MOCHA: A Tool for Mobility Characterization

Fabrício R. Souza
Federal University of Minas Gerais
Belo Horizonte, Brazil
fabricio.souza@dcc.ufmg.br

Augusto C.S.A. Domingues
Federal University of Minas Gerais
Belo Horizonte, Brazil
augusto.souza@dcc.ufmg.br

Pedro O. S. Vaz de Melo
Federal University of Minas Gerais
Belo Horizonte, Brazil
olmo@dcc.ufmg.br

Antonio A. F. Loureiro
Federal University of Minas Gerais
Belo Horizonte, Brazil
loureiro@dcc.ufmg.br

ABSTRACT

There are many mobility models in the literature with diverse formats and origins. Besides the existence of studies that analyze and characterize these models, there is a need for a framework that can compare them in an easy way. MOCHA (Mobility framework for CHaracteristics Analysis) is a tool that characterizes and makes possible the comparison of mobility models without any hard work. We implemented 9 social, spatial and temporal characteristics, which were extracted from various (real and synthetic) distinct mobility traces. MOCHA has a classifying module that attributes each characteristic the statistic distribution that better describes it. As a validation process, all the traces were compared using the T-SNE method for data visualization, resulting in the approximation of similar traces. One of the advantages of using MOCHA is its ease of use, being able to read diverse traces formats and converting them to its standard format, allowing that different types of traces, such as check-in, GPS, contacts, and so on, to be compared. The metrics used in the tool can become a standard for trace analysis and comparison in the literature, allowing a better vision of where one trace belongs related to others. MOCHA is available for download at https://github.com/wisemap-ufmg/MOCHA

KEYWORDS

Mobility Characterization, Mobility Models, Mobile Networks

ACM Reference Format:
Fabrício R. Souza, Augusto C.S.A. Domingues, Pedro O. S. Vaz de Melo, and Antonio A. F. Loureiro. 2018. MOCHA: A Tool for Mobility Characterization. In *21st ACM International Conference on Modelling, Analysis and Simulation of Wireless and Mobile Systems (MSWIM '18), October 28-November 2, 2018, Montreal, QC, Canada.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3242102.3242124

MSWIM '18, October 28-November 2, 2018, Montreal, QC, Canada

ACM ISBN 978-1-4503-5960-3/18/10…$15.00
https://doi.org/10.1145/3242102.3242124

1 INTRODUCTION

There is a variety of mobility models available in the literature with different formats, origins and purposes. Such models have important applications on the simulation of mobile networks, as a representation of users behavior, knowing that it is possible to assess the quality of a model by how faithful it is to real users' mobility. Small World in Motion (SWIM) [18], Self-similar Least Action Walk (SLAW) [16], and Working Day Model (WDM) [9] are mobility models frequently used in the literature. SWIM uses complex networks metrics to create "small-world" graphs among users. SLAW was one of the first mobility models to consider truncated power-laws as statistical distributions to users' flights, waiting time, etc. WDM is a model based on a person's regular working day: they wake up around the same time every day, leave for work around the same time, stay at their work location for the same amount of time and return home. Even though we can use all of these mobility models to represent a person's mobility, there are different contexts and modeling decisions applied to each one, making it difficult to directly compare them. This being said, it is important and necessary to be able to compare different mobility models. For instance, we can compare a synthetic trace generated using some mobility model with a real trace to verify how close to reality the mobility pattern created is, which speaks to the quality of the model. Also, we can compare a real trace with a synthetic one allowing us to verify how close to other existing models the current one is. This ability allows for a better understanding of the relationships between the different mobility models without the need to process large data. However, there is not a common procedure in the literature to compare mobility traces.

There are many factors to consider when performing trace comparison. First, the data collected is not always the same, making it hard to compare certain information between two traces if it is absent in one of them. In contact traces, such as DARTMOUTH [14] and SWIM [18], the samples represent encounters between pairs of network users on a given location. This location can be geographic coordinates or a point of interest. This type of trace is known to be chronologically sparse, i.e. we have no information about what happened with the users between two encounters. On the other hand, in GPS traces, such as the San Francisco Taxi trace [20], the samples show the precise position (longitude and latitude) of a user in a given moment in time. However the occurrence of users encounters are not clear. Nevertheless, this data can be mined from the trace. Another type of trace is the check-in one, such as [26],

which logs all the times a user literally checked-in a point of interest, via social networks. Thus, we can see that the spatial precision, the time granularity and contextual information make it harder to compare different types of traces.

It is clear that there is a need to compare different traces. In order to address this problem, we introduce MOCHA (**MO**bility framework for **CH**aracteristics **A**nalysis), a tool that allows for the extraction of characteristics from mobility traces in different formats, making it easy to compare them. The tool has been divided in four modules: the first converts the traces into a normalized format in which they can be processed by the following steps without loss of information; the second one extracts different social, spatial and temporal characteristics from the data; the third one classifies each of those characteristics according to their statistical distribution; and the fourth one uses visualization methods to compare and cluster the traces according to their characteristics distribution. Therefore, the contributions of this work are twofold:

(1) Providing a free tool to the research community that can be used to analyze and compare mobility traces;
(2) Defining a set of spatial, social and temporal metrics which can be used to analyze mobility data.

To validate MOCHA, we gathered and compared different real and synthetic mobility traces available in the literature. The results present valuable insights, such as the clustering of traces with similar purposes. MOCHA is an open-source tool[1], available for use under GPL v3 license. By being open-source anyone can add new metrics and new visualization methods to the tool.

