Will you accept my job? A new approach towards predicting human participation in mobile crowdsensing

Tanveer Ahmed* and Abhishek Srivastava

Indian Institute of Technology Indore, India
Email: phd12120101@iiti.ac.in
Email: asrivastava@iiti.ac.in
*Corresponding author

Abstract: The exponential growth of wireless devices have paved the way for several new and innovative mobile driven paradigms. One area that has started to receive attention in literature is mobile crowdsensing. Though promising, one specific aspect of mobile crowdsensing that has been mostly ignored in literature is human participation. In this paper, we take on this issue and present a novel probabilistic approach to predict whether we can expect a response from the crowd or not. The proposed candidate selection algorithm takes its inspiration from statistics and handles the typical uncertainties in human behavior. Validation of the proposed framework is done in two ways: 1) we implement a prototype and deploy it over an enterprise service bus; 2) we perform numerical testing on real datasets. With this experimental testbed, we show the efficacy of the framework in actual deployment scenarios.

Keywords: mobile computing; human participation prediction; mobile crowdsensing.

Reference to this paper should be made as follows: Ahmed, T. and Srivastava, A. (2018) 'Will you accept my job? A new approach towards predicting human participation in mobile crowdsensing', *Int. J. Social and Humanistic Computing*, Vol. 3, No. 1, pp.1–19.

Biographical notes: Tanveer Ahmed is a PhD student in Indian Institute of Technology Indore. He is working in the area of mobile crowdsensing. His areas of interest are mobile computing, mobile crowdsensing and web services. He also works in the areas of cognitive psychology and uncertainty quantification.

Abhishek Srivastava is an Associate Professor and the Dean of Student Affairs in Indian Institute of Technology Indore. He holds a doctorate from University of Alberta, Canada. Prior to his career in Indian Institute of Technology Indore, he was an Assistant Professor in Rose-Hulman Institute of Technology, USA.

1 Introduction

The recent advances in mobile computing has created an insatiable environment, compelling business organisations to focus more on mobile oriented application development. This competitive environment present numerous possibilities to explore

previously unthinkable research dimensions and take advantage of the opportunities that have presented themselves. Realising the potential behind this vision, several new paradigms have emerged in practice as well as in academic literature, for instance internet of things (Tan and Wang, 2010), mobile vehicular technologies (Gerla and Kleinrock, 2011), mobile healthcare (Li et al., 2004) to name a few. Following this trend, one such paradigm that has given equal freedom to industry and academia to explore new research venues is mobile crowdsensing. Designed as a sensing framework to empower individuals with the capability of consuming and providing ubiquitous mobile services, the paradigm allows normal people to transform raw data into useful information through their companioned devices (Guo et al., 2015). This information allows society to achieve several objectives, thereby taking one more step towards achieving the goal of providing a much better and a much smarter future, for instance pollution monitoring (Kim et al., 2011), traffic update (Pan et al., 2013), crime reporting (Geoffrey and Schectman, 2013). Having a foundation in mobile services, the notion of crowdsensing has been refined by literature into personal and societal sensing (Ganti et al., 2011). In personal sensing, the underlying mechanisms directly influence the daily life of a person, for instance diet management for people with diabetes (Holtz and Lauckner, 2012), sensing one's routine (Ranvier et al., 2015), etc., whereas, community sensing pertains to sensing activities involved in achieving societal objectives, for example finding the lost child (Liu and Li, 2014). Furthermore, literature has also subcategorised these activities into two different dimensions. According to Ganti et al. (2011), crowdsensing is further sub-classified into two categories, participatory sensing and opportunistic sensing. In participatory sensing, a human being explicitly and deliberately provides sensing information, whereas in opportunistic sensing a human being automatically and unknowingly acts a sensing apparatus. Although the notion of crowdsensing has been refined into multiple branches, but the idea that a human being can turn into a data source providing real-time sensing information, is indeed one of the most lucrative points that has led to significant research in this domain (Ra et al., 2012; Hu et al., 2013; Agarwal et al., 2013). Though attractive and promising, but, the work in this area is still in its infancy and require significant efforts on several fronts. According to Srivastava et al. (2012), the main area of research for any crowdsensing system are battery consumption, context inference, participant recruitment, privacy, data quality.

Taking a step back and focusing on these issues, there is a growing body of work that has tried to pursue some of the research directions. In this regard, literature has recognised the importance of the technical factors of a mobile device, and correspondingly there is plethora of work that addresses a few of these challenges. For instance, literature has given importance to task segregation and outsourcing the discrete process steps (of a workflow) to the mobile device of a person (Ra et al., 2012; Hu et al., 2013). Further, there also exist a vast body of work dealing with mobile-based context inference and prediction (Gomes et al., 2013), context aware preference management (Krause et al., 2006), preserving the battery power of a mobile device (Xiong et al., 2015) and so many others. However, to the very heart of the paradigm there lies a problem that, to our surprise, has been left untouched. We believe the core to any area where the '*crowd*' is involved is the *human factor*. Though, literature has very well appreciated and accepted the issues related to the technical capabilities of a mobile device, but it has turned a blind eye towards the main contributer of information: *the person that owns the device*. We argue, ignoring the crowd in crowdsensing is not only infeasible, but it is also a slippery slope. To support the argument, we present an example that is commonly

followed by literature (Liu and Li, 2014; Hu et al., 2013; Ra et al., 2012; Agarwal et al., 2013). Consider a requester wants to know – "what is crowd size at Wall Street in New York City today?". To aid the requester in his/her query, the server deployed on the cloud select some users who are currently available at Wall Street and outsources the task to their mobile device. If the worker complies with the request, he/she will get suitable reward. With respect to this recurring example, we ask a question: is the person surely going to reply to the request no matter what? Will incentives be enough to make a person comply to one's request? Obviously, the answer to both the two questions is: no. In contrast to the similar sensing paradigms, e.g., internet of things, where the producer of the information is a machine, crowdsensing has to deal with human beings and their erratic behaviour. Therefore, considering human ideology, can we expect a response from the crowd to the posted requests all the time? This basic issue, in a paradigm where the main contributors are human beings, is one of the shortcomings that literature has failed to address so far. Hence, the wise choice in such a mixture of the physical and the digital world is to approach the technical capabilities of a device and the issue of human participation simultaneously.

