Citation for published version


DOI

https://doi.org/10.1145/2837614.2837633

Link to record in KAR

http://kar.kent.ac.uk/51097/

Document Version

UNSPECIFIED
From MinX to MinC: Semantics-Driven
Decompilation of Recursive Datatypes

Ed Robbins
University of Kent, UK
er209@kent.ac.uk

Andy King
University of Kent, UK
a.m.king@kent.ac.uk

Tom Schrijvers
KU Leuven, Belgium
tom.schrijvers@cs.kuleuven.be

Abstract
Reconstructing the meaning of a program from its binary executable is known as reverse engineering; it has a wide range of applications in software security, exposing piracy, legacy systems, etc. Since reversing is ultimately a search for meaning, there is much interest in inferring a type (a meaning) for the elements of a binary in a consistent way. Unfortunately existing approaches do not guarantee any semantic relevance for their reconstructed types.

This paper presents a new and semantically-founded approach that provides strong guarantees for the reconstructed types. Key to our approach is the derivation of a witness program in a high-level language alongside the reconstructed types. This witness has the same semantics as the binary, is type correct by construction, and it induces a (justifiable) type assignment on the binary. Moreover, the approach effectively yields a type-directed decompiler.

We formalise and implement the approach for reversing MinX, an abstraction of x86, to MinC, a type-safe dialect of C with recursive datatypes. Our evaluation compiles a range of textbook C algorithms to MinX and then recovers the original structures.

Categories and Subject Descriptors D.3.1 [Programming Languages]: Formal Definitions and Theory

Keywords reverse engineering, decompilation, recursive datatypes

1. Introduction
Reverse engineering is the activity of reconstructing the behaviour of a program from its binary executable. Quite apart from illicit purposes such as removing copyright protection, reversing has legitimate applications which include, but are not limited to: understanding the operation of, and threat posed by, viruses and malware; exposing flaws and vulnerabilities in commercial software, especially prior to deployment in government or industry; checking for license infringement; and interfacing with legacy software.

Reversing amounts to searching for a high-level meaning that is consistent across a binary. Typing, likewise, checks the elements of a program combine in a consistent, meaningful way, yielding a valuable program abstraction for the reverse engineer [15]. For example, datatypes can guide test-generation in fuzzing [27], help locate information in a core dump in memory-based forensics [4], and support program reconstruction [6][21]. Unfortunately, type recovery has been more make-shift and make-do than a discipline shaped by formal principles. IDAPro, the leading commercial disassembler, applies heuristics to assign simple types to locals [9]. REWARDS [13] recovers types from a single execution trace, which sheds no light on other traces. TIE [15] badges itself as being principled, but does not relate its type judgements to the semantics of the binary (which remain unspecified). Yet a firm semantic footing for these type systems is essential: a type recovery system which derives an incorrect type can easily mislead a reverse engineer undertaking a security audit, or misdirect a fuzzer into the wrong search space. Furthermore, these existing type systems [9][15][18] are unable to recover recursive datatypes.

The consensus is that types assigned to the binary should correspond to the original types of the source. But, needless to say, this is unavailable. This begs the question: what does it mean for the types to be correct, if there is no source to check correctness against?

We answer this question from a semantic perspective by constructing a witness program in a type-safe high-level language. The witness is not an arbitrary program, but carefully constructed to semantically coincide with the binary. Then, by proving that the witness is type-correct, we unequivocally establish that the binary inhabits the recovered types. Structured operational semantics (SOS) define our exemplar low-level and high-level languages, inspired by x86 and C respectively. The centrepiece of our formalisation is a decompilation relation that defines how the witness faithfully mimics the executable, and under what conditions. Together these components add up to a semantically-justified type-based decompiler. In summary, we make the following contributions:

1. We present a novel semantics-driven approach to type recovery that validates the types inferred for a low-level (MinX) program with a high-level witness (MinC) program. Unique to our work is a rigorous connection from our MinX binary to our MinC witness, founded on three key semantic components:
   (a) an SOS for MinX, designed as an abstraction of x86 to elucidate crucial control-flow details, such as the argument passing convention, needed to show that the MinX and MinC memories remain truly in sync;
   (b) an SOS and static type system for MinC, a type-safe dialect of C designed to illustrate decompilation of pointer arithmetic and the recovery of recursive structures;
   (c) a decompilation relation, that conservatively specifies when a MinX program corresponds to a MinC program.

2. In two steps we formally prove that the MinX program inhabits the recovered types:
   (a) First we show that the witness program is type-correct for the derived types.

3. The MinX Language

We introduce a lean instruction set, called MinX, to illustrate our semantic construction. MinX, by design, abstracts away from control-flow details that distract from the main goal of recovering recursive datatypes and that would otherwise make the presentation unmanageable.

Hence, the instruction set supports an unbounded number of registers. This means that register spilling does not need to be considered and SSA conversion [29] can be avoided. Also MinX fixes a single abstract calling convention, which avoids the need for argument detection techniques and control-flow recovery [11][2]. These restrictions can be relaxed in actual binaries by preprocessing steps.

3.1 Memory

An important challenge that we do not sidestep is that of differently-sized values. At the binary level, the field of a structure is accessed by a byte offset, which depends on the sizes of the data objects that precede it. This layout problem is intrinsic to type recovery and thus we assume memory is organised into bytes, and permit data objects to straddle contiguous bytes. We thus let $\text{Bit} = \{0, 1\}$ and define $\text{Byte} = \text{Bit} \cup \{\perp\}$ where $\perp$ denotes a single byte of uninitialised (random) memory. A word is then a vector of bytes, that is, $\text{Word} = \text{Byte}^\mathbb{N}$. The operator : concatenates vectors of bytes (and single bytes).

Register Bank The set of registers, each of which is of word size, is denoted $\text{Regs} = \{r_0, r_1, \ldots\}$. A function $R : \text{Regs} \rightarrow \text{Word}$ maps registers to the words they contain. To manipulate 1 byte and 2 byte objects within a register, we introduce an accessor function $R_{i,j} : \text{Regs} \rightarrow \text{Byte}^\mathbb{N}$ which slices from byte $i$ (counting from zero) up to but not including byte $j$ of a given register $r$. This is defined by $R_{i,j}(r) = \vec{b}$ where $R(r) = \vec{b} : \vec{b}^\prime$ and $\vec{b}_i \in \text{Byte}^i$, $\vec{b}_j \in \text{Byte}^{j-i}$, $\vec{b}_j^\prime \in \text{Byte}^{4-j}$.

