Automated Theorem Proving for the Two-Variable Fragment in the First Order Logic

Author: Patrick Cătălin Alexandru Sava

Supervisor: Dr. Ian Pratt-Hartmann

Abstract

First-order logic has few fragments which are very interesting in terms of decidability. One of them is the two-variable fragment with no function symbols. Even though the two-variable fragment could be very rich in terms of potential discoveries, there are few papers written in this regard by now. One of them is Hans de Nivelle and Ian Pratt-Hartmann’s paper of 2001 which, by no coincidence, represents namely the goal of the project. Being the first of its kind, this paper describes a decision procedure for the two-variable fragment with equality (prior works describe only the decision procedure for the case without equality). Implementing this algorithm embodies the project’s aim, nonetheless, the roadmap for achieving this objective involves moreover the implementation of a theorem prover for the general case.

implement decision procedure of HDV and IPH paper

Potential new abstract:

The two-variable fragment with no function symbols is a fragment of first-order logic, known as being decidable. This report presents the implementation of the decision procedure for the two-variable fragment with equality. This procedure was firstly described in Hans de Nivelle and Ian Pratt-Hartmann’s paper of 2001. The implementation is presented in the form of a theorem prover. Finally, the implementation will be compared in terms of correctness and implementation against other similar tools, specifically Vampire.

Acknowledgements

The project would not have been possible without the continuous help and support of my supervisor, Dr. Ian Pratt-Hartmann. I am grateful for the experience of working with him on such a challenging and rewarding project. Thus, I would like to extend my appreciation for the excellent advice I was receiving from him throughout the entire project.

Further, I would like to thank the programming community, my teachers and friends who facilitated my self-development, knowledge and capabilities.

Finally, I would like to thank my family (especially my mother, my grandmother and my partner, Yoana) for all of their support and love during all of these years.

[**Introduction**](#_4mskwzd6nvm5) **5**

[Background](#_2szn5g88nwaf) 5

[Motivation](#_8k4aygsvrx9q) 5

[Project aim](#_ax4updv1mq93) 5

[Project Roadmap](#_r2ujw6ayztbf) 6

[Methodology](#_oizawpbvlrdq) 6

[Report Structure](#_2oarfns66zbh) 6

[Impact of Covid-19](#_xwhv18ieh44j) 7

[**Context**](#_9zd57b4duhd5) **7**

[First-Order Logic](#_dy6konftaet3) 7

[Definitions](#_djdrgmdkxmng) 7

[Operators](#_2vspvqs2p7le) 8

[Special case: equality](#_wnin7kxxh15l) 8

[Quantifiers](#_fi1ejy2x2bsw) 8

[Precedence](#_dqleivz3fmzk) 8

[Model](#_b2ymtzfccxmi) 9

[Validity and Satisfiability](#_fmzc8s7c2bok) 9

[Soundness and Completeness](#_tu40xscopkem) 9

[Decidability](#_4f0rsqujbia0) 9

[Clausal Normal Form (CNF)](#_r0vwamua472s) 10

[Basic Reduction](#_jj13vywtq8bx) 10

[Skolemization](#_soolxzd7jy5g) 10

[Automated Theorem Proving](#_f3q4g59ixn98) 11

[Unification](#_c8vaumr6n4pg) 11

[Resolution rule](#_aqodhlrppqkq) 12

[Factoring rule](#_4p8p51x1sbok) 12

[Refinements](#_ifehuyerqdp6) 13

[Tautology Removal](#_2uipkhcq6lh9) 13

[Subsumption](#_1tbv1tqm568p) 13

[Unification within the same clause, on literals having the same sign](#_77q9rj31lmho) 13

[Depth-ordered resolution](#_e7rwvwdaw2yp) 13

[Two-Variable fragment](#_gioufz95cz1z) 14

[Without equality](#_xq9n8zjme282) 14

[With equality](#_8lahq96r0pgp) 15

[**Design**](#_xxpev9n2lwjw) **17**

[**Implementation**](#_gfjtingyzhcz) **18**

[Environment and technology stack](#_l1vcfc3nbwqf) 18

[Front-end](#_d5si54q21z42) 19

[Tokenizer](#_cn3wt633rot0) 19

[Parse Tree](#_l6i6odkebjow) 20

[Intermediate Representation](#_e3c2dgf365go) 20

[The approach for disambiguating the given formulas](#_vrbua8pgemx8) 21

[Basic reduction](#_xa71vjnr2u09) 21

[The approach for resolving the precedence](#_oja1m8ihg9eo) 21

[The approach for eliminating double Implications](#_ebd5oqbkz1ac) 22

[The approach for eliminating implications](#_hfjgzsolr963) 24

[The approach for pushing the operator not](#_o5k49jl3p72w) 24

[Skolemization](#_bxgow04gybs6) 25

[Simplified Clausal Normal Form](#_abdm34xrxnfz) 26

[Back-end](#_61zrhg2p0mft) 28

[Clausal Normal Form and Unification](#_apo4ra7ygpa6) 28

[Term Unification](#_vmh3sd9x2znq) 28

[Literal Unification](#_p6mijx9wvw29) 30

[Clause Unification](#_vrcfqfte9hhj) 30

[Basic Theorem Prover](#_5shpuwpg832v) 32

[Factoring Rule](#_gjbi0aoldqn4) 32

[Subsumption Rule](#_c7rxg2pyk711) 34

[Resolution Rule and Theorem Proving](#_qwoqx2j13lw2) 34

[Depth-Ordered Theorem Prover](#_kqflx93o42gg) 35

[Two-Variable Theorem Prover](#_2qdj0scw4dg3) 36

[Variable renaming](#_4f1p2ldfm34v) 36

[Without Equality](#_wmmcmnyjbk25) 37

[With Equality](#_e5cregib9bw3) 38

[**Evaluation**](#_wb83b0x8anmn) **39**

[Unit testing](#_xmp6tw7l5j9v) 39

[End-to-End testing](#_b1rwdin7qwlu) 39

[Using the problems found on tptp.org](#_mlszyojem44) 39

[Using general formulas](#_2bj5s1nzowqv) 39

[Testing against Vampire](#_gs6mdb6o2163) 39

[**Experiments**](#_w8eibc75csft) **39**

[**Reflection**](#_vct7big78d3b) **39**

[**Conclusion**](#_l2sgo3qyqxj8) **39**

[**Appendix 1**](#_witszixm8988) **39**

[**References**](#_j1y5fgxid3et) **39**

# 

# Introduction

## Background

First-order logic is very rich in terms of expressivity. Currently the world is aware of its presence in many fields such as Mathematics, Philosophy or Computer Science. In Computer Science specifically, it could be present as a part of sophisticated solutions in the subfield of databases or the one of natural language processing.

In comparison with propositional logic, first-order logic is more powerful due to the predicates and the quantifiers, also being likely to the fact that it can operate on larger domains than just the boolean one. One could argue that this is the reason why it can be used in the field of Computer Science (for example, to prove that a programming language has a sound type system). Symmetrically, higher-order logic has a greater power than first-order logic, but it does not provide a sound and complete proof calculus. [reference <https://en.wikipedia.org/wiki/Higher-order_logic>]

In essence, first-order logic lies at the intersection of high expressiveness and the combination of soundness and completeness and this is one of the reasons why it is valued by the research world.

## Motivation

During the course COMP24412 I became particularly attracted to using Prolog and Vampire. I have then realised that this is an area of research which I have not had the chance to explore well enough so I was looking for a third-year project liying in this sub-field.

At the same time, I was looking to grasp my knowledge in the sub-field of compilers, which the implementation of a theorem-prover was able to offer to me since I strongly preferred to write the front-end of it by myself.

Ultimately, the curiosity of seeing how well I could implement a piece of software which is going to compete directly with an absolut champion like Vampire, encouraged me to accept this challenge.

## Project aim

The main goal of the project was to implement the first ever decision procedure for the two-variable fragment with equality of the first order-logic which was priorly described in a paper written by Hans de Nivelle and Ian Pratt-Hartmann (2001), published in Gore et al’s “Automated Reasoning” (2001), but for which there is no evidence up to this point that it was actually implemented by someone else. Besides, an additional goal of the project was to compare the algorithm with Vampire and to produce some satisfiability charts in the case in which the input is a randomly generated formula.

## Project Roadmap

The project roadmap was well-defined:

* Semester one
  + The first six weeks (including week 0) were dedicated to familiarising myself more with the topic, implying mostly reading papers, books and additional materials
  + The next three weeks (up to week 8) were mostly invested in writing the parser for the formulas
  + The last weeks up to the Christmas holiday were spent coding a general theorem prover
  + The Christmas holiday and the exam session were dedicated to testing the existing functionality, coding the depth-ordered theorem prover and the two-variable theorem prover for the case without equality
* Semester two
  + The first three weeks were spent on reading some papers, coding the two-variable theorem prover for the case with equality, optimizing, testing and debugging the code
  + The next three to four weeks were mostly invested in optimizing, testing and debugging the code. On top of that, two random generators for formulas and some scripts for comparing my work against Vampire were written.

## Methodology

One of the most important aspects of the project was definitely time management. Coordinating with my supervisor, while both researching and implementing the theorem prover was challenging and required good prioritization abilities. My supervisor suggested keeping track of everything I am doing in a log book, which I found to be very useful, not only for writing this report, but also for organizing my thoughts and ideas better. Even though I did not feel it as being helpful at the beginning, this idea proved to be brilliant in time, especially when the size of the codebase for the project hit 5000 lines of code and when even remembering what purpose my own lines of code were serving was problematic.

## Report Structure

The structure of the report is chosen in a way in which the reader is presented with the context first, then with the high-level idea of the project and further with details of smaller granularity. Hence, the following section is reserved for reiterating the basic notions of first-order logic and the formal tools which are going to be part of the algorithm. The section afterwards is dedicated to the design decisions, including, but not limited to, the high-level presentation of the two-variable theorem prover. Afterwards, the report introduces the section assigned for the implementation which indeed underlines the implementation’s low-level details. Following this, the report presents methods of evaluating the project, subsequently discussing the experiments and reflections. The final section concludes the report.

## Impact of Covid-19

Fortunately, I am in a position in which I could say that the whole situation caused by Covid-19 did not create any major disruptions for me, and that’s mostly because my project was completely independent of any University equipment. The communication with my supervisor was excellent and I feel grateful that the overall quality of the project was not affected by this matter.

# Context

For the following sections I will assume that the reader is familiar with the first-order logic, even though I will perform a high-level iteration on all of the aspects which are of interest to my project.

