Discovering Mobility Patterns of Instagram Users through Process Mining Techniques

Claudia Diamantini*, Laura Genga†, Fabrizio Marozzo‡, Domenico Potena* and Paolo Trunfio‡
*Dipartimento di Ingegneria dell'Informazione, Universita Politecnica delle Marche, Italy
Email: {c.diamantni,d.potena}@univpm.it
†Eindhoven University of Technology, Eindhoven, The Netherlands
Email: l.genga@tue.nl
‡DIMES, University of Calabria, Italy
Email: {fmarozzo,trunfio}@dimes.unical.it

***Abstract*—Every day a huge amount of data is generated by users of social media platforms, like Facebook, Twitter and so on. Analyzing data posted by people interested in a given topic or event allows inferring patterns and trends about people behaviors on a very large scale. These posts are often geotagged, this way enabling mobility pattern analysis. In this work, we investigate the use of *Process Mining* techniques to support the discovery and the analysis of mobility patterns of social media users. We discuss the results obtained analyzing posts of Instagram users who visited EXPO 2015, the Universal Exposition hosted in Milan, Italy, from May to October 2015.**

I. Introduction

Social media analysis is an emerging research area aimed at extracting useful knowledge from the huge volume of data generated by users of online networks such as Facebook, Twitter and Instagram. Examples of applications are the analysis of collective sentiments, understanding the behavior of groups of people, or discovering the dynamics of public opinion [1]. In many cases, social media posts are geotagged, thus allowing the extraction of trajectories of nomadic social users employing trajectory mining techniques [2]. In this paper, we investigate the use of *Process Mining* (PM) techniques [3] to discover mobility patterns of social media users attending large-scale events. The event under analysis is EXPO 2015, the Universal Exposition held under the theme "Feeding the Planet, Energy for Life," that was hosted in Milan, Italy, from May 1st to October 31st, 2015. The data source is composed of the geotagged posts published by the Instagram users who visited EXPO. Instagram is a fast-growing service for social networking, mobile photo-sharing, which allows users to share a photo, which can be accompanied by text, hashtags, mentions to other users and a location field. Data was collected as part of a previous work in which classical trajectory mining techniques were employed [4]. It contains geotagged posts published by about 238,000 Instagram users who visited EXPO, resulting in more than 570,000 posts published during the visits.

This work has been partially funded by the ITEA2 project M2MGrid (No. 13011)

PM is a relatively young research area whose goal consists in analyzing *event logs* recording data related to past process executions to understand how a process is actually performed within an organization. With respect to other business intelligence techniques, mainly aimed at evaluating processes by means of synthetic indicators, PM is focused on inferring the *structure* of end-to-end process executions. PM techniques have been initially developed to analyze business processes, which usually are well-defined, structured processes. Nevertheless, during the last years several research efforts have been performed to apply PM techniques in domains characterized by complex, and often unstructured, processes. As an example, we can mention applications related to exploring learning data from Coursera [5], inferring clinical pathways in Health Care processes [6], monitoring devices data to fault diagnosis [7] and so on.

In the present work, we explore the use of PM techniques to gather relevant insights about visitor's behaviors. We consider the visit of each user as a single execution of a general "Expo visiting" process, whose activities are represented by the photos posted on Instagram. By doing so, we are able to obtain a unified perspective on visitor's behaviors, represented in the form of a process model describing their trajectories, starting from which interesting knowledge regarding most relevant mobility patterns can be inferred. To the best of our knowledge, this process-oriented perspective of mobility analysis is novel in the literature and this is the first attempt to apply PM techniques to this task.

The plan of the paper is as follows. Section II provides some definitions and the objectives of the study. Section III describes the methodology employed for data collection and analysis. Section IV presents the main results of the analysis. Section V discusses related work. Section VI concludes the paper.

II. Preliminary definitions

We provide here some definitions of the main concepts that will be used in the remainder of the paper.

A *process* can be generally defined as a flow of *activities*, i.e. atomic pieces of work, performed in a certain order to achieve a given goal. Each execution of a process is named process *instance*. In turn, the execution of an activity is called an *event*. Process instances are typically recorded in event logs in the form of *traces*, storing events in their stricit order of occurence. Relevant information typically stored in the event log includes the timestamp of the event, the name of the executed activity, the resource who executed it and so on. The minimum requirements that have to be fulfilled by an event log in order to apply PM techniques are i) every event has to be related to its corresponding activity and process instance, and ii) events in a process instance have to be ordered.

