Cluster Aware Mobility Encounter Dataset Enlargement

Rajarshi Haldar[†], Salih Safa Bacanli[‡], Moayad Aloqaily[⋆], Adel Ben Mnaouer[σ], and Damla Turgut[‡]

[†]University of Illinois Urbana-Champaign, Champaign, IL, USA
[‡]Department of Computer Science, University of Central Florida, Orlando, FL, USA
[σ]Canadian University of Dubai, UAE
[⋆]Gnowit Inc., Ottawa, ON, Canada.
Emails: [‡]{sbacanli, turgut}@cs.ucf.edu, [⋆]maloqaily@ieee.org, [†]Rhaldar2@illinois.edu

***Abstract*—The recent emerging fields in data processing and manipulation has facilitated the need for synthetic data generation. This is also valid for mobility encounter dataset generation. Synthetic data generation might be useful to run research-based simulations and also create mobility encounter models. Our approach in this paper is to generate a larger dataset by using a given dataset which includes the clusters of people. Based on the cluster information, we created a framework. Using this framework, we can generate a similar dataset that is statistically similar to the input dataset. We have compared the statistical results of our approach with the real dataset and an encounter mobility model generation technique in the literature. The results showed that the created datasets have similar statistical structure with the given dataset.**

***Index Terms*—IoT, dataset, data processing, synthetic data, mobility**

I. Introduction

Realistic network simulations require the use of an acquired dataset. An acquired dataset may contain the GPS or local positions of the mobile wireless nodes or the encounter durations of the mobile nodes. Unlike the position based datasets, the encounter based datasets simply state which two nodes encountered each other and between which timestamps. Encounter datasets hide the real location information of the nodes. In some cases, the real location of the node might be important in order to make a decision and we would be losing this piece of information. On the other hand, encounter structure contains just the needed piece of information for the mobile wireless network. For instance, node degree, inter-contact times and inter-contact durations are examples for these types of information. Since the dataset syntax is simple, the encounter dataset structure can also be used for other datasets containing social events which has start and finish time besides the actors (i.e. phone calls or continuous messaging).

Although the encounter datasets are easy to understand and simple to work on, acquiring an encounter dataset is not easy and it is mostly a laborious work. The researcher needs to provide wireless devices that can connect to each other and record the connection data. Then, the recruit4ed volunteers would need to carry these devices. After some period of time, the researcher needs to collect all the devices from the volunteers with the hope that the devices did not break down or got lost. After collecting the devices, the connectivity data from each device should be extracted and inconsistent data should be removed from the dataset. The network's representation level increases when the number of devices in the dataset increases. In this case, the researcher should have as many devices as possible resulting in additional expensive.

We hoped to generate a synthetic encounter dataset that has the same statistical properties with the input datasets in terms of distribution of the inter-contact times and contact durations. The university campus human mobility trace datasets used in the literature contains participants from different areas. The dataset collected at University of Milano [1] contains traces from graduate students, faculty staff, and technical personnel, totaling 44 participants. University of St Andrews [2] and University of Cambridge [3] datasets contain traces of graduate and undergraduate students. These datasets are also relatively small in terms of encounter records and participant size.

The mobility patterns of different types of participants may not be similar. Faculty personnel may not go to as many buildings or classrooms or as frequently as the students. Graduate students might be going to one or two research centers and may not be taking as many classes as undergraduates. Creating a mobility model generally may prevent researchers from understanding the differences these groups of people. It is still hard to create encounter mobility models of the stated groups as most of the datasets are still small to see the statistical distributions of encounter times and durations. By dividing these datasets into various groups, it becomes even harder to get a statistical distribution as the data size decreases substantially.

One of the most challenging part of creating an encounter based mobility model is that no detailed information about participants are stated either in the description or in the papers that used a specific dataset. The number of graduate students or staff forming the dataset is stated in the dataset description but examined datasets did not provide such information. The authors of this paper have contacted the faculty members who have provided the University of Milano dataset and acquired the information about which nodes are graduate student, academical or technical staff. MIT Reality dataset contains the participant information and answers of survey data for each participant; however, the dataset format is unlike the datasets from the University of Milano [1] or University of St Andrews [2].

