Towards Human-Aware D2D Communication

Rafael Lima Costa, Aline Carneiro Viana, Artur Ziviani, Leobino Nascimento Sampaio

Towards Human-Aware D2D Communication

Rafael Lima Costa*†, Aline Carneiro Viana‡, Artur Ziviani§, Leobino Nascimento Sampaio*
*Computer Science Department, Federal University of Bahia, Salvador, Brazil
†École Polytechnique, Université Paris-Saclay, Palaiseau, France
‡Inria, Palaiseau, France
§National Laboratory for Scientific Computing, LNCC, Petrópolis, Brazil
{rlimacosta,leobino}@ufba.br, aline.viana@inria.fr, ziviani@lncc.br

***Abstract*—Mobility, social interactions, and other human characteristics shall support Future Mobile Networks in routine prediction and resource management. This work investigates human-aware metrics supporting services or protocols leveraging opportunistic communication. These metrics represent different types of knowledge extracted from people routine present in their movements. Because of the strong routine component of human mobility, such metrics capture different but recurrent behaviors on wireless encounters between mobile users. We report the experience through a case study with a real-world dataset along with results from trace and metrics analysis. The results show heterogeneity in metric coefficients and contact occurrence and duration in different periods of the day, highlighting the need for characterising traces before their use.**

***Index Terms*—Human-centered computing, social computing, data management systems, network architectures**

I. Introduction

Connectivity intermittence, the strong dependence on user mobility patterns, and devices with limited resources are characteristics that make research on mobile networks quite challenging. Studies on the effectiveness and performance of network solutions for mobility scenarios, in general, are supported by simulation-based experiments relying on mobility models for the generation of synthetic traces. More recently, researchers have made efforts to use real datasets (also referred to as traces) for obtaining more robust and realistic results.

Using real datasets has become even more relevant for studying mobile network architectures of the future, which will be increasingly focused on the behavioral requirements of users [1]. In other words, exploring architectures in which characteristics and habits from the human behind a communication device are taken into account to serve him better. For these reasons, the study of human behavior [2] is essential for research development in the area. An individual has characteristics such as mobility, personality, and socioeconomic traits, social interactions, character, humor, traffic profile, and context that can be studied to offer a more suitable network service [3]. Historically, several studies have examined user characteristics such as mobility [4], [5]. However, evaluation metrics and other factors must be closer to human behavior inherent aspects. The use of this information in prediction and for identifying routines more precisely also needs initiatives.

The study of human-aware solutions through real datasets is essential, given the need to bring networks and their users closer. In addition to the availability of these real traces, several challenges need to be addressed, such as temporal gaps and inconsistent data caused by data collection errors. Criteria related to processing, modeling, knowledge extraction, and analysis need to be taken into account until the raw data becomes useful for research. In this paper, we apply methods in each step mentioned to extract human-aware mobility metrics. This work aims to answer *"What type of knowledge extracted from human mobility routines can be useful to services or protocols leveraging opportunistic communication?"*

We discuss user data extraction and manipulation, including steps, and techniques for solving problems. Through knowledge extracted from MACACO[1] (private European Dataset), we discuss human-aware metrics to support opportunistic communication. This dataset is obtained through an Android mobile phone application. This app collects the data related to the user's digital activities, such as generated uplink/downlink traffic, available network connectivity, and visited GPS locations. We present a case study from the trace manipulation, including our methodology applied for users-selection, home and workplace inference, error filtering, and for filling time gaps. Finally, we bring results from dataset manipulation and metrics analysis, and discuss metrics application.

We organized the rest of this work as follows: in Section II-A, related work; in Section II-B we discuss the framework for extracting human-context data linking to our metrics; in Section III, we propose the metrics and a human-aware time division; in Section IV we bring the case study with MACACO, our results analysis, and discussions; in Section V we discuss metrics applications; in Section VI, conclusions, future objectives, and research opportunities.

II. Rationale

In this Section, we review related work and discuss a framework for datasets manipulation and knowledge extraction.

A. Related Work

Over the years, different works extracted user mobility metrics to assist routing algorithms and forwarding strategies. In [6], authors introduced BUBBLE Rap, an algorithm exploiting centrality, and community detection for delay-tolerant networks forwarding. The idea came from the intuition that

