

Generation and Analysis of a Large-scale Urban Vehicular Mobility Dataset  

Sandesh Uppoor, Student Member, IEEE, Oscar Trullols-Cruces, Student Member, IEEE, Marco Fiore, Member, IEEE, Jose M. Barcelo-Ordinas, Member, IEEE

Abstract—The surge in vehicular network research has led, over the last few years, to the proposal of countless network solutions specifically designed for vehicular environments. A vast majority of such solutions has been evaluated by means of simulation, since experimental and analytical approaches are often impractical and intractable, respectively. The reliability of the simulative evaluation is thus paramount to the performance analysis of vehicular networks, and the first distinctive feature that has to be properly accounted for is the mobility of vehicles, i.e., network nodes. Notwithstanding the improvements that vehicular mobility modeling has undergone over the last decade, no vehicular mobility dataset is publicly available today that captures both the macroscopic and microscopic dynamics of road traffic over a large urban region. In this paper, we present a realistic synthetic dataset, covering 24 hours of car traffic in a 400-km$^2$ region around the city of Köln, in Germany. We describe the generation process and outline how the dataset improves the traces currently employed for the simulative evaluation of vehicular networks. We also show the potential impact that such a comprehensive mobility dataset has on the network protocol performance analysis, demonstrating how incomplete representations of vehicular mobility may result in over-optimistic network connectivity and protocol performance.

Index Terms—Vehicular mobility, scenario generation, network connectivity, epidemic dissemination.

I. INTRODUCTION

Privately owned cars and public transport vehicles are envisioned to become actual communication hubs in the near future, as heterogeneous network interfaces become integral part of the car equipment and the seamless Internet connection capabilities offered by tablets and smartphones lure passengers’ attention. As a result, vehicular environments have recently emerged as a promising area of research to the telecommunication networking community. The introduction of enhanced infrastructure-based systems, involving the likes of WiMAX and LTE-A technologies, and novel communication paradigms, such as ad hoc and opportunistic networking, has paved the way for the proposal of a flurry protocols specifically designed for the forthcoming communicating vehicles and covering all the layers of the network stack.

The performance evaluation of most of these vehicular networking solutions requires large-scale scenarios, making direct experimental assessments impractical due to their cost and complexity. Simulation becomes then the tool of choice in the validation of new network architectures and protocols for vehicular environments.

Unfortunately, simulative performance evaluation of vehicular networks are often biased by the underlying mobility representation. As a matter of fact, as repeatedly proven in the past [1]–[3], the movement of vehicles can dramatically affect the behavior of network protocols, and an incorrect representation of car traffic can lead to misleading conclusions, even in presence of a flawless network-level simulation. As a result, it is today acknowledged that, for the results of a vehicular simulative campaign to be credible, mobility traces must be employed that capture the unique macroscopic and microscopic dynamics of car movement patterns.

Such considerations have led to substantial progress in the quality of car movement traces for vehicular networking research over the last few years. The simplistic stochastic models employed in early works [1] have been replaced by random mobility over realistic road topologies [4] at first, and by microscopic vehicular models borrowed from transportation research [5] later on. These features were then included in dedicated simulation environments, and integrated with road signalization [6], [7]. Ever since, vehicular mobility simulators have been growing their complexity and features, allowing to accurately simulate the individual movement of vehicles over realistic road topologies [8]. Moreover, in parallel with the evolution of synthetic traces of vehicular mobility, real-world car traffic dataset have grown in number and scale.

In this paper, we go a step further in assessing the impact that realism in the vehicular mobility representation has on the design and evaluation of networking solutions. To that end, we provide a threefold contribution. First, we present in Sec.II a concise yet comprehensive survey of the state of theart in road traffic modeling and tracking, discussing the strengths and weaknesses of the vehicular mobility traces currently available for networking research. Second, we introduce an original synthetic dataset of urban vehicular mobility, that is characterized by an unprecedented combination of scale, detail and realism, and publicly available [9]. We detail the generation process of this new trace through a set of open-source state-of-art tools in Sec.III, discuss the challenges it poses and how to solve them in Sec.IV, and provide an analysis of its features in Sec.V. Third, we show in Sec.VI the impact that such a comprehensive representation of vehicular mobility can have on the evaluation of networking technologies. To that end, we compare our dataset against several traces commonly used in vehicular network simulation, in terms of pure connectivity properties as well as in a practical networking use case. Our analysis evidences the vehicular network performance bias that mobility traces with limited microscopic and macroscopic detail can induce, and leads to the future research directions outlined in Sec.VII.

S. Uppoor and M. Fiore are with INSA Lyon and INRIA, Lyon, France. O. Trullols-Cruces and J. M. Barcelo-Ordinas are with the Universitat Politècnica de Catalunya, Barcelona, Spain. This work was supported by the joint lab between INRIA and Alcatel-Lucent Bell Labs on Self Organizing Networks, and by the Spanish grants TIN2010-21378-C02-01 and SGR2009-1167.
II. RELATED WORK

The relevance of mobility modeling to the simulation of vehicular networks has been long acknowledged, a factor that has pushed the research community to seek for an ever-increasing realism in road traffic traces fed to network simulators. In this section, we overview the body of work on vehicular mobility traces for network simulation. We categorize the datasets based on the nature of their macroscopic traffic data, i.e., the sources employed to determine the time and routes of trips traveled by individual vehicles in the dataset. The most relevant features of the different mobility traces are summarized in Tab. I. For more details on the simulation environments mentioned in the table, we refer the interested reader to [8], [10].

A. Perception and small-scale measurements

A number of synthetic vehicular mobility traces were generated by feeding real-world road topologies and perception-based macroscopic traffic information to microscopic-level simulators, such as SUMO [6] or VanetMobiSim [7]. Notwithstanding the high level of detail granted by the use of such simulators, these traces yield simplistic large-scale features. Indeed, the macroscopic traffic data they employ are based on the authors’ perception of the road traffic in the simulated area, as in the traces of Porto, Portugal [11], of several areas of Turin, Italy [12], and of downtown Karlsruhe, Germany [13]. As an alternative, simple assumptions are made, as in the vehicular mobility trace of the city of Zurich, Switzerland [14], where larger roads attract more traffic. Finally, small-scale measurements conducted by the authors themselves are also used to complement intuition, as in a trace of car traffic in the center of Berlin, Germany [15]. However, all these approaches lack the statistical rigor needed for a realistic representation of the macroscopic traffic distribution, and, as such, they can hardly capture the complexity of traffic flows in urban areas or their evolution over long time periods. Also for these reasons, these traces only cover modest geographical surfaces of few tens of square kilometers, or have a limited time duration in the order of tens of minutes.