## First-Order Logic

As mentioned above, first-order logic has a stronger expressiveness than the propositional logic due to the presence of quantifiers and more verbose predicates, in the sense that those predicates could have arguments, which could be constants, variables or even functions. I will define these in the next subsection.

From now on, I will be referring only to first-order logic, so it can be assumed that any reference to logic from this report refers to first-order logic.

### Definitions

The highest level of granularity I will be using throughout this report will be the logic formula, which could be any string accepted by the grammar of the logic described in Appendix 1.

**Definition 1.** A predicate is a n-ary symbol containing n terms as arguments. Conversely the arity of a predicate is the number of arguments.

**Definition 2**. The sign of a predicate is positive if it is not negated and negative otherwise.

**Definition 3**. An atomic formula or an atom, is a formula containing only a positive predicate.

**Definition 4**. A formula containing only a positive predicate or only a negative predicate is called literal.

**Definition 5.** A term is either a function, a variable or a constant.

**Definition 6.** A function is a fixed (or pre-defined) mapping from a variable or a constant to a variable or a constant.

**Definition 7**. A variable is a mapping from a variable to another variable (including itself) or to a constant.

**Definition 8.** A constant is a mapping exclusively to itself.

Due to the fact that function symbols break the decidability (the reason why this holds is explained in the later subsections), then the input will be restricted so as to not contain such terms.

### Operators

With the risk of repeating the basics, the operators in which are of interest are: double implication (⇔), implication (⇒), negation (ㄱ), and (∧), or (∨), equality (=) and finally inequality (!=).

Round brackets are considered as well part of the operators, since they improve the clarity in the majority of cases, especially when used together with quantifiers.

Sometimes, the operator and is referred to as conjunction and the operator or as disjunction.

#### Special case: equality

The reader could assume that up to the point when the two-variable theorem prover algorithm is presented, both the equality and inequality are not part of our language. The reason why the report will proceed this way is mostly because dealing with equality (and inequality) in the general case (that is, outside the two-variable fragment) is out of the scope of the project, involving different techniques such as paramodulation.

### Quantifiers

The quantifiers with which the report deals are the following:

* for all (∀)
* there exists (∃)
* there exists an unique (∃!)

The uniqueness quantifier will not be of interest up to the last subsection of this section. Moreover, the grammar will not contain this symbol and it will be encountered only at an abstract level, when the report will be dealing with equality.

### Precedence

In terms of the operator precedence the following conventions are adopted which should be the same as the ones of Vampire (this is assumed for convenience, to be easier to generate random formulas which imply the same thing for both Vampire and our theorem prover). These conventions are the following:

* the round brackets enforce precedence in the same way they do in the majority of mathematical systems
* the operator not and both the existential and universal quantifiers are applying strictly to the next:
  + predicate, if they precede one
  + subformula enclosed in a pair of round brackets, if they precede the corresponding open bracket
* the operator and has a strictly higher precedence than the operators corresponding to or, implication and double implication
  + this is analogous for the operator or and the (double) implications, and for the implication and the double implication

A very important aspect which should be taken into consideration as well is the fact that the implication is right associative.

### Model

**Definition 9.** The domain of discourse is the set of entities over which the variables from a first-order logic formula are ranging. [reference <https://en.wikipedia.org/wiki/Domain_of_discourse>]

**Definition 10.** A model (or an interpretation) of a formula refers to a mapping from each variable from the formula to an element of the domain of discourse. [reference <https://en.wikipedia.org/wiki/First-order_logic>]

### Validity and Satisfiability

**Definition 11.** A formula is valid iff it is true in all of the possible models. Conversely, a formula is invalid if and only if it is false in all models (Wilfrid, 1977, p. 242).

**Definition 12.** A formula is satisfiable iff there exists at least one model in which it is true. Conversely, a formula is unsatisfiable if and only if it is false in all models [reference wilfried book?]

As a consequence, a formula is invalid iff it is unsatisfiable. Moreover a formula is satisfiable iff its negation is invalid.

### Soundness and Completeness

Let the knowledge base be the set of axioms initially existing. This could be seen also as a set of true hypotheses or formulas.

As true things are proved using the knowledge base, these will be added to it.

Some deductive systems of first-order logic are sound and complete [reference https://en.wikipedia.org/wiki/First-order\_logic]. The intuition behind soundness is that only true things can be inferred from the knowledge base (as an analogy one could think to the definition of the implication, that a true fact can imply only another true one) while the one behind completeness is that all of the true things which could be possibly derived from the knowledge base will be eventually being reached. (as an analogy one could imagine a graph containing initially one node for each axiom, in which nodes and edges are getting dynamically added; when adding a node containing a brand new fact (derived using only the existing nodes of the graph at that point), undirected edges from the nodes corresponding to the facts which inferred the latter are added additionally; when time tends to infinity, the graph will contain all of the facts which could possibly be derived from the knowledge base)

### Decidability

**Definition 13.** A decision problem is a problem which could be reduced to a yes-no question.

Decidability in logic refers to the existence of an effective method for approaching a decision problem [reference <https://en.wikipedia.org/wiki/Decidability_(logic)>].

For example, the satisfiability problem of first-order logic is undecidable [reference <https://en.wikipedia.org/wiki/First-order_logic>], whereas the same problem for the two-variable fragment without function symbols, but with the existence of uniqueness quantifiers, is decidable [reference <https://en.wikipedia.org/wiki/Two-variable_logic>]. Consequently, the two-variable logic has to be sound and complete.

## Clausal Normal Form (CNF)

For approaching the automated theorem proving, a more convenient form to deal with the formulas is needed. Firstly, the input format has to be underlined. It is assumed that the input contains N formulas, the first N - 1 of them forming the knowledge base, and the N-th one representing the query (that is, the formula of which satisfiability has to be verified assuming that all of the formulas from the knowledge base are axioms).

Secondly, it would be better to reduce everything to a single formula. This could be done easily by negating the N-th formula, wrapping each formula in a pair of round brackets and then adding N - 1 conjunctions between the N formulas. The resulting formula will be acknowledged thereafter as F.

As described in Leitsch (1997, p.12-18), F could be transformed into a sub-language of logic called conjunctive normal form (sometimes being referred to as clausal normal form). The final form will be a conjunctions of disjunctions, in which no quantifiers are present and will be referred to as C. Even though F might not have contained functions symbols, C could contain and this is explained in the next subsections.

### Basic Reduction

We will repeat the following seven steps until the formula does not change anymore:

* break double implications in implications
* break implications in disjunctions
* push the operator not further on conjunctions
* push the operator not further on disjunctions
* simplify two consecutive not operators to nothing
* push the operator not past universal quantifiers
* push the operator not past existential quantifiers

Let F’ be the resulting formula after all of the steps above have been exhausted. It is obvious that F’ contains only quantifiers, disjunctions, conjunctions, brackets and literals.

### Skolemization

The name of this procedure comes after the Norwegian mathematician Thoralf Skolem [reference <https://en.wikipedia.org/wiki/Skolem_normal_form>]. The intuition behind this procedure is that the existentially quantified variables could be always seen as the result of a brand-new function which has as arguments all of the universally quantified variables having in their scope the former. By “scope” the report uses the same meaning as the one from programming. (i.e. one could make an analogy with the scope of local variables in C++). Obviously, if there are no such universally quantified variables, our function becomes a constant, precisely a Skolem constant. Otherwise, it remains a function, precisely a Skolem function. Hence, all of the occurrences of existentially quantified variables will be replaced with Skolem functions or Skolem constants. Afterwards, all of the quantifiers (both existential and universal) could be safely disposed of. The reason for being able to do so is because, from now on, the terms on sub-types (functions, variables, constants) will be differentiated by keeping track of that while doing the Skolemization. It could be easily seen that the quantifiers are part of redundancy now.

For the sake of simplicity it could be assumed for any variable x in F’ that all of the variables which are in the scope of x have unique names. (i.e. a substring like ∀x∀y∀x or one like ∀x∃y∃y would never be present in F’). The approach for resolving this issue will be further described in the implementation details.

Let F’’ be the resulting formula. The only thing which has to be done to transform F’’ in C is to apply the distributivity laws over conjunctions and disjunctions in order to end up with C. Having in mind the preparation for the next step, C could be transformed into a set of disjunctions (or clauses) since it is uniquely determined which operator stays between them (the operator and). Hence, the name of clause form is probably better justified now.

Even though here explicitly the distributivity laws are used, the way they are going to be implemented is going to differ from the traditional way. This is because implementing the distributivity laws naively could lead to an exponential growth of the size of the formula.

## Automated Theorem Proving

In this subsection the report is going to present the core ideas and tools behind the automated theorem proving. All of these are part of the codebase of the project.

### Unification

The key of the automated theorem proving is in the ability of the software to detect whether two literals are the same if the sign would be ignored. In other words, they are the same predicate-wise, term-wise and order-wise. (i.e. P(f(x)) and ㄱP(f(x)) respect that while P(f(x)) and P(f(y)) does not).

**Definition 14.** We call substitution a mapping from the set of variables to the set of terms. Moreover we could compose two substitutions in the same way we could compose two functions.

**Definition 15.** An invalid substitution is one in which there exists a mapping from a variable to a term containing the same variable (and by that it is meant that it could be either itself or a function having the variable nested in one of its arguments). A valid substitution is obviously a substitution which is not invalid.

**Definition 16.** Two terms are unifiable if there exists a valid substitution such that after its appliance they become the same.

**Definition 17.** Two n-ary predicates P(t1, t2, …, tn) and Q(r1, r2, …, rn) are unifiable if P and Q are identical and ti and ri are unifiable for 1 ≤ i ≤ n.

**Definition 18.** Two literals are unifiable if their corresponding predicates (letting aside the sign) are unifiable.

**Definition 19.** Let P(t1, t2, …, tn) and Q(r1, r2, …, rn) be two n-ary unifiable predicates. Let S1, S2, …, Sn be the valid substitutions for the terms ti and ri  with 1 ≤ i ≤ n. We will call S the result of the composition of S1, S2, …, Sn. If the appliance of S results unifies P and Q (that is, makes P and Q identical) then S is the most general unifier of P and Q. Note that we could refer to most general unifiers for literals in the same way we do for predicates since the signs do not matter.

Having explained unification, the two core rules for theorem proving: resolution and factoring can be presented.

