Modelling Rust's Reference Ownership Analysis Declaratively in Datalog

Albin Stjerna

June 1, 2019

Abstract

This thesis is about something.

Contents

Chapter 1

Introduction

Rust is a young systems language originally developed at Mozilla Research [1]. Its stated intention is to combine high-level programming language features like automatic memory management and strong safety guarantees, in particular in the presence of concurrency or parallellism, with predictable performance and pay-as-you-go abstractions in the style of C++ and similar systems languages.

One of its core features is the memory ownership model, which enables compile-time safety guarantees against data races, unsafe pointer dereferencing, and runtime-free automatic memory management, including for dynamic memory allocated on the heap.

This report describes the implementation of Rust's memory safety checker, called the borrow checker, in an embedded Datalog engine, as well as its optimisation. In practice, the full analysis encompasses a variable liveness analysis, initialisation and deinitialisation tracking, and may-reference analysis for validation of Rust's memory safety guarantees.

mention results.

Chapter 2

Background

Whenever a reference to a resource is created in Rust, its borrowing rules described in Section 2.1 must be respected for as long as the reference is alive, including across function calls [2]. In order to enforce these rules, the Rust language treats the scope of a reference, called its lifetime, as part of its type, and also provides facilities for the programmer to name and reason about them as they would any other type.

Since its release, the Rust compiler has been extended through proposal RFC 2094 to add support for so-called non-lexical lifetimes (NLLs), allowing the compiler to calculate lifetimes of references based on the control-flow graph rather than the lexical scopes of variables [3]. During the spring of 2018, Nicholas Matsakis began experimenting with a new formulation of the borrow checker, called Polonius, using rules written in Datalog [4]. The intention was to use Datalog to allow for a more advanced analysis while also allowing for better compile-time performance through the advances done centrally to the fixpoint solving provided by the Datalog engine used for the computations [5].

Datalog, and other types of logic programming has been previously employed for program analysis, in particular pointer analyses such as may-point-to and must-point-to analysis, both similar to what is described in this report in that they require fix-point solving and graph traversal, often with a context sensitive analysis (i.e. respecting function boundaries) like the one described here[6]–[17]. These systems employ a wide variety of solver technologies and storage back-ends for fact storage, from Binary Decision Diagrams (BDDs) to explicit tuple storage, as used in this study. Some of them, like Flix, also extends Datalog specifically for static program analysis[16].

In addition to being context-sensitive, Rust's borrow checker is also flow-sensitive (i.e. performs analysis for each program point), like the system described by Hardekopf and Lin, and whose form is very similar to the analysis performed in practice by Polonius [18].

A 2016 study uses the Soufflé system, which synthesizes performant C++ code from the Datalog specifications, similar to how Datafrog embeds a minimal solver as a Rust library, to show promising performance for analysis of large programs [19]. The Doop system, developed by Smaragdakis and Bravenboer, also shows that explicit tuple storage sometimes vastly outperforms BDDs in terms of execution time [14], as do sparse bitmaps [12].

Formally, the semantics of Rust's lifetime rules has been captured in the language Oxide, described by Weiss, Patterson, Matsakis, *et al.* [20]. This report will therefore not concern itself with the semantics of the borrow rules, except where necessary to explain the rules.

The contributions made within the scope of the thesis project specifically includes the implementation of liveness and initialisation calculations (Sections 3.1.2 and 3.1.3 respectively), as well as work on region inference for higher-ranked types. Finally, the report also evaluates the runtime performance of the system and suggests some potential optimisations in Section 3.3, and performs a field study of the shape of input data in Section 3.2. The core rules of Polonius for the region constraints were already written when the project started, but are described in Section **??** for completeness.

2.1 The Borrowing Rules

Most of these examples are borrowed from Weiss, Patterson, Matsakis, *et al.* [20].

