

# Optimism in Equality Saturation

RUSSEL ARBORE, University of California, Berkeley, USA

ALVIN CHEUNG, University of California, Berkeley, USA

MAX WILLSEY, University of California, Berkeley, USA

Equality saturation is a technique for program optimization based on non-destructive rewriting and a form of program analysis called e-class analysis. The current form of e-class analysis is pessimistic and therefore ineffective at analyzing cyclic programs, such as those in SSA form. We propose an abstract interpretation algorithm that can precisely analyze cycles during equality saturation. This results in a unified algorithm for optimistic analysis and non-destructive rewriting. We instantiate this approach on a prototype abstract interpreter for SSA programs using a new semantics of SSA. Our prototype can analyze simple example programs more precisely than clang and gcc.

## 1 Introduction

Optimizing compilers transform an input program into a “better” program. Most compilers implement a transformation-based approach—an input program is modified by a sequence of individual passes. A challenge with this approach is that the order of the individual passes matters, because each pass destroys the previously known program (often referred to as the phase ordering problem).

*Equality Saturation.* Equality saturation is an alternate approach to organizing compilation, where transformations are implemented as term rewrites and a data structure called an e-graph keeps track of all intermediate programs and known equivalences between them [47]. This approach enables *non-destructive* rewriting, which bypasses the phase ordering problem. Recent implementations of equality saturation also include program analysis capabilities (called e-class analysis) [51]. Rewriting and analysis cooperate: rewrites can conditionally depend on analysis facts, and equalities from rewriting combine analysis facts into more precise ones [12]. These systems have been applied in many areas, including floating point accuracy [37], circuit synthesis [13], tensor and linear algebra [50, 52], 3D CAD [33], and imperative program compilation [17, 47].

*Optimism.* Many interesting programs contain loops, which typically correspond to cycles in data flow. A subset of program analyses, called “optimistic” analyses in the compilers literature [6, 48], are capable of precisely analyzing cyclic program representations. Other analyses, called “pessimistic” analyses, can only reason about programs inductively, but cyclic programs admit no inductive argument. As a result, pessimistic analyses are typically too conservative for analyzing programs with loops; they often return the least precise result. Optimistic analyses, as their name suggests, will optimistically assume a potentially unsound analysis fact about program fragments in a cycle and then later refine the analysis fact until it can no longer be disproved—this is a co-inductive argument [6]. Temporary unsoundness prevents interleaving rewriting with optimistic analysis; one cannot safely act on the results of optimistic analyses until a fixpoint is reached.

*Optimism in Equality Saturation.* Incorporating optimistic analyses into equality saturation has not been achieved previously. Existing e-class analyses are inherently inductive, meaning cyclic programs cannot be precisely analyzed. Performing optimistic analyses on e-graphs after rewriting is fraught; we show that the straightforward approach leads to unsoundness. Additionally, as optimistic analyses are not incrementally sound, optimistic analyses cannot be interleaved with rewriting in equality saturation. Prior work in equality saturation has identified a specific need for

optimism: de-duplicating isomorphic cycles in an e-graph [47, 57, 58]. A technique from traditional compilers, optimistic global value numbering, would solve this problem [6, 41], but is not applicable due to the aforementioned difficulties.

We incorporate optimistic analyses into an equality saturation system for the first time. Our approach can soundly and precisely analyze cyclic programs. As a demonstration, we prototyped a combined equality saturation engine and abstract interpreter for imperative programs, based on a sea-of-nodes style program representation with a novel semantic formulation. Our prototype can analyze example programs more precisely than clang or gcc. It also presents a solution to the cycle de-duplication problem described above. In summary, our contributions are as follows:

- We identify the core issue preventing optimism in equality saturation—equality saturation creates ill-formed represented graphs which poison optimistic analyses (Section 3).
- We propose a SSA form program representation that 1) can be easily embedded into an e-graph and 2) has a simple semantics amenable to abstract interpretation [9] (Section 4).
- We describe optimistic analyses over SSA e-graphs—discovered equalities make analyses more precise, but also create ill-formed represented graphs. We propose a dataflow analysis algorithm that computes an abstraction that is both precise in the presence of well-formed cycles and sound in the presence of ill-formed cycles (Section 5).
- We build a prototype program optimizer in Rust that combines equality saturation and optimistic analyses and evaluate it on example programs requiring both rewriting and optimistic analysis to fully optimize; the production compilers clang and gcc cannot fully optimize these examples, while our tool can (Sections 6 and 7).

## 2 Background

This paper is concerned with performing abstract interpretation over e-graphs representing SSA programs during equality saturation. We give a brief background on each of these components.

### 2.1 Equality Saturation

Equality saturation is a technique for performing non-destructive rewriting [47]. Several implementations exist, the two most popular being egg [51] and egglog [55].

**2.1.1 E-Graphs.** Equality saturation implementations use *e-graphs* to store terms modulo an equivalence relation. Conceptually, an e-graph consists of two components: a *term bank*, which is a set of terms in some language, and an *equivalence relation*, which identifies which terms in the e-graph are equivalent. Normally, this relation is also a *congruence* relation, which in addition to obeying reflexivity, symmetry, and transitivity, is also closed under congruence. New terms can be *inserted* into the e-graph, which start as dis-equal to other terms. New equalities can be *asserted*, at which point the equivalence relation becomes coarser (fewer, larger equivalence classes). New equalities may imply further equalities due to congruence—congruence is explicitly propagated during *rebuilding*. Finally, a key operation for the equality saturation use-case is *e-matching*, where the term bank is searched for terms matching a syntactic pattern. More formally:

*Definition 2.1 (E-Graphs).* An e-graph is pair of a set of *e-nodes*  $\mathcal{N}$  and *e-classes*  $\mathcal{C}$  where:

- An e-node  $n \in \mathcal{N}$  is a *function symbol*,  $f$ , paired with a list of input e-classes  $c_1, c_2, \dots, c_k \in \mathcal{C}^k$ , where  $k$  is the arity of  $f$ .
- An e-class  $c \in \mathcal{C}$  is a subset of  $\mathcal{N}$  where the e-nodes in the set have been asserted equal.
- The e-classes  $\mathcal{C}$  partition the e-nodes  $\mathcal{N}$ . The e-class of an e-node  $n \in \mathcal{N}$  is written  $[n] \in \mathcal{C}$ . This also defines an equivalence relation  $\equiv$ , where  $n_1 \equiv n_2 \iff [n_1] = [n_2]$ .

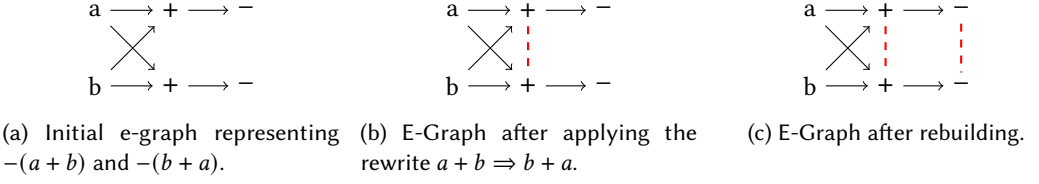


Fig. 1. An example e-graph during equality saturation. Symbols are e-nodes, solid edges connect e-nodes to argument e-classes, and dashed red edges connect e-nodes in the same e-class. (Note that some e-graph literature connect e-nodes  $\rightarrow$  argument e-classes. We go the other way, e-nodes  $\leftarrow$  argument e-classes, indicating the flow of the data, aligning with the convention of drawing SSA graphs.)

The prior definition of “term bank” is informal, and is not mentioned in Definition 2.1. Prior work defines the term bank as the set of “represented terms” in the e-graph [51]. We instead give a definition for *represented graphs*, since we will use e-graphs to represent cyclic terms in this paper (represented terms are simply represented graphs that are also trees).

**Definition 2.2 (Cyclic Terms).** A cyclic term is a graph with nodes  $\mathcal{V}$  where a node  $v \in \mathcal{V}$  is a function symbol,  $f$ , paired with a list of input nodes  $v_1, v_2, \dots, v_k \in \mathcal{V}^k$ , where  $k$  is the arity of  $f$ . We call this a *cyclic term* because the graph may contain cycles (mutually dependent nodes in  $\mathcal{V}$ ).

**Definition 2.3 (Represented Graphs).** A cyclic term  $\mathcal{V}$  is a represented graph of an e-graph  $(\mathcal{N}, \mathcal{C})$  when there is a map  $m \in \mathcal{V} \rightarrow \mathcal{N}$  such that for all nodes  $v \in \mathcal{V}$ ,  $v$  and  $m(v)$  have the same function symbol and  $m(v)_i = [m(v_i)]$  for all  $i \in 1..k$  for arity  $k$  node  $v$ . In other words,  $m$  is a homomorphism from the cyclic term into the e-graph over dependency structure, modulo equivalence. We say that a particular e-class  $c \in \mathcal{C}$  represents a cyclic term  $\mathcal{V}$  if for some  $v \in \mathcal{V}$ ,  $[m(v)] = c$ .

Figure 1 shows three example e-graphs. The e-nodes and e-classes of an e-graph are mutually recursive—e-nodes have e-classes as children rather than other e-nodes, since it doesn’t matter which e-node in an e-class is being referred to (they’re all equivalent). After performing rebuilding, the relation represented by the e-classes is also a congruence relation.

**2.1.2 Batched Rewriting and Rebuilding.** A typical equality saturation workflow has three steps:

- (1) The e-graph is seeded with an initial term (the term we are interested in transforming).
- (2) Rewrites rules are iteratively applied to the e-graph. This may reach a fixpoint (called saturation), but is not guaranteed to—applications often stop rewriting after a timeout [46].
  - (a) Rewrites query the e-graph using the *e-matching* procedure to find terms represented by the e-graph matching the left hand side of the rewrite, like the  $x + 0$  in  $x + 0 \Rightarrow x$ .
  - (b) Matches yield substitutions, which are used to instantiate the right hand side pattern—these new terms are inserted and asserted equal with matched terms.
  - (c) New equalities added by rewrites may imply further equalities by congruence, so *rebuilding* is run to discover these new equalities.
- (3) *Extraction* picks a single term from the e-class containing the input term that is optimized with respect to some metric, such as size or a performance cost model [4, 19, 47, 51, 53].

**2.1.3 E-Class Analysis.** The *e-class analysis* framework adds program analysis capabilities into equality saturation without bespoke extensions, which enables the safe application of rewrite rules that are conditioned on some analysis result [51]. E-class analysis associates a semi-lattice element with each e-class and propagates facts in a bottom-up fashion. egglog uses a Datalog-like architecture to associate multiple semi-lattice facts with each e-class, but the fundamental mechanism is the same [55]. Facts derived for e-nodes inside the same e-class can be combined to

increase precision. Rewrites may depend on analysis results for soundness, and rewrites inflate e-classes, making the combined fact for each e-class more precise.

Consider the following example of an interval e-class analysis from Coward et. al. [12]. Let our language be arithmetic expressions, and the analysis domain be the set of intervals. Consider the following expressions over variables  $x$  and  $y$ , where  $x \in [0, 1]$  and  $y \in [1, 2]$ .

$$\frac{x - y}{x + y} \in [-2, 0] \quad \frac{2x}{x + y} - 1 \in [-1, 1]$$

Since the variables  $x$  and  $y$  have known intervals, the intervals for the terms can be derived as  $[-2, 0]$  and  $[-1, 1]$ . E-class analysis reacts to asserted equalities by combining the facts of the merged e-classes. If rewriting discovers that the two above expressions are equivalent (they are), the e-classes will be merged, and the fact for the new e-class will be their *intersection* (if two equivalent terms have intervals, they both must lie in both intervals). Thus, the new fact for the merged e-class will be  $[-1, 0]$ , which is more precise than either of the original intervals.

**2.1.4 Cycles in E-Graphs.** Rewrites can create cycles in e-graphs when some term is discovered equivalent to one of its sub-terms. Most languages embedded into e-graphs are acyclic, meaning that these cycles do not correspond to valid singular terms in the language, but rather an infinite family of acyclic terms. An important property of e-class analysis is that all intermediate analysis results are sound, meaning the analysis can always be terminated early. This allows analyses of e-graphs containing cycles to always be capable of terminating, even if the analysis around the cycle does not reach a fixpoint (this is a form of narrowing [9, 12]). In this work, we explicitly embed cyclic terms into the e-graph, which requires extending the notion of representation to graphs (Definition 2.3), though this does not mechanically change the e-graph [47].

