Who to query? A two step querying technique for tracking real-time variant/unknown event distributions

Mai ElSherief
Dept. of Computer Science
UC Santa Barbara
mayelsherif@cs.ucsb.edu

Ramya Raghavendra
IBM T. J. Watson Research Center
rraghav@us.ibm.com

Elizabeth Belding
Dept. of Computer Science
UC Santa Barbara
ebelding@cs.ucsb.edu

ABSTRACT
Some abstract sentences...

1. INTRODUCTION
Some introductory sentences...

2. RELATED WORK
- Spatial task distribution (maximizing task assignment)
- Applications: emergency scenarios, safety applications, etc. - crowd sourcing and crowd sensing

3. RESEARCH QUESTION AND PROPOSED TECHNIQUE
In this paper, we envision a world where users can be probed to contribute to an unanswered question. An unanswered question can be related to a phenomenon that needs to be tracked under constraints of N resources. Examples include disastrous and safety applications. In particular, we have a two-dimensional grid and a number of objects that can sense the environment around them. These objects can be humans, artificial sensors, mobile phones or even robotic sensors. If we are interested in answering the question "How are things going in this grid?", we can basically ask or query all the objects in the two-dimensional space and aggregate their findings. In this paper, we assume that to answer this question, you can only query N objects. Hence, the question becomes: *Given N resources, who should you select to track a real-time phenomenon?* Answering this question becomes essential in the case of a limited bandwidth of resources. This is particularly important in emergency scenarios when a network's performance degrades and preserving energy and other resources become important.

If we attempt to tackle this question from a probabilistic point of view, then the straightforward answer would be to try to select objects/users with the same probabilistic distribution as the phenomenon. For instance, if we know that a certain phenomenon occurs at different places in the two-dimensional grid uniformly, then we would have no bias in selecting the users to query i.e. each user/object would have the same distribution of being selected to be queried. On the other hand, if we know the phenomenon we are interested in querying is more prevalent in certain areas of the grid as opposed to other areas, we would take that into consideration when we are selecting the users and select more users to query in this are and fewer users in other areas where there is a smaller probability of occurrence.

But what if you do not know the distribution or what if the distribution of the phenomenon is time variant? The aforementioned question becomes more interesting in this case and we can then inquire if there is a systematic algorithm that can be used for querying/selecting users to track a phenomenon regardless of the probabilistic distribution or time variations.

In this paper, we introduce a two-stage technique that can be used to select N users to track a real-time phenomenon with no prior information about the event distribution. The technique outperforms the random user selection by a percentage of $20-63\%$ on average in terms of number of users chosen that were close in the events and outperforms the dispersion maximization technique by a percentage of $20-68\%$ on average.

3.1 Technique Description
We assume that we have M users in our two-dimensional grid and that the system that selects a user to query is bounded by N resources where $N < M$. Each of the M users has a specific location in the grid determined by a two-dimensional system e.g. (x, y) or a (lat, long). We also assume that the users selected will participate in answering the question of interest to the system and fully co-operate. A pre-selection phase can be used to eliminate users that aren't likely to co-operate or users who can provide false information using a system of building trust over time. The ways to rule out users based on trust or refusal to co-operate is not the main focus of this paper. Instead, we focus on how to select N out of M users to where $N < M$ to keep track of events occurring in the two-dimensional grid.

Our technique combines K nearest neighbor (KNN) queries with querying users to maximize the dispersion of their location in the grid as depicted in Algorithm 1. We devise the

selection of users into two stages. In the first stage, our goal is selecting users with the aim of maximizing the dispersion of their locations. Based on the crowd feedback in the first stage, we go into a more fine-grained selection. The users that provide a positive feedback (i.e. they witness an event/ emergency in their location) are called the pivot users. In the second stage, we aim to get the K nearest neighbors for the pivot users. We assume that because the pivot users witness an event, the K nearest neighbors will witness another event of the same type in a neighboring area.

The aforementioned technique assumes full trust in the first stage users to respond and provide unfalsified responses. To remedy that, we can explore dividing the selection of the second phase users into two groups: a group comprising of the KNN of the pivot users and another group that aims to maximize the dispersion. In this section, we will focus on studying our two-stage querying technique with the assumption of having full trust in the crowd and discuss the remedy in subsequent sections.

