Strategies for crowdsourcing for disaster situation information

E. T.-H. Chu[1], Y.-L. Chen[1], J. W. S. Liu[2] & J. K. Zao[3]
[1]*Department of Computer Science and Information Engineering, National Yunlin University of Science and Technology, Taiwan, R.O.C.*
[2]*Institute of Information Science, Academia Sinica, Taiwan, R.O.C.*
[3]*Department of Electronic and Computer Science, National Chiao Tung University, Taiwan, R.O.C.*

Abstract

When existing surveillance sensors used by a disaster warning and response system cannot provide adequate data for situation assessment purposes, crowdsourcing information collection can be an effective solution: People armed with wireless devices and social network services can be used as mobile human sensors. Their eye-witness reports can complement data from in-situ physical sensors and provide the system with more extensive and detailed sensor coverage. The crowdsourcing strategy used by the system can be random, relying solely on mobility of individuals for coverage of the threatened area; or crowd-driven, with the system providing situation updates as feedback to aid the crowd; or system-driven with individuals moving in response to directives from the system. The relative merits of the strategies clearly depend on the disaster scenario and the characteristics of the crowd.

This paper presents a general crowd model for characterizing individuals within a crowd and the crowd as a whole and an abstract mobility model of crowd movements in the threatened area. The models can be specialized to characterize different disaster scenarios and crowds, and used in the simulation of the crowdsourcing strategies for evaluation purposes. Data on relative performance of different strategies for two types of disasters were thus obtained.

Keywords: crowdsourcing, disaster management, crowd mobility model.

WIT Transactions on the Built Environment, Vol 119, © 2011 WIT Press
www.witpress.com, ISSN 1743-3509 (on-line)
doi:10.2495/DMAN110231

1 Introduction

A disaster surveillance and response system must estimate boundaries of threatened area(s), assess the threat potential and acquire situation awareness to support decisions on what alerts and warnings to issue when a disaster seems imminent and how to handle emergencies and calamities during and after the disaster. In-situ sensors and sensor networks and remote sensor systems used to collect data for this purpose may not always provide the system with sufficiently complete and detailed view of the threatened area. The area coverage of a sensor network (or networks) and density of sensors deployed are often limited by costs. This is the primary reason that Gulf of Mexico coast and Southern California region were not adequately monitored by surveillance cameras and other sensors during the 2010 BP oil spill and 2009 California wildfire disasters [1, 2]. Other reasons for inadequate surveillance sensor coverage include that some in-situ sensors may be damaged just when they are needed and thick clouds, vegetations, buildings, etc. can render remote sensors (e.g., surveillance satellites and unmanned aerial vehicles) ineffective. The resultant blind regions in sensor coverage can leave responders ill informed of imminent dangers to hundreds of people. This was what happened during Typhoon Morakat in 2009 in Taiwan [3].

A way to get fuller and more detailed coverage than what physical sensors can provide is crowdsourcing data collection. People using wireless devices and Web 2.0 services are in essence mobile human sensors. Their eye-witness reports of conditions at different locations can complement data from physical sensors to eliminate blind spots and mend fragmentation in sensor coverage.

This paper focuses on alternative strategies used by a disaster surveillance system to manage *crowdsourcing data collection* (*CDC*) *processes*. To keep our discussion concrete without loss of generality, we assume hereafter that the system triggers a CDC process by broadcasting a data collection request to a crowd. The process ends when the system has collected enough data to construct a sufficiently complete view of the threatened area. Possible strategies used by the system can be divided roughly into three types: random, crowd-driven and system-driven. One can say that a *random strategy* is a minimal strategy. After broadcasting a CDC request, the system does nothing other than collecting and processing reports from the crowd, relying solely on mobility of individuals for coverage of the threatened area. According to the *crowd-driven strategy*, the system updates the observed current conditions of the threatened area based on reports it has collected and processed and provides the information as feedback to the crowd. Otherwise, it lets the crowd guide themselves in their exploration efforts after broadcasting a request. According to a *system-driven strategy*, the system issues directive(s) to all individuals or a selected subset of the crowd who has responded to its initial CDC request. Each directive to targeted individuals guides them in their exploration. The directive is also a new request, leading to new responses from the crowd. The communication between the system and the crowd repeats until the system has a complete view of the threatened area and the CDC process ends.

