# Duplicate Report Detection in Urban Crowdsensing Applications for Smart City

Jize Zhang
Department of Civil & Environmental Engineering & Earth Sciences
University of Notre Dame
Notre Dame, US
jzhang14@nd.edu

Dong Wang
Department of Computer Science and Engineering
University of Notre Dame
Notre Dame, US
dwang5@nd.edu

***Abstract*—Crowdsensing has become an emerging data collection paradigm for smart city applications. A new category of crowdsensing-based urban issue reporting systems have been developed to enable pervasive and real-time monitoring of urban infrastructure malfunctions. A key challenge exits in such systems is the *duplicate report problem* where uncoordinated users may submit redundant reports about the problem caused by the same underlying issue. The duplicate report problem has a significant economic impact on the municipal government. This paper develops a new *duplicate report detection* scheme that accurately detects the duplicate reports using an Expectation Maximization (EM) framework. The new duplicate report scheme has been evaluated on both synthetic and real world datasets collected from Chicago smart city applications. The results showed that our scheme significantly improves duplicate report detection accuracy compared to the state-of-the-arts.**

***Keywords—Duplicate Report Detection, Smart City, Crowdsensing, Urban Infrastructure, Expectation Maximization***

## I. INTRODUCTION

This paper presents a new analytical framework to address the duplicate report detection problem in urban crowdsensing applications for *smart city*. Urban infrastructures (e.g., roads, street lights, sidewalks, traffic lights) are an essential part of a city and monitoring their conditions has recently received a significant amount of attentions [1-2]. With the proliferation of smartphones and the ubiquitous Internet connectivity, crowdsensing has become an emerging data collection paradigm for smart city applications (e.g., noise monitoring, earth quake detection, green navigation) [1-4]. Along this trend, a new category of crowdsensing-based *urban issue reporting systems* have been developed to allow common citizens to report, track, and comment on the urban infrastructure malfunctions they encountered, which leads to a more efficient repair and issue fixing process [5-6]. However, a key challenge exists in such issue reporting systems: more than one report could be submitted by uncoordinated participants about the same underlying issue/fault. We refer to this challenge as *duplicate report problem*.

Crowdsensing brings rich human intelligence into the sensing process, but its has also introduced a significant drawback in the urban issue reporting systems: the same issue might be reported by a user repeatedly, or by different users at different time/locations. The duplicate reports would require inspectors to be sent out multiple times for the same issue, which would result in an increase of the inspection cost and decrease of the response efficiency. For example, the cost for Chicago municipal to send multiple inspectors to the same issue due to duplicate reports was *$6.9 million* in streetlight issues alone per year [5]. Duplicate reports account for *32%* of total reports documented by Chicago municipal on *See-Click-Fix* datasets [8]. Therefore, an efficient and accurate *duplicate report detection* scheme could help the government to greatly reduce the management cost, shorten the response time and optimize the maintenance troop allocation. In this paper, we develop a *new duplicate report detection scheme* that is able to accurately identify the duplicate reports collected from a real world urban issue reporting system for smart city [1].

There exist two important challenges in designing a duplicate detection scheme for the urban issue reporting systems. The *first challenge* is how to determine if two reports are duplicates or not. Two street-light-broken reports with different addresses would be tagged as duplicate if they result from the same malfunctioning electricity grid. Thus, a trivial duplicate detection scheme that only considers reports with the exact same physical address as duplicates would result in a low accuracy. The *second challenge* is the lack of suitable datasets and ground truth. Most related datasets do not have labels indicating if a report is a duplicate or not, and even for datasets that have such labels, they do not contain information about the pairwise relationship between the duplicate reports [6].

In this paper, we develop a new duplicate report detection scheme to detect duplicate reports based on its categorical, temporal and spatial features. In particular, we formulate the duplicate detection problem as an Expectation Maximization (EM) problem and develop a *fully unsupervised* binary classification approach to identify duplicate reports. We evaluate the new duplicate report detection scheme on both synthetic datasets and real world datasets collected from Chicago smart city applications [8]. The results showed that our approach outperforms the state-of-the-art baselines by significantly improving the duplicate report detection accuracy. The results of this paper is important because it allows crowdsensing based urban issue reporting system to accurately identify the duplicate reports with little prior knowledge about the crowd and the urban infrastructure. This scheme can be readily used in future smart city and urban sensing applications to improve the city management efficiency and reduce the infrastructure maintenance cost.

