Crowd Management: A New Challenge for Urban Big Data Analytics

Clayson Celes[1,2], Azzedine Boukerche[1], and Antonio A. F. Loureiro[2]
[1]PARADISE Research Laboratory, EECS - University of Ottawa, Canada
[2]Department of Computer Science, Federal University of Minas Gerais, Brazil
{claysonceles, loureiro}@dcc.ufmg.br, boukerch@site.uottawa.ca

Abstract

The increasing availability of tremendous amounts of data generated by people, vehicles, and things have provided unprecedented opportunities for understanding human behavior in the urban environment. At the same time, crowd management systems can benefit city planning, emergency control, and mobile network design. In this work, we exploit urban data as a way of analyzing crowd behavior. We analyze the types of crowd situations, describe the major types of urban data, and highlight their strengths and weaknesses. We then discuss the key research challenges and opportunities in the analysis of urban environments. Moreover, through case studies, we explain how to apply urban data for spatial, temporal and semantic observations of crowd situations.

Index Terms

Crowd management, Urban data, Urban sensing, Data analysis, Human behavior, Mobility, Big data

I. Introduction

According to the United Nations, 68% of the world's population will live in urban areas by 2050. With population growth in large urban centers, a natural tendency is the formation of crowds of people in several situations. A constant concern of the authorities is the formation, management, and control of crowds to prevent circumstances of disorder and provide efficient solutions to mitigate problems related to urbanization [1].

Crowd management and crowd control are operationally different [2]. The former refers to the set of measures that must be taken to facilitate the movement and enjoyment of people. The latter refers to the actions taken when the crowd of people behaves differently from expected. In this work, we focus on using data from the urban environment as resources to study crowd situations, given the popularization of connected devices. For instance, Ericsson estimates that there will be around 3.5 billion cellular Internet of Things connections by 2023.

As initial steps to provide solutions for crowd management and control in the context of smart cities, two fundamental building blocks are urban sensing and urban data analytics [3]. Urban sensing allows us to obtain data from the cyber-physical world and urban data analytics understand the city dynamics.

Traditional sensing methods use surveillance camera, road sensors, and wireless sensor networks. Usually, these data sources cover only particular regions. Moreover, the deployment and maintenance of the infrastructure in a citywide scale is very expensive [4]. For these reasons, the exploration of humans as sensors emerges as an opportunity to overcome the problem of covering the whole city [5]. Given the increasing availability of urban data extracted by different infrastructures of crowdsensing and participatory sensing [6], we have witnessed a growing interest in the investigation of properties of human behavior based on a huge volume of sensed data with the goal of improving the city operation systems.

Fig. 1 illustrates the key blocks of a data-driven process for crowd management and control. Initially, raw data from various sources are collected from the cyber-physical world and submitted to a preprocessing for the treatment of imperfections. Afterwards, they are submitted to an integration, storage and processing step aiming at greater completeness and computational optimization. Mining, analysis, and visualization are fundamental to getting insights and adding value. This step allows discovering knowledge from the

