Short-Term Scientific Mission Grant
- APPLICATION FORM[1] -

Action number: CA20111

Applicant name: STEFANIA DUMBRAVA

Details of the STSM

Title: *Towards Reliable Distributed Graph Databases: Automated Verification of a Conflict-Free Replicated Property Graph Data Structure*

Start and end date: 12/07/2022 to 21/07/2022 (must end before October 31)

Travel cost: 150 Euros

Accomodation cost: 850 Euros

Living cost: 300 Euros

Requested grant: 1300 Euros

Goals of the STSM

Purpose and summary of the STSM.

(max.200 word)

Stefania Dumbrava, the visiting researcher, is an associate professor at ENSIIE and TelecomSud Paris (SAMOVAR). She is an expert in formal methods and databases, as well as a member of LDBC's Property Graph Schema Working Group, designing the standard language for graph databases.

Mario Pereira, the host, is an assistant professor at Universidade NOVA, Lisbon. He is an expert in software verification and functional programming. He is the lead developer of Cameleer verification framework and an active member of the Why3 development team.

Stefania and Mario met as PhD candidates, working at the Laboratory for Formal Methods in Orsay. Interested in exploring interactions between our research topics, we propose this visit to kick-start a collaboration aimed at:

1. holding a joint seminar to outline the open challenges and opportunities of using current verification tools to improve graph database reliability. The seminar will target researchers and students at the Universidade NOVA.

2. investigating the usage of automated tools for distributed graph database reliability.

We will specify a property graph conflict-free replicated data type (PG-CRDT) and turn it into a verified prototype. This use-case will highlight the interoperability of different OCaml verification tools and the challenges of automated reasoning in this setting.

[1] This form is part of the application for a grant to visit a host organisation located in a different country than the country of affiliation. It is submitted to the COST Action MC via-e-COST. The Grant Awarding Coordinator coordinates the evaluation on behalf of the Action MC and informs the Grant Holder of the result of the evaluation for issuing the Grant Letter.

Working Plan

Description of the work to be carried out by the applicant.

(max.500 word)

The visit will last for 10 days, from July 12th 2022 to July 21st 2022 and will contribute to WG2: Automated Theorem Provers and WG3 (applications): Program Verification.

The work plan is structured in two parts, following the outlined goals.

1. **July 12th - July 14th: Joint Seminar** “Towards Improving the Safety and Reliability of Graph Database Systems”.

The first session will be devoted to an introductory course on graph databases, using the Neo4j system as a core tool. The course targets people with general knowledge of computer science, but no previous knowledge of graph databases. The second session will give an overview of automated deductive verification tools. We will focus particularly on verification tools targeting the OCaml ecosystem, i.e., Cameleer and Why3. The third session consists of discussions on the research challenges inherent to applying automated verification in the distributed graph database setting. To this end, we will identify a list of open problems related to reasoning about the correctness of non-centralised processing of semi-structured data.

2. **July 15th - July 19th: Research Discussions.**

We will discuss how novel database systems, such as graph databases, can benefit from the strong guarantees provided by formal verification tools. Starting from the challenging distributed setting, we will focus on identifying a methodology that can push the boundaries of automation in deductive verification and provide a usable framework for database practitioners.

Indeed, automated deductive verification, normally relying on SMT solvers to discharge proof obligations, has made tremendous progress in the course of the last decades. The so-called SMT revolution had a crucial impact in scaling up automated verification to industrial-level challenges. One can cite the Verisoft XT project (verification of Microsoft’s Hypervisor and PikeOS, using VCC program verifier and the Z3 SMT solver) and the Everest project (verification of cryptographic software, using the F* proof assistant and the Z3 SMT solver) as successful applications of automated verification technologies in real world applications.

It is the overarching objective of this short-term mission to investigate the applicability of such tools in the context of processing real-world interconnected data, using the state-of-the-art graph database technologies.

We hope to structure the discussion along 3 axes:

• [15-16 July] Challenges of data storage and management in the distributed setting. Specifying and designing a prototype implementation of a PG-CRDT structure for distributed graph databases.

• [17 July] Survey of the current OCaml verification environment and discussion on automation and interoperability.

• [18-19 July] Challenges of formal verifying the OCaml PG-CRDT structure.

• **July 20th. 9th: Wrap up.** We will reserve the last day of the visit to gather course feedback and questions to lay out a plan for potential future collaborations, e.g., applying to common grants at the national and international level.

• **July 21st: Return to Paris.**

Expected outputs and contribution to the Action MoU objectives and deliverables.

Main expected results and their contribution to the progress towards the Action objectives (either research coordination and/or capacity building objectives) and deliverables.

Working groups to which this mission contributes: WG2 and WG3

(max.500 words)

At the end of the action, we aim to have drafted an experience report on the capabilities and limitations of employing SMT solvers to automatically verify OCaml code in Why3, in the context of designing a reliable specification for distributed property graph processing.

As deliverables, we set to design a draft specification of PG-CRDT structures and a formal library containing base building blocks for verifying distributed graph databases in Why3.

Other than contributing to the work done in WG2 and WG3, we expect to also contribute to advancing the overall objectives of the EuroProofNet COST action at several levels.

Concerning the **research coordination objectives**, we will contribute to the following points:

• Point 2 - “Promote the output of detailed, checkable proofs from automated theorem provers”: this follows from working on the verification of the PG-CRDT structures using Why3 and the integrated automated theorem proves.

• Point 3 - “Make techniques for program verification more effective and more accessible to all stakeholders”: this will be a consequence of working on the interoperability of Cameleer and Why3, in the context of developing a verified library for distributed graph database processing.

Concerning the **capacity building objectives**, we expect to contribute to the following points:

• Point 1 - “Bring together members of the different communities working on proofs in Europe”: our open seminar will provide an exchange forum between researchers from Universidade NOVA working in different formal methods approaches, namely behavioral type systems and information control flow.

• Point 2 - ”Act as a stakeholder platform in the field of formal proofs from its theoretical grounds to its industrial applications': the topic of our seminar will be an useful opportunity to interact and gain feedback from interested industrial partners, notably the developers of AntidoteDB (in Lisbon) and industrial users of CRDTs (e.g., Tarides in France).

• Point 3 - “Create an excellent and inclusive network of researchers in Europe with lasting collaboration beyond the lifetime of the Action”: the visit will provide an occasion to plan further bi-lateral collaborations targeting scientific publications, writing a joint project, and applying to funding at the national and international level.

• Point 4 - “Ease access to formal verification techniques in education and other areas of science”: for dissemination purposes, the seminar will be recorded and the use-case targeting distributed graph processing data structures and algorithms can serve as a starting point for further pedagogical purposes (inclusion in the gallery of verified Why3 programs, scientific publications, and demos).

• Point 5 - “Actively support young researchers, the under-represented gender, and teams from regions with less capacity”: both the visiting researcher and the host are young professors that do not currently have their own research grants to fund exchange visits. This mission would be a very useful opportunity to advance on an interdisciplinary topic and plan the submission of common projects.