[1] MACACO Dataset. Available at: http://macaco.inria.fr/

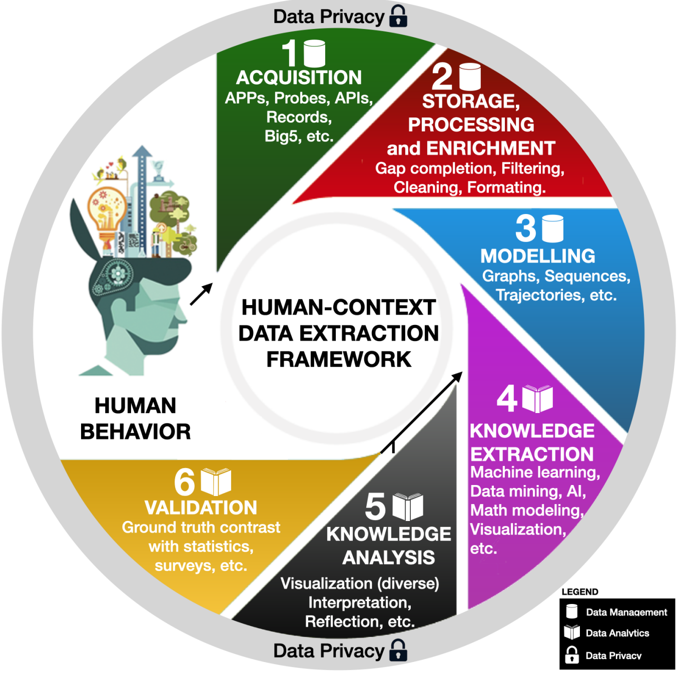

Data Privacy
1 ACQUISITION
APPs, Probes, APIs, Records, Big5, etc.
2 STORAGE, PROCESSING and ENRICHMENT
Gap completion, Filtering, Cleaning, Formating.
3 MODELLING
Graphs, Sequences, Trajectories, etc.
4 KNOWLEDGE EXTRACTION
Machine learning, Data mining, AI, Math modeling, Visualization, etc.
5 KNOWLEDGE ANALYSIS
Visualization (diverse) Interpretation, Reflection, etc.
6 VALIDATION
Ground truth contrast with statistics, surveys, etc.
HUMAN BEHAVIOR
HUMAN-CONTEXT DATA EXTRACTION FRAMEWORK
Data Privacy
LEGEND
Data Management
Data Analytics
Data Privacy

Fig. 1. Framewok for extracting human-context raw data to assist future networking solutions.

individuals of the same community are more likely to interact than with a randomly chosen user. BUBBLE Rap achieved a similar delivery ratio, but much lower resource utilization than flooding, control flooding, PROPHET, and Simbet. In [7], authors apply mobile users' trajectories, scenario interactions, and traffic demands in the WiFi hotspot deployment problem. Their results show a higher offload ratio with reduced hotspots number in comparison with a state-of-art solution. In [8], authors introduced social group meetings (instead of static social communities from BUBBLE Rap), defined as a group of people together, in space and time, for some social reason. Their algorithm, GROUPS-NET, forwards the messages through the most probable group-to-group path. In large-scale scenarios, they achieved approximately the same delivery ratio of Bubble Rap, but with 40% less overhead. In [9], the authors developed SAMPLER, an opportunistic forwarding strategy combining mobility, PoIs, and social-awareness. Their experiments showed improved delivery ratio, reduced overhead, and faster message delivery.

From the related initiatives herein described, we consider that understanding human-behavior and context from mobility is important for networking solutions. Different from such studies, here we discuss metrics with a more granular link with time. We focused on identifying different human activities and context, linked to the time of the day. Through our metrics and dataset analysis, we found that they vary according to the moment observed. Trace analysis and characterization are essential to understand the population studied and to propose proper solutions for each scenario. Metrics and insights based on generic aspects might not apply to all populations. Facets such as node density, proximity, transportation mode, and even cultural factors can change contact dynamics. Finally, from these traces characterizations, we believe in the possibility of identifying novel mobility metrics for different applications.

B. Datasets Manipulation and Knowledge Extraction

The human being is complex, and so identifying his behavior, routines, and other traits hidden into his mobility requires a series of procedures. In this context, we discuss the framework described in Figure 1 for extracting and manipulating human-context data. This framework was essential for identifying features of the metrics introduced in this work.

1) Data Management: Features (i) Acquisition, (ii) Storage, Processing and Enrichment, and (iii) Modeling:

- **Acquisition**: Human-behavior analysis requires data availability, usually from different sources. Among the options, collecting from wireless networks measurement infrastructures, specific service APIs, or social networks (e.g., Foursquare) and mobile devices. The last (also applied in MACACO) is one of the primary sources, given their ubiquity and diversity of equipped sensors.
- **Storage, Processing, and Enrichment**: The possible large amount of generated data requires secure, scalable, and fault-tolerant storage platforms that enable parallel, real-time requests. In processing, data association, and integration may also be necessary: several data sources grouping different data types are explored simultaneously to extract useful information. Besides, cleaning techniques, enrichment by normalization, false entries detection, geographic and temporal interpolation, among others, must be applied to compensate gaps and reduce raw data inconsistency [10]. In Section IV-A, we apply some of these procedures. Finally, dimensional reduction of multidimensional data may also be necessary before analysis. For this, selecting useful resources is valid.
- **Modeling**: Data modeling format must allow extracting space-time information and different components relationship. Graphs have been the most used format to model social ties or people's spatiotemporal behaviors within environments. In this context, the vertex in a graph can represent users in a network or their visited locations [9]. The edges connect vertices when a meeting occurs, or when locations are visited sequentially by an individual. Besides graphs, using space-time trajectories of chronologically ordered points are habitual. Section IV-B describes our modeling process.

2) Data Analysis: Features extraction and knowledge analysis, as well as data validation:

- **Extraction and knowledge analysis**: Among the types of knowledge extraction, we can mention: detection of maximum and average user displacement, mobility metrics calculation (e.g., centrality), patterns detection and modeling, correlation, and causality between involved entities, detection of behavioral profiles, classification or data grouping, and detection of changes or irregularities in the data. Knowledge extraction from user activity on social networks is common, as their posts may contain valuable data. Visualization techniques, machine learning, artificial intelligence, human-computer interaction methods, time series modeling, sophisticated network metrics, statistics, and empirical analysis are useful examples. Sections III and IV-C feature our knowledge extractions and analysis.

- **Validation**: Validation consists of verifying data correctness and usefulness, providing guarantees of adequacy, accuracy, and consistency. In statistical analysis, cross-validation is a technique used to assess how the results will be generalized to an independent dataset. Another common technique is crossing the used data (usually incomplete or reduced) with what is called *ground truth* data (usually official or complete data).

