Mobile Agent-Based Approach for Modeling the Epidemics of Communicable Diseases

M. Saravanan, P. Karthikeyan
[1]Ericsson Research India
Chennai, India
m.saravanan@ericsson.com, p.karthikeyan@ericsson.com

A. Arathi, M. Kiruthika and S. Suganya
Meenakshi Sundararajan Engineering College
Chennai, India
msecarti@gmail.com, mailkrithie@gmail.com,
mailme.s.suganya@gmail.com

Abstract—**The increase in the use of mobile phones generates the formation of mobile social networks which can make use of various purposes including education, public health and controlling epidemics. Social networks consist of the basic building blocks called as the communities within which the social interactions are intensive, but between which they are very weak. Everyone could observe that the spread of infectious disease inside communities often has the ability to cross countries borders and spread rapidly. With the widespread of diseases causing major public health problem, we argue that human mobility patterns not only influence the spreading, but are also useful for preventing and creating awareness of the diseases. In this paper, we present new opportunities offered by the field of mobile social networks for understanding the spread of infectious diseases. For this purpose we propose two models namely MABM (Mobile Agent Based Model) and SDC (Spread Discovery Control) model to understand the spread of communicable diseases between different regions. The proposed SDC model is used to comprehend the spread of diseases by extracting the community structures and the analysis of mobility pattern of each agent (user) within the mobile network. Moreover, the understanding of spread details helps us to propose the control strategy to avoid the spread of the epidemic disease on the specific region. To realize our proposed models in a better way, we have modeled one such communicable disease usually spreading every year in West African region**

Keywords—**Social Networks, MABM, SDC, Community Mining, Influential user.**

I. INTRODUCTION

Mobile social network analysis based on mobile communication services has been used to map the interaction between groups of people. Today, mitigating the effect of an epidemic has become the major public health concern for all countries. The outbreak of an epidemic is triggered due to person-to-person contact while the human contact networks in turn exhibits a strong community structure. The call detail records are very useful in depicting the human contact and mobility traces and hence in finding the social community structures that are available in specific time period on particular location. The novelty in our approach is to explore the possibilities for understanding the principles for telecom community structures that can lead to the prediction of the disease spread.

Agent based models (ABM) were developed to effectively simulate the spread pattern of diseases by considering the uniqueness in each agent's behavior [1]. However, most agent based models have not combined the real geographic environment with the agent environment [2]. In this paper, we propose a Mobile ABM that uses the Voice calls, SMS and MMS, etc related to each individual Call detail Records (CDRs), and derives the social interaction from the individual's mobility patterns. This pattern identification has led to realization that each individual is being characterized by a particular travel distance and they travel to the set of few places frequently, which means they show temporal and spatial regularity [3].

The subdivision of the social network by considering the spatial and temporal aspects provides an accurate partitioning of entire network into smaller communities, in the best possible way. After the extraction of the communities, the structural properties of each community is analyzed to find the related role of *strong and weak ties* in information spread among the members [5, 6]. Analysis of stronger ties allows in identifying unique users called *influencers*. They are the agents playing a key role in spread. In this study we would like to focus them to control the spread of communicable diseases.

Hence, the main contribution of our research is the use of mobile phone call detail records (CDRs) for evaluating different communities in each region and determining the mobility pattern of each user located in different communities. This therefore helps in tracking the disease spread pattern across the communities based on the degree of connectivity between each user and thus can be used to understand and thereby control the pandemic diseases spread.

To initiate the control strategy, we have proposed Spread Discovery Control (SDC) model that can provide insights into preventing the spread of disease between the communities by informing the highly influential persons who have high probability of spreading the alert to a large number of persons in his or her community. The influential users in mobile networks are the ones who can communicate to more number of other users. The identification of such influential users can help us to compute the various infectious and control rates possible among communities, based on whether the carrier is an ordinary agent or an influencer.

II. Related Work

We have focused on the following studies that are relevant to disease spread and agent-based models. It helps us to initiate new methods that model the epidemics of communicable diseases through mobile CDRs.

A. Compartmental Epidemic disease models

The SIR model [1] was proposed for studying the disease spread (disease diffusion) and it involves three distinct agents- Susceptible, Infected and Recovered. This model divides the population into three states: S, I, R referring to the susceptible, infective and recovered groups of people, respectively. R is referred to as, removed or refractory as well. The Compartmental models don't consider the individualistic nature of humans. They assume that all persons in a particular compartment are identical. The compartmental models suffer from behavioral generalizations.