We can measure the relative performance of the strategies along multiple dimensions, including the accuracy of the estimated threatened area boundary and resolution of the view of the area, response time of the CDC process (i.e., time required to obtain the estimate and view), the costs and rewards of each CDC process and so on. We will return to define and discuss the figures of merit of our choice. Regardless the figures of merit used, the relative performance of strategies clearly depends on the disaster scenario and the crowd characteristics.

This paper makes three contributions to studies on strategies for crowdsourcing sensor data collection. The first is a general crowd model for characterizing each individual within a crowd and the crowd as a whole. Rather than some qualitative attributes, our model characterizes each individual in the crowd quantitatively in terms of his/her contributions to the CDC process. The quantitative nature of our model resembles the concept of crowd quality for quantification of the quality of crowdsourced spatial data and software testing [4, 5]. The second is an abstract and formal mobility model of crowd movements. The mobility model is also quantitative. It complements existing human mobility models such as the ones described in [6, 7] that were developed to characterize movements of people in their normal daily lives. Our models are meant to be specialized to characterize different disaster scenarios and crowds and used in simulation of crowdsourcing strategies for evaluation purpose.

The third contribution of this paper is a general methodology for evaluating strategies for crowdsourcing sensor data collection. A search of Internet for crowdsourcing strategies usually returns numerous entries on the subject, too numerous to list as citations in this paper. None of them addresses effectiveness of strategies for managing CDC processes, however.

Following this introduction, Section 2 presents definitions and underlying assumptions. Section 3 present our models of the threatened area, crowd and crowd mobility and discusses how the models can be specialized to model different crowds (e.g., official responders, NGO volunteers, unknown crowds) and their mobility for different types of disasters (e.g., oil spill, earthquake, landslide, flood and wildfire) at different locales. Section 4 presents parameters of simulation experiments for the purposes of evaluating different system-crowd interaction and crowd mobility strategies. Section 5 defines figures of merits used to measure their performance and simulation data on two types of disasters as case studies. Section 6 summarizes the paper and presents future work.

2 Definitions and assumptions

Clearly, the quality of the human sensor data collected by a CDC process and response time of the process critically depends on crowd quality [4]. According to their skills and motivation, we divide all participants of a CDC process roughly into types I, M and U.

(A) Participant Types

Participants of I-type are *ideal human sensors*. A *type-I individual* may have been trained or have practiced to be a human sensor. At each step during a CDC process, he/she moves to the right location promptly, makes a right observation and sends an accurate report. Ideal human sensors are likely to be government disaster responders (e.g., policemen, firemen, and soldiers) and some volunteered responders from NGOs (e.g., Red Cross), local communities, etc.

An *M-type participant* is highly motivated and hence, is reasonably responsive: he/she may be a registered volunteer, a person affected by the disaster, and so on. The participant is known to the system and can be rewarded in someway afterwards. However, the sensor data collected and reported by him/her may not be accurate.

U-type participants are unknown to the system. A U-type individual may take a longer time to respond to request or not respond at all. Moreover, the data collected by him/her may not be accurate. Nevertheless, past experiences have shown that unknown crowds can help in many ways during major disasters.

(B) Sub-strategies

A strategy for managing CDC processes can be divided into three parts. They are sub-strategies for participant selection, result quality assurance and system-crowd interaction. We present here an overview of the participant selection part. To do so, we note that in general, a strategy for managing CDC processes needs to take into account of not only crowd composition but also the fact that sensor data on some regions of the threatened area may be more critical than data on other regions. Take the Gulf Coast during the BP oil spill as an example. We are more concerned with protecting regions that are frequented by tourists and/or have rich varieties of vegetations and wildlife. It makes good sense to direct high quality participants to check those regions for tar balls and other early signs of oil than other parts of the coast. This aspect of sensor coverage can be accounted for by giving an *importance value* v_i (or simply *value*) to each region in the threatened area: The more critical the region, the higher its value.

The problems solved by a *participant selection* (*crowd composition*) *strategy* include how to make best use of the available high quality participants to achieve specified goals subject to various constraints. To illustrate, we consider the simple case where the system uses only type-I participants to explore a threatened area that has n regions with values v_i, for $i = 1, 2, \ldots, n$. Suppose that the goal of a budget constrained CDC process is to maximize the total value of all the explored regions in the threatened area, under the condition that each region is to be fully explored or not explored at all. Then, the problem P_1 to be solved by the participant selection sub-strategy can be stated as follow:

$$P_1 : \quad \textit{Maximize} \qquad \sum x_i v_i \tag{1}$$

$$\textit{Subject} \quad \text{to} \qquad \sum x_i c_i \le B, \tag{2}$$

$$x_i \in \{1,0\}, \quad i = 1 \cdots n,$$

In (1) and (2), B is the total budget available for each CDC process, c_i is the cost of sending a sufficient number of responders to collect data for region i, and variable x_i is 1 if the region is to be explored and is 0 otherwise.