Variables must be provably initialised before use Whenever a variable is used, the compiler must be able to tell that it is guaranteed to be initialised:

```
let x: u32;
let y = x + 1; // ERROR: x is not initialised
```

A move deinitialises a variable Whenever ownership of a variable is passed on (*moved* in Rust parlance), e.g. by a method call or reassignment, the variable becomes deinitialised:

```
struct Point(u32, u32);

let mut pt = Point(6, 9);
let x = pt;
let y = pt; // ERROR: pt was already moved to x
```

There can be any number of shared references A shared reference, also called a *borrow* of a variable, is created with the `&` operator, and there can be any number of simultaneously live shared references to a variable:

```
struct Point(u32, u32);

let mut pt = Point(6, 9);
let x = &pt;
let y = &pt; // This is fine
```

There can only be one simultaneous live unique reference Whenever a unique reference is created, with `&mut`, it must be unique:

```
struct Point(u32, u32);

let mut pt = Point(6, 9);
let x = &mut pt;
let y = &mut pt; // ERROR: pt is already borrowed

// code that uses x and y
```

This error happens even if the first borrow is shared, but not if either `x` or `y` are dead (not used).

A reference must not outlive its referent A reference must go out of scope at the very latest at the same time as its referent, which protects aganst use-after-frees:

```
struct Point(u32, u32);

let x = {
    let mut pt = Point(6, 9);
    &pt
};

let z = x.0; // ERROR: pt does not live long enough
```

In this example, we try to set `x` to point to the variable `pt` inside of a block that has gone out of scope before `x` does.

2.2 The Borrow Check in the Rust Compiler

The logic of the borrow check as described in Section 2.3 is calculated at the level of an intermediate representation of Rust called the Mid-Level Intermediate Representation (MIR), corresponding to the basic blocks of program control flow. The input data to the Polonius solver is generated in the Rust compiler by analysing this intermediate representation. This means that we can safely assume to be working with simple variable-value assignment expressions, of the type `_1 = _2`.

2.3 From Lifetimes to Reference Provenance

As lifetimes are a part of a variable's types, they can be referred to by name like any other type using the syntax `&'lifetime`. In the literature, the terms "region" [4], "(named) lifetime" , and "reference provenance" [20] (RefProv) are all employed. As the section heading suggests, I will use the last one of them as I believe it best captures the concept.

From a type system perspective, the RefProv is part of the type of any reference and corresponds to the borrow expressions that might have generated it. For example, if a reference `r` has the type `&'a Point`, `r` is only valid as long as the terms of the loans in `'a` are upheld. Take for example the code of Listing 2.1, where `p` would have the type `&'a i32` where `a` is the set $\{L_0, L_1\}$.

Listing 2.1: An example of a multi-path borrow.

```
let x = vec![1, 2];

let p: &'a i32 = if random() {
  &x[0] // Loan L0
} else {
  &x[1] // Loan L1
};
```

If a reference is used in an assignment like `let p: &'b i32 = &'a x`, the reference, `p`, cannot outlive the assigned value, `x`. More formally the type of the right-hand side, `&'a i32`, must be a subtype of the left-hand side's type; `&'a i32 <: &'b i32`. In practice, this establishes that `'b` lives at most as long as `'a`, which means that the subtyping rules for regions establishes a set membership constraint between the regions, as seen in Rule 2.1.

$$\frac{a \subseteq b}{\texttt{\&'a u32} <: \texttt{\&'b u32}} \tag{2.1}$$

How to incorporate the fact that this is with respect to the current point?

Finally, when talking about the *liveness* of a RefProv r, we will mean that r occurs (what does occurs mean? Does it encompass exploded subset-types etc?) in the type of a variable that is live at some point in the control-flow graph p, with the meaning that any of the loans in r might be dereferenced at control-flow points reachable from p, and thus that the terms of the loans in r must be respected at that point.

2.4 Deallocation As a Special Case of Variable Use

When Rust's variables go out of scope, they are implicitly deallocated, or dropped in Rust parlance. Explicit deallocation is also possible by calling the function `drop()`, which takes ownership of a variable and performs deallocation, or, for complex objects, calls the `drop()` method. For some types such as integers, deallocation is not necessary and the compiler generates no actual `drop()`:s in the MIR.

Rust provides a default deallocator for data structures, but it can be overridden. This has repercussions on liveness calculations, because while the default deallocator for an object never needs to access its fields, a custom deallocator might access any of

them. This means that any conditions of a loan that resulted in a reference r stored in a `struct` s instance a must only be respected as far as `a.drop()` is concerned if s implements a custom deallocator. Otherwise the loan of r may be safely violated, as the default deallocator never dereferences r and thus doesn't require r to be valid. An illustration of this can be seen in Listing 2.2.