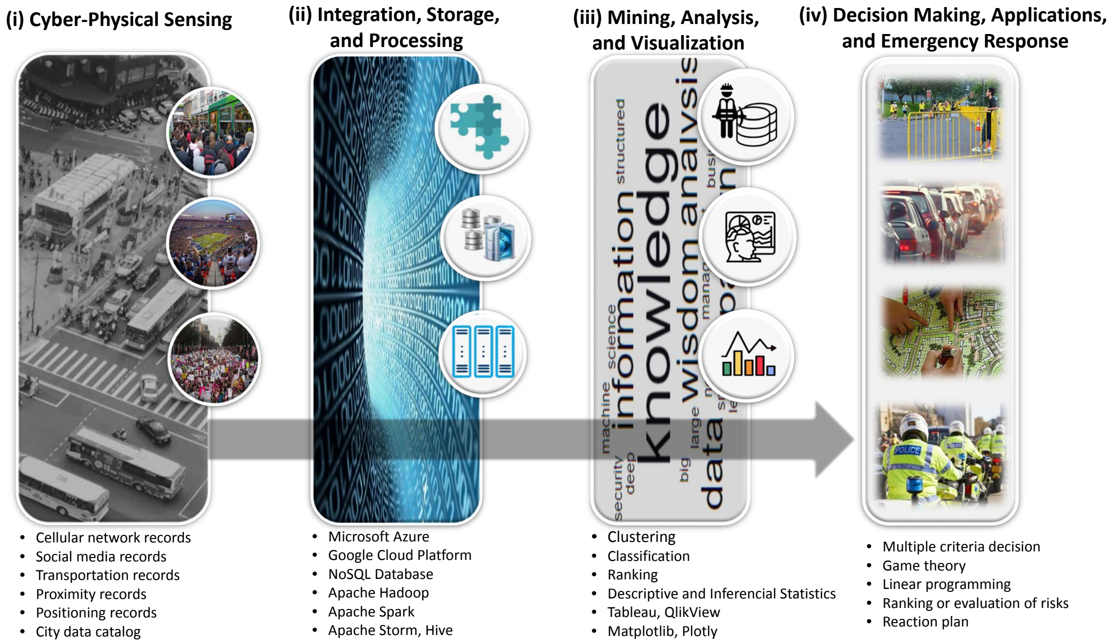

(i) Cyber-Physical Sensing
(ii) Integration, Storage, and Processing
(iii) Mining, Analysis, and Visualization
(iv) Decision Making, Applications, and Emergency Response
security deep machine science information knowledge big large wisdom data analysis structured
POLICE
• Cellular network records
• Social media records
• Transportation records
• Proximity records
• Positioning records
• City data catalog
• Microsoft Azure
• Google Cloud Platform
• NoSQL Database
• Apache Hadoop
• Apache Spark
• Apache Storm, Hive
• Clustering
• Classification
• Ranking
• Descriptive and Inferencial Statistics
• Tableau, QlikView
• Matplotlib, Plotly
• Multiple criteria decision
• Game theory
• Linear programming
• Ranking or evaluation of risks
• Reaction plan

Fig. 1: Big picture of the data-driven process for crowd management and control.

collected data in order to assist the decision making, provision of applications and emergency response that are performed in the last step. As we are interested in discussing the potentialities of urban data to detect, characterize, and analyze crowd situations, we will focus on steps (i) and (iii) and discuss how they relate to the others.

The organization of this work and our contributions are summarized in the following. Section II presents the types of crowd situations considered in this work. Section III identifies and reviews the main types of urban data discussing their strengths and weaknesses, highlighting their applicability in the study of crowd situations. Section IV presents a detailed discussion about the main challenges and opportunities related to the detection, characterization, and analysis of crowd situation and their relationships with urban data. Section V presents two case studies that show how urban data can be applied for spatial, temporal and semantic observation in the study of crowd situations. Finally, Section VI presents our final remarks.

II. Types of Crowd Situations

We define crowd as a collective situation in which individuals come together, with a shared purpose, around a specific location, during a time interval. In Sociology, Blumer [7] has classified crowds into four categories: casual, conventional, expressive, and acting. In the security area, the US Army Field Manual on Civil Disturbance Operations considers four types of crowds: casual, sighting, agitated, and mob-like.

We have classified crowds in specific situations that occur within a city based on spatial, temporal, and semantic aspects as: causal, scheduled, and protest crowds. Casual crowd means people who happen to be at the same place at the same time. In this case, the individuals have no common bond and normally exhibit short-term purpose. It happens routinely in large centers when people are waiting for public transportation or during their leisure time in the park. Scheduled crowd means people who come together for a well-defined purpose. This type of crowd represents a scheduled event where people share a common goal such as concert, parade, exhibition and game. Protest crowd can contain the characteristics found in the other types of crowds, with the addition that people are claiming something, as in a march or rally, or in a violent case as in a riot.

For all those cases, we observe some perspectives that must be considered to characterize them. The temporal perspective indicates the time window (beginning and end) of such a scenario, even for unforeseen situations. The spatial perspective consists in characterizing how the crowd geographically occupied the region of the city. Finally, the theme refers to the semantic description of that situation.