Algorithm 1 Two-stage querying algorithm

```
1: function selectUsersFromGrid (firstStageRatio, N)
2:     selectedUsers = {}
       usersFeedback.size == 0
3:     selectedUsers = y
4:     a = b
5:     secondStageUsersCount = M − ⌊(firstStageRatio ∗ N)⌋
       Attempt (k= secondStageUsersCount/ firstStageUsers)
6:     Divide it
7:     quota = y
8:     append k
9:     return selectedUsers
```

4. EXPERMIENTS

In order to quantify the performance of our technique, we test it under different scenarios. We investigate the technique using three types of data spread: clustered, uniform and real datasets. In our experiments, we compare our algorithm in the selection of users to two policies as follows.

- Random user selection: For this policy, we select N users randomly based on a uniform distribution.
- Selection based on dispersion maximization: The selection of users in this policy depends on selecting N users from the crowd who maximize the dispersion of their locations.

4.1 Experiments Variables

There are multiple variables that can be controlled to test the behavior of the two-stage querying technique. Table 1 explains the most important variables.

The environment settings are related to the size of the $2D$ matrix, the number of incidents and their distribution across the matrix and the number of resources to choose from. In all of our experiments, except the case study, we set up the $2D$ matrix as a 10 by 10 matrix. We show results for incident

Environment settings:
• matrix dimension: represents the length and the width of the $2D$ spatial matrix. We model the spatial area under investigation as a $2D$ square matrix.
• incident count: number of incidnets distributed across the cells of the spatial matrix
• resources or crowd count: the M resouces from which N, where $N < M$, will be chosen to query
Query settings:
• N: the number of resources the system is limited by to query/sense
• first stage percentage: the percentage of users/sensors of the N resources that will be selected to query in the first phase. In our analysis, we test the cases of selecting $20\%, 40\%, 60\%$ and 80% of the N resources in the first stage.
• k setting: used to identify the KNN crowd individuals/sensors to an incident
Approximation settings:
• Maximization trials: number of attempts to maximize the dispersion of selected individuals/sensors from the crowd

Table 1: Different parameters of the two-stage querying technique.

count of 50 and number of resources or M of 100. We varied the environment settings in our experiments and no noticeable differences were observed in performance. Instead, we focus on varying the query settings to better understand our querying technique. In this section, we will focus on varying the first stage percentage and leave the variation of the k setting to the following section. We also show results for $t\_setting = 30$ which constitutes 30% of the available resources (M). We notice that the gap between the performance of our technique and the other techniques increase when $t\_setting$ decreases and all the techniques converge in performance when $t\_setting$ approaches M.

To measure the performance of our technique in comparison to other forms of selection, we utilize two different metrics namely count of people/sensors queried in the KNN of incidents and the number of incidents covered by the people queried. The two metrics are formally defined as follows.

- Close people count: This is measured as the absolute number of people/resources in the KNN of each incident for all incidents. This is formally represented as follows:

$$Closepeoplecount = \sum \forall_{incident i} |(KNN_i \cap QU)| \quad (1)$$

  where QU (the "Queried Users" set) is the set of users selected for querying.

- Coverage: measured as the number of incidents covered out of the total number of incidents occurring in the $2D$ matrix. We assume an incident is covered if at least one of the people/resources in the incident's KNN was queried. This is formally measured as:

$$Coverage = \sum \forall_{incident i} (Coverage_i) where, \quad (2)$$

First stage percentage	20%	40%	60%	80%
Surge over Random	62.5%	58.89%	35%	20%
Surge over Dispersion Maximization	67.8%	62.32%	39.8%	20.64%

Table 2: Surge of Two Stage technique in comparison to Random and Dispersion Maximization techniques.

$$Coverage_i = \begin{cases} 1, & \text{if } (KNN_i \cap QU) \neq \phi \\ 0, & \text{otherwise} \end{cases}$$

4.2 Clustered data experminets

In this subsection, we aim to test our technique in a scenario where the events take a clustered form. Geographer Waldo R. Tobler's stated in the first law of geography: "Everything is related to everything else, but near things are more related than distant things." In this subsection, we assume that the incidents are related to each other and that they take a clustered form i.e. they form clusters across the $2D$ spatial matrix as seen in Fig 1.

For these type of experiments, we vary the number of clusters in our $2D$ matrix from one cluster to ten clusters while fixing the resources or crowd count to be 100. To ensure data variability, we model the size of each cluster as a random variable while making sure that the aggregated size of all the clusters is equal to the crowd count. For each case of varying the number of clusters, we average over 100 different configurations. The objective in this section is to measure the effect of variation of the first stage percentage on our performance metrics.