C. Data Privacy

Despite not being the focus of this work, user-data privacy is a must to support applications and innovation, without jeopardizing individual rights and security [11]. We argue that using appropriate techniques to guarantee user privacy according to the data type is necessary. Below, we show metrics extracted from user mobility and a human-aware time division.

III. Metrics for Human-Aware Applications

This Section presents metrics to extract and to interpret different routine features from human-mobility behavior.

A. Identifying routines from human mobility

According to other works, such as [12], the existing regularity in human mobility contributes to the low entropy of these movements. As a result, aspects such as spatial, temporal, and social regularity (meetings), have become essential in mobility studies. Due to our routines, it is possible to predict meetings, points of interest, and displacements[2]. Those factors can leverage network resource management and communication types (e.g., device-to-device). However, depending on the time of day, our mobility profile generally changes (e.g., restricted when we are at home, and more mobile when traveling from home to work). Therefore, we propose here a more granular temporal observation. This idea is opposed to previous initiatives in which they calculate mobility metrics from constant time windows (e.g., 24 hours or the previous 6 hours observed) [13]. In Table I, we present our temporal approach, dividing the day into six non-uniform periods, with different durations. We emphasize that this was a division adopted in MACACO to reflect different user mobility-profiles throughout the day. We argue that temporality must fit the studied population to bring more accurate results in research.

We justify our proposal, relating the periods to the instants the population travels and has longer confinements. The "EM" period, for example, accounts for most individuals displacement from home to work, while in the "M" the confinement at work (also in the "A" period) and smaller displacements during lunchtime. With this temporal division, the intention is to extract from the mobility traces, metrics, and decision factors increasingly accurate and closer to the real activities and human routines in the appropriate periods of the day. Reshaping these periods is possible whenever necessary. Below, we present metrics extracted from the MACACO Dataset, but which can be obtained or used through other mobility traces.

TABLE I
TEMPORAL DIVISION PROPOSED

#	Period	Time Interval
EM	Early Morning	06:00:00 - 09:59:59
M	Morning	10:00:00 - 13:59:59
A	Afternoon	14:00:00 - 17:59:59
EE	Early Evening	18:00:00 - 20:59:59
E	Evening	21:00:00 - 23:59:59
N	Night	00:00:00 - 05:59:59

B. Social-Awareness as Centrality Degree

The Centrality Degree (CD) measures the social bonds of a user (human), that is, his encounters. Someone with a higher degree is more "popular" (he has further encounters).

We calculate metric coefficients by a period of the day and learn from user mobility on a week k to apply this knowledge in week $k+1$. Therefore, Equation 1 represents the average *CD* of a node u in a period $p \in (EM, M, A, EE, E, N)$ as $(\Delta_{CD_p(u)})$. The sum of $i = 1$ to d represents the consecutive days of the week k and d the number of previous days considered (5, excluding weekends). Further, n is the number of network nodes, and $e_{(u,v)}$ is an index of value one if there is an edge e between u and v nodes in the period p. Considering our network as a dynamic contact graph $G_t = (V, E_t)$, where V is the set of users (mobile nodes), and E_t is the set of users edges (contacts) detected, where $t \in (1, 2, ..., a)$ and $a \leq 432,000$ seconds (working days duration in seconds of the week k). There is an edge $e \in E_t$ between two nodes if they are in the communication range defined at the time t, that is, they are in contact. In this work, we define our communication range as 30 meters (average range of WiFi Direct). This range ensures good certainty in the contact probability. Any range bigger than 30 meters would increase the uncertainty of the contact probability. On the other hand, although we believe lower ranges than 30 meters will not change such contact certainty, we plan to examine smaller ranges in future works.

$$\Delta_{CD_p(u)} = \frac{\sum_{i=1}^{d} CD_p^i(u)}{d}, where\ CD_p(u) = \frac{\sum_{v=1}^{n} e_{(u,v)}}{n-1} \tag{1}$$

C. Coverage Area as Radius of Gyration

The Radius of Gyration (RG), quantifies an individual's mobility related to a center of mass, calculated from his movements. In our strategy, this metric is for selecting users

[2] Facebook Filed A Patent To Calculate Your Future Location. https://www.buzzfeednews.com/article/nicolenguyen/facebook-location-data-prediction-patent

who have made the most displacements within a network cell (details in the next Section). Here we also do the learning per period $p \in (EM, M, A, EE, E, N)$ of a week k to apply the knowledge in a week $k+1$. Therefore, in Equation 2, we calculate the average *RG* of a node u in a period p as $\Delta_{RG_p}(u)$. The sum from $i = 1$ to d, represents the days of the week k, where d is the number of previous days considered (5). Further, N is the number of positions (coordinates) recorded, l_j is a location at index j, and l_{cm} is the center of mass. For each user $u \in V$, we consider that there is a set $L_p = (l_1, l_2, ..., l_n)$ of places found in the period p of the week k. Each location $l = (x, y)$, where x, y are coordinates associated with a time instant. With this approach, we seek a more precise metric, reflecting the movements in each period of the day.

$$\Delta_{RG_p}(u) = \frac{\sum_{i=1}^{d} \; RG_p^i(u)}{d}, \; whereRG_p(u) = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(l_j - l_{cm})^2} \; and \; l_{cm} = \frac{1}{N}\sum_{j=1}^{N} l_j \tag{2}$$