This becomes even more relevant when considering routing algorithms in Delay-Tolerant networks. For instance, social network aware routing [4] has been suggested as a viable alternative to traditional DTN-routing protocols sun as epidemic routing when network resources are limited. In social network aware routing, the nodes are socially connected and are clustered into communities.

The larger dataset can be used in Internet of Things (IoT) related applications. The opportunistic network routing algorithm can use the internet and ad hoc communication together. Another application case might be that the nodes may connect to internet to download some specific information about the encountered node. Based on the downloaded information, the node may decide to send all the messages or some of them to the encountered nodes. Using larger dataset that has similar statistical properties with the current dataset will be useful for testing IoT related opportunistic network routing algorithms.

The created datasets from this encounter mobility model can be used in network simulations while testing the routing techniques. In addition, understanding the human mobility may help researchers and entrepreneurs to use this information for new applications or messaging systems. The metrics about the dataset may give information about the network, therefore the structure of specific community in that environment. If we can compare mobility models of different communities on different environments, we can learn additional interesting information about the environment (e.g., people in the university campuses stop and talk to each other more frequently than urban areas).

II. Related Work

There exists research on simulating encounters between mobile radios. EMO [5] characterizes the node encounters using parameters such as inter-contact time, encounter durations, and the number of nodes encountered by a node. Using these parameters, it simulates encounters. However, it does not employ any clustering on the nodes and considers each node individually. This can be problematic if there are not sufficient number of encounters for each node, so it would be harder to synthetically generate new encounters.

Karagiannis et al. [6] have examined the fundamental properties that determine the basic performance metrics for opportunistic communications. They have found, as an invariant property that there is a characteristic time, order of half a day, beyond which the distribution decays exponentially. Up to this value, the distribution in many cases follows a power law, as shown in recent work. This power law finding was previously used to support the hypothesis that inter-contact time has a power law tail, and that common mobility models are not adequate. More specifically, the CCDF of inter-contact time between the two nodes closely follows a power law decay up to a characteristic time; however, the decay becomes exponential beyond this.

Conan et al. [7] studied the distributions of inter-contact time and contact durations and concluded that each pair of nodes in a delay tolerant network does not share the same inter-contact time distribution as every other node pair. They provided the first formal analysis of the impact of heterogeneous exponential inter-contact time distributions on simple single-copy routing schemes. We have taken into account this aspect by using community detection and clustering.

There has been previous work on analyzing the mobility models on routing performance benchmarks. Thakur et al. [8] found that the performance of the mobility models is not analogous to realistic trends. There were dramatic deviations from realism indicating serious flaws in existing models and their inadequacy as testbed tools for any type of performance evaluation purposes. Riascos and Mateos [9] have analyzed the human urban mobility in New York and Tokyo based on the data harvested from Foursquare application.

Many studies have been performed to specify the nature of the inter-contact time distribution. Jahromi et al. [10] observed that aggregate contact times are best fitted with Pareto distributions, and aggregate inter-contact times followed a power law distribution to a certain point and after that followed an exponential distribution. They concluded that contact times and inter-contact times do not strictly follow a power law distribution.

There are several advantages of social-based routing protocols, by identifying socially similar nodes, and utilizing the additional context information such as shared interests or community affiliations. Schurgot et al. [11] presented the evolution of DTN routing protocols and expressed that solutions using the social graph form a new class of routing protocols well suited for the opportunistic networks.

Bulut et al. [12] noted that only few of the proposed delay tolerant network routing algorithms take into account the effect of social structure of the network on the design of the routing algorithm. It is noted in many studies that the movement of nodes in a mobile network and the interactions between these nodes are not purely random and homogeneous but rather present a mixture of homogeneous and heterogeneous behaviors. In a real mobile network, we always see grouping of nodes into communities such that the nodes within the same community behave similarly and the nodes from different communities show different behaviors. They demonstrate that the consideration of the underlying social structure of a delay tolerant network can help in designing better routing algorithms. In fact, analyzing the community structure of these networks has improved both the message delay as well as the number of message copies formed in multi-copy based routing algorithms.