B. Road traffic imagery

An original approach to the derivation of the macroscopic traffic information is adopted in [16], where stereoscopic aerial photography is leveraged to capture the vehicle distribution in the city of Porto, Portugal. A private aircraft was flown over the city for two hours in the early afternoon of a weekday, and photographs were shot from the plane every 5 seconds. The flight followed a parallel row pattern so as to cover the whole geographical area of 41.3 km² corresponding to the surface of Porto. By studying the aerial imagery, the authors were able to reconstruct a single snapshot of the positions of 10566 vehicles in the urban area. Given that pictures of different city zones were taken at different moments, car positions in the snapshot refer to different instants: the time error is of 23 minutes between two cars, on average. Although this appears as an interesting way to derive static macroscopic data, its applicability to the generation of mobility traces is not immediate, due to time error above and to the cost of running the aerial photography campaign for a time sufficient to derive an actual mobility trace rather than a single snapshot.

Another recent attempt at using imagery to estimate the macroscopic behavior of car traffic is presented in [17]. There, the authors exploit the pervasiveness of road surveillance cameras to infer traffic densities in ten different urban areas, including London, Sydney and Toronto. This approach provides coarse information on the traffic flows, and could be used for the calibration of microscopic vehicular mobility, similar to the roadside detectors discussed next. However, no actual mobility trace leveraging such data is available to date.

C. Roadside detectors

Induction loops, infrared counters and roadside sensors represent the traditional way to measure vehicular traffic flows in both freeways and urban road networks. In [18], two sets of empirical data are used, obtained from dual-loop and metal detectors from sections of the I-80 Freeway in Berkeley, CA, USA, and of the Gardiner Expressway in Toronto, Canada. The detector information covers a span of 24 hours and allows to determine the per-lane inter-vehicle arrival time and spacing. The data can be fed to a microscopic simulator to derive the position of individual vehicles over time, however its validity is limited to highway environments.

Roadside detector are instead employed in an urban environment within the iTetris project [19]. Synthetic vehicular mobility traces of several areas of the city of Bologna were generated accounting for macroscopic traffic data acquired through 636 induction loops spread over the road network, and complemented by user surveys on usual commuting trips. The main trace covers 20.6 km² in the city center for a period of one hour, featuring the movement of 10333 vehicles. Thanks to the real-world nature of the macroscopic data they are built upon, these traces reach an unprecedented level of realism. Unfortunately, they do not cover large surfaces nor long time periods. Moreover, public accessibility to the mobility traces is not granted yet by the project consortium.

A similar approach has also been taken in [20], where the authors calibrate the microscopic mobility simulation of the city of Luxembourg through traffic counter information gathered by the local Ministry of Transport. As such real-world data only covers major traffic arteries, it is complemented by driver routes inferred from the different nature of geographical zones in the area under study, and used to define traffic flows on medium- and small-sized roads. The resulting mobility trace covers a very large area of 1700 km² and features 150000 car trips. Although a very interesting dataset, the Luxembourg trace focuses on highway and major road traffic, is limited to the morning period, and only accounts for in-bound flows, i.e., traffic moving towards the city center.

D. Socio-demographic surveys

Socio-demographic surveys represent a significant source of information for the derivation of vehicular traffic data. The seminal work in [21] presents a synthetic mobility trace whose macroscopic model is derived by knowledge of drivers’ activity in downtown Portland, OR, USA. The resulting mobility dataset is acknowledged to be very realistic, but only covers 15
minutes of car traffic in an area of 21 km², for a total of 16529 simulated cars. Similar issues affect the trace of Braunsvichweig, Germany, employed in [22], that features highly realistic road traffic description, but it is limited to a small region of 12 km². Moreover, both these datasets were obtained through commercial mobility simulators and are not publicly available.

The largest vehicular mobility trace generated to date reproduces the car traffic in the whole Canton of Zurich, a 65000 km² region of Switzerland [23]. There, 24 hours of car traffic for the whole region are obtained from the Swiss Regional Planning Authority and complemented using the 1994 Swiss National Travel Survey. The resulting synthetic mobility dataset, as well as subsets of the same, are widely employed in the vehicular networking literature. However, the size of the road topology forces the authors to limit the detail of the microscopic-level simulation. Thus, they resort to a queue-based Multi-agent Microscopic Traffic Simulator (MMTS) [24], significantly less accurate than standard fine-grained vehicular mobility simulators based on car-following models, employed in all the synthetic traces previously mentioned. Moreover, and again for scalability reasons, the road topology is pruned down to major traffic arteries, and only the morning and afternoon traffic peaks hours are modeled.

E. Real-world tracking

Most of the previously discussed traces are synthetically generated by injecting macroscopic data into a microscopic mobility simulator. However, a number of traces have been recorded directly in the real world, by logging the position of vehicles during their movements. Traces of this kind with a meaningful scale (i.e., comprising more than a few units) mostly come from mobile fleets, e.g., buses or taxis.

The first of such traces was recorded in Seattle, WA, USA [25], and reproduces the movement of 1200 buses over an area of 5100 km² during a period of two weeks. The data was retrieved through the Automatic Vehicle Location (AVL) system, that is commonly deployed in recent public transport systems, with the position of each bus updated every 2 minutes. A similar technique has been more recently employed in [26] to extract a mobility trace of 1648 buses traveling in the urban area of Chicago, IL, USA. This second dataset also spans over a long period of time, namely 17 days, and has a time granularity of 20-40 seconds, thus higher than that of the Seattle bus trace. Both these traces, apart from being limited to buses, are affected by a low time granularity, which constrains their use to delay-tolerant networking studies.

A similar technique to obtain real-world traces consists in directly retrieving vehicle positions from GPS receivers. Seminal work in such a direction was conducted in Boston, MA, USA [27]. The DieselNet testbed consists of 30 buses traveling over a geographical area of 241.40 km². The size of the testbed makes it of small interest from a mobility viewpoint: in fact, DieselNet is employed to evaluate the communication capabilities of the buses, not their mobility.

A large-scale collection of vehicular GPS recordings is instead the aim of the Shanghai Grid project, where the movement patterns of approximately 4000 taxis and 2000 buses were recorded for several months over an area of 240 km² in Shanghai, PRC. From the overall dataset, mobility traces were extracted for taxis [28] and buses [29]. In the first case, the trace captures the movement of 1171 cabs in the Shanghai inner urban area for three months, with a granularity of 1 minute approximately. In the second, the trajectories of 700 buses over a geographical area of 150 km² are recorded. Although these traces involve a larger number of vehicles with respect to the DieselNet testbed, they still remain limited to a minor subset of the overall traffic in the urban area considered, and are affected by coarse time granularity.
Similar features characterize a taxi mobility trace in Beijing, PRC [30], where the positions of 2927 taxis are updated every minute for 24 hours. The scale is instead much larger in the T-Drive dataset [31], that includes 30000 taxis moving around the same city during three months, although the available refined mobility trace is limited to 10357 cabs and one week of duration. Finally, the routes of 500 taxis San Francisco, CA, USA, can be fetched through the Cabspotting website [32], where the data is constantly updated. As for all the other GPS traces, also these datasets suffer from limited update frequency and road traffic penetration rate.