In light of this issue, the objective of this paper is to employ principles of probability to handle the problem of human participation prediction. To that end, we propose a novel probabilistic method to recruit a human from the crowd. We show the step-by-step derivation and present the necessary details of the method to validate its theoretical feasibility. We use the basic principles of data engineering to show that the proposed method is efficiently able to handle uncertain human participation habits. To demonstrate the viability of the proposed framework in actual deployment, we engineer a prototype and use real world datasets provided by stackoverflow. We deploy the prototype on an enterprise service bus, thereby demonstrating the feasibility of the proposed method in current cloud-based computational environments. Through testing performed with this experimental setting, we show the practical implication of our approach.

Throughout this paper, mobile crowdsensing is often referred to as crowdsensing.

2 Related work

Recently, the paradigm of crowdsensing has generated a lot of interest in literature. Ra et al. (2012) is one the most mature platforms that provides a programming framework for crowdsensing. It also allows breaking a complex task into smaller tasks, thereby reducing the complexities in crowdsensing tasks. Vita (Hu et al., 2013) is also a cloud-based proposal that tries to outsource tasks to handheld devices via service oriented architecture. Agarwal et al. (2013) focuses a utility driven framework for society-based crowdsensing exercises. Xiao et al. (2013) tries to handle scalability by instantiating a virtual machine for every mobile device participating in crowdsensing exercises. Further, there is also the application oriented line of research, where work has targeted specific applications of mobile crowdsensing. For example, the work presented in Liu and Li (2014) focuses on finding the lost child via community-based crowdsensing. Geoffrey and Schectman (2013) uses mobile crowdsensing to gather dynamic sensing feeds during the Boston bombing. Work in Kim et al. (2011) uses citizen science for reporting pollution in an area. Although, the work discussed here is limited, but, an exhaustive review of crowdsensing techniques is available in Guo et al. (2015). In this paper, we did

not find any method that uses a specific combination of statistics and technology to handle the issue highlighted in this paper. That being said, there are, however, a few method that focus on *trust-based* candidate selection (Amintoosi and Kanhere, 2014, 2013; Amintoosi et al., 2015). However, these models do not dig down to the details of uncertainty in human behaviour. They use 'minimum variance unbiased' methods. As will be discussed in the following section, this method is not feasible and falls short on several instances. On the other hand, we go deep into statistics, and use a mathematical framework that incorporates typical uncertainty in human participation. Moreover, we show the mathematical derivation of the method. We proceed step-by-step with all the details for the problem. In doing so, we propose a general method that can be utilised by any crowd-based paradigm, therefore, we not only build upon existing work, but also try to complement research in literature.

As crowdsensing has an inherent dependency on mobile devices, therefore, there are several constraints, specially with a mobile device, that have to be addressed. In this regard, work has identified and focused on multiple constraints in candidate recruitment. For example, in Hassani et al. (2015), a context aware recruitment framework is proposed. In this work, the authors allocate the tasks to users who are familiar with the desired location. At much the same time, they also preserve the energy of the mobile device. Similarly, the work of Hachem et al. (2014) focuses on the criterion of crowdsensing task allocation under the coverage constraints. Further, Xiong et al. (2015) also focuses the same criterion by adding the energy constraint. There are also proposals focusing the criterion of financial constraints (Jaimes et al., 2014a, 2014b). To sum the work in constraint dependent crowdsensing, literature has indeed realised the importance of a few constraints. In some instances, there are even a few combinations of these constraints (Wang et al., 2016). But, work is not focused on handling the '*human participation constraint*'. That is, work does incorporate the willingness of the user to contribute. In this paper, we have focused on this criterion only. Therefore, the work builds upon the foundation laid by these papers, and complements the idea by handling the participation constraint using a novel statistical method.

3 Predicting human participation

In this section, we propose a method to recruit a candidate from the crowd. The method draws its inspiration from statistics. However, before getting into the details, we present the problem from existing literature's point of view. The issue of recruiting a volunteer from the crowd in mobile crowdsensing is: we have to select a person p from the crowd set C, so that the person is at the location l at time t, where $l \in L$ and $t \in T$. Here T, L are the time and location sets respectively. T can denote the hour of the day, the day of the week, or the month in a year. L is the set of all the locations a user has visited. This basic setting is followed in literature. Correspondingly, and according to the current standing of literature, the probability of selecting a person for a particular crowdsensing exercise is mathematically expressed as:

$$P_{p(i)}(sel) = P_{p(i)}(l) \times P_{p(i)}(t^l) \tag{1}$$

where, $P_{p(i)}(sel)$ represents the probability of selecting a person $p(i)$, $P_{p(i)}(l)$ is the probability that the person is at location l, and $P_{p(i)}(t^l)$ is the probability that the person is

at location l at time t. We modify this basic equation to accommodate the proposal that the probability of selecting a person is not only dependent on the person being at location l at time t, but the probability that the person is also willing to provide a response r. Therefore, the equation is rewritten as:

$$P_{p(i)}(sel) = P_{p(i)}(l) \times P_{p(i)}(t^l) \times P_{p(i)}(r) \tag{2}$$

where $P_{p(i)}(r)$ represents the probability of getting a response from the person $p(i)$. The problem of estimating the values of $P_{p(i)}(l)$, and $P_{p(i)}(t^l)$ has been thoroughly investigated in literature. Therefore, we will focus our attention to the problem of predicting the value of $P_{p(i)}(r)$ only.

3.1 Proposed method

To accomplish the objective of selecting a candidate, we employ prior information available to the system, and predict the most likely estimate of the posterior. To achieve this, we derive the probability of getting a response from the person. To explain the method, lets assume that the person has responded k times out of a total of N requests in the past.