The contributions of the papers are summarized as follows:

- To the best of our knowledge, the proposed method is the *first* attempt to solve the duplication

[1] http://www.seeclickfix.com/

detection problem in crowdsensing-based urban issue reporting systems using a *fully unsupervised* classification approach.

- The proposed duplicate detection scheme could flexibly incorporate different cost models.
- We showed significant performance gains achieved by our approach through a real world urban crowdsensing application (e.g., the new scheme achieves the highest duplicate report detection accuracy compared to the state-of-the-art baselines).

The rest of this paper is organized as follows: in Section II, we review the related work. In Section III, we formulate duplicate detection problem in an urban crowdsensing scenario and present the duplicate detection scheme. Evaluation results are shown in Section IV. We conclude the paper in Section V.

## II. Related Work

Crowdsensing has emerged as a new smart city application paradigm that empowers common citizens to voluntarily contribute their observations about a city at an extremely large scale [11]. A comprehensive overview of crowdsensing applications can be found in [12]. Some important challenges in crowdsensing include energy efficiency [4,13], data quality [14], resource allocation [15,16], privacy preservation [17], and incentives design [18]. In this paper, we study a new challenge in urban crowdsensing applications: *duplicate report detection*, which has not been quite well addressed in the current literature on urban crowdsensing.

Extensive works have been carried out in the field of mining the urban infrastructure sensing data. For example, Zha et al. proposed a random forest based approach to accurately predict the volume of issue reports in New York City [19]. In [20], a fine-grained noise mapping of NYC has been inferred using 311 complaint data, social media, road network, and points of interest data. The municipal benefits of adopting urban crowdsensing has also been studied using a simulation based approach in [21].

Specifically, Budde et al. studied the duplicate detection problem in urban crowdsensing applications [8], which is the closest to our work. Their approach transformed the sensing data into a graph and applied existing community detection techniques to cluster the graph. All reports that belong to the same cluster were considered as duplicates. They also showed that their method outperformed other density-based techniques for spatial clustering (e.g., ST-DBSCAN and ST-GRID). However, their solution is very heavyweight and requires sufficient training data to produce reasonable results.

In contrast, we formulate the duplicate report detection problem as a binary classification problem where duplicate reports are classified based on the categorical, temporal and spatial information. Our binary classification approach explicitly exploits the pairwise distance information and is able to identify the duplicate reports with the "same root cause" that normally cannot be found by the density based solutions [8].

## III. Framework of EM-based Duplicate Detection

### A. Problem Formulation

First let's define the terms for the problem formulation. A ***report*** is an observation of malfunctioning urban infrastructure issues, which has one temporal dimension (i.e., creating time), two spatial dimensions (i.e., latitude, longitude) and a ***category*** (e.g., street lights, traffic signs, potholes) description. The user who creates the report is referred as a ***citizen***, while the entities processing and responding to the report are referred to as ***municipal***. By submitting reports via mobile apps or phone calls, citizens could act as sensors to report malfunctioning urban infrastructure issues. Such citizen-as-sensor approach normally has a good coverage as citizens are naturally distributed across the city. However, a major problem of this approach is the *duplicate reports*. If the municipal's capacity to process and respond to the reports are limited, the efficiency of municipal would be significantly decreased when multiple reports are submitted about the same issue that only needs to be processed once [8].

In this paper, we consider the report data available in the form of events, $E = \{E_1, \ldots, E_N\}$ where $E_i = \{e_i^1, \ldots, e_i^{n_i}\}$ and $e_i^j$ is the $j$th report in the $i$th category, $1 \le i \le N$. Each report $e_i^j$ consists of a tuple $< s(e_i^j), t(e_i^j) >$, where $s(\bullet)$ is the pair of longitude and latitude, and $t(\bullet)$ is a timestamp indicating the creation time of the report.

We define a report to be *duplicate* if the municipal identifies the existence of earlier reports about the same urban infrastructure issue. Following the definition in [5], the possibility of two reports in different categories referring to the same issue are not considered. Specifically, for an arbitrary report, we define its *relevant report set* as the other reports in the same category created in a certain time period before it: $R(e_i^j) = \{e_i^k \in E \mid 0 \le t(e_i^j) - t(e_i^k) \le \Delta T_i\}$. If no report exists in a report's relevant report set, the report is defined to be *original*; otherwise, it is a *duplicate*.

