Towards Large Scale Technology Impact Analyses: Automatic Residential Localization from Mobile Phone-Call Data

Vanessa Frias-Martinez and Jesus Virseda and Alberto Rubio and Enrique Frias-Martinez

***Abstract*—Studies to understand the impact that demographic and socio-economic factors have in the use of cell phones have been traditionally carried out by social and technical researchers through the use of questionnaires and personal interviews. In recent years, and due to the pervasiveness of cell phones in emerging and developing economies, large datasets with millions of interactions are generated, anonymized and stored in real time by telecommunication and internet companies. However, these datasets do not typically contain any socio-economic information that characterizes the users. As a result, in order to understand the impact of socio-economic parameters on the use of mobile phones at larger scales, researchers have typically correlated the behavioral analyses drawn from the anonymous cell phone usage datasets to aggregated demographic or socio-economic parameters compiled by institutions like the National Statistical Institutes (NSI) or the World Bank (WB). In order to compute these correlations, the approximate *residential location* of the anonymized users is required. In general, carriers only have such information for users with a contract, which in emerging economies accounts for less than a 5%. In this paper, we propose a new technique to automatically predict the approximate *residential location* of anonymized cell phone users based on their calling behavior, assuming that we have a small set of users for whom their approximate *residential location* is known (the subscribers with a contract). Our results indicate that we can correctly predict the residential location of up to 70% of users with a coverage of 50%. By automatically associating cell phone users to geographical areas, we aim to provide a tool that facilitates the analysis – at a national or global scale – of the impact that socio-economic factors might have in the use of cell phones.**

***Index Terms*—Cell phones in emerging and developing economies, socio-economic analyses, behavioral modeling and characterization, data mining**

I. Introduction

The adoption of information technologies in emerging and developing economies [1] has attracted the interest of many social and technical researchers trying to understand the impact that demographic and socio-economic factors have in the use of technologies [1], [2], [3]. Such analyses are of interest to both policy makers interested in the assessment of technology-based programs, as well as technologists focusing on the development of personalized services for emerging economies. Studies that analyze demographic or socio-economic disparities in the access to technology, are typically based on personal interviews that correlate individual parameters to personal technology use and experiences. Although personal interviews offer important insights that can be helpful towards the characterization of technology usage, these are generally limited to a small number of individuals that are either interviewed in person, or have answered a questionnaire about cell phone usage. Despite the best efforts, the limited number of users that participate in the study may introduce an implicit bias in the analysis.

Due to the pervasiveness of cell phones in emerging economies, large datasets with millions of interactions and cell-phone usage traces are currently generated, anonymized and stored in real time. Telecommunication companies as well as internet companies with mobile services have increasing access to such data. These rich datasets facilitate a large variety of cell phone use analyses in the areas of behavioral analysis [4], human mobility [5], social networks [6], and SMS or web-based m-services [7], [8] at a national scale. In addition to widely expanding the samples of individual interviews, dataset analyses are far less intrusive as individual behavior can be studied without interfering with the users. Such techniques provide a complementary research tool over traditional qualitative approaches [9].

Unfortunately, most large datasets with cell phone usage patterns do not contain any demographic or socio-economic information about individual users. To better understand the impact of socio-economic parameters on the use of mobile phones, researchers have typically correlated the behavioral analyses drawn from the cell phone usage datasets to aggregated demographic or socio-economic parameters compiled by social researchers and ethnographers at institutions such as the World Bank, United Nations or country-based National Statistical Institutes (NSI).

For instance, Eagle [10] studied the correlation between communication diversity and its index of deprivation in the UK. The communication diversity was derived from the number of different contacts that users of a UK cell phone network had with other users. Eagle combined two datasets: (i) a behavioral dataset with over 250 million cell phone users whose geographical location within a region in the UK was known, and (ii) a dataset with socio-economic metrics for each region in the UK as compiled by the UK Civil Service. The author found that regions with higher communication diversity were correlated with lower deprivation indexes. Although this result represents an important first step towards understanding the impact of socio-economic parameters on mobile use at a

[1] Based on the 2009 International Monetary Fund (IMF) World Economic Outlook Country Classification www.imf.org/external/pubs/ft/weo/2009/02/weodata/groups.htm

region level, we seek to elaborate more fine-grained impact analyses that can draw correlations between socio-economic parameters and behavioral models at even smaller scales e.g. cities, neighborhoods.

To accomplish the latter goal, we require a more precise *residential location* (or an approximation) of the set of anonymized users under study. However, telecommunication carriers only obtain the *residential location* information for subscribers that have a permanent contract with the provider, which in the case of emerging and developing economies accounts for less than a 5% of the total customer base (the vast majority generally uses the pre-paid option).

While the evaluation of the impact of socio-economic factors could be restricted to the cell phone users for which residential location is known, we believe that such analysis would not fairly span the large array of socio-economic backgrounds present in emerging economies. In other words, by considering a small fraction of the sample, any study would bias the results towards individuals that have a contract with the telecommunication company.

In this paper, we propose a novel technique to approximate the *residential location* of anonymized cell phone users based on cell phone usage behavior starting from a small set of users for whom their *residential location* is known. Although we demonstrate the process using Call Detail Records (CDRs) from a telecommunications company, the technique presented here could be potentially used to identify the residential location from other types of web or SMS-based application-specific records *e.g.*, consulting a map, checking your email, reading the news or checking the weather for the next day. By associating cell phone users to geographical areas, we open the field to compute and understand the impact that socio-economic factors may have on the way people use mobile phones at a country or a planetary scale.

The paper is organized as follows: Section II summarizes the related work; Sections III and IV formalize the problem of residential location classification and describe our solution based on a genetic algorithm. Section V presents experimental results using Call Detail Records of $100,000$ anonymized individuals from an emerging economy; and section VI summarizes the most important conclusions and future work.

II. Related Work

To the best of our knowledge, there are no previous documented efforts to identify the residential location of an individual based on its cell phone behavioral fingerprint. The lack of research in this area is understandable given that until recently, the combination of behavioral and geo-location information for cell phone users has only been available to telecommunication companies. As new location-based services become available, and the amount of datasets with behavioral and geo-location information increases, we expect to see an increase on identification algorithms for residential and work location. This paper constitutes a first step in that direction.