## 2.2 Abstract Interpretation

Abstract interpretation is a framework for describing approximations of mathematical objects, which usually have lattice structures [9, 10]. Abstract interpretation often refers specifically to the analysis of programs—in this setting, “abstractions” are developed that characterize what possible sets of concrete executions are possible in a particular program [9, 27–31, 40, 42, 49]. Often, abstractions characterize what concrete values a program variable may store at some point during program execution—these are called non-relational abstractions, and include intervals [9] and known-bits [49]. Abstractions can also characterize what concrete values a pair or tuple of program variables may store simultaneously—these are called relational abstractions, and include difference bounds [29], octagons [31, 42], pentagons [28], two variables per inequality [40], and equality [5].

Typically, abstract interpretation is described in terms of a *Galois connection* between a *concrete lattice*  $C$  and an *abstract lattice*  $\mathcal{A}$ , which consists of an abstraction function  $\alpha \in C \rightarrow \mathcal{A}$  and a concretization function  $\gamma \in \mathcal{A} \rightarrow C$ . Given partial order relations  $\sqsubseteq_C$  and  $\sqsubseteq_{\mathcal{A}}$ ,  $\alpha$  and  $\gamma$  form a Galois connection if and only if for all  $c \in C$  and  $a \in \mathcal{A}$ ,  $\alpha(c) \sqsubseteq_{\mathcal{A}} a \iff c \sqsubseteq_C \gamma(a)$ . An abstraction  $a \in \mathcal{A}$  is a *sound* approximation of  $c \in C$  if and only if  $c \sqsubseteq_C \gamma(a)$ , and is a *complete* approximation if and only if  $\gamma(a) \sqsubseteq_C c$ . These definitions are mostly used in proofs— $\gamma$ , for example, is often incomputable ( $\alpha$ , on the other hand, is often computable). Given a function  $f \in C \rightarrow C$ , which can model a step of program semantics given some concrete state, we can lift  $f$  to operate instead on abstract states:  $\alpha \circ f \circ \gamma \in \mathcal{A} \rightarrow \mathcal{A}$ . Since  $\gamma$  is often incomputable, we say an *abstract transformer*  $g \in \mathcal{A} \rightarrow \mathcal{A}$  soundly approximates  $f$  if and only if for all  $a \in \mathcal{A}$ ,  $f(\gamma(a)) \sqsubseteq_C \gamma(g(a))$ .

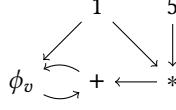
An abstract interpretation can be expressed as a series of equations describing the abstract states at program locations in terms of abstract transformers applied to abstract states at predecessor program locations (sometimes called *dataflow* equations). When programs contain cycles in their control flow, the equations are cyclically dependent—a fixpoint operator of some kind can be used

```

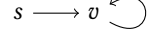
1 x = 1;
2 while 1 {
3   x = x + (1 * 5);
4 }

```

(a) A simple program with a loop.



(b) SSA graph of the program.



(c) CFG of the program.

Fig. 2. A program in pseudo-code and its compilation into a SSA program.

to find a solution to the equations [18]. Assuming the involved abstract transformers are sound with respect to the program operators, any fixpoint solution to the equations is sound with respect to the program semantics [9]. Other conservative program analysis frameworks, such as monotone analysis frameworks, can be described as fixpoints of abstract equations in a similar manner [21]. Most prior work in abstract interpretation is concerned with computing a *least* fixpoint, as this is the fixpoint that is the most precise approximation of the concrete program semantics (in contrast to the *greatest* fixpoint, which is a less precise over-approximation). In the monotone analysis frameworks literature, the equivalent terms are “optimistic” and “pessimistic” analyses, respectively. It has been shown that these analysis results only differ when programs contain loops [48]. Additionally, while optimistic analyses produce precise results when programs contain loops, they are not *incrementally* sound—that is, only the final analysis result is sound, but intermediate analysis results are not sound. This is in contrast to a pessimistic analysis, where all intermediate analysis results are sound, but loops cannot be analyzed precisely [6, 48]. These results naturally extend to the least and greatest fixpoints referred to in the abstract interpretation literature.

### 2.3 Single Static Assignment Form

*Single static assignment* (SSA) form is a category of program representation where 1) every variable has a single definition and 2) every definition is always executed before any of its uses (which is called *dominance*). In fact, variables in SSA form are often instead called *values* to emphasize that they do not change and are unambiguously defined. A special  $\phi$  instruction is used to join data flow at control flow points[38]. Some SSA form representations (such as the sea-of-nodes) define data flow *graphs*, where nodes represent operators and edges represent dependencies, rather than instruction lists—SSA values are identified by nodes, and  $\phi$  nodes are used to join results at control flow points [7, 15]. Figure 2 shows an example program, a corresponding SSA graph (sea-of-nodes style), and a corresponding control flow graph (CFG).

For sea-of-nodes style representations, not all data flow graphs are in SSA form. The first condition cannot be violated—values are uniquely identified by nodes in the graph and edges unambiguously represent dependencies. The second condition *can* be violated—if there is a dependency cycle consisting of non- $\phi$  operations in a graph, there is no order in which the operations can be evaluated to concretely evaluate the program (in every serialization of the cycle, at least one definition will not dominate all of its uses). A program that is in SSA form can be referred to as being “well-formed”. Typically, analyses and transformations inside compilers using SSA-based representations assert well-formed-ness as both a pre-condition and a post-condition.

## 3 Challenges in Abstract Interpretation over E-Graphs

In this section, we describe in detail why performing precise abstract interpretation over e-graphs, specifically in the presence of cycles, is challenging. We identify a key problem inherent to embedding a cyclic program representation into e-graphs and discuss possible solutions to this problem.

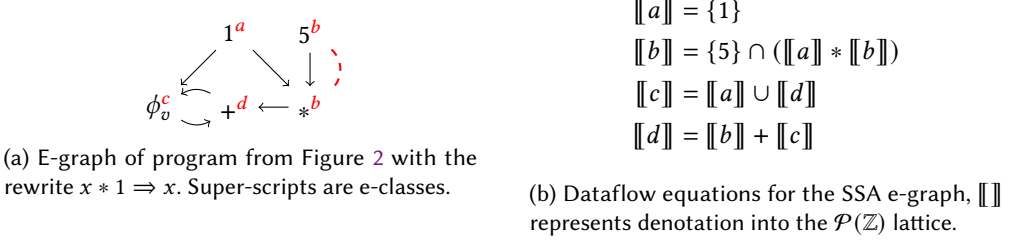


Fig. 3. Flow insensitive dataflow analysis (with the  $\mathcal{P}(\mathbb{Z})$  lattice) of the program from Figure 2.

### 3.1 The Problem

Recall from Section 2.2 that we can express an analysis of a program graph as a set of equations over an abstract lattice. We can express an analysis over an e-graph in a similar fashion. Analysis facts are derived for e-classes, rather than individual nodes, since we know that all nodes in the same e-class evaluate to the same concrete value. The fact derived for an e-class is the meet of the facts derived for the constituent e-nodes [12]. Figure 3 shows an e-graph and its corresponding dataflow equations over the  $\mathcal{P}(\mathbb{Z})$  lattice.

A nice property of dataflow equations, in normal abstract interpretation, is that *every* solution to the equations is a *sound* approximation of the semantics of the represented program. Thus, much prior work in abstract interpretation computes a *least* fixpoint, which is the most precise solution to the equations (or in presence of widening, some relatively small fixpoint). Unfortunately, this assumption does not hold as cleanly in the case of e-graphs.

Let us consider solutions to the equations shown in Figure 3b. There are three distinct solutions:

- (1)  $\llbracket a \rrbracket = \{1\}, \llbracket b \rrbracket = \{5\}, \llbracket c \rrbracket = \mathbb{Z}, \llbracket d \rrbracket = \mathbb{Z}$
- (2)  $\llbracket a \rrbracket = \{1\}, \llbracket b \rrbracket = \{5\}, \llbracket c \rrbracket = \{1 + 5z \mid z \in \mathbb{Z}_{\geq 0}\}, \llbracket d \rrbracket = \{1 + 5z \mid z \in \mathbb{Z}_{> 0}\}$
- (3)  $\llbracket a \rrbracket = \{1\}, \llbracket b \rrbracket = \emptyset, \llbracket c \rrbracket = \{1\}, \llbracket d \rrbracket = \emptyset$

Intuitively, solution #2 is the one we “want”—it correctly captures that the induction variable in the loop is never negative, and it correctly captures that  $\llbracket b \rrbracket = \{5\}$ . However, this solution is neither the greatest nor the least fixpoint of the equations (those would be solutions #1 and #3, respectively). Solution #1 is imprecise with respect to the loop in the original program—nothing of note can be derived for the induction variable (this is because the greatest fix point “gets stuck” at  $\llbracket c \rrbracket = \llbracket d \rrbracket = \mathbb{Z}$ ). Solution #3 is unsound—during the execution of the program, the induction variable will be equal to 6 after one loop iteration, but  $6 \notin \llbracket c \rrbracket = \{1\}$ .

We argue that this situation arises from the two cycles that appear in the e-graph—the cycle containing the  $\phi_v$  and  $+$  nodes and the cycle containing the  $*$  node (the cycle is formed by one input being the  $b$  e-class, which is the e-class of the  $*$  node). In Section 4.2, we describe the notion of a “well-formed” SSA graph, which is a SSA graph that can be given a concrete semantics. In short, the first cycle is well-formed, while the second is not. The first cycle contains a  $\phi$  node and corresponds to a control flow loop in the original program. The second cycle corresponds to the cyclic represented graph  $x = 0 + x$ , which cannot be given a semantics in a normal SSA representation (it violates dominance). We are only interested in abstracting *well-formed* represented graphs, not all represented graphs. However, the dataflow equations do not encode well-formed-ness, and their incorporation of the ill-formed represented graph  $x = 0 + x$  proves problematic. Additionally, the ill-formed cycle represents a common case in equality saturation—rewrite rules often equate a term to one of its sub-terms. Assuming the original SSA graph was well-formed, these ill-formed cycles



only appear once rewrite rules have been applied. This is the core issue frustrating a combination of optimistic analyses (a least fixpoint solution to the dataflow equations) and equality saturation.

### 3.2 Addressing the Problem

We identify four possible approaches to addressing the aforementioned problem:

- (1) Always use the greatest fixpoint solution of the equations.
- (2) “Fix” the equations so that all solutions are sound.
- (3) Remove ill-formed cycles from the e-graph.
- (4) Design an algorithm that computes a sound and precise (not always greatest) solution.

The first solution is what current e-class analysis techniques are based on [12, 51, 55]. E-class analysis computes an incrementally sound set of known facts until no new facts can be derived inductively. This results in computing a greatest fixpoint solution. When there are no cycles in a program, the greatest and least fixpoint solutions are always the same [48]. However, in the presence of cycles, the greatest fixpoint solution may be imprecise. In particular, when well-formed cycles are present in an SSA graph, the greatest fixpoint solution is often not the most precise sound solution (as is the case for the equations in Figure 3b).

The second solution is tempting—the equations ought to model the semantics of the program, and any solution of the equations ought to be sound. However, we believe this is challenging during equality saturation. Consider the e-class  $b$  in Figure 3a. An equation representing the semantics of  $b$  should represent the intersection of *well-formed* represented graphs of  $b$ . Since  $b$  represents an infinite number of well-formed terms, the equation for  $b$  either 1) has to be infinitely large or 2) be cyclic. The first option is not finitely representable and the second option admits the ill-formed cyclic term  $x = 1 * x$ , which is what caused the least fixpoint to be unsound in the first place.

The third solution forces every represented graph to be well-formed by breaking cycles that admit ill-formed represented graphs. This implies that the equations characterizing the e-graph have only sound solutions. A similar approach has been proposed in prior work for extraction [52]. Breaking cycles requires removing an e-node along each cycle. This explicitly removes represented graphs from the e-graph, which may hurt rewriting or analysis precision. Breaking cycles requires care—an e-class can only have an e-node removed if it has at least two e-nodes (otherwise, empty e-classes may be queried during extraction but have no e-node to give).