We further introduce the following two types of distinct misclassification cost:

- **False-positive cost ($c_+$):** The cost associated with an original report being labelled as duplicate (e.g., the cost to address the complaints due to the lack of response).
- **False-negative cost ($c_-$):** The cost associated with a duplicate report being labelled as original (e.g., the cost of unnecessary inspection).

*Our goal is to build a duplicate report classifier that minimizes the misclassification cost associated with reports*. Given the report data set $E$, the false-positive $c_+$ and false-negative cost $c_-$, the goal of the duplicate classifier is to minimize the total misclassification cost:

$$\begin{aligned} & c_+ \times |\{e_i^j \in E \mid \tilde{z}(e_i^j) = 0 \bigcap z(e_i^j) = 1\}| \\ & + c_- \times |\{e_i^j \in E \mid \tilde{z}(e_i^j) = 1 \bigcap z(e_i^j) = 0\}| \end{aligned} \quad (1)$$

where $|\bullet|$ represents for the number of reports in a set. $z$ and $\tilde{z}$ represent the classification and ground truth label of an event

respectively (where the label 1 represents the report is original and 0 otherwise).

### B. Notations

To make the analytical model and solution clear, we summarize the notations we used in Table I.

TABLE I. NOTATIONS

| Symbol | Definition |
|---|---|
| N | The number of categories |
| $n_i$ | The number of reports with ith category |
| $e_i^j$ | The jth report in ith category |
| $t$ | The creation time |
| $s$ | The spatial attribute |
| $z$ | The classification label |
| $\tilde{z}$ | The ground truth label |
| c- | False-positive cost |
| c+ | False-negative cost |

### C. Model Specification

In this subsection, we describe our assumptions and the resulting probabilistic model. For convenience, let's denote the *minimum squared distance* of a report as its squared nearest-neighbour pairwise Euclidian distance with other relevant reports: $x(e_i^j) = \min_{e_i^k \in R(e_i^j)} \| s(e_i^j) - s(e_i^k) \|^2$

Our probabilistic model is built based on two modeling assumptions:

- **Discriminative spatial distribution:** we assume for reports in the same category, the minimum squared distance follows exponential distributions with different parameters:

$$p_i[x(e_i^j) \mid z(e_i^j) = k] = p_{\exp}[x(e_i^j), \lambda_{i,k}] = \lambda_{i,k} \exp[-\lambda_{i,k} x(e_i^j)], k = 0,1$$

- **Category-dependent Prior Probability:** we assume constant prior probability for reports in category *i* to be labeled as duplicate $p_{i,1} = p_i(z = 1)$. Naturally, the original prior probability would be $p_{i,0} = 1 - p_{i,1}$

Under such model, the conditional probability that a report is duplicate or original given the minimum squared distance $x(e_i^j)$ can be computed using the Bayes' theorem:

$$p_i[z(e_i^j) = k \mid x(e_i^j)] = \frac{p_{i,k} \times p_i[x(e_i^j) \mid z(e_i^j) = k]}{\sum_{k=0,1} p_{i,k} \times p_i[x(e_i^j) \mid z(e_i^j) = k]} \quad (2)$$

If the parameter $\lambda$ and $p$ are known, an arbitrary report's conditional probability of being original or duplicate could be easily calculated from Equation (2) and used for classification. Thus, the remaining question is how to infer the parameters. For the labelled data set, i.e., $\tilde{z}$ is known, the maximum likelihood estimates of parameters could be derived as:

$$p_{i,k} = \frac{|e_i^j : \tilde{z}(e_i^j) = k, e_i^j \in E|}{\sum_{k=0,1} |e_i^j : \tilde{z}(e_i^j) = k, e_i^j \in E|}, k = 0,1$$

$$\lambda_{i,k} = \{\frac{\sum_{\tilde{z}=k} x(e_i^j)}{|e_i^j : \tilde{z}(e_i^j) = k, e_i^j \in E|}\}^{-1}, k = 0,1 \quad (3)$$

where $|\bullet|$ represents the count of reports in the set.