The remainder of this section details previous studies that may benefit from the residential location classification technique presented here. Kwon *et al.* [3] conducted a study to understand the impact of demographics and socio-economic factors on the technology acceptance of mobile phones. The authors circulated a four-page survey with 33 questions to 500 cell phone subscribers. They found that older subscribers felt more pressure to accept the use of mobile phones than their younger counterpart. In fact, the cell phones were generally given as presents by family members for security purposes.

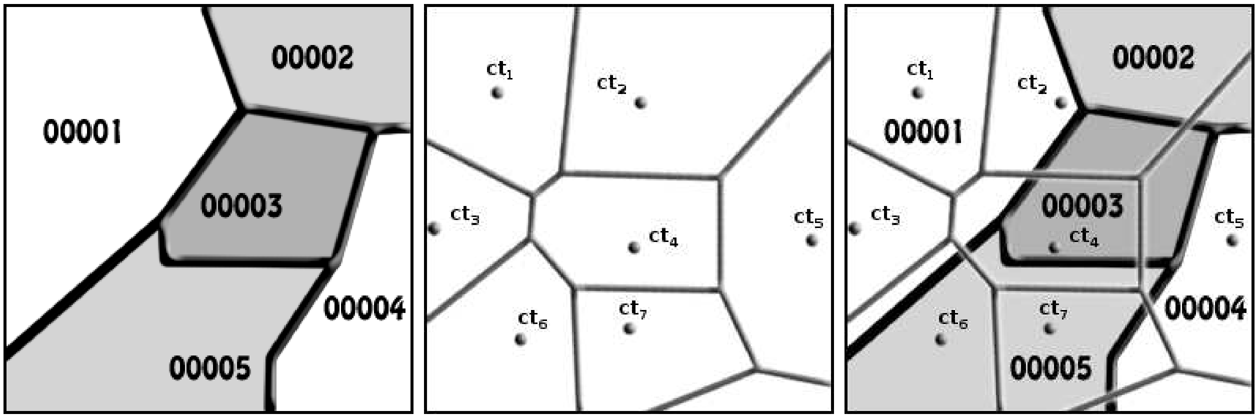
Donner [11] presented a survey of 277 microentrepreneurs and mobile phone users in Kigali, Rwanda, to understand the types of relationships with family, friends and clients, and its evolution over time. Among other findings, the author discovered an inverse correlation between the age of the user and the probability of adding new contacts to its mobile-based social network. The author also claims that users with higher educational levels were also more prone to add new contacts to their social networks.

Reuben [12] studied the economic impact of the use of cell phones on a fishing community in Kerala, India. Through the use of cell phones, fishermen were able to respond quickly to market demand and prevent unnecessary wastage of catch. The author observed that mobile phones helped to coordinate supply and demand, and merchants and transporters took advantage of the free flow of price information. The areas where fishermen largely used cell phones experimented, according to the author, an economic growth.

Many of these studies that correlate cell-phone based behavioral trends with socio-economic parameters, are typically based on interviews or questionnaires to small groups of individuals. Our technique offers the ability to expand these studies to millions of users (without interfering with them) by identifying their residential location and correlating their cell phone behavioral usage to public datasets with geo-referenced socio-economic information.

III. Problem Formulation: Identifying Residential Location

Cell phone networks are built using a set of base transceiver stations (BTS) that are responsible for communicating cell phone devices within the network. Each BTS or cellular tower is identified by the latitude and longitude of its geographical location. The area covered by a BTS can be approximated with Voronoi diagrams [13]. Call Detail Records (CDRs) are generated whenever a cell phone connected to the network makes or receives a phone call or uses a service (e.g., SMS, MMS). In the process, the BTS details are logged, which gives an indication of the geographical position of the user at the time of the call. As mentioned earlier, although we focus on the use of CDRs to model cell phone usage, records from other application-specific services based on web or SMS could also be used. From all the information contained in a CDR, our study only considers the encrypted originating number, the encrypted destination number, the time and date of the call, the duration of the call, and the BTS that the cell phone was connected to when the call was placed.



(a) Distribution of zip codes for an urban area.

(b) Voronoi Diagram showing cell tower coverage areas for the same urban area.

(c) Overlapping of the Zip code map with the Voronoi Diagram.

Fig. 1. Correspondence between Zip Codes and Cellular Towers.

Additionally, the subscribers that have a contract with the carrier, have an indication of their residential location. Throughout this paper, we assume that the residential indication corresponds to a zip code label that approximates the *residential location* of each anonymized subscriber. However, the proposed technique would also work for other formulations of location. While in advanced economies the percentage of subscribers with a contract account for more than 50%, in the case of emerging economies, only $\sim$ 5% of the total population have a contract.

Given a CDR set (containing user cell phone calls for a period of time T) and provided that the residential location of the anonymized users is known *a priori* (in a zip code format), we seek to understand a *residential calling pattern* that characterizes the behavior of users placing or receiving calls from their residential location. We characterize the *residential calling pattern* by the days of the week as well as the specific times of the day at which users make or receive calls at their residential location. In the case of emerging economies, where penetration rates of land lines are much lower than the ones for mobile phones, it makes sense to assume that a vast majority of the population also uses their cell phones while being at home. The underlying assumption is that there exists a shared behavioral fingerprint in each specific society that characterizes social and cultural customs *e.g.,* suburban individuals tend to be home by $6pm$ whereas people in larger cities tend to stay out longer hours and reach home later at night.

We frame the *residential location problem* as a classification problem in which each individual is assigned a BTS that represents her/his residential location. The construction of the classifier is formalized as an optimization problem that seeks to find the best combination of days of the week and times of the day that characterize the calling pattern from residential locations for the set of anonymized users (and their calls) for whom their residential location is known. In other words, we aim to discover the *residential calling pattern* that maximizes the the percentage of users for whom the BTS assigned as residential location is correct. There is a wide variety of techniques to solve optimization problems, and we have selected Genetic Algorithms [14].

The *residential calling pattern* obtained as solution after processing the calls dataset, can then be used to systematically identify the residential location of all the other pre-paid customers lacking any information about their approximate residential location.

Given that cell phone communications are handled by cell towers (BTSs), the residential locations of the users are computed as BTSs by the classification algorithm. However, because the residence of the users known *a priori* is typically expressed using zip codes, we need to first compute the correspondence between BTSs and zip codes before solving the optimization problem.