The remaining of this work is organized as follows. Section 2 presents the related work. The ensemble of metrics used in MOCHA is listed and explained in Section 3. The data collection steps used in this work and the resulting tool are introduced in sections 4 and 5, respectively. The validation results are presented in Section 6. Finally, the conclusions are drawn in Section 7.

2 RELATED WORK

The characterization of mobility models and traces is a hot topic in the literature [2], given its applications in the development of novel technologies for the mobile networks. In this section, we provide a brief overview of the relevant works in the literature, stating how the proposed tool could increase the achievements in the characterization and generation of mobility traces.

2.1 Mobility Models Characterization

Some works propose frameworks for the analysis of mobility traces. In [24], the authors present a method of mobility benchmarking, which includes a multi-dimensional mobility metric space and a framework for mobility analysis and performance of protocols in traces. Additionally, a new mobility model is proposed and analyzed together with other models in the literature. Besides presenting a framework for the analysis of different mobility traces and models, this work doesn't consider the classification of traces in relation to their characteristics. IMPORTANT (Impact of Mobility on the Performance Of RouTing protocols in Adhoc NeTworks), a framework to analyze the impact of mobility in the performance of routing protocols in ad hoc networks is presented in [3]. It focuses on mobility models, mobility metrics, and graph properties in networks. However, only a few mobility models are considered in the framework.

The SLkit [11] is a tool capable of automatizing the process of extraction and analysis of spatial and temporal metrics in *Sensing Layers*, aiming to speed up the investigation of multiple datasets. However, the tool does not consider the extraction of social characteristics from traces. In [25] the authors present a structured framework for extracting the mobility characteristics from a WLAN trace in order to derive parameters to a proposed mobility model. Nevertheless, the characteristics extracted are specific for WLAN traces and thus cannot be applied in the characterization of other types. Finally, [4] proposes a framework to simulate generic mobility that allows the comparison of mobility models considering spatial, temporal, and social metrics, considering the models *Random Waypoint*, *Manhattan*, and *GIS*.

Additionally, there are studies focused on the validation of mobility traces. [5] validates mobility models regarding their proximity to the reality. To do so, the authors compare the ground truth of users behavior and the synthetic data generated by the mobility models. Furthermore, they propose a technique for fine tuning of the mobility models' parameters based on the values obtained from real samples.

From the analysis of the literature, it is possible to observe that there is a need for a tool that allows for the comparison of different mobility traces and models. MOCHA is capable of comparing traces in different formats (such as GPS, contacts, and check-ins), extracting mobility metrics that consider temporal, social, and spatial characteristics. The visual comparison feature considers the statistical distributions and its estimated parameters for each metric. Thus, it is possible to group similar traces, which offers insights on a Cartesian space of mobility features. This allows the validation of generated synthetic traces. The availability of this tool to the scientific community decreases the effort needed for the analysis of mobility models, as well as contributes for setting a standard for the metrics used in the characterization of real and synthetic traces.

3 METRICS

Mobility traces can be used on many different applications, which themselves can be more sensitive to different metrics. One of the types of metrics are the spatial ones, that describe how a user deals with the space that they occupy. The next type is temporal metrics, the ones regarding the user's relationship with time. There are also the social metrics, that illustrate how the user deals with others around themselves. Considering the existence of different types of metrics, it is interesting to consider their collective when analyzing and comparing mobility models. This section presents the metrics extracted from the traces.

3.1 Spatial metrics

3.1.1 Spatial Variability (SPAV). Given a user u and its set of visited locations $L = \{l_1, l_2, \ldots, l_n\}$, its spatial variability S_v can be defined as the Shannon's Entropy [22] S_v of the locations' visit probabilities $P(l_i)$. The more balanced the distribution of the user samples over

[1] MOCHA is available for download at https://github.com/wisemap-ufmg/MOCHA

the locations, the higher the value of S_v. Consequently, an user that visits most of the time a same location will have a low variability. We can define $\bar{S_v}$ the populations' spatial variability, computed as the mean of the individual values:

$$\bar{S_v} = \frac{1}{n} * \sum_{i=0}^{n} S_{v_i}$$

The spatial variability tells a lot about the movement patterns of a population. For example, if we consider a taxi trace, such as [20], we expect this variability to be as high as possible, given the fact that taxicabs move under the will of their passengers. On the other hand, if we consider a check-in trace, such as [26], there is lower variability, since users tend to visit frequently their home and work office, and only occasionally visit other locations such as restaurants and malls.

3.1.2 Radius of gyration (RADG). When dealing with social mobility it is common to assume that every user has a location considered as their home. In [15] the authors define the home as a random point chosen uniformly on the available space. The authors in [10] consider the home as the place where a user's activities begin each day. The latter can be affected by factors such as the frequency of collection of an user and the precision from the device, causing an user to have multiple homes. Similarly, the random choice on the former can set as home a point where an user may never visit. Therefore, in this paper we consider the home of a user as their most visited location, which favors the idea of a recurrent place.