In this paper, we explore the fourth solution. We accept that the dataflow equations for an SSA e-graph admits unsound solutions (with respect to the semantics of well-formed represented graphs of the e-graph). We describe an algorithm that computes a solution to the dataflow equations that is also sound in Section 5. The algorithm is relatively simple and takes inspiration from classic dataflow analysis algorithms [1]. Unlike the first and third solutions, this solution does not sacrifice analysis precision due to the presence of cycles.

## 4 Semantics and Abstract Interpretation of SSA Programs

We describe a simple semantics for a program representation consisting of 1) a SSA graph consisting of data operations,  $\phi$  nodes, and their dependencies, and 2) a CFG. This representation is most similar to the sea-of-nodes representation [7, 15], however, to our knowledge the formalization of its semantics is novel. It consists of an operational component for describing possible walks through the CFG and a denotational component for evaluating nodes in the SSA graph, given a particular control flow walk (rather than given a particular evaluation for  $\phi$  nodes, as in prior work [15, 24]). We develop this representation and its semantics specifically for performing abstract interpretations on it and embedding the data flow portion of the representation into an e-graph

in Section 5. We discuss how flow insensitive or flow sensitive analyses can both be described as abstractions of the proposed semantics.

#### 4.1 Syntax of SSA Programs

An SSA program is a pair  $(\mathcal{S}, \mathcal{G})$  of a SSA graph  $\mathcal{S}$  and a CFG  $\mathcal{G}$ . An SSA graph  $\mathcal{S} = (N, L, I)$  consists of a set of nodes  $N$ , a labeling function  $L \in N \rightarrow F$  that maps each node to a function symbol from the set of function symbols  $F$ , and a input function  $I \in N \rightarrow N^*$  that maps each node to the ordered list of its input nodes ( $\forall n \in N$ , the arity of  $L(n)$  must equal  $|I(n)|$ ). The set of function symbols  $F$  contains both data operations and  $\phi$  operations.  $\phi$  operations are parameterized by a vertex in the CFG, described next (that is,  $\forall v \in V \setminus s, \phi_v \in F$ ).

A CFG  $\mathcal{G} = (V, s, P, C)$  is a graph with vertices  $V$ , a distinguished entry node  $s \in V$ , a predecessor function  $P \in V \rightarrow V^*$  that maps each vertex to an ordered list of its predecessor vertices ( $|P(s)| = 0$ ), and a condition function  $C \in V \rightarrow N^*$  that maps each vertex to an ordered list of SSA graph nodes representing whether the corresponding control flow edge can be traversed ( $\forall v \in V, |P(v)| = |C(v)|$ ). We label control flow edges with predicates, rather than use branch instructions or nodes, because it makes specifying the semantics more straightforward (for deterministic programs, these two forms are equally as expressive).

The SSA graph and the CFG of a program are interconnected. The condition function  $C$  determines what control flow walks are possible based on evaluating nodes in the SSA graph, and the  $\phi_v \in F$  function symbol is parameterized by a block  $v \in V \setminus s$ . No  $\phi$  operation is parameterized by  $s$ , since this vertex has no predecessors. This interaction is standard in sea-of-nodes style IRs [7, 15].

#### 4.2 Concrete Semantics

We describe the semantics of SSA programs in two components: an operational component for describing possible walks through the CFG and a denotational component for describing the possible domain values of each node in the SSA graph for each possible walk in the CFG. We assume the data operations in the SSA graph operate on some domain  $\mathcal{D}$ . We additionally assume some  $\mathcal{T} \subseteq \mathcal{D}$  which correspond to “true” values in the domain. A walk  $W$  in a CFG  $\mathcal{G} = (V, s, P, C)$  is a finite sequence of vertices  $v_1, v_2, \dots, v_k \in V^k$  such that  $v_1 = s$  and  $\forall i \in 2..k, v_{i-1} \in P(v_i)$ . Note that a walk may visit the same vertex multiple times. The set of all walks in a CFG is  $\mathcal{W}$ .

Given a SSA graph  $\mathcal{S} = (N, L, I)$ , we define the denotation of individual nodes. The denotation of a node is a partial function from CFG walks to sets of domain values ( $\llbracket \cdot \rrbracket \in N \rightarrow (\mathcal{W} \rightarrow \mathcal{P}(\mathcal{D}))$ )<sup>1</sup>. The domain of the semantics is only a partial function because  $\phi$  nodes whose vertex has not been visited in a walk yet do not have a defined value.

The operational semantics of the CFG describes what walks are *possible*, that is which walks correspond to an execution of the SSA program. We define a relation  $\rightarrow_{\text{WALK}}$ , written  $W \rightarrow_{\text{WALK}} W'$ , which states that if  $W$  is possible, then  $W'$  is possible.

**Definition 4.1** ( $\rightarrow_{\text{WALK}}$ ). Given  $(W; v) \in \mathcal{W}$ ,  $v' \in V$ , then  $W \rightarrow_{\text{WALK}} (W'; v')$  if and only if:

- (1)  $v \in P(v')$ .
- (2)  $\llbracket C(v')_k \rrbracket (W; v) \in \mathcal{T}$ , where  $v = P(v')_k$ .

Intuitively, this definition means that a control flow walk can progress to another control flow walk if that means taking a single step in the CFG and the predicate on the taken edge evaluates to a “true” value. We define the reflexive and transitive closure of  $\rightarrow_{\text{WALK}}$  as  $\rightarrow_{\text{WALK}}^*$ . The set of possible

<sup>1</sup>The only 0-arity node that may denote to a non-singleton set is a function parameter node, which denotes to  $\mathcal{D}$  for all walks. Data operations on domain values are lifted to operating on sets of domain values.



walks over the CFG is  $\mathcal{W}_{\mathcal{P}} = \{W \in \mathcal{W} \mid [s] \rightarrow_{\text{WALK}}^* W\}$ , which is the set of walks that are possible starting from the entry vertex.

Next, we describe the denotational semantics of SSA graph nodes. The denotation of a SSA node is a partial function from a control flow walk to a domain value.

**Definition 4.2 (Denotation of SSA Nodes).** The denotation of a SSA node  $n \in N$  is given by  $\llbracket n \rrbracket \in \mathcal{W} \rightarrow \mathcal{P}(\mathcal{D})$ , and is defined as:

$$\llbracket n \rrbracket(W; v) = \begin{cases} \llbracket n \rrbracket(W) & L(n) = \phi_{v'} \wedge v' \neq v \\ \llbracket I(n)_k \rrbracket(W) & L(n) = \phi_v \wedge W = (W'; P(v)_k) \\ \{L(n)(i_1, \dots, i_{|I(n)|}) \mid i_k \in \llbracket I(n)_k \rrbracket(W; v)\} & L(n) \in \mathcal{D}^* \rightarrow \mathcal{D} \end{cases}$$

Note that this function is only well defined for certain SSA graphs. For example, consider the SSA graph  $(\{x\}, [x \rightarrow +], [x \rightarrow [x, x]])$ , which represents the cyclic term  $x = x + x$ . The node  $x$  does not have a well defined semantics—according to Definition 4.2,  $\llbracket x \rrbracket(W) = \llbracket x \rrbracket(W) + \llbracket x \rrbracket(W)$ , which is not well defined. We can restrict which SSA graphs admitted by the syntax are *well-formed*.

**Definition 4.3 (Well-formed SSA Programs).** A SSA program  $(\mathcal{S}, \mathcal{G})$  is well-formed if and only if:

- (1) Every cycle in  $\mathcal{S} = (N, L, I)$  (formed by nodes from  $N$  and edges from  $I$ ) contains at least one node labeled by  $L$  with a  $\phi$  operation.
- (2) For every node  $n \in N$  where  $L(n) = \phi_v$  and for every input index  $1 \leq i \leq |I(n)|$ ,
  - (a) There is exactly one  $v' \in V$  such that  $P(v)_i = v'$ .
  - (b) For all walks  $W \in \mathcal{W}$ ,  $v' \in W \implies \Phi_{\text{pred}}(I(n)_i) \subseteq W$ .
  - (c) For all walks  $W \in \mathcal{W}$ ,  $v' \in W \implies \Phi_{\text{pred}}(C(v)_i) \subseteq W$ .

where:

$$\Phi_{\text{pred}}(n) = \begin{cases} \{v\} & L(n) = \phi_v \\ \bigcup_i \Phi_{\text{pred}}(I(n)_i) & L(n) \in \mathcal{D}^* \rightarrow \mathcal{D} \end{cases}$$

The first condition enforces that data operations can always be evaluated in some order, given values for predecessor  $\phi$ s. The second condition is the classic strictness condition of SSA form [38], which states that every value is defined before it is used along every control flow walk<sup>2</sup>. Given a well-formed SSA program and some control flow walk, the semantics of a node can be easily evaluated (or shown non-existent<sup>3</sup>)—it is always well defined. In fact, it is computable for any walk in  $\mathcal{W}$ , not just possible walks in  $\mathcal{W}_{\mathcal{P}}$ . This is convenient, since  $\mathcal{W}_{\mathcal{P}}$  itself depends on the semantics of a SSA program (which is what we are trying to define), while  $\mathcal{W}$  just depends on the syntax.

### 4.3 Abstraction of Concrete Semantics

Next, we define abstract interpretations for SSA programs. An abstraction of the concrete semantics of a SSA program will over-approximate the values from  $\mathcal{D}$  a node may evaluate to on any possible control flow walk. We define abstract interpretations over SSA programs as follows:

**Definition 4.4 (Abstract Interpretation of SSA Programs).** An abstract interpretation of a SSA program is a tuple  $\mathcal{A} = (\Sigma, \gamma, \perp, \mathcal{F}, \sqcup, \nabla)$ , where:

- $\Sigma$  is the set of abstract values.
- $\gamma \in \Sigma \rightarrow (\mathcal{W}_{\mathcal{P}} \rightarrow \mathcal{P}(\mathcal{D}))$  is the concretization function.
- $\perp \in \Sigma$  is a distinguished element where  $\gamma(\perp)(W) = \emptyset$ .

<sup>2</sup>More precisely, that all  $\phi$  nodes that a  $\phi$  node immediately transitively depends on have had their vertices walked over before the  $\phi$  node is evaluated.

<sup>3</sup>If a  $\llbracket n \rrbracket$  invocation inductively evaluates a  $\phi$  node on the walk  $[s]$ , then the node  $n$  is not defined on the original walk.

- $\mathcal{F} \in (F \setminus \{\phi_v | v \in V\}) \rightarrow (\Sigma^* \rightarrow \Sigma)$  is the transfer function, mapping data operations on the concrete semantics into abstract transformers.
- $\sqcup \in \mathcal{P}(\Sigma) \rightarrow \Sigma$  is a join operator over  $\Sigma$ , computing an over-approximation of the union of the concretizations of a set of abstractions.
- $\nabla \in \Sigma \times \Sigma \rightarrow \Sigma$  is a widening operator used to ensure termination.

An abstract interpretation is *sound* when:

- $\{f(i_1, i_2, \dots, i_{|s|}) | i_k \in \gamma(s_k)(W)\} \subseteq \gamma(\mathcal{F}(f)(s)(W))$  for all data operations  $f$ , abstract inputs  $s \in \Sigma^*$ , and walks  $W$  (abstract transformers over-approximate the concrete semantics).
- $\bigcup_{s \in S} \gamma(s) \subseteq \gamma(\sqcup(S))$  for all  $S \subseteq \Sigma$  (joining abstract values over-approximates taking the union of concrete states).
- $\gamma(s_1) \cup \gamma(s_2) \subseteq \gamma(s_1 \nabla s_2)$  (same condition for widening).