However, as we mentioned before, obtaining the labelled datasets is a time consuming and expensive process. To reduce the need for labelled data, we develop an Expectation Maximization (EM) scheme to accurately estimate the required parameters from *unlabelled* data.

### D. Expectation Maximization

To infer the hidden parameters $\Theta$ from unlabelled data, we develop a new Expectation Maximization (EM) algorithm. The EM algorithm is an iterative method to obtain the maximum likelihood estimates of parameters in models which depend on some hidden/latent variables [9-10]. In general, it guesses the hidden variables, and then re-estimates the parameters as if the hidden variables are fully observed.

To fit our problem into the EM framework, we should identify the set of latent variable Z and a vector of unknown parameters $\Theta$ given the observed data set $X$, and then formulate the likelihood function as $L(\Theta; X, Z) = p(X, Z \mid \Theta)$. EM algorithm could then be applied to find out the maximum likelihood estimate of the unknown parameters $\Theta$:

$$L(\Theta; X) = p(X \mid \Theta) = \sum_Z p(X, Z \mid \Theta)$$

$$\Theta = \arg\max_\Theta L(\Theta; X)$$

For our duplicate detection problem, let us take a single category *i* as an example and the entire process could be similarly repeated for all other categories. In such case, the observed data set $X$ represents the set of corresponding minimum squared distances $X = \{x(e_i^j) \mid e_i^j \in E_i\}$. The binary label of a report is the hidden variable $Z = \{z(e_i^j) \mid e_i^j \in E_i\}$. The parameter is $\theta_i = [\lambda_{i,0}, \lambda_{i,1}, p_{i,0}, p_{i,1}]$. We could then formulate the likelihood function $L$ as:

$$L(\theta_i; X) = \prod_{j=1,Ni} \{\sum_{k=0,1} p_{i,k} \times p_i[x(e_i^j) \mid z(e_i^j) = k, \theta_i]\} = \prod_{j=1,Ni} \{\sum_{k=0,1} p_{i,k} \times p_{\exp}[x(e_i^j), \lambda_{i,k}]\} \quad (4)$$

The EM algorithm iterates between two key steps (i.e., E-step and M-step) until the estimation converges, where the converged values of parameters are valid at the local maxima of $L(\theta_i; X)$:

- **E-step (Expectation):** Compute the conditional class probability of observation $x(e_i^j)$ given current estimated parameter $\theta_i^t$ as follows:

$$p_i^t[z(e_i^j)=k \mid x(e_i^j),\theta_i^t]=\frac{p_{i,k}^t \times p_{\exp}[x(e_i^j),\lambda_{i,k}^t]}{\sum_{m=0,1} p_{i,m}^t \times p_{\exp}[x(e_i^j),\lambda_{i,m}^t]} \quad (5)$$

- **M-step (Maximization)**: Find the parameters that maximize the expected likelihood function $L$, and use them as the estimation parameters for the next iteration.

$$\theta_i^{t+1}=\arg\max_{\theta} L(\theta \mid \theta_i^t)$$

The M-step is derived as:

$$\lambda_{i,k}^{t+1}=\{\frac{\sum_{j=1,Ni} p_i^t[z(e_i^j)=k \mid x(e_i^j),\theta_i^t]\times x(e_i^j)}{\sum_{j=1,Ni} p_i^t[z(e_i^j)=k \mid x(e_i^j),\theta_i^t]}\}^{-1}$$

$$p_{i,k}^{t+1}=\frac{1}{N_i}\sum_{j=1,Ni} p_{i,k}^t[z(e_i^j)=k \mid x(e_i^j),\theta_i^t]$$

$$\theta_i^{t+1}=[p_{i,0}^{t+1},p_{i,1}^{t+1},\lambda_{i,0}^{t+1},\lambda_{i,1}^{t+1}] \quad (6)$$

The final EM algorithm is summarized in Algorithm 1.

**Algorithm 1** EM Algorithm

**Required Input:** The set of reports $E_i$ for category $i$

1: Initialize the parameters $\theta_i^0$
2: **while** parameters $\theta_i$ do not converge **do**
3: **for** $i = 1 : N_i$ **do**
4: Compute $x(e_i^j)$ according to its definition
5: Compute $p_i^t[z(e_i^j)=k \mid x(e_i^j),\theta_i^t]$ based on (5).
6: Update $\theta=\theta_i^{t+1}$ based on Eq. (6)
7: **end for**
8: **end while**
9: Return computed MLE estimates of hidden parameters $\theta_i$

Note that the estimated parameters are for the specific category $i$ only, thus computing the whole set of $\Theta=[\theta_1,...,\theta_N]$ needs N runs of the above EM algorithm. After $\Theta$ converge, the estimated parameters are passed to the cost-sensitive classifier to classify the reports.