B. Agent Based Epidemic disease models

An agent is autonomous, adaptive and self directed. Episimdemics algorithm [5], an agent-based approach, simulates the spread of diseases in a realistic social contact network. This work develops a simulator for agent-based modeling and it uses a DES (Discrete Event Simulation) algorithm which changes its state based on the occurrence of event. This approach fails to combine the real geographic environment with the agent environment.

III. Algorithms used in SDC Model

In order to determine the influential disease spreading agent, their mobility pattern and the infectious rate of each agent, we have introduced SDC model that includes the following algorithms.

A. Gravity model algorithm

We have adopted the following expression as a model for the number of consumers C_{ij} travelling between two nearby locations i and j:

$$C_{ij} = P \frac{N_i^{\alpha} N_j^{\gamma}}{\exp(\beta d_{ij})} \qquad (1)$$

where d_{ij} is the distance between the two locations in kilometers and N_i and N_j are the number of mobile phones users of the locations i and j, respectively. The above expression has four free parameters: the exponents α and γ, the inverse characteristic distance β and the proportionality constant P. By applying the logarithmic transformations we get the expression as follows

$$\log(C_{ij}) = \alpha \log N_i + \gamma \log N_j - \beta d_{ij} + \log P \qquad (2)$$

Hence higher the value of C_{ij}, we obtain the maximum number of people moving between the two locations.

B. Community Identification Algorithms

There are many algorithms to find the number of communities in a given network. Based on our past experience

[4], we have applied fast unfolding algorithm for identifying the communities on mobile social networks. Fast unfolding algorithm is an alternate way of finding communities using modularity. The community formation algorithm is divided into two phases. In the first phase, each node has its individual community. Then the modularity is found with all its neighbors and the change in modularity value is evaluated. If there is a positive gain, the communities of nodes are merged into one. This first phase stops when a local maxima of the modularity is attained, i.e. when no individual move can improve the modularity. In second phase, each community is taken as node and the process is repeated, until all the weights are summed. These two phases are iteratively performed unless stabilized value is reached.

C. Applying Shapley value in identifying Influential Spreaders

The game theoretic network centrality assists in finding out the importance of each node in terms of its utility when combined with the other nodes [10]. Given a telecom network, the game theoretic network centrality indicates the coalition value of every combination of nodes in the network. We have used the *Dijkstra's algorithm* to efficiently track the shortest distance between a given node and its neighbor's in calculating Shapley value [12] for each node.

D. Clustering using K-Means algorithm

The K-Means [11] clustering algorithm is also called as the filtering algorithm. We have used K-Means clustering to understand the tie strength among the members in the community and to propose a strategy to simulate spread and control environments.

E. Calculating Infectious Rates

Once the influential agent is identified based on the above algorithm, we determine the community in which the influential agent is present. We identify the infectious rate of the influential agent based on the susceptible people for the influential agent in each level. The infectious rate is determined based on the formula

$$P_i = 1 - \exp(\tau \sum_{r \in R} N_r \ln(1 - rs_i \rho)) \qquad (3)$$

where P_i is the probability that an infection is triggered in a susceptible agent i, τ is the duration of exposure; R is the number of levels that the agent covers the network; N_r is the number of susceptible agent in each level r ; s_i is susceptibility of individual i and ρ is the basic transmissibility of the disease. The value of ρ is identified as follows:

$$\rho = R_o/\eta \qquad (4)$$

where R_o is the average number of calls done by the agent and η is the infectious period. Every agent is considered to be susceptible to the given disease at any point of time. Hence, the susceptibility of each agent is considered as 1. The other parameters are relevant to the specific data set used.

IV. MODELLING THE EPIDEMICS

The daily activities of the people and their role in spreading of disease in the real world are simulated by creating a spatial and temporal model of disease transmission. The spreading pattern can be determined by extracting the dataset using the movement of people.

A. Dataset

In this study we have used the datasets based on anonymized CDRs of voice calls and SMS exchanges between four million customers in West African operator. The dataset which we used has the following:

(a) Antenna-to-antenna traffic on an hourly basis

(b) Individual trajectories for 50,000 customers for two week time windows with antenna location information

(c) The entire observation period with sub-prefecture location information

B. Mobile Agent Based Model (MABM)

MABM is a new approach to the modeling of mobile telecom users as agents. The behavior of the mobile agent is the interactions with the surrounding communities forming the network of connections using the mobility pattern traced with the help of CDRs. An agent can function independently in its environment.

We have understood that Meningitis disease [14] spread has proven vitally fatal in the recent years in the regions of West Africa. Hence we investigate the feasibility and effectiveness of the collected CDR details of mobile phone to derive the temporal spatial distribution of the social community to make decision on the Spread, Discover and Control (SDC) model of our study.