An abstract interpretation can be computed as follows:

- $V_\nabla \subseteq V$  is a set of widening points in  $\mathcal{G}$  (every cycle in  $\mathcal{G}$  contains at least one vertex in  $V_\nabla$ ).
- $\mathcal{F}_S \in (N \rightarrow \Sigma) \rightarrow (N \rightarrow \Sigma)$  is the main iteration function, defined as:

$$\mathcal{F}_S(\sigma) = [n \mapsto \begin{cases} \sqcup_i \sigma(I(n)_i) & L(n) = \phi_v \wedge v \notin V_\nabla \\ \sigma(n) \nabla (\sqcup_i \sigma(I(n)_i)) & L(n) = \phi_v \wedge v \in V_\nabla \\ \mathcal{F}(L(n))(i_1, i_2, \dots, i_{|I(n)|}) & L(n) \in \mathcal{D}^* \rightarrow \mathcal{D} \wedge i_k = \sigma(I(n)_k) \end{cases}, \forall n \in N]$$

- If  $\sigma_0 = [n \mapsto \perp, \forall n \in N]$  and  $\sigma_{i+1} = \mathcal{F}'_{S_\#}(\sigma_i)$ , then the final analysis result is  $\sigma_\infty$ .

Note that the co-domain of  $\gamma$  is not the same as the co-domain of  $\llbracket \cdot \rrbracket$  ( $\mathcal{W}_\mathcal{P} \rightarrow \mathcal{P}(\mathcal{D})$  vs.  $\mathcal{W} \rightarrow \mathcal{P}(\mathcal{D})$ ). The former only defines possible values for possible control flow walks, while the latter is defined over all control flow walks. We are only interested in abstracting over possible control flow walks, so we can treat the domain of the concrete semantics of a node as the set of possible control flow walks. An abstraction  $\Sigma$  of a node  $n$  is sound if  $\forall W \in \mathcal{W}_\mathcal{P} \cap \text{dom}(\llbracket n \rrbracket), \llbracket n \rrbracket(W) \subseteq \gamma(\Sigma)(W)$ .

#### 4.4 Flow Sensitivity in SSA Graphs

Non-relational abstractions are usually thought of as objects that abstract a set of concrete values, and a separate object is calculated for different program points and different variables. However, in our SSA graph representation, every value is *global* to the entire function, and the fact that a node may evaluate differently on different control flow walks is lifted into the semantics (this is why the co-domain of  $\llbracket \cdot \rrbracket$  is  $\mathcal{W} \rightarrow \mathcal{P}(\mathcal{D})$ , rather than just  $\mathcal{P}(\mathcal{D})$ ). Thus, abstract objects must not only abstract sets of concrete values, but sets of concrete values per control flow walk. Given an abstract domain  $\mathcal{D}_\mathcal{A}$ , which abstracts sets of concrete values, there are *at least* three ways we can build  $\Sigma$ :

- (1)  $\Sigma = \mathcal{D}_\mathcal{A}$ , meaning an abstract object is a single abstraction of concrete values that is sound on every possible walk. This is often called a *flow insensitive* analysis.
- (2)  $\Sigma = V \rightarrow \mathcal{D}_\mathcal{A}$ , meaning an abstract object is an abstraction of concrete values per vertex a possible walk can *end* in. This is often called a *flow sensitive* analysis.
- (3)  $\Sigma = V^k \rightarrow \mathcal{D}_\mathcal{A}$ , meaning an abstract object is an abstraction of concrete values per *suffix* of vertices in a possible walk (up to some maximum length  $k$ ).

Note that none of these options store an abstraction of concrete values per possible walk (which would correspond to  $\Sigma = \mathcal{W}_\mathcal{P} \rightarrow \mathcal{D}_\mathcal{A}$ <sup>4</sup>), since  $\mathcal{W}_\mathcal{P}$  is, in general, infinite.

Abstract interpretations over SSA programs can be computed easily via a dataflow analysis computing a fixpoint [9]. A standard technique in abstract interpretation is to compute widening points by determining a weak topological order (WTO) [3] over the CFG. We mark all  $\phi$  nodes

<sup>4</sup>This is the required domain to calculate a meet-over-paths solution, which is in general incomputable [21].

whose block is a component head in the aforementioned WTO as widening points. The important property of the WTO of a CFG is that the WTO identifies a subset of vertices in the CFG such that all cycles contain at least one node in this subset (called the component heads). All data flow cycles in a well-formed SSA graph go through  $\phi$  nodes, and by strictness, those cycles always correspond to cycles in the CFG. Thus, every cycle in the SSA graph contains a  $\phi$  node whose CFG vertex is marked as the head of some component of the WTO.

## 5 Abstract Interpretation over SSA E-Graphs

SSA graphs can be converted into e-graphs, enabling the efficient representation of equivalences in a SSA program. By treating the equivalence relation as a relational abstraction, we can combine equivalence with other abstractions to improve precision. We describe how sound and relatively precise abstract interpretations can be performed over SSA programs stored in e-graphs—in particular, our dataflow analysis algorithm can analyze well-formed cycles with more precision than a greatest fixpoint solution, while not letting ill-formed cycles poison the solution.

### 5.1 SSA E-Graphs

Recall that SSA programs consist of a SSA graph  $\mathcal{S}$  as well as a CFG  $\mathcal{G}$ . We are concerned with representing equivalences over values in the program, which are embodied by  $\mathcal{S}$ . Thus, we will modify  $\mathcal{S}$  to store an equivalence relation over nodes. More specifically, a SSA e-graph  $\mathcal{S}_\equiv = (N, L, I, \equiv)$  consists of a set of nodes  $N$ , a labeling function  $L \in N \rightarrow F$ , a input function  $I \in N \rightarrow (N/\equiv)^*$ , and an equivalence relation  $\equiv$  over  $N$ , which is often also a congruence relation. Note that the inputs to nodes are now equivalence classes  $(N/\equiv)$  rather than nodes directly—this is because every node in the class is known to be equivalent, so it doesn't matter which one is used as the input. The nodes  $N$  form the e-nodes of the e-graph, while  $N/\equiv$  forms the e-classes. A compatible CFG  $\mathcal{G}_\equiv = (V, s, P, C)$  is slightly different from the definition given in Section 4.1, specifically  $C \in V \rightarrow (N/\equiv)^*$ . An equivalence relation  $\equiv$  is *sound* over a SSA program if for every pair of nodes  $n_1$  and  $n_2$  that are equivalent under  $\equiv$ ,  $\llbracket n_1 \rrbracket$  and  $\llbracket n_2 \rrbracket$  are defined over the same set of possible walks and  $\llbracket n_1 \rrbracket(W) = \llbracket n_2 \rrbracket(W)$  for all walks  $W \in \mathcal{W}_{\mathcal{P}} \cap \text{dom}(\llbracket n_1 \rrbracket)$ .

A SSA e-graph does not directly have semantics—rather, the represented graphs (Definition 2.3) of an e-graph are SSA graphs that may have semantics. However, not all represented graphs of an e-graph are well-formed SSA graphs (as shown in Figure 4a). We are only interested in the semantics of *well-formed* represented graphs of the SSA e-graph (and thus, we are only interested in abstracting well-formed represented graphs). If two well-formed represented graphs each have a node that maps to the same e-class in the SSA e-graph and the equivalence relation is sound, then those two nodes have the same semantics on all possible walks. We will assume that the equivalence relation stored in a SSA e-graph is always sound.

### 5.2 E-Graphs Store a Relational Abstraction

A SSA e-graph stores an equivalence relation  $\equiv$ , which constrains the possible concrete value pairs of SSA nodes may evaluate to along any given control flow walk. This constraint represents a *relational* abstraction, which instead of over-approximating the set of concrete values a single node may evaluate to, over-approximates the set of pairs of concrete values a pair of nodes may evaluate to simultaneously. More concretely, for a relational abstraction, the concretization function  $\gamma \in \Sigma \rightarrow (\mathcal{W}_{\mathcal{P}} \rightarrow \mathcal{P}(\mathcal{D} \times \mathcal{D}))$  returns a function from control flow walk to possible pairs of domain values for the pair of nodes the abstraction describes. Additionally, the transfer function  $\mathcal{F} \in (F \setminus \{\phi_v | v \in V\})^2 \rightarrow (\Sigma^* \rightarrow \Sigma)$  maps every pair of data operations to an abstract transformer. The final result of the abstract interpretation  $\sigma_\infty \in (N \times N) \rightarrow \Sigma$  assigns abstractions to pairs of nodes. The join and widen signatures remain the same.

*Definition 5.1 (Equivalence Abstraction).* The equivalence relational abstraction can be defined as:

- $\Sigma_{\equiv} = \{\perp, \top\}$
- $\gamma_{\equiv}(\perp) = [W \mapsto \{(v, v) | v \in \mathcal{D}\}, \forall W \in \mathcal{W}_{\mathcal{P}}]$
- $\gamma_{\equiv}(\top) = [W \mapsto \mathcal{D} \times \mathcal{D}, \forall W \in \mathcal{W}_{\mathcal{P}}]$
- $\mathcal{F}_{\equiv}(f, f) = [(\perp)^* \mapsto \perp, x \mapsto \top, \forall x \in \Sigma^*, x \neq (\perp)^*]$  (equal inputs to a function imply equal output).
- $\mathcal{F}_{\equiv}(f_1, f_2) = [x \mapsto \top, \forall x \in \Sigma^*]$  (when  $f_1 \neq f_2$ ).
- $\sqcup_{\equiv}(S) = \begin{cases} \top & \top \in S \\ \perp & \text{otherwise} \end{cases}$
- $s_1 \nabla_{\equiv} s_2 = s_2$

### 5.3 Reduced Product with Other Abstractions

Abstract interpretations can be combined via product operators to gain additional precision (in a combined analysis), relative to running the analyses separately [9, 10]. Several product operators exist [8], the classic example being the reduced product [10]. The general gist of product operators is to over-approximate the intersection of the concretizations of multiple abstractions, given that multiple abstractions are separately known. The reduced product in particular will iteratively improve individual abstractions using other known abstractions—while the total concretization is not allowed to shrink, the individual abstractions become more precise, which is useful in practice.

Combining non-relational abstractions with the equivalence abstraction is relatively straightforward. We require that in addition to a join operator, a non-relational abstraction additionally defines a *meet* operator  $\sqcap \in \mathcal{P}(\Sigma) \rightarrow \Sigma$ , which computes an over-approximation of the intersection of input abstractions (this operator is sound when  $\bigcap_{s \in S} \gamma(s) \subseteq \gamma(\sqcap(S))$  for all  $S \subseteq \Sigma$ ). Additionally,  $\sqcap(\emptyset) = \top$ . To utilize equivalences in the non-relational analysis, we can create a modified iteration function  $\mathcal{F}'_{\mathcal{S}}(\sigma) = [n \mapsto \sqcap_{m \equiv n} \mathcal{F}_{\mathcal{S}}(\sigma)(m), \forall n \in N]$  to incorporate equivalence information. At each step, this iteration function takes the meet of all derived abstractions of all nodes in each equivalence class and assigns that abstraction to all of the nodes in each equivalence class. Assuming the original abstraction is sound, this combined abstraction is sound.

Alternatively, we can lift abstract interpretations from assigning abstractions to nodes to assigning abstractions to equivalence classes. In an e-graph, all nodes in an equivalence class are known to be equal—thus, we can just derive abstractions per class, rather than per node. More specifically:

*Definition 5.2 (E-Lifted Abstract Interpretations).* We can execute an e-lifted analysis as follows:

- $\mathcal{F}_{\mathcal{S}_{\equiv}} \in ((N/\equiv) \rightarrow \Sigma) \rightarrow (N \rightarrow \Sigma)$  is the node iteration function, defined as:

$$\mathcal{F}_{\mathcal{S}_{\equiv}}(\sigma) = \mathcal{F}_{\mathcal{S}}([n \mapsto \sigma([n]), \forall n \in N])$$

- $\mathcal{F}'_{\mathcal{S}_{\equiv}} \in ((N/\equiv) \rightarrow \Sigma) \rightarrow ((N/\equiv) \rightarrow \Sigma)$  is the class iteration function, defined as:

$$\mathcal{F}'_{\mathcal{S}_{\equiv}}(\sigma) = [c \mapsto \sqcap_{n \in c} \mathcal{F}_{\mathcal{S}_{\equiv}}(\sigma)(n), \forall c \in N/\equiv]$$

- If  $\sigma_{\equiv_0} = [c \mapsto \perp, \forall c \in N/\equiv]$  and  $\sigma_{\equiv_{i+1}} = \mathcal{F}'_{\mathcal{S}_{\equiv}}(\sigma_{\equiv_i})$ , then the final analysis result is  $\sigma_{\equiv_{\infty}}$ .