Through comparing three different kinds of routing strategies [13], the results show that the performance of routing algorithms can be improved by leveraging information obtained from social properties of the nodes. Encounter datasets are used to propose efficient routing techniques. Bacanli et al. [14] proposed State-based Campus Routing (SCR) strategy by using

University of Milano encounter dataset. Xu et al. [15] have used the UAV to collect the information from the clusterhead nodes on the ground for animal monitoring application.

There have been attempts at creating routing protocols that utilize social network properties. One such example is SimBetAge [16], a routing algorithm that takes into account temporal changes in a social network. This demonstrates that respecting the age of the node relations leads to a better performance than a binary representation.

III. The Proposed Approach

In our approach, we need to note that we aim to increase the dataset size. We create a dataset that has the same statistical properties as the input datasets. In order to create a larger dataset based on a given dataset, we need two input parameters beside the input dataset.

The number of clusters, the the number of nodes in each cluster, as well as the cluster association of all the nodes must be known in a given dataset. As an example, considering that the current dataset has only 15, 10, and 10 nodes in each of the corresponding three clusters, the user may wish to create a dataset that has 100 nodes in each cluster. This means that the user can request to have any arbitrary number of nodes in the generated dataset. In addition to these input values, we also need to know when to start and stop modeling the environment.

A. Dataset

We have used the Milano dataset, obtained from the CRAWDAD [1]. The encounter mobility trace of 44 participants are collected at University of Milano campus. The dataset contains encounter traces of graduate students, faculty, and technical staff. It contains a large dataset of contacts within a university campus, which has been chosen as a working context representing many other places where people usually move and meet in their professional/social life. Each node in the network is represented as a graduate student, faculty, or technical staff.

B. Parameters

For each encounter, we store the participants, duration and the inter-contact time. The inter-contact time is the duration between the end of this encounter and the beginning of the next one involving the same participants. There is no inter-contact time with respect to the last encounter.

C. Procedure

1) Preprocessing: To analyze the dataset, we first converted the data into a social network, where each node is a participant, and the weighted edges contain the number of contacts between the two nodes. The social network is then clustered into three categories: graduate student, faculty, or technical staff. For each pair of clusters, we store a list of the duration and inter-contact time of all encounters taking place between the nodes belonging to those two clusters.

2) Creation of synthetic network: We generate a synthetic network that takes the total number of nodes and the total duration as the input. It randomly assigns nodes to the three clusters with the same ratio as the number of nodes in the clusters of the original dataset.

After the creation of the nodes, we assign each of them a weight by sampling from the degree distribution of the nodes in the Milano dataset. We then randomly assign edges between each pair of clusters in the network. The number of edges between any two clusters is chosen such a way that the density of connections between the two clusters in the synthetic network follows the same trend as in the Milano dataset. Here, the density of edges is defined as the ratio of the number of edges between all the node pairs in the clusters and the total number of edges possible between the two clusters. The pair of nodes in which an edge exists is chosen randomly using the weights assigned above.

3) Simulation of encounters: After the nodes and the edges of the synthetic network has been created, we begin to simulate encounters. Between each node pair, we simulate the individual encounters. For each encounter, we assign a duration and an inter-contact time. These values are sampled from the distribution of duration and inter-contact time among all encounters between the two clusters the nodes belong to. We accomplish this using inverse transform sampling. Here, we are assuming that the distribution of the duration and the inter-contact time of all encounters between the two clusters are homogeneous. This process continues for each node pair until the time period exceeds the given duration input.