### Resolution rule

**Notation 1.** Let C be a clause and S a substitution. Then {C}S is the resulting clause by applying the substitution S to all of the literals in C. Similarly we can define the same on a literal L or on an atom A (i.e. that would result in {L}S and {A}S, respectively)

**Notation 2.** Let C be a clause and L be a literal which occurs in C. Then C\L is the resulting clause obtained by removing L from C exactly once.

**Definition 20.** Let C1 and C2 be two clauses having in common two unifiable literals L1 (part of C1) and L2 (part of C2) of opposite signs. Let S be the most general unifier of L1 and L2. Then C1 ∧ C2 ⇒ {C1\L2 ∧ C2\L2}S. {C1\L2 ∧ C2\L2}S is often referred as being the resolvent of C1 and C2, while the atom A which is common to L1 and L2 is referred as the resolved atom.

### Factoring rule

**Definition 21.** Let C be a clause and L be a literal occuring more than once in C. If L occurs k times in C, then we could remove k-1 occurences of it from C.

Once the resolution and factoring rules have been implemented, it can be safely claimed that we have a basic theorem prover. Since the satisfiability problem for first-order logic is undecidable, our program could never terminate. If it terminates, this means that either it reached saturation (this is the case when there is no new clause inferred by the current set of clauses which has not been seen priorly) or it derived the empty clause. The former implies that the C was counter satisfiable while the latter implies the satisfiability.

From now on the report will discuss a few refinements which could speed up the basic theorem prover and ultimately bring us closer to understanding the design of the theorem prover for the two-variable fragment.

### Refinements

#### Tautology Removal

**Definition 22.** Let C be a clause containing two literals L1 and L2 which are identical except their signs (which are opposite). This clause is obviously true in any model, so it can be dropped from the current set of clauses.

#### Subsumption

**Definition 23.** Let C1 and C2 be two clauses such that the set of literals of C1 is strictly included in the set of literals of C2. Then C1 could be seen as a “more general” clause than C2. Therefore C2 could be dropped since it is subsumed by C1.

This strategy makes sense because at any point in time it is assumed that all of the clauses from the current set of clauses are axioms. So the intuition behind subsumption is that C2 is no longer needed in the set since C1 was anyway implying it and since C1 has to be true, then C2 will certainly be true.

Note that “strictly” was used in the definition because if C1 and C2 are identical, it would be anyway preferred not to add whichever came last in the set of clauses.

#### Unification within the same clause, on literals having the same sign

This is a heuristic I created during the implementation process. It looked promising, resulting in speeding up a few tests I was testing against my implementation. Intuitively having a clause C and two unifiable literals L1 and L2 which are part of the clause, the literals could be unified and the factoring rule could be applied. This way a less general clause is obtained, which is going to have a shorter size than the former.

#### Depth-ordered resolution

**Definition 24.** A ground term is a term which does not contain variables. In other words, it is a constant, a function symbol or a function symbol having nested only function symbols or/and constants.

**Definition 25.** The depth of a variable and the depth of a ground term are both 0.

**Definition 26.** The depth of a function which is not a ground term is the maximum number of nested function symbols up to a variable. (i.e. f(x, g(y)) has depth 2, f(x, y) has depth 1 and f(f(x, g(y)) has depth 3).

**Definition 27.** The depth of an atom is the maximum depth of its terms (i.e. P(f(x, g(y)), f(x, y)) has depth 2, P(f(x, y)) has depth 1 and P(f(x, y), f(f(x, g(y))) has depth 3). Similarly we could define the depth of a literal as the depth of its corresponding atom, since the sign does not have any role in the calculation of it.

As described before, the satisfiability problem for first-order logic is undecidable. From that we could deduce that the basic theorem prover we have discussed might not halt in some cases. One of the reasons for which that would happen is the unbounded nesting of function symbols. (i.e. we don’t currently have any mechanism to avoid expansions like P(f(f(f(f(f(….)))))) to happen)

As described in Leitsch (1997, p. 99), a binary relation <A could be defined on the set of all atoms which is having the following properties:

* is reflexive
* is transitive
* for any two atoms A and B and a substitution S, A <A B implies {A}S <A {B}S

**Definition 28.** The depth-ordered theorem prover will be working in the same was as the basic theorem prover, with the restriction that we will apply the resolution rule on two clauses C1 and C2 only when there is no literal in L in the resolvent C (of C1 and C2) such that B <A L, where B is the resolved atom of the resolution (Leitsch, 1997, p. 99).

The depth-ordered resolution is sometimes referred to as ordered resolution and it is proved as being complete (Joyner Jr. et al., 2002).

## Two-Variable fragment

Moreover, the two-variable fragment theorem prover will be discussed. Since this section describes the algorithm at a theoretical level, it could be assumed that wherever references are not in place and the difficulty of (sub-)topics presented exceed the basics of logic, the credits for them are owned by my supervisor. I mention this because even though I have found the same notions in academic papers or books, I understood best the principles of these notions during the tutorial classes.

For the rest of the next two subsections it will be assumed that a logic formula is given in two variables which is going to be passed further as input to the two-variable theorem prover. Moreover, this formula is in the CNF format.

For the sake of simplicity it could be additionally assumed that the set of variables occurring in our formula is {x, y}.

### Without equality

The report will start by adding a restriction to the logic formula: for the rest of this subsection, it will be assumed that the logic formula does not contain equality or inequality. Following that, the algorithm will be revisited for the case with equality in the next subsection.

With the risk of breaking the consistency, this time, the report will be labeling the steps of the algorithm, in order to make it easier to revisit for the case with equality:

1. We are going to execute the depth-ordered resolution on our formula, only on the literals involving exactly two variables. Having said that, the resolved atom will always have exactly two different variables.
   1. If the empty clause is derived, this means that our formula is satisfiable.
   2. Otherwise, the algorithm should halt by saturation, since the ordered resolution is used.
2. If the algorithm did saturate then we will be left with a set of clauses. From them, all of the clauses containing literals in exactly two variables could be safely disposed of.
3. Now we are left with a set of clauses with the following property: if we would group each clause on the set of distinct variables it is having, the maximum number of groups we would get per clause would be three. This is because, in the worst case, the group corresponding to following the will exist:
   1. empty set (i.e. all of the literals in this group do not hold variables; they are ground instances as they are sometimes referred)
   2. the set {x}
   3. the set {y}

This whole procedure is referred to in De Nivelle and Pratt-Hartmann (2001) as the splitting rule and it could be applied to a clause which does not have overlapping variables, which is the case for ours.

1. Let C be our set of clauses and |C| be the size of it. Further, |C| group choices will be made, one for each clause. In other words, at most 3|C| auxiliary sets of clauses will be built containing exactly one group from each clause. Further, the depth-ordered theorem prover will be run on each of these auxiliary sets.
   1. If all of these runs will derive the empty clause, this means that the input formula is unsatisfiable.
   2. Otherwise, that would imply that the depth-ordered theorem prover halted by saturation on at least one of these auxiliary sets. Hence, the input formula is satisfiable.

### With equality

Before revisiting the algorithm above, the report will detail how to deal with:

* the equality
  + dealing with the equality operator is trivial in our case, since the equality can be simply replaced operator with a special predicate of arity two called “Equality”
  + in case the equality operator occurs multiple times, all of these occurrences will be replaced by the same predicate called “Equality”
* the inequality
  + dealing with the equality could be seen as a fairly similar process as dealing with equality with the consideration that the inequality operator will be replaced with a brand-new predicate of arity two called “Inequality”
    - as in the case of equality, different occurrences of the inequality will be mapped to the same predicate called “Inequality”
  + in order to preserve the logical equivalence, a new clause in one variable has to be added, containing the negated “Inequality” predicate to our CNF (i.e.ㄱInequality(x, x) has to be true)

We could treat the “Inequality” like any other predicate in our set of clauses from now on, but the report will further expand the “Equality” in our algorithm.

After the report has established how to deal with these two operators, the algorithm can be revisited by mentioning the changes that need to be make:

* Step 1 will now be doing resolution on literals involving exactly two variables, which are not the “Equality”.
* Step 2 will now dispose of all clauses containing literals different from the “Equality” which are in exactly two variables.
* Before proceeding to step 3, the equality has to be expanded in the same way as it is described in Lemma 5 and Lemma 6 in De Nivelle and Pratt-Hartmann (2001). The report will present the high-level idea of this expansion letting the group of the ground instances aside:
  + Let R an arbitrary clause in two variables from our set of clauses, containing equality. We will define T(x) as being all of the subformulas in variable x from R and U(y) being all of the subformulas in variable y from R. Hence, we can write R the following way:

The aforementioned formula is logically equivalent with the following:

For the following formula

there exists as well the following logical equivalence

The problem which arises with the logical equivalence from above is that our algorithm will enter into an infinite loop, since each attempt of replacing the equality would result in a new formula containing equality again.

The solution for this case is explained as part of Lemma 6 in De Nivelle and Pratt-Hartmann (2001) and it involves the introduction of a brand new constant for each clause containing equality. The construction presented in the paper takes into account unary and binary predicates.

However, if the point in which the algorithm is currently taken into consideration, the binary predicates can be safely omitted, since we cannot have any. Our formula will have after the disposal from step 2 only literals in one variable, which could be safely seen as predicates of arity one. (even though this has to be done with caution, since, for example, P(constant, x) and P(x, constant) are representing different predicates of arity one, even though the share the same predicate name)

Combining the observation above with the Lemma 6 from De Nivelle and Pratt-Hartmann (2001), we have the following:

and

are equisatisfiable, where e is a brand-new constant symbol.

* The only thing left to do before safely proceeding to step 3 would be to execute once more the clausal normal form algorithm on the resulting equisatisfiable formula, since this formula is not anymore a valid CNF.

The two subsections from above conclude the second section of the report, which was dedicated to the theoretical part of the project. The next section will present the big picture of the algorithms and how they fit together.

# Design

The codebase has three main components: the front-end, the intermediate representation and the back-end.

The front-end is responsible for parsing the input, reporting any syntactic and some semantic errors. Furthermore, it also builds the parse tree of the input formula. With the risk of creating confusions, the report would not be following precisely the definitions from the compiler terminologies, our parse tree being a hybrid between what a parse tree and an abstract syntax tree are. The reason for that is because all of the operations up to the point in which our formula is in CNF are performed on this tree, which changes it dynamically. In terms of algorithms, this component uses mostly ad-hoc algorithms (which will be further described) or depth-first search. This component is executing all of the operations in polynomial time.