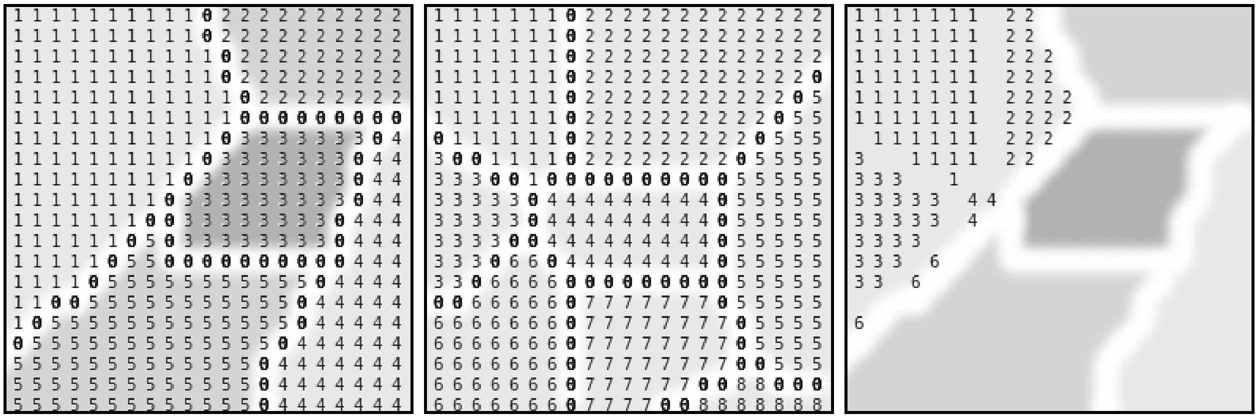
IV. Residential Localization Algorithm

The residential location classification algorithm consists of two main steps: (i) compute the correspondence between the residential location zip codes and the cellular towers, and (ii) solve the optimization problem to identify the residential calling pattern using Genetic Algorithms (GA).

From Zip Codes to Cellular Towers

As discussed previously, the residential location of cell phone users with a contract is known *a priori*. Specifically, the residential location is provided as a zip code, although any other geographical formalization could be used. Since the calls made or received by the users are placed on cellular towers, the network only allows us to identify as residential location a cellular tower (or a set of cellular towers). Thus, we first need to derive the geographical correspondence between zip codes and cellular towers. With the transformation at hand, we will be able to assign a specific set of BTSs to the zip code where the individual claims to live.

To better illustrate the problem, Figure 1(a) shows a map with the zip codes of an urban area. Figure 1(b) shows the corresponding cellular towers that offer coverage to the area,



```
1 1 1 1 1 1 1 1 1 1 0 2 2 2 2 2 2 2 2 2 2
1 1 1 1 1 1 1 1 1 1 0 2 2 2 2 2 2 2 2 2 2
1 1 1 1 1 1 1 1 1 1 1 0 2 2 2 2 2 2 2 2 2
1 1 1 1 1 1 1 1 1 1 1 0 2 2 2 2 2 2 2 2 2
1 1 1 1 1 1 1 1 1 1 1 1 0 2 2 2 2 2 2 2 2
1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0
1 1 1 1 1 1 1 1 1 1 1 0 3 3 3 3 3 3 3 0 4
1 1 1 1 1 1 1 1 1 1 0 3 3 3 3 3 3 3 3 0 4 4
1 1 1 1 1 1 1 1 1 0 3 3 3 3 3 3 3 3 3 0 4 4
1 1 1 1 1 1 1 1 0 3 3 3 3 3 3 3 3 3 3 0 4 4
1 1 1 1 1 1 1 0 0 3 3 3 3 3 3 3 3 3 0 4 4 4
1 1 1 1 1 1 0 5 0 3 3 3 3 3 3 3 3 0 4 4 4
1 1 1 1 1 0 5 5 0 0 0 0 0 0 0 0 0 0 4 4 4
1 1 1 1 0 5 5 5 5 5 5 5 5 5 5 5 0 4 4 4 4
1 1 0 0 5 5 5 5 5 5 5 5 5 5 5 0 4 4 4 4 4
1 0 5 5 5 5 5 5 5 5 5 5 5 5 5 0 4 4 4 4 4
0 5 5 5 5 5 5 5 5 5 5 5 5 5 0 4 4 4 4 4 4
5 5 5 5 5 5 5 5 5 5 5 5 5 0 4 4 4 4 4 4 4
5 5 5 5 5 5 5 5 5 5 5 5 5 0 4 4 4 4 4 4 4
5 5 5 5 5 5 5 5 5 5 5 5 5 0 4 4 4 4 4 4 4
```

(a) Numerical representation of the zip code map shown in Figure 1a computed by a scan line algorithm.

```
1 1 1 1 1 1 1 0 2 2 2 2 2 2 2 2 2 2 2 2 2
1 1 1 1 1 1 1 0 2 2 2 2 2 2 2 2 2 2 2 2 2
1 1 1 1 1 1 1 0 2 2 2 2 2 2 2 2 2 2 2 2 2
1 1 1 1 1 1 1 0 2 2 2 2 2 2 2 2 2 2 2 2 0
1 1 1 1 1 1 1 0 2 2 2 2 2 2 2 2 2 2 2 0 5
1 1 1 1 1 1 1 0 2 2 2 2 2 2 2 2 2 2 0 5 5
0 1 1 1 1 1 1 0 2 2 2 2 2 2 2 2 2 0 5 5 5
3 0 0 1 1 1 1 0 2 2 2 2 2 2 2 2 0 5 5 5 5
3 3 3 0 0 1 0 0 0 0 0 0 0 0 0 0 5 5 5 5 5
3 3 3 3 3 0 4 4 4 4 4 4 4 4 4 0 5 5 5 5 5
3 3 3 3 3 0 4 4 4 4 4 4 4 4 4 0 5 5 5 5 5
3 3 3 3 0 0 4 4 4 4 4 4 4 4 4 0 5 5 5 5 5
3 3 3 0 6 6 0 4 4 4 4 4 4 4 4 0 5 5 5 5 5
3 3 0 6 6 6 6 0 0 0 0 0 0 0 0 0 5 5 5 5 5
0 0 6 6 6 6 6 0 7 7 7 7 7 7 7 0 5 5 5 5 5
6 6 6 6 6 6 6 0 7 7 7 7 7 7 7 7 0 5 5 5 5
6 6 6 6 6 6 6 0 7 7 7 7 7 7 7 7 0 5 5 5 5
6 6 6 6 6 6 6 0 7 7 7 7 7 7 7 7 0 0 5 5 5
6 6 6 6 6 6 6 0 7 7 7 7 7 7 0 0 8 8 0 0 0
6 6 6 6 6 6 6 0 7 7 7 7 0 0 8 8 8 8 8 8 8
```

(b) Numerical representation of the areas covered by Voronoi diagrams shown in Figure 1b computed by a scan line algorithm.

```
1 1 1 1 1 1 1   2 2
1 1 1 1 1 1 1   2 2
1 1 1 1 1 1 1   2 2 2
1 1 1 1 1 1 1   2 2 2
1 1 1 1 1 1 1   2 2 2 2
1 1 1 1 1 1 1   2 2 2 2
  1 1 1 1 1 1   2 2 2
3     1 1 1 1   2 2
3 3 3     1
3 3 3 3 3   4 4
3 3 3 3 3   4
3 3 3 3
3 3 3   6
3 3   6

6
```

(c) Output of the zip codes to cell towers algorithm for zip code 0001 as shown in Figure 1.