A different approach is taken by the ongoing German project simTD, which aims at deploying a first large-scale testbed for car-to-X communication [33]. There, the movement of 120 cars owned by the consortium and 300 other vehicles is tracked. Interestingly, the project also targets the generation of a synthetic trace calibrated on the real-world data collected by the probe vehicles. Although the project appears promising, the GPS data is still limited to a small subset of the overall traffic and, at the present moment, it is not publicly available.

F. Discussion

Overall, the ultimate vehicular mobility trace for network simulation should feature all of the following:

- represent the integrality of road traffic;
- have a high time granularity, the position of each vehicle being traced with a order-of-second precision at least;
- compass very large regions (i.e., whole urban areas), with a faithful description of the road layout and signalization;
- provide a realistic representation of the microscopic behavior of individual drivers, accounting for their interactions with other drivers and considering road regulations;
- be realistic also from a macroscopic point of view, by faithfully mimicking the movement of large-scale traffic flows across a metropolitan area, over long time periods.

The first two requirements rule out current traces obtained though real-world tracking, as they are limited to subsets of the vehicles, i.e., buses or taxis, and exhibit reduced temporal detail, i.e., order-of-minute position updates. These traces are today mainly employed for the performance evaluation of delay-tolerant or opportunistic vehicular networks, and remain interesting for applications that fit such scopes. However, significantly higher penetration rates and granularity are needed for more general use cases. The diffusion of recent navigation systems [34], [35], that periodically communicate the car location to traffic management centers, may help real-world tracking to scale up. However, it is also easy to foresee that market rules and privacy concerns will hinder the public disclosure of such data, similarly to what happens today with logs collected by mobile network operators.

The limitations of real-world datasets force us resort to synthetic mobility traces. Indeed, the generation of such traces can accommodate, at the cost of computational complexity, any volume of traffic and time granularity, thus fulfilling the first two requirements above.

In the context of synthetic trace generation, also the third constraint can be easily met, thanks to the availability of accurate real-world road map services. Similarly, recent vehicular mobility simulators allow to satisfy the fourth requirement, as they implement realistic microscopic vehicular mobility models, borrowed from transportation theory and validated through real-world observations. We refer the interested readers to [36] for a detailed discussion of microscopic mobility modeling.

Today’s challenge lies indeed in the last aspect, i.e., attaining macroscopic-level realism in the simulation of road traffic. In transportation theory, the problem is separated into two phases. First, one has to identify the traffic demand, i.e., the start time, the origin and the destination of each car trip in the simulated region, which are stored in a so-called Origin/Destination (O/D) matrix. Then, a traffic assignment model is ran on the O/D matrix, so as to identify the realistic route followed by each driver to reach his/her destination.

The synthetic traces presented before address the issue of macroscopic realism by using different data sources, as clearly outlined by our proposed classification. Some sources are hardly credible due to their own nature (e.g., perception, small-scale measurements) or not scalable due to their complexity (e.g., aerial photography). In the end, and quite unsurprisingly, the most reliable and scalable sources are those commonly employed for transportation engineering (e.g., road traffic detectors and population surveys).

However, none of the traces making use of realistic macroscopic traffic data sources covers large urban regions for long periods of time, and provides at the same time high microscopic-level accuracy. The Berkeley and Toronto traces [18] are limited to one freeway segment. The Bologna [19], Portland [21] and Braunschweig [22] datasets focus on small urban areas of a few tens of km². The Luxembourg [20] and Canton of Zurich [23] traces are those closer to compassing all of the desirable features we outlined above. However, they mainly model traffic on major roads, and are limited to in-bound flows (Luxembourg) or to traffic peak hours (Canton of Zurich).

The dataset we introduce in the next sections overcomes the limitations of existing traces, and thus represents a step forward in the description of vehicular mobility for the network simulation of large-scale urban areas.

III. THE TAPASCologne dataset

The vehicular mobility dataset we introduce in this paper is mainly based on data made available by the TAPASCologne initiative [37] of the Institute of Transportation Systems at the German Aerospace Center (ITS-DLR), that aims at reproducing car traffic in the greater urban area of the city of Köln, Germany, with the highest level of realism possible.

To that end, different state-of-art data sources and simulation tools are brought together, so as to cover all of the specific aspects required for a proper characterization of road traffic. In this section, we detail such tools and the process through which they are combined to generate the mobility dataset.

A. Road topology

The street layout of the Köln urban area is obtained from the OpenStreetMap (OSM) database [38]. The OSM project provides freely exportable maps of cities worldwide, which are contributed and updated by a vast user community. Maps
include information on road layouts and signalization, railways, buildings, and Points of Interests (PoI) such as parks, commercial centers, leisure centers and commercial activities.

In particular, the OSM road information is generated and validated by means of satellite imagery and GPS traces, and it is commonly regarded as the highest-quality road data publicly available today. Indeed, the accuracy of OSM street layouts, comprising highways, major urban arteries and minor roads, often matches that of proprietary ones such as, e.g., Google Maps or Mappy, especially in large metropolitan areas.

We employ the Osmosis tool [39] to filter the OSM data and extract the road topology information for an area of approximately 400 km² around the urban agglomeration of Köln, including almost 4500 km of roads. We then resort to the Java OSM Editor (JOSM) [40] to repair the OSM data file and make it compatible with the microscopic mobility simulator, as detailed in Sec. IV.

B. Microscopic vehicular mobility

The microscopic mobility of vehicles is simulated with the Simulation of Urban Mobility (SUMO) software [41]. SUMO is an open-source, space-continuous, discrete-time traffic simulator developed by the German Aerospace Center (DLR), capable of accurately modeling the behavior of individual drivers, accounting for car-to-car and car-to-road signalization interactions. More precisely, SUMO can import road maps and information on traffic lights, roundabouts, stop and yield signs from multiple formats, including OSM. The microscopic mobility models implemented by SUMO are Krauss’ car-following model [42] and Krajzewicz’s lane-changing model [43], that respectively regulate each driver’s acceleration and overtaking decisions, by taking into account a number of factors, such as the distance to the leading vehicle, the traveling speed, and the acceleration and deceleration profiles. These models have been long validated by the transportation research community, a fact that, jointly with the high scalability of the simulator, makes of SUMO the most complete and reliable among today’s open-source microscopic vehicular mobility generators. The version we employed for the dataset generation is 12.3.

C. Traffic demand

The traffic demand information on the macroscopic traffic flows across the Köln urban area is derived through the Travel and Activity PAtterns Simulation (TAPAS) methodology [44]. This technique generates the O/D matrix by exploiting information on (i) the population, i.e., home locations and socio-demographic characteristics, (ii) the points of interests in the urban area, i.e., places where working and free-time activities take place, and (iii) the time use patterns, i.e., habits of the local residents in organizing their daily schedule [45]. Within the context of the TAPAS-Cologne project, the aforementioned TAPAS methodology is applied on real-world data collected in the Köln region by the German Federal Statistical Office, including 30700 daily activity reports from more than 7000 households [46], [47]. The resulting O/D matrix faithfully mimics the daily movements of inhabitants of the area over 24 hours, for a total of 1.2 million individual trips.