The intermediate representation is mostly dealing with the parse tree, implementing all of the operations necessary for building the clausal normal form efficiently. In terms of the implementation, the reader will probably notice further that many of the methods implemented for reducing the formula are making assumptions on the structure of the input instead of enforcing that structure (this is because enforcing that would imply more work in terms of the resources than the respective method is supposed to do). The majority of the functionality, if not all, executes the reducing operations using a depth-first search. In the same way as the front-end, all of the operations are executed in polynomial time.

The back-end is probably the biggest part of the codebase, holding the implementation for the unification, basic theorem prover, depth-ordered theorem prover and the two-variable theorem prover. Moreover, the majority of work for the theorem prover is parallelized using multithreading. Implementation-wise, this component makes the most usage of both ad-hoc algorithms and classical algorithms (breadth-first search, depth-first search, backtracking, 2-colorability). Moreover, sophisticated data structures (such as persistent data structures) are as well involved. In terms of time complexity, the majority of the operations are happening in polynomial time, excluding the part involving the two-variable fragment which happens in exponential time, due to the nature of the Cartesian product which is implemented as a backtracking algorithm.

In terms of the language, a few changes have been made since the logic symbols are not present on (probably) none of the existing keyboard layouts. Hence, we will have the following mapping: double implication (<->), implication (->), negation (~), and (^), or (|), equality (=), inequality (!=), universal quantifier (@) and finally the existential quantifier (?). On the top of that, the following conventions are adopted:

* the terms and predicates will always contain letters of the Latin alphabet only
* the predicates will always start with uppercase letter, containing further only lower case letters
* all of the terms will always contain lowercase letters only

As mentioned before, the report will assume that the input formula does not contain function symbols, even though the CNF of it could end up having (due to the Skolemization).

Below it is presented a diagram which gives an idea about how the three aforementioned components interact with each other. The names of the subcomponents are also revealed, in order to offer a hint to the reader in regards to what will follow in the next section.

{insert image here}

# Implementation

This section will be dedicated fully to the implementation details of the algorithms discussed during the course of the previous sections.

## Environment and technology stack

As an operating system, both Windows 10 Pro and Ubuntu 20.04 were used. Ubuntu was actually run as a Windows subsystem, enabling the programmer to utilise both the Linux infrastructure and the Unix-like commands.

In terms of hardware, a computer having a 7th generation Intel i7 processor and 32GB of RAM memory was used. During the implementation process, the generous memory resources were quite useful, since an unoptimized version of the theorem prover was using more than 18GB of RAM memory on some input formulas which were saturating. After noticing that, I optimized the code in such a way that the amount of memory used by the theorem prover was drastically reduced (below 1GB of RAM memory on similar or the same tests).

In terms of programming languages utilized, the choice was to use C++ Standard 17 for the implementation (since I had the most experience with this programming language but also for runtime speed C++ is offering). Standard 17 was a must since the implementation makes extensive use of std::variant, which is a type safe union (CPP Reference, 2021) and of std::execution (CPP Reference, 2021) which provides the execution policies which could be used when parallelizing code. However, the latter is not fully implemented by some C++ 17 compilers so I had to import it manually by linking my code with the libtbb library (Debian, 2021) which provides the implementation. In order to unit-test some parts of the code, the programmer used Google's C++ testing framework (Googletest, 2021). Other programming languages which have been used during the implementation of this project are Python3.8 and Unix Bash. The former was used mostly for writing input generators while the latter was used for End-to-End testing or for making comparisons before our theorem prover and Vampire.

As a building system, CMake was strongly preferred due to the powerful set of options but also due to the fact that Google's C++ testing framework was available to build the easiest for CMake-based projects.

## Front-end

This section is going to present the algorithms behind the front-end. As probably noticed, the workflow of the (sub)components typically tends to get an input and produce an output. The front-end uses exactly this mechanism: the tokenizer will be receiving the raw input and produce an ordered list of tokens; further, the parse tree will be passed this list of tokens and will produce the initial parse tree.

### Tokenizer

As described before, the purpose of the tokenizer is to take the raw input and to parse it into tokens. For better understanding, we will define what a token could be:

* an atom
  + that would be any string matching the following: it starts with uppercase letter, then it has an arbitrary number of lowercase characters (possibly zero) and it further has a pair of round brackets within a set of variables (lowercase strings having a positive length) separated by commas is nested
  + since our convention is that predicates are always starting with capital letter and there are no other entities which could possibly start in the same way, the detection of atoms is trivial: each time when a capital letter is seen, we start scanning up to the first closed bracket
* an operator
  + this is trivial as well, since each operator could be detected with one symbol lookahead
    - the only two particular cases are happening for equality and inequality: they have to be replaced with their corresponding predicates (called “Equality” or “Inequality”)
* a quantifier
  + as for operators, this would be quite easy to detect as well: the current pointer in our raw input has to point to either @ or ? and be followed further by a positive number of lowercase letters (forming the variable)

Apart from detecting the tokens above and eventually reporting errors, the tokenizer will have to potentially add manually a conjunction in the ordered list of tokens. (the conjunction which needs to be added for preserving the logical equivalence for the inequality)

### Parse Tree

Our parse tree will look like a normal tree, with the mention that some information will be stored in each node, which in our code is referred to as an entity. A node might not store any entity. For the nodes storing entities, these could encapsulate 4 types of entities:

* quantified variable (i.e. @x or ?y)
* literal (i.e. P(x) or ~R(f(x)))
* operator (and, or, not, implication, double implication; note that the tokenizer will have replaced both the equality and inequality with predicates by this point)
* normal form (could be ignored it for the moment because it will be analyzed it in the section dedicated to the intermediate representation implementation)

It could be also the right time to reveal the reasons for choosing to store our formula this way:

* for some nodes, the following property holds: the subtree rooted in each of them represents a subformula of our initial formula
* as the reader has probably noticed above, the entity which encapsulates an operator does not include the brackets; this is because the pre-order traversal on the tree can be used for representing the precedence enforced by the brackets
  + hence, our design choice already reduced drastically the initial language for the formula; we managed to get rid of equality, inequality and brackets already

Now the only thing left is to do is to present the algorithm which makes the transition from the ordered list of tokens to the parse tree. The high-level idea of the algorithm will be presented. Note that some details are intentionally omitted for the sake of simplicity:

* we will declare a stack, called fatherChain
* we will dynamically add the first node to the parse tree, which is the root
* we will push the root to fatherChain
* let tokens be the ordered list of tokens
* for each token in tokens
  + if the token is open bracket, an operator, a quantifier or a literal
    - let newNode be a new node added to the parseTree
    - add the newNode in the adjacency list of the node from the top of fatherChain
    - push newNode in the fatherChain
    - if the token is an operator, a quantifier or a literal, assign a new entity to the node newNode, encapsulating whichever type of entity it is
  + if the token is closed bracket
    - pop the top from fatherChain

This algorithm runs in O(|Tokens|) time and consumes O(|Tokens|) memory, where |Tokens| is the number of tokens.

## Intermediate Representation

We will now present the functionality of the intermediate representation, which would be expected to seem much more complex to the reader, since quite a lot of time invested in the implementation of the project was dedicated to polishing this component. This component is represented in the code under the name “Reducer”, which acts as a friend class for the class Parse Tree (which the report has just presented). The name “Reducer” is consistent with its functionality, keep reducing the parse tree until it holds the CNF in the root.

### The approach for disambiguating the given formulas

Similarly to what happens in many programming languages, it is perfectly legal to have two variables with the same name, one nested in the scope of another, being bounded by different quantifiers. This might represent a challenge for us later unless action is taken now. Hence, what will be done would be to run a depth-first search maintaining an accumulator with all of the variables which have been seen so far. If we are currently in a node holding a quantified variable, the respective variable should be added in the accumulator. In case the variable already exists in the accumulator, it has to be mapped to a brand-new variable (which does not exist in the accumulator). Furthermore, at this step a mapping has been added in order to remove it after having traversed the entire subtree rooted in this node. Moreover, while going down the traversal, any occurence of the variable which has been mapped at this step will have to be substituted with the corresponding mapping.

In order to implement this part efficiently, the accumulator was modelled as an std::unordered\_set (CPP, 2021), in order to be able to insert and search in constant time.

The time complexity of this part is O(N), where N is the number of nodes our parse tree is having by the time of traversal.

### Basic reduction

This section will be presenting all of the tools needed in order to achieve the “Basic reduction” discussed in the theoretical part of the report.

#### The approach for resolving the precedence

Another interesting and challenging aspect of our parse tree is how to deal with the precedence of the remaining operators. As already mentioned earlier, the pre-order traversal takes care of the precedence enforced by the brackets. This is our main tool of dealing with the precedence for the other operators as well.

We will discuss the first case, the one of operator and. The remaining cases could be treated analogously as the main idea does not change at all. The report will assume that we are at an arbitrary node in our parse tree, and the neighbours of our parse tree are having various types of entities (literals, operators, quantifiers, including not having any).

Before proceeding, there are few observations that have to be made about our parse tree, assuming that it represents a syntactically valid logic formula:

* the subtree rooted in a neighbour having no entity associated represents a syntactically valid logic formula
* the order in which the neighbours occur is very important, since it uniquely determines the formula
* no two consecutive neighbours could have operators as entities, because otherwise our formula would be invalid syntactically

Having these observations in mind, we could deal with the precedence by adding new nodes in the parse tree, which would be equivalent with adding parentheses. Precisely, some of the current neighbours will be replaced by brand-new nodes having into their adjacency list only the node they are replacing.

For resolving the case for the operator and, we need to take maximal groups of consecutive neighbours with the property that the only operator occurring in the multiset of their entities is the operator and. For each group, we will add all of the nodes belonging to it in the adjacency list of a brand-new node, by preserving the relative order between them. Finally, each group of nodes in the original adjacency list will be replaced by the brand-new node corresponding to the respective group.

After having done this successfully, our parse tree will have the precedence for the operator and resolved, so we could move on to the operator or. The procedure will be identical, since we will not interfere with the operator and at all. The case for implication and double implication could be resolved analogously.

Even though resolving precedence might seem complex, the way it was implemented was quite straightforward, by using four calls to a depth-first search function, one for each operator.

The time complexity of this part is O(N), where N is the number of nodes our parse tree is having by the time of traversal.