Fig. 2. Zip Codes to Cell Towers Algorithm: detecting the fractions of zip codes covered by each cellular tower.

which is approximated using Voronoi diagrams [13]. Every time a user makes a phone call, the call is served by the cellular tower that covers the geographical area where the user is located. Figure 1(c) shows the intersection between the zip code map and the Voronoi diagrams map.

We seek to associate to each zip code the set of cellular towers (BTSs) whose Voronoi diagrams are partially (or totally) included in the geographical area enclosed by the zip code. With this approach, we can represent each zip code zci as $zci = s * ct_a + v * ct_b + ... + w * ct_d$ where $s, v, ...w$ represent the fractions of the cellular towers Voronoi diagrams ct_a, ct_b,...,ct_d that partially cover a certain zip code zci. The final output will associate a list of cell towers to each zip code *i.e.*, $zci = \{ct_a, ct_b, ...ct_d\}$. For example, as can be seen in Figure 1(c), zip code 0001 could be represented as the list of cell towers that cover its geographical area *i.e.*, $zc0001 = 0.5ct_1 + 0.2ct_2 + 0.2ct_3 + 0.05ct_4 + 0.05ct_6$. In our formalism, a user with a zip code 0001 associated to its residential location, will now have it labeled as $\{ct_1, ct_2, ct_3, ct_4, ct_6\}$.

The process to compute the correspondence between zip code areas and Voronoi diagrams is executed as follows. We assume that we initially have the zip code map for the area under study. As a first step, we compute the Voronoi coverage map for the set of cellular towers within that area following Voronoi's algorithm [13]. Next, for both zip code and Voronoi coverage maps, we use a *scan line* algorithm [15] to compute a numerical representation of each map. These representations are then used to calculate the correspondence between each zip code and the coverage of each cellular tower (see Figure 2). Figure 2(a) shows the zip code map previously shown in Figure 1(a), coded by the scan line algorithm. The pixels in each zip code area are represented by a number that indicates its membership to a specific zip code; the borders between zip codes are represented with a 0. Figure 2(b) represents the Voronoi coverage map shown in Figure 1(b), where the scanline algorithm assigns to each pixel a number based on its membership to a specific Voronoi coverage area. Finally, using the numerical representations from Figures 2(a) and 2(b) we compute for each zip code, the Voronoi areas included within the zip code's geographical limits and the corresponding coverage fractions. Figure 2(c) shows an example of the output to compute the Voronoi areas that cover zip code 0001. Pseudocode 1 shows the details of the zip codes to cell towers algorithm.

Pseudocode 1 From Zip Codes to Cell Towers.

```
obtain Zip Code Map for Urban Area
compute number-coded Zip Code Map $< ZCmap >$
compute Voronoi Diagrams for the Cell Towers in Urban Area
compute number-coded Voronoi Diagrams $< VDmap >$
**for** each zip code $zci$ in $< ZCmap >$ **do**
  $counter[ct] = 0 \ \forall \ ct \ \in < VDmap >$
  $total\_pixels\_counter = 0$
  **for** each pixel $pj$ in $< VDmap >$ **do**
    **if** $pj \ \in \ zci$ **then**
      $counter[VDmap(pj)] + +$
      $total\_pixels\_counter + +$
    **end if**
  **end for**
  **for** each cellular tower $ctk$ in $< VDmap >$ **do**
    **if** $counter[ctk] > 0$ **then**
      $zci.add(ctk, counter[ctk]/total\_pixels\_counter))$
    **end if**
  **end for**
**end for**
```

The Optimization Problem

We have formalized the *residential location problem* as a classification problem that assigns to each user a BTS representing her/his residential location. The identification of the calling pattern that assigns users to residential BTSs is formalized as an optimization problem solved with a Genetic Algorithm (GA). The GA focuses on finding the combination of days of the week and times of the day that best characterize the *residential calling pattern* of all the anonymized users for whom both their residential location (zip code) and cell phone calls (CDRs) are known. The residential location of the users is transformed from zip codes to lists of cellular towers (BTSs) using the algorithm described in the previous section. The optimization problem is solved using the JGAP implementation of GAs [16]. Next, we give a brief introduction to Genetic Algorithms and describe its main components for the *residential location problem*.

Introduction to Genetic Algorithms: Genetic Algorithms (GA) are search algorithms based on the mechanics of natural selection tailored for vast and complex search spaces [17]. A GA starts with a population of abstract representations (called $chromosomes$) of candidate solutions ($individuals$) that is forced to evolve towards improved sets of solutions ($populations$). A chromosome is composed of several genes that code the value of a specific variable of the solution. Each gene is typically represented as a string of 0s and 1s. During the simulated evolution, individuals from one generation are used to form a new generation, which is (hopefully) closer to the optimum solution. The idea of survival of the fittest is of great importance to genetic algorithms. GAs use a fitness function in order to evaluate the quality of the solution represented by a specific individual. The fittest individuals will be used to create new, and conceivably better, populations. In each generation, the GA creates a new set of individuals obtained from recombining the fittest solutions of the previous generation (crossover), occasionally adding random new data (mutation) to prevent the population from stagnating. This generational evolution is repeated until some condition (for example number of populations or improvement of the best solution) is satisfied. Hereby, we will refer to *individual* as a candidate solution being evaluated by the GA, and to *user* as a subscriber with cell phone calls whose residential location we want to identify.

In our context, the Genetic Algorithm takes as input the set of cell phone calls (CDRs) made by the users in the dataset, and their residential locations each expressed as a list of cellular towers. Each individual (candidate solution), designed to capture the *residential calling pattern*, is evaluated by a fitness function that computes the number of users for whom the residential location is correctly assigned. After stability is reached, the optimal solution will contain the values that best characterize the *residential calling pattern*. In the next subsections, we describe in detail the chromosome and its genes, the fitness function, the GA architecture and its configuration.