It is safe to assume that a user tends to move to places closer to their home, to run errands, shop, visit friends etc. The radius of gyration is a metric that measures the distance between a user's home and the other places they have visited. We consider the RADG as the maximum distance between the set of visited locations L by a user u and the user's home, in this paper we used the euclidean distance between points as the distance function. It can be formally defined as:

$$RADG_u = max(distance(u.home, l) \forall l \in L)$$

3.1.3 Travel distance (TRVD). According to [10, 13], the action of moving from a place to another can be described as a trip, jump or flight. Moreover, the travel distance can be defined as the distance between two consecutive places. There are many ways to calculate this travel distance, such as euclidean distance, geodesic distance, hop count, and so on. By definition, we want the distance between two consecutive points o and d, computed through a distance function:

$$TRVD_{o,d} = distance(o, d)$$

The analysis of this characteristic allows for a better understanding of how a user moves from place to place. That understanding can be used to leverage the dissemination of messages on vehicular and opportunistic networks.

3.2 Temporal

3.2.1 Visit time (VIST). Visit time can be defined as the time spent in each location by a user [13]. It can be formally defined as the difference between the time instant where an user u departed location t and the time instant where u arrived at t:

$$VIST_{u,t} = departure_{u,t} - arrival_{u,t}$$

3.2.2 Travel time (TRVT). Travel time can be defined as the time spent moving between two places. Given two consecutive locations l_a and l_b, where l_a is the origin and l_b the destination of user u, we have:

$$TRVT_{u,l_a,l_b} = arrival_{u,l_b} - departure_{u,l_a}$$

3.3 Social

3.3.1 Contact entropy (CONEN). The contact entropy of an user describes, just like the Spatial Variability, how distributed are the contacts of an user u among his set of contacted users. Users with small values of entropy have most of their contacts with few other users, which can be seen as his friends. On the other hand, users with high values of entropy have their encounters well-distributed among his set of peers. Detecting users with high variability of encounters can be helpful to the routing of messages in opportunistic networks.

Given an user u and its set of contact peers $C = c_1, c_2, \cdots, c_n$, its contact entropy CE_u can be defined as the Shannon's Entropy of the probabilities $P(c_i)$ of user u contacting user c_i.

3.3.2 Inter-contact time (INCO)). Inter-contact time can be defined as the time interval between two consecutive encounters between a pair of users [15]. In opportunistic scenarios, INCO represents the frequency of encounters between each pair of users, which represents an opportunity for message transmission [23].

3.3.3 Contact duration (CODU). Contact duration can be defined as the amount of time two users spent inside each others transmission range, without interruptions. Similarly to the INCO, the CODU represents an opportunity to transmit a message in an opportunistic network. However instead of showing the frequency of encounters the CODU shows the amount of data that can be delivered in each encounter. For each entry of the normalized trace, the algorithm writes the TDC on the CODU file.

3.3.4 Encounter regularity (EDGEP). The encounter regularity is a complex networks metric that maps the regularity of a social relation. An user has a group of other users that they tend to meet more often than other random users [7]. The more contacts two users have in a given time window, the stronger is the social tie between them. When considering opportunistic scenarios, EDGP represents how often we can deliver a message between a specific pair of users.

The encounter regularity (EDGEP) represents how long an encounter lasts when considering a set time window. This time window represents the maximum amount of time two devices can stay in contact during the trace. For example if the total time of the trace is 15 days, and we are considering a daily encounter regularity, the time window is 15, but if we want to consider a hourly encounter regularity, the time windows should be 15*24. MOCHAS' standard is to consider a daily time window.

3.3.5 Topological overlap (TOPO). The topological overlap represents the social overlap between each pair of nodes when considering every encounter they had. TOPO can be used to determine opportunities to deliver messages inside a community. It is easier

to deliver a message to a person of a set group if the message is relayed via users from that same group.

3.3.6 Social Correlation (SOCOR). The social correlation represents the Pearson's correlation value between the EDGEP and the TOPO values. This value represents how stable the relationship between the users a set user meets along their trips and the amount of times they saw these users in a set time window. This metric can be used to evaluate how efficient the data dissemination on opportunist networks would be for a group of users, if they have a high SOCOR value it means that they meet regularly.

4 DATA COLLECTION

There are numerous mobility models and traces available in the literature, e.g. SWIM [18], SLAW [16], SMOOTH, DARTMOUTH [14], serving different purposes of analysis. In this work, we collected a diverse amount of those and used them to validate MOCHA by extracting their metrics and comparing the results. This subsection describes each mobility trace used.

SWIM [18] is a synthetic mobility model that uses small-world graphs, from complex networks, to create mobility. The model uses two parameters as configuration of such mobility, the first one, α, dictates how far a user is willing to go from their home and the second one, β describes how long a user stays put in a place. Seven different traces were generated. Three of those had the β value set to 0.7 and the α value varied from 0.6 to 0.8 on 0.1 increments. The other four traces had the α value set to 0.7 and the β value varied from 0.6 to 0.75 on 0.05 increments.

The WDM mobility model [9] tries to replicate the mobility patters of a person that works in a traditional nine to five job, which means that the person wakes up, goes to work, stays there for a long period of time, after work the person can run some errands, and then go home. The One Simulator[2] offers pre-configured WDM scenarios and four of those were used in this work. The first one is called ProbShopping0 simulates a scenario where all users go straight home from work, never going shopping. The second one is called ProbShopping100 is a scenario where all users go shopping after work and then go home. The third one is ShoppingGroupMax10 where the users form groups to go shopping after work. The size of the group in random between 2 and 10 and the probability to go shopping is 50%. The last one is called Taxi100, where all users go to and from work via taxi, never using the public transport system.