Figures 2a, 2b, 2c, 2d depict the results for Close People count when varying the first stage percentage from 20% to 80%. We notice that our two-stage querying technique always outperforms the Random crowd/sensors selection and the selection based on maximizing the dispersion only. Table 2 depicts the amount of surge in Close people count in comparison to Random and Dispersion Maximization techniques. We notice that as the amount of resources queried in the first stage decreases, the number of close people queried increases. This is due to the fact that when the first stage percentage decreases, the second stage resources increase under limited resources constraints which focuses on resources close to incidents detected in the first stage. On the other hand, incident coverage tends to increase as the first stage count increases. This is depicted in Figure 3. We can also notice that both Close people count and Incident coverage tends to increase with number of clusters till the number of clusters is around four or five and then decreases.

4.3 Uniformly distributed data experminets

- put uniformly distributed results here

4.4 Long Tail Distribution

- put long tail results here

4.5 Case Study: Hollaback harassment data set

After applying the two-stage querying technique to the previously mentioned three distributions (clustered, uniform and long-tail), we wish to examine the technique under real incident distributions. In order to do that, we test our querying technique on a global street harassment dataset provided by Hollaback [1].

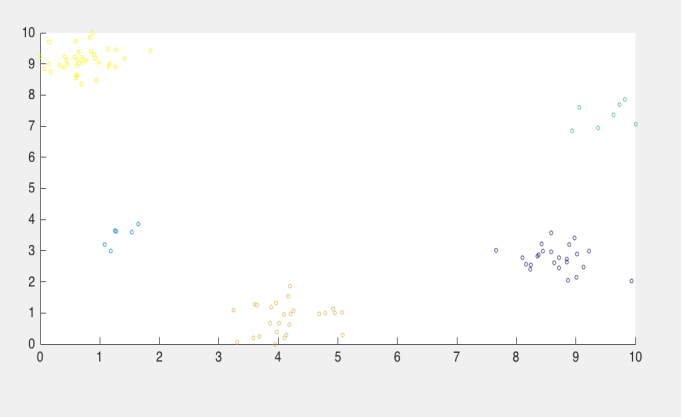


Figure 1: An example of a $2D$ spatial matrix with 5 clusters

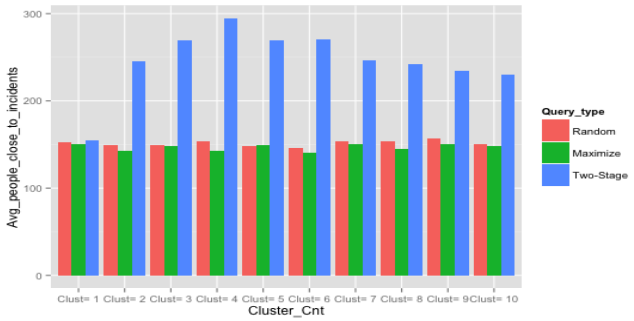
4.5.1 *Data Overview*

Hollaback [1] is a non-profit movement powered by local activists in 92 cities and 32 countries to end street harassment. The Hollaback project collects data on street harassment events worldwide. Through the Hollaback phone app and the online platform, users can report stories of street harassment to share with the Hollaback community. This empowers victims to speak out about everyday harassment and spread the word about the prevalence of these events. In some communities, local governments are informed in real-time about street harassment so that there is a system-wide level of accountability. In addition, the Hollaback app uses GPS to record a data set representing the locations of street harassment events as a means of improving the collective understanding of street harassment and how it can be prevented. As of January 2016, over 8000 street harassment incidents have been recorded in their dataset since February 2011. It is on this data set that we wish to test the two-stage querying technique.

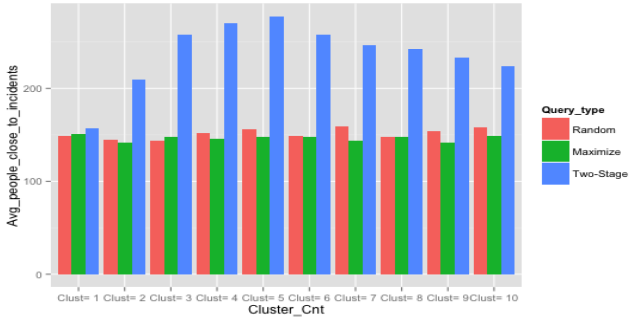
4.5.2 *Analysis*

Going through the Hollaback dataset, we select different cities for which we have enough harassment samples for statistical significance (i.e. more than 30 samples). We test the performance of Random selection, Dispersion maximization selection and the two-stage querying on six different cities: Paris in France, Brussels in Belguim, Berlin in Germany, Baltimore, Maryland in the US, Buenos Aires in Argentina and Istanbul in Turkey. In this paper, we show results for Paris, Brussels and Istanbul.