### E. Cost-Sensitive Classification

After obtaining the maximum likelihood estimates of the estimation parameter $\Theta$, the remaining task is to build the cost-sensitive classifier. The conditional misclassification cost associated with one report would be computed as:

$$c_-[e_i^j \mid x(e_i^j),\theta_i]=\mathrm{c}_-\times p_i[z(e_i^j)=0 \mid x(e_i^j),\theta_i]$$
$$c_+[e_i^j \mid x(e_i^j),\theta_i]=\mathrm{c}_+\times p_i[z(e_i^j)=1 \mid x(e_i^j),\theta_i] \quad (7)$$

where the false positive and false negative cost per report is denoted as $c_+$ and $c_-$ respectively.

The Bayes-optimal classifier that minimizes the expected cost would simply classify the incoming report based on:

$$z(e_i^j)=\begin{cases}0 & c_-[e_i^j \mid x(e_i^j),\theta_i]>c_+[e_i^j \mid x(e_i^j),\theta_i]\\ 1 & c_-[e_i^j \mid x(e_i^j),\theta_i]<c_+[e_i^j \mid x(e_i^j),\theta_i]\end{cases} \quad (8)$$

In fact, when the inequality becomes equality, both decisions will be optimal.

Finally, we note that given inferred $\theta_i$, the computational cost of labelling a report only depends on the procedure of obtaining $x(e_i^j)$, which requires finding the minimum value of the squared pairwise distance between the labelled report and all other reports in its *relevant sets*. Therefore, the *complexity* of our cost-sensitive classifier to classify a report only depends on the size of its relevant report set: $O(|R(e_i^j)|)$.

## IV. Evaluation

In this section, we evaluate our EM based classification scheme on both synthetic and real-world datasets. The evaluation results show that our approach achieves the highest duplicate report detection accuracy compared to the state-of-the-art baselines.

### A. Synthetic Scenario

In the synthetic scenario, we generate the ground truth information and labels in simulation and evaluate the performance of different schemes under different parameter settings. To simulate the real-world situation in Chicago, we simulated more than 300 data points on a unit square area and generate different categories of issue reports to test our scheme.

Synthetic datasets are generated according to a clustered spatio-temporal point process considering both global and local clustering effects as follows:

- **Global process:** the center of reports, denote by $E^c$, follow a uniform Poisson point process with a constant intensity function $\lambda(t,l_1,l_2)$.
- **Local additive process:** for one center $e^c$, a cluster of more than one report $e_i^j$ is independently generated based on an additive Gaussian point process. The numbers of duplicate reports in such cluster follow a Poisson distribution with parameter $\lambda_{dup}=p_{i,0}/p_{i,1}$.

  For every report inside the cluster, the temporal indexes are distributed as follows:

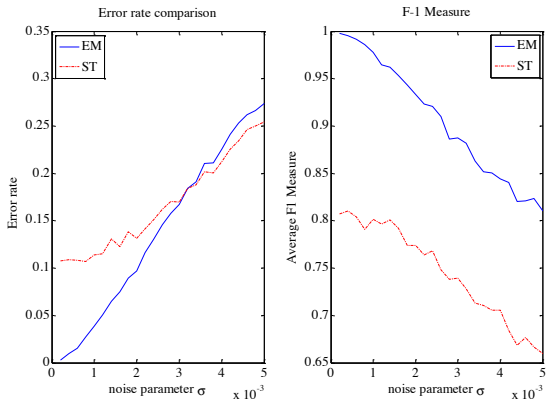
  $$s(e_i^j)\sim Gaussian[s(e^c),\sigma^2]$$
  $$t(e_i^j)\sim Uniform[t(e^c),t(e^c)+\Delta T]$$

  $\lambda_{dup}$ is fixed at 1. To simulate different categories, the spatial noise parameter σ varies from 0 to $5*10^{-3}$.