We define a random variable $R(i)$ for the person $p(i)$ as:

$$R(i) = 1 \{\text{if the person responds to the } i^{\text{th}} \text{ request.}\} \tag{3}$$

And

$$T(N) = \sum_{j=1}^{N} R(i) \ \{\text{The total number of responses}\} \tag{4}$$

Lets assume that the probability that a person $p_{(i)}$ is willing to comply to a request r is π.

$$\Rightarrow P\big(R(i) = 1\big) = \pi;\ f(\pi) \tag{5}$$

where, $f(\pi)$ is the distribution function of the parameter π. Similarly, the probability that the person $p(i)$ is not willing to provide a response is expressed as:

$$P\big(R(i) = 0\big) = 1 - \pi \tag{6}$$

As crowdsensing applications are initiated by any number of requesters at any time. Therefore, we assume that each request is initiated independently. Therefore, the distribution of $T(N) = k$ follows binomial distribution. Mathematically, we get:

$$\Rightarrow P\big(T(N) = k\big) = \binom{N}{k} \pi^k (1-\pi)^{N-k} \tag{7}$$

In this formulation, our objective is to calculate the probability of getting a response at the $(N + 1)^{\text{th}}$ request, given a history of N requests (out of which k responses are obtained). In simple words, given a history of participation, we have to predict whether we can expect future participation from the same person or not.

According to Bayes theorem, posterior information is proportional to prior time likelihood. Therefore, using this property, we express the issue mathematically as:

$$P\big(R(N+1)=1\big|T(N)=k\big)=\frac{P\big(T(N)=k\big|R(N+1)=1\big)\times P\big(R(N+1)=1\big)}{P\big(T(N)=k\big)} \tag{8}$$

We know:

$$\Rightarrow P\big(T(N)=k\big|R(N+1)=1\big)=P\big(T(N)=k\big) \tag{9}$$

In this paper, we use the assumption that $T(N) = k$ and $R(N + 1)$ are independent given π (Sun et al., 2006). Therefore:

$$P\big(T(N)=k\big)=\int_0^1 P\big(T(N)=k|\pi\big)\,f(\pi)d\pi \tag{10}$$

We know that the probability that a person is going to comply to the next incoming request is π. Therefore, we have:

$$P\big(R(N+1)=1\big)=\pi \tag{11}$$

Substituting (9), (10), (11) in (8), and simplifying the equations, we get:

$$P\big(R(N+1)=1\big|T(N)=k\big)=\frac{\int_0^1 \pi P\big(T(X)=k|\pi\big)\,f(\pi)d\pi}{\int_0^1 P\big(T(X)=k|\pi\big)\,f(\pi)d\pi} \tag{12}$$

Substituting the expression from (7) in (12) and simplifying, we get:

$$P\big(R(N+1)=1\big|T(N)=k\big)=\frac{\int_0^1 \pi^{k+1}(1-\pi)^{N-k}\,f(\pi)d\pi}{\int_0^1 \pi^{k}(1-\pi)^{N-k}\,f(\pi)d\pi} \tag{13}$$

To solve the above equation, we require a particular distribution for the probability of a person responding to a request. In other words, we need a distribution function ($f(\pi)$) for the parameter π. As crowdsensing has to deal with human beings, therefore, there is always an element of uncertainty. Consequently, the probability of success at each trial is not fixed but random. As a result, an ideal candidate for this particular situation is conjugate beta prior density function. Thus, we use the conjugate beta prior to define the probability density function of π. This function is defined as:

$$f(\pi;\alpha,\beta)=\frac{\pi^{\alpha-1}(1-\pi)^{\beta-1}}{B(\alpha,\beta)} \tag{14}$$

where, α and β are the two parameters, $B(\alpha, \beta)$ is the beta function defined as:

$$B(\alpha,\beta)=\int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt \tag{15}$$

The mean and variance of the beta distribution are well known and are as follows:

$$E[\pi]=\frac{\alpha}{\alpha+\beta} \tag{16}$$

$$\text{var}(\pi) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \tag{17}$$

Representing equation (14) in terms of gamma function, we have:

$$f(\pi;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\left(\Gamma(\beta)\right)}\pi^{\alpha-1}(1-\pi)^{\beta-1} \tag{18}$$

Substituting the above expression from equation (18) in equation (13), we get:

$$P\left(R(N+1)=1\middle|T(N)=k\right) = \frac{\int_0^1 \pi^{k+1}(1-\pi)^{N-k}\,\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\left(\Gamma(\beta)\right)}\pi^{\alpha-1}(1-\pi)^{\beta-1}d\pi}{\int_0^1 \pi^{k}(1-\pi)^{N-k}\,\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\left(\Gamma(\beta)\right)}\pi^{\alpha-1}(1-\pi)^{\beta-1}d\pi} \tag{19}$$

Simplifying the above equation and manipulating the numerator and denominator, we get:

$$P\left(R(N+1)=1\middle|T(N)=k\right) = \frac{\int_0^1 \pi^{k+\alpha}(1-\pi)^{N-k+\beta-1}d\pi}{\int_0^1 \pi^{k+\alpha-1}(1-\pi)^{N-k+\beta-1}d\pi} \tag{20}$$

This equation has similarity to the beta function described above in equation (15). Therefore, rewriting the equation in terms of beta function, we have:

$$P\left(R(N+1)=1\middle|T(N)=k\right) = \frac{B(k+\alpha+1,\,N-k+\beta)}{B(k+\alpha,\,N-k+\beta)} \tag{21}$$

After observing the partial result, we go back to the assumption of equation (14), where we assumed the conjugate prior as a beta distribution. Using the property of conjugate priors, the posterior distribution of π is also beta distribution. As a result, we substitute the values of α and β as $\alpha + k$ and $\beta + N - k$ respectively (Box and Tiao, 2011). Therefore, using these values and by simplifying the fraction, we get:

$$P\left(R(N+1)=1\middle|T(N)=k\right) = \frac{B(2k+\alpha+1,\,2N-2k+\beta)}{B(2k+\alpha,\,2N-2k+\beta)} \tag{22}$$

The equation denotes the probability of a person responding to the $(N + 1)^{\text{th}}$ request, given a history of N requests. The importance of this equation is that it is much more precise and a lot more feasible than minimum variance unbiased estimation shown below:

$$\text{Unbiased probability, } P\left(R(N+1)\right) = \frac{k}{N} \tag{23}$$

The above equation is simple and does not include uncertainty. This uncertain factor is important because we are dealing with humans, and we can never accurately predict an individual's behaviour. In this paper, we work with this typical constraint.

3.2 Parameter estimation

The derivation presented in the previous subsection provides a method to estimate the probability of getting a response from the person. However, from equation (22), we see that the probability is dependent on two unknown parameters:

1 α

2 β.