D. Traffic assignment

The actual assignment of the vehicular traffic flows described by the TAPAS-Cologne O/D matrix over the road topology is performed by means of Gawron’s algorithm [48]. This traffic assignment technique computes the fastest route for each vehicle, and then assigns to each road segment a cost reflecting the intensity of traffic over it. By iteratively moving part of the traffic to alternate, less congested paths, and recomputing the road costs, the scheme finally achieves a so-called user equilibrium. Additionally, since the intensity of the traffic demand varies over a day, the traffic assignment model must also be able to adapt to the time-varying traffic conditions. Indeed, Gawron’s algorithm satisfies such a requirement, thus attaining a so-called dynamic user equilibrium. Gawron’s is one the most popular traffic assignment techniques developed within the transportation research community, and allows to reach a road capacity utilization close to reality and significantly higher than that obtained with, e.g., a standard weighted Dijkstra algorithm.

E. Simulation

The individual components presented above are combined as depicted in Fig. 1 in order to generate the vehicular mobility dataset. First, the information contained in the TAPAS-Cologne O/D matrix are used to identify the boundaries of the exact simulation region, extract the associated map from OSM and filter it so as to remove unneeded content that does not concern the road layout. Then the OSM map is converted to a format readable by SUMO, and fed to the microscopic mobility simulator. The TAPAS-Cologne O/D matrix is also used as an input to Gawron’s algorithm, which, in turn, determines an initial traffic assignment and provides it to SUMO. Then, a first vehicular mobility simulation can be run with SUMO, and, once finished, a feedback on the resulting traffic density over the road topology is sent back to Gawron’s algorithm. Based on such new information, a new traffic assignment is computed, and a second SUMO simulation is run. The process is repeated until a traffic assignment is generated that allows to sustain the whole volume of the traffic demand.

IV. REPAIRING THE DATASET

Making all of the previous components work together is not a trivial task. Indeed, when simply running the SUMO simulation with the data sources made available by OSM and TAPAS-Cologne, the result is plain unusable. In Fig. 2(a), we plot the time evolution of the number of vehicles that (i) are traveling on the road topology, (ii) have successfully ended their trip by reaching the their destination, (iii) are waiting to enter the road topology, which they cannot presently do due to an excessive congestion of the road segment they are supposed
From the plot, we can note how the number of traveling vehicles present in the simulation rapidly grows up to exceed a hundred thousands units, a figure completely unrealistic for a city the size of Köln. Additionally, such a number does not tend to decrease as one could expect once the morning traffic peak is exhausted; instead, it keeps growing indefinitely. It is also possible to observe that the number of vehicles that end their trip grows very slowly over time: in fact, from the values portrayed in the figure, only a very small fraction of the cars that are present on the road topology can reach their destination. Finally, the number of vehicles that are waiting to enter the road topology, which we would like to stay as close as possible to zero, grows to hundreds of thousands of units.

The mean travel time, in Fig. 2(b), also shows a quite unrealistic behavior, as more than half an hour is required, on average, for a driver to reach its destination at 10:00 am, when the traffic should be sparse. Similarly, the average speed of vehicles, in the same plot, tends to zero as the time elapses.

These results are clear symptoms of how the road topology cannot sustain the volume of cars injected according to the traffic demand model. Indeed, when looking at a snapshot of the car traffic in the region, it is evident how the simulation quickly reduces to a huge traffic jam. As an example, Fig. 3 depicts the map of the road traffic at 7:00 am: the road topology is mostly covered by bright red dots, representing cars stuck in heavily congested traffic. Next, we discuss the reasons for such a result, and present solutions to them.

A. Over-comprehensive and bursty traffic demand

The original TAPASCologne O/D matrix yields the traffic demand volume depicted in Fig. 4(a), which shows the number of vehicles injected in the whole road network every second. By analyzing the O/D matrix, its source data, and its effect on the microscopic mobility simulation, we identified and fixed the following three problems.

First, the demand in the O/D matrix is not limited to the vehicular traffic; rather, it includes information on the daily trips of all Köln inhabitants, independently from whether their walk to their destination, or employ public transports, or take a car as either passengers or drivers. Clearly, we are only interested in the latter kind of mobility, since the volume of vehicular traffic directly maps to that of car drivers. According to [45, Fig. 4], car drivers account for approximately 50% of the overall trips in the TAPASCologne O/D matrix: thus, we adjusted the O/D matrix by only considering that one trip every two concerns the movement of a vehicle.

Second, the original demand presents an unrealistic variability in the injected traffic over short time scales. This can be observed in Fig. 4(a), where, within the span of a few minutes, peaks up to 200 vehicles/s in the injected traffic alternate with instants of reduced injected traffic as low as 10 vehicles/s. Such an excessive burstiness is hardly observable in the reality, especially considering that the injected traffic is aggregated over a very large area. Indeed, we observe these peaks of traffic to be a major cause of congestion, forcing large masses of cars to try entering the road topology at the same time, and thus creating sudden traffic jams. In order to address this issue, we smooth down the original O/D matrix, by adding to the departure time of each vehicle a random offset uniformly distributed in the interval $[-5, 5]$ minutes. This allows to remove the injection bursts, yet retaining the traffic demand properties over larger time scales.

Third, the legacy TAPASCologne demand only includes trips starting or ending within the 400-km$^2$ simulated region. Therefore, vehicles crossing the area entirely, mainly highway traffic around the urban center, are not accounted for. We resort to data made available by the traffic information system of the Nordrhein-Westfalen Ministry of Transport [49] and introduce the missing highway traffic in the repaired demand. The daily evolution of the final traffic demand is depicted in Fig. 4(b).
Fig. 5. Example of wrong restriction enforcement in OpenStreetMap data.

B. Inconsistent road information

A second source of errors in the simulation was identified in the OSM road data. Although very complete from a topological viewpoint, the OSM map embeds information at times inconsistent with respect to reality. The impact of such inconsistencies, albeit negligible on most of the usages of OSM, can be dramatic for the simulation of vehicular mobility.

A first type of inconsistency is represented by wrong traffic movement restrictions enforced on some road segment. Consider the situation in Fig. 5: there, a no left turn restriction (bottom of the left image) is present in the OSM road information, for the East-West lane of the horizontal road (right plot). This prevents a car traveling along such a lane to turn left, as in the example in the figure. The OSM data contains at times restrictions of this kind which are not actually there in the real world: the shape of the street layout is not affected by such errors, however the microscopic traffic simulation is, since they can cause vehicles to perform long detours or to get stuck by waiting indefinitely for a possibility to turn and continue their journey. We also identify wrong restrictions by checking the features of congested intersections and T-junctions against the real-world road signalization through the Google Street View service: when necessary, we correct the OSM data via visual inspection. This allowed us to fix approximately one thousand erroneous restrictions in the area under study.