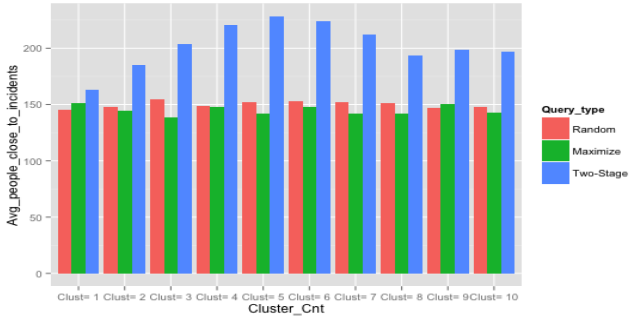
In order to seperate the collected Hollaback dataset into different cities, we used bounding box coordinates. After that, we drew the border lines for the different cities and removed any outliers from our datasets. Figure 4 shows the distribution of events for the different cities. The Paris dataset contains 197 harassment incidents and covered an area of 28.2 sq mi while the Brussels dataset contains 154 incidents covering a geographic area of 28.4 sq mi. Istanbul



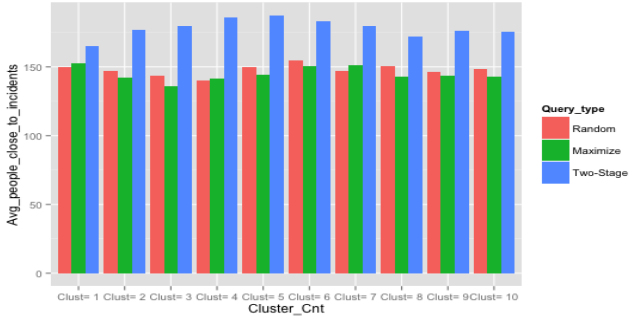
(a) Average number of people close to the incidents (Close people count) when maximizing the dispersion with 20% of available resources.



(b) Average number of people close to the incidents (Close people count) when maximizing the dispersion with 40% of available resources.



(c) Average number of people close to the incidents (Close people count) when maximizing the dispersion with 60% of available resources.



(d) Average number of people close to the incidents (Close people count) when maximizing the dispersion with 80% of available resources.

Figure 2: Varying the first stage percentage with different number of clusters.

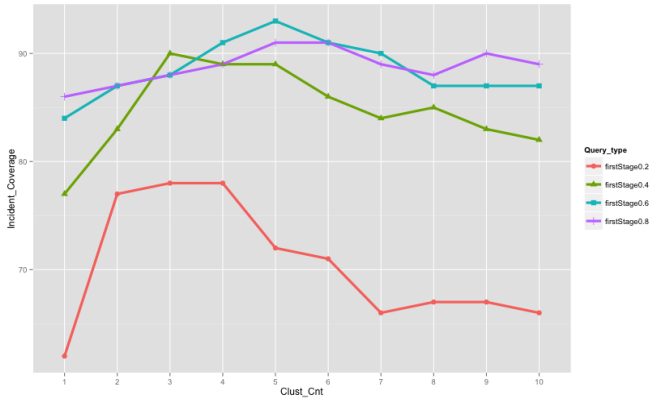


Figure 3: Incident coverage for different values of First stage percentage.

had 87 reported incidents covering an area of 138 sq mi on the left of Bosporus Strait and 69 sq mi on the right.

For each of the cities, we generate different variations of uniformly distributed crowd ($M = 100$) across the city. In this kind of analysis the parameters, matrix dimension and incident count, are not controlled by our analysis but rather enforced by the dataset. We measure the Close people count and the Incident Coverage for all three querying techniques and plot the results in Figure 5 and Figure 6 respectively. We notice that the Two-stage technique outperforms both the Random and Dispersion Maximization in terms of Close People count for all three cities. In terms of incident coverage, Figure 6 shows that dispersion maximization achieves maximum incident coverage. The figure also shows that the two-stage technique can achieve this maximum by setting the first stage percentage to be 80%. The aforemntioned two figures suggest that there is an inherent tradeoff between accuracy and coverage under constrained resources which we will discuss in detail in later sections. The figures also suggest that the two stage technique under setting the first stage percentage to be 80% can achieve a balance between accuracy and coverage.

4.6 Stressing the two stage querying technique (k=1)

4.7 Technique Variations

- prior knowledge variation - second stage division

5. DISCUSSION

- Our assumptions and limitations...