Data	**Advantages**	**Limitations**
Cellular network records	Coverage of large areas (e.g., cities, countries); Large volume of users; No additional costs or infrastructure for data collection	Spatial imprecision of user positioning; CDR entry is logged only when a user performs a telecommunication operation; Availability of data is restricted
Social media records	Coverage of large areas (e.g., cities, countries); Large volume of users; Availability of APIs and web crawlers for data collection	User decides whether to share location; Both spatial and temporal inaccuracies for data analysis
Transportation records	Vehicle traffic data can be obtained by API or Web crawlers; Public vehicles have no privacy issues	Area of coverage of the infrastructure is conditioned to transport logistics; Availability of data is restricted; Vehicle traffic data might be outdated
Proximity records	Reduced cost because it uses the infrastructure itself to collect data; It is useful for internal and external monitoring	Recruitment of volunteers to participate in the data collection; Generally restricted to controlled regions and environments; A smaller scale wrt number of users, when compared to CDR
Positioning records	Spatially accurate and regular data, tracking users more precisely; It is applied for tracking people and vehicles	Few traces available; Privacy and battery consumption make it difficult for users to create them; A smaller scale wrt users, when compared to CDR
City data catalog	Most are general data without privacy restrictions; Open data initiatives are increasingly popular	Data is generally poorly formatted; Depends on public policies to encourage open data

TABLE I: Overview of main advantages and limitations of urban big data for crowd management.

III. Overview of Urban data

Urban data refers to the set of data generated by several sources in the urban environment. In this section, we describe some of the urban data types that are adequate to study crowd. Table I gives an overview of the main advantages and limitations of big urban data for crowd management. Those data sources can be obtained from open-data initiatives, crawlers, crowdsourcing, or making a request to a particular organization.

a) ***Cellular network records:*** Call Detail Records (CDRs) are the best example of cellular network records that contain information about the location of users. CDR entries are metadata recorded when a mobile user is involved with a telecommunication operation (e.g., call, text message). Generally, each CDR record contains the identifier of the users involved in the operation, the timestamp of the operation, and the identifiers of the base stations the mobile users are registered.

b) ***Social media records:*** Location-based social media systems enable the study of the city dynamics and its social behavior in a large scale, providing urban data much more quickly than traditional sensing methods. In this type of social media, whenever users make a post, they have the possibility of sharing their geo-location, which is known as “check-in”.

c) ***Transportation records:*** It consists of all data acquired from the public and private transportation systems of a city. In this group, we have individual and collective mobility data of people and vehicles. For instance, transport monitoring camera can provide real-time and historical data of the traffic situation. In many cities, the access to buses and train stations is controlled by smart cards. This type of card allows us to identify the flow of people among regions within a city. Other data sources that provide traffic data and traffic incidents are also interesting.

d) ***Proximity records:*** These are records that capture the connectivity between mobile devices. There are two common ways of collecting this data: (i) mobile devices (e.g, smartphones) monitor encounters (records) using their communication technology (e.g., Bluetooth, RFID, Wi-Fi Direct); and (ii) users connect to Wi-Fi access points within a controlled environment (e.g., campus university). In this latter case, proximity records are created by observing the spatiotemporal overlapping of users within the range of an access point.

e) ***Positioning records:*** Currently, a large number of devices are equipped with a Global Navigation Satellite System (GNSS) receiver (e.g., GPS) to determine their location. Positioning records contain positioning and timing data received by GNSS receivers. These records usually have very absolute and highly periodical location information, making it easier to keep track of users' trajectories. However, they are subject to the occurrence of long periods/distance between two consecutive entries when tracking mobile entities [8].

f) ***City data catalog:*** Official data obtained from government institutions are important to analyze crowd situations. Demographic data, road network, public parking spots, weather information, points of interest (PoIs), historical data about traffic incidents are examples of useful information that can provide insights to crowd analysis.

IV. Key Research Challenges

Inspired by types of crowds presented in Section II, we point out the main challenges to detect, characterize, and analyze crowd in a city domain when we are using urban data. By mining these ubiquitous data, we can take a step towards to respond contextual question like where people are and go, when, why, and how.

a) ***Data preprocessing and management:*** All data types discussed above have constraints such as limited amount of monitored users, and temporal and spatial coverage. A strategy to overcome these problems and have a better representativeness is to combine different data sources, adding more value in terms of scale, spatiotemporal precision, and semantic meaning. Thus, heterogeneous data fusion is a promising technique to deal with problems that appear in crowd management using urban data, such as data quality, data imperfection, outliers, conflicting data, and data sparsity.