In this work, we consider visitor's trajectories as instances of the process 'visiting Expo'. Therefore, in our process each activity corresponds to visit an EXPO pavilion. Data related to each visit are stored in Instagram posts published by the visitor, which collects the name of the pavilion, the time of the visit and so on. The sequence of posts published by the same visitor represents a trace, where events correspond to the posts and are totally ordered on the basis of the timestamps. In the following we formalize notions related to our event log. Given:

- IP: Set of georeferenced Instagram posts published inside EXPO in the period May 1st - October 31st, 2015.
- VIS: Set of EXPO visitors who published at least one of the Instagram posts in IP.
- PAV: Set of EXPO pavilions, including both country pavilions and organization/company ones.
- PN: Set containing the position name of every country pavilion ("USA Pavilion", "China Pavilion", etc.) and of every organization/company pavilion ("UN Pavilion", etc.) in PAV.
- gn: Generic position name ("EXPO 2015") for all the places that are inside EXPO but are not associated with any pavilion in PAV.

we define event, trace and event log as follows.

Definition 1 (Event): An *event* is defined as an n-uple $pt_i = (v_i, pn_i, ts)$ where $pt_i \in IP$, $v_i \in VIS$, $pn_i \in PN$ and *ts* is the timestamp of the post.

Definition 2 (Trace, Event log): A *trace* is a sequence of posts published by the Instagram users who visited EXPO, i.e. $l = \langle pt_1, pt_2, ..., pt_m \rangle$ such that $v_1 = v_2 = ... = v_m$. An *event log* L is a multiset of traces.

Note that an event log is defined as a multiset of traces because it can involve two or more identical traces, corresponding to visitors who followed exactly the same paths.

III. Methodology

The whole data collection and analysis process is composed of five steps: A) identification of the Instagram locations inside EXPO; B) collection of Instagram posts inside EXPO and identification of visitors; C) creation of the input dataset; D) Process mining. A description of the steps is given in the remainder of the section.

A. Identification of the Instagram locations

The goal of this step is to identify the set EL of Instagram locations inside the EXPO area. The Instagram API provides a service that, given a geographical point c and a distance r, returns the set of Instagram locations falling in the circle centered at c of radius r. The EXPO area was partitioned in a grid, where C = {c_1, c_2, ..., c_n} are the grid intersection points laying within the EXPO area. For each c_i in C, the Instagram API were asked to return the set L_i of locations within a given distance r from c_i.

The set EL of Instagram locations inside EXPO is the union of all L_i. EL was then cleaned to filter out the locations with coordinates outside the EXPO area, which were returned starting from grid points at the borders of the area. At the end of data collection and cleaning, EL contained 2,890 locations. Every l_i in EL is a tuple <*id*, *name*, *latitude*, *longitude*>, where *latitude* and *longitude* are the geographical coordinates of l_i.

B. Collection of Instagram posts inside EXPO and identification of visitors

The goal of this step is to collect the set IP of Instagram posts published inside the EXPO area in the period May 1st - October 31st, 2015, and to identify the set VIS of Instagram users who visited EXPO. To this end, given the set EL of Instagram locations inside EXPO identified in the previous step, we proceeded as follows. For each l_i in EL, we submitted to the Instagram API a query that, given l_i and the period specified above, returns the set PT_i of Instagram posts published from l_i during the period. By making the union of all the PT_i, we obtain the set IP of all the Instagram posts published inside EXPO. Then, we extracted the set VIS of the distinct users who published at least one post in IP. The number of posts in IP is equal to 570,973, while the number of users in VIS is equal to 238,196, with an average of 2.4 posts per user.

C. Creation of the input dataset

The goal of this third step is the creation of the dataset T used as input for the analysis. We recall that the i-th tuple T_i of T is a pair <v_i, {P_{i1}, ..., P_{ik}}>, where P_{ij} contains *position name* and *timestamp* of the j-th post by user v_i.

To assign a position name to the posts in IP, two dictionaries have been used: First Level Dictionary (FLD) and Second Level Dictionary (SLD). FLD receives the Instagram location of a post and returns the corresponding position name in PN, if a mapping is found. To this end, FLD contains a set of tuples <$term_a$, $term_b$>, where $term_a$ is an Instagram location in EL, and $term_b$ is a position name in PN. Using FLD, we were able to assign a

position name to 160,307 posts, which represent the 28.08% of the posts in IP. The remaining posts were passed to SLD, which receives the text field of a post and returns the corresponding position name in PN, if a mapping is found. SLD contains a set of tuples $<term_a, term_b>$, where $term_a$ is a word or phrase that may be found in the text of a post, and $term_b$ is a position name in PN. Using SLD, we were able to assign a position name to 195,699 posts, which represent an additional 34.27% of the posts in IP.

The last step of data preparation aims at obtaining an event log fitting the minimum requirements needed to apply PM techniques. More precisely, the event log has to be built in such a way that $a)$ for each event both its corresponding activity and the process instance it belongs to are known, and $b)$ the order of the events in a trace reflects the order with which corresponding activities have been executed in the corresponding process instance. In our case, we decided to consider the visit of a single person as a process instance; hence, we associated to each person a different id, which became the trace id. Then, we considered the act of posting a photo on Instagram as a process activity; hence, each Instagram post represents an event. Note that Instagram posts have a timestamp that allows ordering them in each process instance. Other information is available as well, like the coordinates where the photo was taken, which will be exploited in deriving process models. The output of the data preprocessing step consists in a well-formed event log, which becomes the input for the last step of the methodology, i.e. the process mining step.