The integer programming problem P_1 assumes that the total number of type-I participants and the response time of the CDC process are unconstrained. In general, these constraints also need to be considered. For major disasters such as BP oil spill and southern California fire [1, 2], the system also needs to use registered but untrained type-M individuals and even unknown participants. We can formulate the constrained optimization problem of allocating I type and M type participants to regions similarly. Due to space limitation, we leave the problems of participant selection to a future paper [8].

Hereafter, we focus on sub-strategies for *result quality assurance* and *system-crowd interactions*. The former is concerned with ways to process sensor data reported by participants, which we will discuss shortly. The latter governs system-crowd interactions. Alternatives include random, crowd-driven and system-driven strategies defined earlier in Section 1. For studying their relative performance, we assume that the numbers of types I and M individuals participating in a CDC process are known and fixed during each CDC process and denote the numbers by N_I and N_M, respectively. The number N_U of type-U individuals is unknown and may vary during the process. Finally, we assume that all regions have the same value except for where it is stated otherwise.

3 Scenario, crowd and crowd mobility models

As it will become evident shortly, specifics about the characteristics of sensors are unimportant. To characterize the threatened area, we can start from the ideal condition. Ideally, the threatened area would be covered by a sufficient number of physical sensors at locations chosen to achieve the required spatial resolution. In other words, data provided by all the sensors would enable the system to generate a complete view of the area, including a sufficiently accurate estimate of the area boundary and fine spatial resolution.

Unfortunately, for reasons including the ones stated in Section 1, σ sensors S_1, S_2 ,…, S_σ are missing or broken. We assume here that the system knows their identities and locations. The goal of the CDC process is to acquire one or more eye-witness reports on the condition around the neighborhood of each missing sensor to complement data from existing physical sensors. Hereafter, we refer to such reports from participants as s*ensor samples* and *sample values*.

(A) Graph Model of Threatened Area

For the sake of managing CDC processes, it suffices for the system to characterize the threatened area by a directed graph containing σ nodes. Each node S_i in the graph represents a neighborhood of a specified size around a missing sensor S_i. There is a directed edge (S_i, S_j) from S_i to S_j if there are one or more paths along which participants can reach S_j directly from S_i. The label $T_{i,j}$ of the directed edge (S_i, S_j) is the minimal time required to go from S_i to S_j and upon arrival at the neighborhood of S_j, make an observation and send a sensor sample.

As an illustrative example, Figure 1(a) shows a part of a coastline threatened by oil pollution. The part should be under the watch of surveillance cameras and other physical sensors (e.g., for water quality) but is not. In the figure, the dots along the coastline mark the ideal locations of missing surveillance sensors. Human sensors, like physical sensors, at those locations can provide the system with needed data for complete coverage. Figure 1(b) shows the graph characterizing the scenario. In this case, the time to travel between two adjacent sensor locations is independent of direction of travel. We can simplify the graph by making it undirected and give each edge one label.

Figure 2 gives another example. The dots labelled S_i for i = 1, 2,…, 7 in part (a) of the figure mark where in a national park an early wildfire warning system should have sensors but does not. When there is a fire within a striking distance away, some combinations of low humidity, high temperature and wind direction and speed around those locations call for the evacuation of park visitors near by. Part (b) shows the graph maintained by the system for this scenario. In this case, the travel time between two sensor locations may depend on the direction of the travel, and not all sensor locations are connected by direct paths.

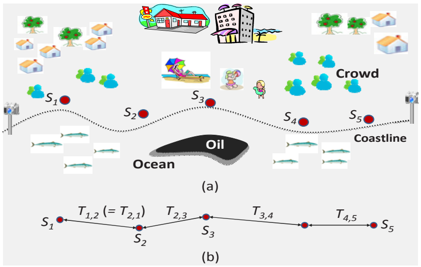

Figure 1: Oil spill disaster scenario.