The GRM mobility model [19] is concerned with maintaining the regularity of the group encounters on the mobility model. The trace used is provided by the authors, it contains 100 users and is openly available. The trace uses the default configuration set by the authors in the article.

The Dartmouth trace [14] is a real data trace that represents the movements of students on the Darthmouth university campus. The data was collected using access points scattered around the campus, and logged every encounter between a student and the access point.

The Gowalla trace [6] is a real data trace similar to the Dartmouth one, but it represents the movement of students by logging their check-ins in places inside and outside campus.

[2]https://akeranen.github.io/the-one/

5 MOCHA

In this work we present MOCHA (MObility CHAracterization framework), a tool to process, extract, classify and compare mobility characteristics in traces and mobility models. One of the advantages of MOCHA is the possibility to compare traces of different types, such as contact, check-in and GPS traces by converting them to a defined standard format. After this pre-processing, we extract 11 social, spatial and temporal metrics, which later are used to classify and compare the traces. Therefore, MOCHA can be used to evaluate the quality and similarity of new real and synthetic traces in relation to those existing in the literature. Figure 1 illustrates the steps executed by the framework, highlighting the inputs needed and the outputs produced by each one. The following subsections describe in details each of MOCHA's modules.

5.1 Processing

Due to the different gathering methods of traces in the literature, such as access points scattered around an area, GPS logs from a smartphone, check-ins on social networks, and son on, each one of them have a different format to record the mobility of their entities. Therefore, we must process these traces first in order to apply the algorithms to extract the metrics.

MOCHA's processing module is responsible to normalize these traces in a unique defined format. In this format, every encounter that happened during the collection of the trace is represented. The fields in the MOCHA format are:

- *ID1* is an unique integer for User 1;
- *ID2* is an unique integer for User 2;
- *TA* is the trace's current time stamp;
- *TI* is the time stamp from when the contact between *ID1* and *ID2* started;
- *TD* is the contact duration time;
- *X1* is User 1's first coordinate;
- *Y1* is User 1's second coordinate;
- *X2* is User 2's first coordinate;
- *Y2* is User 2's second coordinate;

Parsing raw traces. Raw traces usually contain at least four different fields per entry, the user's ID, two coordinates and a time stamp. The format MOCHA uses as standard represents all contacts that happened during the trace, to find these contacts we need to parse the whole trace and, for each entry verify how the movement described in it alters the whole context of contacts.

In this work, we consider a contact every time two devices are within communication range of each other. A value R needs to be defined as such range, and it is used to gather all the encounters.

The algorithm evaluates each entry on the raw mobility trace. A graph is created and if a vertex is already on the graph it means that it is not the first time the device has been found on the trace, which means the algorithm needs to check if any of the device's former neighbors are not within communication range anymore. If they are not, that means an encounter has ended and needs to be logged on the parsed trace, according to the format specified earlier, and the edge that existed between the two vertexes needs to be removed. If the vertex is not present in the graph, the algorithm needs to add it and check if any other vertex is within communication range of

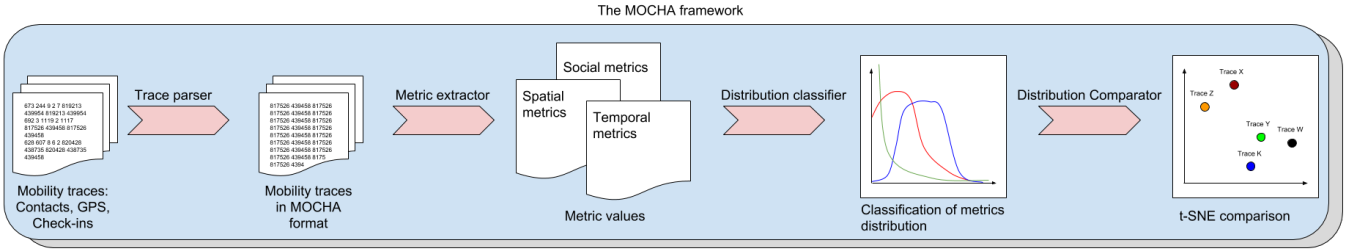

Figure 1: Steps taken in MOCHA

the vertex. If they are, an edge is added between the two of them with the weight of the current time. This weight is used to calculate the duration of the encounter.

To be able to find neighbors, we created a cell mesh that represents the area where the devices are moving. To identify the cell a device belongs to, the algorithm calculates a value $K = \lfloor \frac{X}{R} \rfloor$ and a value $L = \lfloor \frac{Y}{R} \rfloor$, meaning the cell the device belongs to is the cell K L. Once the device's cell has been assigned, the algorithm check all the adjacent cells for users. Adjacent cells are cells that are in the interval K - 1 to K + 1 and L - 1 and L + 1.

Another type of mobility trace MOCHA can have as an entry is a contact trace. Contact traces usually have more information than raw mobility traces, making it easier to extract the information needed to normalize the mobility trace. For example, considering the SWIM format for contact mobility traces, we extract all information needed. SWIM mobility traces entries' are in the following format:

TCU TYPE ID1 ID2 X1 Y1 X2 Y2

. The inputs are the same used in the normalized format for the traces, with the addition of the TYPE input, which is irrelevant for the parsing process.

The algorithm evaluates each entry on the SWIM mobility trace. A graph is created. If an edge between ID1 and ID2 already exists, it means that this is not their first encounter, so the algorithm needs to log a new entry on the parsed trace, according to the normalized format, and the edge's weight is set to the current trace time. If the edge does not exist, it means that it is their first encounter, so vertex are added as needed to the graph, and the edge is added with the current trace time.