Listing 2.2: The custom deallocator for `OwnDrop` enforces the loan of `data`, but the loan in `OwnDrop` is effectively dead and thus can be violated.

```
struct OwnDrop<'a> {
    data: &'a u32,
}

struct DefaultDrop<'a> {
    data: &'a u32,
}

impl<'a> Drop for OwnDrop<'a> {
    fn drop(&mut self) {
        // might access self.data
    }
}

fn main() {
    let mut x = 13;
    let a = OwnDrop { data: &x };

    let mut y = 12;
    let b = DefaultDrop { data: &y };

    // let mutrefa = &mut x;
    // ERROR: the loan of x must be respected...

    // ...but the loan of y need not be!
    let mutref = &mut y;
    *mutref = 17;

    // all variables are implicitly dropped here
}
```

2.5 Datalog and Datafrog

This thesis uses the notation of Soufflé [19]. Datalog is a derivative of the logic programming language Prolog, with x guarantees about execution. It describes fixpoint calculations over logical relations as predicates, described as fixed input *facts*, computed *relations*, or *rules* describing how to populate the relations based on facts or other relations. For example, defining a fact describing that an individual is another

individual's parent might look like `parent(mary, john).`, while computing the `ancestor` relation could use the two rules `ancestor(X, Y) :- parent(X, Y).` and `ancestor(X, Y) :- parent(X, Z), ancestor(Z, Y).`, reflecting the fact that ancestry is respectively either parenthood or transitive parenthood (example from the Wikipedia article on Datalog [21]).

Datafrog [5] is a minimalist Datalog implementation embedded in Rust, providing an implementation of a worst-case optimal join algorithm as described in [22]. The fact that Datafrog is embedded in Rust means that standard Rust language abstractions are used to describe the computation. Static facts are described as `Relations`, while `Variables` are used to capture the results of computations, both of which are essentially sets of tuples. Rules are described using a join with either a `Variable` or an `Relation`, with an optimised join method for joins with the variable itself. Only single-step joins on the first tuple element are possible, which means that more complex rules must be written with intermediary variables, and manual indices created whenever a relation must be joined on a variable which is not the first in the tuple.

Listing 2.3: The implementation of `var_live(V, P)` in Datafrog

```
var_live_var.from_leapjoin(
    &var_live_var,
    (
        var_defined_rel.extend_anti(|&(v, _q)| v),
        cfg_edge_reverse_rel.extend_with(|&(_v, q)| q),
    ),
    |&(v, _q), &p| (v, p),
);
```

As an example, the Datafrog code for `var_live(V, P)` of Listing 3.1 becomes the code in Listing 2.3, and the corresponding join used for the first half of `region_live_at(R, P)` of Listing 3.2 can be seen in Listing 2.4.

Listing 2.4: The first half of the implementation of `region_live_at(R, P)` in Datafrog

```
region_live_at_var.from_join(
    &var_drop_live_var,
    &var_drops_region_rel,
    |_v, &p, &r| {
        ((r, p), ())
    });
```

Chapter 3

A Datalog Model for the Rust Borrow Checker

3.1 The Borrow Checker in Datalog

3.1.1 Starting Facts

Verify the terminology for this in Datalog parlance!

`borrow_region(R, B, P)` the region R may refer to data from borrow B starting at the point P (this is usually the point *after* the right-hand-side of a borrow expression).

`universal_region(R)` for each named region$R$ supplied to the function. R is considered live in every point.

`cfg_edge(P,Q)` whenever there is a relation $P \rightarrow Q$ in the control flow graph.

`killed(B,P)` when some prefix of the path borrowed at B is assigned at point P.

`outlives(R1, R2, P)` when $R_1 \subseteq R_2$ must hold at point P, a consequence of subtyping relationships as described in Rule (2.1).

`invalidates(P, L)` when the loan L is invalidated at point P.

`var_used(V, P)` when the variable V is used for anything but a drop at point P.

`var_defined(V, P)` when the variable V is assigned to (killed) at point P.

`var_drop_used(V, P)` when the variable V is used in a drop at point P.

`var_uses_region(V, R)` when the type of V includes the RefProv R.

`var_drops_region(V, R)` when the type of V includes the RefProf R, and V also implements a custom drop method which might need all of V's data, as discussed in Section 2.4.

3.1.2 Reference Liveness

The basic liveness of a variable (Listing 3.1) is defined as follows: if a variable v is live in some point q and q is reachable from p in the control-flow graph, then v is live in p too unless it was overwritten; in essence liveness is propagated backwards through the CFG.