For the register map, $R$, we define a notion of partial update in which only the least significant $w$ = $|\vec{b}|$ bytes of register $r$ are updated with the bytes $\vec{b}$ as follows:

$$R_{w}(w) \{ r \mapsto \vec{b} \} = R \circ \{ r \mapsto \vec{b} : R_{w-4}(r) \}$$

Heap Memory The heap is modelled as a (partial) function $H : \text{Word} \rightarrow \text{Byte}$ and therefore is byte addressable. To read stored objects that straddle $w$ consecutive bytes we define a function $H^w : \text{Word} \rightarrow \text{Byte}^w$ that reads and amalgamates $w$ bytes of the heap into a single vector as follows:

$$H^w(a) = \left\{ \begin{array}{ll} \{\perp\}^w & \text{if } \perp \in a \\ H(a + 4w) : \cdots : H(a) & \text{otherwise} \end{array} \right.$$  

where the operations $+$ and $-4$ denote addition and subtraction in 4 byte bit-vector arithmetic, and $1$ denotes a 4 byte bit-vector.

3.2 Syntax

The syntax of MinX programs consists of four syntactic categories. The first, $w ::= 2 | 4$, denotes the width, in bytes, of the primitive data objects that are supported by the instruction set. The second, $i$, defines the instructions themselves:

$$i ::= \text{mov}_w r, c \mid \text{mov}_w r, [r] \mid \text{mov}_w r, [r] + c \mid \text{eq}_w r, r, r_k \mid \text{op}_w r, r_j \mid \text{alloc} r, r_j$$

where $c$ denotes a numeric constant and $a \in \text{Word}$ is the location of the function that is to be invoked. (A Harvard architecture is assumed throughout). Square brackets indicate indirection. The instructions $\text{op}_w$ are themselves parameterised by the

Figure 1: System overview
\[
\begin{align*}
\text{Figure 2: Structured Operational Semantics of M1N X programs}
\end{align*}
\]

categories \( \oplus \in \{+,-\} \) and \( \otimes \in \{+,\,-,\,*,\,/,\,.,\,\ldots\} \), and the width \( w \) of their operands. The third category, \( b \), defines blocks:

\[
b ::= \iota; b \mid (\text{if}_w r_i \text{ goto } a); b \mid \text{ goto } a \mid \text{ret}
\]

Observe that blocks are terminated by control instructions, but conditional jumps only arise within a block. The fourth category, \( d_x \), defines how functions are declared:

\[
d_x ::= (\text{if}_w r_i \text{ goto } a); b \mid \text{ goto } a \mid \text{ret}
\]

where \( \lambda_x : \text{Word} \rightarrow b \) is a partial mapping from addresses to blocks. Moreover, if

\[
\text{labels}_x(b) = \begin{cases} 
\{a\} & \text{if } s = \text{labels}_x(\text{goto } a) \\
\{a\} \cup \text{labels}_x(b') & \text{else if } b = (\text{if}_w r_i \text{ goto } a); b' \\
\emptyset & \text{else if } b = \iota; b' \\
\end{cases}
\]

otherwise.
we require \( \bigcup \{ \text{labels}_a(b) \mid a \rightarrow b \in \mathcal{L} \} \subseteq \text{dom}(\mathcal{L}_a) \) so that jump targets are contained within \( \mathcal{L}_a \). Finally, \( r_{\text{arr}} \) and \( r_{\text{loc}} \) denote vectors of (distinct) registers and \( a_0 \in \text{dom}(\mathcal{L}_a) \).

### 3.3 Structured Operational Semantics

Figure 3 presents our (mostly small-step) SOS for MINC as two judgements: \( \mathcal{R} \vdash (H, R, a) \rightarrow (H', R') \) and \( \lambda_a: \mathcal{R} \vdash (H, R, b) \xrightarrow{a} (H', R', b') \). The former details the behaviour of single instructions and the latter of blocks. Both are parameterised by a vector \( \mathcal{R} \) of register assignments (needed solely in the proofs) to state that data accessible from these (shadowed) registers is not mutated by a call. For reasons of space, we comment only on noteworthy details.

For brevity, in \( \text{ex-mov-rc} \), we write \( c \) for the vector of \( w \) bytes that represents it. Notice in \( \text{ex-mov-rr} \) how the least significant \( w \) bytes of the register \( r_i \) are mutated while the remaining \( 4 - w \) bytes are left intact. In the \( \text{ex-mov-ri} \) rule, the \( r_i \) register is set to 0 uninitialised values if \( r_j \) contains one uninitialised value. The rule \( \text{ex-mov-ir} \) requires the 4 bytes of \( r_i \) to be initialised (otherwise computation gets stuck). Note how the \( w \) low bytes of register \( r_j \) are copied to \( w \) consecutive bytes of the heap. Rule \( \text{ex-}+\mathbf{r}^\mathbf{r} \) is included to support pointer arithmetic, where the data objects are \( c \) bytes in size. Rules \( \text{ex-rrc} \) and \( \text{ex-rrr} \) get stuck on, for instance, division by zero.

Observe how \( \text{ex-alloc} \) initialises allocated memory (to indicate garbage). If the register \( r_i \) itself contains an uninitialised byte, then rule \( \text{ex-alloc-}^\mathbf{b} \) sets \( r_i \) to be uninitialised. The rule \( \text{ex-alloc-}^\mathbf{r} \) is employed to allocate an array of objects, each of size \( c \).

The \( \text{ex-call} \) is worthy of note, as it is particularly unusual. Four bytes of each of the registers \( r_i \) are copied to \( r_{\text{arg}} \), which is abbreviated to \( r_{\text{called}} \), and the return register, \( r_{\text{ret}} \), are set to uninitialised values. The entry block \( \mathcal{L}_a(a_0) \) is executed, and any subsequent block that leads on from it, until \( \text{ret} \) is encountered, whichupon the register values are restored and register \( r_j \) updated with the return value.