#### The approach for eliminating double Implications

After an unambiguous formula in our parse tree is reached, the next step would be to reduce it to an even simpler sub-language: one which does not contain double-implications. It is well known that the double implication could be splitted into a conjunction of two implications, but the problem which would arise if this approach is taken is that the size of our formula could grow exponentially (i.e. P(x) ⇔ (R(x) ⇔ Q(x)) will further transform in P(x) ⇔ ((R(x) ⇒ Q(x)) ∧ (Q(x) ⇒ R(x))), finally becoming (P(x) ⇒ ((R(x) ⇒ Q(x)) ∧ (Q(x) ⇒ R(x)))) ∧ ((R(x) ⇒ Q(x)) ∧ (Q(x) ⇒ R(x))) ⇒ P(x)); as it can be noticed, the initial formula has one occurence of Q, the next one has two while the final one has four).

In order to prevent this issue from happening, we will be always replacing the deepest occurence of a double implication with a brand-new predicate (for the purpose of this section the reader could assume that the replacement of a double implication consists of replacing the operator, the left-hand side and the right-hand side, respectively). In order to preserve the logical equivalence while doing that, a new conjunction has to be added to our formula, containing a double implication between the brand-new predicate introduced and the subformula it is replacing. Repeating this procedure until there is no double implication operator nested in a subformula which is further part of a double implication will finally produce a logical equivalent formula on which it is safe to apply the expansion described in the first paragraph without having to worry about any exponential blow-up.

Implementing the replacement of the double implications with predicates would be quite trivial to perform on the parse-tree:

* we visit all the nodes using a depth-first search
* whenever we detect an occurrence of the double implication, it is checked whether it is the deepest in the parse tree and if it is, proceed as follows:
  + we will take the three nodes forming it (left-hand side, double implication, right-hand side) and we will append them to the adjacency list of a brand-new parse tree node, called R; note that appending the three nodes refers to moving the entire subtrees rooted in them
  + we will create two brand-new nodes in our parse tree, both holding the same brand-new predicate as an entity; these two nodes will be referred to as T and P
  + we will now replace the three nodes forming the initial double implication with T
  + finally we will append at the end of the adjacency list of the root the nodes C, P, D, R in this order; C and D are brand-new nodes holding the operator and and the double implication, respectively; these four nodes simulate the appendance of the conjunction needed for preserving the logical equivalence
    - we omitted here some details which take care of the precedence for the sake of simplicity

The report will detail as well the algorithm of getting rid completely of double implications having performed the algorithm above by this point:

* we visit all the nodes using a depth-first search
* whenever an occurence of the double implication is detected, having the nodes L, D, R, the following is performed:
  + we create a deep copy of node L in node L’
  + we create a deep copy of node R in node R’
  + we create a brand-new node I holding an implication as an entity
  + we change the entity held on D to an implication
  + we create a brand-new node C holding the operator and
  + we will now replace the nodes L, D, R with the nodes L, D, R, C, R’, I, L’ in this order
    - we omitted here some details which take care of the precedence for the sake of simplicity

Because there is a polynomial number (in respect with the size of the formula) of double implications in the parse tree, the complexity of performing this transformation is polynomially bounded. In fact, if N is the number of nodes in our parse tree, it would be expected to quadruple in the worst case. (it would double for performing the optimization for replacing the double implications with brand-new predicates and it would double further once more for performing the expansion)

Hence, the time complexity of this reduction is O(N), with N defined above.

#### The approach for eliminating implications

In the previous section the report has shown how to eliminate the double implications. Hence, the only operator left which has to be eliminated is the implication. Due to its nature, the same challenges will not be faced for the double implication. The reason why this holds is because the implication from left-hand side to the right-hand side is equivalent with the disjunction between the left-hand side negated and the right-hand side, which does not double the number of occurrences of any predicate which is part of any of the sides.

The algorithm is therefore straightforward:

* we visit all the nodes using depth-first search
* whenever we find for a node Q three consecutive neighbours matching the pattern left-hand side (referred as L), implication (referred as I), right-hand side (referred as R) proceed as follows:
  + create a brand-new node K, holding the operator not as an entity
  + add L in the adjacency list of K (as being the only direct son)
  + replace the occurrence of L in the adjacency list of Q with K
  + change the operator node I is holding to disjunction
  + leave R unchanged

Since all of the operations which are executed when finding an implication could be performed efficiently in O(1), the time complexity for this part is O(N), where N is the number of nodes our parse tree is having by the time of traversal.

#### The approach for pushing the operator not

The only outstanding task after having implemented all of the previous approaches is to push the operator not as far as possible. In other words, it has to be pushed past quantifiers, disjunctions of conjunctions. Moreover, simplifying consecutive occurrences has to be done as well.

Due to the very convenient sub-language of which our parse tree is now part, we can go ahead by just implementing the algorithm as it is:

* we visit all the nodes using depth-first search
* whenever we are located in a node of which assigned entity is the operator not we iterate through its neighbours and do the following depending on the entity held by the neighbour:
  + no entity transforms in the operator not
  + operator
    - not, transforms in no entity
    - and, transforms in or
    - or, transforms in and
  + quantifier
    - existential transforms in universal and moreover, an intermediary brand-new node below the neighbour and its son is created, holding the operator not
    - universal transforms in existential and moreover, an intermediary brand-new node below the neighbour and its son is created, holding the operator not
  + literal
    - the sign gets changed and the operator not will not go past it
  + normal form
    - nothing, because by this point our parse tree does not have any node holding this sort of entity
* one aspect which has to be highlighted is that perform all of the changes described above before going down the depth-first search traversal
  + hence, implementing the same changes but using a breadth-first search would have probably been more natural

Since each node is visited exactly once and the changes which take place to the entities of the nodes are executed in constant time, the time complexity for this approach is O(N), where N is the current number of nodes from the parse tree.

### Skolemization

Once only conjunctions are obtained, disjunctions and the negations are sticked to the predicates, implementing the Skolemization is performed:

* we visit all the nodes using depth-first search, by maintaining a stack for the universally quantified variables which are lying on the path from root to the current node; moreover, we will also maintain a std::map [reference <https://en.cppreference.com/w/cpp/container/map>] (called M) containing the mapping for the existentially quantified variables to their corresponding functions (and their respective arguments)
* in case the current node holds as an entity an universally quantified variable, the corresponding variable are pushed to the stack; when we are done with visiting the subtrees rooted in all of its neighbours, that variable will be popped off the stack
* in case the current node holds as an entity an existentially quantified variable K, a brand-new function symbol F will be created and an entry in the map M will be added for the key K, with the value corresponding to a tuple holding F and a copy of the current state of the stack (which holds the variables within K nests)
* in case the current node holds a literal as an entity, we would iterate through the terms of it and for each variable having an entry in the map M, we will replace its occurrence with the corresponding value from the map (that is, the Skolem function or constant)

After the algorithm is performed, all universal quantifiers could be disposed of. This will not affect the correctness since during the execution of the Skolemization since we kept track of the types of terms (a detail which was omitted in the presentation of the algorithm for the sake of simplicity). Having a mechanism to uniquely determine the type of the terms enables us to execute the removal of the universal quantifiers without losing any information.

The report will now analyze the time complexity of our implementation for Skolemization. The operations performed on the map M will be done efficiently, on a logarithmic complexity. However, the amount of variables stored in M could be quite large in the worst case scenario (i.e. a formula which has a very high number of nested quantifiers), precisely a number of variables bounded by N2 (where N is the number of nodes from the parse tree). Hence, the time complexity of our algorithm is O(N2log2N).

### Simplified Clausal Normal Form

After having performed all of the algorithms described in this section, the construction of the clausal normal form can be finalized.

Before continuing, it has to be established that the simplified version of the clausal normal form will be attempted. That is, the CNF will be retrieved in a convenient format for now, but the way of representing it will be changed after the retrieval (this is going to be detailed in the section dedicated to the back-end). Hence, the report will model the CNF as an array of arrays (each containing literals). This representation is completely accurate assuming that between the inner arrays there are conjunctions and between the elements of an inner array disjunctions are laying.

The report will now present the algorithm which merges two arbitrary CNF (we will call it “merge(A, B)”) having between them an operator (either conjunction or disjunction; both of the cases will be discussed):

* declare toBeAdded globally, as being a set of disjunctions which need to be appended to our formula as conjunctions
* let A and B be two CNFs, each stored as discussed, as an array of arrays containing literals
* if the operator between A and B is a conjunction, create C as a concatenation of A and B and return C
* otherwise the operator is a disjunction and there are three cases:
  + if either of A or B are empty, return the another one
  + the trivial case, both the sizes of A and B are 1
    - create C having a single element, the concatenation of the first (and only) element of A and the first (and only) element of B
    - return C
  + otherwise, this means that either A or B violate the condition from above
    - we will therefore apply the following algorithm on each of them twice:
      * let be D the input of the algorithm, passed by reference (i.e. the modifications on it are visible in the calee)
      * if all of the elements from D are having size 1, then:
        + create a brand-new predicate F
        + for each element P of D add (ㄱF ∨ P) in toBeAdded (note that what we are actually adding is an implication)
        + D becomes [[F]], where a pair of square brackets represent an array
      * otherwise, perform the following:
        + for each element P of D

create a brand-new predicate Fi

add (ㄱFi ∨ P1 ∨ P2 … ∨ Pj-1 ∨ Pj) in toBeAdded, where j is the size of P (note that we have added an implication)

P becomes F

* + - * after the algorithm has been applied twice, we can call once more merge(A, B)

Finally, after the algorithm is completed, all of the elements from the set toBeAdded have to be appended as actual conjunctions to the resulting CNF. The report will come back to this later, after all of the CNF is built, apart from this detail.

After having presented the algorithm which merges two CNFs on conjunction or on disjunction, the report will move ahead in detailing the algorithm for retrieving CNF from the parse tree:

* we will visit all of the nodes, using a depth-first search
* when we are about to come back from the recursive call for a node P, we are going to make some changes to the entity held by it:
  + if it does not hold any entity, it will now hold an empty CNF (i.e. [ ])
  + if it holds a literal L, it will now hold a CNF containing only that literal (i.e. [[L]])
  + if we are located in any of the two cases presented above, we will iterate once more through the neighbours of P, calling subsequently the merge function between its corresponding CNF and the CNF of each of its neighbours which are holding one as well
    - finally, the node P will hold as an entity the CNF corresponding for the its whole subtree
* after executing the algorithm for all of the nodes, the CNF of the whole parse tree will reside in its root

As described before, we have now to append the disjunctions from toBeAdded as conjunctions to the CNF we have retrieved from our parse tree. This could be done straightforwardly by concatenating the former to the latter.