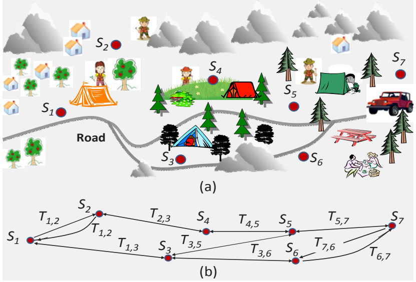

Figure 2: Wildfire surveillance scenario.

(B) Participant and Crowd Model

Similar to numerous details about disaster scenarios, many attributes of individual participants are unimportant for the purpose of managing CDC processes. Neither is what accurate sensor sample values are. The system can character each participant *k* abstractly by the following two sets of parameters.

Sample Errors The first set specifies the accuracy of sensor samples reported by the participant: Let η denote the number of sample values that participant *k* is to report at every sensor location.

$$\boldsymbol{\Theta}_k = (\Theta_{k,1}, \Theta_{k,2},\ldots, \Theta_{k,\eta}) \qquad (3)$$

is the *error* in each sensor sample. The error $\Theta_{k,i}$ in the *i*-th sample value is a random value with distribution function $F_{k,i}(x)$ (i.e., the probability that $\Theta_{k,i}$ is less or equal to *x*). Take the scenario illustrated by Figure 2 as an example. Participant *k* is requested to report temperature, humidity, wind direction and speed at each sensor location. In this case, each sensor sample contains four sample values. Errors in the values have different distribution functions.

Throughout this paper, we assume that these random variables are statistically independent. We also ignore the effects of such factors as technical problems, mob behavior, etc. and assume that sample errors of different participants are statistically independent. In case studies presented in subsequent sections, we do not further divide participants beyond types I, M and U. So, sample errors of participants of the same type are identically distributed. With a slight abuse of the notations, we index the distribution functions of these random variables by participant type and write them as $F_{I,i}(x)$, $F_{M,i}(x)$ and $F_{U,i}(x)$ (for $i = 1, 2, \ldots, \eta$) for types I, M, and U participants, respectively. Again, sample values reported by I-type participants are accurate (i.e., $F_{I,i}(x) = 1,\ x \geqq 0$, for all i).

How the system uses sample values reported by multiple participants of other types to improve result qualities depends on the types of sample values. For numerical sample values, the system can take average of sample values returned by the participants, knowing that the variance of error in the average decreases with the number m of reported values. For sample values that assumes binary values (e.g., presence of tar ball(s) detected or not detected), the system can also take average of the reported values. This is just a way of voting, with an average larger than 0.5 indicating that major participants reported TRUE or 1. An alternative is to take maximum (or minimum) of all sample values. For the oil spill scenario, this means that the system would take action to investigate as soon as some participant detected some sign of oil.

Response time The second set of parameters Δ_k and Π_k give the *response time per sample* of participant k:

$$R_k(i, j) = \Delta_k + \Pi_k T_{i,j} \tag{4}$$

Specifically, R_k (i, j) is the amount of time required by the participant k to travel from location S_i to S_j and take and report a sample of S_j upon arrival at neighborhood of S_j. Δ_k is the *delay per sensor*: Upon receiving a CDC request, or after reporting a sample at a location, the participant may not move on to the next destination location immediately. This random variable accounts for this delay. Here, we assume that the distribution function $G_k(t)$ of Δ_k is not a function of sensor location. A more detailed model may use different distribution functions for different locations and different steps during a CDC process.

The *efficiency factor* Π_k in (4) accounts for the extra time above the minimal time per sample taken by participant k. It is a random variable of value equal to or larger than 1. Its distribution function $H_k(x)$ is identically equal to zero for $x < 0$. Similar to errors in sample values, delays per sample and efficiency factors of participants of each type are statistically independent, identically distributed.

In addition to being accurate, Type-I participants are also prompt. For them, $\Delta_k = 0$, and $\Pi_k = 1$. We use $G_M(t)$, $G_U(t)$, $H_M(x)$, and $H_U(\mathrm{x})$ to denote the distribution functions Δ_k and Π_k, respectively, for types M and U participants. These distribution functions, together with distribution functions of sample errors and the numbers N_I, N_M, and N_U of participants of types I, M and U, respectively, completely characterize the composition of the crowd.

