Fine-grained Reductions Around CFL Reachability

ALEKSANDRA ISTOMINA, Saint Petersburg State University, Russia and JetBrains Research, Russia

RESEARCH ADVISOR: SEMYON GRIGOREV, Saint Petersburg State University, Russia and JetBrains Research, Russia

1 INTRODUCTION

CFL Reachability problem finds application in different fields of research: static code analysis (e.g. type-based flow analysis [15] or points-to analysis [18, 19]), graph databases, bioinformatics. It's goal is to detect valid paths between vertices in a graph.

There are several cubic [16, 21] and slightly subcubic [5] (with time $O(n^3/polylog(n))$) algorithms for CFL rechabilty. The big open question is whether a truly subcubic (with time $\tilde{O}(n^{3-\epsilon}) = O(n^{3-\epsilon} \cdot polylog(n))$) algorithm exists.

One of the ways to answer that question is given by fine-grained complexity. If one makes a fine-grained reduction from other problem that is known or strongly believed to have some lower bound to CFL Reachability problem, then CFL Reachability problem will also have lower bound. Currently there are several results in the area, but they are scattered and have no structure.

In this paper we give an overview on these existing results: what is already achieved and what questions are still to be answered. We have focused mostly on static problems and connections between them, dynamic problems are omitted from this paper.

2 PRELIMINARIES

Context-free grammar (CFG) is a four $G = (N, \Sigma, P, S)$, where N is a set of nonterminals, Σ is a set of terminals, P is a set of productions of the followings form: $A \to \alpha$, $\alpha \in (N \cup \Sigma)^*$ ans S is a starting nonterminal. Denote a context-free language of words derived from the starting symbol as $L(G)$.

CFG Recognition problem is to decide whether $w \in L(G)$ given a CFG G and a string $w \in \Sigma^*$. This problem is closely related to *CFG parsing problem* where we want a possible derivation sequence, if $w \in L(G)$. It is known [17] that CFG recognition is as hard as CFG parsing up to logarithmic factors.

Let $D = (V, E)$ be a directed graph which edges are labelled with symbols from P. We call a path from vertex v to vertex u an S-path if concatenation of labels on that path is a word from $L(G)$. *Context-free-language (CFL) Reachability* problem [12] is to determine if there exists an S-path between a given sets of vertices A, B. In *single source/single target (s-t)* CFL Reachability $A = \{s\}, B = \{t\}, s, t \in V(D)$. In *all-pairs* CFL Reachability $A = B = V(D)$.

Dyck-k reachability problem is a CFL reachability problem where G defines a Dyck language on k types of parentheses. The corresponding grammar is $G = (N, \Sigma, P, S)$, where $N = \{S\}, \Sigma = \{(_i,)_i\}, \forall i = 1, \ldots, k$, productions rules are $S \to \epsilon |SS|(_1 S)_1| \ldots |(_k S)_k$, where ϵ is the empty string.

For problems P, Q and time bounds t_P, t_Q, a *fine-grained reduction* [2] from (P, t_P) to (Q, t_Q) is an algorithm that, given an instance I of P, computes an instance J of Q such that:

- I is a YES-instance of P if and only if J is a YES-instance of Q,
- for any $\epsilon > 0$ there is a $\delta > 0$ such that $t_Q(|J|)^{1-\epsilon} = O(t_P(|I|)^{1-\delta})$,
- the running time of the reduction is $O(t_P(|I|)^{1-\gamma})$ for some $\gamma > 0$.

Aleksandra Istomina: graduate student, email: aleksandra2999@mail.ru, ACM student member number: 4678238.