D. Sojourn Time

With the *Sojourn Time* (ST), we quantify a user's stay in a network cell. In our strategy, this metric aims to identify nodes that remain in the same cell as a potential user interested in some content (more details in the next Section). We consider a geographical space divided into different cells of an operator. We investigate the mobility of each user $u \in V$ in a week k, calculating his length of stay per period p in each cell $c \in (c_1, c_2, ..., c_n)$ to apply in the week $k+1$. Therefore, in the equation 3, $\Delta_{ST_p^c}(u)$ is the average *ST* in minutes of u in c during p. Also, $i = 1$ to d represents the days of the week k, and d is equal to the number of previous working days considered (5). The duration (in minutes) comes from the timestamps associated with at least two consecutive pairs of coordinates $(x, y) \in L_p$; and $(x, y) \in c$ (c's geographic domain). $\Delta_t^p(u)_c$ is the total time t in the period p that node u stayed inside cell c.

$$\Delta_{ST_p^c}(u) = \frac{\sum_{i=1}^{d} \; ST_p^c(u)_i}{d}, \; where \; ST_p^c(u) = \Delta_t^p(u)_c \tag{3}$$

E. Destination Proximity as Geographical Awareness

With this new metric of geographic science, we try to answer how close a user gets to a cell (given its coverage and location). As in the previous metrics, we examined user mobility by period p in a week k to apply this knowledge in the following week $(k+1)$. The goal is to find users who can approach the edge or reach the next cell. Thus, in Equation 4, we calculate the average *Maximum Proximity* (MP) as $\Delta_{MP_p^c(u)}$, that is, a proximity coefficient of a user u related to a cell $c \in (c_1, c_2, ..., c_n)$ during a period p. The days of the week k range from $i = 1$ to d, which is equal to the number of previous days considered (5). We apply the traditional geodesic formula to calculate the distance between the cell and node u. Thus, in min, we obtain the shortest distance between u and c, considering all pairs of coordinates of u available in the period p. We emphasize the need for applying a proper distance formula for each dataset format, depending on the coordinates system.

$$\Delta_{MP_p^c(u)} = \frac{\sum_{i=1}^{d} \; MP_p^c(u)_i}{d}, \; where \; MP_p^c(u) = min|geodesic(u, c)| \tag{4}$$

F. Geographic Direction Awareness

This new metric relates to instant user mobility. In it, we extract directions from the recent displacement of a node u to check if his movement direction is to a specific network cell c. For this, we calculate the geodesic direction between pairs of coordinates, considering the last 30 minutes of displacement. The result is the mode (the most common direction from a set). This on-demand metric applies if we want to extract movement direction. Each node checks its new recent direction related to a given cell c and compares its results. Following, we present the MACACO case study, as well as results from the analysis of the dataset, metrics, and time division.

IV. Experiments, Results, and Evaluation

A. Case Study: MACACO Dataset

We evaluate the methodology of this work on a spatiotemporal personal dataset collected through an Android mobile phone application. The application collects the data related to the user's digital activities such as available network connectivity (i.e., Bluetooth, APs, cell towers), battery availability, and visited GPS locations. These activities are logged with a fixed periodicity of 5 minutes for 99% of the measurements. The 1% occasional longer gaps are due to system background jobs and settings of the specific operating system version. We remark that the 5-minute sampling approach differs from those employed by popular GPS tracking projects, such as MIT Reality Mining [14] or GeoLife [15], where users' positions are sometimes irregularly sampled.

Concerning such previous efforts, the regular sampling in MACACO data grants a neater and more comprehensive overview of a user's movement patterns. The data covers about 190 users who live in six countries (including Brazil) spanning from 2014 to 2018. Throughout our experience with the MACACO dataset, we applied several preliminary steps for finally extracting metrics and features from human mobility. Following, we summarize these efforts.

- **Selecting spatial coverage and filtering out "bad" users:** In the processing and enrichment phases, the first challenge was to select a subset of users with common spatial characteristics, since we aimed to analyze their routines and contact dynamics. We identified a population of 62 users (Students, researchers, and administrative staff) in two universities (UFMG and PUC-MG in Brazil) where lectures were held during the day and at night. For the sake of simplicity in the performed spatial analysis, we choose to use here users from the same country, i.e., 62 users. Our selection occurred by mapping user mobility trails to geographical coordinates from the city of Belo Horizonte and the Metropolitan Region. We removed 28 users because they had little (sparse) data or inconsistent information, for example, zeroed GPS measurements, and excessive duplicated entries.

- **Grouping observation periods:** Each user collection period (in which he used the application) is unique. As mentioned, the *trace* has data between 2014 and 2018. We discarded data before 05/22/2015, as the request interval changed from 1 to 5 minutes on that date. Aiming to select four weeks of data (20 working days), we filtered the "best" working days for each user, excluding holidays. This technique gives the chance to compare activities by days of the week since the selected population shares a social context (attends one of the Universities).
- **Identifying important locations:** We inferred a "HOME LOCATION" selecting users who stayed in the city of Belo Horizonte and the Metropolitan region during the night, from *02:00* to *06:00*, that is, when most people are at home. We also checked the data from *10:00* to *21:00* to identify presence on the UFMG or PUC-MG ("WORK LOCATION") campus from morning till night.
- **Filling spatiotemporal Gaps:** During the dataset acquisition, temporal gaps come from disabling the app (intentionally, or by lack of battery). In order to fill these gaps, we performed mobility inference according to different criteria. For those in the period from 02:00 to 06:00, we applied each user "HOME LOCATION" (the most frequent during the 20 days on that time window). We applied linear geodesic interpolation in the remaining gaps throughout the day. Given two pairs of coordinates, each pair associated with a time instant, the intermediate points (latitude and longitude) are calculated recursively.