A second type of inconsistency is that of correct movement restrictions of this kind which are not actually there in the real world: the shape of the street layout is not affected by such errors, however the microscopic traffic simulation is, since they can cause vehicles to perform long detours or to get stuck by waiting indefinitely for a possibility to turn and continue their journey. We also identify wrong restrictions by checking the features of congested intersections and T-junctions against the real-world road signalization through the Google Street View service: when necessary, we correct the OSM data via visual inspection. This allowed us to fix approximately one thousand erroneous restrictions in the area under study.

A second type of inconsistency is the presence of road information not recognized by the conversion tool. E.g., attributes with two values are considered as incorrect by the converter, and the associated roads are not included in the toplogy used for the simulation. An example is shown in Fig. 7: the double value of the source field (left) causes SUMO not to account for the associated road in simulation (right, a road should connect the North and South branches). We corrected the OSM data so to make all attributes compatible with the SUMO converter.

C. Flawed road topology conversion

The OSM road information is natively imported by SUMO through an automated conversion process that proves not to be error-free. A first cause of problems is the presence of road information not recognized by the conversion tool. E.g., attributes with two values are considered as incorrect by the converter, and the associated roads are not included in the topology used for the simulation. An example is shown in Fig. 7: the double value of the source field (left) causes SUMO not to account for the associated road in simulation (right, a road should connect the North and South branches). We corrected the OSM data so to make all attributes compatible with the SUMO converter.

A second critical aspect is the fact that the topological information in OSM is, at times, simply unfit to be directly converted to the SUMO street layout. An example is depicted in Fig. 8, where the real-world aerial photography of a rather complex intersection (top left), the associated Google map information (top right), and the OSM road topology (bottom left) match. However, the conversion of the latter within SUMO results in an exceedingly intricate intersection, where vehicles get stuck and rapidly form a permanent traffic jam (bottom right). The reason for such a simulation result is that, since two segment links (white dots in the bottom left plot) are present, SUMO interprets the OSM topology as if two road junctions co-existed, one next to the other. As a consequence, the number of traffic lights that regulate the car flows into the crossroad is doubled, and yield signs are placed right at the middle of the intersection: the result is the impossibility for vehicular in-flows to correctly merge at the intersection.
In order to fix such a problem, we acted directly on the OSM road information, by joining road segment links that refer to the same physical intersection. Such an operation allows then a correct conversion by SUMO, so that no traffic jams are observed anymore at the road junctions. Additionally, we corrected in several cases the number of lanes entering and leaving an intersection, so as to match aerial photography data.

The third problem we remarked in the OSM-to-SUMO conversion lies in the traffic light deployment. OSM road information already includes data on the presence or absence of traffic lights at road junctions, and SUMO automatically sets the green/red periodicity according to the priority of the roads entering each junction. However, the SUMO converter also employs by default a technique to place additional traffic lights over the street layout. After having verified the negative impact of such a traffic light guessing, we disabled it. In addition, we identify a number of situations where the presence of traffic lights is not beneficial, and indeed does not correspond to reality: in particular, this is often the case for intersections formed by peripheral roads with identical priority but very unbalanced traffic. Indeed, the similar traffic light periodicity assigned to each road in such a context leads to long queues on the trafficked roads. We thus removed such traffic lights from the OSM data so as to be consistent with the real world.

**D. Simplistic default traffic assignment**

Running the microscopic mobility simulation, with the traffic demand corrected as from Sec. IV-A and the road topology fixed as from Sec. IV-B and Sec. IV-C, still results in large congestion and continuous traffic jams all over the street layout. The reason lies in the traffic assignment, i.e., the way drivers choose the route to reach their intended destination. Indeed, SUMO employs a simple Dijkstra’s algorithm on the road topology graph, by weighting edges, i.e., road segments, on their length, as well as on the maximum speed they allow: clearly, shorter and faster roads are preferable, and thus are associated with smaller weights. Unfortunately, this means that drivers having similar origin and destination points all choose the same routes for their trips: as a result, they concentrate on major roadways, which are rapidly filled to their maximum capacity, whereas slower or minor roads remain unused. Obviously, high-speed roads alone cannot handle the whole demand, and thus the traffic assignment needs to be improved.

To that end, we resort to the traffic assignment solution proposed by Gawron and presented in Sec. III-D. Such a technique achieves a dynamic user equilibrium by iterating until no significant difference is observed between subsequent simulations. Fig. 9 shows the evolution of traffic while iterating the assignment algorithm. The number of vehicles traveling at the same time over the road topology, in Fig. 9(a), tends to explode during the first iterations, as it happens before patching the demand and road topology. However, as Gawron’s algorithm iterates, the car traffic is progressively reduced, since drivers tend to employ the different available routes and thus better exploit the capacity of the road network. Fig. 9(b) confirms that iterations significantly improve the traffic conditions, as they increase the number of vehicles that reach their destination, successfully ending their trip. Similar trends are observed for the other traffic metrics, and in all cases iterations after the 35th do not produce any noticeable improvement. Therefore, in the following, we consider the traffic assignment obtained at the 35th iteration.

**V. ANALYSIS OF A LARGE-SCALE URBAN MOBILITY TRACE**

The resulting dataset comprises seven hundred thousand car trips in the Köln larger metropolitan area, over a period of 24 hours. The simulated traffic now mimics the normal daily road activity in the region, as the fixed road topology can accommodate the updated traffic demand and assignment.

Evidences of the correct behavior of the simulated mobility are given in Fig. 10(a). By comparing it to the equivalent plot before repair, in Fig. 2(a), it is clear that the number of traveling cars now follows the traffic demand, with peaks during the morning (from 7:00 am to 9:00 am) and afternoon
reach their destinations, and the number of vehicles waiting ended trips now grows over time, as more and more drivers at around noon, can also be observed. Also, the number of such as very low traffic at night and a lower traffic peak}

maximum of 15,000 vehicles travel at the same time over (from 4:30 pm to 6:00 pm) rush hours. An approximate maximum of 15,000 vehicles travel at the same time over the road topology, at around 8:00 am. Real-world behaviors, such as very low traffic at night and a lower traffic peak at around noon, can also be observed. Also, the number of ended trips now grows over time, as more and more drivers reach their destinations, and the number of vehicles waiting to enter the simulation is reduced to values close to zero. The average travel time and speed recorded during the morning, in Fig. 10(b) confirm the previous results, as we observe quite constant behaviors, only modified during the peak hours.

As a result, the road traffic at 7:00 am, in Fig. 11, looks significantly better than the original one, in Fig. 3. Indeed, highways are denoted by a bright blue, corresponding to speeds higher than 90 km/h. Large portions of the urban roads are in violet, indicating fluid traffic conditions. The traffic appears only congested in the city center, where dark red regions are visible: however, no significant bright red area is present, meaning that vehicles move at slower speeds, from 30 to 50 km/h, but are not stuck as it was previously the case.