C. Spread Discovery Control (SDC) Model

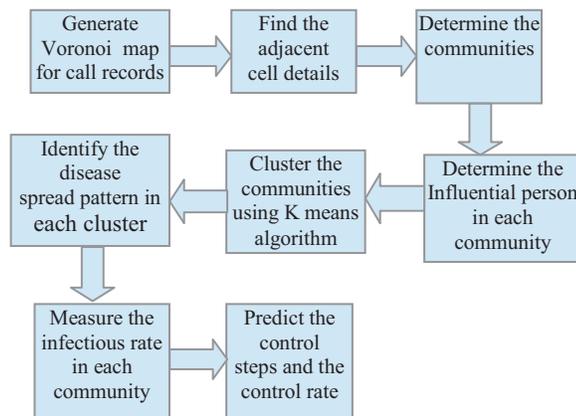There are three different categories of epidemiological modeling were considered in our approach.

Fig. 1.   General Structure of SDC Model

1) Spread Model

The Spread model focuses on fine-grained modeling of the spread of disease (e.g. Meningococcal Meningitis) in a large real-world social network. The spread has been initiated from one region in Western Africa. We carry out the simulations and measures with the real time data of the CDRs of 4 million customers of Western African Operators. Based on their call patterns we have identified the users who are located in Region I. This is because the disease Meningococcal Meningitis that we study originated from Region I. The influential agent of the spread model is capable of spreading the disease faster in short span of time. In this study, we have determine the influential agent based on two criteria: i) the **total time** that two people spend talking to each other as one of the measure and the **number of times** the person has contacted, as greater the number of calls greater the probability that the agent might meet the people he has contacted; ii) the **number of people** the agent has called, as greater the number of people contacted wider the disease spread.

2) Discover Model

It is characterized by two main components such as mobility identification and the formation of social network.

a) Mobility Identification

Every time the call is made, the details of calls like calling party, called party, BTS id, sub- prefecture id, connection date time etc are stored. Each time a CDR is generated we get a sample about the users location details at the time of call. Using this statistics we identify the users who move form Region I to various other regions. With this we get an idea of how the disease is spread from Region I to the other regions and also the persons who are responsible for the disease spread. The general structure of our model is as shown in Fig 1.

b) Social network formation

The disease spread across the locality occurs when the agent moves to the new locality physically which is determined from the CDR details. When the agent uses the antenna id on the new locality for his calling it implies that there is a physical movement of agent. Hence this helps us to track the mobility pattern of the agent from his base location. It is identified that maximum number of people have travelled from Region I(the originating point) to a Region II. Also, the number of calls that have been made from the people with home location as Region I to people with Region II is the largest when compared to other locations..

3) Control Model

The ultimate goal of any research work is to create mass awareness. Hence we consider the CDR's to identify the influential person from the mobile social network. The geographical distribution of any place along with the cell towers located in the regions bordering the place can help tourist and other persons entering the disease prone areas, to gain access to preventive measures, symptoms, treatments and medical facilities available.

V. EVALUATION

A querying tool Tora [8] is used to select mobile phone users of Region I by using the antenna id coverage in the region from the dataset. The dataset consisted of the people whose home location is Region I and also the people who have traveled to Region I as there is a high probability that the

individual would contract the disease from the infected persons in Region I. Fig 2 shows the distribution of Antenna IDs of Region I formed using Voronoi Tessellation.
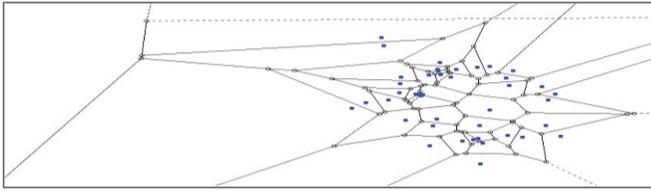

Fig. 2.   Region I Antenna distribution

A.   Tracking user movements

The movement of every user is traced based on the individual mobility pattern using the gravity model [7]. This mobility pattern of mobile phone users is used to identify the various other localities to which these users have travelled and thus estimated nearly to 23 regions. Fig 3 shows the details of mobile user's movement from Region I to other regions inside West Africa. It helps us to simulate the spread of disease pattern to other regions.
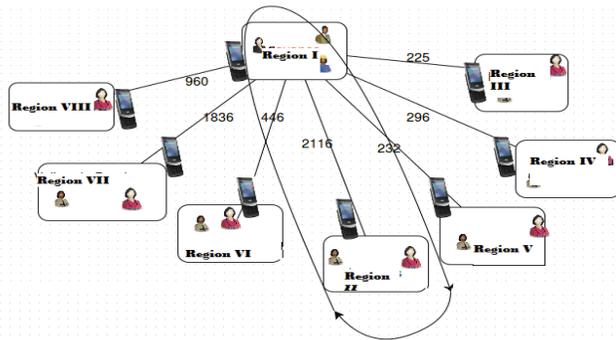

