Stefano Zacchiroli
Télécom Paris
19 place Marguerite Perey
91120 Palaiseau, France
stefano.zacchiroli@telecom-paris.fr

May 19, 2022

Bachelor Thesis Results Approbation Act

Graduate last name, first name: Starodubtsev, Andrey Igorevich

Title of thesis project: Implementing TinkerPop infrastructure for WebGraph

Initial problem statement:

- *Context:* Software Heritage[1] is an ambitious research project whose goal is to collect, preserve in the very long term, and share the whole publicly accessible Free/Open Source Software (FOSS) in source code form.
- *Description:* Software Heritage uses the WebGraph[2] framework for graph compression. This allows to manipulate the huge archive Merkle DAG in RAM efficiently, via the swh-graph[3] component. The current RPC API to navigate the graph is however very limited and ad hoc. We would like to exploit the current compressed graph representation using a standard graph traversal language such as the Gremlin graph traversal language. The goal of this work was to design, implement, and experiment with a backend for Apache TinkerPop[4] (a popular open source implementation of Gremlin) that sits on top of WebGraph. If successful it will allow to traverse the huge Software Heritage graph with both the current efficiency and the convenience of a high-level and expressive graph traversal language.

List of solved problems: Implemented TinkerPop infrastructure for WebGraph with support of unmodifiable data, and vertex/edge properties. Added support for different vertex/edge property sources. Assessed the performance of the solution by implementing Gremlin queries for a real dataset (the Software Heritage graph) and profiled their execution. Implemented several TinkerPop queries corresponding to real-world Software Heritage use cases.

Assessment of the amount of work done: in terms of code: 3000 lines of code, released

[1] https://www.softwareheritage.org/

[2] https://webgraph.di.unimi.it/

[3] https://docs.softwareheritage.org/devel/swh-graph/

[4] https://tinkerpop.apache.org/

as open source software (see the swh-graph-tinkerpop[5] and webgraph-tinkerpop[6] repositories); complemented by all the associated design, debugging, and benchmarking effort contributed by the candidate.

Degree of results approbation in enterprise activities: at Software Heritage we are very pleased of the work conducted by the candidate. It has allowed us to evaluate the usefulness of an expressive, general purpose graph traversal language on top of our huge graph, which was previously queried via an ad-hoc traversal API only. The achieved performances are really good, consider the extra overhead that is naturally incurred by having to bridge two worlds (TinkerPop and WebGraph).

Description of the impact of innovation integration: this practical experiment has allowed us to evaluate the appropriateness of using TinkerPop, as a language, to query and exploit the Software Heritage graph at scale. Based on this practical assessment we have determined that the language (TinkerPop) is a bit too verbose for being used as-is to address our practical use cases. As a result of the work conducted in this thesis we now have a solid implementation basis for the next language-design steps. In particular we are looking into developing higher-level (as in: more abstract) graph traversal operators that can be plugged into TinkerPop to make query writing more natural (for our specific domain). This work is now actionable thanks to the foundational work conducted as part of this bachelor thesis.

Sincerely,

Stefano Zacchiroli
Professor
Télécom Paris
Polytechnic Institute of Paris

[5] https://github.com/andrey-star/swh-graph-tinkerpop

[6] https://github.com/andrey-star/webgraph-tinkerpop