D. Process mining

The last step of the methodology aims at inferring a process model describing visitors' path among EXPO pavilions. It is worth noting that the process we intend to analyze is characterized by a high degree of heterogeneity. Indeed, EXPO attracted many different visitors, who were totally free in choosing the order they preferred to visit pavilions. As a result, the event log we get from the previous step involves a plethora of heterogeneous behaviors, describing multiple, different possible visiting paths. From PM literature, it is well known that directly mining a process model from this kind of event logs result in so-called *spaghetti-like* models, i.e. chaotic models, which provide barely or no support to the process analyst in exploring process behavior.

To address this issue, in this work we combine a set of *abstraction*, *clustering* and *pattern discovery* techniques, proposed in several PM approaches to deal with the complexity of spaghetti-like models. These techniques work on different elements of the event log and can be applied before, during or after the model discovery phase. The assumption underlying abstraction techniques is that the complexity of the process derives from considering too fine-grained process activities, with the result that the inferred model provides a detailed and low-level perspective on the process. This complexity can hence be reduced by grouping activities which can be considered homogeneous under some criteria (e.g., they belong to the same subprocess, they occur very close, they are performed by the same actor). In this way, a simplified, high-level version of the process is obtained which, although losing in accuracy, is able to highlight the main process behaviors. In this work, we apply abstraction both during the preprocessing and during the process discovery phase. Since we are interested in exploring visitors' trajectories, we employ spatial-related criteria to group process activities; namely, we group together pavilions that are close with respect to their spatial coordinates.

Figure 1 shows how the EXPO locations (e.g., pavilions) have been grouped into 11 clusters using the K-Means algorithm [8]. The name of each cluster reflects the grid positions of the locations it contains. For example, cluster named *GH06-09* contains all the locations from G06 to G09 and from H06 to H09, where prefix G is used for the positions facing the North side of Decumano street (i.e., the west-east avenue), while prefix H is used for the South side.

We adopt *filtering* process discovery techniques to derive the simplified process model. These techniques abstract from less relevant process behaviors, discarding them directly during the mining of the model. More precisely, we derive the models described in this work by exploiting the *infrequent Inductive Miner* [9] (iIM) and the *Fuzzy miner* (FM) algorithms [10]. The iIM algorithm adopts an iterative procedure. At the beginning, it generates a directed graph representing the direct follows relations between activities of the event log. Then, it defines a "cut" of the graph, i.e., a partition between the nodes in disjoint sets such that all the activities in a subset have the same relation with activities in the other subsets. Relations like sequences, loops, parallelism are taken into account. The log is then split to reflect the partitioning of the activities and the procedure is then repeated on each sublog, until sublogs contain only one activity. iIM introduces a set of heuristics aimed at filtering infrequent behaviors.

In contrast with iIM, the FM algorithm does not return a formal process model showing activities control-flow. Instead, it returns a schema involving just the most relevant activities (i.e. schema nodes), displayed as single activities or aggregated in clusters, and their sequences (i.e. schema edges). To this end, the algorithm exploits two parameters: significance and correlation. The former represents both the importance of each activity and of the activities sequences, the latter how closely related are subsequent activities. The less relevant activities are simply deleted from the final output or, if they belong to a set of highly correlated activities, they are clustered in a single node. Similarly, less relevant sequences are not displayed in the outcome. Several possible metrics can be used for significance and correlation; e.g. the frequency for the significance and the proximity for

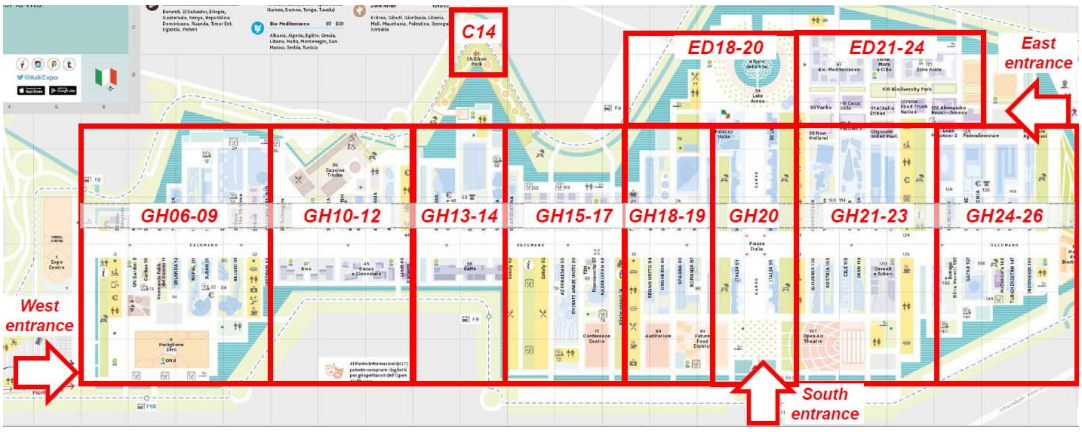

Figure 1. High-level clustering of pavilions.

correlation, that is how temporally close two events are.