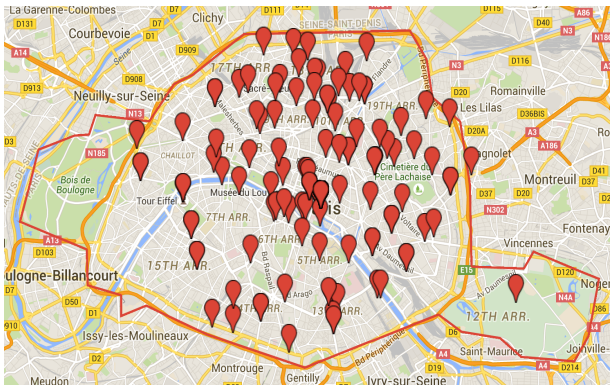
6. CONCLUSIONS

In this paper, we introduced....

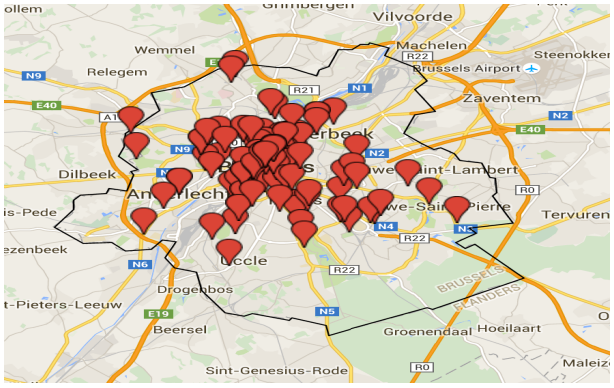
7. ACKNOWLEDGMENTS

8. REFERENCES

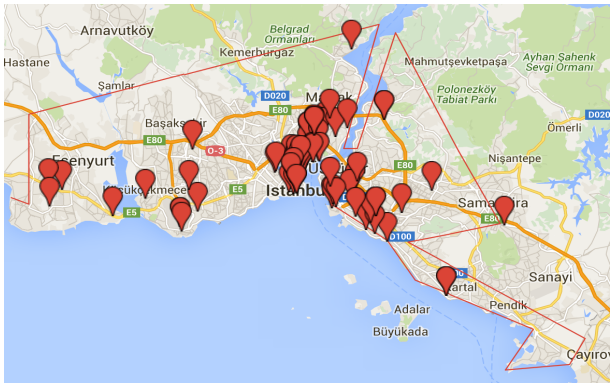
[1] Hollaback. Read and Share Stories. When it comes to street harassment, you are not alone. `http://www.ihollaback.org/share/`, 2015. [Online; accessed July-2015].



(a) Hollaback harassment reports in Paris.



(b) Hollaback harassment reports in Brussels.



(c) Hollaback harassment reports in Istanbul.

Figure 4: Distribution of harassment incidents across Paris, Brussels and Istanbul.

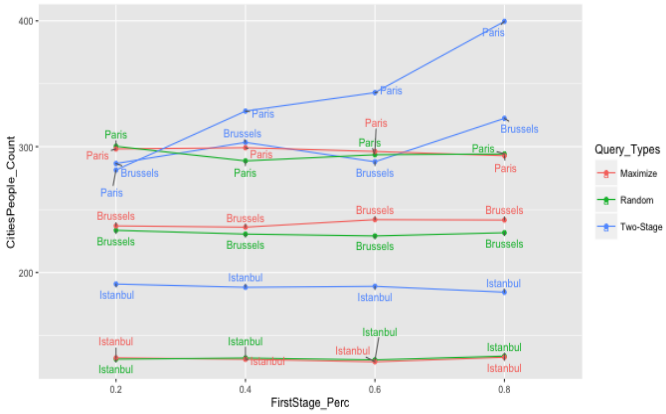


Figure 5: Close people count for different values of First stage percentage.

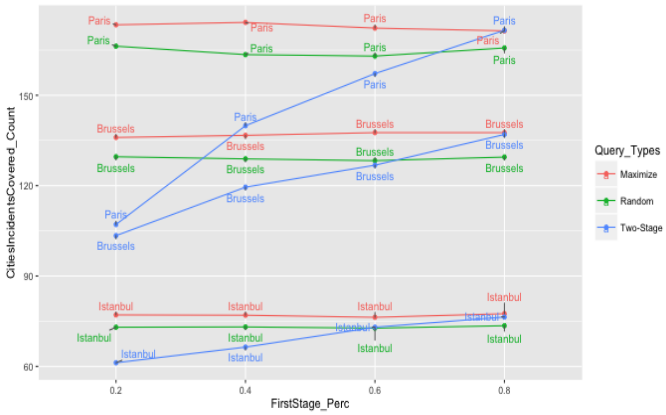


Figure 6: Incident coverage for different values of First stage percentage.