Ref. Ares(2023)1456088 - 28/02/2023

MASTER

Multiple ASpect TrajEctoRy management and analysis

Project Acronym	***MASTER***
Project Full Name	***M**ultiple **AS**pects **T**raj**E**cto**R**y management and analysis*
Project Number	***777695***
Deliverable Title	***Privacy and Big Data issues in holistic trajectory construction and management***
Deliverable No.	***D3.2***
Contract Delivery Date	***28/02/2023 (M60)***
Actual Delivery Date	***28/02/2023(M60)***
Responsible Authors	***Karine Zeitouni (Unit Manager of UVSQ)***
	This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie- Slodowska-Curie grant No 777695

DOCUMENT INFORMATION

GRANT AGREEMENT N.	777695
PROJECT ACRONYM	MASTER
PROJECT FULL NAME	Multiple Aspects Trajectory Management and Analysis
STARTING DATE (DUR.)	01/03/2018 (70 months)
ENDING DATE	31/12/2023
PROJECT WEBSITE	http://www.master-project-h2020.eu
COORDINATOR	Chiara Renso (CNR)
WORKPACKAGE N. | TITLE | START – END MONTH	WP3 | **Holistic trajectories construction and management** | M1 – M60
WORKPACKAGE LEADER	UNIVE
DELIVERABLE N. | TITLE	D3.2 | Privacy and Big Data issues in holistic trajectory construction and management
RESPONSIBLE AUTHOR	Karine Zeitouni (Unit Manager of UVSQ)
DATE OF DELIVERY (CONTRACTUAL)	28 February 2023 (M60)
DATE OF DELIVERY (SUBMITTED)	28 February 2023 (M60)
VERSION | STATUS	1.0
NATURE	REPORT
DISSEMINATION LEVEL	PUBLIC
AUTHORS (PARTNER)	CNR, HUA, UPRC, UFC, UFSC, DAL
CONTRIBUTORS	Antonios Makris (HUA), Chiara Renso (CNR), Christos Sardianos (HUA), Eleftherios Kofidis (UPRC), Leopoldo Soares de Melo Junior (UFC), Panagiotis Nikitopoulos (UPRC), Raffaele Perego (CNR), Stan Matwin (DAL), Michela Rial (CNR)

ACRONYM LIST

MASTER	Multiple Aspects Trajectory Management and Analysis
ICT	Information and Communication Technologies
ISTI	Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo"
CNR	Consiglio Nazionale delle Ricerche
UNIVE	Ca' Foscari University of Venice
UVSQ	University of Versailles Saint-Quentin
UFC	Federal University of Ceara'
UPRC	University of Pireaus Research Center
HUA	Harokopio University of Athens
PUC	Pontificial University of Rio de Janeiro
DAL	Dalhousie University
THIRA	Municipality of Thira
ER	Experienced Researcher
ESR	Early Stage Researcher
AIS	Automatic Identification System (AIS)

TABLE OF CONTENTS

1. INTRODUCTION

The objective of this deliverable (D3.2) is to "*report about privacy concerns and scalability issues posed in MASTER by the representation, management, and storage of holistic trajectories*" as stated in the MASTER Grant Agreement Annex 1 Part A pages 4. According to the GA, this deliverable reports on the scientific activities of two different task, one related to the privacy issues of holistic trajectories Task 3.3 and another related to the big data issues for large amount of holistic trajectories, Task 3.4. Both tasks ends at M60 and the activities were accomplished during the secondments linked to WP3.

The objective of Tasks 3.3 is to "*cope with privacy concerns posed in MASTER by the representation, management, and storage of holistic trajectories*". Indeed, holistic trajectories, due to their rich semantic annotations, pose the privacy of individuals at risk. It is therfore important to identify the issues and possible solutions. This task started at M6 and finished at M60. Objectives of this task have been achieved as detailed in Section 2.