Data management should look for scalable and distributed solutions based on high-performance computational architectures, as we are dealing with big data volume, variety and velocity.

b) ***Incentive mechanism and publicly available data:*** A key issue is to encourage people to make their information available so that people can act as mobile sensors in the cities. One possibility is to explore crowdsourcing where a large group of people feeds proactively information about events. This task can be interesting in obtaining data for analyzing both casual and scheduled crowds. A direction is to recruit people with common interests and social relationships to enter meaningful samples, avoiding false information and manipulations [9]. Moreover, initiatives that encourage open data from different city operations are also important (e.g., data from public transportation and access to Wi-Fi networks). In all cases, incentive mechanisms can be related to monetary return, entertainment, or services to users while they provide data.

c) ***Privacy and security in crowdsensing:*** In general, the obtained data are strongly related to the human participation, and, thus, we need to preserve the people's privacy and security. On the other hand, we need to motivate their participation and make data publicly available while ensuring their privacy and security. There are some solutions based on anonymization, multiparty computation and data perturbation, but they fail in several respects depending on the data type. For example, how to guarantee privacy without losing the representativeness and quality of the data. Almost all data types need a high attention to privacy and security. The city data catalog type requires low attention because it involves general information.

d) ***Detect crowd situation:*** A key issue in crowd management is to detect a crowd situation, which depends on the data type. For instance, using check-ins from Foursquare, we can identify PoIs that people visit daily; or we can discover unexpected large-scale events when we analyze users call frequency in CDR. In both cases, we need different techniques to detect the crowd, such as statistical tools (e.g., density estimation), machine learning algorithms, and anomaly detection techniques.

e) ***Crowd density estimation:*** Crowd density estimation infers the amount of people in a crowd. This task is very important to control evacuation and avoid crowd disasters. Recent efforts have concentrated on estimating density using surveillance camera images [10]. This needs a previously installed infrastructure and additional cost. Using urban data, we can explore their pervasiveness and availability to infer the crowd density, using proper techniques.

*f) **Localization and spatial coverage**:* We can use urban data to discover the location and the spatial extent of crowd situations. For instance, users posting geolocalized messages on Twitter or making a cell phone call close to an event work as human sensors to describe the location of this event.

Some problems to identify the user's location may exist due to the data type. For example, CDR data contains a location as a function of a base station, a social media message may contain no location or users at different locations may be discussing an event. Thus, to locate individuals become a challenge. In social media, natural language processing (NLP) techniques are paramount for obtaining a relative location. Also, relevant localization and spatial coverage problems might occur when individuals attend different events at the same region, and, thus, we will need to disambiguate them.

*g) **Spatial and temporal tracking**:* In some situations, the crowd does not remain concentrated at the same location. In this case, we are interested in understanding the crowd dynamics and observing its coverage area. Most data types allow us to do spatial and temporal tracking, but each one has its challenges and limitations. For instance, in case of a meeting point for gathering people, they might begin to move in groups, similar to a flock or swarm behavior, or exhibit a spatial fragmentation.

We can also observe the mobility dynamics of a crowd in relation to time. For instance, we can establish the moments of the formation and dispersion of the crowd, and its characteristics along the time.

*h) **Semantic enrichment**:* Semantic enrichment consists in applying methods to data to gain new knowledge, in this case from a crowd situation. Using this technique, we can have a contextual understanding of what people are doing and why they are moving, their emotional behavior [11] and a description of the situation. Therefore, we are faced with the challenge of how to effectively integrate urban data to obtain meaningful knowledge of everyday situations. This task is fundamental to assist authorities in decision-making. Social media records and city data catalog contain the main information to perform semantic enrichment.

Usually, raw data is not enough to understand a scenario. In this case, we can apply semantic enrichment such as NLP and use a named-entity recognition method to social media data to extract specific information such as location, time, and agents. For a deeper and more representative knowledge, semantic Web concepts (e.g., Linked Open Data and ontology) can be used to perform disambiguation and coreference.

V. Case Studies

Below, we present two case studies that explore different types of urban data. For these studies, we integrate data for the same period, clean imperfect data, and transform data to suit the data mining techniques (e.g., clustering and detection of anomalies).