Description of the Chromosome and Genes: We define a chromosome composed of three different genes (see Figure 3). The first two genes represent the *starting time* and the *finishing time i.e.,* the range that defines the time period under which users make cell phone calls from their residential location. Each time variable is composed of seven bits, which divides the day in fractions of 11.25 minutes each. Finally, the third gene represents the *days of the week* when users typically make cell phone calls from their residential location. Each bit of this field represents one day of the week *e.g.,* 1000000 is Sunday, 0100000 is Monday, and 1000001 comprises Saturday and Sunday.

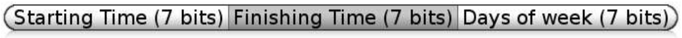


Fig. 3. Scheme of the chromosome and its genes.

Each individual (candidate solution) is evaluated by computing, for each user, the list of cellular towers that comply with the requirements established by the values of the genes. For example, if an individual has the values $(22:11:00, 07:33:00, 1000001)$, we compute for each user the cellular tower that handled calls on Saturdays and Sundays during the time range $22:11:00-07:33:00$.

It may be the case that more than one cellular tower complies with the requirements of the candidate solution. If that is the case, the cell tower with the highest number of calls is selected. Finally, if the resulting cellular tower is included in the list of BTSs that cover the user's zip code, the residential location is considered correct. On the other hand, it may be the case that some users do not make phone calls during the days or times specified by the candidate solution, in which case, that specific user is not assigned any cell tower as residential location.

Fitness Function: In order to evaluate the overall *quality* of each candidate solution, we define the fitness function using the accuracy and the coverage of the *residential calling pattern* described by the individual. We define *accuracy* as the percentage of users for whom the calling pattern correctly assigns as residential location one of the cellular towers in the user's cellular towers list associated to its zip code. On the other hand, *coverage* is defined as the percentage of users from the dataset that are assigned a cell tower (correct or incorrect) as residential location.

The fitness function is defined as $fitness = p*coverage+q*accuracy$ where the values of p and q are weights assigned to each of the two measures depending on the significance we want to give to the accuracy and the coverage of the algorithm. The optimal values for these weights are computed by testing the performance of the Genetic Algorithm across different ranges. We refer the reader to the experimental section V for further details on the calibration of these weights.

GA Configuration & Architecture: In order to evaluate the fitness function, we need to process all the calls made or received by each user. Given that CDR datasets contain millions of cell phone calls that account for gigabytes of data, we have designed an architecture that allows for the evaluation in parallel of large populations of individuals.

Specifically, we have chosen the *shared-pool model* architecture (see Figure 4), where each process (or island) executes a local genetic algorithm and periodically exchanges candidate solutions with other islands through the shared pool [18].

In our architecture, each process is initialized with a randomly generated population of 50 individuals. At every generation, the reproduction is carried out for a 90% of the total population; the crossover is executed with a 35% of pairs of the selected population by randomly selecting a gene in each

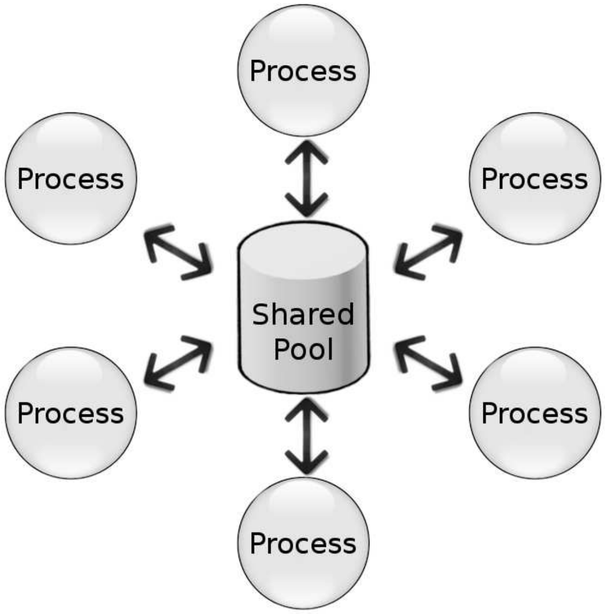


Fig. 4. GA Distributed Architecture: Shared-Pool Model.

individual and exchanging its content with its partner; and the mutation is executed for each gene with a probability of $1/12$ and by randomly creating a new gene. Next, the fitness value for each individual is evaluated and the population is ordered by decreasing fitness values. The fittest individual is always moved to the next generation, and all the other individuals have a probability of being brought to the next generation proportional to their fitness value.

Each process is executed on one core and runs in parallel with the other processes in our architecture of four dual-core Intel processors [2]. In order to increase heterogeneity in the population explored by each process, every 20 generations five individuals *migrate* to the *shared pool*. The migrants are selected using the roulette wheel selector *i.e.,* the better the fitness of the individual is, the higher the probability of migrating is. Upon migration, the process retrieves from the shared pool another five individuals and replaces the previous population using an inverse roulette wheel *i.e.,* the worse the fitness of the individual is, the higher the probability that it will be substituted by a migrant from the shared pool.

V. Experimental Results

In this section we introduce the dataset, we describe the setup of the shared-pool model architecture and discuss the evaluation of the results obtained, both globally and by age groups.

CDR Dataset and Setup

Our initial CDR dataset contains 5 months of cell phone calls collected from $100,000$ residential subscribers in a city from an emerging economy. We exclude users that do not meet an average of at least two calls per day in an attempt to eliminate subscribers that use cell phones sporadically and minimize systematic uncertainties due to calling behaviors based on very few calls. All subscribers have a cell phone contract with the same carrier, and both their zip code residential location and age are known.

The shared-pool model architecture was setup with four computers of eight cores each, which allows us to run up to 32 different processes (islands). Since each island is initiated with 50 randomly generated individuals, the architecture can explore 1600 individuals per generation *i.e.,* we can explore a large number of parameter combinations in little time.

In order to determine a *residential calling pattern*, we first apply the algorithm to match each zip code in the city to a list of cell towers that offer coverage to the associated geographical area (see Pseudocode I). This pre-processing associates to each subscriber a list of cell towers that represent their residential locations. After that, we start the genetic algorithm in each one of the 32 islands with 50 randomly generated individuals and make them evolve until a stable state (a solution) is reached.