Observe how \( \text{ex-if-true} \) and \( \text{ex-if-false} \) get stuck if the decision register contains an uninitialised byte.

### 4. The MINC Language

MINC (Minimal C) is the language of our witness programs and the target of our decompilation. We do not use C directly because, even though C is expressive enough to capture the semantics of machine instructions, it lacks one crucial ingredient: type safety. In contrast, we designed MINC to be type-safe, and still be close enough to C to recover MINX programs.

#### 4.1 Syntax

MINC features three different kinds of types:

\[
\begin{align*}
t &::= \text{short} \mid \text{long} \\
\theta &::= t \mid \tau \mid \tau \rightarrow \theta \mid \theta \mid N
\end{align*}
\]

A primitive type \( t \) is either a short or a long integer. A compact type \( \theta \) is a type that can be assigned to a program variable, it is either a primitive type \( t \) or a pointer type \( \tau \). Finally, a general type \( \tau \) is either a compact type \( \theta \), an array type \( \theta[j] \) or a struct type identified by its name \( N \).

MINC programs themselves are defined in terms of the categories of statements \( s \), expressions \( e \) and values \( v \) as follows:

\[
\begin{align*}
s &::= (l ::= e); s \mid \ell ::= x \mid e ::= \ell \mid c \mid kx \mid f(x) \\
goto l &::= x \rightarrow c \mid \text{new } \theta \mid \text{new } \theta[e] \\
\text{return} &::= (e_1 \oplus e_2) \mid (e_1 \otimes e_2)
\end{align*}
\]

where \( \oplus \) and \( \otimes \) are defined as in MINX. Function declarations,

\[
d_c ::= f(x : \theta)(y : \theta, l, e_1, e_2)
\]

include a vector of arguments and their types \( \rightarrow \theta \), a vector of locals and their types \( \theta \), an entry label \( l \in \text{Label} \), a partial mapping of labels to statements, \( \lambda_c : \text{Label} \rightarrow b \), and an index \( j \) into \( y \) indicating which local variable holds the return value. Labels must be contained within \( \text{dom}(\lambda_c) \), analogous to what was previously defined.

### 4.2 Type System

Figure 4 defines the MINC type system as well-typing judgements \( \Sigma \vdash d, \Gamma_c : \Sigma \vdash s, \Gamma_c : \Sigma \vdash \ell : \theta \) and \( \Gamma_c : \Sigma \vdash e : \theta \) for the four syntactic sorts. As well as the variable environment \( \Gamma_c \), the judgements make use of \( \Sigma \), which maps struct names \( N \) to vectors \( \theta \) of their field types. We also assume a global environment \( \phi_c \), which contains all function definitions.

Because it admits type safety, the type system is stricter than that of C. For example, C allows any integer to be added to the address of a primitive type, which is not permitted in MINC. Unlike C, arrays are first-class types, which means that it is possible to pass them as arguments to functions and return them, without demoting them to simple pointers. The upshot of this is that in MINC it is possible to distinguish between \( \theta \) and \( \theta^\mathbf{r} \).

To regain some of C’s flexibility, without compromising on type safety, we have equipped MINC’s type system with subtyping (see Figure 3). For instance, an array \( \theta[^*] \) is a subtype of \( \theta^\mathbf{r} \), which allows the assignment of an array to a compatible pointer.

As evident in the rules \( t^{-} \) and \( t^{-} \), MINC literals are tagged: short values are tagged \( c_s \) and long values \( c_l \). Although in general pointer literals do not exist in MINC, an exception is the special value \( \bot \), equivalent to \( \bot \text{ULL} \) in C.

#### 4.3 Semantics

Figure 5 presents the semantics of MINC, which are defined (and related to MINX in Fig 7) in terms of some auxiliary functions: \( \text{sizeof}(\theta) \) gives the size of an object of type \( \theta \) in bytes:

\[
\text{sizeof}(\text{short}) = 2 \quad \text{sizeof}(\text{long}) = 4 \quad \text{sizeof}(\tau^\mathbf{r}) = 4
\]

The functions \( \text{untag}(\tau) = \tau \) and \( \text{tag}(\tau) = \tau \) remove and add a tag \( \mathbf{r} \). Furthermore, given a set of non-empty ranges \( \pi \subseteq \{[l, u] \mid 0 \leq l \leq u \leq 2^{32} - 1\} \) that represents a set of (disjoint) memory regions, addition of a short and an address is defined as:

\[
u +_{u} v = \begin{cases} 
\bot & \text{if } u = \bot \lor v = \bot \\
\text{tag}^*(s) & \text{else if } 3r \in \pi, \{m, s \} \subseteq r \\
\text{err} & \text{otherwise}
\end{cases}
\]
where \( s = \text{untag}_s(u) +_4 m \) and \( m = \text{untag}_m(v) \). Addition of address/short, long/address, and long/address are analogously defined, so that the sum of an integer and an address \( n_\ast \) must fall within the same range of \( \pi \) as \( n_\ast \). The sum of two longs is simply
\[
\vec{\rho}(f) = f(x:\theta)(y:\theta',l,\lambda_c,j), \quad \forall \epsilon_i \in c, \theta'_j \in \vec{\theta} : \Gamma_c: \sigma \vdash \epsilon_i : \theta'_j, \Sigma \vdash \theta' < : \theta, \Sigma \vdash \theta'_j < : \theta_j
\]

Figure 4: Type-correct MinC programs

4.4 Type Safety

The MinC language is a type-safe variant of C. This means that well-typed MinC programs do not get stuck. We can be formally precise about this property in the usual way, stating preservation and progress properties for the syntactic terms of MinC.

Safe type depends on the notion of a store typing \( \Psi \) that associates a type \( \theta \) with every address \( a \) in the store \( \sigma \). A store \( \sigma \) is well-typed, denoted \( \Gamma_c; \Sigma; \Psi \vdash \sigma; \pi \vdash a : \theta \)

Figure 6 defines the auxiliary judgements for well-typed addresses and values. Similar to a store, an environment \( \rho \) is well-typed, denoted \( \Gamma_c; \Sigma; \Psi \vdash \rho \), iff

\[
\forall(x:\theta) \in \Gamma_c : \Sigma; \Psi \vdash \rho(x) : \theta \text{ and } \forall(x:\theta) \in \Gamma_c : \Sigma; \Psi \vdash \rho(x) \neq 0
\]

For the sake of brevity, we state the two properties here only for expressions. The other propositions and all proofs can be found in the on-line appendix, which is available at [http://kar.kent.ac.uk/51459/](http://kar.kent.ac.uk/51459/).