Therefore, we need a data driven method to estimate their numerical values. Moreover, an additional issue is: how to select the values of these two parameters so that the resulting probability remains unaffected to reparametrisation. To understand this statement, if k and N denotes the number of responses and requests per day, then a particular value of α and β should not effect the probability when k and N denotes the number of responses and requests per hour. As a result, we have to formulate an estimate of these two parameters so that the result is insusceptible to reparametrisation. To address this issue, we look into statistical literature. In that matter, it is a known fact that Jeffrey's prior is invariant to the effect of reparametrisation (Clarke and Barron, 1994). Therefore, we will use Jeffrey's prior and will estimate the values of these variables. Jeffery's prior is defined as:

$$\phi_J(\pi) \propto \sqrt{I(\pi)} \tag{24}$$

where, $I(\pi)$ is the fisher's information, defined as:

$$I(\pi) = -E\left[\frac{d^2 \log p(K|\pi)}{d^2\pi}\right]$$

In our case, $k \sim binomial(N, \pi)$ and:

$$p(k|\pi) = \binom{N}{k}\pi^k(1-\pi)^{(n-k)}$$

Taking the log of above expression and differentiating twice, we get:

$$\log\left(p(k|\pi)\right) = \log\binom{N}{k} + k\log(\pi) + (N-k)\log(1-\pi)$$

$$\frac{d\log\left(p(k|\pi)\right)}{d\pi} = \frac{k}{\pi} - \frac{N-k}{(1-\pi)}$$

$$\frac{d^2\log\left(p(k|\pi)\right)}{d\pi^2} = -\frac{k}{\pi^2} - \frac{N-k}{(1-\pi)^2}$$

We know that the expected value ($E[K]$) of binomial distribution is $N\pi$, therefore, substituting k as $N\pi$ in above equation, and using Fisher's information, we get:

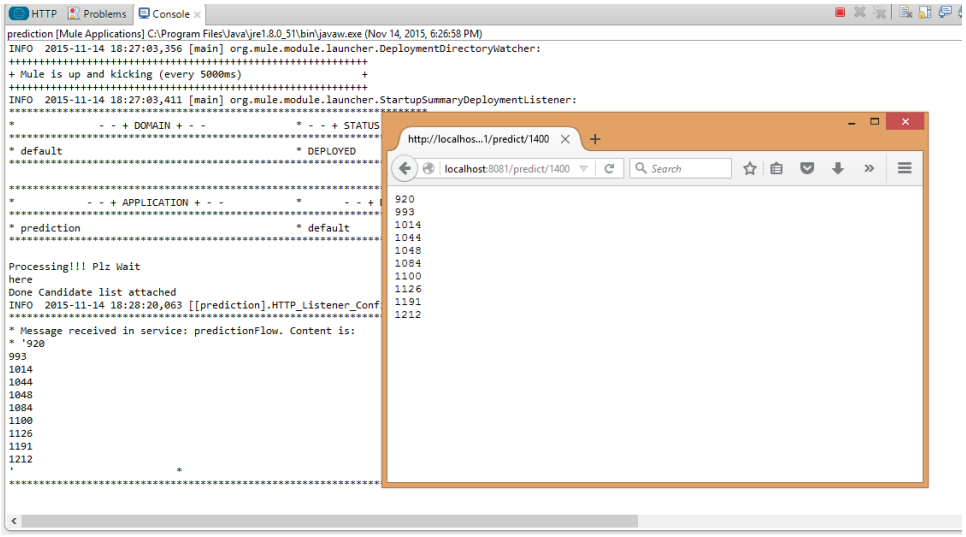
$$I(\pi) = -E\left[\frac{d^2 \log p(K|\pi)}{d^2\pi}\right]$$
$$= \frac{N\pi}{\pi^2} + \frac{N - N\pi}{(1-\pi)^2}$$
$$= \frac{N}{\pi} + \frac{N}{1-\pi}$$
$$= \frac{N}{\pi(1-\pi)}$$

where, N is a constant. Therefore, from equation (24) we have:

$$\phi_I(\pi) \propto \pi^{-1/2}(1-\pi)^{-1/2} \tag{25}$$

This expression follows a beta distribution $B(\alpha, \beta)$ with parameters 0.5 and 0.5 [for details see Box and Tiao (2011)]. Thus, for the proposed framework we choose these values. An advantage of using these particular values is that they represent *non-informative priors*, thereby following the principle of – '*let the data do the talking*'. This is important as we have an element of objectivity in the system.

Figure 1 A snapshot of the developed application deployed over MuleESB (see online version for colours)



3.3 Complexity analysis

The algorithm used in the paper to predict whether or not we can expect a response from the crowd is shown in Algorithm 1. The method terminates if it find suitable candidates whose probability of selection is greater than a certain threshold defined by the parameter λ. To analyse the time and space complexity, we will assume that the system has a set of

U registered users, where each user has a set of L locations, T time intervals, R responses for each time interval and location $l \in L$. To this end, the loops in the algorithms run for every user $i \in U$, at the location $l \in L$, time $t \in T$, and response $r \in R$. Using this method, the space complexity for the method is $O(|U| |L| |T| |R|)$. This is because the system has to store the data for every user providing a response at a particular time and location. However, for a particular crowdsensing application, specifically characterised by spatial and temporal requirements, the location, time, and the previous number of responses is constant. Therefore, they can be ignored. Hence, the search complexity reduces to $\Theta(|U|)$.

Algorithm 1 Algorithm for selecting a candidate from the crowd

for $j = 1$ to $|L|$ **do**

Compute probability $P_{p(i)}(l)$ that a person is at location l.

end for

for $j = 1$ to $|T|$ **do**

Compute probability $P_{p(i)}(t^l)$ that a person is at location l at time t.

end for

for $i = 1$ to $|U|$ **do**

Compute probability $P_{p(i)}(r)$ of getting a response from person i at location l and time t.

end for

for $i = 1$ to $|U|$ **do**

$$P_{selection(p_i)} = P_{p(i)}(l) \times P_{p(i)}(t^l) \times P_{p(i)}(r)$$

if $\left(P_{selection(p_i)} > \lambda\right)$ **then**

Select the person p_i

end if

end for

4 Results

4.1 Data collection and prototype development

To validate the viability of the proposed method in actual deployment, we have developed a prototype. The prototype deployed as a web-based application was implemented using Java, and is deployed over an enterprise service bus, MuleESB. We chose MuleESB for two reasons:

1 it is open source and freely available

2 by deploying the proposed framework on an ESB, we show the feasibility of the method in current cloud-based computational environments.