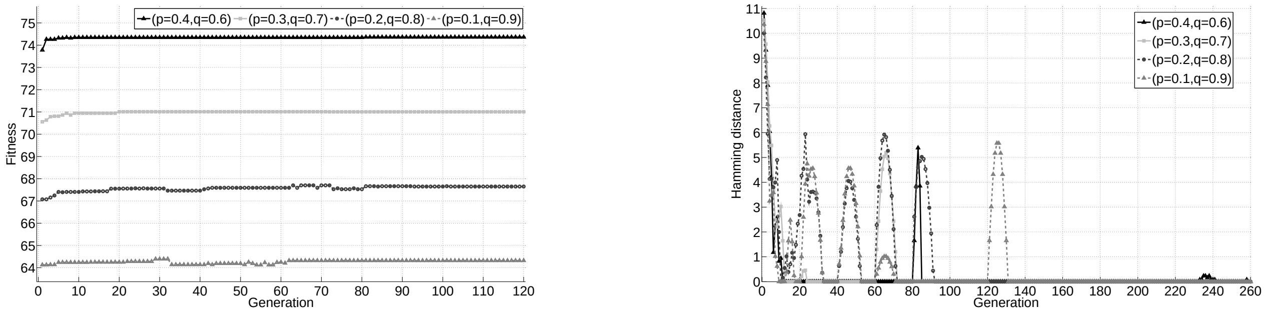
Analysis of Results

In order to understand the evolution of the genetic algorithm, we compute both the fitness and the Hamming distance at each generation. The fitness function helps us understand the quality of the solutions being explored: the higher the value, the better the solution in terms of accuracy and coverage. The Hamming distance is used to measure the *diversity* of the individuals being explored at each generation [19]. As the GA gets closer to the optimal solution, the individuals explored are expected to be more similar among themselves and thus have Hamming distances closer to zero.

As explained in Section IV, the fitness function can give different weights to the accuracy and the coverage of the individuals. To explore the impact that values p and q may have in the fitness function, we evaluate the fitness values for the following combinations $(p = 0.4, q = 0.6)$, $(p = 0.3, q = 0.7)$, $(p = 0.2, q = 0.8)$, $(p = 0.1, q = 0.9)$. Figure 5(a) shows the evolution of the fitness function values for each pair (p, q). Each pair is evaluated using 8 cores (islands) that are booted with 50 randomly generated individuals. For each generation, the fitness value shown per pair (p, q) is the best across all 8 islands (400 individuals) explored. We can observe that the best fitness values are associated to the combination $(p = 0.4, q = 0.6)$ reaching fitness values of up to 75 after 10 generations, and showing small incremental improvements over time. Other combinations reach stability later (after 20 to 60 generations) with smaller fitness values.

Figure 5(b) shows the evolution of the Hamming distance over time. For each generation and combination (p, q), we plot the Hamming distance of the individuals from the island that reached the best fitness function. We observe that after approximately 10 generations the Hamming distance decreases significantly showing that a first *good* solution has been found. Every 20 generations, when the islands exchange individuals,

[2] Dell Precision T7400 with 3.4GHz, 64-bits, and 32GB of memory



(a) Fitness function values per generation to evaluate when a stable state is reached.

(b) Hamming distances between individuals in each generation to evaluate the diversity of the explorations.

Fig. 5. Evolution of the Genetic Algorithm over time for different weight combinations (p,q) in the fitness function.

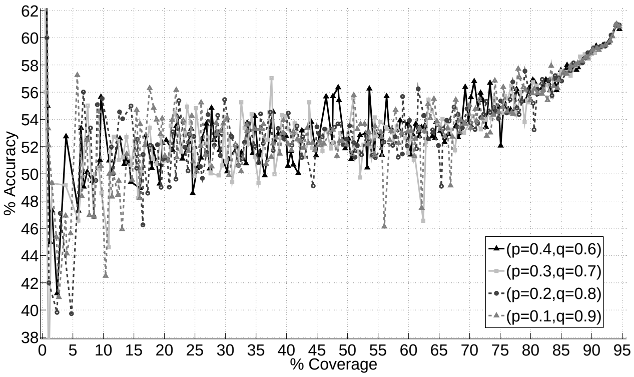


Fig. 6. Accuracy and Coverage values for each combination of weights (p,q) in the fitness function. The pairs (accuracy,coverage) represent the quality of individuals evaluated over time.

we observe a periodic increase in the Hamming distance due to the diversity introduced by the migrants. However, after approximately 60 generations, the Hamming distance reaches a value of 0, thus confirming that a stable state has been reached. Small bumps in the distance are observed subsequently, these mainly relate to randomly generated individuals that are periodically explored by each individual island.

To better understand the meaning behind the fitness function, Figure 6 depicts the accuracy versus the coverage values for each combination (p, q). Each pair in the plots, represents the coverage and the accuracy of the solutions (individuals) explored by any of the islands associated to each (p, q) over time until a stable state is reached (in terms of fitness values and Hamming distances). If two individuals share the same coverage, the best accuracy value is plotted. As confirmed in Figure 5(a), we observe that the best combination of weights is $(p = 0.4, q = 0.6)$, yielding accuracies of up to 61% with a coverage of nearly 95%. Thus, we have found a *residential calling pattern* that reveals the residential location for almost all subscribers with an accuracy of 61%. The *residential calling pattern* was found for days Monday, Wednesday, Saturday and Sunday from $15:45:00$ to $11:03:45$.

Although the genetic algorithm reaches high coverage values, the accuracy does not seem to increase beyond 61%. This means that there exists a 39% of subscribers for whom their residential location was wrongly predicted. In an attempt to dig into the causes of these wrong classifications, we studied the list of cell towers that the candidate solutions (individuals) assigned to the users as potential residential locations. Recall that the algorithm selects as the residential location, the cell tower that has the highest number of cell phone calls among all the cell towers that comply with the requisites specified in the genes. We observed that for many of the users that were wrongly classified, there existed very little difference between the number of calls in the first and the second most used towers. Thus, we decided to add an extra condition in the assignment of the residential location. A user is assigned a cell tower as residential location, if and only if the difference in percentage of total calls between the first and the second cell towers represents a minimum percentage r. If the difference between these two towers is smaller than the percentage r required, the user is not assigned a residential location.

This extra condition attempts to *clean* the behavioral signal related to the *residential calling pattern* by not stating a residence location unless the differences between the two most predominant calling behaviors in a user are sufficiently large. Although this condition lowers the coverage, we aim to increase the accuracy of the residential location classification. We explored values for r from 0% to 30%, where 0% represents the case initially explored that yielded accuracies of up to 61%.