3 MAIN RESULTS

This section is organised as follows. Firstly, we define necessary problems. Secondly we present a map of existing fine-grained reductions and make an overview of the most interesting ones. After that we state some open problems and possible ways to solve them.
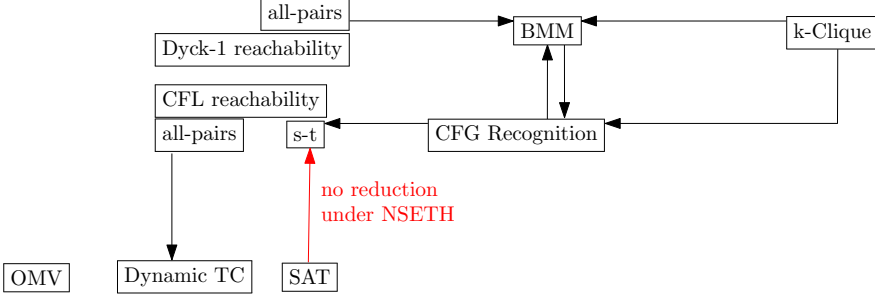

Fig. 1. Existing reductions concerning CFL Reachability and CFG Recognition. Black arrow $a \rightarrow b$ represent existing reduction from a to b. Red arrow analogously represent non-existence of the reduction.

3.1 Existing problems and hypotheses

There are several problems that are connected with CFL Recognition and Reachability.

Boolean satisfiability problem (SAT, k-SAT) is to determine if there exists an interpretation of variables that satisfies a given Boolean formula on n variables written in k-CNF, $k > 3$. The hypothesis about SAT, that we are interested about, is NSETH [3] which proposes that there is no $\epsilon > 0$ such that k-SAT can be solved co-nondeterministically in time $2^{(1-\epsilon)n}$ for any k.

In *Boolean Matrix Multiplication (BMM)* problem it is needed to calculate matrix product of the two given $n \times n$ matrices over (AND, OR). BMM hypothesis states that there is no $O(n^{3-\epsilon})$ combinatorial algorithm for that.

Orthogonal Vectors (OV) problem decides whether two sets X, Y of n boolean d-dimensional vectors contain a pair $x \in X, y \in Y$ which dot product equals zero. Hypothesis states that OV problem can not be solved in $O(n^{2-\epsilon} \cdot poly(d))$ time.

Given an undirected graph U the k-*Clique* problem seeks the clique on k vertices in U.

The incremental *Dynamic Transitive Closure (DTC)* [7] problem asks to maintain reachability information in a directed graph between arbitrary pairs of vertices under insertions of edges.

In the *Online boolean Matrix-Vector multiplication (OMV)* [9] problem we are given an $n \times n$ boolean matrix M, we receive n boolean vectors $v_1, \ldots, v_n$ one at a time, and are required to output Mv_i (over the boolean semiring) before seeing the vector v_{i+1}, for all i.

3.2 Existing reductions

Here we present fine-grained complexity results about CFL Reachability and Recognition problems.

First of all we need to mention the interconvertibility of CFL Reachability problems and a class of set-constraint problems [13] as this result allows us to reformulate our problem if we wish so.

One of the interesting reductions was a reduction from all-pairs Dyck-1 Reachability problem to BMM problem [1], [11]. It was firstly proved by Bradford via algebraic matrix encoding and then combinatorially by combining Dyck-1 path from bell-shaped paths.

In one of the recent papers [14] the reduction from all-pairs CFL Reachability to incremental DTC have been proven.

The following results are the most interesting ones concerning the existence of the truly subcubic algorithm for CFL Reachability.

The problem has been shown to be 2NPDA-complete [8]. It means that subcubic algorithm for CFL Reachability would lead to subcubic algorithms for the whole 2NPDA class and cubic upper bound has not been improved since discovery of the class in 1968.

Non-existence of the truly subcubic combinatorial algorithm for s-t CFL Reachability under BMM hypothesis was proved combining two reductions: the combinatorial reduction from CFG recognition to s-t CFL reachability [4] and combinatorial reduction from BMM to CFG Recogniiton [10]. This means that if one wants to find the truly subcubic algorithm for CFL Reachability, it should not be combinatorial one.

