# Parsing techniques for graph analysis

Semyon Grigorev
Saint Petersburg State University
7/9 Universitetskaya nab.
St. Petersburg, 199034 Russia
Semen.Grigorev@jetbrains.com

Ekaterina Verbitskaia
Saint Petersburg State University
7/9 Universitetskaya nab.
St. Petersburg, 199034 Russia
kajigor@gmail.com

Nowadays input data for parsing algorithms are not limited to be linear strings, and context-free grammars are used not only for programming languages specification. One classical example is context-free path querying for graph data bases where an input is a graph and path constraints are specified by a grammar. Graph parsing may find an application in different areas: in software engineering for dynamically generated strings analysis, in graph data bases for paths querying, etc. The idea of multiple input GLL parsing, presented at Parsing@SLE-2016 by Elizabeth Scott and Adrian Johnstone, is also a particular case of graph parsing: a set of token-with-extent can be treated as a directed graph where extents are vertices and tokens label the edges. Thus, graph parsing can be considered as a great connection (!!!!!!) of multiple computer science areas: formal languages theory, parsing algorithms, data bases, graph theory.

Our group is working on several questions posed in this area [3, 10] which still do not have satisfying solutions. Our efforts are mostly aimed at improving performance, lifting up limitations on an input and finding new fields of application for graph parsing.

We already developed several graph parsing algorithms and applied them to different problems. First of all, we created a RNGLR-based algorithm and applied it to the analysis of dynamically generated SQL queries [5]. GLL-based context-free path querying algorithm [2], implemented by the authors, runs faster than the solution presented at ISWC-2016 [6]. Our algorithm based on matrix multiplication [1] allows one to utilize GPGPU for graph processing, and it is faster than the GLL-based, but it does not construct a parsing forest.

Currently, we are working on an extension for Meerkat [11] library which allows one to use parser-combinators for graph parsing and integrates context-free querying into the programming language with no need to use designated DSLs. Another direction of work is extending the algorithm based on matrix multiplication with the support for conjunctive grammars [7]. This will make it possible to execute more complex queries which can be utilized for pseudoknots finding. By mechanization of the GLL-based algorithm in Coq and proving its correctness, we hope to build a foundation for formal reasoning about the extensions under development.

The most exciting area of graph parsing application for us now is a problem of context-free pattern search in metagenomic assemblies. An assembly may be presented as a graph whereas a secondary structure of a sequence to search for can be specified in terms of a grammar. Some structures in biological sequences, such as pseudoknots, require conjunctive grammars for structure description, which motivates moving towards grammar classes which are more descriptive than context-free grammars.

One of future direction of our research is an adoptation of advanced matrix multiplication techniques, such as approximated matrix multiplication and sparse matrix multiplication, for graph parsing. We hope to get more effective algorithms for huge graphs processing. Also, we want to apply the algorithm based on matrix multiplication for boolean grammars [7]. Unfortunately, this means that we will only be able to get an approximation of the result since the problem of graph parsing is undecidable even for conjunctive grammars. Moreover, as parsing with respect to boolean grammars is not monotonic, it is impossible to naively extend the solution for conjunctive grammars. Another research direction is an effective algorithms for non-recursive context-free grammars intersection [8, 9] which can be used in speech recognition or for compressed strings processing, and finding of other types of grammars with decidable intersection prolem. In the practical way, the one goal of our work is to create an abstract framework for parsing based on the generalization of GLL parsing algorithm [4]. Finally, we want to investigate practical areas of application and to create solutions based on our framework to demonstrate its practical value.

## 1. REFERENCES

[1] Azimov, Rustam, and Semyon Grigorev. “Graph Parsing by Matrix Multiplication.” *arXiv preprint arXiv:1707.01007* (2017).

[2] Grigorev, Semyon, and Anastasiya Ragozina. “Context-Free Path Querying with Structural Representation of Result.” *arXiv preprint arXiv:1612.08872* (2016).

[3] Hellings, Jelle. “Querying for Paths in Graphs using Context-Free Path Queries.” *arXiv preprint arXiv:1502.02242* (2015).

[4] Izmaylova, Anastasia, Ali Afroozeh, and Tijs van der Storm. “Practical, general parser combinators.” *Proceedings of the 2016 ACM SIGPLAN Workshop on Partial Evaluation and Program Manipulation.* ACM, 2016.

[5] Nederhof, Mark-Jan, and Giorgio Satta. “Parsing non-recursive context-free grammars.” *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics.* Association for Computational Linguistics, 2002.

[6] Nederhof, Mark-Jan, and Giorgio Satta. "The language intersection problem for non-recursive context-free grammars." *Information and Computation* 192.2 (2004): 172-184.

[7] Okhotin, Alexander. "Conjunctive and Boolean grammars: the true general case of the context-free grammars." *Computer Science Review* 9 (2013): 27-59.

[8] Scott, Elizabeth, and Adrian Johnstone. "GLL parsing.", *Electronic Notes in Theoretical Computer Science*, 253.7 (2010): 177–189.

[9] Verbitskaia, Ekaterina, Semyon Grigorev, and Dmitry Avdyukhin. "Relaxed Parsing of Regular Approximations of String-Embedded Languages." *International Andrei Ershov Memorial Conference on Perspectives of System Informatics.* Springer International Publishing, 2015.

[10] Yannakakis, Mihalis. "Graph-theoretic methods in database theory." *Proceedings of the ninth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems.* ACM, 1990.

[11] Zhang, Xiaowang, et al. "Context-free path queries on RDF graphs." *International Semantic Web Conference.* Springer International Publishing, 2016. 632–648.