In Fig. 12, we analyze some interesting features of the trips in the dataset. The length of routes traveled by drivers in the simulated region, in Fig. 12(a), appears to be mostly in the order of a few kilometers, as one would expect in a urban area. Yet, trips longer than 10 km are not uncommon, as one driver over five has to travel such a long road: those are mostly commuters living in the suburbs of Köln. Also, note the peaks around 15 and 20 km: these are the contributions of the traffic crossing the whole region, whose route length is thus constrained to the exact length of the highway segment in the simulated area. The route length has a clear impact on the distribution of trip durations, in Fig. 12(b), where we find the same shape of Fig. 12(a) and we can observe how half of the trips last less than 8 minutes.

Interestingly, we found the macroscopic traffic simulated in the final TAPASCologne dataset to nicely match that observed in the real world, through real-time traffic information services. In Fig. 13, we compare the simulation output, at 5:00 pm, with the road traffic information retrieved through the ViaMichelin and GoogleMaps live traffic services at the same hour. This represent a critical period of the day, in the middle of the afternoon traffic peak, and key features of real-world mobility patterns are faithfully reproduced in the dataset: e.g., the congestion on the highways around the city, where commuters merge with long-distance travelers passing through the region, or generalized but discontinuous heavy traffic in the city center, especially along major roads. We regard the result as very encouraging, although we acknowledge that more rigorous tests are needed to fully prove the realism of the dataset. The latter would require the use of fine-grained real-world data, including (i) road traffic detector data such as induction loops, infrared counters or traffic camera imaging, (ii) cellular handoff information from phones onboard vehicles, and (iii) GPS records collected by automobile service providers. To some extent, such information could be compared against traffic counts and travel times observed in the synthetic trace. Unfortunately, none of these data is publicly available, as transportation authorities, telecom operators and car manufacturers keep them highly reserved for security, privacy and industrial secrecy reasons.

VI. IMPACT OF MOBILITY ON VEHICULAR NETWORKING

Having introduced a new large-scale vehicular mobility dataset, deemed to yield a higher level of realism than currently available traces, we are now interested in understanding which impact the additional realism brought by our dataset has on the evaluation of vehicular network solutions.

To that end, in Sec. VI-A, we first introduce the vehicular traces commonly employed in the literature that we take as a
reference. Then, in Sec. VI-B, we analyze the connectivity of the vehicular ad hoc network in the TAPAS-Cologne dataset, and compare it with those observed in the reference scenarios above: by focusing on the underlying network topology, we draw results that are independent from the service model or network protocol stack that build above it. Finally, in Sec. VI-C, we observe the effect that the mobility representation has on a specific networking use case, i.e., the epidemic opportunistic communication.

In the following, in order to characterize the vehicular communication capabilities, we assume a simple disc model, with a range of 100 m, set according to experimental results on vehicle-to-vehicle communication [50]. We reckon that the disc model is a very simplistic approach, however this choice allows us to point out the impact of mobility, which is the objective of this paper, and avoid biases due to the irregular signal propagation of urban environments. Introducing a realistic signal propagation for vehicular communication in urban environments is out of the scope of this paper, but it is part of our future goals, as discussed in Sec. VII.

A. Reference scenarios

We compare our dataset to the following references. The Zurich region trace is a 400-km² subset of the Canton of Zurich dataset introduced in [23]. As discussed in Sec. II, the level of detail of this trace is relatively low, from several viewpoints. First, the road map is coarse, only accounting for highways and main traffic arteries in Zurich: as shown in Fig. 14(a), the resulting street layout is significantly less detailed than that of the Köln area in Fig. 11. Second, the Multi-agent Microscopic Traffic Simulator (MUMTS) used for the trace generation employs a queuing approach [24], faster but less accurate than the car-following model SUMO adopts. Third, the traffic demand is not as accurate as the one we dispose of, as portrayed in the left plot of Fig. 15. There, we compare the traffic volume recorded in our dataset with that in the Zurich region scenario. Although the general trend is the same, with traffic peaks early in the morning and in the middle of the afternoon, the traffic demand in the Zurich region scenario unrealistically drops to zero between 10:00 am and 2:00 pm, as well as after 8:00 pm. Clearly, the Zurich region dataset models the traffic peak hours only.

The Köln pruned trace is a simplified version of our TAPAS-Cologne dataset. It is generated using the Köln OSM road topology presented in Sec. III-A, the TAPAS travel demand of Sec. III-C and Gawron’s traffic assignment introduced in Sec. III-D. However, it only features major road arteries, as show in Fig. 14(b). The trace adopts a simplified microscopic model, namely the Constant Speed Motion with Pauses model [7] calibrated so that the evolution of the average speed over time is identical to that recorded in the TAPAS-Cologne dataset. Finally, the travel demand is limited to the morning and afternoon traffic peaks, as displayed in Fig. 15. The rationale behind the Köln pruned trace is to have a dataset that yields similar characteristics to the Zurich region one (i.e., real-world road layout limited to major traffic arteries, simplistic representation of the microscopic car movement, realistic macroscopic traffic flows limited to the rush hours) in the same urban area of the TAPAS-Cologne dataset.

The third reference scenario is named Zurich downtown, and corresponds to the Zurich dataset [14] of Sec. II. It thus covers around 12 km² of the Zurich road network, depicted in Fig. 14(c). The car traffic was simulated using the Generic Mobility Simulation Framework (GMSF) [14], which includes accurate road topology information from the Swiss Geographic Information System, and high-detail car-following microscopic mobility via the Intelligent Driver Model (IDM). However, as already stated in Sec. II, the traffic demand used in the Zurich downtown trace is limited in terms of realism and extent, only comprising 20 minutes of mobility. These limitations are evident in the plot of Fig. 14(c), where the traffic demand leads to exceedingly low speeds, around 20 km/h, even along major urban arteries. The right plot of Fig. 15 is even more clear, comparing the traffic volume of the Zurich downtown scenario to that recorded in the inner 20 km² of the whole Köln region described in the TAPAS-Cologne dataset. This subset of our dataset, introduced as a fair comparison term to the smaller reference traces, will be referred to as Köln downtown. It is clear that the Zurich downtown scenario only captures a very marginal fraction of the daily traffic.

Similar observations apply to the fourth reference scenario, whose traffic demand was built based on direct observations by the authors. This trace, hereinafter Turin downtown, maps
to the Turin trace [12] of Sec. II. It describes the road traffic in an area of 20 km$^2$ of Turin, Italy, as depicted in the right plot of Fig. 14(d): there, we can observe the localized heavy congestion that characterized the mobility in the area. The dataset has also a limited duration, covering one hour of moderate traffic, as displayed in the right plot of Fig. 15.