Recently it was discovered [6] that there exist subcubic certificates for s-t CFL Reachability (for existence and non-existence of the valid paths). From this fact it follows that there are no reductions under NSETH from SAT problem to CFL Reachability problem.

3.3 Open problems

The equivalence under subcubic reduction of s-t and all-pairs CFL reachability is not yet discovered. However it is known [20] that detecting one triangle of a total negative edge weight in a weighted graph and up to $n^{3-\epsilon}, \epsilon > 0$ such triangles are subcubic equivalent tasks.

Currently there is no non-trivial lower bound on Dyck-1 reachability problem. Yet we know that there exists a reduction from OV problem to Andersen pointer analysis [11] where as a part of reduction appears slightly modified Dyck-1 reachability with additional if-condition for edge existence in a graph. If similar reduction exists to pure Dyck-1 reachability problem it would give conditional quadratic lower bound.

Finding a fine-grained reduction from all-pairs shortest paths (APSP) problem or OMV problem to CFL reachability problem would give a conditional lower bound on it's complexity. We highlight APSP problem and it's subcubic equivalent analogues [20] as APSP is connected to problems on paths. We highlight OMV problem as it is connected to dynamic problems as is CFL reachability problem.

4 CONCLUSION AND FUTURE WORK

In this paper we have collected existing results in fine-grained complexity concerning CFL Reachability and CFG Recognition problems. We have presented existing reductions and some open problems with possible ways of their solution.

CFL reachability is a popular problem strongly connected with many areas. It has several cubic algorithms and no truly subcubic combinatorial one under BMM hypothesis. It has been proved that we can't get cubic lower bound on CFL reachability using reduction from SAT under NSETH.

Still other reductions may be possible. For example, from APSP or OMV problems. Getting cubic lower bound through them is a possible way of future work as is closing other open problems.

This overview was focused on static problems and reductions between them. We have reached the DTC problem in a big field of dynamic problems, where are many reductions of our interest. We plan to extend our research there.

5 ACKNOWLEDGMENTS

REFERENCES

[1] Phillip G Bradford. 2017. Efficient exact paths for Dyck and semi-Dyck labeled path reachability. In *2017 IEEE 8th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON)*. IEEE, 247–253.

[2] Karl Bringmann. 2019. Fine-Grained Complexity Theory. In *36th International Symposium on Theoretical Aspects of Computer Science*. 1.

[3] Marco L. Carmosino, Jiawei Gao, Russell Impagliazzo, Ivan Mihajlin, Ramamohan Paturi, and Stefan Schneider. 2016. Nondeterministic Extensions of the Strong Exponential Time Hypothesis and Consequences for Non-Reducibility. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science* (Cambridge, Massachusetts, USA) *(ITCS '16)*. Association for Computing Machinery, New York, NY, USA, 261–270. https://doi.org/10.1145/2840728.2840746

[4] Krishnendu Chatterjee, Bhavya Choudhary, and Andreas Pavlogiannis. 2017. Optimal Dyck Reachability for Data-Dependence and Alias Analysis. *Proc. ACM Program. Lang.* 2, POPL, Article 30 (Dec. 2017), 30 pages. https://doi.org/10.1145/3158118

[5] Swarat Chaudhuri. 2008. Subcubic Algorithms for Recursive State Machines. In *Proceedings of the 35th Annual ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages* (San Francisco, California, USA) *(POPL '08)*. Association for Computing Machinery, New York, NY, USA, 159–169. https://doi.org/10.1145/1328438.1328460

[6] Dmitry Chistikov, Rupak Majumdar, and Philipp Schepper. 2021. Subcubic Certificates for CFL Reachability. *arXiv preprint arXiv:2102.13095* (2021).

[7] Kathrin Hanauer, Monika Henzinger, and Christian Schulz. 2020. Faster Fully Dynamic Transitive Closure in Practice. *ArXiv* abs/2002.00813 (2020).