The application was developed via Anypoint Studio v5.3.0. The inbuilt server package was deployed on a machine with i7 processor, 8 GB ram, and 2.4 Ghz processing speed

with Windows 8 as the operating system. The application developed using RESTful principles provided a uniform method of accessing the information stored at the middleware. Thus, using this type of a methodology, we provided a universal strategy to invoke the application from any entity in the real world. A snapshot of the application deployed using this settings is shown in Figure 1. In this figure, a third party application requested the middleware to return a list of possible candidates who will be recruited for a crowdsensing exercise.

As crowdsensing lacks a mature platform, therefore, to numerically test the performance of the proposed recruitment algorithm, we experiment with datasets provided by crowdsourcing platforms. This is because similar to crowdsensing, crowdsourcing also relies upon the efforts of the volunteers to get the job done. Therefore, our motive of testing the participation habits is perfectly aligned with this mature platform. In this regard, we have collected data from public data repositories of stackoverflow. We assumed that a question posted by a user is analogous to a request by a requester, and the answer is analogous to a response provided by the worker. We have collected the details of 10,000 users for one year.

In Figure 4, we have shown the response time of the prototype when the software processed the data consisting of 10, 50, 100, 1,000 and 10,000 users respectively. In this test, the database had the details for the previously specified number of users. It is visible from the figure that when the number of users is high, the execution time is also high. This is expected as the software has to process data of several users simultaneously.

Figure 2 Flow diagram used by mule (see online version for colours)

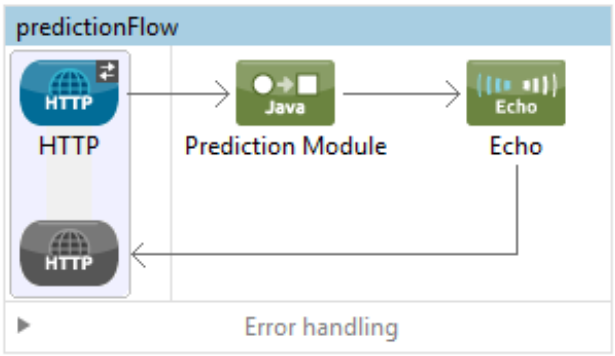


Figure 3 Memory, CPU, disk access of the prototype (see online version for colours)

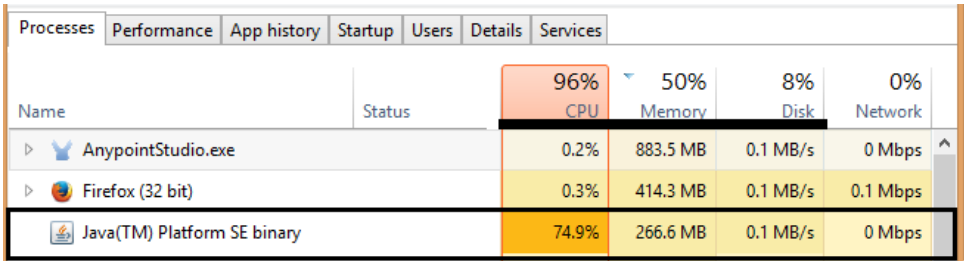
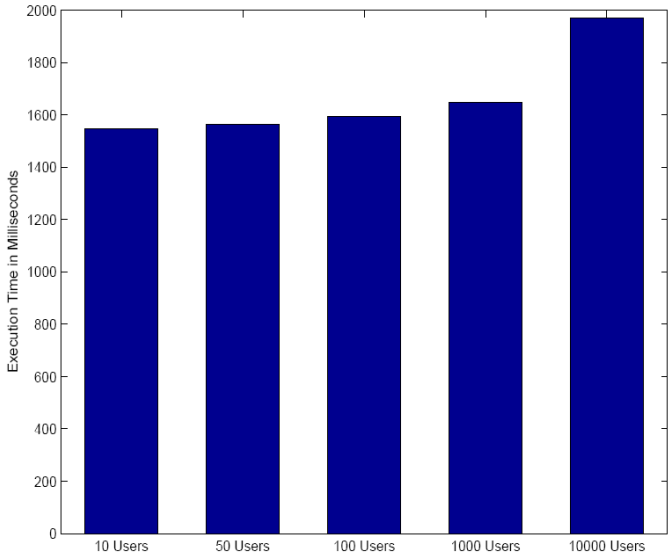


Figure 4 Response time of the prototype with different number of users (see online version for colours)