We compared the performance our scheme with the *ST-DBSCAN* algorithm, which is one of the state-of-the-art algorithms for clustering spatial-temporal data [22]. It requires

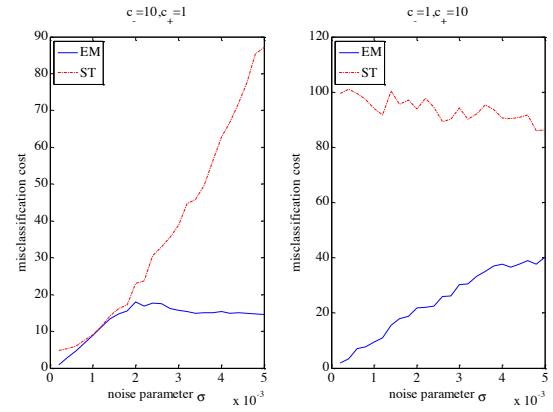
several ad-hoc input parameters, such as maximum spatial and temporal distance *Eps1 Eps2*, and the minimum number of point *minPts* in the aforementioned threshold. In our experiment, we set *minPts = 2*, $Eps1 = 3\sigma$, and $Eps2 = \Delta T$.

We evaluate the performance of our approach (EM-based classification) and the ST-DBSCAN approach based on the average error rate and F-1 measure that were computed over 100 experiments. To evaluate the performance of two schemes under *cost-sensitive* scenarios, we also implement two cost models: $c_- = 10c_+$ and $c_- = 0.1c_+$, and compare the total misclassification cost of two schemes using the two models.




(a) Error Rate (b) F-1 Measure

Fig. 1 Error rate and F-1 Measure Comparison Between EM-based Classification and ST-DBSCAN.




(a) $c_- = 10c_+$ (b) $c_- = 0.1c_+$,

Fig. 2 Total Misclassification Cost Comparison Between EM-based Classification and ST-DBSCAN

The results are shown in Figure 1 and Figure 2. We observe that our scheme clearly outperforms the ST-DBSCAN in F-1 measure and achieves better error rate with small noise parameter σ. We also observe that our scheme has lower total misclassification cost under different noise parameter σ. The results are encouraging considering that fact that our EM scheme is fully *unsupervised* while ST-DBSCAN is a *semi-supervised* approach.

## B. *Real-world Scenario*

For the real-world scenarios, we used collections of data extracted from the *SeeClickFix* Chicago tracking platform [6]. All the 34690 reports were submitted between February 2013 and 2014. Duplicate reports in the SeeClickFix Chicago dataset have been manually labeled, which provide us the *ground truth* for the evaluation. We choose 3 categories of reports with higher duplicate ratio to investigate: *Traffic Signal, Potholes, and Street Light All Out.* For simplicity, we assume false positive and false negative costs are equal and focus on classification accuracy evaluation in this section.

*Baselines:* We compared our method to three baseline methods. The first one: *CD* is a community-detection based scheme proposed in [8]. In this scheme, each pair of reports that are in the same category and within an ad-hoc spatial-temporal distance are linked by an undirected unweighted edge. The second one: *EC* is a naïve yet well-performed approach by simply taking an incoming report with the same text description of the address as a previous "open" report as a duplicate. The third one: *ST* is the ST-DBSCAN algorithm as mentioned before [8]. Note that both *CD* and *ST* schemes are semi-supervised approaches that require sufficient training data to turn their model parameters.

*Evaluation Metric:* Accuracy, F-1 Measure, Precision, and Recall Rate.

*Evaluation Result:* Table II-IV show the comparison results of evaluation metrics for the three categories of reported issues (i.e., street lights, traffic signal and potholes). We observed that the EM-based classification approach achieves overall good performance compared to the baselines: it achieves the *highest accuracy score* in all three test categories and the *highest F-1 measure* in the street light and traffic signal category and the second highest F-1 score in the potholes category. The EC scheme has similar or even better precision results in some categories yet *far worse recall rate* compared to our algorithm because it is too selective in labeling the duplicates. Other two approaches (CD and ST) both require some input parameter, which are usually obtained by manually labeling the training dataset. However, our EM scheme requires no training data and its performance is generally robust and accurate. These results validate the performance gains achieve by our new EM-based classification scheme in solving the duplicate report detection problem in a real world smart urban crowdsensing scenario.