**Proposition 1** (Preservation of MinC Expressions). If \( \Gamma_c; \Sigma; \Psi \vdash \rho, \Sigma; \Psi \vdash \sigma; \pi \) and \( \Gamma_c; \Sigma; \Psi \vdash e : \theta \) then for some \( \Psi^\prime \supseteq \Psi \):
\[
\Gamma_c; \Sigma; \Psi^\prime \vdash \rho \land \Sigma; \Psi^\prime \vdash \sigma; \pi \land \Gamma_c; \Sigma; \Psi \vdash e : \theta
\]

**Proposition 2** (Progress of MinC Expressions). If \( \Gamma_c; \Sigma; \Psi \vdash \rho, \Sigma; \Psi \vdash \sigma; \pi \) and \( \Gamma_c; \Sigma; \Psi \vdash e : \theta \) then:
\[
\Gamma_c; \Sigma; \Psi \vdash \rho \land \Sigma; \Psi \vdash \sigma; \pi \land \Sigma; \Psi \vdash e : \theta
\]
Figure 5: Structured Operational Semantics of MiNC programs.
5. Decomposition Relation

This section relates MINX and MINC programs through a decompilation relation. The relation is spread out over three judgements, one for each of the three main syntactic categories of MINX.

**Instruction Decomposition** Figure 7 defines the decompilation judgement \( Σ \vdash d_a \rightsquigarrow d_c \) which explains how to decompile a MINX instruction \( d_a \) into a MINC assignment \( d_c \). Additional parameters to the judgement are a variable mapping \( \Theta \) that relates MINX registers to MINC local variables, a MINC typing environment \( Γ \), for those local variables, and a set of MINC struct definitions \( Σ \).

We do not have space to explain every rule in detail, so we highlight a number of important aspects:

- The judgement is defined by syntax-directed rules, as is usually the case for compilation and elaboration relations. The main difference is that, as the latter usually define (partial) functions; they are deterministic with typically one rule per syntactic construct. Our decompilation judgement is non-deterministic and features multiple rules per syntactic construct, one for each distinct typing that can be assigned to the instruction. For instance, the three rules tr-mov-r1, tr-mov-r2 and tr-mov-r3 compile instruction \( \text{mov}_w, [r_i], [r_j] \) into either \( x := \ast y \), \( x := y[0] \) or \( x := y \rightarrow 0 \) depending on whether \( y \) has type \( \theta \ast θ \) or \( θ \ast N \ast θ \).

The instruction widths \( w \) play an important role in restricting the possibilities for the recovered MINC types. For instance, rule tr-rc-rc ensures that the width \( w \) of the arithmetic operation \( op_w \) is identical to the size of the recovered primitive type \( t \). Hence, a width of 4 gives rise to and a width of 2 to short.

On several occasions the rules have to make up for the difference in memory granularity between MINX and MINC. In particular, in MINC the stride between two array elements is always 1. However, in MINX, the same stride depends on the size of the elements. Hence, rules like tr-rc for array pointer arithmetic convert between a MINX stride of \( c = m \ast \text{sizeof}(θ) \) and a corresponding MINC stride of \( m \).

When allocating a statically known amount of memory the rules tr-alloc-rc1, tr-alloc-rc2 and tr-alloc-rc3 also exploit the sizes of types to determine whether a primitive type, a particular struct or an array is allocated. We only support the decompilation of dynamic memory allocation for the instruction alloc \( r_i, r_j \ast c \) where we can statically verify that the amount of allocated memory is a multiple of the memory size. This, among others, makes our decompilation relation conservative and incomplete. It is a price we gladly pay in order to provide strong guarantees about the validity of the recovered types.

**Function Definition Decomposition** The bottom half of Figure 8 defines the judgement \( Σ \vdash d_x \rightsquigarrow d_c \) that decomposes a MINX function definition \( d_x \) into a MINC definition \( d_c \). The single rule of this judgement sets up the variable and label maps, and decomposes the basic blocks with respect to an appropriate typing environment.

5.1 Meta-Theoretical Properties

Our decompilation relation satisfies two strong properties that justify its relevance: 1) the produced MINC witness program is well-typed, and 2) the witness has the same operational semantics as the original MINX program. Taken together these two properties give meaning to the statement that the original MINX program inhabits the recovered types.

5.1.1 Well-Typing

The first claim states that the recovered witness program is well-typed. This is asserted as three propositions, one for each of the judgements.

**Proposition 3 (Well-Typed Instruction Decomposition).** If \( Σ \vdash d_a \rightsquigarrow d_c \) then for some \( θ_1 \) and \( θ_2 \):

\[
Γ \vdash d_c : θ_1 \quad \wedge \quad Γ \vdash e : θ_2 \quad \wedge \quad Σ, l : θ_2 <: θ_1
\]

**Proposition 4 (Well-Typed Block Decomposition).** If \( Σ \vdash d_a \rightsquigarrow d_c \) then \( Σ \vdash s \).

**Proposition 5 (Well-Typed Definition Decomposition).** If \( Σ \vdash d_a \rightsquigarrow d_c \) then \( Σ \vdash d_e \).

For the proofs of these propositions, we refer to the appendix. Due to the type safety of MINC, it follows that the witness program is operationally well-behaved.

5.1.2 Memory Correspondence

The main semantic effect of both MINX and MINC programs is a transformation of program memory. However, because MINX and MINC programs act on very different memory structures, the semantic correspondence is that the local variables are store-mapped whereas the registers are not.

**Value Correspondence** Figure 9 defines the judgement \( Σ \vdash d_a : d_c \rightsquigarrow d_v \) that states the basic correspondence between a MINX byte sequence \( d_b \) and a MINC value \( d_v \). This judgement is parameterised by an address map \( μ_a \) that relates MINX and MINC addresses. The rules are obvious, relating 0 pointers, addresses, bottoms, and numeric values of the appropriate byte sizes.