The objective of Task 3.4 is to address the challenges raised by the consolidation of voluminous and heterogeneous movement data in different formats collected from various sources. The goal is to propose big data solutions for the management of a mass of holistic trajectories. The achieved objectives are detailed in Section 2.

The role of CNR researchers has been to work on Task 3.3 *Privacy issues in holistic trajectories* in order to define an approach for privacy on Multiple Aspects Trajectories, and Task 3.4 *Big Data solutions for holistic trajectories,* to identify and tackle issues such as unbalanced classification, in collaboration with UFC reaserchers, and ranking big data in maritime applications with researchers from DAL.

The role of UPRC researchers has been collaborating with DAL on Task 3.4 *Big Data solutions for holistic trajectories,* to devise a method of trajectories hot spots Identification and leveraging tensor structure for multi-aspect trajectory modeling.

The role of HUA researchers has been to work on Task 3.4 *Big Data solutions for holistic trajectories* to investigate scalable trajectory query processing, trajectory similarity analysis, and personalized trajectory recommendation.

2. WP3 OBJECTIVES AND TASKS

According to GA Annex I Part A, the objective of WP3 is: "[…] focus on the development of novel techniques of semantic enrichment to properly build holistic trajectories thus enabling the analysis carried out in WP4 and exploited in the application scenarios of WP5. This WP will investigate data management issues including methods to store and query holistic trajectories by considering Big Data and Privacy issues. Training is embedded in each secondment as seconded ESRs will attend seminars and courses at the hosting institution and includes the events like the Summer School and Dagstuhl Seminar."

WP3 consists of the following tasks:

T3.1: Holistic trajectory enrichment methods (Leader CNR) (M1-M24)

T3.2: Holistic trajectories data management (Leader UNIVE) (M6-M60)

T3.3: Privacy issues in holistic trajectories (Leader UVSQ) (M6-M60)

T3.4: Big Data solutions for holistic trajectories (Leader HUA) (M6-M60)

T3.5: Training on holistic trajectories building and management (Leader CNR) (M1-M60)

The work developed in preparing this deliverable is connected to the secondments linked to WP3 and specifically for Tasks 3.3 and Task 3.4. Tasks T3.1 and T3.2 are presented in deliverables D3.1 and D3.3. Results of task T3.5 on training will be presented in Deliverable D2.2 "Training and Networking activities report and material" due at M70 (31/12/2023).

Table 1 below shows the secondments linked to tasks T3.3 and T3.4 at M60, described in the present deliverable. Based on the Researcher Declarations submitted to the SyGMA system, the total number of secondments (or splits) related to these tasks is 8. The total number of person months, indicated in Table 1 column "PMs" is 16.26.

This deliverable is based on activities carried out during these secondments as stated in the GA. The research activity carried out for these tasks has produced 3 publications listed in Section 4.

Table 1: Secondments executed from M6 to M60 linked to WP3 "Holistic trajectories construction and management" tasks T3.3 and T3.4

RD #	Secondment#	Fellow ID	Profile	Secondee Name	Sending Institution	Hosting Institution	From	To	PM	Task
2	77	2	ER	Leopoldo Soares de Melo Junior	Universidade Federal do Ceará	Consiglio Nazionale Delle Ricerche	01-08-2018	19-02-2019	6.63	T3.4
13	36	14	ESR	Panagiotis Nikitopoulos	University of Piraeus Research center	Dalhousie University	01-09-2018	31-10-2018	2.00	T3.4
14	2	3	ER	Raffaele Perego	Consiglio Nazionale Delle Ricerche	Dalhousie University	26-09-2018	08-10-2018	0.43	T3.4
16	66	15	ESR	Antonios Makris	Harokopio University	Universidade Federal do Ceará	01-10-2018	30-11-2018	2.00	T3.4
17	65	16	ER	Christos Sardianos	Harokopio University	Universidade Federal do Ceará	01-10-2018	30-11-2018	2.00	T3.4
22	106	4	ER	Chiara Renso	Consiglio Nazionale Delle Ricerche	Universidade Federal De Santa Catarina	26-10-2019	30-11-2019	1.17	T3.3