Hence, it has been shown how the retrieval of the CNF is done. However, a good question one would ask will be why the report has proceeded in such a convoluted manner (easier implementations being trivial to spot at a glance). The reason which justifies our complex implementation is the avoidance of an exponential blow-up (as in the case of the double implications). While merging the CNFs, distributivity laws are actually applied efficiently. If this is done naively, by doing the cartesian product of the two CNFs, the size of the CNF will evidently face exponential growth.

By analysing the complexity of the merge algorithm, it could be concluded that it introduces a polynomially (in respect with the size of CNF) bounded amount of additional conjunctions. In the worst case, the time complexity of this part will be O(N), where N is the number of nodes from the parse tree.

The depth-first search algorithm runs in polynomial time, precisely O(N2). Even though the programmer could have came with much more sophisticated solutions (such as one involving the small-to-large trick, optimizing the time complexity to O(N log2N) overall), it was considered that it is not necessary, since that solution will represent only a micro-optimization (being dominated by O(N2log2N), from the Skolemization step).

Hence, the overall time complexity for the retrieval of the Simplified Clausal Normal Form is O(N2).

## Back-end

During the course of this section the report will discuss the theorem proving part of the project. First, the data structure chosen for representing the clause form will be described. Following that, the implementation of the three theorem provers (basic, depth-ordered and two-variable) will be expressed in detail.

### Clausal Normal Form and Unification

Since the theorem proving might require complex representations of the literals from the CNF, the way of representing the clausal normal form will be changed. Instead of seeing it as an array of arrays, as used and described in the intermediate representation, a much more convenient way of storing (in respect with the unification) will be required.

Hence, the CNF will now encapsulate an array of clauses. A higher level of granularity is needed, so each clause will encapsulate an array of literals. Finally, a literal will store behind its representation an array of terms. A term could be quite complex following the definition (i.e. for example a function symbol having nested other terms). With the risk of creating confusion, a term will further represent an array of terms. The figure from below will hopefully help the reader in gaining a better understanding.

{insert image here}

In order to present the unification, a bottom-up approach will be utilized (i.e. the algorithm will be presented from terms, then the transition to literals and finally the transition to clauses; the last part will be later plugged in the basic theorem prover).

#### Term Unification

The first tool which will be described as part of this section is the recursive function called “findPartialSubstitution”. This is going to take as an input two terms, A and B. It will try to unify A with B and return true in case it succeeded, false in case it fails and the substitution in case it partially succeeded. In case of failure, we will try to unify B with A as well. Finally, in case of failure A and B are not unifiable. The algorithm for findPartialSubstitution is the following:

* if A and B are identical, then return true
* otherwise, check the following:
  + if A is a constant, return false
  + if A is a variable, called V
    - if B contains that variable, return false
    - otherwise return a tuple containing V and the whole term of B, signifying that V is mapped to B (which could be a function, a constant or another variable
  + if A is a function symbol, then
    - if B is not a function symbol, return false
    - otherwise, do the following:
      * if A and B are not having the same function name, return false
      * otherwise, for each position in the list of arguments of A and B, try to unify their corresponding terms (i.e. for f(x, y, z) and f(a, b, c) we will try first x with a, then if successful y with b and finally if successful z with c)
        + if the return value R for the current unification is true, continue to the next pair of terms (because this means that the pair of terms we have tried to unify was identical)
        + otherwise, return R
      * finally, if nothing was returned yet, return true

In order to establish the time complexity of the aforementioned algorithm, the function treeSize has to be introduced, having the domain in the set of terms (let it be called T) and the co-domain in the set of natural numbers (i.e. treeSize : T → Ｎ), with the following recursive definition for treeSize(K):

* if K is a variable or a constant, treeSize(K) is 1
* if K is a n-ary function symbol with the name g, then K could be written as g(t1, t2, …, tn); therefore,

Having defined the treeSize, the time complexity for findPartialSubstitution can be established: O(M2) where M is the minimum between treeSize(A) and treeSize(B). This is owing to the fact that at each call of the function the verification for equality of A and B traverses the tree of each of them and for highly nested A and B, we would perform this check as many times as there are nodes in the tree of term with the lowest number of nodes.

The second tool which the report will present is a function called “augmentUnification”, which takes as an input two terms, A and B. This function is way simpler than the previous one, representing just a “wrapper” for findPartialSubstitution. Its purpose is to call findPartialSubstitution for A and B. In case of failure, it will call the same function but for B and A (so the arguments are just reversed). Finally, if both of the calls fail, the function will be returning false. Conversely, if both of the calls returned true, this function will be returning true. Otherwise, the function will be returning a substitution (revisit the previous algorithm for better understanding). If the reader is familiar with the Python programming language, one could make an analogy between this function and a Python generator. (i.e. if it does not return false at the initial call, it will keep returning substitutions until either of true or false is reached; if false is reached after a couple of substitutions were returned, this means that the operations have to be rolled back --- this will be explained later). The function augmentUnification has the same time complexity as findPartialSubstitution.

Finally, the third tool is a function called “createDeepCopy”, having as input one term. This function will create a deep copy of the input and return it. The implementation uses a variation of the depth-first search and the report will not present the details of it (since the algorithm is trivial and does not contain anything sophisticated). This function will be required in the next subsection. The time complexity for it is O(R), where R is the value of the treeSize for the input term.

#### Literal Unification

In the previous section three tools (or functions) were discussed that will be used in the back-end. In order to avoid confusion, we will refer to functions “augmentUnification” and “createDeepCopy” as “augmentUnificationTerm” and “createDeepCopyTerm” from now on.

The report is going to describe the function “augmentUnification”, which applies the algorithm of unification on literals. This gets as an input two literals. Similarly, after this section it will be referred to as “augmentUnificationLiteral” for better clarity.

This function will do the following:

* it will compare the literals predicate-wise
  + if the predicates are different, returns false
  + otherwise, it will call augmentUnificationTerm for each argument of the predicates
    - if we are at a pair of terms and the algorithm returns false, we will return false further
    - otherwise, if it returns true, we move to the next pair of terms
    - otherwise, if it returns a substitution, we return it
  + finally, if we did not return yet, we will return true

The time complexity of this function is O(), where E is the maximum between the length of the predicates, N is the number of arguments the predicate is having and Mi is defined as being the minimum between the values of treeSize for the ith pair.

Similarly with the previous section a function called “createDeepCopyLiteral” will be introduced. This function will create a deep copy of a given literal. The complexity of this function is O(), where F is the length of the corresponding predicate, N is the number of terms (arguments) the predicate is having and Ri corresponds to the treeSize value of the ith term.

#### Clause Unification

This functionality was implemented using C++ templates [reference https://en.cppreference.com/w/cpp/language/templates] and it represents the core of theorem proving. Once it has been detailed, it would be reasonable to expect way shorter sections describing the theorem proving. It is also worth mentioning that this part makes use of most modern C++, std::variant (CPP Reference, 2021) and lambda expressions [reference <https://en.cppreference.com/w/cpp/language/lambda>] being heavily used.

We are going to describe the function called “attemptToUnify”, which is a templated function receiving as parameters the following:

* the first clause, F
* the second clause, S
* a function returning a boolean, called literalCheck
  + this function expects two literals as parameters
* a function returning a boolean, called resolventCheck
  + this function expects a literal and a clause as a parameter

In the actual implementation, both literalCheck and resolventCheck are implemented using lambda expressions, which justifies the reason for which attemptToUnify is a templated function (since the types of the functions have to be passed as template arguments).

At this point, the algorithm behind attemptToUnify will be presented. Some of the details might be omitted, for the sake of simplicity and clarity. The algorithm does the following:

* declares an array of clauses, called result; this is going to return all of the possible clauses derived using the factoring rule from the clauses F and S
* for each literal X in F
  + for each literal Y in S
    - if literalCheck of X and Y returned true, then
      * assign to F’ a deepCopy of F
      * assign to S’ a deepCopy of S
      * make sure that F’ and S’ do not have common variables
        + in case they do, rename the variables in such a way such that they do not anymore
      * while augmentUnification applied on F’ and S’ returns a substitution
        + apply the substitution on both F’ and S’
      * finally if augmentUnification returns true, then
        + commit this unification, by removing X (which in the meantime might have gotten substituted) from F’ and Y from S’ (analogous as for X) and concatenating S’ to F’
        + append F’ to result, only if resolventCheck of X and F’ succeeded

Note that the appliance of substitution represents only a mapping from one term to another one and it could be implemented in O() for an N-ary predicate, with Ri corresponding to the value of treeSize for the ith term of it.

Similarly, the step of renaming all of the variables from F’ and S’ could be implemented using the subsequent appliance of multiple substitutions.This could be done by only retrieving all of the distinct variables from F’ and S’ in two sets, VF’ and VS’, respectively. Once we have these two sets, their intersection has to be computed, and then we should assign for each variable which is part of the intersection a brand-new variable (which is not part of neither of VF’ and VS’). This assignment will be performed on only one of the clauses F’ or S’. The complexity of this step is O(), assuming that C is the cardinal of the intersection of VF’ and VS’ and R and Q correspond to F’ and S’, respectively. The first sum corresponds to the traversal of F’ (in order to retrieve all of the variables), the second part corresponds to the traversal of S’ (analogous as for F’) and the last part corresponds to the number of variables which will be substituted multiplied by the size of the term trees of F’ and S’ (since no criteria is applied of choosing the clause which will be substituted when assigning new variables for those which are part of the intersection).

Finally, the complexity of the attemptToUnify algorithm is O( ), where:

* |F| and |S| are the size of F and S, respectively
* LC is the number of operations literalCheck is performing for a pair of literals in the worst case (this cannot be determined clearer since this function is passed as a parameter to the attemptToUnify, so we do not know its complete implementation)
* RC is the number of operations resolventCheck is performing for a literal and a clause in the worst case (as for the literalCheck function, this can be only determined at the runtime, since the function is passed as a parameter of attemptToUnify)
* the last part of our time complexity analysis is the same as for the variable renaming step, with the mention that in the worst case, it has to be renamed in addition the maximum between the cardinals of VF’ and VS’

### Basic Theorem Prover

During the course of this section the report will present the details of the basic theorem prover, which is the algorithm which stays behind all of the theorem proving algorithms implemented as part of this project.

This part is by far the most convoluted from the entire project. Because of that, the report will attempt to present a simplified version of the implementation of the algorithms (for the sake of simplicity and clarity).

