

Type Flattening Obfuscation

Ta Thanh Dinh

tathanhdinh@gmail.com

Abstract—Beside data and control flow, high-level types are important in binary code analysis, particularly in decompilation. Some research papers have introduced methods to map machine-dependent objects into types of some C-like type system. For the obfuscation/anti-decompilation purpose, we present a technique which bypasses existing type recovery approaches. We have implemented a prototype obfuscating C compiler to demonstrate the technique, the compiler is given open source.

Index Terms—type recovery, decompilation, obfuscation

1. Introduction

Binary code *decompilation* [5] is to transform the low-level, machine-dependent code of a program into a high-level form, like code of a high-level language. In almost all academic research papers and commercial products, the target language is C. Similar to compilers, a modern binary code decompiler consists of many phases [5, 11]: disassembly, function boundary detection, immediate representation (IR) lifting, control-flow graph (CFG) recovery, high-level variables detection, type (i.e. variable types and function signatures) recovery, etc. Each phase requires particular but not independent [8] analysis techniques: the results of one can affect another. The analyzed program is transformed gradually into a higher-level, more abstract and more understandable representation.

In the opposite direction, binary code *obfuscation* is a method to protect the low-level code from being decompiled, or from being analyzed in general. Because the code analysis contains of different interdependent phases, the obfuscation [13, 26] can proceed at any of them, e.g. anti-disassembly (binary packer, self-modifying code), binary stripping, control-flow flattening, virtualization (for both data and control obfuscation)... just name a few. Basically, each obfuscation method consists of one or several *potent transformations* [9, 13] which hide certain properties of the code.

Context and problem. An optional feature of binary code decompilation is *type reconstruction*, namely to recover high-level types from machine-dependent objects [6, 11]. This is the research objective of some research papers [15, 18, 21, 22], and killing feature of commercial [32, 33] as well as open source [31] binary code analysis tools. Beside decompilation, types and particularly *function signatures* are also essential in numerous applications, e.g. static binary rewriting [14, 17] and raising [27, 29], see for example [19] for a more completed list. Thus the knowledge about types expand the attack surface since more analysis can be applied on the programs.

Contribution. Despite of successes in binary type reconstruction and the need of protecting function signatures, to the best of our knowledge there is no explicit effort in hiding type information. This paper presents a method for type obfuscation, the principal idea is based on the fact that the compiler does not need to preserve all information about high-level types (type erasure), then with specific tricks we can exploit the *semantics gap* between the high-level language and machine code to make some information very hard if not impossible to be recovered. We do not claim that all type information can be hidden, the attacker can eventually know some but it would be hard to distinguish the concrete underlying types from one to another, thus the proposed notion of *type flattening*.

We implement the tricks in *uCc* [35], an open source obfuscating C compiler which obfuscates function signatures. The functions in binaries generated by *uCc* can be perfectly analyzed by classical procedures (boundary detection, disassembling, CFG recovery, etc), only their signatures are obfuscated. That way, we can evaluate the effectiveness of type obfuscation tricks on function signatures while excluding unwanted obfuscation effects that may come from (bad) results of other analysis phases. We find that Mixed Boolean Arithmetic (MBA) expressions [12, 24] are a good match for the goal.

In summary, our contributions are as follows:

- We introduce the notion of *type flattening*, it aims at protecting a high-level property (types) of the program in contrast with classical methods which focus on lower properties as data or control flow.
- We implement an open source prototype compiler *uCc* to realize the ideas of obfuscation. We introduce in *uCc* a method to generate pointer aliases which would be hard to reverse thanks to MBA [12].
- We give also an implementation for the permutation polynomials of MBA while other open source state-of-the-art obfuscators (e.g. Tigress [36]) give only basic arithmetic encoding expressions. Other deobfuscation tools (e.g. Syntia [23], QSynth [30]) can profit *uCc* to test their capabilities of MBA simplification.
- We evaluate the binaries generated by *uCc* against decent decompilers, the results show that no one can detect correctly the underlying types of arguments on function signatures: the original types are indistinguishable from the highest types in the C's integer conversion rank.

2. Brief history of binary type inference

In statically typed languages, the compiler does not need preserve source code level type information in the gener-

ated machine code (type erasure), then type recovering requires special techniques. Before presenting the type obfuscation, we give a brief discussion about how current methods on binary type inference work, that gives some intuition about our bypassing technique.