Figure 7 shows, for $(p = 0.4, q = 0.6)$, the coverage and the accuracy reached by the genetic algorithm when we require a minimum percentage r of difference between the first and the second cell towers in order to assign a residential location to each user. As in Figure 6, each pair (accuracy, coverage) represents the quality of the solutions evaluated by any of the individuals explored across all islands until a stable state is reached (in terms of fitness values and Hamming distances). We observe that very good values are obtained when $r = 25\%$ with accuracies of over 72% and a coverage of up to 40% or when $r = 20\%$ with accuracies of approximately 70% and

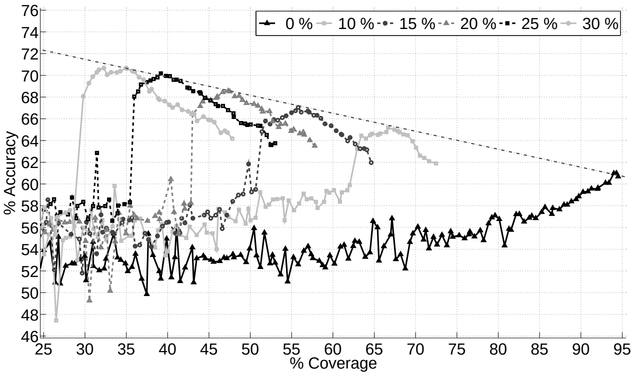


Fig. 7. Accuracy and Coverage values for various percentages r of difference in the number of total calls between the first and the second most used cell towers. The pairs (accuracy,coverage) represent the quality of individuals evaluated over time.

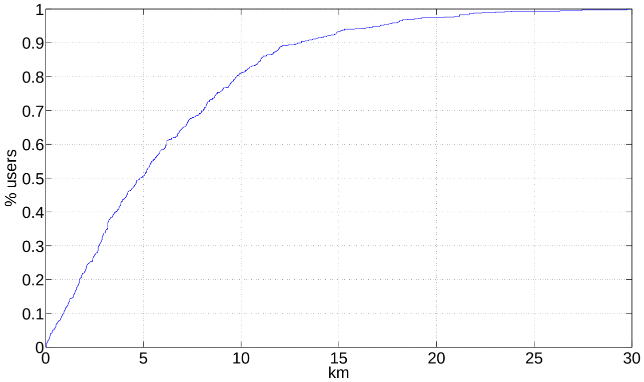


Fig. 8. CDF showing the percentage of misclassified users whose error distance between the real and the detected residential location is smaller than a certain value.

a coverage around 50%. The *residential calling pattern* for the latter was represented by the following candidate solution: days Monday, Tuesday, Friday, Saturday and Sunday from $17:15:00$ to $8:26:15$.

Impact of the Misclassification in the Correlation Analysis

The final aim of the residential location algorithm is to assign a residential location to each user. These locations are then used to associate users to geographical regions that have been characterized by specific socio-economic factors. Potential correlations between these socio-economic parameters and cell phone use might give us an insight into human behavior and technology usage. It may be argued that while the residential location of a 70% of the users is correctly classified, there might a noticeable effect on the correlation analyses from the remaining 30% users that have been misclassified.

In order to quantify such effects, for each misclassified user we computed the distance between the cell tower incorrectly identified as residential location by the algorithm and the centroid of the group of cell towers associated to her/his residential zip code. Figure 8 shows a CDF that represents the percentage of users for whom the error distance between real and predicted residential location was smaller than a certain distance measured in km. It can be observed that for 82% of the misclassified users, the error distance is smaller than $10km$. These erroneous residential classifications might be related to traffic being forwarded through other close-by cell towers or to users who failed to notify changes of residency. In any case, the impact of this small error in further correlation analyses will be minimal.

Analysis by Age Ranges

In an attempt to improve our classification rates and to study the impact of age in the identification of residential location, we run a separate genetic algorithm (with its shared-pool architecture) for each age range considered: (18–24), (25–34),(35–44), (45–54), (55–69).

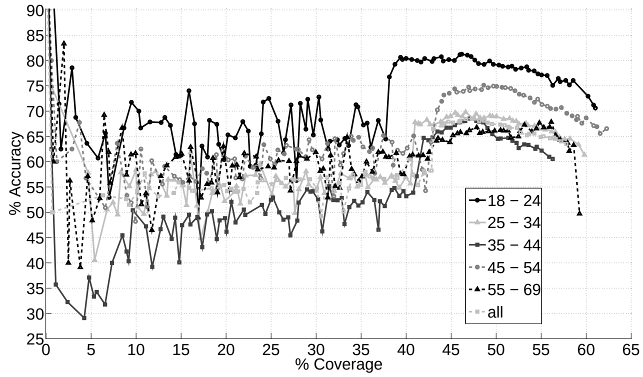


Fig. 9. Accuracy and Coverage values per age range obtained by the individuals explored by the genetic algorithm across all islands until stable states are individually reached.

Figure 9 shows, for $(p = 0.4, q = 0.6)$ and $r = 20\%$, the coverage and accuracy for each age range. Each pair (coverage, accuracy) specifies the quality of the solution represented by the individuals explored by each genetic algorithm until stable states are individually reached (in terms of fitness values and Hamming distances). In fact, by dividing users by age ranges, we observe an improvement in the accuracy for certain age groups. This result also implies that different age groups have distinct *residential calling patterns* and as a result some age groups might be more predictable than others. For the ranges considered, the age range (18–24) showed the highest improvement, reaching accuracies of up to 80% with a 40% coverage, with the *residential calling pattern* being Monday, Wednesday and Friday from $17:37:30$ to $7:18:45$.

Solution Degradation

Given that genetic algorithms are computationally expensive, minimizing the cost of the fitness function evaluation is particularly relevant. To this end, we want to understand to which extent the accuracy and coverage of the solutions

DataSet Size	Accuracy	Coverage	Processing Time
4 months	67.21%	49.03%	3.59s
3 months	66.57%	48.89%	2.67s
2 months	65.46%	48.32%	1.94s
1 month	64.51%	46.56%	1.15s
2 weeks	62.04%	45.61%	0.73s
1 week	59.00%	44.51%	0.49s

TABLE I
ACCURACY, COVERAGE AND PROCESSING TIME VALUES FOR CDR DATASETS OF DIFFERENT SIZES.

degrade depending on the size of the CDR dataset being processed. With smaller CDR datasets, the time it takes to evaluate each subscriber's residential location and the final solution will be reduced, but the information available to identify cell phone behavioral patterns will be more limited.