In addition to the application of abstraction techniques, we also exploit trace clustering. These techniques assume that the complexity is due to the presence of heterogeneous traces in the event log describing different process *variants*. Hence, they aim at clustering similar traces in order to generate simpler process models, describing the same (or similar) variants. In this case, we exploit again spatial criteria to cluster the traces. In particular, we clustered together those traces where the first visited pavilion belongs to one of the three entrances of EXPO. In Figure 1 we have the *West Entrance* in *GH06-09*, the *South Entrance* close to *GH20*, *GH18-19* and *GH21-23*, and the *East Entrance* between *ED21-24* and *GH24-26*. Indeed, it is reasonable to expect that visitors who started the visit from the same entrance show more homogeneous paths than the others.

Finally, we exploit pattern discovery techniques on the event logs. Instead of inferring a complete model, these techniques highlight the most relevant paths of the process, providing useful insights during process analysis. More precisely, we exploit the technique proposed in [11] whose algorithm is tailored toward the mining of portion of process behaviors. Authors define a set of pattern templates representing "repeats" constructs, in particular loops, represented by multiple consecutive occurrences of an activity/sequence of activities, and subprocesses, represented by sequences of activities repeated among the traces. Patterns are detected by exploring the event log to identify those sequences that fit the patterns templates; then, they are filtered on the basis of their support. It is worth noting that this approach also returns so-called *alphabets*, where the alphabet of a pattern consists in the distinct set of the labels of its activities, with the aim of aiding the analyst in detecting possible parallelisms.

IV. Results

In this section we report the results obtained analyzing Instagram data related to EXPO visits. As process discovery technique, we use the infrequent Inductive Miner (iIM) with a threshold of 0.2. Figure 2 shows the model inferred from the event log abstracted at highest level. As we can note, this model, differently from spaghetti-like models, is very easy to understand. Arcs of the model are labeled by the number of traces (i.e., visitors) following that arc. In the Figure, only parallel gateways (i.e., split/merge AND) are explicitly represented, while arcs that split into multiple arcs represent an exclusive gateway (i.e., split/merge XOR). Paths that are in parallel (i.e., occurring in different branches generated by a parallel gateway) describe behaviors in which pavilions have been visited in the same trace, but the Process Mining algorithm has not been able to determine an order among them; e.g., clusters *GH20*, *C14* and *GH10-12*.

From the model, we recognize three macro behaviors, corresponding to branches enclosed by three pairs of parallels gateways, denoted in Figure 2 by A, B and C respectively. The upper branch of the model says that cluster *ED21-24* is executed in parallel to activities belonging to B and C blocks. This means that people who visited *ED21-24* have also visited some pavilions within B and C. However, *ED21-24* has been visited by very few visitors (4565), corresponding to the 8.14% of traces. This can be explained by analyzing the original activities belonging to this cluster; indeed, they involve non-national pavilions, e.g. the "Biodiversity Park", and pavilions of companies like "Coca Cola" and "Alitalia".

Pavilions of the B block, which correspond to clusters *GH20*, *C14* and *GH10-12*, are always visited before pavilions belonging to the C block. However, they have been visited by a small percentage of visitors: 10.02% for *GH20*, 6.71% for *GH10-12* and 0.13% for *C14*. From the map, it turns out that *C14* corresponds only to the "Children Park", thus suggesting that this zone of EXPO was not very frequented by visitors (or, anyway, most of them didn't post photos from this zone). Cluster *GH20* contains the "Italian" pavilion, which is formed by four buildings and is crossed by the Cardo (i.e., the north-south avenue). Although