5.2 Extraction

Once the data is in the MOCHA format, we can proceed to extract the desired properties. The objective of the extractor is to mine from the trace every mobility property described in Section 3. After the extraction, we can analyze the statistic curve generated by the aggregation of all the samples of a property, for example.

It is important to highlight some suppositions which must be considered. First, not every trace analyzed by MOCHA have all the possible characteristics for extraction. In a check-in trace without geographic coordinates of its points of interests (PoI), it is impossible to calculate properties like Travel Distance. Similarly, a contact trace without geographic coordinates allows only the extraction of social metrics. However, we can see that it is always possible to extract all social metrics.

Additionally, there are some properties in which their extractions depend on configuration parameters sensible to the simulation context. For example, all the contacts in MOCHA are evaluated based on a 50 meters radius. However, this parameter can be tuned according to the context, if needed. The 50 meter radius is considered the default value is this work because traces largely used in the literature, like *Dartmouth* [14], are based on Wi-fi routers proximity, in which a contact is considered every time two distinct users are connected to a same router. Setting the range radius of Wi-fi networks to approximately 50 meters is common in the literature [8].

Finally, considering that the process of extraction is executed only for traces normalized to the MOCHA format, every trace must be chronologically sorted and each records must be in the format described in 5.1.

5.3 Classification

With the extracted characteristics it is possible to classify them according to their statistical distributions. To do this, we used the methods of Maximum Likelihood Estimation (MLE) [21] together with the Akaike Information Criteria (AIC) [1].

The MLE is a method that identifies how close to a given distribution a dataset is. The quadratic error between the data and the distribution are calculated, and the distribution with the lowest quadratic error is the closest one to the original data. However, different distributions use a set of different parameters, which is of extreme importance to the classification. It is not possible to considerate that a distribution that uses more parameters than other has the same level of details in the classification. To solve this problem, we apply the AIC. The distribution with the smallest AIC value is the one closer to the real data. MOCHA compares every extracted characteristic, except SOCOR, with seven different distributions, being them Gamma, Weibull, Exponential, Normal, Log-normal, Pareto (Power-Law) and Log-logistic [12].

5.4 Comparison

After classifying the metrics according to their statistical distributions, we proceed to the step of comparing the traces. To do this, we use the metrics classification with the intention to verify which traces have similarities between themselves. This information is of great importance to works that use mobility traces as a source. First, it is possible to verify how similar a new trace (real or synthetic) is in relation to those used in the literature, allowing researchers to consider knowledge obtained from previous traces analysis to new

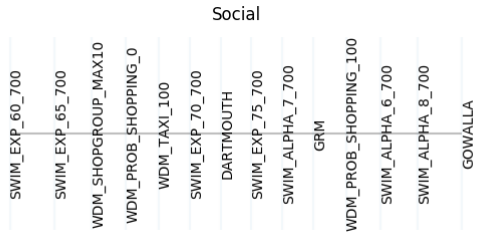

Figure 2: 1D t-SNE visualization considering metrics in the social spectrum. The more distant the traces, the more they are different

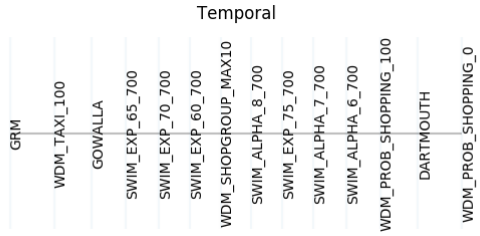

Figure 3: 1D t-SNE visualization considering metrics in the temporal spectrum. The more distant the traces, the more they are different

ones, thus reducing the time spent with simulations and analysis. Second, it is also possible to evaluate the quality of a new trace through comparison. Finally, the use of this tool favors the elaboration of a map of traces' characterization, which can help the selection of traces to be applied in a study.

To perform the comparison, we use t-SNE (t-Distributed Stochastic Neighbor Embedding) [17], a technique to visualize data with high dimensionality in a plot of two or three dimensions. t-SNE produces data visualizations that are significantly better than those produced by other data aggregation techniques. MOCHA uses this technique because the main objective of its comparison module is to offer a visual guide of relations between traces.

6 RESULTS

In this section, we show the results of applying MOCHA to real and synthetic mobility traces in the literature. We apply t-SNE to each category of metrics (spatial, social and temporal), instead of to the entire set, and present visualizations considering various combinations of categories. Therefore, it is possible to look at the differences in a specific spectrum, if needed, allowing for an easier evaluation and tuning of new traces.

We start by looking at the results in each dimension. Figures 2, 3, and 4 show the result of t-SNE comparison in the social, temporal, and spatial spectrum, respectively. Each trace is represented over a line, with the distance in this line between two traces representing their similarity. The closer, the more similar. In the first one (Figure 2), it is possible to see that the SWIM trace with 0.6 value

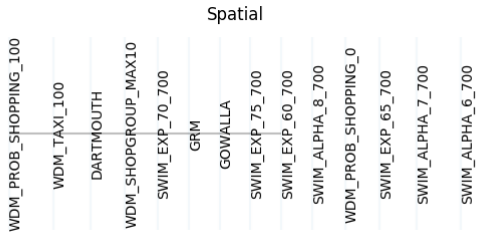

Figure 4: 1D t-SNE visualization considering metrics in the spatial spectrum. The more distant the traces, the more they are different

for β is not a good representation of the Gowalla trace, for the social spectrum of the traces, that can be inferred because they are 13 spaces distant from one another in the space. However the the SWIM trace with α value 0.8 is the closest one to Gowalla, showing that the selection of parameters is very important when utilizing a synthetic trace. On the over hand when we shift the spectrum to the temporal one, Figure 3, we can see that the similarities have changed. The SWIM trace that was the closest one to the Gowalla one is now not as close, and the SWIM that was the farthest one is now closer. We can see that each metric spectrum describes the traces in its own way. On the spatial spectrum we can again see another shift.