This component makes use of multithreading and it is also persistent (in the sense that clauses could be dynamically appended and popped while a timeline of these operations is maintained). For the persistence part, we will be keeping track of a timestamp named “currentTimestamp” (in fact, this is more complex on the implementation side, but presenting the mechanism this way will give a solid enough understanding to the reader).

Because the basic theorem prover might not terminate, and hence, the termination is not in our control, we might sometimes decide to skip the complexity analysis.

Apart from the current set of clauses (named “AllClauses”), which is always maintained chronologically (in order to not break the persistence mechanism), the basic theorem prover keeps track of all of the clauses seen so far (by maintaining a set of clauses called “HSet”). Because the clauses could be quite large, a hashing function which relabels the variables was used (i.e. ~P(x, f(x), y, x, g(y)) becomes ~P(v1, f(v1), v2, v1, g(v2))). Again, we try to reduce the risk of overloading the reader with implementation details, so the implementation of the hashing function will not be described.

#### Factoring Rule

This algorithm implements the factoring rule described in the theoretical part of this report. For better integration of the multithreading, I decided to implement the tautology removal and the clause unification (with itself) as part of this procedure.

As it could be easily noticed, this part is executed on one single clause. Hence, multiple clauses could be treated completely independent. This is the reason why the programmer decided to make use of multithreading here, by parallelizing this procedure.

We will further describe this procedure using a top to bottom approach (i.e. the high-level overview first):

* let C be the input clause
* if C is marked as deleted, then return
* let H be the current hash of C
* perform removeDuplicates procedure on C
  + if C was changed then
    - compute H’ as the new hash of C
    - remove H from HSet
    - insert H’ to HSet
    - assign H’ to H
* perform isTautology for the clause C
  + if C was detected as being a tautology
    - mark C as deleted and save additionally the timestamp of this moment; then return
* let unificationResult be a list of clauses
* assign to unificationResult the list of all clauses resulted by unifying two literals of the same sign from C
* iterate through unificationResult
  + if the current clause currentClause is not a tautology then
    - perform removeDuplicates on currentClause
    - compute currentHash as the hash of currentClause
    - if currentHash is does not exist in HSet then
      * append currentClause in the AllClauses

The procedure removeDuplicates takes as input a clause and removes literals from it such that a literal does not occur multiple times.

The procedure isTautology takes as input a clause and searches for two atoms L and L’ such that they are identical apart from the fact that they have opposite signs. If it finds such a pair it will then report that a tautology has been found.

Finally, the procedure unification result is a variation of the function attemptToUnify. This variation could be regarded as an actual call to the attemptToUnify procedure with the following mentions:

* F and S will point to the same thing
* the step in which the intersection of VF’ and VS’ is enforced to be empty will be omitted
* literalCheck will return true only if the two literals are having the same predicate name and the same sign
* resolventCheck will always return true
* the time complexity will be the same as for attemptToUnify with the mention that C (the cardinal of the intersection of VF’ and VS’) will be equal to 0

#### Subsumption Rule

The subsumption is executed on a single clause, in the same way as the factoring rule does, independently from the other clauses from AllClauses. Similarly, since this isolation is enforced by the logic of the algorithm, the code corresponding to the implementation could be parallelized.

We will now describe the algorithm:

* let C be the input clause
* if C is marked as deleted, return
* let CHS be the ordered set of the hashes for its literals (i.e. for the clause (~P(x) ∨ L(x, y) ∨ R(z, y, x)), the corresponding the set is {L(v1, v2), ~P(v1), R(v3, v2, v1)})
* we will now iterate through all of the other clauses which have not been deleted
  + for each clause E
    - we compute EHS
    - if EHS is included in CHS, then C is subsumed by E
      * C gets marked as deleted and we remember the timestamp at which it got marked as deleted
    - if CHS is included in EHS, then E is subsumed by C
      * E gets marked as deleted and we remember the timestamp at which it got marked as deleted

The algorithm for subsumption works for ground clauses and it might work as well for clauses very dense in terms of variables. The reason for the latter is because relabelling the variables in the way described in one of the clauses might not match the relabelling from another clause. However, a correct implementation for non-ground clauses might be performed in exponential time, which would outweigh the benefits.

#### Resolution Rule and Theorem Proving

The report will describe the resolution rule together with the previous two rules, as part of the basic theorem proving. Similarly with the subsumption and factoring rules, this rule could be implemented as well using multi-threading.

The report will now present the basic theorem proving, which makes use of the previously discussed function, attemptToUnify. Hence, the algorithm is the following:

* literalCheck applied on two literals will return true if the corresponding two predicates are the same, but the literals are of opposite signs
* resolventCheck will return true no matter what is the input
* let repeat be a boolean initialized with true
* while repeat is true do
  + set the variable repeat to false
  + for all clauses C in AllClauses do the following (this step creates separate threads in the actual implementation):
    - if C is marked as deleted, then continue
    - for all clauses G in AllClauses do the following
      * if G is marked as deleted, then continue
      * let unificationResult be a list of clauses
      * assign to unificationResult the result of attemptToUnify, having as arguments C, G, literalCheck and resolventCheck
      * if unificationResult is not empty, then do
        + iterate through all of the clauses J from unificationResult

if the clause J is a tautology, then continue

perform removeDuplicates on J

let H be the hash of J

if H exists in HSet, then continue

otherwise do the following

insert H to HSet

if J is the empty clause, then

return true

append J to AllClauses

set the variable repeat to true

* finally, if this point is reached, return false (we have saturated)

This concludes the presentation of the basic theorem prover. The next section will be presenting the depth-ordered theorem prover, which is a very slight variation of the former.

### Depth-Ordered Theorem Prover

The depth-ordered theorem prover is identical with the basic theorem prover described in the previous section with the exception that it has a different implementation for the resolventCheck. Hence, solely the implementation of this function will be detailed during the course of this section.

As before, a top-to-bottom approach for presenting the resolventCheck function will be utilized. The following implementation will make use of a function called isAOrdering, which is going to be detailed later.

Hence, the resolventCheck function does the following:

* let L and C be the input of the function, the resolved literal and the clause, respectively
* let LC be the set of all the literals which are part of the clause C
* for each literal T in LC
  + if isAOrdering applied on L and T (in this order, because the order matters) returns true then
    - return false
* return true

We will now define the function isAOrdering:

* let A and B be two literals, the input of the function
* let AMax and BMax be the maximum depth of a variable in A and B, respectively
* if A does not have variables, return false
* if B does not have variables, return false
* if AMax ≤ BMax, return false
* let AVars and BVars be the set of variables from A and B, respectively
* for each variable V in AVars
  + let AV be the maximum depth of V in A
  + let BV be the maximum depth of V in B
  + if AV ≥ BV, return false
* for each variable V in BVars
  + let AV be the maximum depth of V in A
  + let BV be the maximum depth of V in B
  + if BV ≤ AV, return false
* return true

The changes mentioned above are the only ones needed to transform the basic theorem prover into a depth-ordered theorem prover. The time complexity of the entire depth-ordered theorem prover is difficult to compute (even though the termination is guaranteed). However the complexity for the resolventCheck and isAOrdering function could be calculated.

The time complexity for the resolventCheck is O() where |C| represents the size of the clause set C and A(L, LCi) represents the maximum number of operations (in the worst case) consumed on running the isAOrdering function on the literal L and the ith literal from the clause C.

Since the definition of A(L, LCi) does not yet have the desired level of verbosity, we will further define the time complexity for A(L, W), where both L and W are literals, having the size of their arity LA and WA, respectively. Finally, the time complexity for A(L, W) is O( ), where Li and Wi are the values corresponding to the treeSize function for the ithterm of the literal L and ith term of the literal W, respectively.

### Two-Variable Theorem Prover

In this section, the only theorem prover left to discuss will be presented: the two-variable fragment theorem prover. The explanation will start with describing the variable renaming approach and continue with explaining the version which assumes that the formula to be proved does not contain equality. Finally, the version for the case with equality will be detailed.

#### Variable renaming

Before moving on to the algorithms, the approach of renaming the variables will be detailed:

* let C be the set of clauses
* let G be an empty graph, in which nodes and edges could be dynamically used
* for each clause Q in C
  + let LVMax be the maximum number of distinct variables across all of the literals of the clause Q
  + if LVMax is greater than 2, then return false (i.e. the given formula is not part of the two-variable fragment)
  + otherwise:
    - for each literal L in Q
      * let V be the set of variables occurring in L
      * if V has size 2:
        + draw an edge from the node corresponding to the first variable to the node corresponding to the second variable in G
      * if V has size 1:
        + add a node corresponding to the variable from V in G
* finally, return true

Finally, if the algorithm above returns true, the algorithm can proceed by 2-colouring the graph G. Since the graph built by the algorithm aforementioned is bipartite, 2-colouring can be performed in polynomial time (in respect with the size of the graph, which is bounded by the number of distinct variables occurring in the given formula, further bounded by the size of the input), using the depth-first search algorithm. Since detailing this algorithm is out of the project scope, the exact implementation will be omitted.

#### Without Equality

Since the two-variable theorem prover inherits from the basic theorem prover, the first step of defining it will present the implementations of the literalCheck and resolventCheck.

The implementation of the literalCheck will do as follows:

* let F and S be the input literals
* if F and S do not have opposite sign, return false
* if F and S do not have the same predicate names, return false
* let VF and VS be the variables contained in F and S, respectively
* if either of |VF| and |VS| are having the size 2 and none of F and S represent the equality (i.e. their predicate names is not called Equality), then return true
* otherwise return false

Please note that even though we check for equality, during the course of this section the respective check could be ignored. The reason for checking for something which would be never encountered is to avoid redefining this function again in the next section.

The implementation of the resolventCheck will be the same as the one for the depth-ordered theorem prover. It could be the same as for the basic theorem prover, but that does not guarantee the termination.

The time complexity for the literalCheck function is O( ), where FP and Sp are the lengths of the predicate names for F and S, respectively and FTi and STi are the values corresponding to the treeSize function for the ithterm of the literal F and ith term of the literal S, respectively.

The basic theorem prover will now be executed using the aforementioned literalCheck and resolventCheck implementation. It will either derive the empty clause or reach saturation. In case the former happens, we found the contradiction and hence, the algorithm can report validity for the given formula. Otherwise, all of the clauses containing literals in two variables (excepting the Equality predicate) will be removed and the algorithm will continue.