**Registers versus Local variables** We denote that MINX register banks \( R \) and MINC local variables \( \bar{R} \) are pair-wise related with:

\[
μ_a; μ_π; Σ \vdash \bar{R} \rightsquigarrow \bar{R}
\]

The relation is parameterized by an address map \( μ_a \), register-variable maps \( μ_π \) and a store \( Σ \). The relation stands for:

\[
\forall (r : x)_w ∈ μ_Γ; : \exists n_a : (x : n_a) ∈ ρ_i : \exists ν : (n_a : v) ∈ σ \wedge \exists δ : \bar{b} = R_{i,0,w}(r) \wedge μ_a \vdash \bar{b} \rightsquigarrow v
\]

The relation expresses that any related register \( r \) and local variable \( x \) have associated values \( \bar{b} \) and \( v \) that are related. The main complication is that the local variables are store-mapped whereas the registers are not.
Heaps versus Stores The MINX heap $H$ and MINC store $\sigma$ are related with:

$$
\mu_a; \nu_a; \pi; \vec{\rho} \vdash H \circ \sigma
$$

This relation summarizes 6 different properties addressing 4 concerns. Firstly, as in the previous cases, this relation is parameterised by an address map $\mu_a$ that relates addresses in $H$ with addresses in $\sigma$. Obviously, these related addresses point to related values.

$$
\forall (a, n_a) \in \mu_a : \exists ! b, v : b = H_{0\cdot w}(a) \land v = \sigma(a) \land \mu_a \vdash b \circ \sigma v
$$

Secondly, we have to contend with the difference in granularity between MINX and MINC: While we can only address values as a whole in MINC, MINX addresses individual bytes and can point into the middle of a value. To bridge this gap, the address map $\mu_a$ only covers the addresses in $H$ that point at the first byte of a value. The complementary header map $\nu_a$ relates each address in $H$ (especially those pointing into the middle of a value) to the address of the first byte of the value and its width.

$$
\forall a \in \text{dom}(H) : \exists a', w : \nu_a(a) = (a', w) \land (a' + w) \geq a
$$

These header addresses are fixpoints of $\nu_a$:

$$
\forall (a, w) \in \text{range}(\nu_a) : \nu_a(a) = (a, w)
$$

Moreover, they are covered by $\mu_a$:

$$
\forall (a, w) \in \text{range}(\nu_a) : \exists n_a : (a : n_a) w \in \mu_a
$$

Thirdly, not all addresses in $\sigma$ are related to an address in $H$. This is a consequence of the discrepancy between registers and local variables: the store-mapped local variables (i.e., those tracked in $\vec{\rho}$ or $\rho$) have no counterpart in $H$. Hence, they need not have a counterpart in the relation.

$$
\forall n_\sigma \in \{\text{dom}(\sigma) - \text{range}(\vec{\rho})\} : \exists a, w : (a : n_\sigma) w \in \mu_a
$$

Finally, adjacent MINX addresses in a range tracked by $\pi$ must be related to adjacent addresses in MINX (taking into account the width $w$ of the value).

$$
\forall (n_\sigma, n_\sigma + c) \in \pi : \forall i \in [0, c - 1] : \exists a, a', w, w' : a + w = a' \land (a, n_\sigma + i) w \in \mu_a \land (a', n_\sigma + i + 1) w' \in \mu_a
$$

5.1.3 Semantics Preservation With the memory relations in place we can state one important aspect of semantics preservation: the original MINX program and the corresponding decompiled MINC program take related memories to related memories.

Proposition 6 (Preservation of Related Memory for Instructions). If

- $\mu_{\text{ho}}; \Gamma; \Sigma; \Psi \vdash l \leftarrow \ell := e$
- $\Gamma; \Sigma; \Psi \vdash \rho$
- $\Sigma; \Psi \vdash \sigma; \pi$
- $\mu_a; \nu_a; \pi; \vec{\rho}; \rho \vdash H \circ \sigma$
- $\mu_a; \vec{\rho}_1; \mu_{\text{ho}}; \sigma; \vec{\rho}_2 \vdash \vec{R}, R \circ \sigma, \vec{\rho}, \rho$
- $\vec{R} \vdash (H, R, i) \leftrightarrow (H', R')$

then for some $\mu'_a \supseteq \mu_a$ and $\nu'_a \supseteq \nu_a$:

- $\mu'_a; \vec{\rho}_1; \mu_{\text{ho}}; \sigma' \circ (a \rightarrow v) \vdash \vec{R}, R \circ \sigma, \vec{\rho}, \rho$
- $\mu'_a; \nu'_a; \pi; \vec{\rho}; \rho \vdash H' \circ \sigma' \circ (a \rightarrow v)$

A second aspect of the semantics preservation is that, if the MINX program does not get stuck, the MINC program may get stuck only through a violation of memory that is guarded by $\pi$.

Proposition 7 (Preservation of Progress for Instructions). If

- $\mu_{\text{ho}}; \Gamma; \Sigma; \Psi \vdash l \leftarrow \ell := e$
- $\Gamma; \Sigma; \Psi \vdash \rho$
- $\Sigma; \Psi \vdash \sigma; \pi$
- $\mu_a; \nu_a; \pi; \vec{\rho}; \rho \vdash H \circ \sigma$
- $\mu_a; \vec{\rho}_1; \mu_{\text{ho}}; \sigma; \vec{\rho}_2 \vdash \vec{R}, R \circ \sigma, \vec{\rho}, \rho$
- $\vec{R} \vdash (H, R, i) \leftrightarrow (H', R')$

then

- $\Sigma; \vec{\rho}; \rho \vdash (\sigma, \pi, \ell) \leftarrow e$ or
- $\Sigma; \vec{\rho}; \rho \vdash (\sigma, \pi, \ell) \leftarrow (\sigma', \pi', \ell) \vdash e$ and $\Sigma; \vec{\rho}; \rho \vdash (\sigma', \pi', \ell) \leftarrow e$ or
- $\Sigma; \vec{\rho}; \rho \vdash (\sigma, \pi, \ell) \leftarrow (\sigma', \pi', \ell) \vdash (\sigma'', \pi'', \ell)$.