Representing the similarities using a 1D graph can be interesting but when we increase the amount of dimensions the inferences we can make are greater. We can see on Figure 5(b) that the WDM_Prob_Shopping0 trace and the WDM_Prob_Shopping100 trace, which differ on the chance of a user going shopping after work, are closer than on Figures 5(a) and 5(b). That can be interpreted as the users sharing the same spatial patterns and social ones, because the office they go to work is the same and the people they see while shopping is what makes the social aspect different. We can also see that that the GRM model represents well the temporal and spatial aspects of the Dartmouth trace, since on Figure 5(a) they are very close, while on figure 5(b), when the social aspect is introduced, the two traces become much more separated, indicating that GRM cannot capture the social aspect of the Dartmouth trace.

Another type of visualization that MOCHA offers is the one shown on Figure 6. In it, similar traces have similar line forms. We can see that the SWIM_ALPHA_8 trace has a different distribution when comparing the INCO (Inter-contact time) metric. The α parameter represents how far the users are willing to go from their houses. If a user goes further it is expected that it will take longer for him to meet people he met in the past, making the INCO parameter have a different distribution. We can also see that changing the β parameter on SWIM traces only affected the INCO metric. When we look at the WDM lines we can see that the different parameters applied to the traces only affected the spatial and temporal metrics, having no impact on the social aspect of the trace.

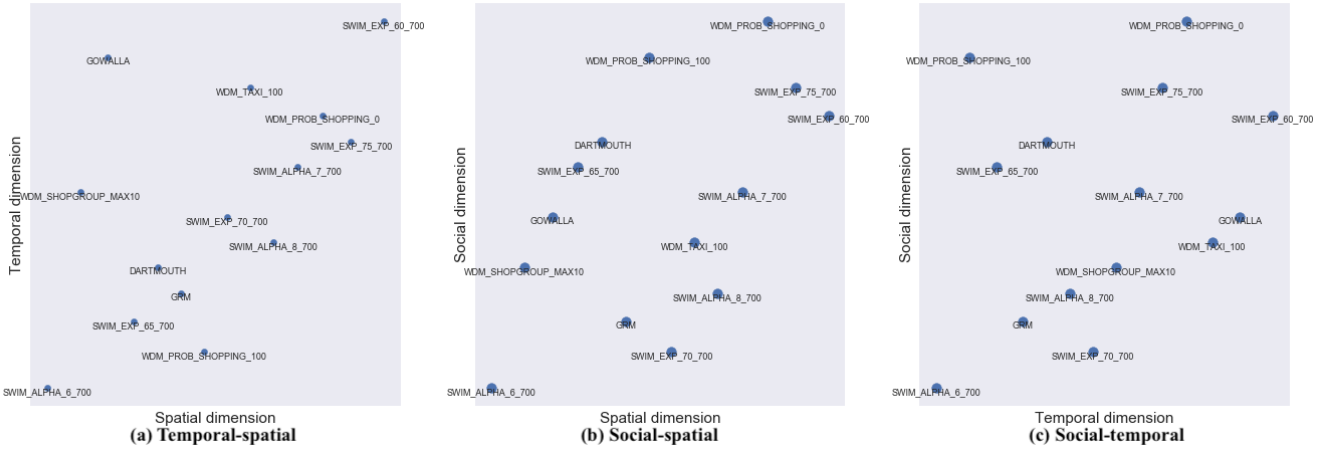

Figure 5: Two-dimensional visualization of differences between traces

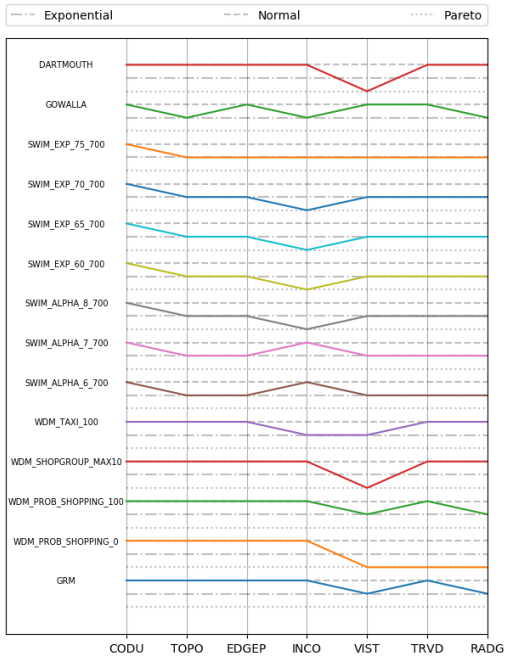

Figure 6: Line graph of some of the metrics in MOCHA. Similar traces have similar lines.