39	93	26	ER	Eleftherios Kofidis	University Of Piraeus Research Center	Dalhousie University	03-02-2020	03-03-2020	1.03	T3.4
1	81	1	ER	Stan Matwin	Dalhousie University	Consiglio Nazionale delle Ricerche	01-03-2020	30-03-2020	1	T3.4

TASK 3.3 PRIVACY ISSUES IN HOLISTIC TRAJECTORIES (LEADER: UVSQ) (M6-M60).

According to the GA Annex 1 Part A this task "copes with privacy concerns posed in MASTER by the representation, management, and storage of holistic trajectories." This task has been executed during secondment #106 Chiara Renso (CNR) seconded to Federal University of Santa Catarina. During such secondment the secondee worked with Fernanda Oliveira Gomes, a PhD student of UFSC, on defining an approach for privacy on Multiple Aspects Trajectories. The proposed approach addresses the issue that the more semantic aspects we have, the more privacy of the tracked individual can be affected. The proposed approach tries to identify the aspects that can put privacy as risk case by case, distinguishing between the type of aspect as defined in the pioneering paper [Mello et al. 2019]. Starting from this, the proposed approach applies a generalization step to each trajectory aspect based on its potential of affecting individual privacy. In addition the approach uses the "swap" technique to swap the aspects while maintaining the data utility. An ongoing joint work of the secondee, Fernanda Gomes and Prof. Monreale from University of Pisa tries to formalize this idea and evaluate it in two datasets to be published as a scientific paper.

[Mello et al 2019] Ronaldo dos Santos Mello, Vania Bogorny, Luis Otavio Alvares, Luiz H. Z. Santana, Carlos Andres Ferrero, Angelo Augusto Frozza, Geomar Andre Schreiner, Chiara Renso. MASTER: A Multiple Aspect View on Trajectories. Transactions in GIS. First published: 09 May 2019 https://doi.org/10.1111/tgis.12526

TASK 3.4 BIG DATA SOLUTIONS FOR HOLISTIC TRAJECTORIES (LEADER: HUA) (M6-M60).

According to the GA Annex 1 Part A this task "focuses on the scalability issues posed in MASTER by the representation, management, and storage of holistic trajectories."

The activity related to this task has been executed during several secondments, as shown in Table 1 (#77, #36, #2, #66, #65, #93, #81), involving partners CNR, HUA, UPRC, DAL, and UFC.

During secondment #77, Leopoldo Soares de Melo Junior seconded to CNR, collaborated with CNR researchers Roberto Trani (an ESR from CNR), Franco Maria Nardini and Chiara Renso (researchers at CNR). He studied binary classification approaches to imbalanced datasets. Imbalanced datasets are datasets where one of the two classes (positive/negative) has considerably more instances than the other class. This generally causes problems in prediction machine learning-based systems. This is a general problem that can be applied to

trajectories datasets as well as non-mobility datasets. First, the proposed approach evaluates the performance of cost-sensitive classifiers and ensembles on imbalanced datasets. Next, they evaluated some stalking combinations to improve the classification performance of imbalanced data. Finally, they studied the combination of sampling techniques, pool generators, and dynamic classification selection to deal with the imbalance problem. As a result, they found that an extension of a dynamic ensemble selection combined with Random Forest consistently improves the classification performance of moderate and highly imbalanced datasets. As a result, a paper has been published [1].