Let D be a brand-new instance of the depth-order theorem prover (which, as mentioned before, is a persistent data structure, being able to efficiently perform appendance and removal from the set of clauses) and C be the set of clauses after the disposal of those mentioned above. The next step of the algorithm will be a backtracking procedure doing the following:

* it will get as an input the set of clauses C (which is 0-indexed), the index current clause which is processed R (initially 0) and the depth-order theorem prover D (which is passed by reference)
* if R is equal with the size of C
  + run D (which holds all of the clauses added so far)
    - if it returned true (which means that saturation was reached), return true
    - otherwise, return false
* otherwise, group the literals from the Rth clause on the set of variables occurring in that respective literal
  + for each such set,
    - create a new clause W containing all of the literals from having that set of variables
    - append W to D, and remember the timestamp T at which W was added
    - run backtracking recursively on the tuple (C, R + 1, D)
      * if it returned true, return true
    - revert D to the timestamp T (that will not only remove W from D, but also all of the clauses inferred past W)
* finally, if this point is reached, return false

Since the set of the variables can be either empty, {x} and {y}, the time complexity of our algorithm is O(3|C| \* DTP), where |C| is the size of the clause set and DTP is the amount of operations the depth-ordered theorem prover will perform in the worst-case for a given set of clauses.

#### With Equality

Assuming that the formula contains equality, the algorithm has to resolve the equality after the disposal was executed and before the initial call of the backtracking.

The approach for implementing this part is quite surprising, involving the conversion of the set of clauses C into a string and subsequently appending to that string the implications described by the Lemmas 5 and 6 in [reference HDNIPH]. After performing these concatenations, the string will be passed further as an input to the Clausal Normal Form algorithm. The output of the latter will overwrite C.

Since this part implements exclusively string manipulations consistent with either the Lemmas 5 and 6 from [reference HDNIPH] or with what was presented earlier in the theoretical part of this report, it will be omitted for the sake of simplicity.

# Evaluation

This section will be dedicated to the evaluation part of the project and will highlight what were the main methods of asserting the correctness of the implementation.

## Unit testing

Google’s testing framework was used in order to unit-test some functionality from the front-end component. However, in the early stages of the development it was noticed that the value unit-testing adds up is not as big as the end-to-end testing would be. This is because the functionality of the theorem proving is distributed across multiple components and the testing should assert more on the communication between these components other than the implementation of each component in part. If the former would fail, the latter might not fail whereas if the former would pass, the latter will pass with a very high chance.

## End-to-End testing

End-to-end testing is a black-box testing technique of verifying the functionality of the code. For the purpose of the project, this translates into a way of expecting an exact output for a specific input.

### Using the problems found on tptp.org

The first way of achieving this kind of testing was to use the problems available on tptp.org. Even though the archive of problems from tptp.org is generous in terms of size, after some filtering only less than a hundred of these problems were representing valid inputs for the use-case of the implementation. This happened because only tests containing not more than two variables, not containing function symbols in the input nor special operators like the arithmetic ones were of interest.

The final archive of tests contained only 47 problems, majority of them being part of the “SYN” (which possibly comes from “syntax”) problem set.

Even though 47 test-cases is not a big amount, this part of the testing detected most of the mistakes which happened during the implementation process. For automating this part, a Linux Bash script was produced and used. (this shall be revisited later)

### Using general formulas

A general formula generator was written as well in order to achieve end-to-end testing. Even though the formulas generated represented as well a good way of detecting potential bugs, it was still not enough. The reason for that was because the implementation of this generator involved a recursive function of whose calls were randomly weighted. Even though the generated formula was a valid first-order logic formula, figuring out whether it is a theorem or not was practically impossible without doing the same for a third-party software which would act as a verifier. Definitely Vampire could have been that software, but generating the same formula for the theorem prover as for Vampire represented a challenge as well. Moreover, even if a lot of effort would have been invested in finding a way of resolving these issues, the generator could not have been adapted such it generates a formula with an exact number of variables and that would have imply that the programmer would be able to produce large formulas part of the two-variable fragment only probabilistically. Finally, since the two-variable fragment is known as being decidable only when it does not contain function symbols, the generator should have been resilient to this problem as well. Approaches for all of the issues above excluding the last one exists. However, the last problem is probably too complex for being solved in polynomial time (because keeping track of the number of variables and the potential arising function symbols while generating a totally unrestricted first-order logic formula is infeasible).

Because figuring out how to couple the things mentioned above always lead to new challenges, the idea of using the general formula generator for black box testing was abandoned. Fortunately, having a third-party verifier proved itself as being a solid starting point and the exact strategy is described in the next section.

## Testing against Vampire

During the course of this section, a test-generator will be presented. The contribution for the idea behind the generator (where not referenced) is of my supervisor.

In order to do end-to-end testing using Vampire, we will have to first define the Scott’s normal form [reference to M.Otto]:

In the formula from above, is a formula in CNF containing A clauses and is an atomic formula. All of the atoms occurring in the aforementioned form are having arity exactly two.

In order to adapt the Scott’s normal form as a starting point for a test-generator for two-variable fragment formulas, the following conventions have to be adopted:

* the predicate names occurring in could occur in and vice versa
  + in other words, there is a global set Q (of cardinality |Q|) of predicate names which includes all of the names from and from which the predicate names are be picked
* each of the A clauses of will have exactly L literals, with L being chosen randomly from the range [LMIN, LMAX], where LMIN and LMAX are parameters for the generator representing the minimum and the maximum number of literals per clause, respectively

Having defined the Scott’s normal form, generating now a formula in the two-variable fragment becomes trivial:

* A, C, LMIN, LMAX get set with a value
* let L be a number randomly pick from the range [LMIN, LMAX]
* let P be the desired value for |Q|
  + P gets set
  + Q gets populated with P unique predicate names (containing letters of the latin alphabet), in increasing order of their lengths (i.e. first all of the predicate names of size 1, then all of the predicate names of size 2 and so on)
* let F be a string initialized with “∀x∀y(“
* let alpha be an empty set of strings
* repeat A times the following
  + let W be a list of literals
  + repeat L times the following
    - let R be an empty string
    - pick a boolean B at random
      * if true, append negation to R
      * otherwise let R unchanged
    - pick a predicate name U at random from Q
    - append U to R
    - pick a boolean B’ at random
      * if true, append “(x, y)” to R
      * otherwise append “(y, x)” to R
    - append R to W
  + join all the elements of W on disjunction (i.e. {“¬Pa(x, y)”, “Pb(y, x)”} becomes ¬Pa(x, y) ∨ Pb(y, x)) and append the result to alpha
* join all the elements of W on conjunction and append the result to F
* append “∨ x = y) ∧” to F
* let beta be an empty string
* repeat C times the following
  + let betai be string initialized with “(∀x∃y(
  + pick a predicate name K at random from Q
  + pick a boolean B’’ at random
    - if true, append “(x, y)” to betai
    - otherwise append “(y, x)” to betai
  + append “∧ x≠y)) ∧” to betai
  + append betai to beta
* remove the last character from beta
* append beta to F
* return F

The algorithm from above runs in polynomial time in respect with A\*LMAX+C, assuming that the lengths of the predicate names are reasonably small in comparison with the parameters which is the case since the following function is exponential:

* t(x) = the number of unique predicate names of length smaller or equal with x

Finally, the test-generator from above has the following advantages:

* it can be adapted to both the theorem prover and Vampire (it does not use any particularity of either of their languages), assuming that they have the same operator precedence (which in theory happens, but extra brackets could be added for safety)
* it enforces the number of variables (which is a good thing, since all of the implemented theorem provers could be tested using it)
* when negated, it does not add any function symbols (it adds only constant) which makes possible the testing on the two-variable fragment without function symbols
* it even tests the front-end and the intermediate representation because when the formula has equality (which would happen in most of the cases of interest)

Finally, a script which generates random two-variables formulas can be implemented. This script will be able to perform end-to-end testing, since the output of Vampire will be considered correct. Such a script was actually implemented as part of the project and in 25000 randomly generated formulas it detected the same verdict for both vampire and the two-variable theorem prover.

## Stress testing

In order to optimize the theorem prover, the decision of making its core infrastructure (involving the factoring, resolution and subsumption rules) multi-threaded was made. A challenge of this design aspect was related to testing, precisely how to make sure that at different runs the theorem prover does not have different outputs in terms of meaning (i.e. derived empty clause versus reaching saturation).

In addressing this issue, the 47 tests from tptp.org were used. The script running those tests for each theorem prover (basic, depth-ordered and two-variable) was run 2000 times and the results were checked for each test-suite of 47 tests in part. None of them has failed, all of them being consistent with the expected output.

The value this kind of testing brought to the project was solid, detecting many bugs during the implementation process, which the other ones did not. The best example is a bug related to a race condition which was detected in 1000 runs, occurring just 6 times (that is 6 out of 47000).

# Experiments

# Reflection

# Conclusion

# Appendix 1

# References

De Nivelle, H. and Pratt-Hartmann, I. (2001). ‘A Resolution-Based Decision Procedure for the Two-Variable Fragment with Equality’, *Automated Reasoning,* pp. 211-225.

Gore, R., Leitsch, A. and Nipkow, T. (2001) ‘Automated Reasoning’, Proceedings of the First International Joint Conference, IJCAR, Siena, Italy, 2001, pp.211-225

Wilfrid, H. (1977). *Logic.* 1st edn. Penguin.

Leitsch, A. (1997). *The Resolution Calculus.* 1st edn. Springer.

Joyner Jr., W.H., Fermueller, C.G., Leitsch, A., Hustadt, U. and Tammet, T. (2002). ‘Ordered Resolution’ [PowerPoint presentation]. Available at: <https://people.mpi-inf.mpg.de/~hillen/documents/4_OrderedResolution.pdf>

(Accessed: 7 April 2021).

CPP Reference (2021). *C++ Utilities Library: std::variant.* Available at: <https://en.cppreference.com/w/cpp/utility/variant> (Accessed: 7 April 2021)

CPP Reference (2021). *C++ Utilities Library: std::execution.* Available at: <https://en.cppreference.com/w/cpp/algorithm/execution_policy_tag_t>

(Accessed: 7 April 2021).

Debian (2021). *Package: libtbb-dev (2020.3-1 and others).* Available at: <https://packages.debian.org/sid/libtbb-dev> (Accessed: 7 April 2021).

Googletest (2021). *Googletest.* Available at: <https://github.com/google/googletest>

(Accessed: 7 April 2021).

CPP Reference (2021). *C++ Utilities Library: std::unordered\_set.* Available at:

<https://en.cppreference.com/w/cpp/container/unordered_set>

(Accessed: 7 April 2021).