4) Creation of Synthetic Dataset: We then convert this synthetic data into an encounter dataset that has the same format as the input dataset. Each encounter is a tuple containing four values: the first node, the second node, the start time, and the end time of the encounter. The first node is chosen randomly by choosing a cluster. Afterwards, based on the density of the cluster, a random node number in that cluster is selected. Then, based on the distribution of the inter-encounter times of the two clusters of the first and second node, an inter-contact time is chosen as start of the contact time. Based on the distribution of the contact durations between the two clusters, a number is picked from the distribution as the contact duration. While selecting the next encounters, the inter-encounter time is chosen and contact start time is calculated by adding the inter-contact time with the finishing time of the last successful encounter duration. Since the encounter times and durations are created contiguously, the data will be sorted in order of start times of the contacts.

As it is shown in the algorithm 1, the ICDBetween and ICTBetween functions are taking the two clusters as parameters. These functions return a random number based on the distribution of inter-contact times and inter-encounter times between the two cluster inputs. At the end, the algorithm returns a set of encounter entries. If these entries are sorted based on their contact start times, the dataset would be usable by simulators and also it would be easier to analyze and follow chronologically.

We need to note that, while creating a synthetic dataset, we have used exactly the same parameters of the real dataset. In other words, the number of nodes in each cluster for the synthetic dataset is the same as the number of nodes in each cluster within the real dataset. The maximum data collection time also matches the duration of the real dataset.

IV. Results

We have compared the statistical distributions of inter-contact time and contact durations of our generated datasets. The datasets generated by EMO [5] using Milano dataset and the real Milano dataset.

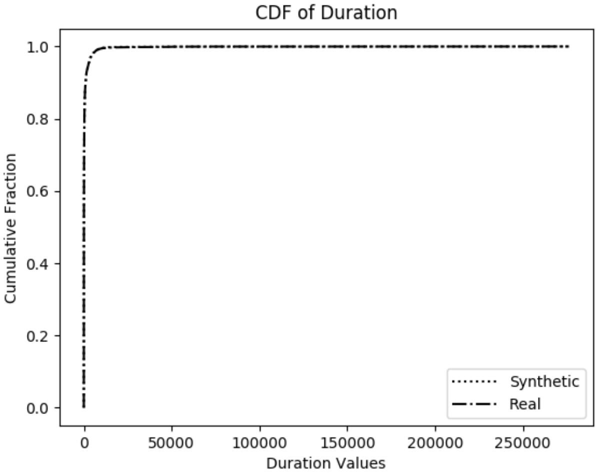

Fig. 1. CDF of Inter Contact Durations of first 250000 seconds

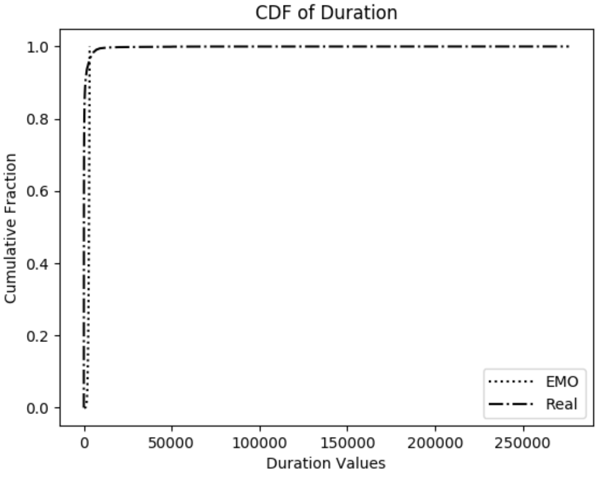

Fig. 2. CDF of Inter Contact Durations of first 250000 seconds

Figure 1 and Figure 2 show the cumulative distribution functions of inter-contact durations comparing our synthetic model and EMO. Both figures show the durations of first 250000 seconds. Our synthetic models seem to be fitting better than EMO. We only showed the results showing less than 250000 seconds as the difference is most visible when the graphs are showing results less than 250000 seconds.