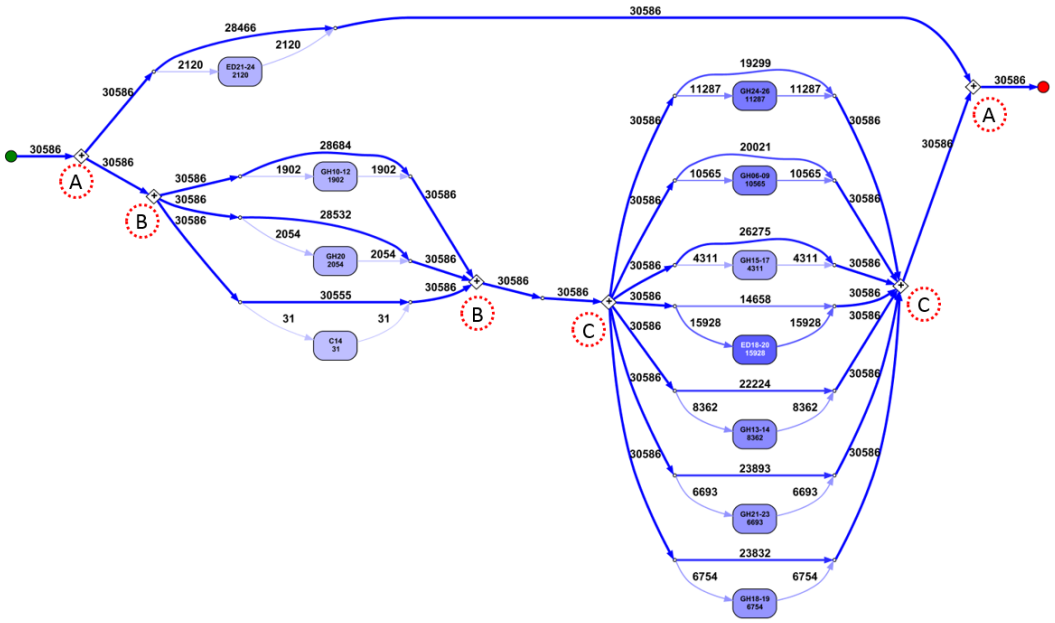

Figure 2. Process model obtained with high-level clustering.

GH20 leads to the highly visited "Tree of Life", it has been infrequently photographed. In the second parallelisms (C block), we have the group of most visited pavilions, among which the most frequent one turns out to be, as expected, the cluster *ED18-20* involving the "Tree of Life", which has been visited by 51.14% of visitors. Note that cluster *ED18-20* has also the "European Union" pavilion and the "Lake Arena", but in total they have been visited by less than 0.7% of visitors of the cluster.

Besides the overall model, we also analyzed some interesting process variants. In particular, we grouped together traces in which the first visited pavilion corresponds to one of the main entrances. These traces represent the 59.46% of visits. Figure 3, 4 and 5 shows the models obtained by using Fuzzy Miner algorithm for South, West and East entrances respectively.

The models represent transitions between clusters of pavilions, without explicitly describe parallelisms. The circular node with the triangle is the starting point of the process, namely people enter the EXPO through the specific gate. The end point is represented by the circular node with a square inside, namely people leave the EXPO.

We like to note that most people have visited only 2 clusters in a trace. Whatever the chosen entrance, it is likely that people have gone directly to a cluster close to another gate or to the "Tree of Life" (*ED18-20*). The former behavior occurs in 20.09% of visits, and the latter in around 17% of cases. This shows that visitors spent a long time in pavilions of clusters close to gates or at the "Tree of Life". Table I shows most frequent patterns for each of the three entrances. Noteworthy, around the 6% of people have visited *GH13-14* before leaving the EXPO. The most visited pavilions in this cluster are the ones of China, Thailand and Colombia.

Table I
MOST FREQUENT PATTERNS.

Gate	Pattern	Occurrence
South Entrance	*GH18-19→ED18-20*	3.04%
	GH21-23→ED18-20	2.93%
	GH21-23→GH24-26	2.62%
	GH18-19→GH06-09	2.14%
	GH20→ED18-20	1.57%
	GH18-19→GH13-14	1.48%
	GH18-19→GH15-17	1.23%
West Entrance	*GH06-09→ED18-20*	5.05%
	GH06-09→GH18-19	3.06%
	GH06-09→GH24-26	2.97%
	GH06-09→GH13-14	2.96%
	GH06-09→GH21-23	2.33%
	GH06-09→GH15-17	1.55%
East Entrance	*GH24-26→ED18-20*	3.53%
	GH24-26→GH21-23	2.52%
	GH24-26→GH06-09	2.26%
	GH24-26→GH18-19	2.19%
	GH24-26→GH13-14	1.53%

V. RELATED WORK

In the last years, several researches have investigated the development of PM techniques able to cope with unstructured, spaghetti-like processes. We can group these techniques in three main groups, i.e. i) *abstraction*, ii) *clustering* and iii) *pattern discovery* techniques. The first group of techniques is aimed to provide a higher level view on the process, able to highlight most important process behaviors. A simple strategy to achieve this simplification consists filtering less relevant elements from the process model, as done for instance by the infrequent Inductive

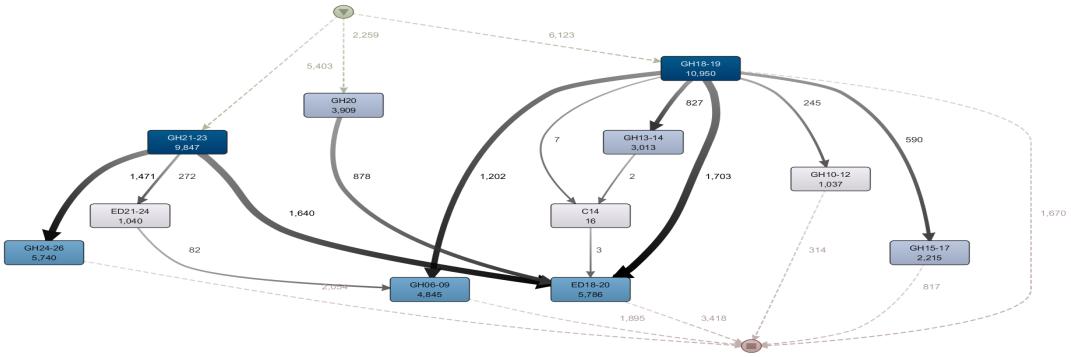

Figure 3. Model of traces in which first visited pavilion corresponds to South entrance.