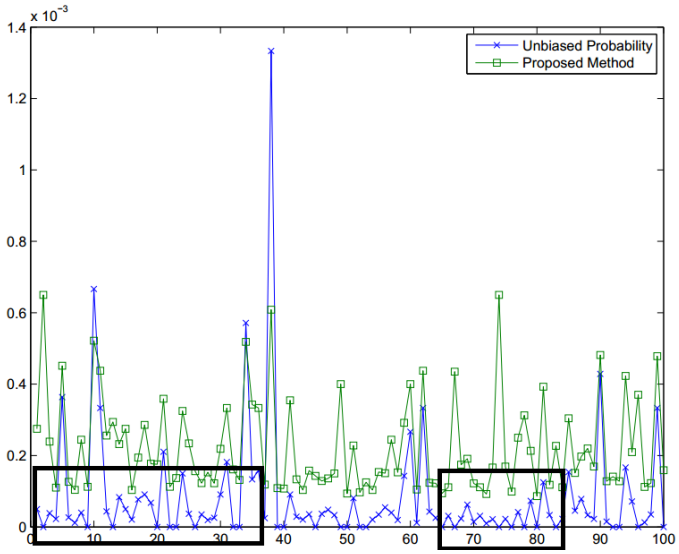


4.2 Comparison with unbiased probability

$$\text{Unbiased probability, } P\left(R(N+1)\right) = \frac{k}{N} \tag{26}$$

To compare the performance of the method with unbiased probability [equation (26)], we have shown the probability of getting a response for 100 days in Figure 5. The data to calculate the probability on the current day was taken from the previous day. In this figure, we have highlighted a few cases when the number of responses from a person is zero. In this scenario, since the person has not responded at all, therefore, the unbiased method is producing a zero numerical value ($k = 0$, $N \neq 0$). In other words, the system is certain that the person is never going to respond to any of the future requests. This is infeasible in practical situations, especially considering the case that we are dealing with a human crowd. With humans, the uncertainty factor is high, consequently, the participation at an exercise can change any time. We know that human behaviour is erratic and can go through several changes. Therefore, if we want to work with human beings, we need information incorporating typical human factors. In this regard, and in contrast to unbiased probability, the proposed method is producing a lower numerical value, i.e., the probability of getting a response is less. This is acceptable because if the users did not responded to any of the request, then we can say that the probability of such users participating in future crowdsensing exercises is also less, but, it is not zero.

Figure 5 Proposed method vs. unbiased probability (see online version for colours)



To further show the importance of the proposed method, consider the case of the cold start problem. By cold start problem, we imply that the user is new to the system, and has neither received nor responded to any of the requests. In that case (k = 0, N = 0), the unbiased probability [equation (26)], will produce $\frac{0}{0}$. In other words, the system is stuck in unstable state. This is problematic in real situations. In contrast, using the derivation shown in Section 3, the proposed model is not stuck at all. To understand this, we take the case of cold start and substitute the values of k, N as zero in equation (22). Using these values, we get:

$$P\left(R(N+1)=1 \middle| T(N)=k\right)=\frac{B(\alpha+1,\beta)}{B(\alpha,\beta)} \tag{27}$$

From the derivation shown in Section 3.2, we know that the values of α and β is 0.5. Therefore:

$$\begin{aligned} P\left(R(N+1)=1 \middle| T(N)=0\right) &= \frac{B(1.5,0.5)}{B(0.5,0.5)} \\ &= \frac{1.57}{3.14} \\ &= 0.5 \end{aligned}$$

Thus, the probability of getting a response in the case when the user is new to the system is 0.5. This is intuitively as well as practically more feasible. In other words, when a user is new to the system, then there is a 50% chance that he/she will comply to a request.

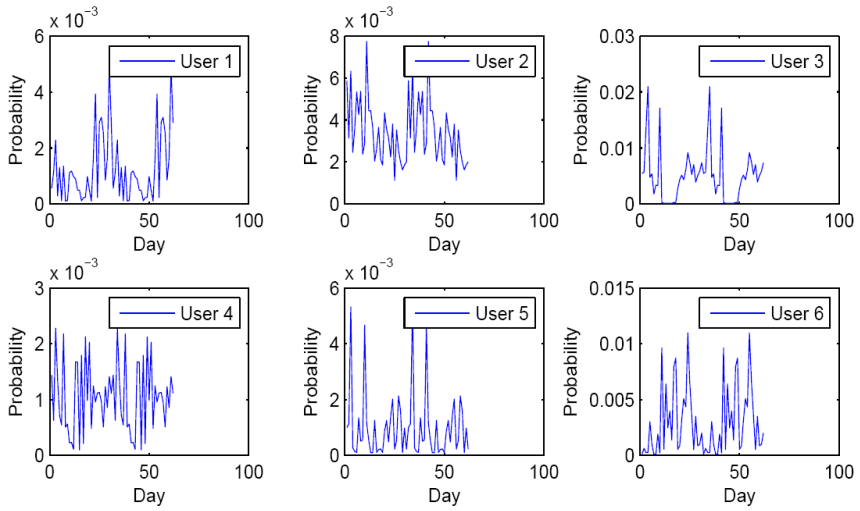
4.3 Predictive capability

To begin with the analysis on the predictive capability of the method, we have shown daily evolution of probability for a few users monitored for a continuous period of 60 days in Figure 6. It is visible from the figure that the probability for these users follow several ups and downs. This type of a pattern is expected as no user from the crowd is going to participate everyday with the same rigor. Owing to certain circumstances in a person's daily routine, these type of situations are expected. However, a pertinent question in this context is: with this erratic and constantly changing behaviour, what is the accuracy of the system. Therefore, in the next series of experiments we test the accuracy of the method. In our experiments, accuracy is defined as follows:

$$Accuracy = \frac{N_{resp}}{N_{recom}}$$

where, N_{resp} is the number of candidates who actually responded, and N_{recom} is the number of candidates who were recommended by the system.

Figure 6 Probability of getting a response for some users (see online version for colours)



To test the method in real scenarios, we have compared the performance with random user selection method and recency first method. By recency first method, we imply selecting a person who has recently provided a response. To begin with the test, we chose each day in the dataset, and calculated the probability of selection for the next day. Therefore, the method automatically selected a few candidates and recommended them to the client. With this type of testing methodology, the result for *each month* is shown in Figure 7. Further, the accuracy values averaged over the year is also presented in Figure 8. It is clear from the figures that recruiting candidates randomly is certainly not the best way forward. In this context, and according to the recommendation of literature (Cardone et al., 2013), selecting a person based on recency first method might seem a good option. However, from the results, the accuracy of the proposed method is much better than the accuracy of recency first method. To be precise, the accuracy of the

proposed method is ~42%, whereas the accuracy of recency first method is ~31%. Thus, the method showed good performance.