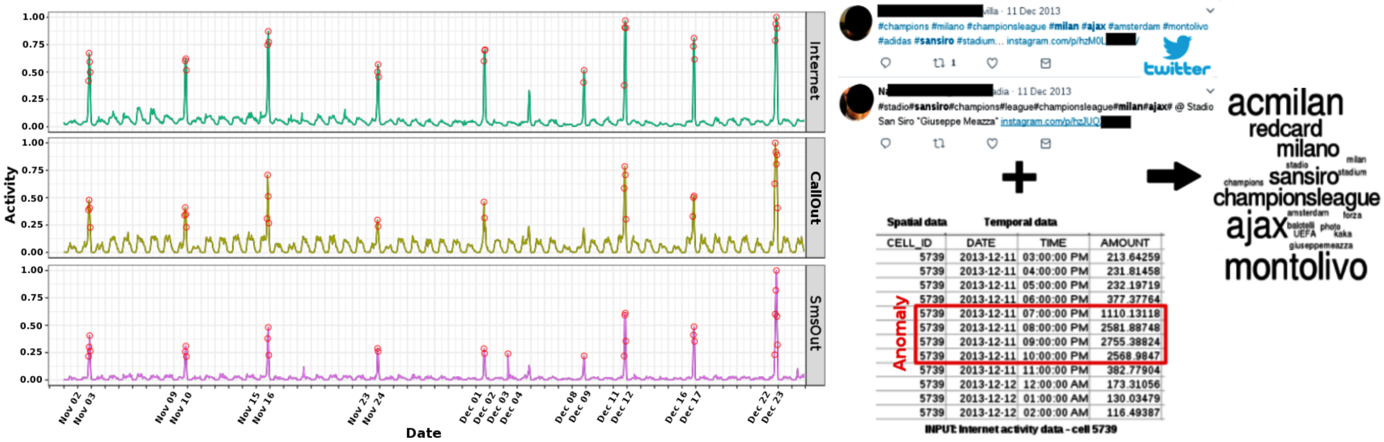
A. Case Study 1

In this study, we show the benefits of using data from telecommunication operations combined with social media and PoI datasets. The telecommunication datasets contain aggregated activity of Short Message Service (SMS), calls and Internet connections [12], for Milan, Italy. The datasets are spatially aggregated using a grid, covering an area of $552\,\mathrm{km}^2$ and each cell has an area of $235 \times 235\,\mathrm{m}^2$. For each cell, and for each 10-minute interval, the overall number of SMS, calls, and Internet connections are expressed as an activity level due to privacy issues.

Our goal is to detect the occurrence of the crowd. For each cell, we model the intensity of activities as a time series, and apply an anomaly detection [13] method, called S-H-ESD[1]. Using this strategy, we can monitor activity peaks in scheduled events or discover abnormal situations of casual events.

Fig. 2a shows the time series obtained by different types of telecommunication data over a two-month period at cell 5739. The red markings in the peaks represent the points of anomalies detected by our algorithm. In general, for each peak, it was noticed the existence of four anomalies that comprise the first hour before the event, the two hours during the event, and the hour after the event. Since we performed

[1] https://github.com/twitter/AnomalyDetection



[redacted] [redacted]villa · 11 Dec 2013
#champions #milano #championsleague **#milan #ajax** #amsterdam #montolivo #adidas **#sansiro** #stadium... instagram.com/p/hzM0L[redacted]/
1

[redacted] Na[redacted]adia · 11 Dec 2013
#stadio#**sansiro**#champions#league#championsleague#**milan**#**ajax**# @ Stadio San Siro "Giuseppe Meazza" instagram.com/p/hzJUQ[redacted]

Spatial data Temporal data

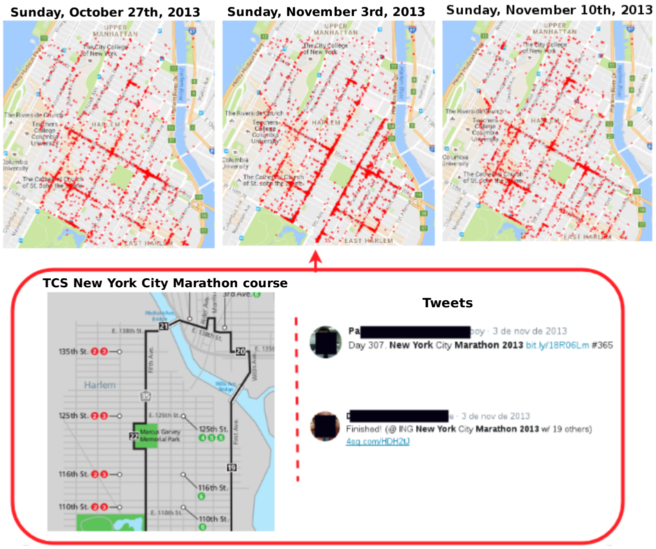
CELL_ID	DATE	TIME	AMOUNT
5739	2013-12-11	03:00:00 PM	213.64259
5739	2013-12-11	04:00:00 PM	231.81458
5739	2013-12-11	05:00:00 PM	232.19719
5739	2013-12-11	06:00:00 PM	377.37764
5739	2013-12-11	07:00:00 PM	1110.13118
5739	2013-12-11	08:00:00 PM	2581.88748
5739	2013-12-11	09:00:00 PM	2755.38824
5739	2013-12-11	10:00:00 PM	2568.9847
5739	2013-12-11	11:00:00 PM	382.77904
5739	2013-12-12	12:00:00 AM	173.31056
5739	2013-12-12	01:00:00 AM	130.03479
5739	2013-12-12	02:00:00 AM	116.49387