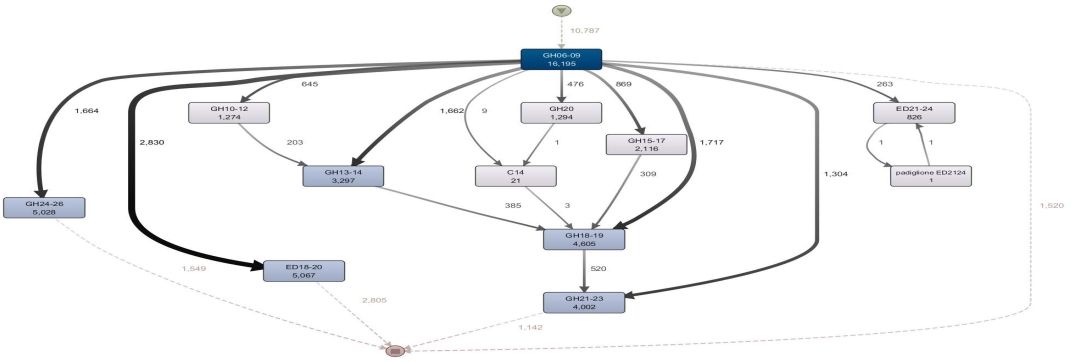

Figure 4. Model of traces in which first visited pavilion corresponds to West entrance.

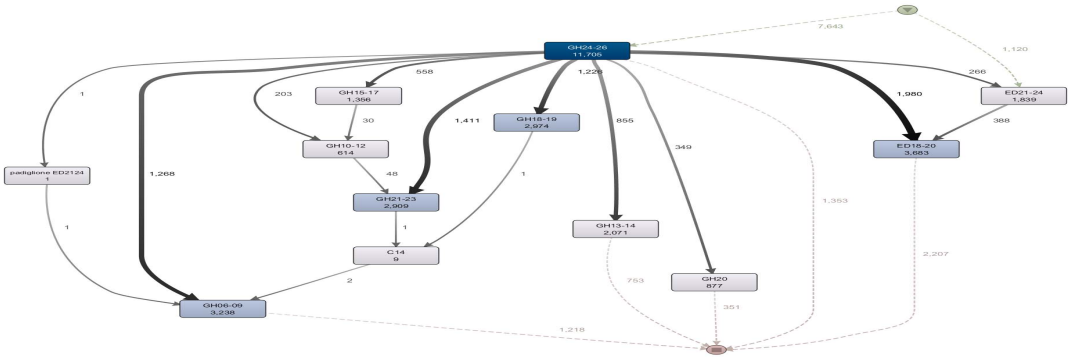

Figure 5. Model of traces in which first visited pavilion corresponds to East Entrance.

Miner exploited in this work and by other process discovery algorithms, like the Heuristic Miner [12]. Another, more sophisticated approach, consists in supporting the analyst in moving from a fine grained to a coarse-grained representation of the process, determining which process elements can be *abstracted*, i.e. filtered or grouped together. The abstraction process can be applied: on the process model, e.g. by grouping low-level events in higher-level activities [13], [14]; during the discovery of the process model, like done by the Fuzzy miner previously introduced; or on a process model previously obtained, like done for instance in [15], [16].

The second category of approaches is based on a divide-and-conquer strategy; namely, the goal consists in clustering log traces, to obtain a set of sublogs involving more homogeneous behaviors. Models mined from each sublog are usually much simpler than the model mined from the entire event log, since they take into account only behaviors which are similar under given criteria. A plethora of approaches have been proposed during last years to address trace clustering issue (see e.g. [17] for a survey on trace clustering techniques).

Finally, approaches from the third category exploit pattern

discovery techniques to mine the most relevant activities execution patterns (i.e., subprocesses), rather than inferring a complete, end-to-end process model. This approach allows the analyst to focus only on the most relevant portions of the process, i.e. its most relevant subprocesses. These approaches can be further refined in two groups, according to the kind of patterns they return. The first group involves approaches which return totally ordered subprocesses, i.e. subprocesses which correspond to portion of traces and whose events are totally ordered on the basis of their order of occurrence in the event log. A well-known example is [11] used in this work. The second group involves approaches aimed at inferring partially ordered subprocesses; among them, we can mention the *episode miner* presented in [18], which models groups of events frequently occurring together with their eventually following relations, stating which event occurred before/after than another, and the approach introduced in [19], which builds for each log the corresponding instance graph (i.e., a directed graph showing the activities control flow for the corresponding process instance) and then apply subgraph mining techniques on the set of instance graphs to extract subprocesses modeling causal and concurrent ordering relations among events.