This approach has the practical advantage of not requiring intermediate storage of abstractions per node, but rather just per equivalence class. The ability for equivalence classes to summarize the analysis information of contained nodes is a central assumption of e-class analysis [12, 51] and has been discovered in prior work on relational abstractions [27], where this technique is referred to as “map factorization”. An interesting property of e-graphs is that not only are analysis results factorized by an equivalence relation, but the program representation itself is also factorized by the equivalence relation—in recent work using e-graphs (and in our formulation), e-nodes refer to e-classes as children, rather than other e-nodes [51].

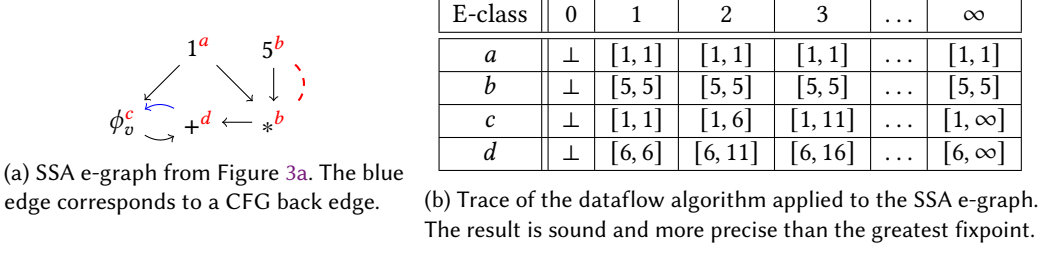


Fig. 4. A trace of the dataflow algorithm performing a flow insensitive interval analysis on the running example program (from Figure 2a).

#### 5.4 A Dataflow Algorithm for Abstracting SSA E-Graphs

An unfortunate property of SSA e-graphs is that not every represented graph will be well-formed. Consider the e-graph shown in Figure 4a. The represented graph consisting of the e-nodes  $*^b$  and  $1^a$  is ill-formed, as it represents the cyclic term  $x = 1 * x$ . Thus, if we were to compute a least fixpoint solution to the dataflow equations of the e-graph, we would arrive at a solution that is unsound with respect to well-formed represented graphs (as described in Section 3), meaning the previous algorithm (Definition 5.2) does not work for e-graphs containing ill-formed cycles.

We propose an algorithm that computes a particular solution of the dataflow equations. This solution is sound and is more precise than the greatest fixpoint (it can analyze well-formed cycles in the SSA e-graph relatively precisely). The general idea is to identify the well-formed cycles in a SSA e-graph and treat them specially. All well-formed cycles contain a  $\phi$  node whose vertex is in a corresponding cycle in the CFG. Thus, we compute a set of control flow edges such that every cycle in the CFG contains an edge from that set. This is an identical requirement as finding widening points in a CFG [9]. We compute a WTO over the CFG, and the set of edges of interest (which we call “back” edges) are edges from vertices inside WTO components to their respective component head vertices (for reducible CFGs, these are back edges in loop nests [3]).

*Definition 5.3 (Abstract Interpretation of SSA E-Graphs).* Given an abstract interpretation  $\mathcal{A} = (\Sigma, \gamma, \perp, \top, \mathcal{F}, \sqcup, \sqcap, \nabla)$ , we can compute a sound abstraction of a SSA program  $\mathcal{S}_{\equiv} = (N, L, I, \equiv)$ ,  $\mathcal{G}_{\equiv} = (V, s, P, C)$  as follows:

- $B \subset V \times \mathbb{N}$  is a set of back edges (edges from vertices inside a component of a WTO to that component’s head, identified by a vertex and predecessor index).
- $\mathcal{DF} \in ((N/\equiv) \rightarrow \Sigma) \rightarrow ((N/\equiv) \rightarrow \Sigma) \rightarrow ((N/\equiv) \rightarrow \Sigma)$  is the “inner” iteration function, which reads the previous “outer” iteration’s abstraction to shrink the current outer iteration’s

abstraction, defined as:

$$\mathcal{DF}(\sigma_{\text{old}}, \sigma) = [c \in (N/\equiv) \mapsto \sqcap_{n \in c} \mathcal{DF}'(\sigma_{\text{old}}, \sigma)(n)] \quad \text{where}$$

$$\mathcal{DF}'(\sigma_{\text{old}}, \sigma) = \left[ n \in N \mapsto \begin{cases} \sqcup_i \sigma(I(n)_i) & L(n) = \phi_v \wedge \text{BACK}(v) = \emptyset \\ \sigma_{\text{old}}([n]) \nabla & \\ (\sqcup_{i \in \text{FORWARD}(v)} \sigma(I(n)_i) \sqcup & L(n) = \phi_v \wedge \text{BACK}(v) \neq \emptyset \\ \sqcup_{i \in \text{BACKWARD}(v)} \sigma_{\text{old}}(I(n)_i)) & \\ \mathcal{F}(L(n))(i_1, \dots, i_{|L(n)|}) & L(n) \in \mathcal{D}^* \rightarrow \mathcal{D}, i_k = \sigma(I(n)_k) \end{cases} \right]$$

$$\text{FORWARD}(v) = \{i \mid \exists v' \in V, P(v)_i = v' \wedge (v, i) \notin B\}$$

$$\text{BACKWARD}(v) = \{i \mid \exists v' \in V, P(v)_i = v' \wedge (v, i) \in B\}$$

- If  $\sigma_{0 \times \infty} = [c \in (N/\equiv) \mapsto \perp]$ ,  $\sigma_{i \times 0} = [c \in (N/\equiv) \mapsto \top]$ , and  $\sigma_{i+1 \times j+1} = \mathcal{DF}(\sigma_{i \times \infty}, \sigma_{i+1 \times j})$ , then the final analysis result is  $\sigma_{\infty \times \infty}$ .

The algorithm given in Definition 5.3 differentiates well-formed from ill-formed cycles. In particular, given a set of back edges  $B$ , which identifies the set of well-formed cycles exhaustively, the algorithm in effect breaks these cycles by computing a two-level fixpoint. The inner fixpoint corresponds to a greatest fixpoint calculation while the outer fixpoint corresponds to a least fixpoint calculation. Abstractions flow along “forward” edges (edges that are not back edges) in the inner fixpoint and flow along back edges in the outer fixpoint. Any ill-formed cycles will have their edges flowed over in the inner fixpoint, which will not allow them to poison the analysis result, due to the inner loop computing a greatest fixpoint. Only edges in well-formed cycles can propagate abstractions across outer fixpoint iterations, which allows the algorithm to compute least fixpoints specifically over well-formed cycles. An example of a trace of this algorithm running can be seen in Figure 4b, where the columns correspond to iterations of the outer fixpoint. Although an ill-formed cycle exists in the processed e-graph, the only edge that can propagate results starting from  $\perp$  (rather than from  $\top$ ) is an edge in a well-formed cycle.

## 6 Combining Abstract Interpretation and Equality Saturation

Now that we can perform abstract interpretations over SSA e-graphs, we consider how we can combine analysis results over programs with equality saturation. We propose a fixpoint algorithm that alternates between phases of abstract interpretation and equality saturation, using the improved precision in one half to improve the precision of the other half.

### 6.1 Using Analysis Results in Equality Saturation

First, we discuss how analysis results can help equality saturation. Recall that equality saturation consists of the repeated application of *rewrite rules* to an e-graph. A rewrite rule performs *e-matching* to find represented graphs matching some pattern—for each matched graph, an *action* is taken, usually to insert a new represented graph and assert that sets of e-classes are now known equal. A rewrite rule may additionally depend on a set of *conditions*—that is, even if the pattern e-matches some represented graph, some analysis fact must also be known about involved e-classes for the rewrite to be sound. For example, in a setting involving bitvector operations (such as a compiler middle-end), the rewrite  $a/b \Rightarrow a \gg \log_2(b)$  is only valid if  $b$  is known to be a power of two. To depend on values of an abstraction in rewrite rules, we require:



- (1) Depended on abstractions must be flow insensitive, meaning  $\Sigma$  must concretize to a single, sound set of concrete values for every possible control flow walk.
- (2) There must exist a computable partial order  $\sqsubseteq$  on the abstract values  $\Sigma$  such that  $\forall a, b, c \in \Sigma, a \sqsubseteq c \wedge b \sqsubseteq c \iff a \sqcup b \sqsubseteq c$  and  $\forall a, b, c \in \Sigma, c \sqsubseteq a \wedge c \sqsubseteq b \iff c \sqsubseteq a \sqcap b$ .

The first requirement is necessary because rewrite rules equate values *globally*—for the equivalence relation stored by an e-graph to be sound, it must be true that nodes in the same class evaluate to the same concrete value on all control flow walks. Thus, if a rewrite rule requires a condition, that condition must be true during all control flow walks, which is a flow insensitive requirement. We discuss a potential relaxation of this requirement in Section 6.3. The second requirement is needed to specify conditions in rewrite rules. A condition in a rewrite rule should only check that an analysis result implies the absence of certain evaluations of a node. For example, the rewrite rule  $a/b \Rightarrow a \gg \log_2(b)$  is only valid if it is known that  $b$  will *never* evaluate to a value that is not a power of two. Analysis facts that are more precise than a required fact (less in the partial order  $\sqsubseteq$ ) imply smaller concretizations, meaning the rewrite rule is still valid. Thus, rewrite rule conditions are only allowed to perform lower threshold tests on analysis facts using the  $\sqsubseteq$  operator<sup>5</sup>.

## 6.2 Combined Abstract Interpretation and Equality Saturation Algorithm

Abstract interpretation and equality saturation can both be described as fixpoint algorithms—a fixpoint algorithm for abstract interpretation over SSA e-graphs is given in Definition 5.3, and equality saturation can be expressed as the fixpoint of applying a set of rewrite rules and rebuilding [51]. A natural first attempt at combining the two might be then to simply run iterations of abstract interpretation and equality saturation in the same fixpoint loop. The issue is that abstract interpretation and equality saturation are incompatible in the following sense:

- Abstract interpretation algorithms usually compute *least* fixpoints, which are *optimistic* analyses, meaning that intermediate abstractions are not necessarily sound approximations of the concrete semantics—only the final abstraction (the computed fixpoint) is sound.
- Equality saturation is a *pessimistic* analysis, meaning that intermediate e-graphs that aren't fully saturated are still sound. This property is key when saturation is not possible, which is often the case in practice. Additionally, equalities discovered at intermediate steps are never thrown away. Indeed, equalities may be *impossible* to discard without removing represented terms, such as with the e-graph shown in Figure 4a—removing the non-trivial equality would mean the terms  $1 * 1 * 5$ ,  $1 * 1 * 1 * 5$ , and so on are no longer represented.

Combining abstract interpretation and equality saturation naively causes unsound intermediate results “derived” by the abstract interpretation to fire unsound rewrites in the e-graph—the resulting equalities are never discarded, as equality saturation assumes all intermediate results are sound.

We instead propose that abstract interpretation and equality saturation be run in two separate halves—each is run in its own fixpoint, and both are run in an “outer” fixpoint to propagate results between the halves. This ensures that only sound results derived from the abstract interpretation are used for rewrites. Algorithm 1 shows pseudo-code for this approach. Note that  $\sigma$  is re-initialized (Line 5) every iteration of the outer fixpoint (Lines 4 to 17). This is because keeping the analysis results of the previous iteration of the outer fixpoint may be unnecessarily imprecise.

<sup>5</sup>Some rewrite rules, such as one inserting an e-node with a constant symbol into an e-class if an interval analysis derives a constant value for the e-class, may perform other kinds of tests on the analysis result than strictly  $\sqsubseteq$ . However, these rules must be equivalent to some (potentially very large) combination of rules fulfilling this requirement. For example, the aforementioned constant reification rule is equivalent to the combination of a rule per possible constant that tests that the interval analysis has derived that particular constant.