B. Data Modeling

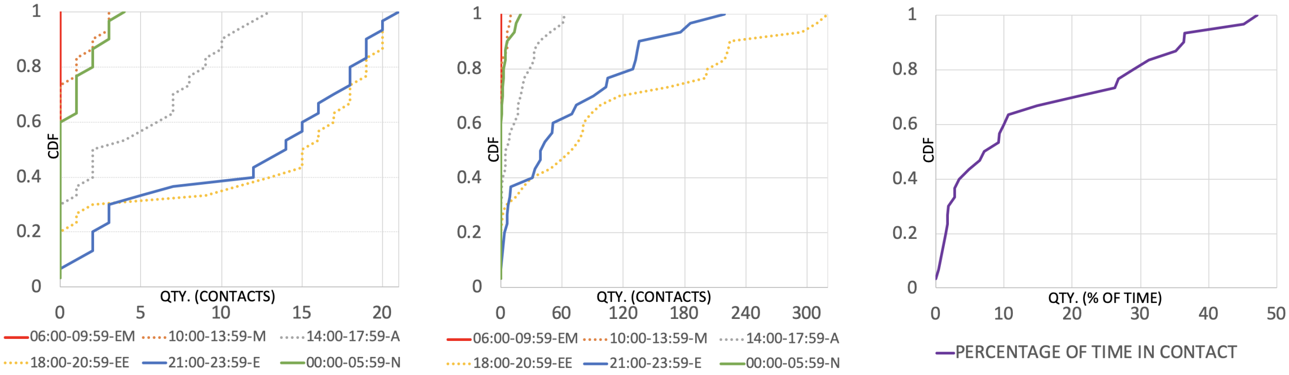
After manipulating the data, we modeled it for our research goals (assisting opportunistic communication services or protocols). Therefore, we built a contact graph to study user behavior along the periods of the day and extracted the metrics introduced in the previous Section. To identify the contacts, we calculate their occurrence and duration through each user location data (coordinates) associated with its timestamps. Based on geodesic calculations, the proximity up to 30m (i.e., the average range of WiFi Direct) is detected. As there is no clock synchronization for each measurement of each node, we applied sliding windows of 5 minutes. For example, the initial window for a user u on a day d is from *00:00:00* to *00:05:00*. If two nodes u and v were in the range within that time frame, and on the same day d, the contact occurred. If they remained in range in the next time window, that is, from *00:05:00* to *00:10:00*, the contact duration increments. As a result of this process, we generated three contact graphs: (i) a graph per period where the vertices are users, and there is an edge between a random user u and v if they were in contact; (ii) a graph per period where the vertices are users, and there is a weighted edge between a random user u and v accounting for the number of times they were in contact; (iii) a graph per period where the vertices are users, and there is a weighted edge between a random user u and v accounting for their total contact duration. From these graphs, we plotted charts with unique and total contacts number, users' percentage of time in contact, average and maximum contact duration, and centrality degree metric (applying the formula).

We extracted the other metrics individually through querying the dataset. For each user u *Radius of Gyration* (per period), we selected his coordinates associated with timestamps belonging to each day period. For the *Sojourn Time*, we did a spatial analysis. Due to the shared social context of the trace, most contacts take place on campus. So we decided to consider only that geographical area to build a cell division. Original cells from local operators are enormous in comparison to the campus area. Therefore, after some experiments with cell sizes, we divided the area into nine cells (each 113m x 113m) ranging from "A" to "I". That was necessary for making valid the metrics relying on geographic science and cell notions. For the *Sojourn Time* (per period) calculation of a user u, we selected his coordinates associated with timestamps, accounting for the amount of time the user spent in each cell geographic domain, according to the defined cell size above mentioned. As for *Destination Proximity*, we first calculated central point coordinates for each cell (ranging from "A" to "I"). Then, for each user u, we selected his coordinates per period and compared it with each cell central-point to obtain his maximum proximity distance concerning the cells. As previously stated in the formula, we applied the geodesic distance calculation. Geographic Direction was not extracted for now, as it refers to instant mobility and will be calculated from the last 30 minutes of user activity, if necessary, according to the strategy.

C. Obtained Results and analysis

In Figure 2(a), we plot a Cumulative Distribution Function (CDF) with the number of unique contacts. The "EM" period has zero contacts (and so removed from some other plots), while the "M" and "N" periods have a shallow contact occurrence. In the same Figure, we see about 70% of users with at least five unique contacts during "EE" and "E". In comparison with Figure 2(b), where we plot the CDF for the total contacts, we see higher coefficients, which is justified by the users' routines (in this case, repeating social links). During "EE", more than 20% of users have 200-320 contacts.

In Figures 2(c) and 3(a), we evaluate user contact duration. In the first, we see about 40% of the trace with less than 5% of the time in contact. The periods of low contact incidence (mentioned earlier) have a significant impact on this graph, so we evaluate this reality per period in Figure 3(a). In this graph, we find higher percentages in the "EE" and "E" periods. In Figures 3(b) and 3(c), we evaluate average and maximum contact duration. We discarded the minimum contact duration