Figure 7 Average accuracy on a monthly basis (see online version for colours)

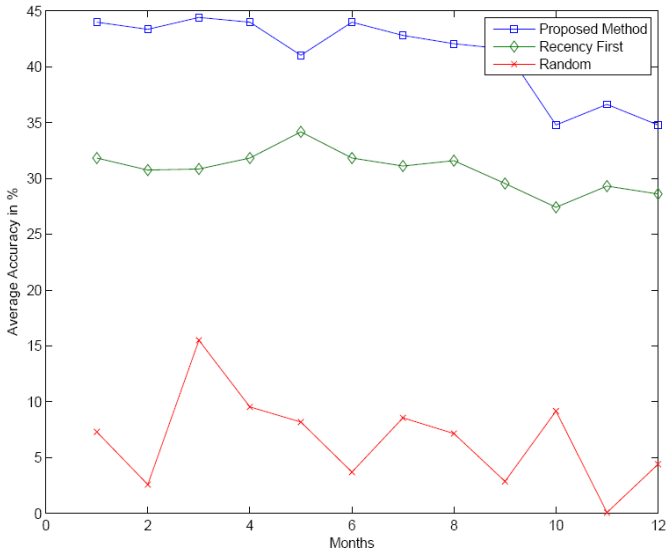
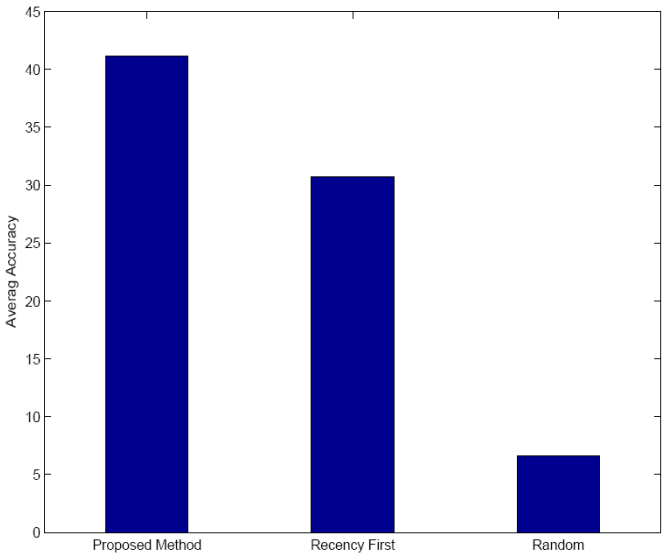


Figure 8 Accuracy averaged over an entire year (see online version for colours)



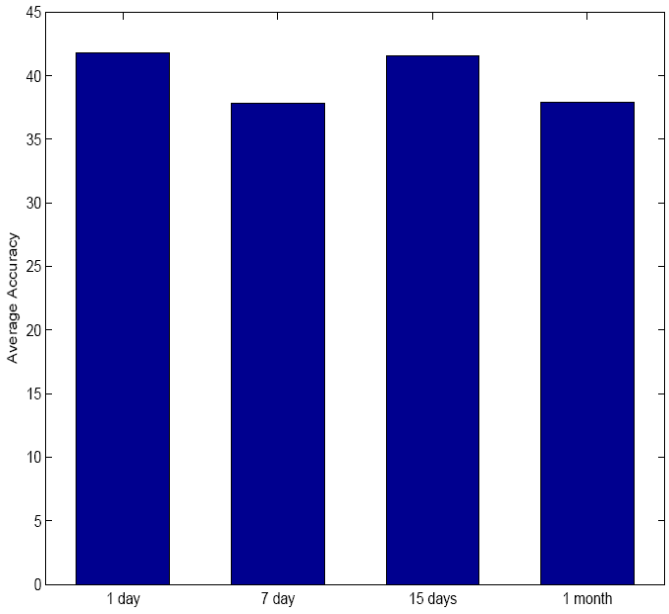
4.4 Importance of history

The next series of tests were conducted to test the behaviour of historical values in predicting the future behaviour of the crowd. Specifically, we wanted to find out the

answers to the following questions: if a person has responded to a request today, then what is the probability that he/she will respond tomorrow? Moreover, what is accuracy? Further, if a person has been active for one week, then what is the probability that he/she will be active tomorrow.

To find the answers to these questions, we conducted a few tests. The test was designed as follows. We wanted to check the predictive capability of the framework by taking in the entire data for the previous one day, previous seven days, last 15 days, and the last one month. The result corresponding to this test is presented in Figure 9. As shown in the figure, we get a high accuracy value when we look into the last one day and the past 15 days. Though, the accuracy is high for the test concerning the last one day, but the difference is not significant. To be precise, the values for one day is 41.76%, and the number for 15 days is 41.54%. The exact reason why the accuracy for these two numbers (one day and 15 days) is high is, however, unknown. But, this result gave a few insights. First, to predict the future behaviour of a person, it is more plausible to look into the recent activity rather than taking into account the entire historical data. This is because the interest to participate in an activity will can change over time. Thus, it is more practical to look into the recent participation habits. Second, this process also has computational advantages. Mining the data to look deep into historical values takes lot of computational time, for example mining the last five years of data. Moreover, as the process is not expected to yield good results, therefore, it not logical to proceed this way. Thus, we recommend using more recent activity for predicting the future participation habits.

Figure 9 Importance of historical values (see online version for colours)



5 Conclusions and future work

In this paper, we revisited the topic of mathematical models capturing human behaviour and their capability to predict future participation habits. We proposed a novel probabilistic method to recruit a candidate to participate in crowdsensing exercises. We focused on a human centric recruitment algorithm and employed data engineering. To demonstrate the feasibility of the method in actual deployment, we engineered a prototype. The prototype was deployed on an enterprise service bus. Further, we used real world datasets to validate the practical application of the method. Through numerical simulations, we found that the method showed good performance.

In this work, we did not took the case of contextual information dictating the actions of an individual. Like many common traits, context information is also one of the important criterion when we deal humans. In the future work, we aim to incorporate this criterion into a statistical model, and make the method more precise.

References

Agarwal, V., Banerjee, N., Chakraborty, D. and Mittal, S. (2013) 'Usense – a smartphone middleware for community sensing', in *2013 IEEE 14th International Conference on Mobile Data Management (MDM)*, IEEE, Vol. 1, pp.56–65.

Amintoosi, H. and Kanhere, S.S. (2013) 'A trust-based recruitment framework for multi-hop social participatory sensing', in *2013 IEEE International Conference on Distributed Computing in Sensor Systems (DCOSS)*, IEEE, pp.266–273.

Amintoosi, H. and Kanhere, S.S. (2014) 'A reputation framework for social participatory sensing systems', *Mobile Networks and Applications*, Vol. 19, No. 1, pp.88–100.

Amintoosi, H., Kanhere, S.S. and Allahbakhsh, M. (2015) 'Trust-based privacyaware participant selection in social participatory sensing', *Journal of Information Security and Applications*, Vol. 20, No. 1, pp.11–25.