---

**Algorithm 1** Abstract interpretation and equality saturation combination
 

---

```

1: procedure COMBINED( $\mathcal{S}_{\equiv}, \mathcal{G}_{\equiv}$ )
2:    $(N, L, I, \equiv) \leftarrow \mathcal{S}_{\equiv}$ 
3:    $B \leftarrow \text{WTOBackEdges}(\mathcal{G}_{\equiv})$  // CFG does not change during analysis or rewriting
4:   while  $\mathcal{S}_{\equiv}$  has changed and before timeout do // does not need to be driven to fixpoint
5:      $\sigma \leftarrow [c \mapsto \perp, \forall c \in N/\equiv]$ 
6:     while  $\sigma$  has changed do
7:        $\sigma_{\text{old}} \leftarrow \sigma$ 
8:        $\sigma \leftarrow [c \mapsto \top, \forall c \in N/\equiv]$ 
9:       while  $\sigma$  has changed do
10:         $\sigma \leftarrow \mathcal{DF}(\sigma_{\text{old}}, \sigma)$ 
11:      end while
12:    end while
13:    while  $\mathcal{S}_{\equiv}$  has changed and before timeout do // does not need to be driven to fixpoint
14:       $\mathcal{S}_{\equiv}.\text{apply\_rewrites}(\sigma)$ 
15:       $\mathcal{S}_{\equiv}.\text{rebuild}()$ 
16:    end while
17:  end while
18:  return  $\sigma, \mathcal{S}_{\equiv}$ 
19: end procedure

```

---

We now state a useful property of this algorithm. Call  $\sigma_i$  the analysis result derived and  $\mathcal{S}_{\equiv_i}$  the SSA e-graph constructed at the end of iteration  $i$  of the outer fixpoint loop (Lines 4 to 17).

**PROPOSITION 6.1.**  $\sigma_{i+1}(c) \sqsubseteq \sigma_i(c)$  for all e-classes  $c \in (N_i/\equiv_i)$ .

This can be seen by examining every action a rewrite rule may perform on an e-graph and showing that none have the capacity to degrade a derived abstraction.

- (1) Adding an e-node to an e-class is an action. Given old e-class  $c$  and e-node  $n$  to add, the class iteration function evaluated on the new e-class is  $\mathcal{DF}(\sigma_{\text{old}}, \sigma)(c \cup \{n\}) = \mathcal{DF}(\sigma_{\text{old}}, \sigma)(c) \sqcap \mathcal{DF}(\sigma_{\text{old}}, \sigma)(\{n\}) \sqsubseteq \mathcal{DF}(\sigma_{\text{old}}, \sigma)(c)$ , meaning any analyzed result of the new e-class will be at least as precise as before adding the e-node.
- (2) Equating two e-classes is an action. Given old e-classes  $c_1$  and  $c_2$ , the class iteration function evaluated on the new e-class is  $\mathcal{DF}(\sigma_{\text{old}}, \sigma)(c_1 \cup c_2) = \mathcal{DF}(\sigma_{\text{old}}, \sigma)(c_1) \sqcap \mathcal{DF}(\sigma_{\text{old}}, \sigma)(c_2)$ , which is at least as precise as both  $\mathcal{DF}(\sigma_{\text{old}}, \sigma)(c_1)$  and  $\mathcal{DF}(\sigma_{\text{old}}, \sigma)(c_2)$ .

Rebuilding just asserts equalities between pre-existing e-classes, and thus also only improves analysis precision. Additionally, because analysis results only improve every iteration and rewrite rules can only perform lower threshold tests against analysis facts, the set of possible rewrites in the SSA e-graph also only grows every iteration. This means that each iteration of the outer fixpoint monotonically improves analysis precision and rewrite applicability. Note that the above proposition does not rely on the SSA e-graph being saturated during every iteration of the outer fixpoint, so this property holds even when rewriting is cut off early (for example, after a timeout).

### 6.3 Using Flow Sensitive Abstractions in Contextual E-Graphs

In order to use flow sensitive abstractions in rewrites, we need to be able to record equivalences in a flow sensitive manner. However, the equivalence abstraction that an e-graph stores is flow insensitive (the equivalences between nodes must hold on *all* possible control flow walks). Conceptually,

we can modify the equivalence abstraction to be flow sensitive by defining  $\Sigma_{\equiv} = V \rightarrow \{\perp, \top\}$  (thus, a separate equivalence relation is maintained per CFG vertex a control flow walk can end in). Then, when a rewrite rule depends on a flow sensitive abstraction, any discovered equalities are asserted in a flow sensitive fashion. In prior work, e-graphs that can store multiple equivalence relations efficiently have been called “contextual e-graphs” [16, 20] or “colored e-graphs” [43, 44]. Equivalence relations are identified by either “contexts” or “colors” and form a hierarchy or lattice relationship. Intuitively, if a context  $A$  is the parent of another context  $B$  in the hierarchy, then any equalities known in  $A$  are automatically known in  $B$ . For implementing a flow sensitive equivalence abstraction with a contextual e-graph, we suggest that dominance in a CFG is a good choice of hierarchy. This is because if a vertex  $A$  dominates a vertex  $B$  in a CFG, any abstraction that is sound for every walk ending at  $A$  will automatically be sound for every walk ending at  $B$  (because any such walk will have traversed through  $A$ —this is what it means for  $A$  to dominate  $B$ ). Implementing contextual e-graphs efficiently is an open research question, and we hope this perspective serves as a motivating use case for future work in this area.

## 7 Implementation and Examples

In this section, we expand on how our techniques apply to a set of example programs. Additionally, we implemented our techniques as a small Rust tool that correctly analyses the example programs.

### 7.1 Rust Implementation

We implemented the described scheme for combining equality saturation and abstract interpretation over SSA e-graphs in a Rust tool. The tool parses code in a pseudo-code syntax and translates each function into a SSA e-graph and accompanying CFG. Equality saturation and abstract interpretation can be run in the aforementioned alternating scheme on the e-graph. E-graphs and analysis facts over e-classes are stored in functional database tables. There is a single table per function symbol and per kind of analysis (in the same style as egglog [55]). Each row in a function symbol table is an e-node mapping input e-classes and other parameters to the e-class the e-node is a member of. Each row in an analysis table maps an e-class to its corresponding analysis fact. A set of manually implemented analysis rules implement  $\mathcal{DF}'$  while the conflict resolution of rows in analysis tables takes the meet of conflicting abstractions, which implements  $\mathcal{DF}$  (both from Definition 5.3).

We implement three analyses—interval, offset, and reachability. Interval analysis is the standard non-relational abstraction of numeric values that represents a set of integers in a range from a low integer to a high integer [9, 10]. Offset analysis is a relational abstraction that relates SSA values  $a$  and  $b$  if it is known that  $a = b + c$ , for some static  $c \in \mathbb{Z}$ . As offsets form a group, we implement this abstraction using the labeled union find approach [27] (note that this union find is separate from the union find implementing the equivalence abstraction in the e-graph). Our techniques can be extended to handle relational abstractions fairly easily—the computed  $\sigma$  stores an abstraction per pair of e-classes, rather than per e-class. Reachability analysis abstracts what control flow walks are possible. We split reachability analysis into two cooperating abstractions—vertex reachability and edge reachability. Figure 5 summarizes the interval and offset abstractions. As reachability analysis characterizes the CFG rather than the SSA e-graph, we implement it as a separate cooperating analysis in a reduced product. The reachability abstraction makes the interval and offset abstractions more precise by eliminating unreachable edge inputs to  $\phi$  nodes. The interval abstraction makes the reachability abstraction more precise by identifying when an edge’s condition will never be met (in our implementation, an edge will never be met if the interval  $[0, 0]$  can be derived for the condition value)—this effectively captures conditional constant propagation [2].

$$\begin{aligned}
\Sigma_I &= \{(l, h) \mid l \in \mathbb{Z} \cup \{-\infty\}, h \in \mathbb{Z} \cup \{\infty\}, l \leq h\} \cup \{\perp_I\} \\
\gamma_I(\perp_I) &= \lambda w. \emptyset \quad \gamma_I((l, h)) = \lambda w. \{z \mid z \in \mathbb{Z}, l \leq z \leq h\} \\
\top_I &= (-\infty, \infty) \quad (l_1, h_1) \sqcup_I (l_2, h_2) = (\min(l_1, l_2), \max(h_1, h_2)) \\
(l_1, h_1) \sqcap_I (l_2, h_2) &= \begin{cases} (\max(l_1, l_2), \min(h_1, h_2)) & \max(l_1, l_2) \leq \min(h_1, h_2) \\ \perp_I & \text{otherwise} \end{cases} \\
(l_1, h_1) \nabla_I (l_2, h_2) &= \begin{pmatrix} -\infty & l_1 > l_2 \\ l_2 & \text{otherwise} \end{pmatrix}, \begin{pmatrix} \infty & h_1 < h_2 \\ h_2 & \text{otherwise} \end{pmatrix} \\
\Sigma_O &= \mathbb{Z} \cup \{\perp_O, \top_O\} \quad \gamma_O(\perp_O) = \lambda w. \emptyset \quad \gamma_O(\top_O) = \lambda w. \mathbb{Z}^2 \\
\gamma_O(z) &= \lambda w. \{(a, b) \mid (a, b) \in \mathbb{Z}^2, a = b + z\} \\
z_1 \sqcup_O z_2 &= \begin{cases} z_1 & z_1 = z_2 \\ \perp_O & \text{otherwise} \end{cases} \quad z_1 \sqcap_O z_2 = \begin{cases} z_1 & z_1 = z_2 \\ \top_O & \text{otherwise} \end{cases} \quad z_1 \nabla_O z_2 = z_2
\end{aligned}$$

Fig. 5. The interval ( $I$ ) and offset ( $O$ ) abstractions. Both abstractions are fairly standard, except that the concretization functions return functions that take and ignore a control flow walk argument.

```

1 fn example1(y) {
2   let x = -6;
3   let z = 42;
4   while y < 10 {
5     y = y + 1;
6     x = x + 8;
7     let lhs = ((x + y) + z) * y;
8     let rhs = 2 * y + (y * y + z * y);
9     if lhs != rhs {
10      z = 24;
11    }
12    x = x - 8;
13  }
14  return z + 7;
15 }

```

Fig. 6. First example program to analyze and rewrite. The goal is to show that the returned value is 49, which requires rewriting and an optimistic interval analysis.

```

1 fn example2(x) {
2   let y = x;
3   while y < 10 {
4     let xt = x;
5     x = y * y + y * 5;
6     y = xt * (y + 5 + 0);
7   }
8   return x - y;
9 }

```

Fig. 7. Second example program to analyze and rewrite. The goal is to show that the returned value is 0, which requires rewriting and an optimistic equivalence analysis.

## 7.2 Example Programs

The first program example is shown in Figure 6. The “goal” is for the final returned value to be discovered equal to 49. At a high level, discovering this fact requires an optimistic analysis to discover that  $x = 2$  at line 9<sup>6</sup>, rewrites to discover that  $((2 + y) + z) * y = 2 * y + (y * y + z * y)$ ,

<sup>6</sup>The reader may have noticed that we said we can only combine flow insensitive analyses with rewriting in Section 6.1, yet here we require we know the value of  $x$  at a particular program location. The discrepancy is accounted for by the fact that we define flow insensitive abstractions over *SSA values*, not *program variables*. This does not fully emulate flow sensitivity, as two different branches may define the same de-duplicated SSA value. However, in this case the SSA value corresponding to the program variable  $x$  at line 9 can be analyzed precisely enough.

a rewrite to discover that  $(2 * y + (y * y + z * y)) \neq 2 * y + (y * y + z * y) = 0$ , and finally an optimistic analysis to discover that line 10 is unreachable.

The second program example is shown in Figure 7. The “goal” is to for the final returned value to be discovered equal to 0. This requires an optimistic analysis to discover that  $x = y$  at line 8. Since  $x$  is a function parameter, its interval is  $[-\infty, \infty]$ —to discover  $x = y$ , the offset analysis must find that the SSA values being subtracted at the end of the function are related by the offset 0. A rewrite rule discovers that  $x_t * (y + 5) = x_t * y + x_t * 5$ . After this rewrite, the optimistic offset analysis can derive that  $x = y$ . The final return constant value is discovered by a rewrite that rewrites  $x - x \Rightarrow 0$  for any e-class  $x$ . Note that we include the rewrite rule  $x + 0 \Rightarrow x$  in our rule set, we means we will discover a ill-formed cycle due to the sub-expression  $5 + 0$ . This cycle does not correspond to a back edge in the CFG, and thus the derived facts for the involved e-classes are sound with respect to the well-formed represented graphs of the SSA e-graph.