Listing 3.1: The Datalog rules for variable use-liveness.

```
var_live(V, P) :-
    var_live(V, Q),
    cfg_edge(P, Q),
    !var_defined(V, P).
```

There is also a very similar relation for drop-liveness with an identical shape and different input relations corresponding to drop-uses. An example of the output from this calculation can be seen in Figure 3.1.

Both of these are then used to calculate the reference liveness relation (Listing 3.2), which serves as input for the rest of the borrow checker. A given RefProv r is live at some point p if it is in the type of a variable v which is either drop-live or use-live at p, with some notable caveats for drop-liveness (discussed in Section 2.4) embedded in the `var_drops_region` relation.

Listing 3.2: The Datalog rules RefProv liveness.

```
region_live_at(R, P) :-
    var_drop_live(V, P),
    var_drops_region(V, R).

region_live_at(R, P) :-
        var_live(V, P),
        var_uses_region(V, R).
```

3.1.3 Variable Initialisation

3.1.4 Loan Constraint Propagation

3.1.5 Loan Violations

3.1.6 Illegal Subset Relations

3.2 A Field Study of Borrow Patterns

3.3 Optimising the Borrow Checker

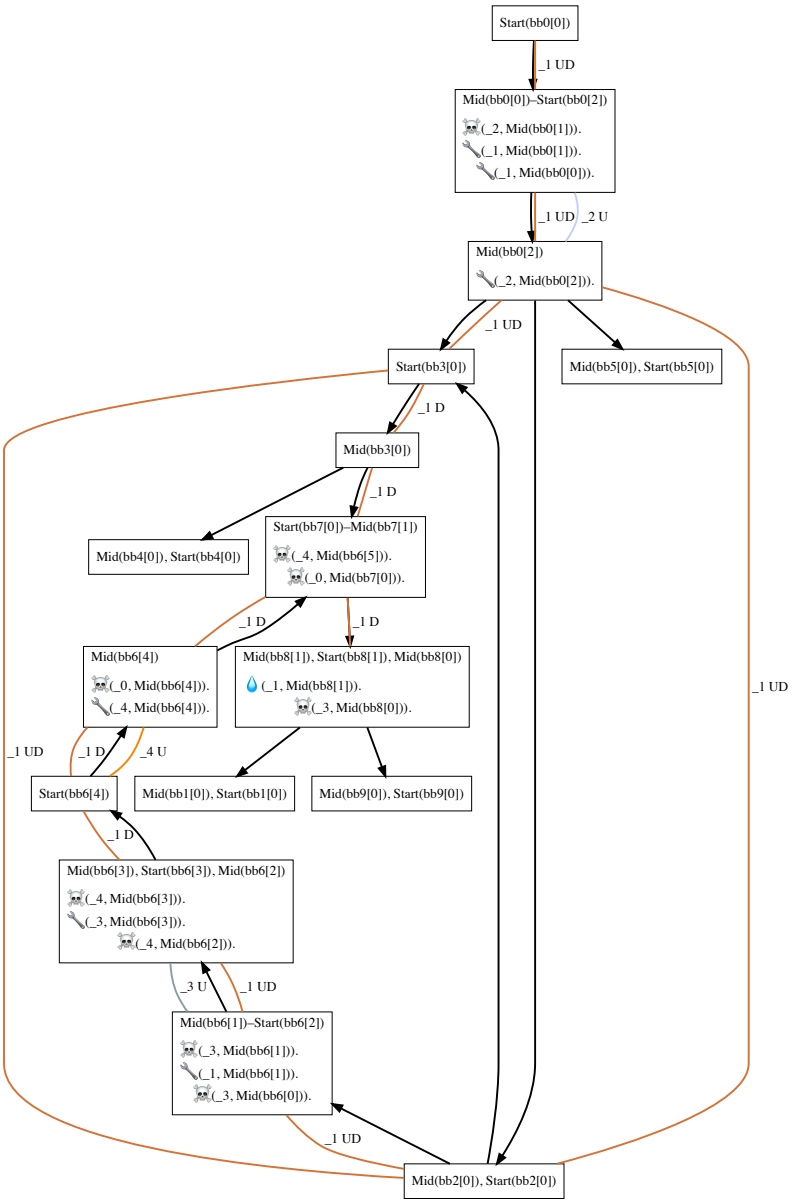

Figure 3.1: A graph representation of the the variable liveness calculation results, with relevant Polonius facts as they occur (a droplet symbolising `var_drop_used`*, a wrench* `var_used`*, and a skull and crossbones symbolising* `var_defined`*). Variables are named by prefixing underscores, and edges annotated with the propagated live variable and its liveness type(s) (**D**rop or **U**se).*