Though a broad survey for research up until 2015 can be referenced in [19], it sustains a storage point of view bias: types are attached always with concrete storage primitives (e.g. registers, memory), there are no essential differences between types and data structures, so are the techniques to recover them. Actually, types are compile-time constraints, they may or may not have runtime storage imprints. An example is C's *type qualifier* (e.g. `const`, `restrict`), in general any *refinement type* should not leave storage traces, the same thing with generics. More concretely, as we will present in the section, the *low-level polymorphism* is a very specific problem that binary type recovery techniques have to deal with. Also, the survey lacks some important papers which are only published until later [21, 22].

We focus only on semantics-based approaches, recent research using machine learning [28] or statistical language model [20] are out of scope of the paper. We omit the phase of variable/function detection, which is an essential step before type recovering, more details on this subject can be referenced in [10]. We avoid also difficulties in disassembling, the binaries are supposed to be perfectly disassemblable.

From now on, unless otherwise stated, the target language is C, this is also the target language of almost all research papers and tools in the domain.

2.1. Initial work

Though earlier ideas have been proposed in another context [4], the research in recovering types from low-level languages may begin with the classic paper of Mycroft [6] for his interest of decompilation. The principal idea is inspired by the work of Damas-Hindley-Milner [1, 2, 3] in the ML language: types of variables and functions are checked/referenced automatically from how they are used in the program's source code. For example, given an expression

$$x + y$$

then at least x or y must have integer type, it is impossible that both of them are pointers since adding two pointers does not type check.

The method of Mycroft has several limits, as pointed out by Van Emmerik [11]. One of them comes from the fact that the low-level languages take care mostly on the value of the computation, then (the result of) an expression can be used in several ways and it behaves as different types (low-level polymorphism). Let's consider an assignment

$$p' = p + n$$

where $\vdash p : \text{ptr}(S)$ (p is of type pointer to a struct S) and $\vdash n : \text{int}$, Mycroft's rules derive $\vdash p' : \text{ptr}(S)$ since $p+n$ is considered as the offset calculation to access some element of an array of S . But $p+n$ can be also an offset calculation to access some field of type, e.g. `int`, of the struct S , then $\vdash p' : \text{ptr}(\text{int})$.

To overcome these problems, Van Emmerik has proposed a *data-flow based* (in contrast with Mycroft's *constraint based*) approach where type information of an object will be refined gradually, instead of binding it early to some fixed type. He proposed using *subtype lattices* to express the preciseness of type information: p' will not

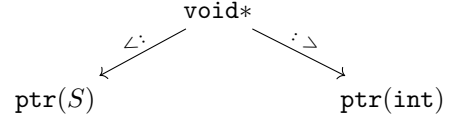


Figure 1. A subtype lattice

be early bound as $\text{ptr}(S)$, instead $\vdash p' : \text{void}^*$ (adding integer to pointer does not always result in pointer of the same type) and $\text{ptr}(S) <: \text{void}^*$. The precise type is only assigned later, when enough constraints are derived from other uses, e.g.:

$$*p' = m$$

where $\vdash m : \text{int}$, it then derives $\vdash p' : \text{ptr}(\text{int})$, finally $\vdash p' : \text{ptr}(\text{int})$ since $\text{ptr}(\text{int}) = \text{ptr}(\text{int}) \sqcap \text{void}^*$.

The lack of an IR with well-defined semantics limits Van Emmerik's work, he had to use ad-hoc type patterns to recognize and propagate type/subtype relations.

2.2. Improvement

Lee et al. in TIE [15] had the same idea of using type lattice for type preciseness, their deduction rules are more detail and support more cases (e.g. calls and dynamic jumps) but basically similar to Van Emmerik's. For example, the previously discussed assignment

$$p' = p + n$$

will generate $\vdash \text{ptr}(T_\beta) <: \tau_{p'}$ where T_β is a type variable and $\tau_{p'}$ means type of p' , but T_β is not free (never used outside the assignment) then this constraint is equivalent with $\vdash p' : \text{void}^*$. The notable improvement is the use of an IR named BIL (BAP Instruction Language), this makes the type analysis simpler and more coherent.

Polymorphism. The approaches discussed until now only consider basic cases of *low-level polymorphism*, e.g. adding a pointer to an integer may result in a pointer of the same type or not, but there are more. For example, `mov` can freely move data between signed and unsigned values, or even a constant can behaves as different types: zero is an integer, but it can be also a `NULL` pointer. Another case is the indistinguishability between a pointer to a struct and a pointer to the first field of this struct. All come from the low-level appearance of *type casting*, more details can be referenced in [7].