TABLE II. STREET LIGHT ALL OUT

| | CD | EC | ST | EM |
|---|---|---|---|---|
| Accuracy | 48.5% | 60.70% | 67.25% | **83.47%** |
| F-1 Score | 62.31% | 51.54% | 68.55% | **79.90%** |
| Precision | 52.93% | **85.66%** | 75.24% | 79.20% |
| Recall | 75.75% | 36.86% | 62.96% | **80.60%** |

TABLE III. TRAFFIC SIGNAL MALFUNCTION

| | CD | EC | ST | EM |
|---|---|---|---|---|
| Accuracy | 32.99% | 89.10% | 76.06% | **89.10%** |
| F-1 Score | 46.12% | 81.11% | 61.60% | **81.11%** |
| Precision | 30.98% | 89.66% | 62.82% | **89.66%** |
| Recall | 90.24% | 74.05% | 60.43% | **74.05%** |

TABLE IV. POTHOLES UNFILLED

| | CD | EC | ST | EM |
|---|---|---|---|---|
| Accuracy | 50.76% | 59.51% | 53.62% | **65.65 %** |
| F-1 Score | 65.90% | 37.83% | **69.34%** | 67.48% |
| Precision | 53.21% | **91.20%** | 53.31% | 66.91% |
| Recall | 86.54% | 23.86% | **99.13%** | 68.05% |

Finally, we also study the convergence property of the EM-based classification scheme. The results of the street light case are shown in Figure 3. We observe that the EM-based classification scheme converges very quickly: the estimation on the likelihood function, classification parameters and duplicate ratio stabilizes within ten iterations. The results for other two cases (i.e., traffic signal and potholes) are quite similar and we do not present them here due to the space limit.

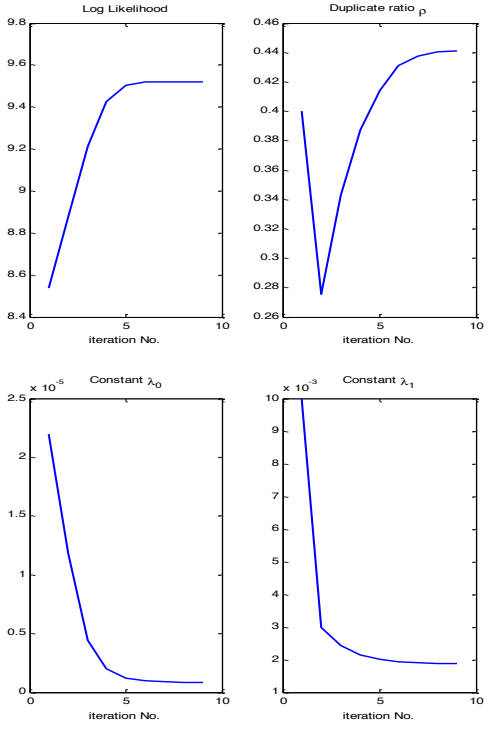



Fig. 3 The iterative convergence procedure for the EM-based classification scheme for Street Light Cases.

## V. CONCLUSIONS

This paper presents a new analytical approach to solve the duplicate report detection problem in crowdsensing based urban issue reporting systems. The proposed solution is a fully unsupervised binary classification approach developed based on the Expectation Maximization framework. We evaluated our new solution through both synthetic and real world datasets collected from smart city applications. The results showed that our approach significantly improves the duplicate report detection accuracy compared to the state-of-the-art baselines. The results of this paper can directly contribute to building more cost-efficient and responsive urban crowdsensing and monitoring systems for future smart cities.

## ACKNOWLEDGMENT

This material is based upon work supported by the National Science Foundation under Grant No. IIS-1447795.

## REFERENCES

[1] D. Estrin, "Participatory sensing: applications and architecture [internet predictions]," *Internet Computing, IEEE,* vol. 14, 2010, pp. 12-42.

[2] R. K. Rana, C. T. Chou, S. S. Kanhere, N. Bulusu, and W. Hu, "Ear-phone: an end-to-end participatory urban noise mapping system," in *Proceedings of the 9th ACM/IEEE International Conference on Information Processing in Sensor Networks*, 2010, pp. 105-116.