Chapter 4

Conclusions and Future Work

Bibliography

[1] N. D. Matsakis and F. S. Klock II, "The rust language," in *Proceedings of the 2014 ACM SIGAda Annual Conference on High Integrity Language Technology*, ser. HILT '14, event-place: Portland, Oregon, USA, New York, NY, USA: ACM, 2014, pp. 103–104, isbn: 978-1-4503-3217-0. doi: 10.1145/2663171.2663188. [Online]. Available: http://doi.acm.org/10.1145/2663171.2663188 (visited on 03/29/2019) (cit. on p. 4).

[2] C. Nichols and S. Klabnik. (2019). The Rust programming language, [Online]. Available: https://doc.rust-lang.org/book/foreword.html (visited on 04/01/2019) (cit. on p. 5).

[3] *RFC 2094: Non-lexical lifetimes*, original-date: 2014-03-07T21:29:00Z, Apr. 1, 2019. [Online]. Available: https://github.com/rust-lang/rfcs (visited on 04/01/2019) (cit. on p. 5).

[4] N. D. Matsakis. (Apr. 27, 2018). An alias-based formulation of the borrow checker, Baby Steps, [Online]. Available: http://smallcultfollowing.com/babysteps/blog/2018/04/27/an-alias-based-formulation-of-the-borrow-checker/ (visited on 04/01/2019) (cit. on pp. 5, 7).

[5] *Datafrog: A lightweight datalog engine in rust*, Mar. 29, 2019. [Online]. Available: https://github.com/rust-lang/datafrog (visited on 04/01/2019) (cit. on pp. 5, 10).

[6] S. Dawson, C. R. Ramakrishnan, and D. S. Warren, "Practical program analysis using general purpose logic programming systems—a case study," in *Proceedings of the ACM SIGPLAN 1996 Conference on Programming Language Design and Implementation*, ser. PLDI '96, Philadelphia, Pennsylvania, USA: ACM, 1996, pp. 117–126, isbn: 0-89791-795-2. doi: 10.1145/231379.231399. [Online]. Available: http://doi.acm.org.ezproxy.its.uu.se/10.1145/231379.231399 (cit. on p. 5).

[7] M. Berndl, O. Lhoták, F. Qian, L. Hendren, and N. Umanee, "Points-to analysis using bdds," *SIGPLAN Not.*, vol. 38, no. 5, pp. 103–114, May 2003, issn: 0362-1340. doi: 10.1145/780822.781144. [Online]. Available: http://doi.acm.org.ezproxy.its.uu.se/10.1145/780822.781144 (cit. on p. 5).

[8] E. Hajiyev, M. Verbaere, O. de Moor, and K. De Volder, "Codequest: Querying source code with datalog," in *OOPSLA Companion*, Citeseer, 2005, pp. 102–103 (cit. on p. 5).

[9] J. Whaley and M. S. Lam, "Cloning-based context-sensitive pointer alias analysis using binary decision diagrams," *SIGPLAN Not.*, vol. 39, no. 6, pp. 131–144, Jun. 2004, issn: 0362-1340. doi: 10.1145/996893.996859. [Online]. Available: http://doi.acm.org/10.1145/996893.996859 (cit. on p. 5).

[10] M. S. Lam, J. Whaley, V. B. Livshits, M. C. Martin, D. Avots, M. Carbin, and C. Unkel, "Context-sensitive program analysis as database queries," in *Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, ACM, 2005, pp. 1–12 (cit. on p. 5).

[11] W. C. Benton and C. N. Fischer, "Interactive, scalable, declarative program analysis: From prototype to implementation," in *Proceedings of the 9th ACM SIGPLAN International Conference on Principles and Practice of Declarative Programming*, ser. PPDP '07, Wroclaw, Poland: ACM, 2007, pp. 13–24, isbn: 978-1-59593-769-8. doi: 10.1145/1273920.1273923. [Online]. Available: http://doi.acm.org/10.1145/1273920.1273923 (cit. on p. 5).