About mobility analysis from social network, several approaches have been proposed in literature [2]. A trajectory pattern mining algorithm to discover mobility patterns, modeled as sequences of visited dense regions with travel time, is proposed in [20] and [21]. The authors of these papers used sequential pattern mining techniques to analyze moving objects over real data and synthetic benchmarks. In [22], a clustering algorithm is proposed to discover local urban area dynamics. By analyzing social media data of residents in urban areas, the algorithm discovers the local assets of a city, such as municipal borders, demographics, development and geographic resources. Reference [23] describes the analysis of geotagged tweets to discover the most frequent movements of fans attending the 2014 FIFA World Cup. By adopting a trajectory pattern mining methodology, several results in terms of number of matches attended by groups of fans, clusters of most attended matches and most frequented stadiums, are reported. Trajectory mining is also performed in [4] on the same data used in this work, but employing algorithms and techniques for associative analysis (associative pattern mining) and sequential analysis (sequential pattern mining). In [24] is studied how large events, e.g. Olympic Games, could influence the flow movement of urban population. Analyzing data from a large location-based social service, the authors created a supervised learning algorithm, exploiting a combination of both geographic and mobility features, for predicting businesses thrive of local retailers during large events. Another challenging task faced in literature is the analysis of geotagged photos published by social media systems, to discover Regions-of-Interest (RoIs) and frequent trajectory patterns for travel recommendation. In [25], tourist photos as held by Flickr are exploited to estimate the probability that a tourist will be visiting a landmark. The system proposed in [26] exploits a trip model based on canonical mobility sequences among touristic place clusters. Reference [27] describes a system that interactively helps users in travel routes planning, by exploiting the popular destinations to visit, the visiting order of destinations, the visiting time, etc. In [28], the authors exploit a Markov chain model to discover tourist routes among different RoIs and propose also an algorithm for topological analysis of personalized travel routes for different tourists. With respect to the discussed literature, we share some similarities with work performing trajectory mining, since our aim is the discovery of a descriptive model for mobility. However, looking at visits as instances of the same mobility process completely changes the perspective, and allows to highlight the presence of cycles or the differences in moving behaviors as parallelisms. To the best of our knowledge, this work is the first attempt to apply process mining techniques in mobility analysis from social network.

VI. Conclusions

In this work, we investigated the use of Process Mining (PM) techniques to discover and analyze mobility patterns arising from geotagged social media data. The proposed methodology involves two main phases. First, a set of posts related to a given event within a predefined temporal interval are collected and transformed in an event log, where each trace represents the sequence of locations visited by a single user. Then, a mining phase is performed, to infer and explore mobility patterns. In order to deal with the high complexity and the low degree of structure of the process, we propose to combine filtering process discovery, location abstraction, clustering and pattern discovery techniques, in order to highlight only the most relevant behaviors. We tested the approach on a collection of Instagram posts published by visitors of EXPO 2015. Results shows the capability of the proposed methodology to support the analysis of mobility patterns and to derive interesting insights, both by means of the analysis of high-level models describing the overall behaviors of the user and the extraction of more specific patterns focused on a certain area of the exposition. As future work, we plan to investigate the introduction of "zooming" mechanisms, allowing the analyst to switch between different levels of granularity when analyzing the model.Another research direction is the coupling of Instagram geotagging with image processing techniques to compare the actual content/position of a photo and enrich the dataset with this kind of information to improve trajectory description.

References

[1] D. Talia, P. Trunfio, and F. Marozzo, *Data Analysis in the Cloud*. Elsevier, October 2015.

[2] Y. Zheng, "Trajectory data mining: An overview," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 6, no. 3, p. 29, 2015.

[3] W. van der Aalst, *Process Mining: Discovery, Conformance and Enhancement of Business Processes*. Springer, 2011.

[4] E. Cesario, A. R. Iannazzo, F. Marozzo, F. Morello, G. Riotta, A. Spada, D. Talia, and P. Trunfio, "Analyzing social media data to discover mobility patterns at expo 2015: Methodology and results," in *The 2016 International Conference on High Performance Computing and Simulation (HPCS 2016)*, Innsbruck, Austria, 18-22 July 2016, pp. 230–237.

[5] P. Mukala, J. C. Buijs, M. Leemans, and W. M. van der Aalst, "Learning analytics on coursera event data: A process mining approach." in *SIMPDA*, 2015, pp. 18–32.

[6] Z. Huang, X. Lu, H. Duan, and W. Fan, "Summarizing clinical pathways from event logs," *Journal of biomedical informatics*, vol. 46, no. 1, pp. 111–127, 2013.

[7] C. Günther, A. Rozinat, W. van der Aalst, and K. van Uden, "Monitoring deployed application usage with process mining," *BPM Center Report BPM-08-11*, pp. 1–8, 2008.