B. Network connectivity

Let us define as cluster a group of vehicles such that a (multi-hop) path exists between any pair of them at a given time instant [2]. Therefore, vehicles belonging to different clusters cannot communicate at that time, neither directly nor passing through other vehicles. Fig. 16 shows the average number of clusters observed in each scenario. The plot also reports the mean and standard deviation of the cluster size, i.e., the number of vehicles that belong to a cluster.

By looking at the larger region scenarios, on the left of the plot, a clear difference emerges between our dataset and that of Zurich: the latter results in a much more connected network than the former, with vehicles grouping in less than one third of the clusters we record in our dataset. One could wonder whether that is the effect of a much higher percentage of singletons, i.e., clusters composed of one isolated node, in the TAPAScologne-based trace: from the figure, however, a similar fraction of singletons is present in both scenarios, accounting for approximately 60% of the overall clusters. The reason for such a difference is instead explained by the extremely high average and standard deviation of the cluster size in Zurich scenario: these are evidences of the coexistence of the many singletons with several giant components that gather a large portion of the vehicles. Such giant components cannot instead be found in the Köln scenario, where clusters tend to be much smaller and more uniform in size.

Such an average value analysis is confirmed by the results in Fig. 17. The left figure, detailing the evolution of the cluster number over time, shows that the behavior previously described is actually not influenced by the daytime. Indeed, the number of clusters in the Zurich scenario is significantly lower than that recorded in our Köln dataset at all times. The plot also confirms that the Zurich trace is unusable between 10:00 am and 2:00 pm, as well as after 8:00 pm.

The right image of Fig. 17 confirms instead our intuition on the presence of giant components in the Zurich trace. As a matter of fact, the cumulative distribution function (CDF) of the cluster size, in the outer plot, proves that the TAPAScologne dataset contains a higher number of small (20 vehicles or less) clusters. Conversely, the complementary CDF, in the inset plot, highlights the tail of the distribution, allowing to observe a much more probable (0.1%) presence of very large clusters (more than 10,000 vehicles) in the Zurich scenario. In the TAPAScologne trace, the largest cluster does not exceeds 4,000 cars, and appears with orders-of-magnitude lower probability (0.001%), even if the number of cars concurrently traveling in the two traces is comparable, see Fig. 15.

At this point, one could rightfully ask whether the connectivity differences between the Kölnc and Zurich region traces are imputable to the diverse urban areas represented in the two mobility datasets. Indeed, the two scenarios are characterized by dissimilar road topologies and road traffic flows, which could justify the non-comparable connectivity results. However, let us observe the behavior of the Kölnc pruned trace in Fig. 16 and Fig. 17. Despite the fact that it portrays road traffic in the Kölnc area, the Kölnc pruned trace results in a vehicular network featuring a time-varying number of clusters nicely matching that of the Zurich region dataset. Moreover, it yields a low average cluster size with a very high standard deviation as well as 10,000-node clusters appearing with 0.1% probability, exactly as in the Zurich region case.

Therefore, we conclude that it is not the underlying urban environment that determines the differences between the TAPAScologne and Zurich region scenarios, rather the diverse level of realism of the mobility description. In particular, the emergence of unrealistically large components in the Zurich trace is imputable to the combination of reduced road topology information and low microscopic mobility detail, the same features we observe in the Kölnc pruned trace. Considering only major roads and a simplistic microscopic approach leads therefore to a very homogeneous traffic over the few traffic arteries, that act then as seamlessly connected backbones for the vehicular ad hoc network.

As far as the downtown scenarios are concerned, their average behavior in terms of clustering is depicted in the right portion of Fig. 16. The number of clusters is significantly lower than that observed in the larger region scenarios, consistently
with the reduced size of the areas. In this case, results are more similar through the different datasets, however the connectivity of the vehicular network in the Köln trace presents a slightly higher variability in both cluster number and size: this is due to the fact that our dataset captures the evolution of the traffic over the day, whereas the other scenarios are only representative of a short time span characterized by quasi-static network clustering properties. The left plot in Fig. 18 supports such a conclusion. There, we also remark that the largest component sizes observed in the Köln region and downtown scenarios match: such a component appears during the rush hours in the trafficked city center.

The right plot in Fig. 18 focuses on the node degree, i.e., the number of communication neighbors of a vehicle, an especially relevant metric in small-scale scenarios. We can note that vehicles in the Köln scenario tend to have smaller 1-hop neighborhoods, with only 5% of the them having more than 30 neighbors, while 60% have less than five nodes within communication range. On the contrary, in the reference downtown scenarios, the fraction of vehicles with large neighborhoods of more than 30 nodes grows to 25%, and only 20 to 30% of the nodes have five or less neighbors.

Summarizing our findings, we can conclude that the topology of a vehicular network built on the car traffic described by our dataset is sensibly different from those obtained with currently available mobility traces. More precisely, when compared with the standard large-scale mobility trace, i.e., Canton of Zurich, our dataset appears significantly more detailed. In turn, such additional detail leads to a less connected and less stable network. Considering instead small-scale traces, our dataset shows an equivalent level of detail, but a much more variegated behavior, as it models a whole day rather than a few tens of minutes of road traffic. As a result, the Köln dataset allows to observe a connectivity variability that is not captured by the other traces. Overall, the discussion brings us to conjecture that evaluating network protocols or architectures through low-detail or spatially- and temporally-limited mobility traces is a risky practice, that can lead to over-optimistic performance results. Network simulation results in the next section substantiate this conclusion.

C. Epidemic dissemination

We now consider a networking application use-case, and evaluate the effect that a high-detail large-scale mobility dataset has on the system performance. More specifically, we focus on the epidemic dissemination of some small content (e.g., information on the current status of road traffic in the area or a map update for the on-board navigation system) throughout the whole network. We assume a susceptible-infected model, where a node that has been reached by the content (infected) can forward the data to the uninformed vehicles (susceptible) it encounters. We consider that the content can be transmitted over the wireless channel in a few milliseconds, and that transmissions occur in a broadcast fashion, so that all 1-hop neighbors of the sender can receive it at once. Finally, we neglect medium contention and channel errors: we recognize these to be strong assumptions, however (i) they make simulations computationally feasible in the very large-scale scenarios we consider, and (ii) they do not significantly impact our comparative evaluation, since they affect in a similar manner the different mobility datasets.

The epidemic dissemination is run in the 400-km² region scenarios, i.e., the TAPAScologne, Zurich region, and Köln pruned datasets. A source, located at the city center, broadcasts the content for the first time at 7:00 am, i.e., during the morning rush hour. We test different technology penetration rates, i.e., ratios of cars equipped with vehicle-to-vehicle communication interfaces and participating in the network.

The left plot of Fig. 19 portrays the dissemination ratio, i.e., the percentage of vehicles reached by the content, versus time, when all the vehicles take part in the dissemination process. We observe that the dissemination is very fast in all scenarios, as almost all vehicles are informed in a few minutes. However, while the curve obtained through the TAPAScologne dataset is more gentle, with two minutes required to reach 80% of the network, the Zurich region and Köln pruned traces lead to a much faster spreading. In fact, 80% of the vehicles are informed in a few seconds in these scenarios. This is an artifact of the unrealistically high connectivity that, as our previous topological analysis unveiled, characterizes the Zurich region and Köln pruned vehicular networks: it is indeed the presence of large clusters including most of the road traffic that makes the spreading exceedingly fast.