There are similar such propositions for basic blocks and for function definitions. Again we refer to the appendix for the proofs.

6. Implementation The decompilation relation is a conceptual device, literally a relation, which details what it means for a MINX program to be in correspondence with a MINC program. An algorithm, however, can be derived for solving a problem by adding control to Horn
\[ \mu_t; \Gamma_c; \Sigma \vdash e \]
one provided they share the same byte offset \( c \), an action that both simplifies the store and performs propagation. This can be specified in CHR as:

\[
\text{struct}(N, C, M1, Ty1) \land \text{struct}(N, C, M2, Ty2) \iff M1 = M2, Ty1 = Ty2.
\]

Furthermore, given a CHR constraint \( \text{sizeof}(\theta, w) \) that holds iff \( w = \text{sizeof}(\theta) \), and two constraints \( \text{struct}(N', c, m, \theta) \) and \( \text{struct}(N, c + w', m', \theta') \) it follows \( m' = m + 1 \). This form of propagation can be realised in CHR using:

\[
\text{struct}(N, C1, M1, Ty1), \text{struct}(N, C2, M2, Ty2) \implies \text{nonvar}(C1), \text{nonvar}(C2), \text{sizeof}(Ty1, W), \text{nonvar}(W), C2 := C1 + W \mid M2 \# M1 + 1.
\]

CHR rules are likewise used to express the subtyping relation. In all, this gives a solver for computing a witness and its type, in less than 900 LOC, but more importantly, derives one that is faithful to the rules of the decomaption relation.

To generate input for the solver, the self-hosting ANSI C89 compiler gcc [30] was retargeted to generate MinX. Our derivative, dubbed minxcc, supports the core features of C89. To stay within MinX, minxcc applies some rather unusual transformations. Each constant string, which is normally encoded as a global variable, is converted into a function that returns a pointer to newly allocated heap memory, that contains the string. malloc (and its friends) are replaced by the alloc instruction, while calls to free are removed completely. Memory thus grows as execution proceeds, exactly as specified in Fig. 7. Mathematical operations such as \( + \) and logical operations such as \( \oplus \) are reduced to MinX operations using equivalences taken from Hacker's Delight [11].

As a sanity check, we wrote a Haskell interpreter for MinX, following the SOS semantics of Fig. 2. The results of interpreting the MinX code, on various inputs, were then checked against those obtained by compiling the benchmarks using gcc, which was taken as ground truth. For the satisfaction of going full circle, a translator was written in Haskell to convert the MinX witness program into C, for testing with gcc. Each benchmark was subject to these two levels of checking.

The solver requires input to be pre-processed and presented in the MinX language. Figure 10 lists (pretty-printed) MinX code for summing the elements of a linked list. The arguments, return register and locals, denoted \( r_{arg}, r_{ret}, r_{loc} \) in Section 5 correspond to \( (r1), r0, (r2) \) respectively in the code listing. The mapping \( \lambda_0 \) is represented using the .BB0 and .BB1 labels to directly tag their corresponding block. Note that blocks can overlap. The label...
<table>
<thead>
<tr>
<th>benchmark</th>
<th>LoC</th>
<th>original</th>
<th>recovered</th>
<th>time</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>structs</td>
<td>solns</td>
<td>structs</td>
</tr>
<tr>
<td>aatree</td>
<td>315</td>
<td>2.734</td>
<td>1 1</td>
<td>1</td>
</tr>
<tr>
<td>avltre</td>
<td>269</td>
<td>2.188</td>
<td>1 1</td>
<td>1</td>
</tr>
<tr>
<td>binheap</td>
<td>184</td>
<td>2.558</td>
<td>2 2</td>
<td>2/1</td>
</tr>
<tr>
<td>binomial</td>
<td>303</td>
<td>3.732</td>
<td>2 2</td>
<td>5/4</td>
</tr>
<tr>
<td>hashsep</td>
<td>256</td>
<td>1.017</td>
<td>2 4</td>
<td>6/5/6/5</td>
</tr>
<tr>
<td>hashquad</td>
<td>260</td>
<td>0.977</td>
<td>2 4</td>
<td>2/3/2/2</td>
</tr>
<tr>
<td>kdtree</td>
<td>112</td>
<td>0.891</td>
<td>1 1</td>
<td>1</td>
</tr>
<tr>
<td>leftheap</td>
<td>182</td>
<td>0.762</td>
<td>1 1</td>
<td>1</td>
</tr>
<tr>
<td>list</td>
<td>262</td>
<td>0.829</td>
<td>1 2</td>
<td>2/1</td>
</tr>
<tr>
<td>mergesort</td>
<td>135</td>
<td>0.644</td>
<td>1 1</td>
<td>1</td>
</tr>
<tr>
<td>pairheap</td>
<td>298</td>
<td>3.216</td>
<td>1 1</td>
<td>2</td>
</tr>
<tr>
<td>queue</td>
<td>188</td>
<td>0.960</td>
<td>1 1</td>
<td>1</td>
</tr>
<tr>
<td>redblack</td>
<td>317</td>
<td>2.918</td>
<td>1 2</td>
<td>3/2</td>
</tr>
<tr>
<td>sets</td>
<td>120</td>
<td>0.355</td>
<td>0 1</td>
<td>0</td>
</tr>
<tr>
<td>skip</td>
<td>239</td>
<td>2.616</td>
<td>1 1</td>
<td>1</td>
</tr>
<tr>
<td>sort</td>
<td>364</td>
<td>2.339</td>
<td>0 1</td>
<td>0</td>
</tr>
<tr>
<td>splay</td>
<td>332</td>
<td>2.648</td>
<td>1 1</td>
<td>1</td>
</tr>
<tr>
<td>stackar</td>
<td>161</td>
<td>1.680</td>
<td>1 1</td>
<td>1</td>
</tr>
<tr>
<td>stackli</td>
<td>140</td>
<td>1.270</td>
<td>1 2</td>
<td>2/2</td>
</tr>
<tr>
<td>treap</td>
<td>288</td>
<td>2.580</td>
<td>1 1</td>
<td>2</td>
</tr>
<tr>
<td>tree</td>
<td>208</td>
<td>1.104</td>
<td>1 1</td>
<td>1</td>
</tr>
</tbody>
</table>

Figure 12: Solutions and Recovered Structures

of the entry block, \(a_0\), is left implicit by adopting the convention that the first block is always the entry block.