[8] S. Lloyd, "Least squares quantization in pcm," *IEEE Trans. Inf. Theor.*, vol. 28, no. 2, pp. 129–137, Sep. 2006. [Online]. Available: http://dx.doi.org/10.1109/TIT.1982.1056489

[9] S. J. Leemans, D. Fahland, and W. M. van der Aalst, "Discovering block-structured process models from event logs containing infrequent behaviour," in *International Conference on Business Process Management*. Springer, 2013, pp. 66–78.

[10] C. W. Günther and W. M. Van Der Aalst, "Fuzzy mining–adaptive process simplification based on multi-perspective metrics," in *International Conference on Business Process Management*. Springer, 2007, pp. 328–343.

[11] R. J. C. Bose and W. M. van der Aalst, "Abstractions in process mining: A taxonomy of patterns," in *International Conference on Business Process Management*. Springer, 2009, pp. 159–175.

[12] A. Weijters, W. M. van Der Aalst, and A. A. De Medeiros, "Process mining with the heuristics miner-algorithm," *Technische Universiteit Eindhoven, Tech. Rep. WP*, vol. 166, pp. 1–34, 2006.

[13] C. W. Günther and W. M. van der Aalst, *Mining activity clusters from low-level event logs*. Beta, Research School for Operations Management and Logistics, 2006.

[14] C. W. Günther, A. Rozinat, and W. M. Van Der Aalst, "Activity mining by global trace segmentation," in *International Conference on Business Process Management*. Springer, 2009, pp. 128–139.

[15] D. Fahland and W. M. Van Der Aalst, "Simplifying discovered process models in a controlled manner," *Information Systems*, vol. 38, no. 4, pp. 585–605, 2013.

[16] A. Polyvyanyy, S. Smirnov, and M. Weske, "Process model abstraction: A slider approach," in *Enterprise Distributed Object Computing Conference, 2008. EDOC'08. 12th International IEEE*. IEEE, 2008, pp. 325–331.

[17] T. Thaler, S. F. Ternis, P. Fettke, and P. Loos, "A comparative analysis of process instance cluster techniques." *Wirtschaftsinformatik*, vol. 2015, pp. 423–437, 2015.

[18] M. Leemans and W. M. van der Aalst, "Discovery of frequent episodes in event logs," in *International Symposium on Data-Driven Process Discovery and Analysis*. Springer, 2014, pp. 1–31.

[19] C. Diamantini, L. Genga, D. Potena, and E. Storti, "Pattern discovery from innovation processes," in *Collaboration technologies and systems (CTS), 2013 international conference on*. IEEE, 2013, pp. 457–464.

[20] F. Giannotti, M. Nanni, F. Pinelli, and D. Pedreschi, "Trajectory pattern mining," in *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '07. New York, NY, USA: ACM, 2007, pp. 330–339.

[21] A. Altomare, E. Cesario, C. Comito, F. Marozzo, and D. Talia, "Trajectory pattern mining for urban computing in the cloud," *Transactions on Parallel and Distributed Systems (IEEE TPDS)*, vol. 28, no. 2, pp. 586–599, 2017, iSSN:1045-9219.

[22] J. Cranshaw, R. Schwartz, J. I. Hong, and N. M. Sadeh, "The livehoods project: Utilizing social media to understand the dynamics of a city." in *ICWSM*, 2012.

[23] E. Cesario, C. Congedo, F. Marozzo, G. Riotta, A. Spada, D. Talia, P. Trunfio, and C. Turri, "Following soccer fans from geotagged tweets at fifa world cup 2014," in *Proc. of the 2nd IEEE Conference on Spatial Data Mining and Geographical Knowledge Services*, Fuzhou, China, July 2015, pp. 33–38, iSBN 978-1- 4799-7748-2.

[24] P. Georgiev, A. Noulas, and C. Mascolo, "Where businesses thrive: Predicting the impact of the olympic games on local retailers through location-based services data," *CoRR*, vol. abs/1403.7654, 2014.

[25] T. Kurashima, T. Iwata, G. Irie, and K. Fujimura, "Travel route recommendation using geotags in photo sharing sites," in *Proc. of the 19th ACM Int. Conf. on Information and Knowledge Management*, ser. CIKM '10. New York, NY, USA: ACM, 2010, pp. 579–588.

[26] K. Okuyama and K. Yanai, "A travel planning system based on travel trajectories extracted from a large number of geotagged photos on the web," in *The era of interactive media*. Springer, 2013, pp. 657–670.

[27] X. Lu, C. Wang, J.-M. Yang, Y. Pang, and L. Zhang, "Photo2trip: generating travel routes from geo-tagged photos for trip planning," in *Proceedings of the international conference on Multimedia*. ACM, 2010, pp. 143–152.

[28] Y.-T. Zheng, Z.-J. Zha, and T.-S. Chua, "Mining travel patterns from geotagged photos," *ACM Trans. Intell. Syst. Technol.*, vol. 3, no. 3, pp. 56:1–56:18, May 2012.