To quantify the degradation, we divided the initial 5 month CDR dataset into different sets of smaller size ranging from 4 months to 1 week each. For each one of these CDR subsets, we ran the genetic algorithm with the shared-pool model and let it evolve until a stable solution (in terms of fitness values and Hamming distance) was reached. Table I shows, for each dataset size, the accuracy, the coverage, and the average time needed to process one candidate solution. The accuracy and coverage values are computed as the average of these measurements throughout all possible temporal datasets with the same size *e.g.*, we compute the accuracy for all the 1 week sets and report the average of these measurements. The processing time represents the average time needed to evaluate one individual (candidate solution) for each CDR dataset.

In general, we observe that as the CDR dataset shrinks in size, the accuracy and coverage of the best solutions also decrease in an almost linear way. Additionally, the time needed to compute the residential locations also decreases linearly. In general, the selection of the optimal CDR-dataset size depends on the computational capabilities available and the expected quality of the solution *i.e.*, better accuracy and coverage over longer processing times or *vice versa*.

VI. CONCLUSIONS AND FUTURE WORK

The pervasiveness of cell phones in emerging and developing economies is generating large datasets with millions of interactions and cell-phone usage traces which are anonymized and stored in real time. These datasets have enabled, for the first time in history, a variety of technology-usage analyses at a country-scale level, increasing the samples available from small groups of interviewed users.

In this paper, we have presented a technique to identify the approximate *residential location* of anonymized cell phone users based on their calling behavior. By associating cell phone users to geographical areas, we have opened the door to compute and understand the impact of socio-economic factors on the way people use mobile phones at very large scales.

Our technique aims to find a *residential calling pattern* that characterizes the behavioral fingerprint of subscribers within a geographical area. We have framed this characterization as an optimization problem that seeks to understand the times of the day and days of the week at which the bulk of subscribers use their phones from their residential locations. Our results show that we can achieve accuracy rates of around 61% with a coverage up to 95%. We have also explored a more constrained technique that aims to improve the accuracy by decreasing the coverage, obtaining an accuracy of 70% with a coverage of 50%. Age based analysis indicates that if the age information is available, a higher accuracy can be obtained by producing independent GAs for each age group. Finally, we have shown that more than 80% of the users whose residential location was misclassified show distance errors smaller than $10km$.

Although different cities will reveal distinct calling patterns, we believe that this paper represents a template solution for the automatic detection of the residential location based on cell phone usage data.

Future work will focus on using the technique presented to compute the residential location of subscribers across different cities in emerging economies. By characterizing each subscriber using the cell phone behavior and the residential location, we will be able to understand whether there exist correlations between socio-economic factors associated to different geographical areas and cell phone usage at country or planetary scales.

REFERENCES

[1] W. Taylor, G. Zhu, J. Dekkers, and S. Marshall, "Socio-economic factors affecting home internet usage: Patterns in central queensland," *Informing Science Journal,* vol. 6, 2003.

[2] A. El-Ghannam, "The influence of demographic and socio-economic factors upon using information technology among more, moderate, and less developed countries in the globe," *International journal of sociology and social policy*, vol. 25, no. 10–11, pp. 37–53, 2005.

[3] H. Kwon and L. Chidambaram, "A test of the technology acceptance model: The case of cellular telephone adoption," in *Proceedings of the 33rd Hawaii International Conference on System Sciences*, 2000.

[4] V. Frias-Martinez, E. Frias-Martinez, and N. Oliver, "A gender-centric analysis of calling behavior in a developing economy using call detail records," in *AAAI Spring Symposium on Artificial Intelligence for Development (AI-D)*, 2010.

[5] A. Rubio, V. Frias-Martinez, E. Frias-Martinez, and N. Oliver, "Human mobility in advanced and developing economies: A comparative study," in *AAAI Spring Symposium on Artificial Intelligence for Development (AI-D),* 2010.

[6] J. Onnela and J. S. et al., "Structure and tie strengths in mobile communication networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 18, 2007.

[7] Grameen, "Mobile technology for community health (MoTeCH)," 2009, www.grameenfoundation.applab.org/section/ghana-health-worker-project.

[8] ZMQ, “mHealth for development: A UN and Vodafone Foundation report,” 2008, www.unfoundation.org/global-issues/technology/mhealth-report.html.

[9] S. Laws, C. Harper, and R. Marcus, “Research for development, a practical guide,” *Save the Children, Sage Publications Ltd.*, 2003.

[10] N. Eagle, “Behavioral inference across cultures: Using telephones as a cultural lens,” *IEEE Intelligent Systems*, vol. 23:4, pp. 62–64, 2008.

[11] J. Donner, “The use of mobile phones by microentrepreneurs in kigali, rwanda: Changes to social and business network,” *Information Technologies and International Development*, vol. 3, no. 2, 2007.

[12] A. Reuben, “Mobile phones and economic development: Evidence from the fishing industry in india,” *Information Technologies and International Development*, vol. 4, no. 1, 2007.

[13] G. Voronoi, “Nouvelles applications des paramètres continus à la théorie des formes quadratiques,” *Journal fur die Reine und Angewandte Mathematik*, vol. 133, pp. 97–178, 1907.

[14] J. H. Holland, “Adaptation in natural and artificial systems,” *University of Michigan Press, USA*, 1975.

[15] J. M. Lane, L. C. Carpenter, T. Whitted, and J. F. Blinn, “Scan line methods for displaying parametrically defined surfaces,” *Communications ACM*, vol. 23, no. 1, pp. 23–34, 1980.

[16] K. Meffert, “Jgap - java genetic algorithms and genetic programming package,” http://jgap.sf.net.

[17] D. Goldberg, “Genetic algorithms in search optimization and machine learning,” *Addison Wesley*, 1989.

[18] G. Royd, H. Lee, J. L. Welch, Y. Zhao, V. Pandey, and D. Thurston, “A distributed pool architecture for genetic algorithms,” in *CEC’09: Proceedings of the Eleventh conference on Congress on Evolutionary Computation*. Piscataway, NJ, USA: IEEE Press, 2009, pp. 1177–1184.

[19] R. W. Hamming, “Error detecting and error correcting codes,” *Bell System Technical Journal*, vol. 26, no. 2, 1950.