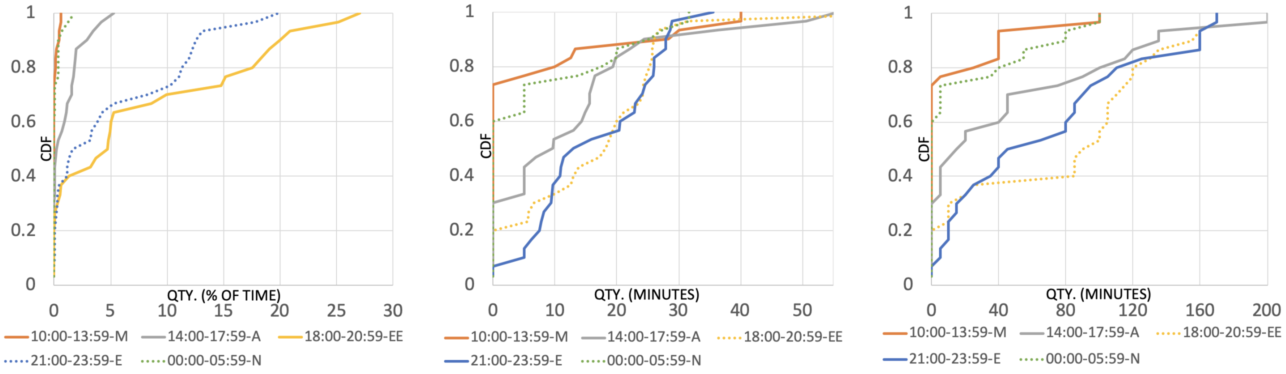


(a) Unique contacts number

(b) Total contacts number

(c) Percentage of time in contact (full trace)

Fig. 2. Contact Analysis of MACACO Dataset.

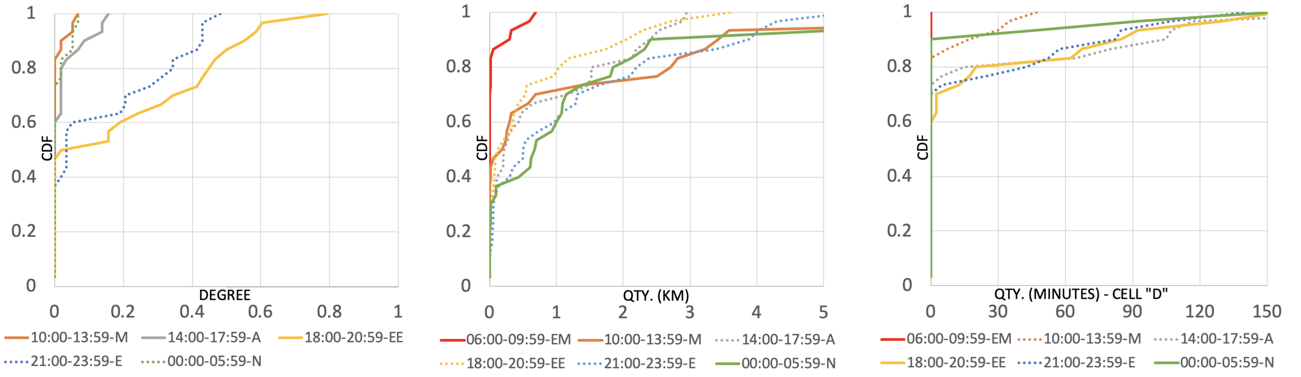


(a) Percentage of time in contact per period

(b) Average contact duration (minutes)

(c) Maximum contact duration (minutes)

Fig. 3. Contact Analysis of MACACO Dataset.



(a) Centrality degree per period

(b) Radius of gyration per period (kilometers)

(c) Sojourn Time in cell "D" per period (minutes)

Fig. 4. Metrics Analysis extracted from MACACO Dataset.

evaluation due to the uncertainty resulting from the lack of clock synchronization. There is a heterogeneity of information in each period, which reinforces our intuition that human beings behave differently according to the context and time of day, therefore directly impacting metrics coefficients. In Figure 3(b), we find almost 80% of users without contacts in the "M" period. There are only about 10% of contacts with a longer duration (15 to 40 minutes). In the "A" period, the duration increases. About 70% of users have contacts. "A" and "EE" periods are interesting to study metrics related to geographic science, as people make higher displacements when they return home or go to University, for example. In the "EE" period, about 80% of users get in contact, with about 50% of contacts lasting at least 20 minutes. The same applies to the next period ("E"), with a higher occurrence of contact number and duration of 5 to 200 minutes (figure 3(c)(b)). In the "N" period, there are fewer contacts incidence and shorter duration, since most people tend to be at home and do not share a social context at this time. This finding would be different in a trace where users live on the same campus, for example. Our claim reinforces here: knowing more about the humans behind the devices can lead to more proper solutions.

In addition to the dataset analysis, we extracted our metrics from MACACO to evaluate their distribution. In Figure 4(a), we see how the user's "popularity" correlates with time. In Figure 4(b), the radius of gyration (in meters) distribution shows more significant displacements in the "M" and "A" periods. As in the previous analysis, the coefficients differ per period, following our intuition. Next, in Figures 4(c) and 5, we plot the CDF of the metrics *Sojourn Time* and *Destination Proximity*, where we randomly use a cell "D" (ranging from "A" to "I"). In the first Figure, more nodes stay in cell "D" during "A", "EE", and "E". In Figure 5, during "EE" and "E", about 40% of users are inside or close to the cell. Proximity can bring interesting insights when aiming to reach a neighboring cell. Throughout the results herein presented, in the following Section, we highlight our main findings as a discussion summary of our results.

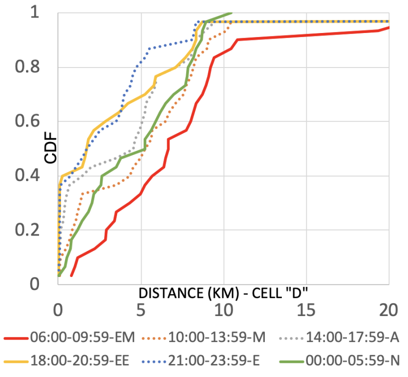


Fig. 5. Destination Proximity to cell "D" per period (kilometers).