This example demonstrates a particular capability of optimistic analyses applied to e-graphs that has eluded prior equality saturation systems: soundly de-duplicating cycles. Our rewrite set includes a rule that merges two e-classes if they are related by the offset 0. Since our offset analysis is optimistic, two well-formed cycles can be discovered to be related by the offset 0 (more precisely, there is a mapping between e-classes in each cycle such that each pair of e-classes in the mapping is related). Additionally, two ill-formed cycles will not be merged, which avoids merging the e-classes  $x = \{5, 0 + x\}$  and  $y = \{3, 0 + y\}$ . In effect, this allows us to merge well-formed cycles, which is sound, while not merging ill-formed cycles, which is unsound<sup>7</sup>. This addresses a known problem in the e-graphs literature [47, 57, 58].

We wrote these examples both in a pseudo-code language our tool ingests (shown in Figures 6 and 7) and in C—we compiled the examples in C with GCC 15.2 and Clang 21.1.0 (with optimization level -O2). Neither GCC nor Clang can fully optimize either example.

## 8 Related Work

*E-Graphs and Equality Saturation.* E-graphs were originally developed for the purpose of computing congruence closure over a set of terms [34] and are a key ingredient in SMT solvers for propagating equalities between theories [35]. Only in later work was applying rewrite rules to e-graphs, via equality saturation, proposed [45, 47, 51]. Applying rewrite rules requires syntactically matching their left hand side, which is called “e-matching”. Most e-matching implementations are based on pattern compilation [32] or relational queries [56]. Another consideration is the representation of the e-graph itself—traditionally, the graph is stored directly as a graph in memory [14, 51], while recent work proposes using database tables with a canonicalization procedure [54, 55]. Our Rust tool implements a version of the latter approach, which has the advantage of presenting a uniform representation for both terms in the e-graph and for analysis facts over e-classes. Tree term languages are the most common target for embedding into e-graphs. Cycles are often created via rewrite rules, but are rarely desired and often filtered away during extraction [52]. The only prior work that we are aware of that encodes an explicitly cyclic program representation into e-graphs is E-PEGs [47]—their compilation tool, called Peggy, does not support lattice-based analyses and only supports purely syntactic rewrites<sup>8</sup>. Neither egg nor egglog<sup>9</sup> support creating an initial e-graph with cycles and both tools treat extracting cycles as a failure mode [51, 55].

<sup>7</sup>Technically, this example only requires an optimistic equivalence analysis, which is less powerful than an offset analysis.

<sup>8</sup>Some simple analyses, such as constant propagation, are re-implemented as pure syntactic transformations in Peggy.

<sup>9</sup>egglog’s relational e-matching can theoretically match graph patterns, not just term patterns [56].

*Abstract Interpretation over E-Graphs.* E-class analysis is a technique for performing lattice-based analyses on e-graphs. A fact from a single lattice is assigned to each e-class in an e-graph [51]. E-class analyses are computed bottom-up, and all facts are conservatively initialized to  $\top$  (computing a greatest fixpoint). Analysis facts are propagated during rewriting, since the analysis is incrementally sound. Later work formalizes e-class analysis as an abstract interpretation and observes the fruitful back-and-forth between analysis and rewriting, but does not produce a precise result for well-formed cycles in a cyclic program representation [12]. egglog allows users to define analyses with functional database tables. Unlike in egg, the database approach supports 1) storing multiple analyses per e-class and 2) relational analyses involving multiple e-classes. However, the restriction of incremental soundness still applies, which prevents analyzing cyclic programs precisely. To our knowledge, no prior equality saturation system admits precise analyses of well-formed cycles.

*Equivalence and Other Relational Abstractions.* Several relational abstract domains have been proposed, including difference bounds [29], difference abstraction [30], octagons [31], pentagons [28], two variables per inequality [40], and equalities (using e-graphs to store the equivalence relation) [5]. As relational abstractions store an abstract value per pair, or even tuple, of variables, the entire abstract state can grow quite large. Some prior work takes advantage of sparseness in the relations between variables [42]. Another approach uses a data structure called a “labeled union find” to store some weakly relational domains by a spanning tree, rather than a fully connected graph [27]. A weakly relational domain can be used to “map factorize” a compatible non-relational domains, meaning the non-relational domain is only stored per relational class, rather than per program variable. The equivalence relation stored in an e-graph can be seen as the maximal example of this, as it is compatible with every other abstraction (including non-relational and relational abstractions). An interesting property of e-graphs is that they not only factor abstractions by equivalence classes, but they also factor *terms* by equivalence classes—this is how e-graphs compactly represent a large (sometimes infinite) set of equivalent terms [34]. An interesting direction for future work could be exploring factoring terms by relations other than equivalence relations—labeled union finds could be used to factor terms that are equivalent modulo a group action [58].

*Combining Analyses and Transformations.* The structure of many compilers consists of phases of analyses and transformations—either the order of analyses and transformations is designed explicitly [36], or a generic “pass” mechanism provides flexible ordering of analyses and transformations [22, 23, 39]. Every intermediate program is well-formed and optimistic analyses are often employed. However, intermediate programs are ephemeral and transformations are run in a specific order, leading to the classic phase ordering problem. Program transformations have also been used specifically as an ad-hoc mechanism to facilitate communication between program analyses [25]. Another approach to combining analyses and transformations is to view transformations themselves as abstract interpretations [11], where the abstraction representing the transformed program concretizes to the same observable semantics as the concrete program (for some definition of observable) [11], or there is some homomorphism between executions of the pre-and-post-transformation programs [24]. Recent work has applied this latter perspective to SSA translation [24] and to simple program simplifications [26]. Formalizing transformations as abstract interpretations allows for straightforward combinations with standard analyses via product operators [26]. In this paper, the equivalence relation stored by an e-graph is an abstraction, but equality saturation itself is not described as an abstract interpretation. We believe a version of “optimistic rewriting” may be possible to arrive at by formalizing our entire method as an abstract interpretation. However, we also believe that the implementation of such a method would be difficult to optimize—a fast implementation of persistent e-graphs is likely necessary.



## 9 Conclusion

This paper tackles the problem of performing optimistic analyses over e-graphs in tandem with equality saturation. We identify a key problem that can cause optimistic analyses to compute unsound results over e-graphs—equality saturation can create ill-formed cycles in the e-graph. In the case of SSA form, this corresponds to cycles that either do not contain  $\phi$ s at all or whose  $\phi$ s do not satisfy strictness. These cycles can poison the analysis results with unsoundness. We identify a technique to compute analyses optimistically while only considering well-formed cycles, rather than all cycles. This technique allows for precise and sound analysis of e-graphs containing cyclic program representations after equality saturation. This technique also enables combining optimistic analyses and equality saturation to discover more conditional rewrites and to refine the abstraction of the analysis via more discovered equalities.

The treatment in this paper is primarily theoretical and discovers a qualitative improvement in capability over prior equality saturation systems. Important future work includes exploring domains where this technique could uncover quantitative gains in optimization potential. As alluded to in this paper, optimizing SSA programs is a candidate for this kind of investigation. However, we believe that reconsidering the program representation itself could be fruitful—our SSA graph representation only embeds data flow into the e-graph, while control flow is left untouched. The program expression graph (PEG) [47] is a representation where control flow can be manipulated by rewrites. In fact, this representation was originally designed for equality saturation.

## Acknowledgments

We are very grateful for discussions with Tyler Hou, Samuel Coward, Cheng Zhang, and Alexandra Silva that motivated and inspired the basis of this work. We are also grateful for broader discussions at the Dagstuhl Seminar 26022, “Program Optimization with E-Graphs”, which helped crystallize much of the thinking that went into this work.

## References

- [1] Alfred V. Aho, Monica S. Lam, Ravi Sethi, and Jeffrey D. Ullman. 2006. *Compilers: Principles, Techniques, and Tools* (2nd Edition). Addison-Wesley Longman Publishing Co., Inc., USA.
- [2] B. Alpern, M. N. Wegman, and F. K. Zadeck. 1988. Detecting equality of variables in programs. In *Proceedings of the 15th ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages* (San Diego, California, USA) (POPL ’88). Association for Computing Machinery, New York, NY, USA, 1–11. doi:10.1145/73560.73561
- [3] François Bourdoncle. 1993. Efficient chaotic iteration strategies with widenings. In *Formal Methods in Programming and Their Applications*, Dines Bjørner, Manfred Broy, and Igor V. Pottosin (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 128–141.
- [4] Yaohui Cai, Kaixin Yang, Chenhui Deng, Cunxi Yu, and Zhiru Zhang. 2025. SmoothE: Differentiable E-Graph Extraction. In *Proceedings of the 30th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 1* (Rotterdam, Netherlands) (ASPLOS ’25). Association for Computing Machinery, New York, NY, USA, 1020–1034. doi:10.1145/3669940.3707262
- [5] Bor-Yuh Evan Chang and K. Rustan M. Leino. 2005. Abstract Interpretation with Alien Expressions and Heap Structures. In *Verification, Model Checking, and Abstract Interpretation*, Radhia Cousot (Ed.). Springer Berlin Heidelberg, Berlin, Heidelberg, 147–163.
- [6] Cliff Click and Keith D. Cooper. 1995. Combining analyses, combining optimizations. *ACM Trans. Program. Lang. Syst.* 17, 2 (March 1995), 181–196. doi:10.1145/201059.201061
- [7] Cliff Click and Michael Paleczny. 1995. A simple graph-based intermediate representation. In *Papers from the 1995 ACM SIGPLAN Workshop on Intermediate Representations* (San Francisco, California, USA) (IR ’95). Association for Computing Machinery, New York, NY, USA, 35–49. doi:10.1145/202529.202534
- [8] Agostino Cortesi, Giulia Costantini, and Pietro Ferrara. 2013. A Survey on Product Operators in Abstract Interpretation. *Electronic Proceedings in Theoretical Computer Science* 129 (09 2013). doi:10.4204/EPTCS.129.19
- [9] Patrick Cousot and Radhia Cousot. 1977. Abstract interpretation: a unified lattice model for static analysis of programs by construction or approximation of fixpoints. In *Proceedings of the 4th ACM SIGACT-SIGPLAN Symposium on Principles*

