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

CCS Concepts: • **Theory of computation → Grammars and context-free languages**; **Program analysis**; **Problems, reductions and completeness**.

1 INTRODUCTION

Context-free language (CFL) reachability is a framework for graph analysis which was introduced by Thomas Reps [16] and allows one to specify path constraints in terms of context-free languages. CFL reachability finds application in such fields of research as static code analysis (e.g. type-based flow analysis [15] or points-to analysis [21, 22]), graph databases [25], bioinformatics [19].

There are several cubic [17, 25] and slightly subcubic [6] (with time $O(n^3/\log n)$) algorithms for CFL reachability. 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 to make a fine-grained reduction from other problem that is known or strongly believed to have some lower bound to CFL reachability problem. In that case CFL reachability problem will have non-trivial conditional lower bound as faster algorithm for it will lead to faster algorithm for other problem. Currently there are several fine-grained complexity results in this 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.

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 [18] that CFG recognition is as hard as CFG parsing up to logarithmic factors.

Let $D = (V, E, L)$ be a directed graph which edges are labelled with symbols from $L \subseteq \Sigma$. 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 [16] 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 $N = \{S\}, \Sigma = \{(_i,)_i\}, \forall i = 1, \dots, k$, productions rules are $S \to \epsilon | SS | (_1 S)_1 | \dots | (_k S)_k$, where ϵ is the empty string.

For problems P, Q and time bounds t_P, t_Q, a *fine-grained reduction* [3] 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 fine-grained complexity problems which we use later. Secondly we present a map of existing fine-grained reductions and make an overview of some selected ones. After that we discuss some open problems.

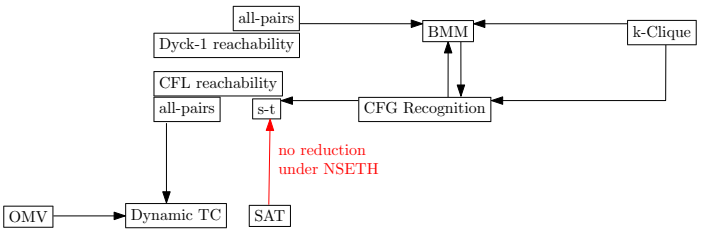

Fig. 1. Existing reductions concerning CFL reachability and CFG recognition. Black arrow $a \longrightarrow b$ represents existing reduction from a to b. Red arrow analogously represents non-existence of the reduction.

3.1 Existing problems and hypotheses

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 [4] 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 on n vertices the *k-Clique* problem [1] seeks the clique on k vertices in U. If $0 \leq F \leq \omega$ and $0 \leq C \leq 3$ are the smallest numbers such that $3k$-Clique can be solved combinatorially in $O(n^{Ck})$ time and in $O(n^{Fk})$ time by any algorithm, for any constant $k \geq 1$, a conjecture is that $C = 3$ and $F = \omega$.

In the *Online boolean Matrix-Vector multiplication (OMV)* problem [11] 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. It is conjectured that there is no algorithm with total time $O(n^{3-\epsilon})$, even with polynomial time to preprocess M.

The incremental *Dynamic Transitive Closure (DTC)* [8] problem asks to maintain reachability information in a directed graph $D = (V, E)$ between arbitrary pairs of vertices under insertions of edges. Conditional lower bound on DTC follows from OMV hypothesis and reduction from it [10]: there is no algorithm with total update time $O((mn)^{1-\epsilon})$ ($n = |V|, m = |E|$) even with $poly(n)$ time preprocessing of the input graph and $m^{\delta-\epsilon}$ query time per query for any $\delta \in (0, 1/2]$ such that $m = \Theta(n^{1/(1-\delta)})$ under OMV hypothesis.

3.2 Existing reductions

First of all we need to mention the interconvertibility of CFL reachability problems and a class of set-constraint problems [14] 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. It was firstly proved by Bradford [2] via algebraic matrix encoding and then

combinatorially by Mathiasen and Pavlogiannis [13] by combining Dyck-1 path from bell-shaped paths. For this and the following reductions see Fig. 1.

In the recent paper of Shemetova et al. [20] the reduction from all-pairs CFL reachability to incremental DTC have been proven. Still this reduction cannot give truly subcubic algorithm for CFL reachability without refuting OMV conjecture [10, 23].

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

The CFL reachability problem has been shown to be 2NPDA-complete [9]. 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 [5] and combinatorial reduction from BMM to CFG Recogniiton [12].

Recently it was discovered by Chistikov et al. [7] 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 that give lower bound stronger than $O(n^{\omega})$.

3.3 Open problems

The equivalence under subcubic reduction of *s-t* and all-pairs CFL reachability is not yet discovered. However the analogous result for triangle detection in a graph is true [24] and, perhaps, similar techniques are applicable in our area.

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 [13] 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 [24] as APSP is connected to problems on paths, and 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 intuition for possible ways of their solution.

To summarize, 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, e.g. from APSP or OMV problems.

Getting cubic conditional lower bound through some hypothesis is a possible way of future work as is closing other open problems. In our overview, we have reached the DTC problem, which lies in a field of dynamic problems, alongside with many other problems and reductions of our interest. In the future, we plan to investigate these directions.

5 ACKNOWLEDGMENTS

The research was supported by the Russian Science Foundation, grant №18-11-00100.

REFERENCES

[1] Amir Abboud, Arturs Backurs, and Virginia Vassilevska Williams. 2018. If the current clique algorithms are optimal, so is Valiant's parser. *SIAM J. Comput.* 47, 6 (2018), 2527–2555.

[2] 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.

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

[4] 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

[5] 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

[6] 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

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

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

[9] 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.

[10] Monika Henzinger, Sebastian Krinninger, Danupon Nanongkai, and Thatchaphol Saranurak. 2015. Unifying and Strengthening Hardness for Dynamic Problems via the Online Matrix-Vector Multiplication Conjecture. In *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing* (Portland, Oregon, USA) *(STOC '15)*. Association for Computing Machinery, New York, NY, USA, 21–30. https://doi.org/10.1145/2746539.2746609

[11] 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.

[12] 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

[13] 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

[14] 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

[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. 1998. Program analysis via graph reachability1An abbreviated version of this paper appeared as an invited paper in the Proceedings of the 1997 International Symposium on Logic Programming [84].1. *Information and Software Technology* 40, 11 (1998), 701–726. https://doi.org/10.1016/S0950-5849(98)00093-7

[17] 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

[18] 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.

[19] Petteri Sevon and Lauri Eronen. 2008. Subgraph Queries by Context-free Grammars. *Journal of Integrative Bioinformatics* 5, 2 (2008), 157 – 172. https://doi.org/10.1515/jib-2008-100

[20] Ekaterina Shemetova, Rustam Azimov, Egor Orachev, Ilya Epelbaum, and Semyon Grigorev. 2021. One Algorithm to Evaluate Them All: Unified Linear Algebra Based Approach to Evaluate Both Regular and Context-Free Path Queries. arXiv:cs.DB/2103.14688

[21] 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

[22] 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

[23] Jan van den Brand, Danupon Nanongkai, and Thatchaphol Saranurak. 2019. Dynamic Matrix Inverse: Improved Algorithms and Matching Conditional Lower Bounds. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*. 456–480. https://doi.org/10.1109/FOCS.2019.00036

[24] 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

[25] 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