6.1 SWIM Case Study

By considering the statistical distribution of each metric and not the values themselves, MOCHA can create a signature for a trace, i.e., it will produce the same results for the same mobility model with different parameters. To demonstrate that, we generated four different traces based on SWIM varying the values for the α parameter, which dictates how far a user is willing to go from their home. The values selected were 0.1, 0.2, 0.8, and 0.9, in which being closer to 1.0 indicates a higher willingness to go further.

Figure 7 shows a histogram for the Travel Distance metric in each one of the generated traces. As we can see, they all follow an exponential distribution, however, the values achieved in each one are different. Additionally, in the trace where $\alpha = 0.1$, the variance is smaller, given the fact that users are less willing to go further, therefore they always travel the same distances. On the other hand, by looking at the trace with $\alpha = 0.9$, the variance is higher, demonstrating that users are taking trips with different sizes.

7 CONCLUSION

With a great amount of works that use mobility traces as data sources for simulations, it is necessary to evaluate the quality of the datasets available in the literature. Two conventional manners of evaluating the quality of a trace are metrics extraction and comparison with well-known traces. However, the lack of a standard set of metrics to perform these steps increases their trustworthiness. Additionally, different traces' formats existing in the literature (contacts, GPS samples, check-in, Wi-fi, Bluetooth, and so on) makes it difficult to perform a direct comparison between them.

In this work, we presented MOCHA, a framework capable of extracting metrics of mobility traces and comparing these traces. First, it is capable of converting traces from different formats to a standard one. From this converted data, social, spatial and temporal metrics are extracted and classified according to their probability distributions. Finally, MOCHA compares the traces according to their distributions, providing a visual guide of how similar they are, as well as providing a signature that stays the same for a given mobility model with different tuning of parameters. This comparison can then be used to evaluate the quality of a new synthetic or real trace. The similarity between traces can help to reduce the number of executed simulations in a study, for example. Last but not least, the tool is open-source and is available for other researchers to make changes adding customized metrics and visualizations.

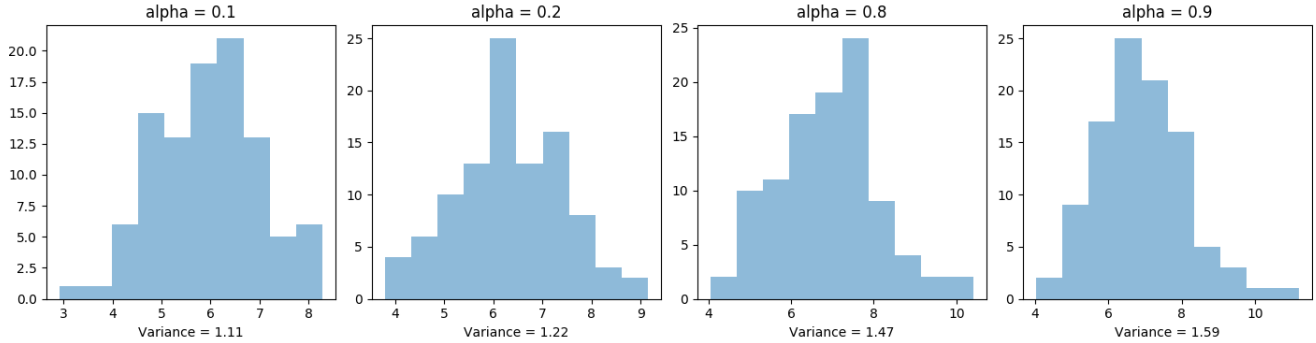

Figure 7: Distribution of Travel Distance values for different SWIM traces generated

As future works we suggest the extraction of more complex metrics that can better characterize the traces and an analysis of how different mobility patterns can affect the dissemination of data in the opportunistic networks scenario. Another suggestion is to be able to create parameter for mobility traces, such as SWIM, GRM and WDM, from a given position on the 2D similarity comparison graph. One more suggestion is to identify exactly which metric makes two, or more, traces different from each other.

ACKNOWLEDGMENT

This work was partially supported by CNPq, CAPES, FAPEMIG, and grant #15/24494-8, São Paulo Research Foundation (FAPESP).

REFERENCES

[1] Hirotugu Akaike. 2011. *Akaike's Information Criterion*. Springer Berlin Heidelberg, Berlin, Heidelberg, 25–25. https://doi.org/10.1007/978-3-642-04898-2_110

[2] L. Alessandretti. 2018. Individual mobility in context: from high resolution trajectories to social behaviour. http://openaccess.city.ac.uk/20077/

[3] Fan Bai, Narayanan Sadagopan, and Ahmed Helmy. 2003. IMPORTANT: A framework to systematically analyze the Impact of Mobility on Performance of RouTing protocols for Adhoc NeTworks. In *INFOCOM 2003. Twenty-second annual joint conference of the IEEE computer and communications. IEEE societies*, Vol. 2. IEEE, 825–835.

[4] Rainer Baumann, Franck Legendre, and Philipp Sommer. 2008. Generic mobility simulation framework (GMSF). In *Proceedings of the 1st ACM SIGMOBILE workshop on Mobility models*. ACM, 49–56.

[5] Rafael L Bezerra, Carlos Alberto Vieira Campos, and Luis Felipe M de Moraes. 2009. Uma proposta de técnica para o ajuste de modelos de mobilidade em redes ad hoc e questionamentos sobre a adequação dos parâmetros envolvidos com base em dados reais. *Simpósio Brasileiro de Redes de Computadores (SBRC)(Brasil, 2009)* (2009).

