specific digital pheromone fields. Finally, our presentation of preliminary experimental results shows promise for improvement of an already strong prediction approach through the integration with the backwards reasoning from past events to common origins. We lay out an ambitious plan for systematic, large-scale evaluation experiments and an agenda for future research that offers further performance gains.

Acknowledgements. This research was conducted with the support of the office of Naval Research (Contract # N00014-08-C-0588). The results presented do not necessarily reflect the opinion of the sponsor.

8 References

- [1] H. V. D. Parunak, J. Sauter, and J. Crossman. Multi-Layer Simulation for Analyzing IED Threats. In *Proceedings of IEEE International Conference on Technologies for Homeland Security (HST 2009)*, 2009.
- [2] Kim Rossmo. An Evaluation of NIJ's Evaluation Methodology for Geographic Profiling Software. In response to National Institute of Justice's *A Methodology*

for Evaluating Geographic Profiling Software: Final Report. March 2005.

[3] Ned Levine. *CrimeStat: A Spatial Statistics Program for the Analysis of Crime Incident Locations* (v 3.1). Ned Levine & Associates, Houston, TX, and the National Institute of Justice, Washington, DC. March 2007.

[4] Yukon Exploration and Geology 2006. D.S. Emond, L.L. Lewis and L.H. Weston (eds.), 2007. Yukon Geological Survey, 2006.

[5] S. Brueckner. Return from the Ant: Synthetic Ecosystems for Manufacturing Control. Dr.rer.nat. Thesis at Humboldt University Berlin, Department of Computer Science, 2000.

[6] H. V. D. Parunak. Real-Time Agent Characterization and Prediction. In *Proceedings of International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS'07)*, Industrial Track, Honolulu, Hawaii, pages 1421-1428, ACM, 2007.

[7] J. A. Sauter, R. Matthews, H. V. D. Parunak, and S. A. Brueckner. Demonstration of Digital Pheromone Swarming Control of Multiple Unmanned Air Vehicles. In *Proceedings of AAAI Infotech@Aerospace*, Arlington, VA, AIAA, 2005.

[8] S. Brueckner, R. Gerth. Applying Distributed Adaptive Optimization to Digital Car Body Development. In *Proceedings of Workshop on Engineering Self-Organising Systems (ESOA 2004)*. Springer LNCS 3464, pages 267-279, 2005.