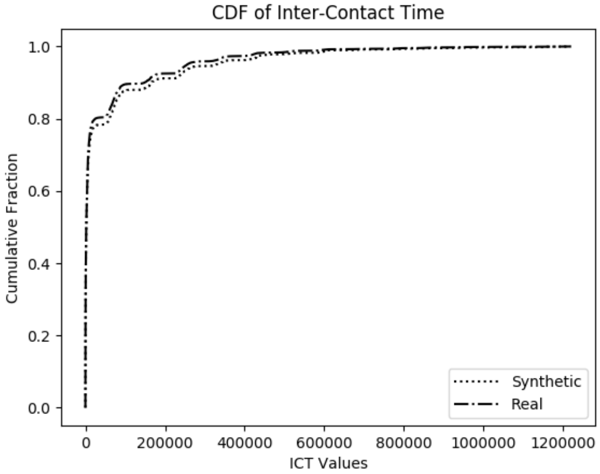

Fig. 3. CDF of Inter-contact Times of first 1200000 seconds

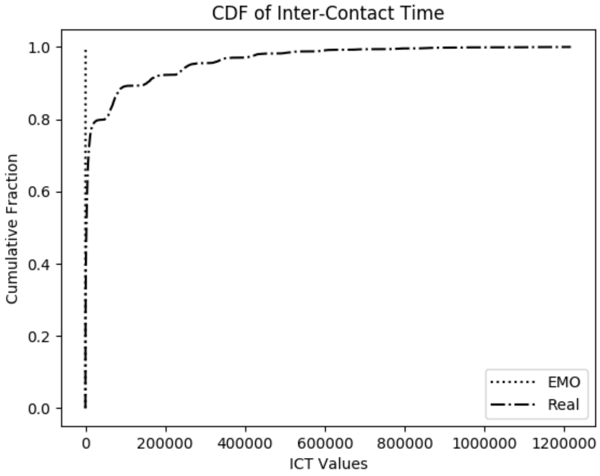

Fig. 4. CDF of Inter-contact Times of first 1200000 seconds

EMO model is developed mainly to create a general encounter mobility model. EMO assumes that all encounter times and duration patterns are homogeneous and it does not take into account any cluster or category of people. Our model on the other hand is aware of the clusters of people in a dataset, achieving improved results. The statistical distribution results were as expected.

V. Conclusion

We have modelled the mobility of graduate students, faculty, and technical staff on a university campus. By doing that we

Algorithm 1 Dataset Enlarger

```
Input: $inputdata$:Encounter dataset $ClusterNodes$:list $MaxTime$:number
Output: Generated Encounter Dataset as list
function EnlargeDataset( $datainput, ClusterNodes, maxtime$ )
    $GeneratedEncounters \leftarrow \{\}$
    ▷ ClusterNodes is an array. Length of ClusterNodes is number of clusters.
    ▷ Each element contains the number of nodes in that cluster
    ▷ MaxTime is maximum number of seconds
    for $i \leftarrow 1$ to $length(ClusterNodes)$ do
        $numberOfNodesInCluster \leftarrow ClusterNodes_i$
        for $node1 \leftarrow 1$ to $numberOfNodesInCluster$ do
            $node2Cluster \leftarrow$ random number between [0, length of $ClusterNodes$)
            $node2 \leftarrow$ random number between [0, $(ClusterNodes_{node2Cluster-1})$)
            $contactEnd \leftarrow 0$
            while $contactEnd < MaxTime$ do
                $interContactDuration \leftarrow ICDBetween(node1Cluster, node2Cluster, inputdata)$
                $interContactTime \leftarrow ICTBetween(node1Cluster, node2Cluster, inputdata)$
                ▷ Curve fitting is done while getting ICT and ICD between clusters
                $contactStart \leftarrow contactEnd + interContactTime$
                $contactEnd \leftarrow interContactTime + interContactDuration$
                $GeneratedEncounters \leftarrow GeneratedEncounters \cup \{node1, node2, contactStart, contactEnd\}$
            end while
        end for
    end for
    return $GeneratedEncounters$                    ▷ The result can be sorted by contactStart times
end function
```

have created a framework where we can generate synthetic dataset for encounters. The created mobility model is compared with EMO [5] and the real dataset. The results show that the mobility model gives similar statistical distributions with real dataset and show more realistic results than EMO [5].