D. Discussion Summary

Judging from the contact (unique, total, and percentage of the time) analysis, we found that the "EE" and "E" periods are the best to evaluate opportunistic communication services or protocols through the MACACO dataset. In those periods, we see not only a higher number of unique contacts but more repeating social links, with a higher duration. On the other hand, the lack of contacts during "EM", and the low incidence in the "N", and "M" periods could result in higher delivery delays, and depending on message TTL, reduced delivery rate. Suppose a message creation happens during "E", but finding the destination node opportunistically is not possible the same day. Due to zero or very low contact-incidence in the upcoming periods, the delivery chance reduces. Further, due to MACACO's amount of users, we must use other datasets to evaluate larger scenarios.

Regarding contact duration, we find a high percentage (about 80% in "EE", and 90% in "E") of users with average contact persistence of at least five minutes. We consider it as an interesting finding for transferring data opportunistically. For example, in a D2D strategy relying on Bluetooth interfaces, transferring 4MB takes approximately 18 sec. (apart from establishing a connection). A reasonable contact persistence also leads to saving device resources (e.g., battery) avoiding re-starting transmissions due to disconnections.

Our centrality degree analysis helped to identify some popular users from the dataset in each period. This metric is frequently applied for delivering messages opportunistically. In a real scenario, this is a problem, as choosing the higher centrality degree nodes too often drains their resources and confronts system feasibility. The Radius of Gyration extracted from MACACO shows confinement depending on the period. We see a higher ROG in periods in which people are possibly going from home to work (or the opposite). In contrast, we see higher confinement during work or class periods (i.e., the dataset features researchers, and students). Understanding these characteristics assists in identifying the best periods for stimulating D2D data offloading. As for the Sojourn Time, we account for *how often* and *how long* the users visited the cells. Through this metric, it is possible to find how many different cells each user visited per period and correlating with ROG.

During more confined periods (e.g., work, or homestay), for most users, the number of visited cells is smaller, the Sojourn Time into the confined cell is higher, while the ROG is restricted. Finally, through the Destination Proximity metric we see the users inside or close to cell "D" during working periods from "M" to "E". Examining the "EE", and "E" periods, we see users getting away from the cell, precisely at the time they suppose to leave the campus. Identifying those users and their HOME LOCATION is relevant if we need to take a content from the campus to a specific area, for example. Looking through the "N", and "EM" periods, we also identify how far the users live from the campus, as they are more likely to stay at "HOME LOCATION" during "N" and part of "EM".

V. Applying the Knowledge into Networking

The purpose of this Section is to show the possibility of applying factors beyond traditional mobility metrics in different problems, but focusing then on a D2D opportunistic communication strategy. In content-centric networks, mobility metrics such as Radius of Gyration, Sojourn Time, and Geographic Direction Awareness may apply for mitigating network flooding. Some initiatives focus on reducing interest and data packets re-transmission by limiting their propagation in a specific geographic area in Vehicular Named Data Network (VNDN) [16]. A second example is applying mobility metrics in models for the destination choice game problem, relevant in migration prediction, global disease mitigation, urban planning, and other situations. Most models ignore possible congestions on the way [17]. User metrics can reveal important facets and assist in predicting congestion.

Among other examples, we discuss below the application of our metrics extracted in an opportunistic D2D communication strategy. The objective consists in taking forwarding decisions based on those metrics for offloading content opportunistically. Therefore, we describe here ideas and intuitions we thought as a part of the strategy's algorithm. The forwarding strategy runs on each node u that carries a content c when u encounters a node v in an instant t. If some algorithm's condition forwards c successfully to v, this also starts running the strategy upon other encounters, up to c's time-to-live. When a node u requests a content c, a central entity knows his current network cell. We assume that each node stores metrics coefficients in a local table divided by period p, and u forwards c to v only if one of the algorithm's conditions is satisfied. Otherwise, u awaits the next encounter.

The first algorithm test is whether v already carries the content. If so, u waits for the next encounter. Otherwise, the algorithm checks whether v is the final destination, forwarding c if true. If not, the strategy checks whether v is in the same cell as the node that requested c. In this case, there is a centrality degree comparison, that is, if $\Delta_{CD_p}(v) > \Delta_{CD_p}(u)$. If true, it means that v has historically encountered more nodes in that period p. So the algorithm tries sending c to v, given his greater capacity for local dissemination. Otherwise, the ROG and ST metrics are compared. If $\Delta_{RG_p}(v) > \Delta_{RG_p}(u)$ and $\Delta_{ST_p^c}(v) > \Delta_{ST_p^c}(u)$ means that v is less "popular" (lower CD). On the other hand, v has the potential to cover a larger area in that cell (higher ROG) and also tends to stay in the cell longer than u. If true, content c is forwarded; otherwise, node u awaits the next encounter. If the consumer node reported a cell different from v 's, the algorithm tests the Geographical Direction. If v is moving towards the consumer cell, c is forwarded. Otherwise, the algorithm checks destination proximity. If $\Delta_{MP_p^c(v)} < \Delta_{MP_p^c(u)}$, the content is forwarded, as there is a greater chance of v meeting nodes at the cell edge or visiting the target cell during the period p. With this strategy, we hope to reach the content destination naturally, basing decisions on human-behavior aspects. Among future works, this strategy will be evaluated through performance metrics and compared with other proposals with a similar purpose.