During secondment #36, Panagiotis Nikitopoulos, ESR from UPRC seconded to DAL, worked with Dalhousie's Big Data Analytics research group on identifying hot-spots on big trajectory data. Hot-spot discovery is based on a spatiotemporal partitioning of the space into predefined areas, followed by some statistical analysis, such as the Getis-Ord statistic. However, the existing algorithms do not consider the peculiarities of trajectories, that is the limited sampling rate, the movement direction, and the multiple types of attribute values. During this secondment, Panagiotis proposed new techniques for discovering such hot-spots over AIS maritime data. First, he employed the well-known "Fast Voxel Traversal Algorithm for Ray Tracing" to interpolate trajectory points between two sequential input samples. He employed a different spatio-temporal partitioning scheme, based on hexagons. Furthermore, he proposed a new formula for assigning attribute values to spatio-temporal areas, regardless of the type of the attribute value. This work has been published paper [3] listed in section 4.

During secondment #2 from CNR to DAL (this is a split) the secondee and the hosting staff discussed how the ranking problem can be adapted for the analysis of large amount of AIS data.

During secondment #66 by a staff member of HUA to Federal University of Ceara', the secondee focused on the following problems:

- Explore the topic of scalable query processing using trajectory data
- Explore the topic of trajectory data compression
- Develop algorithms for trajectory data compression from points dataset
- Visualize the results

The emphasis was put on the first three items. The secondment research outcome has been published in a journal paper [2] listed in Section 4.

Secondment #65 from HUA to UFC focused on the following topics:

- Explore the topic of personalised trajectory recommendations
- Explore the topic of trajectory enrichment with semantic geographical information
- Explore the topic of Location-Based Social Networks (LBSNs) based on sub-trajectory similarities enriched with Facebook profile information
- Develop a library for measuring sub-trajectory similarity, filtering user's different sub-trajectories and recommending sub-trajectories or intermediate stop points for a root.

Secondment #93 from UPRC to DAL focused on the intrinsic multi-dimensionality of multimodal data in holistic trajectories and how it can be exploited through a modelling with multi-way arrays (aka tensors) and their (coupled) decomposition. During this secondment, an extensive review of the related literature and contributed proposals of alternative tensor models and associated method has been completed.

Secondment #81 from DAL to CNR focused on methods to use for the analysis of large amount of AIS data, considering the ranking approach and how to adapt these methods for large amount of AIS data.

3. CONCLUSIONS

In summary, main contributions related to this deliverable are in the topic of Privacy and Big Data.
For Privacy we have developed:

- An approach to define a privacy preserving framework for multiple aspects trajectories (Sec #106)

And for the big data challenge we have reached the following results:

- Machine learning for heavily imbalanced data (Sec #77)
- Identification of trajectories hot spots (Sec #36)
- Ranking algorithms for big vessels data (Sec #2 and #81)
- Scalable query processing on trajectory data (Sec #66)
- Personalized Trajectory Recommendation and trajectory similarity (Sec #65)
- Multi ways arrays for multiple aspects trajectories (Sec #93)

These scientific results are reported also in the 3 publications listed in Section 4.
The research activities carried out during secondments also included training activities of ESRs, including participation to seminars and courses, as specified in the GA.

4. PUBLICATIONS

[1] Leopoldo Melo Junior, Franco Maria Nardini, Chiara Renso, Roberto Trani, José Antônio Fernandes de Macêdo: A novel approach to define the local region of dynamic selection techniques in imbalanced credit scoring problems. Expert Syst. Appl. 152: 113351 (2020). This publication is n 44 in SyGMA.

[2] Antonios Makris, Camila Leite da Silva, Vania Bogorny, Luis Otávio Alvares, José Antônio Fernandes de Macêdo, Konstantinos Tserpes: Evaluating the effect of compressing algorithms for trajectory similarity and classification problems. GeoInformatica 25(4): 679-711 (2021) This publication is n. 34 in SyGMA.

[3] P. Nikitopoulos, A.-I. Paraskevopoulos, C. Doulkeridis, N. Pelekis, Y. Theodoridis, Hot Spot Analysis for Big Trajectory Data. Proceedings of the 2018 IEEE Int'l Conference on BIG DATA (BigData2018), Seattle, WA, USA, 2018. IEEE CPS. This publication is n. 10 in SyGMA.