- of *Programming Languages* (Los Angeles, California) (POPL '77). Association for Computing Machinery, New York, NY, USA, 238–252. doi:10.1145/512950.512973
- [10] Patrick Cousot and Radhia Cousot. 1979. Systematic design of program analysis frameworks. In *Proceedings of the 6th ACM SIGACT-SIGPLAN Symposium on Principles of Programming Languages* (San Antonio, Texas) (POPL '79). Association for Computing Machinery, New York, NY, USA, 269–282. doi:10.1145/567752.567778
  - [11] Patrick Cousot and Radhia Cousot. 2002. Systematic design of program transformation frameworks by abstract interpretation. *SIGPLAN Not.* 37, 1 (Jan. 2002), 178–190. doi:10.1145/565816.503290
  - [12] Samuel Coward, George A. Constantinides, and Theo Drane. 2023. Combining E-Graphs with Abstract Interpretation. In *Proceedings of the 12th ACM SIGPLAN International Workshop on the State Of the Art in Program Analysis* (Orlando, FL, USA) (SOAP 2023). Association for Computing Machinery, New York, NY, USA, 1–7. doi:10.1145/3589250.3596144
  - [13] Samuel Coward, Theo Drane, and George A. Constantinides. 2024. ROVER: RTL Optimization via Verified E-Graph Rewriting. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* (2024), 1–1. doi:10.1109/TCAD.2024.3410154
  - [14] Leonardo De Moura and Nikolaj Bjørner. 2008. Z3: an efficient SMT solver. In *Proceedings of the Theory and Practice of Software, 14th International Conference on Tools and Algorithms for the Construction and Analysis of Systems* (Budapest, Hungary) (TACAS'08/ETAPS'08). Springer-Verlag, Berlin, Heidelberg, 337–340.
  - [15] Delphine Demange, Yon Fernández de Retana, and David Pichardie. 2018. Semantic reasoning about the sea of nodes. In *Proceedings of the 27th International Conference on Compiler Construction* (Vienna, Austria) (CC '18). Association for Computing Machinery, New York, NY, USA, 163–173. doi:10.1145/3178372.3179503
  - [16] Alexandre Drewery, Thomas Jensen, and David Pichardie. 2025. Contextual Equality Saturation. In *SAS 2025 - 32nd Static Analysis Symposium*. Singapore, Singapore, 1–26. <https://inria.hal.science/hal-05226543>
  - [17] Chris Fallin. 2023. ægraphs: Acyclic E-graphs for Efficient Optimization in a Production Compiler. <https://pldi23.sigplan.org/details/egraphs-2023-papers/2/-graphs-Acyclic-E-graphs-for-Efficient-Optimization-in-a-Production-Compiler>
  - [18] H. Gericke. 1957. Tarski Alfred. A lattice-theoretical fixpoint theorem and its applications. *Pacific journal of mathematics*, Bd. 5 (1955), S. 285–309. *Journal of Symbolic Logic* 22 (1957), 370 – 370. <https://api.semanticscholar.org/CorpusID:119521721>
  - [19] Amir Kafshdar Goharshady, Chun Kit Lam, and Lionel Parreaux. 2024. Fast and Optimal Extraction for Sparse Equality Graphs. *Proc. ACM Program. Lang.* 8, OOPSLA2, Article 361 (Oct. 2024), 27 pages. doi:10.1145/3689801
  - [20] Tyler Hou, Shadaj Laddad, and Joseph M. Hellerstein. 2025. Towards Relational Contextual Equality Saturation. arXiv:2507.11897 [cs.PL] <https://arxiv.org/abs/2507.11897>
  - [21] John B. Kam and Jeffrey D. Ullman. 1977. Monotone data flow analysis frameworks. *Acta Inf.* 7, 3 (Sept. 1977), 305–317. doi:10.1007/BF00290339
  - [22] Chris Lattner and Vikram Adve. 2004. LLVM: A Compilation Framework for Lifelong Program Analysis & Transformation. In *Proceedings of the International Symposium on Code Generation and Optimization: Feedback-Directed and Runtime Optimization* (Palo Alto, California) (CGO '04). IEEE Computer Society, USA, 75.
  - [23] Chris Lattner, Mehdi Amini, Uday Bondhugula, Albert Cohen, Andy Davis, Jacques Pienaar, River Riddle, Tatiana Shpeisman, Nicolas Vasilache, and Oleksandr Zinenko. 2021. MLIR: scaling compiler infrastructure for domain specific computation. In *Proceedings of the 2021 IEEE/ACM International Symposium on Code Generation and Optimization* (Virtual Event, Republic of Korea) (CGO '21). IEEE Press, 2–14. doi:10.1109/CGO51591.2021.9370308
  - [24] Matthieu Lemerre. 2023. SSA Translation Is an Abstract Interpretation. *Proc. ACM Program. Lang.* 7, POPL, Article 65 (Jan. 2023), 30 pages. doi:10.1145/3571258
  - [25] Sorin Lerner, David Grove, and Craig Chambers. 2002. Composing dataflow analyses and transformations. In *Proceedings of the 29th ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages* (Portland, Oregon) (POPL '02). Association for Computing Machinery, New York, NY, USA, 270–282. doi:10.1145/503272.503298
  - [26] Dorian Lesbre and Matthieu Lemerre. 2024. Compiling with Abstract Interpretation. *Proc. ACM Program. Lang.* 8, PLDI, Article 162 (June 2024), 26 pages. doi:10.1145/3656392
  - [27] Dorian Lesbre, Matthieu Lemerre, Hichem Rami Ait-El-Hara, and François Bobot. 2025. Relational Abstractions Based on Labeled Union-Find. *Proc. ACM Program. Lang.* 9, PLDI, Article 195 (June 2025), 26 pages. doi:10.1145/3729298
  - [28] Francesco Logozzo and Manuel Fähndrich. 2008. Pentagons: a weakly relational abstract domain for the efficient validation of array accesses. In *Proceedings of the 2008 ACM Symposium on Applied Computing* (Fortaleza, Ceara, Brazil) (SAC '08). Association for Computing Machinery, New York, NY, USA, 184–188. doi:10.1145/1363686.1363736
  - [29] Antoine Miné. 2001. A New Numerical Abstract Domain Based on Difference-Bound Matrices. In *Proceedings of the Second Symposium on Programs as Data Objects* (PADO '01). Springer-Verlag, Berlin, Heidelberg, 155–172.
  - [30] Antoine Miné. 2002. A Few Graph-Based Relational Numerical Abstract Domains. In *Proceedings of the 9th International Symposium on Static Analysis* (SAS '02). Springer-Verlag, Berlin, Heidelberg, 117–132.

- [31] Antoine Miné. 2006. The octagon abstract domain. *Higher Order Symbol. Comput.* 19, 1 (March 2006), 31–100. doi:10.1007/s10990-006-8609-1
- [32] Leonardo Moura and Nikolaj Bjørner. 2007. Efficient E-Matching for SMT Solvers. In *Proceedings of the 21st International Conference on Automated Deduction: Automated Deduction* (Bremen, Germany) (CADE-21). Springer-Verlag, Berlin, Heidelberg, 183–198. doi:10.1007/978-3-540-73595-3\_13
- [33] Chandrakana Nandi, Max Willsey, Adam Anderson, James R. Wilcox, Eva Darulova, Dan Grossman, and Zachary Tatlock. 2020. Synthesizing structured CAD models with equality saturation and inverse transformations. In *Proceedings of the 41st ACM SIGPLAN Conference on Programming Language Design and Implementation* (London, UK) (PLDI 2020). Association for Computing Machinery, New York, NY, USA, 31–44. doi:10.1145/3385412.3386012
- [34] Charles Gregory Nelson. 1980. *Techniques for program verification*. Ph. D. Dissertation. Stanford, CA, USA. AAI8011683.
- [35] Greg Nelson and Derek C. Oppen. 1979. Simplification by Cooperating Decision Procedures. *ACM Trans. Program. Lang. Syst.* 1, 2 (Oct. 1979), 245–257. doi:10.1145/357073.357079
- [36] Michael Paleczny, Christopher Vick, and Cliff Click. 2001. The java hotspot™ server compiler. In *Proceedings of the 2001 Symposium on Java™ Virtual Machine Research and Technology Symposium - Volume 1* (Monterey, California) (JVM'01). USENIX Association, USA, 1.
- [37] Pavel Panchekha, Alex Sanchez-Stern, James R. Wilcox, and Zachary Tatlock. 2015. Automatically improving accuracy for floating point expressions. In *Proceedings of the 36th ACM SIGPLAN Conference on Programming Language Design and Implementation* (Portland, OR, USA) (PLDI '15). Association for Computing Machinery, New York, NY, USA, 1–11. doi:10.1145/2737924.2737959
- [38] Fabrice Rastello. 2016. *SSA-based Compiler Design* (1st ed.). Springer Publishing Company, Incorporated.
- [39] Amit Sabne. 2020. XLA : Compiling Machine Learning for Peak Performance.
- [40] Axel Simon and Andy King. 2010. The two variable per inequality abstract domain. *Higher Order Symbol. Comput.* 23, 1 (March 2010), 87–143. doi:10.1007/s10990-010-9062-8
- [41] Taylor Simpson, Keith Cooper, and L. Simpson. 1997. SCC-based value numbering. (02 1997).
- [42] Gagandeep Singh, Markus Püschel, and Martin Vechev. 2015. Making numerical program analysis fast. In *Proceedings of the 36th ACM SIGPLAN Conference on Programming Language Design and Implementation* (Portland, OR, USA) (PLDI '15). Association for Computing Machinery, New York, NY, USA, 303–313. doi:10.1145/2737924.2738000
- [43] Eytan Singher and Shachar Itzhaky. 2023. Colored E-Graph: Equality Reasoning with Conditions. arXiv:2305.19203 [cs.PL] <https://arxiv.org/abs/2305.19203>
- [44] Eytan Singher and Shachar Itzhaky. 2024. Easter Egg: Equality Reasoning Based on E-Graphs with Multiple Assumptions. In *2024 Formal Methods in Computer-Aided Design (FMCAD)*. 70–83. doi:10.34727/2024/isbn.978-3-85448-065-5\_13
- [45] Michael Stepp, Ross Tate, and Sorin Lerner. 2011. Equality-based translation validator for LLVM. In *Proceedings of the 23rd International Conference on Computer Aided Verification* (Snowbird, UT) (CAV'11). Springer-Verlag, Berlin, Heidelberg, 737–742.
- [46] Dan Suciu, Yisu Remy Wang, and Yihong Zhang. 2025. Semantic Foundations of Equality Saturation. In *28th International Conference on Database Theory (ICDT 2025) (Leibniz International Proceedings in Informatics (LIPIcs), Vol. 328)*, Sudeepa Roy and Ahmet Kara (Eds.). Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl, Germany, 11:1–11:18. doi:10.4230/LIPIcs.ICDT.2025.11
- [47] Ross Tate, Michael Stepp, Zachary Tatlock, and Sorin Lerner. 2009. Equality saturation: a new approach to optimization. *SIGPLAN Not.* 44, 1 (Jan. 2009), 264–276. doi:10.1145/1594834.1480915
- [48] Todd Veldhuizen and Jeremy Siek. 2003. On Combining Program Improvers. (04 2003).
- [49] Harishankar Vishwanathan, Matan Shachnai, Srinivas Narayana, and Santosh Nagarakatte. 2022. Sound, Precise, and Fast Abstract Interpretation with Tristate Numbers. In *2022 IEEE/ACM International Symposium on Code Generation and Optimization (CGO)*. 254–265. doi:10.1109/CGO53902.2022.9741267
- [50] Yisu Remy Wang, Shana Hutchison, Jonathan Leang, Bill Howe, and Dan Suciu. 2020. SPORES: sum-product optimization via relational equality saturation for large scale linear algebra. *Proc. VLDB Endow.* 13, 12 (July 2020), 1919–1932. doi:10.14778/3407790.3407799
- [51] Max Willsey, Chandrakana Nandi, Yisu Remy Wang, Oliver Flatt, Zachary Tatlock, and Pavel Panchekha. 2021. egg: Fast and extensible equality saturation. *Proc. ACM Program. Lang.* 5, POPL, Article 23 (Jan. 2021), 29 pages. doi:10.1145/3434304
- [52] Yichen Yang, Phitchaya Mangpo Phothilimtha, Yisu Remy Wang, Max Willsey, Sudip Roy, and Jacques Pienaar. 2021. Equality Saturation for Tensor Graph Superoptimization. In *Proceedings of Machine Learning and Systems*. arXiv:2101.01332
- [53] Jiaqi Yin, Zhan Song, Chen Chen, Yaohui Cai, Zhiru Zhang, and Cunxi Yu. 2025. e-boost: Boosted E-Graph Extraction with Adaptive Heuristics and Exact Solving. *arXiv preprint arXiv:2508.13020* (2025).
- [54] Yihong Zhang. 2022. PLDI: U: Towards a Relational E-graph. <https://api.semanticscholar.org/CorpusID:250122313>

- [55] Yihong Zhang, Yisu Remy Wang, Oliver Flatt, David Cao, Philip Zucker, Eli Rosenthal, Zachary Tatlock, and Max Willsey. 2023. Better Together: Unifying Datalog and Equality Saturation. *Proc. ACM Program. Lang.* 7, PLDI, Article 125 (June 2023), 25 pages. [doi:10.1145/3591239](https://doi.org/10.1145/3591239)
- [56] Yihong Zhang, Yisu Remy Wang, Max Willsey, and Zachary Tatlock. 2022. Relational e-matching. *Proc. ACM Program. Lang.* 6, POPL, Article 35 (Jan. 2022), 22 pages. [doi:10.1145/3498696](https://doi.org/10.1145/3498696)
- [57] Philip Zucker. 2024. Co-Egraphs: Streams, Unification, PEGs, Rational Lambdas. <https://www.philipzucker.com/coegraph/>
- [58] Philip Zucker. 2025. Omelets Need Onions: E-graphs Modulo Theories via Bottom-up E-matching. [arXiv:2504.14340](https://arxiv.org/abs/2504.14340) [cs.PL] <https://arxiv.org/abs/2504.14340>