Anomaly

INPUT: Internet activity data - cell 5739

(a) (b)

Fig. 2: (a) It shows the anomalies found in telecommunication data; (b) It represents the enriching raw telecommunication data with semantic information.


Sunday, October 27th, 2013
Sunday, November 3rd, 2013
Sunday, November 10th, 2013
UPPER MANHATTAN
HARLEM
EAST HARLEM
TCS New York City Marathon course
135th St.
125th St.
116th St.
110th St.
Harlem
Fifth Ave.
First Ave.
Marcus Garvey Memorial Park
Tweets

[redacted] Pa[redacted]boy · 3 de nov de 2013
Day 307. **New York** City **Marathon 2013** bit.ly/18R06Lm #365

[redacted] D[redacted]e · 3 de nov de 2013
Finished! (@ ING **New York** City **Marathon 2013** w/ 19 others) 4sq.com/HDH2tJ

Fig. 3: Drop-offs in part of Manhattan in three days and contextual information.

the temporal monitoring at the cell level, we can identify the spatial and temporal dimensions of the occurrence of the event.

Besides the spatial and temporal perspectives, we can explore the semantic aspect when we use other data sources. Looking at the points of interest in Milan, we can see that cell 5739 is located at the San Siro stadium. By applying named-entity recognition method based on NLP[2] in geolocated social media data, it was possible to identify the anomalies as soccer games. Fig. 2b shows the main topics discovered from the social media for that region during the observed time. We noticed that an event was taking place (the soccer match AC Milan vs. Ajax – UEFA Champions League) at the San Siro Stadium, Milan. It is worth mentioning that other data sources can be used in this context (e.g., illustrations and photos from Instagram or Flickr) to confirm the analyzes obtained from other sources.

[2] https://nlp.stanford.edu/software/

B. Case Study 2

In this study, we will show the potentialities of using vehicle positioning data, taxi pick-up and drop-off coordinates, combined with other contextual data (e.g., social media and data from official sources). The taxi dataset contains 1.3 billion trips in New York from Jan. 2009 through June 2016 with GPS coordinates for starting and endpoints, for the NYC Taxi and Limousine Commission[3].

This type of data provides relevant information about people's mobility, besides identifying points of crowd, as shown in Fig. 3. The plot on the top shows the drop-offs in Manhattan, NY, on Oct. 27th, Nov. 3rd, and Nov. 10th, 2013. Importantly, there is a variability on Nov. 3rd when compared to other days[4]. This behavior is justified when we have obtained contextual sources. In this case, the anomaly was the New York City marathon. Therefore, this type of positioning shows the spatial coverage structure of an event in the city, being important for both security control and traffic decision-making. The use of other data sources allows us to increase the knowledge about a crowd situation, as data from social media and official information of the route, in this case. This data combination allows authorities to monitor the situation with a broader view.

VI. Final Remarks

The pervasive availability of urban data provides unprecedented opportunities for analyzing several situations in the cities. The understanding of crowd situations based on urban data brings many opportunities such as urban planning, business, intelligent transportation system and traffic optimization. For instance, traffic flow can be optimized in a region according to events and the number of people nearby; a urban transport system may be adapted depending on commuting behavior; and patterns of collective behavior of city dynamics and urban land use may be discovered. All these examples have several challenges that stem from what we presented and discussed in this work and can be further explored when considering the various types of crowd scenarios.

Acknowledgment

This work was partially supported by the NSERC CREATE TRANSIT, NSERC DIVA Strategic Research Network, Canada Research Chairs Program, CNPq, FAPEMIG, CAPES, and grants 15/24536-2 & #15/24494-8, São Paulo Research Foundation (FAPESP). In addition, we would like to thank Microsoft Research for the Microsoft Azure Research Award.

References

[1] C. Martella *et al.*, "On current crowd management practices and the need for increased situation awareness, prediction, and intervention," *Safety science*, vol. 91, pp. 381–393, 2017.