[3] M. Faulkner, M. Olson, R. Chandy, J. Krause, K. M. Chandy, and A. Krause, "The next big one: Detecting earthquakes and other rare events from community-based sensors," in *Information Processing in Sensor Networks (IPSN), 2011 10th International Conference on*, 2011, pp. 13-24.

[4] R. K. Ganti, N. Pham, H. Ahmadi, S. Nangia, and T. F. Abdelzaher, "GreenGPS: a participatory sensing fuel-efficient maps application," in *Proceedings of the 8th international conference on Mobile systems, applications, and services*, 2010, pp. 151-164.

[5] S. F. King and P. Brown, "Fix my street or else: using the internet to voice local public service concerns," in *Proceedings of the 1st international conference on Theory and practice of electronic governance*, 2007, pp. 72-80.

[6] I. Mergel, "Distributed democracy: Seeclickfix. com for crowdsourced issue reporting," *Com for Crowdsourced Issue Reporting (January 27, 2012),* 2012.

[7] Motorola. (2003). *The City of Chicago's 311 Systems*. Available: http://www.motorolasolutions.com/content/dam/msi/docs/business/products/software_and_applications/public_sector_applications/customer_service_request/_documents/static_files/chicago_fct_sht_final.pdf

[8] M. Budde, J. De Melo Borges, S. Tomov, T. Riedel, and M. Beigl, "Leveraging spatio-temporal clustering for participatory urban infrastructure monitoring," in *Proceedings of the First International Conference on IoT in Urban Space*, 2014, pp. 32-37.

[9] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the royal statistical society. Series B (methodological),* pp. 1-38, 1977.

[10] R. O. Duda, P. E. Hart, and D. G. Stork, "Pattern classification and scene analysis 2nd ed," ed: Wiley Interscience, 1995.

[11] M. Srivastava, T. Abdelzaher, and B. Szymanski. "Human-centric sensing." Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences 370.1958 (2012)

[12] D. Wang, T. Abdelzaher, and L. Kaplan. "*Social Sensing: Building Reliable Systems on Unreliable Data*." Morgan Kaufmann, 2015.

[13] D. Wang, T. Abdelzaher, B. Priyantha, J. Liu, and F. Zhao. "Energy-optimal Batching periods for asynchronous multistage data processing on sensor nodes: foundations and an mPlatform case study." *Real-Time Systems* 48, no. 2 (2012): 135-165.

[14] D. Wang, H. Ahmadi, T. Abdelzaher, H. Chenji, R. Stoleru, & C. Aggarwal, "Optimizing quality-of-information in cost-sensitive sensor data fusion." *Distributed Computing in Sensor Systems and Workshops (DCOSS), 2011 International Conference on*. IEEE, 2011.

[15] D. Wang, Q. Kun, and L. Wang. "Design of DBA algorithm in EPON upstream channel in support of SLA." *JOURNAL-CHINA INSTITUTE OF COMMUNICATIONS* 26.6 (2005): 87.

[16] J. Wang, D. WANG, K. TIMO, and Y. ZHAO. "A Novel Anti-Collision Protocol in Multiple Readers RFID Sensor Networks [J]." *Chinese Journal of Sensors and Actuators* 8 (2008): 026.

[17] I. Boutsis, and V. Kalogeraki. "Privacy preservation for participatory sensing data." Pervasive Computing and Communications (PerCom), 2013 IEEE International Conference on. IEEE, 2013.

[18] R. Sasank, et al. "Examining micro-payments for participatory sensing data collections." Proceedings of the 12th ACM international conference on Ubiquitous computing. ACM, 2010.

[19] Y. Zha, and M. Veloso. "Profiling and Prediction of Non-Emergency Calls in NYC." Workshops at the Twenty-Eighth AAAI Conference on Artificial Intelligence. 2014.

[20] Y. Zheng, et al. "Diagnosing New York City's noises with ubiquitous data." Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing. ACM, 2014.

[21] T. Winkler, Z. Holger, and M. Weinberg. "Municipal benefits of participatory urban sensing: A simulation approach and case validation." Journal of theoretical and applied electronic commerce research, 2014, pp. 101-120.

[22] D. Birant, and A. Kut. "ST-DBSCAN: An algorithm for clustering spatial–temporal data." Data & Knowledge Engineering 60.1 (2007): 208-221.