Box, G.E.P. and Tiao, G.C. (2011) *Bayesian Inference in Statistical Analysis*, Vol. 40, John Wiley and Sons, Canada.

Cardone, G., Foschini, L., Bellavista, P., Corradi, A., Borcea, C., Talasila, M. and Curtmola, R. (2013) 'Fostering participaction in smart cities: a geosocial crowdsensing platform', *Communications Magazine*, IEEE, Vol. 51, No. 6, pp.112–119.

Clarke, B.S. and Barron, A.R. (1994) 'Jeffreys' prior is asymptotically least favorable under entropy risk', *Journal of Statistical Planning and Inference*, Vol. 41, No. 1, pp.37–60.

Ganti, R.K., Ye, F. and Lei, H. (2011) 'Mobile crowdsensing: current state and future challenges', *Communications Magazine*, IEEE, Vol. 49, No. 11, pp.32–39.

Geoffrey, F. and Schectman, J. (2013) 'Citizen surveillance helps officials put pieces together', *The Wall Street Journal* [online] https://www.wsj.com/articles/SB10001424127887324763404578429220091342796.

Gerla, M. and Kleinrock, L. (2011) 'Vehicular networks and the future of the mobile internet', *Computer Networks*, Vol. 55, No. 2, pp.457–469.

Gomes, J.B., Phua, C. and Krishnaswamy, S. (2013) 'Where will you go? mobile data mining for next place prediction', in *Data Warehousing and Knowledge Discovery*, pp.146–158, Springer, Prague, Czech Republic.

Guo, B., Wang, Z., Yu, Z., Wang, Y., Yen, N.Y., Huang, R. and Zhou, X. (2015) 'Mobile crowd sensing and computing: the review of an emerging human powered sensing paradigm', *ACM Computing Surveys (CSUR)*, Vol. 48, No. 1, p.7.

Hachem, S., Pathak, A. and Issarny, V. (2014) 'Service-oriented middleware for large-scale mobile participatory sensing', *Pervasive and Mobile Computing*, Vol. 10, Part A, pp.66–82.

Hassani, A., Haghighi, P.D. and Jayaraman, P.P. (2015) 'Context-aware recruitment scheme for opportunistic mobile crowdsensing', in *2015 IEEE 21st International Conference on Parallel and Distributed Systems (ICPADS)*, IEEE, pp.266–273.

Holtz, B. and Lauckner, C. (2012) 'Diabetes management via mobile phones: a systematic review', *Telemedicine and e-Health*, Vol. 18, No. 3, pp.175–184.

Hu, X., Chu, T., Chan, H. and Leung, V. (2013) 'Vita: a crowdsensing-oriented mobile cyber-physical system', *IEEE Transactions on Emerging Topics in Computing*, Vol. 1, No. 1, pp.148–165.

Jaimes, L.G., Vergara-Laurens, I. and Chakeri, A. (2014a) 'Spread, a crowdsensing incentive mechanism to acquire better representative samples', in *2014 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*, IEEE, pp.92–97.

Jaimes, L.G., Vergara-Laurens, I. and Raij, A. (2014b) 'A crowd sensing incentive algorithm for data collection for consecutive time slot problems', in *2014 IEEE Latin-America Conference on Communications (LATINCOM)*, IEEE, pp.1–5.

Kim, S., Robson, C., Zimmerman, T., Pierce, J. and Haber, E.M. (2011) 'Creek watch: pairing usefulness and usability for successful citizen science', in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, pp.2125–2134.

Krause, A., Smailagic, A. and Siewiorek, D.P. (2006) 'Context-aware mobile computing: learning context-dependent personal preferences from a wearable sensor array', *IEEE Transactions on Mobile Computing*, Vol. 5, No. 2, pp.113–127.

Li, C-J., Liu, L., Chen, S-Z., Wu, C.C., Huang, C-H. and Chen, X-M. (2004) 'Mobile healthcare service system using RFID', in *2004 IEEE International Conference on Networking, Sensing and Control*, IEEE, Vol. 2, pp.1014–1019.

Liu, K. and Li, X. (2014) 'Finding nemo: finding your lost child in crowds via mobile crowd sensing', in *2014 IEEE 11th International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*, IEEE, pp.1–9.

Pan, B., Zheng, Y., Wilkie, D. and Shahabi, C. (2013) 'Crowd sensing of traffic anomalies based on human mobility and social media', in *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, ACM, pp.344–353.

Ra, M-R., Liu, B., La Porta, T.F. and Govindan, R. (2012) 'Medusa: a programming framework for crowd-sensing applications', in *Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services*, ACM, pp.337–350.

Ranvier, J-E., Catasta, M., Vasirani, M. and Aberer, K. (2015) 'Routinesense: a mobile sensing framework for the reconstruction of user routines', in *2nd International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services, number EPFL-CONF-208793*.

Srivastava, M., Abdelzaher, T. and Szymanski, B. (2012) 'Human-centric sensing', *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, Vol. 370, No. 1958, pp.176–197.

Sun, Y.L., Yu, W., Han, Z. and Liu, K.J. (2006) 'Information theoretic framework of trust modeling and evaluation for ad hoc networks', *IEEE Journal on Selected Areas in Communications*, Vol. 24, No. 2, pp.305–317.

Tan, L. and Wang, N. (2010) 'Future internet: the internet of things', in *2010 3rd International Conference on Advanced Computer Theory and Engineering (ICACTE)*, IEEE, Vol. 5, pp.V5–376.

Wang, W., Gao, H., Liu, C.H. and Leung, K.K. (2016) 'Credible and energy aware participant selection with limited task budget for mobile crowd sensing', *Ad Hoc Networks*, Vol. 43, No. 6, pp.56–70.

Xiao, Y., Simoens, P., Pillai, P., Ha, K. and Satyanarayanan, M. (2013) 'Lowering the barriers to large-scale mobile crowdsensing', in *Proceedings of the 14th Workshop on Mobile Computing Systems and Applications*, ACM, p.9.

Xiong, H., Zhang, D., Wang, L. and Chaouchi, H. (2015) 'EMC 3: energy-efficient data transfer in mobile crowdsensing under full coverage constraint', *IEEE Transactions on Mobile Computing*, Vol. 14, No. 7, pp.1355–1368.