The right plot of Fig. 19 summarizes the initial spreading performance of the dissemination, by reporting the latency in reaching different quantiles of the dissemination ratio. For each penetration rate, in abscissa, we compare the time required to reach 5%, 25%, 50%, 75% and 95% of the vehicles, in the region scenarios. The result shows that the content is successfully disseminated throughout the whole network in all cases, however it is also clear that the very high connectivity granted by the Zurich trace leads to latencies that, under the different penetration rates, are from two to six times lower than those recorded in the Köln scenario. Interestingly, the same is true for the Köln pruned dataset: this confirms that it is the realism of the mobility description that leads to the network performance difference, rather than the specificities of the urban areas considered.

Not only the latency, but also the content survivability is a metric of interest in the study of epidemic dissemination. The latter represents the capacity of the content to self-sustain in the network, so that new vehicles starting their trips can immediately receive it. The left plot of Fig. 20 portrays the evolution of the dissemination ratio over the whole day,
after the initial injection at 7:00 am. The curves refer to the TAPAScologne and Zurich datasets, since the Köln pruned trace performs once more close to the Zurich one and is omitted for the sake of clarity. Penetration rates of 1.0 and 0.1 are considered. The figure clearly shows how the content dies out at around 10:00 am in the Zurich scenario, whereas it is able to self-sustain during the whole day in the Köln dataset. Indeed, the sudden drop in the dissemination ratio observed in the first scenario is imputable to the disappearance of road traffic after the morning peak: the complete absence of vehicles after 10:00 am makes it impossible for the content to survive.

However, as long as traffic is present in the Zurich network, the elevate connectivity leads to extremely good survivability performance under any penetration rate. In the right plot of Fig. 20, we provide a zoom on the morning rush hours, that outlines how the dissemination rate is constant at 100% in the Zurich trace, when all vehicles participate in the system. In fact, such performance is only slightly affected by the penetration rate in the Zurich scenario, as limiting the network to 10% of the nodes still allows to maintain the content alive throughout 95% of the network. It is interesting to note that the dissemination self-sustains in a comparable manner in the Zurich scenario with 10% of the nodes and in the Köln one with 100% of the nodes. In the latter scenario, reducing the penetration rate to 0.1 results instead in a dramatic loss of performance, the content surviving in 60% of the network only.

These results confirm that limited realism in the representation of the vehicular mobility can have a significant impact on the performance evaluation of networking solutions. Indeed, even a dataset such as the Canton of Zurich one, widely adopted in the literature and deemed to be realistic enough, can introduce significant performance bias. As we showed, its limited microscopic precision and incomplete road traffic demand can lead to dissemination results that are overly positive on a time span of a few hours and excessively poor on a daily basis. The dataset we introduce in this paper raises the bar in that sense, as its very detailed yet large-scale representation of road traffic represents a more consistent ground on which to test vehicular network solutions.

VII. CONCLUSIONS AND FUTURE WORK

We presented a novel large-scale urban vehicular mobility dataset. The dataset is obtained by considering a realistic road topology, microscopic driver behaviors and macroscopic traffic flows, over an area of 400 km² during 24 hours. We discussed the dataset generation process, and analyzed the salient features of the resulting vehicular mobility. We compared our dataset with traces commonly employed in the literature, in terms of connectivity features of the vehicular network and of epidemic dissemination performance. The results showed that simplistic assumptions on the macroscopic or microscopic dynamics of road traffic can greatly affect the properties of network topology. This, in turn, proved to introduce a significant risk of biasing the performance evaluation of network protocols for vehicular environments. These results demonstrate that the levels of macroscopic and microscopic realism granted by our dataset are necessary to networking studies.

To the best of our knowledge, the dataset we present in this paper constitutes the most complete vehicular mobility trace to date, and is freely available through the project website [9]. In particular, the SUMO configuration data is publicly available, which makes our dataset usable with bidirectional simulation tools, such as TraNS [51], Veins [52], or iTetris [19], that allow to dynamically modify the road traffic as a consequence of messages exchanged in the vehicular network.

We remark however that a number of issues remain open. First, real-world data allowing a more rigorous validation of the mobility patterns is required, as discussed in Sec. V. Secondly, realistic signal propagation in urban vehicular environments is to be integrated with mobility datasets, so as to grant an additional layer of realism in simulation. To that end, we invite the interested reader to refer to [53] and references therein, that well summarize how radio propagation models could be implemented within mobility datasets such as ours. As a final aspect, it is desirable that multiple traces of similar or higher quality are generated for different cities, and evaluated under diverse vehicular network protocols: this would allow to assess (i) the actual impact of the road layout on the vehicular network performance, and (ii) whether the level of realism and detail of the Köln mobility dataset is actually sufficient to networking studies.

REFERENCES


Sandesh Updoo (S’10) is a PhD candidate within the SelNet action of the Inria–Alcatel Lucent Bell Labs joint laboratory. He is part of the Inria UrbaNet team, hosted by the CITI Research Laboratory in Lyon, France. He received a Master of Technology degree from the Network Engineering Department of the Manipal Institute of Technology, India, in 2009, and he worked as an early stage researcher in FORTH, Greece, in 2010. His research focuses on vehicular networks and mobility modelling.

Oscar Trullols-Cruces (S’09) received both B.Sc. and M.Sc degrees in computer engineering from the Technical University of Catalonial, Barcelona, Spain, in July 2008. He joined the the Computer Networking Research Group in July 2007, as a PhD candidate. He visited the Royal Institute of Technology (KTH), Stockholm, Sweden, during spring semester in 2009. His research interests include vehicular networks, mobile ad hoc networks, delay tolerant networks and mobility modeling.

Marcio Fiore (S’05, M’09) is an Associate Professor at INSA Lyon and a researcher within the Inria UrbaNet team hosted by the CITI Lab. He received M.Sc degrees from the University of Illinois at Chicago and Politecnico di Torino, in 2003 and 2004, respectively, and a PhD degree from Politecnico di Torino, in 2008. He has been a visiting researcher at Rice University and Universitat Politècnica de Catalunya. His research focuses on mobility characterization and its impact on networking.

Jose M. Barcelo-Ordinas (M’04) is an Associate Professor at Universitat Politècnica de Catalunya (UPC). He received his PhD and M.Sc at UPC (1998 and 1991). He joined the CompNet research group in 1993. He has participated in several European projects (EXPLOIT, BAP, EXPERT, NET-PERF, MOEBIUS, WIDENS) and Networks of Excellence (EuroNGI, EuroNFI, EuroNF). His research areas are mobile networks, delay tolerant networks, VANETs and wireless sensor networks.