Figure 11 presents the MiNX witness program generated by the solver, again pretty-printed for human comprehensibility since the solver represents the witness as an abstract syntax tree. The local variables, denoted \(y : \theta\) in Section 4 are given immediately before the entry block. The mapping \(\lambda_c\) is represented, again by using labels to tag the blocks. The index \(j\), used to identify which local variable is returned in Section 4, is identified by printing each return statement with the variable \(y_j\).

7. Evaluation

The solver was deployed on a suite of textbook [31] programs, chosen because of their use of data-structures. Figure 12 lists the benchmarks, complete with LoC for the C and the MiNX assembly files. The solns column records the number of type assignment and witness program pairs generated by the solver. The original structs column indicates the number of struct types defined in the benchmark, whereas recovered structs records the number of struct definitions in each of the solutions. Thus 6/5/6/5, for example, indicates that the first solution has 6 recovered structs, the second has 5, the third 6, and the fourth 5; backtracking enabling all solutions to be enumerated. The benchmarks were run on a single core Intel Atom Z540 at 1.86GHz with 2GB of RAM. The benchmarks, assembly files, and witnesses (which embed the recovered types) are all available in a second on-line appendix, that is available at http://kar.kent.ac.uk/51448/.

Observe that some benchmarks have more than one solution and some solutions have a different number of struct definitions than the original program. This is due to several factors: Homogeneous structs, where every (accessed) field has the same type, cannot be distinguished from arrays, and therefore can be typed either as an array or a struct. This issue is exhibited by the binheap benchmark.

When the same struct type is used in separate parts of a program (e.g., in functions that are never called) our decompiler generates distinct copies of the struct. In most cases the definitions are identical (as in the tree benchmark), however a function may not assign every field of a struct, leading to an under-constrained type assignment problem and a struct definition that omits the unaccessed fields, as in the treap and avltre benchmarks. Combining these issues can result in multiple solutions that differ in their types and number of structs, which arises in the binomial benchmark.

The key point, however, is not visible from the table: there exists one witness program whose regenerated types are identical to those of the original benchmark. Moreover, for benchmarks with multiple solutions, every witness has types compatible with those of the original program (in the sense that arrays are compatible with homogeneous structs). In addition, all recursive types in every witness are present in the original, and every recursive type present in the original appears in every witness. Furthermore, when translated back into C, each witness behaves as the original benchmark.

8. Related Work

A self-contained introduction to type recovery is given in [29], which summarises the problem as “The ... problem for a decompiler is to associate each piece of data with a high-level type”. The author, like others [6, 28], introduces a dataflow analysis over type lattice of primitive types, but accepted wisdom is to formulate type inference as constraint solving because dataflow analysis classically deals with unidirectional flows.

Dynamic type recovery

Dynamic techniques have been suggested for type recovery [15], in which types are reconstructed from an execution trace. Each memory location accessed by the program is tagged with a timestamp because the same location can store values of different types over its lifetime. Each location and timestamp pair is then assigned a type, in either the on-line or off-line phases of the type recovery algorithm. In the on-line phase known types are propagated to the pairs, as a value which inhabits a type, is stored in a location. However, the type may remain unknown until the control encounters a system call or a library call, or some machine instruction, whose arguments or operands expose the type. The on-line phase is thus augmented with an off-line phase which propagates type assignment against the control to resolve any unassigned pair. The method requires an oracle to supervise the selection of the trace, and an examination of one trace will fail to infer types that hold universally across the whole program.

Somewhat surprisingly, Bayesian unsupervised learning has been applied to recognise structure in memory images [5]. The memory image is scanned, looking for all pointers, which are then used to locate the positions of objects and their size, which are bounded by the distance to the next object. Unsupervised learning is then used to classify malware according to its memory layout, a technique that could be taken further with static type recovery.

Recurse data types

Mycro [21] recognised that type reconstruction could rule out inconsistent decompilation steps and thereby aid program reconstruction. This link is formalised in our decompilation relation that is the centrepiece of our formalisation. A self-contained introduction to type recovery is given in [29], which summarises the problem as “The ... problem for a decompiler is to associate each piece of data with a high-level type”. The author, like others [6, 28], introduces a dataflow analysis over type lattice of primitive types, but accepted wisdom is to formulate type inference as constraint solving because dataflow analysis classically deals with unidirectional flows.

Dynamic type recovery

Dynamic techniques have been suggested for type recovery [15], in which types are reconstructed from an execution trace. Each memory location accessed by the program is tagged with a timestamp because the same location can store values of different types over its lifetime. Each location and timestamp pair is then assigned a type, in either the on-line or off-line phases of the type recovery algorithm. In the on-line phase known types are propagated to the pairs, as a value which inhabits a type, is stored in a location. However, the type may remain unknown until the control encounters a system call or a library call, or some machine instruction, whose arguments or operands expose the type. The on-line phase is thus augmented with an off-line phase which propagates type assignment against the control to resolve any unassigned pair. The method requires an oracle to supervise the selection of the trace, and an examination of one trace will fail to infer types that hold universally across the whole program.

Somewhat surprisingly, Bayesian unsupervised learning has been applied to recognise structure in memory images [5]. The memory image is scanned, looking for all pointers, which are then used to locate the positions of objects and their size, which are bounded by the distance to the next object. Unsupervised learning is then used to classify malware according to its memory layout, a technique that could be taken further with static type recovery.

Recurse data types

Mycro [21] recognised that type reconstruction could rule out inconsistent decompilation steps and thereby aid program reconstruction. This link is formalised in our decompilation relation that is the centrepiece of our formalisation. His work was inspired by the desire to synthesise datatypes from register transfer language (RTL) code generated from BCPL, which itself is untyped, not distinguishing between arrays and structures. He discussed the issue of padding, which arises when some of the fields of a structure, but not all, can be inferred, as well as proposing a type unification algorithm for synthesising a recursive datatype when a type variable is unified with a term containing it.