[6] Eunjoon Cho, Seth A. Myers, and Jure Leskovec. 2011. Friendship and Mobility: User Movement in Location-based Social Networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '11)*. ACM, New York, NY, USA, 1082–1090. https://doi.org/10.1145/2020408.2020579

[7] Pedro OS Vaz de Melo, Aline Carneiro Viana, Marco Fiore, Katia Jaffrès-Runser, Frédéric Le Mouël, Antonio AF Loureiro, Lavanya Addepalli, and Chen Guangshuo. 2015. Recast: Telling apart social and random relationships in dynamic networks. *Performance Evaluation* 87 (2015), 19–36.

[8] Savio Dimatteo, Pan Hui, Bo Han, and Victor OK Li. 2011. Cellular traffic offloading through WiFi networks. In *Mobile Adhoc and Sensor Systems (MASS), 2011 IEEE 8th International Conference on*. IEEE, 192–201.

[9] Frans Ekman, Ari Keränen, Jouni Karvo, and Jörg Ott. 2008. Working Day Movement Model. In *Proceedings of the 1st ACM SIGMOBILE Workshop on Mobility Models (MobilityModels '08)*. ACM, New York, NY, USA, 33–40. https://doi.org/10.1145/1374688.1374695

[10] Frans Ekman, Ari Keränen, Jouni Karvo, and Jörg Ott. 2008. Working day movement model. In *Proceedings of the 1st ACM SIGMOBILE workshop on Mobility models*. ACM, 33–40.

[11] Fabrício Ferreira, Thiago Henrique Silva, and Antônio Alfredo Ferreira Loureiro. 2017. SLkit: An R package for property extraction and analysis of multiple Sensing Layers. In *Computer Networks and Distributed Systems (SBRC), 2017 XXXV Brazilian Symposium on*. IEEE, 1184–1191.

[12] Z. Q. John Lu. 2010. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 173, 3 (2010), 693–694. https://doi.org/10.1111/j.1467-985X.2010.00646_6.x

[13] Dmytro Karamshuk, Chiara Boldrini, Marco Conti, and Andrea Passarella. 2011. Human mobility models for opportunistic networks. *IEEE Communications Magazine* 49, 12 (2011), 157–165.

[14] M. Kim, D. Kotz, and S. Kim. 2006. Extracting a Mobility Model from Real User Traces. In *Proceedings IEEE INFOCOM 2006. 25TH IEEE International Conference on Computer Communications*. 1–13. https://doi.org/10.1109/INFOCOM.2006.173

[15] Sokol Kosta, Alessandro Mei, and Julinda Stefa. 2014. Large-scale synthetic social mobile networks with SWIM. *IEEE Transactions on Mobile Computing* 13, 1 (2014), 116–129.

[16] K. Lee, S. Hong, S. J. Kim, I. Rhee, and S. Chong. 2009. SLAW: A New Mobility Model for Human Walks. In *IEEE INFOCOM 2009*. 855–863. https://doi.org/10.1109/INFCOM.2009.5061995

[17] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9, Nov (2008), 2579–2605.

[18] A. Mei and J. Stefa. 2009. SWIM: A Simple Model to Generate Small Mobile Worlds. In *IEEE INFOCOM 2009*. 2106–2113. https://doi.org/10.1109/INFCOM.2009.5062134

[19] Ivan O. Nunes, Clayson Celes, Michael D. Silva, Pedro O.S. Vaz de Melo, and Antonio A.F. Loureiro. 2017. GRM: Group Regularity Mobility Model. In *Proceedings of the 20th ACM International Conference on Modelling, Analysis and Simulation of Wireless and Mobile Systems (MSWiM '17)*. ACM, New York, NY, USA, 85–89. https://doi.org/10.1145/3127540.3127570

[20] Michal Piorkowski, Natasa Sarafijanovic-Djukic, and Matthias Grossglauser. 2009. CRAWDAD dataset epfl/mobility (v. 2009-02-24). Downloaded from https://crawdad.org/epfl/mobility/20090224. https://doi.org/10.15783/C7J010

[21] F. W. Scholz. 2004. *Maximum Likelihood Estimation*. John Wiley Sons, Inc. https://doi.org/10.1002/0471667196.ess1571.pub2

[22] Claude Elwood Shannon. 2001. A mathematical theory of communication. *ACM SIGMOBILE mobile computing and communications review* 5, 1 (2001), 3–55.

[23] Libo Song and David F Kotz. 2007. Evaluating opportunistic routing protocols with large realistic contact traces. In *Proceedings of the second ACM workshop on Challenged networks*. ACM, 35–42.

[24] Gautam S Thakur and Ahmed Helmy. 2013. COBRA: A framework for the analysis of realistic mobility models. In *INFOCOM, 2013 Proceedings IEEE*. IEEE, 3351–3356.

[25] Cristian Tuduce and Thomas Gross. 2005. A mobility model based on WLAN traces and its validation. In *INFOCOM 2005. 24th Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings IEEE*, Vol. 1. IEEE, 664–674.

[26] Dingqi Yang, Daqing Zhang, Longbiao Chen, and Bingqing Qu. 2015. NationTelescope: Monitoring and visualizing large-scale collective behavior in LBSNs. *Journal of Network and Computer Applications* 55 (2015), 170–180.