[12] B. Hardekopf and C. Lin, "The ant and the grasshopper: Fast and accurate pointer analysis for millions of lines of code," in *Proceedings of the 28th ACM SIGPLAN Conference on Programming Language Design and Implementation*, ser. PLDI '07, San Diego, California, USA: ACM, 2007, pp. 290–299, isbn: 978-1-59593-633-2. doi: 10.1145/1250734.1250767. [Online]. Available: http://doi.acm.org.ezproxy.its.uu.se/10.1145/1250734.1250767 (cit. on pp. 5, 6).

[13] Y. Smaragdakis, M. Bravenboer, and O. Lhoták, "Pick your contexts well: Understanding object-sensitivity," in *Proceedings of the 38th Annual ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages*, ser. POPL '11, Austin, Texas, USA: ACM, 2011, pp. 17–30, isbn: 978-1-4503-0490-0. doi: 10.1145/1926385.1926390. [Online]. Available: http://doi.acm.org/10.1145/1926385.1926390 (cit. on p. 5).

[14] Y. Smaragdakis and M. Bravenboer, "Using datalog for fast and easy program analysis," in *International Datalog 2.0 Workshop*, Springer, 2010, pp. 245–251 (cit. on pp. 5, 6).

[15] G. Balatsouras, K. Ferles, G. Kastrinis, and Y. Smaragdakis, "A datalog model of must-alias analysis," in *Proceedings of the 6th ACM SIGPLAN International Workshop on State Of the Art in Program Analysis*, ACM, 2017, pp. 7–12 (cit. on p. 5).

[16] M. Madsen, M.-H. Yee, and O. Lhoták, "From datalog to flix: A declarative language for fixed points on lattices," in *Proceedings of the 37th ACM SIGPLAN Conference on Programming Language Design and Implementation*, ser. PLDI '16, Santa Barbara, CA, USA: ACM, 2016, pp. 194–208, isbn: 978-1-4503-4261-2. doi: 10.1145/2908080.2908096. [Online]. Available: http://doi.acm.org.ezproxy.its.uu.se/10.1145/2908080.2908096 (cit. on p. 5).

[17] M. Eichberg, S. Kloppenburg, K. Klose, and M. Mezini, "Defining and continuous checking of structural program dependencies," in *Proceedings of the 30th International Conference on Software Engineering*, ser. ICSE '08, Leipzig, Germany: ACM, 2008, pp. 391–400, isbn: 978-1-60558-079-1. doi: 10.1145/1368088.1368142. [Online]. Available: http://doi.acm.org.ezproxy.its.uu.se/10.1145/1368088.1368142 (cit. on p. 5).

[18] B. Hardekopf and C. Lin, "Semi-sparse flow-sensitive pointer analysis," in *Proceedings of the 36th Annual ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages*, ser. POPL '09, Savannah, GA, USA: ACM, 2009, pp. 226–238, isbn: 978-1-60558-379-2. doi: 10.1145/1480881.1480911. [Online]. Available: http://doi.acm.org.ezproxy.its.uu.se/10.1145/1480881.1480911 (cit. on p. 5).

[19] B. Scholz, H. Jordan, P. Subotić, and T. Westmann, "On fast large-scale program analysis in datalog," in *Proceedings of the 25th International Conference on Compiler Construction*, ACM, 2016, pp. 196–206 (cit. on pp. 6, 9).

[20] A. Weiss, D. Patterson, N. D. Matsakis, and A. Ahmed, "Oxide: The essence of rust," Mar. 3, 2019. [Online]. Available: https://arxiv.org/abs/1903.00982v1 (visited on 04/21/2019) (cit. on pp. 6, 7).

[21] Wikipedia contributors, *Datalog — Wikipedia, the free encyclopedia*, https://en.wikipedia.org/w/index.php?title=Datalog&oldid=899388108, [Online; accessed 1-June-2019], 2019 (cit. on p. 10).

[22] H. Q. Ngo, E. Porat, C. Ré, and A. Rudra, "Worst-case optimal join algorithms," in *Proceedings of the 31st ACM SIGMOD-SIGACT-SIGAI symposium on Principles of Database Systems*, ACM, 2012, pp. 37–48 (cit. on p. 10).