[8] Nevin Heintze and David McAllester. 1997. On the Cubic Bottleneck in Subtyping and Flow Analysis. In *Proceedings of the 12th Annual IEEE Symposium on Logic in Computer Science (LICS '97)*. IEEE Computer Society, USA, 342.

[9] Kasper Green Larsen and Ryan Williams. 2017. Faster Online Matrix-Vector Multiplication. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms* (Barcelona, Spain) *(SODA '17)*. Society for Industrial and Applied Mathematics, USA, 2182–2189.

[10] Lillian Lee. 2002. Fast Context-Free Grammar Parsing Requires Fast Boolean Matrix Multiplication. *J. ACM* 49, 1 (Jan. 2002), 1–15. https://doi.org/10.1145/505241.505242

[11] Anders Alnor Mathiasen and Andreas Pavlogiannis. 2021. The Fine-Grained and Parallel Complexity of Andersen's Pointer Analysis. *Proc. ACM Program. Lang.* 5, POPL, Article 34 (Jan. 2021), 29 pages. https://doi.org/10.1145/3434315

[12] David Melski and Thomas Reps. 1997. Interconvertbility of Set Constraints and Context-Free Language Reachability. *SIGPLAN Not.* 32, 12 (Dec. 1997), 74–89. https://doi.org/10.1145/258994.259006

[13] David Melski and Thomas Reps. 1997. Interconvertbility of Set Constraints and Context-Free Language Reachability. In *Proceedings of the 1997 ACM SIGPLAN Symposium on Partial Evaluation and Semantics-Based Program Manipulation* (Amsterdam, The Netherlands) *(PEPM '97)*. Association for Computing Machinery, New York, NY, USA, 74–89. https://doi.org/10.1145/258993.259006

[14] Egor Orachev, Ilya Epelbaum, Rustam Azimov, and Semyon Grigorev. 2020. *Context-Free Path Querying by Kronecker Product*. 49–59. https://doi.org/10.1007/978-3-030-54832-2_6

[15] Jakob Rehof and Manuel Fähndrich. 2001. Type-Base Flow Analysis: From Polymorphic Subtyping to CFL-Reachability. *SIGPLAN Not.* 36, 3 (Jan. 2001), 54–66. https://doi.org/10.1145/373243.360208

[16] Thomas Reps, Susan Horwitz, and Mooly Sagiv. 1995. Precise Interprocedural Dataflow Analysis via Graph Reachability. In *Proceedings of the 22nd ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages* (San Francisco, California, USA) *(POPL '95)*. Association for Computing Machinery, New York, NY, USA, 49–61. https://doi.org/10.1145/199448.199462

[17] Walter L. Ruzzo. 1979. On the Complexity of General Context-Free Language Parsing and Recognition (Extended Abstract). In *Proceedings of the 6th Colloquium, on Automata, Languages and Programming*. Springer-Verlag, Berlin, Heidelberg, 489–497.

[18] Manu Sridharan and Rastislav Bodík. 2006. Refinement-Based Context-Sensitive Points-to Analysis for Java. *SIGPLAN Not.* 41, 6 (June 2006), 387–400. https://doi.org/10.1145/1133255.1134027

[19] Manu Sridharan, Denis Gopan, Lexin Shan, and Rastislav Bodík. 2005. Demand-Driven Points-to Analysis for Java. *SIGPLAN Not.* 40, 10 (Oct. 2005), 59–76. https://doi.org/10.1145/1103845.1094817

[20] Virginia Vassilevska Williams and R. Ryan Williams. 2018. Subcubic Equivalences Between Path, Matrix, and Triangle Problems. *J. ACM* 65, 5, Article 27 (Aug. 2018), 38 pages. https://doi.org/10.1145/3186893

[21] Mihalis Yannakakis. 1990. Graph-Theoretic Methods in Database Theory. In *Proceedings of the Ninth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems* (Nashville, Tennessee, USA) *(PODS '90)*. Association for Computing Machinery, New York, NY, USA, 230–242. https://doi.org/10.1145/298514.298576