(C) Mobility Models

A mobility model characterizes the movement of a participant from sensor location to sensor location during a CDC process in conformance the system-crowd interaction strategy used by the system. Possible models include the ones listed below. With the exception of the shortest-time-tour (STT model), the models assume that every participant, regardless of his/her type, plans one move at a time without looking ahead. The descriptions of the models below are in terms of the graph model of the threatened area: We say that a participant is at node S_i when we mean that he/she is in the neighborhood around the location represented by the node. By that the participant has visited node S_i, we mean that he/she has already taken and reported one or more sensor sample at that location. We say that he/she chooses an outgoing edge of S_i when we mean that he/she chooses to go next to the location represented by the sink node of the edge. As stated earlier, types I and M individuals who responded to the CDC request at the start a CDC process remain to be participants until the CDC process terminates. In contrast, at any step of the process, a type U individual may drop out with probability $D > 0$. The statements below are conditioned on that the participant at S_i does not drop out after visiting the node.

- *Random Walk* (*RM*) *Model*: After visiting $S_{i,}$ the participant is equally likely to choose any of the outgoing edges of S_{i}.
- *Random Walk Forward-Only* (*RMFO*) *Model:* The participant first discard from consideration all outgoing edges of S_i leading to sink nodes he/she has already visited and then chooses one edge among the remaining outgoing edges with equal probability.
- *Random-Least-Visited-Firs*t (*RLVF*) *Model*: The participant first marks the outgoing edges leading to sink nodes that have been visited fewest times by all participants and then with equal probability chooses an edge from the marked edges.
- *Global-Least-Visited-First (GLVF) Model:* The participant chooses with equal probability an outgoing edge among the outgoing edges in path(s) to the node (or nodes) that have been visited the least number of times among all nodes in the graph.
- *Shortest Time Tour* (*STT*) *Model*: The model assumes global knowledge of the graph model of the threatened area. Each participant follows a tour computed for him/her so that every node is visited by a specified number of participants in the shortest time.

RM and RMFO models are the only mobility models that are applicable when the system uses the random system-crowd interaction strategy. Again, RM is a pure random walk model. Take the scenario in Figure 1 as an example. A participant who is at S_2 when a CDC process starts is equally like to go left and right, back and forth until the system terminates the process. He/she is likely to have visited all the nodes if the process runs a sufficiently long time. A shortcoming is that he/she is also likely to visit some nodes (e.g., S_2, S_3 and S_4 in this example) many more times than other nodes.

Now, suppose that the participant chooses each move according the RMFO model and chooses S_3 from S_2. From S_3, the only choice is S_4, and at S_4 the only choice is S_5. At S_5 he/she has no node to visit and hence essentially drop out of the process. It is easy to see that unless the graph for the threatened area is nearly fully connected, participants should not follow this movement model.

The RLVF and GLVF models are applicable when the system uses the crowd-driven strategy and provides participants with the current numbers of visits of all the nodes. They appear to be better alternatives than RM and RMFO models. According to RLVF model, preference is given to adjacent nodes that have been visited the least number of times at the time. The model still has the common shortcoming of all mobility models which do not make use global information on connectivity of the sensor locations. GLVF model is a possible remedy, but each participant must consider all nodes for every move.

The STT model assumes that the CDC process is system directed. After receiving responses to its CDC request, the system computes for each responded participant a tour through the threatened area such that all sensor locations are explored by a specified number of participants in the shortest time. We note that if there is only one participant, the tour sought by the system is a solution of the well-known travelling salesman problem, which is known to be NP-hard. Given multiple participants, each of them only needs to visit nodes in a sub-graph. We want the maximum of the minimum lengths of their tours to be as short as possible. We need efficient heuristics to solve this variant of the travelling salesman problem and will present the heuristics in [8].

4 Experiment setup

To determine the relative performance of different system-crowd interaction and crowd movement strategies, we conducted several simulation experiments based on two disaster scenarios: oil spill disaster and wildfire surveillance. For the oil spill disaster, we used Figure 1(b) to represent the threatened area but added two nodes to represent two more ideal locations for missing surveillance sensors along the coastline. In total, we have seven locations. From left to right, they are $S_{1,}$ $S_{2,}$ …, and $S_{7.}$ Any two adjacent locations are connected. In the wildfire surveillance scenario, we used Figure 2(b) to represent the threatened area. To keep the number of parameters small, we experimented with only the case where the distances between all adjacent locations are equal.