VI. Conclusions and Future Work

Real datasets availability and knowledge extraction will leverage mobile network solutions. In this work, we investigated human-aware mobility metrics to support opportunistic communication. Further, we linked a framework for human-context data extraction with a case study of a real-world dataset manipulation. Finally, we presented results from dataset manipulation, metrics, and a novel temporal resolution analysis. An opportunistic D2D strategy will feature this knowledge.

Among future work, we will compare the strategy's performance through delivery rate, overhead, and delay with other similar intent proposals and other datasets. Further, it is also necessary to discuss novel mobility metrics (or different resolutions from traditional ones, e.g., visitation frequency and trip length) to mitigate opportunistic networks challenges and suit different scenarios. Finally, investigating hidden Markov Models [18], [19] to predict next-destination could be useful to transfer a message from one location to another.

Acknowledgment

This work is partially supported by CAPES, Inria, CNPq, FAPESB, FAPERJ, and FAPESP.

References

[1] R. Costa, L. Sampaio, A. Ziviani, and A. Viana, "Humanos no ciclo de comunicação: facilitadores das redes de próxima geração," in *Livro de Minicursos do XXXVI Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC) 2018*, Campos do Jordão, SP, may 2018.

[2] K. Thilakarathna, A. C. Viana, A. Seneviratne, and H. Petander, "Design and analysis of an efficient friend-to-friend content dissemination system," *IEEE Transactions on Mobile Computing*, vol. 16, no. 3, pp. 702–715, March 2017.

[3] E. M. R. Oliveira, A. Viana, K. P. Naveen, and C. Sarraute, "Mobile data traffic modeling: Revealing temporal facets," *Computer Networks*, vol. 112, p. 176–193, 01 2017.

[4] C. P. Lau, A. Alabbasi, and B. Shihada, "On the analysis of human mobility model for content broadcasting in 5g networks," in *2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, Oct 2017, pp. 1–7.

[5] F. Xia, J. Wang, X. Kong, Z. Wang, J. Li, and C. Liu, "Exploring human mobility patterns in urban scenarios: A trajectory data perspective," *IEEE Communications Magazine*, vol. 56, no. 3, pp. 142–149, MARCH 2018.

[6] P. Hui, J. Crowcroft, and E. Yoneki, "Bubble rap: Social-based forwarding in delay-tolerant networks," *IEEE Transactions on Mobile Computing*, vol. 10, no. 11, pp. 1576–1589, Nov 2011.

[7] E. M. R. Oliveira and A. C. Viana, "From routine to network deployment for data offloading in metropolitan areas," in *2014 Eleventh Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*. NY: IEEE Press, June 2014, pp. 126–134.

[8] I. O. Nunes, P. O. S. V. de Melo, and A. A. F. Loureiro, "Leveraging d2d multihop communication through social group meeting awareness," *IEEE Wireless Communications*, vol. 23, no. 4, pp. 12–19, August 2016.

[9] I. O. Nunes, C. Celes, I. Nunes, P. O. S. Vaz de Melo, and A. A. F. Loureiro, "Combining spatial and social awareness in d2d opportunistic routing," *IEEE Communications Magazine*, vol. 56, no. 1, pp. 128–135, Jan 2018.

[10] Y.-A. Montjoye, C. Hidalgo, M. Verleysen, and V. Blondel, "Unique in the crowd: The privacy bounds of human mobility," *Scientific reports*, vol. 3, p. 1376, 03 2013.

[11] J. Andrew, J. Karthikeyan, and J. Jebastin, "Privacy preserving big data publication on cloud using mondrian anonymization techniques and deep neural networks," in *2019 5th Int. Conference on Advanced Computing Communication Systems (ICACCS)*, March 2019, pp. 722–727.

[12] P. O. V. de Melo, A. C. Viana, M. Fiore, K. Jaffrès-Runser, F. L. Mouël, A. A. Loureiro, L. Addepalli, and C. Guangshuo, "Recast: Telling apart social and random relationships in dynamic networks," *Performance Evaluation*, vol. 87, pp. 19 – 36, 2015, special Issue: Recent Advances in Modeling and Performance Evaluation in Wireless and Mobile Systems.

[13] A. C. S. A. Domingues, F. A. Silva, and A. A. F. Loureiro, "Space and time matter: An analysis about route selection in mobility traces," in *2018 IEEE Symposium on Computers and Communications (ISCC)*, June 2018, pp. 00 958–00 963.

[14] N. Eagle and A. (Sandy) Pentland, "Reality mining: Sensing complex social systems," *Personal Ubiquitous Comput.*, vol. 10, no. 4, pp. 255–268, Mar. 2006.

[15] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma, "Mining interesting locations and travel sequences from gps trajectories," in *Proc. of the WWWW*, ser. WWW '09. ACM, 2009, pp. 791–800. [Online]. Available: http://doi.acm.org/10.1145/1526709.1526816

[16] A. M. de Sousa, F. R. C. Araújo, and L. N. Sampaio, "A link-stability-based interest-forwarding strategy for vehicular named data networks," *IEEE Internet Computing*, vol. 22, no. 3, pp. 16–26, 2018.

[17] Y. Xiao-Yong and Z. Tao, "Destination choice game: A spatial interaction theory on human mobility," *Scientific Reports*, vol. 9, 02 2018.

[18] M. Yin, M. Sheehan, S. Feygin, J. Paiement, and A. Pozdnoukhov, "A generative model of urban activities from cellular data," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 6, 2018.

[19] S. Hasan and S. V. Ukkusuri, "Reconstructing activity location sequences from incomplete check-in data: A semi-markov continuous-time bayesian network model," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 3, pp. 687–698, 2018.