This framework can be potentially used by the researchers to predict the next inter-contact time for a person on campus or average contact duration for that person. The datasets that contain encounter mobility in the literature are either recorded for short amount of time or the number and variety of the participants are a few. This framework also allows researchers to generate synthetic encounter mobility dataset for use in network analysis or routing protocols. For instance, the simulations on larger synthetic dataset can be leveraged to assess the efficiency of the opportunistic routing techniques.

References

[1] P. Meroni, S. Gaito, E. Pagani, and G. P. Rossi, "CRAWDAD dataset unimi/pmtr (v. 2008-12-01)," Downloaded from http://crawdad.org/unimi/pmtr/20081201, Dec. 2008.

[2] I. Parris, T. Henderson, and F. B. Abdesslem, "CRAWDAD dataset st_andrews/locshare (v. 2011-10-12)," Downloaded from https://crawdad.org/st_andrews/locshare/20111012, Oct. 2011.

[3] J. Scott, R. Gass, J. Crowcroft, P. Hui, C. Diot, and A. Chaintreau, "CRAWDAD dataset cambridge/haggle (v. 2009-05-29)," Downloaded from https://crawdad.org/cambridge/haggle/20090529, May 2009.

[4] R. Misra and S. Shukla, *Social Network Aware Routing for Delay Tolerant Networks*. Springer Berlin Heidelberg, 2011, pp. 101–110.

[5] F. Tan, Y. Borghol, and S. Ardon, "EMO: A statistical encounter-based mobility model for simulating delay tolerant networks," in *WoWMoM*, June 2008, pp. 1–8.

[6] T. Karagiannis, J.-Y. Le Boudec, and M. Vojnovi, "Power law and exponential decay of inter contact times between mobile devices." *ACM Mobicom*, p. 183, 2007.

[7] V. Conan, J. Leguay, and T. Friedman, "The heterogeneity of inter-contact time distributions: its importance for routing in delay tolerant networks," 2006.

[8] G. S. Thakur, U. Kumar, A. Helmy, and W. Hsu, "Analysis of spatio-temporal preferences and encounter statistics for DTN performance," *CoRR*, vol. abs/1007.0960, 2010.

[9] A. P. Riascos and J. L. Mateos, "Emergence of encounter networks due to human mobility," *PLOS ONE*, vol. 12, 10 2017.

[10] K. K. Jahromi, F. Meneses, and A. Moreira, "Analyzing temporal scale behaviour of connectivity properties of node encounters." *Procedia Computer Science*, vol. 40, no. MoWNet, pp. 57 – 65, 2014.

[11] M. R. Schurgot, C. Comaniciu, and K. Jaffres-Runser, "Beyond traditional DTN routing: social networks for opportunistic communication," *IEEE Communications Magazine*, vol. 50, no. 7, pp. 155–162, July 2012.

[12] E. Bulut, Z. Wang, and B. K. Szymanski, "Impact of social networks on delay tolerant routing," in *IEEE GLOBECOM*, Nov 2009, pp. 1–6.

[13] H. Gong and L. Yu, "Study on routing protocols for delay tolerant mobile networks," *International Journal of Distributed Sensor Networks*, vol. 9, no. 1, 2013.

[14] S. Bacanli, G. Solmaz, and D. Turgut, "Opportunistic message broadcasting in campus environments," in *IEEE GLOBECOM*, December 2015, pp. 1–6.

[15] J. Xu, G. Solmaz, R. Rahmatizadeh, D. Turgut, and L. Bölöni, "Animal monitoring with unmanned aerial vehicle-aided wireless sensor networks," in *IEEE LCN*, October 2015, pp. 125–132.

[16] J. A. B. Link, N. Viol, A. Goliath, and K. Wehrle, "Simbetage: Utilizing temporal changes in social networks for delay/disconnection tolerant networking," in *ACM MobiQuitous*, July 2009, pp. 1–2.