In both scenarios, a CDC request is broadcast to start a process of collecting sensor readings at all locations. The data reported in the next section were obtained from an experiment where the system uses different strategies to interact with different types of participants: Specifically, it uses the crowd-driven strategy to interact with types I and M participants and provides them with feedback so that they can move according to the Random-Least-Visited-First (RLVF) model. The system uses random strategy in its interaction with unknown crowd. Without feedback and guidance, type-U participants have no choice but to move according to the Random-Walk (RW) model. Regardless of his/her type, each participant moves to an adjacent location based on his/her mobility model after reporting a sample at each location.

To simplify our experiments, the number η of sample value taken at every location was set to one. A node is considered *visited* (and is marked as such) when a sufficiently accurate estimate of the sensor value for the location represented by the node has been obtained. By definition of type I, a node (location) is marked as visited immediately after a type-I participant has visited the node. For types M and U participants, we used *sample mean* (i.e., the average of sample values reported by these types of participants) at a location as a *sensor value estimate* (i.e., an estimate of the accurate sensor reading) for that location. The node is marked visited as soon as the standard deviation of the sample values reported by all types M and U participants who have visited the node become equal to or less than a specified threshold percentage of the estimate. We refer to this threshold as the *acceptance threshold.* The CDC process stops when all locations are marked as visited.

Parameters of the simulation experiments include distribution functions of error in sample value, delay per sensor and efficiency factor for types M and U participants. The data presented in Section 5 were taken in an experiment where the sample value reported by each type-M participant was randomly generated in the range [5, 45] with uniform distribution, and the sample value reported by each type-U participant was randomly generated in the range [0, 50] with uniform distribution. In other words, sample errors are uniformly distributed in ranges [-20, 20] and [-25, 25], for type-M and type-U participants, respectively. The acceptance threshold was set to 15%.

We set the minimum time per sample (i.e., the values of all edge labels) taken by a participant to 10 for all experiments. The efficiency factor of a type-I participant is 1, by definition. The efficiency factor of a type-M participant is randomly generated in [1, 2] with uniform distribution and of a type-U participant is randomly generated in [1, 10] with uniform distribution. For simplicity, we set to zero the delay per sensor for all participants and the drop-out probability of type-U participants. Finally, to remove the effect of initial locations of participants, we let all participants start at location S_4.

5 Performance measures and simulation results

We compare different strategies along two dimensions. The first is the *response time of the CDC process:* it is the length of time between when the system issues a CDC request to the instant when all nodes are marked visited.

The second performance measure is the spatial resolution of sensor coverage achieved by the process. To define this performance measure, let *h* denote the required time for a type-I participant to visit all locations: The s*patial resolution* achieved by a strategy is defined as the ratio of the number of visited locations in the duration *h* to the number of total locations. The higher the spatial resolution is, the better the strategy is along this dimension.

Figure 3 and Figure 4 show spatial resolution as a function of the number of participants. The data were taken in an experiment series in which there was only one type of participants for each experiment. As the figures show, the more type-M participants take part during a CDC process, the better the spatial resolution. In particular, Figure 3 shows that the spatial resolution becomes 100% when there are three or more type-M participants. In other words, in this oil spill disaster scenario, it is not necessary to use any trained type-I participant to patrol the area when there are more than three type-M participants. In contrast, type-U participants are not always helpful. As we can see from Figure 4, the spatial resolution does not improve noticeably when the number of type-U participants increases from three to six.

Figure 5 and Figure 6 show the dependency of the response time of the CDC process on the number of participants. Again, in each experiment, there is only one type of participants. We can see that in general, the more participants are involved, the shorter the response time is. When the number of participants becomes large, the difference among crowds becomes small. As Figure 5 shows, the response time of six type-M participants is almost the same as that of six type-I participants. In addition, both Figure 5 and Figure 6 show that when all participants are of type I, the response time of the CDC process does not improve much when the number of participants increases from two to six. Our results also indicate that participants following RLVF model perform much better than the ones following the RM model. For example, in Figure 6, the response time of a type-U participant is almost five times longer than that of a type-M participant.

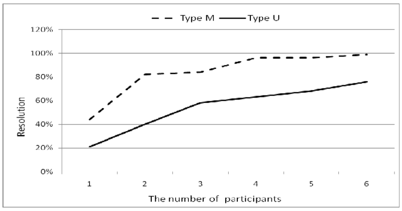

Figure 3: Oil spill disaster scenario.

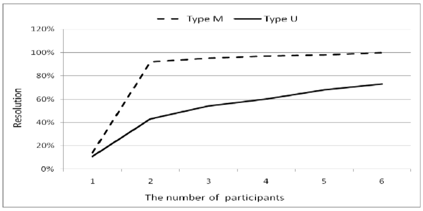

Figure 4: Wildfire surveillance scenario.