Fig. 3.   *Tracking the user movements from CDR's*

According to the data provided by WHO [14], it was found that the disease Meningitis spreads maximum from Region I to Region II. While studying the pattern of movements, we also found that greater number of people moved from Region I to Region II and hence we identified these two regions for applying our SDC model to figure out the controlling strategy in discussion section.

VI.   EXPERIMENTATION AND RESULTS

We analyzed the data set given to us in all the possible ways by conducting various experiments and thereby deriving different attributes. These attributes were derived from the data set using the querying tool, Tora [8]. To get a clear picture, these data were visualized in a network graph using the visualization tool Gephi [9].

A. Detecting Communities in different regions

Provided with the details of 4 million dataset of West African Operator, we have identified the disease prone region as Region I which matches with the information provided by WHO [14]. After processing the call detail records, we have identified that most of the people travelled from Region 1 to

Region II. Hence we approach this problem in three different directions as follows: (a) Analysis of the spread pattern within the localities of Region I; (b) Analysis of the spread pattern for the people travelling from Region I to II; (c) Analysis of the spread pattern within the localities of Region II.

B.   Impact of influential users

In each of the above said stages, we have identified list of influential user and its connections based on their respective Shapley values.
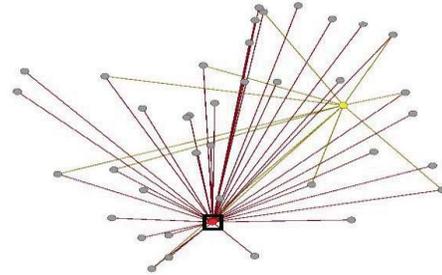

Fig. 4.   Identify influential user using Shapley Value

The squared node depicts the most influential node in the community in Fig 4. It has connections with almost all the other nodes in that community and has the highest Shapley value in that community. Or conversely, greater the Shapley value, the more influential the node is. If in case the influential node is also the infected node, then there is high probability that the entire community will be infected in a shorter span of time. Hence more the number of influential users, greater the disease spread. Table 1 depicts the total number of influential users in each stage.

TABLE I.   IDENTIFICATION OF INFLUENTIAL USERS IN UNCOVERED COMMUNITIES

	Total number of users	Total number of influential user	Percentage
Region I to Region I	2762	179	6.4808
Region I to Region II	2147	81	3.7727
Region II to Region II	7526	329	4.3715

The percentage of influential users is higher within the Region I and Region II when compared to the set of users who have shown high mobility patterns (Region I to Region II).

C. Finding Infectious Rates

After generating the influential users on the emerged communities with different hop of nodes, we derive the infectious rate for each node. The value of expected fraction of infected members after specific (r) rounds is computed, and the number of persons infected in each round increases exponentially. The infectious rates of the influencers is calculated for each hop and compared with that of ordinary users. This is done cluster-wise. Based on various experiments using different values for k, it was found that k=3 is more apt as it contains sufficient number of samples in each cluster.

Three clusters are generated i.e., high, medium, low. Table 2 shows the infectious rates calculated for different clusters.

TABLE II.   INFECTIOUS RATES CALCULATED FOR DIFFERENT CLUSTERS

REGION I TO REGION II	SINGLE INFLUENCER	SINGLE NORMAL USER	AVERAGE OF NORMAL USERS
HIGH CLUSTER	1.0000	0.1976	0.1976
MEDIUM CLUSTER	0.9996	0.1222	0.1222
LOW CLUSTER	0.9993	0.0720	0.0557

Here, based on the number of nodes, the communities are all divided into clusters. As the size of the cluster increases, the proximity between the nodes in that particular community increases; thereby increasing the infectious rate.

VII.   DISCUSSION

The Compartmental models don't consider the individualistic nature of humans. They assume that all persons in a particular compartment are identical. Unlike this our model considers the behavior of every individual in tracking the disease spread. The SDC model is implemented to determine the areas in which the disease spreads, by dividing the larger network into smaller communities, which on analyzing provides a wider perspective of determining the disease spread pattern and helps the concerned authorities to implement the control strategies to avoid the happening of major catastrophe. To prove the significance of our approach of identifying influential users in comparison with normal users, we have performed paired t-statistics [13] on the calculated infectious rates as part of SDC model implementation. T-statistics is used to determine the level of mean score difference between influential user spread rate with one single normal user and five normal users in different regions. Table 3 shows the results of t-values with significant level to strengthen our proposal of start discussing the control measures from Influential users rather than the normal users.