[2] A. E. Berlonghi, "Understanding and planning for different spectator crowds," *Safety Science*, vol. 18, no. 4, pp. 239–247, 1995.

[3] Y. Zheng *et al.*, "Urban computing: Concepts, methodologies, and applications," *ACM TIST*, vol. 5, no. 3, pp. 38:1–38:55, 2014.

[4] B. Guo *et al.*, "Mobile crowd sensing and computing: when participatory sensing meets participatory social media," *IEEE Commun. Mag.*, vol. 54, no. 2, pp. 131–137, 2016.

[5] T. H. Silva *et al.*, "Users in the urban sensing process: Challenges and research opportunities," *Pervasive Computing: Next Generation Platforms for Intelligent Data Collection*, pp. 45–95, 2016.

[6] A. Draghici and M. V. Steen, "A survey of techniques for automatically sensing the behavior of a crowd," *ACM CSUR*, vol. 51, no. 1, p. 21, 2018.

[7] H. Blumer, "Collective behavior," *New outline of the principles of sociology*, p. 166, 1951.

[3] http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml

[4] We only show 3 days, but the same behavior is observed in other days.

[8] C. Celes *et al.*, "Improving vanet simulation with calibrated vehicular mobility traces," *IEEE Trans. on Mobile Computing*, vol. 16, no. 12, pp. 3376–3389, 2017.

[9] I. B. Amor *et al.*, "Discovering best teams for data leak-aware crowdsourcing in social networks," *ACM TWEB*, vol. 10, p. 2, 2016.

[10] J. M. Grant and P. J. Flynn, "Crowd scene understanding from video: a survey," *ACM TOMM*, vol. 13, no. 2, p. 19, 2017.

[11] A. Giachanou and F. Crestani, "Like it or not: A survey of twitter sentiment analysis methods," *ACM CSUR*, vol. 49, no. 2, p. 28, 2016.

[12] G. Barlacchi *et al.*, "A multi-source dataset of urban life in the city of milan and the province of trentino," *Scientific data*, vol. 2, p. 150055, 2015.

[13] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM CSUR*, vol. 41, no. 3, p. 15, 2009.

BIOGRAPHIES

Clayson Celes is currently working toward the Joint/Dual PhD degree in Computer Science at the University of Ottawa, Canada and the Federal University of Minas Gerais (UFMG), Brazil. He received the bachelors degree in computer science from the State University of Ceara (UECE), Brazil, in 2010 and the MSc degree in computer science from UFMG, in 2013. His research areas are vehicular networks, data analysis, and mobile computing.

Azzedine Boukerche (FIEEE, FEiC, FCAE, FAAAS) is a Distinguished University Professor and holds a Canada Research Chair Tier-1 position at the University of Ottawa (uOttawa). He is founding director of the PARADISE Laboratory and the DIVA Strategic Research Centre at the uOttawa. He has received the C. Gotlieb Computer Medal Award, Ontario Distinguished Researcher Award, Premier of Ontario Research Excellence Award, G. S. Glinski Award for Excellence in Research, IEEE Computer Society Golden Core Award, IEEE CS-Meritorious Award, IEEE TCPP Leaderships Award, IEEE ComSoc ASHN Leaderships and Contribution Award, and University of Ottawa Award for Excellence in Research. He serves as an Associate Editor for several IEEE transactions and ACM journals, and is also a Steering Committee Chair for several IEEE and ACM international conferences. His research interests include wireless ad hoc and sensor networks, mobile computing, performance evaluation and modeling of large-scale distributed and mobile systems.

Antonio A. F. Loureiro is a full professor at UFMG, where he leads the research group on mobile ad hoc networks. He received his B.Sc. and M.Sc. degrees in computer science from UFMG, and his Ph.D. degree in computer science from the University of British Columbia, Canada. He was the recipient of the 2015 IEEE Ad Hoc and Sensor (AHSN) Technical Achievement Award and the Computer Networks and Distributed Systems Interest Group Technical Achievement Award of the Brazilian Computer Society. He is a regular visiting professor and researcher at the PARADISE Research Laboratory at the University of Ottawa and is an international research partner of DIVA Strategic Research Networks. His main research areas include wireless sensor networks, mobile computing, and distributed algorithms. In the last 15 years, he has published regularly in international conferences and journals related to those areas, and has also presented tutorials at international conferences.