This approach has been reexamined though the perspective of SMT [25, 26] using the theory of rational trees (cyclic unification) [10], so that the solved form can directly encode the inferred recursive datatypes. Operations such as addition can be assigned one of three possible types, depending on whether the operands are an integer and a pointer, or vice versa, or simply two integers. This case splitting can be encoded propositionally and the theory used.
SecondWrite \cite{balakrishnan2007analyzing} extends work on variable recovery \cite{balakrishnan2004divine} with so-called best effort pointer analysis \cite{blazy2013verified} Section 5.2 to infer some datatypes: they “dig into the points-to set to discover if it is pointing to an address which is declared as the starting point of a structure”. Generic types are used for symbols for which they cannot infer types, and type casts are introduced to convert the generic type to the actual types used in an operation.

Following the idea that “well-typed programs cannot go wrong”, type recovery has been muted as a check for the validity of low-level code \cite{dijkstra1972self-modification}. Recursive types are recovered using a rational-tree solver from the low-level typeless template code of a graph reduction machine. If the solver fails, the code is judged unsafe.

**Verified decompilation and disassembly**  Decompilation is not always to C: Java bytecode has been decompiled into recursive functions, based on type theory \cite{blazy2012verified}, which is amenable to formal reasoning. Worthy of particular note, is the decompilation of machine code into the language of HOL4 \cite{lindley2014trustworthy}. With a view to proving full functional correctness, machine code is decompiled into tail-recursive functions. These functions describe the effect of the machine code, yet offer a layer of abstraction above it. Properties proved for the function are, by an automatically derived theorem, related to the original machine code, so the decompiler does not need to be proved correct. Recursive predicates could be defined in HOL4 to assert in that memory conformed to a recursive datatype, but for the purposes of engaging with the reverse engineer, it seems more natural to decompile to a type-safe dialect of C.

Disassembly, the act of decoding the bit patterns of machine instructions into a textual representation, is itself non-trivial for self-modifying code. Self-modification is used to disguise malware but also arises in JIT compilation. In general, disassembly requires indirect jump targets to be computed, which can be approximated by abstract interpretation \cite{blazy2013verified}. In the case of self modifying code, each memory write needs to be checked to determine how it modifies the code base. Modifications to the code base themselves entail a form of abstract decoding in which the analyzer does not recover the exact instruction, but a collection of applicable instructions. Nevertheless disassembly has been formalised \cite{blazy2014verified}, though we consider self-modification to be beyond the scope of our study.

**Trusted compilation**  Further afield, is the wide body of work on trusted compilation, most notably represented by the CompCert project that produced a fully certified optimising C compiler \cite{blazy2014trusted}. The CompCert compiler transforms source code into machine code incrementally through a large number of intermediate languages, each of which is designed to handle a specific compilation stage such as common subexpression elimination, register allocation or control flow linearisation. Correctness is verified at each successive level of transformation with proofs created using a proof assistant. Where CompCert aims at proving semantic correctness of these optimisations, our interest is in type preservation and any witness, no matter how closely it mirrors the binary, is sufficient for our type correctness argument.

Typed Assembly Language (TAL) \cite{blazy2014trusted} represents another approach to trusted compilation, where the aim is to prove that type consistency is maintained through compilation to an assembly language that can be type checked to prove safety properties of the executable code. The limitation is that no machine exists that can execute TAL, so as a compromise the code is type checked either when assembled to machine code or by a runtime loader that recognises a special typed object format. In contrast we have provided a type inference algorithm for assembler that allows type checking without the presence of any explicit type information in the binary.

**9. Discussion**  System and library calls provide a rich source of type information, from both their arguments and return types \cite{dijkstra1972self-modification}. In the short term, we plan to harvest these types and exploit them in our solver. We will also relax the restrictions on the MiNX calling convention, using dataflow analysis to identify arguments, that is, what is written before and what is read after a function call \cite{balakrishnan2007analyzing, balakrishnan2004divine}. To aid readability, we will also recognise the shape of certain (recurring) structures to refine the auto-generated structure and field names.

We only recover types if there exists a well-typed witness that corresponds semantically to the binary. This is not a limitation; it is a deliberate design choice. While not a problem for our benchmarks, the binary could be compiled from source that includes cast conversions, which are erased in the binary, but induce type conflicts. With a view to deployment, we intend to extend MiNC to union types but preserve the type safety of MiNC by raising a type error if a union field value is out of range. Recent work on the automatic localisation of type errors \cite{blazy2013verified} has shown how (weighted) MaxSMT \cite{blazy2013verified} can be applied to compute type conflicts that are minimal, subject to a compiler-specific ranking criterion. The idea is to minimise the sum of the weights of the unsatisfied (type) constraints. Since type recovery can be formulated as SMT \cite{lindley2014trustworthy}, we intend to apply error localisation to introduce a union type whenever a conflict is encountered, but do so in a way that minimises impact on the witness. This is a medium term goal.

The step beyond inferring an arbitrary well-typed witness for the binary, is to infer a witness that is comprehensible to the reverse engineer. Enriching the decompilation relation with more rules, possibly even for sequences of instructions, would increase the class of MiNC programs that can be regenerated, and provide the solver with latitude to select one decompilation over another. As a long term research goal, we intend to explore how preferences \cite{blazy2013verified} can steer the solver towards emitting a more intelligible witness, though it is far from clear how this can be quantified.

**10. Conclusions**  We have answered the fundamental question of how to derive types from a binary executable that truly have semantic meaning. Our answer is both principled and unique in that it derives a high-level witness program in concert with the types (provided one exists). We prove that the witness inhabits the inferred types, and establish a notion of memory consistency with the binary, thereby showing that the binary conforms to the inferred types. Apart from establishing type correctness, which is a first step towards certified decompilation, the construction also yields a type-based decompiler. We have evaluated the decompiler on more than 20 textbook programs and, for all, have recovered a witness program in a type-safe dialect of C, complete with the original recursive datatypes.

**Acknowledgments**  This work was supported by grant EP/K031929/1 funded by GCHQ in association with EPSRC, and partly funded by the Flemish Fund for Scientific Research (FWO).

**References**