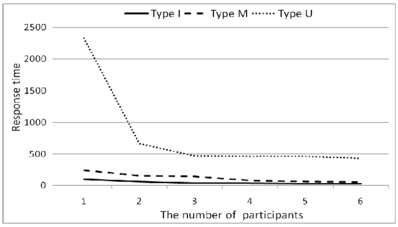

Figure 5: Oil spill disaster scenario.

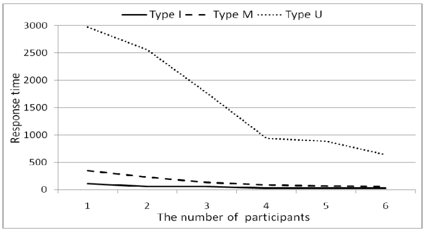

Figure 6: Wildfire surveillance scenario.

To further investigate the effect of crowd composition, we mix different types of participants in an experiment: Each crowd includes six participants. The response times for different crowd models are listed in Table 1. The 3-tuples in the first row of the table represent crowd models. Numbers in the 3-tuples gives the numbers of types I, M and U participants in the crowd, respectively. We can see that a crowd with type-I participants solely invariably achieves a short response time. Moreover, a crowd without type-I participants always performs worse than a crowd with type-I participants. The response time achieved by a crowd without type-I participants depends on the number of type-M participant As an example, in the wildfire surveillance scenario, the response time of crowd (0, 3, 3) is about 1.39 times longer than that of crowd (0, 6, 0).

Table 1: Response time of different crowd model.

	(6, 0, 0)	(3, 3, 0)	(2, 2, 2)	(3, 0, 3)	(0, 6, 0)	(0, 3, 3)	(0, 0, 6)
Oil spill disaster	30	42	49	54	56	220	431
Wildfire surveillance	30	37	49	52	61	85	642

6 Summary and future work

Previous sections presented general models that can be used to represent different disaster scenarios and characterize individual participants in a crowd and their movements in the threatened area for the purpose of studying strategies for crowdsourcing sensor data. The models abstract away irrelevant details about the disaster scene, in-situ physical sensors, individual participants and their movements so that we can focus on the characteristics of elements that are important for managing crowdsourcing sensor data collection.

The preliminary data presented above show that in general, the more participants are involved in a CDC process, the shorter the response time of the process and a crowd of type-I participants only usually has a short response time. As stated earlier, the response time may not improve noticeably when the number of type-I participants increases beyond some value, however. This fact indicates a need for effective methods to properly allocate trained responders for data collection function. Developing such methods is a part of our future work. For a crowd without type-I participants, our simulation results show that the number of type-M participant can have a significantly impact on the response time. On the other hand, type-U participants are not always helpful.

We made many simplifying assumption in studies done thus far. Most important ones are that the random variables characterizing relevant attributes of participants and sample errors are statistically independent. This assumption is clearly not always valid and will be removed in our future studies. As stated earlier, we leave problems that are theoretical in nature to a technical report [8] on theoretical foundation of the CDC process management.

References

[1] Deep Horizon Oil Spill, http://en.wikipedia.org/wiki/Deepwater_Horizon_oil_spill and http://www.google.com/crisisresponse/oilspill/

[2] 2009 California Wildfire, http://en.wikipedia.org/wiki/2009_California_wildfires

[3] Typhoon Morakot Aftermath, http://en.wikipedia.org/wiki/Typhoon_Morakot#Taiwan_3

[4] van Exel, M, E. Dias and S. Fruitier, "The impact of crowdsourcing on spatial data quality indicators," The 6th International Conference on GIS, September 2010, www.giscience2010.org/pdfs/paper_213.pdf

[5] "Crowd quality and normal testing," Testing of the Future, January 2010.

[6] Hui, P. and J. Crowcroft, “Human mobility model and opportunistic networks,” *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences,* June 2008.

[7] Schlink, U. et al, “Evaluation of human mobility models, for exposure to air pollutants,” *Sci. Total Environment*, August 2010.

[8] Chu, E T.-H., J. W. S. Liu and J. K. Zao, “Theoretical foundation of strategies for crowdsourcing sensor data collection,” Tech. Report No. TR-IIS-11-001, Institute of Information Science, Academia Sinica, Taiwan, 2011.