TABLE III.   T-STATISTICS FOR MEAN SCORE DIFFERENCE BETWEEN INFLUENTIAL USERS WITH NORMAL USER

No: of communities	Regions	Influential Vs Single normal user	Influential Vs Average of five normal users
71	Region I to Region I	132.93**	129.22**
35	Region I to Region II	92.76**	90.53**
159	Region II to Region II	163.72**	163.12**

** - Significant at p<0.1 level

VIII.   CONCLUSION

This research project thereby addresses the organized research in the area of multifarious systems models design, with a specific aim to improve the accuracy of predicting the disease prone areas and alert the customers and tourists with appropriate recommendations by incorporating real-world human mobility patterns from call detail records. The models proposed in this paper also help us to capture the complexity of people movements and the way of spread of disease to others.. The proposed models could impact all phenomena driven by human mobility, from epidemic prevention to emergency response, urban planning and Mega city development projects.

Future enhancements could involve network modeling and usage of real time data's. Also an application can be developed so as to curb the disease spread. Using this application the users who are connected to the infected persons can be notified of the disease. The infected user informs the operator about the status of his health condition. The operator checks whether the user is an influential user or whether he is connected to an influential user based on his presence in a community. This process of notifying the sub network of the infected user is done without affecting the privacy of the former. This is one way the disease spread can also be controlled in an effective way

REFERENCES

[1] Enrique Frıas-Martınez, Graham Williamson, Vanessa Frıas-Martinez, "An Agent-Based Model of Epidemic Spread using Human Mobility and Social Network Information", Telefonica Research, Madrid – Spain,School of Computer Science , University College, Dublin – Ireland

[2] Jiasheng Wang, Jianhong Xiong, Kun Yang, Shuangyun Peng, Quanli Xu, "Use of GIS and Agent-based Modelling to Simulate the Spread of Influenza", In Proceedings of 18th International Conference on Geoinformatics, 2010

[3] Marta C.Gonzalez , Ceasar A. Hildalgo and Albert-Laszlo Barabasi, "Understanding individual human mobility patterns". Nature, 453, 779-782 (5 June 2008) | doi:10.1038/nature06958; Received 19 December 2007; Accepted 27 March 2008.

[4] Saravanan M and Prasad, G "Labelling Communities using structural properties", In Proceedings of DBKDA, 2010.

[5] Christopher L. Barrett,Keith R. Bisset, Stephen G.Eubank,Xizhou Feng, Madhav V.Marathe, "EpiSimdemics: an Efficient Algorithm for Simulating the Spread of Infectious Disease over Large Realistic Social Networks", In Proceedings of the 2008 ACM/IEEE conference on supercomputing, 2008.

[6] Bakshy E., Rosenn I., Marlow C., Adamic, " The role of social networks in information diffusion", WWW 2012, April 16-20, 2012, Lyon, France.

[7] Duygu Balcan, Vittoria Colizza,Bruno Goncalves,Hao Hu,Jose J.RAmasco and Alessandro Vespignani, "Multiscale mobility networks and the large scale spreading of infectious diseases", PNAS 106, 21484 (2009)

[8] http://torasql.com/

[9] Mathieu Bastian and Sebastien Heymann Mathieu Jacomy , "Gephi : An Open Source Software for Exploring and Manipulating Networks ",Gephi, WebAtlas,Paris, France

[10] Aadithya, K. V. and Ravindran, B. (2010) "Game Theoretic Network Centrality: Exact Formulas and Efficient Algorithms". In the Proceedings of the Ninth International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2010), pp.1459-1460.

[11] Ian Witten and Eibe Frank, Data Mining: Practical MachineLearning Tools and Techniques, 2nd Edition, Morgan Kaufmann, ISBN 0120884070, 2005.

[12] Saravanan, M., and Vijay Raajaa G.S. "A Graph-based Churn Prediction Model for Mobile Telecom Networks", In Proceedings of the 2012 International Conference on Advanced Data Mining and Applications (ADMA 2012), Nanjing, China, 15-18 December 2012

[13] Siegal S, Castellan NJ (1988) Nonparametric statistics for the behavioral sciences. McGraw Hill, Berkeley.

[14] http://www.who.int/csr/don/2009_03_25/en/index.html