Noonan et al. handled these problems in Retypd [21] by first using subtyping in almost all derived constraints. The effect of data moving $x = y$ will be represented by $\vdash \tau_y <: \tau_x$. More importantly, they proposed a *type capability* model: each type variable is attached with several labels representing it capabilities. For example, the pointer dereference and assignment

$$x = *p$$

will result in $\vdash \tau_p.\text{load} <: \tau_x$, means p is a readable pointer (`.load` label), and the type of the dereferenced value is a subtype of type of x . The labels on τ_p allows to represent constraints on the inner structure of p (if exists) and p itself. Retypd used lattices for subtype relations, and type analysis is proceeded on an IR, similar with TIE.

2.3. Existing implementations

Only Van Emmerik gives an open source implementation of type recovery in his Boomerang decompiler, Lee et al and Noonan et al. do not. Published recently, Ghidra [31] is an open source decompiler which has type recovery, we do not know how it works yet. Other open source decompilers, Snowman [34] or RetDec [25], do not seem focus much on this kind of analysis. There are also commercial tools whose methods are not published, most notably Hex-Rays [32] and JEB [33].

3. Type obfuscation

In this section, we present our proposal for type flattening obfuscation and techniques used to obtain that notion. We focus only on obfuscating scalar types (pointers, integers, but not floats), supporting aggregate types (e.g. struct, nested struct) is still the ongoing work.

3.1. Type flattening

The type recovery approaches proposed several techniques to deal with the problem of low-level polymorphism, but a common point is to use some *subtype lattice* for the preciseness of inferred types. In the lattice, the bottom type \perp means that the variable violates some constraints in the type system [15]. Ideally, \perp should not occur since in the worst case, the decompiler can simply simulate the “weak” type system of the low-level language, we consider only \top .

The top type \top means universal or any, intuitively if a variable is of type \top then we only know the most trivial information about its type. The idea of *type flattening* is similar, removing useful information about type of an object means making the type recovery algorithm infer the object’s type as \top .

Definition 1. A high-level object is called *type flattened* up to a type inference algorithm with subtyping if its type is inferred as \top in the subtype lattice of the algorithm.

Uncertainty. Unsurprisingly, under some real world conditions, \top type does not mean we do not know anything, we actually know some properties. Recall that in our context, the binaries are disassemblable, function boundaries can be recognized correctly. Thus the binary, as our goal is to make it reusable, must respect the ABI (Application Binary Interface). For example, AMD64 System V ABI specifies that the first parameter of a function is passed via `rdi` register, thus in the worst case of the binary type inference, the type of the first argument is `size64`. The actual type may be `char*`, `signed32`, `unsigned16`, etc. but it is always subtype of `size64` (see fig. 2). This is actually what are being performed by some binary raising projects [27, 29] and decompiler [16].

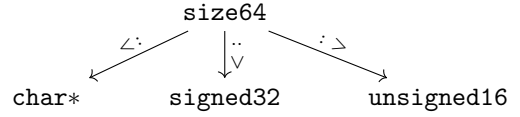


Figure 2. \top as `size64`

Goal. Our goal in type flattening obfuscation is to make any inferred type become \top . In our specific context, under AMD SystemV ABI, it is `size64`.

3.2. Data-flow disrupting

Though quite different about type systems, constraint generation rules and constraint solvers; later type recovery techniques follow the *data-flow based* approach proposed by Van Emmerik. Actually, most of implementation code in TIE is for the data-flow analysis and CFG building. Retypd uses an external abstract interpreter which does the data-flow analysis, this interpreter contributes greatly to the preciseness of the type inference.

A direct technique for type flattening is to disrupt the data-flow analysis, that is done by introducing *pointer aliases*. Let’s look at the piece of code in listing 1, for illustration purpose we have used a pseudo-C syntax but only put type annotation on the variables whose types are supposed to be already known.

```
foo(signed32 x) {
    p = &x;
    y = *p;
    return y;
}
```

Listing 1. Direct data movement

Any type recovery method can find the return type of `foo`. For example, with TIE: $\vdash p: \text{ptr}(\text{signed32})$ from the first assignment, then the second infers $\vdash y: \text{signed32}$. Retypd gives a slightly different result $\vdash \text{signed32} <: y$ where $y <: \text{size32}$, which is also more precise.

In listing 2, we insert an alias q for p to hide the data-flow. TIE derives $\vdash y: \text{size32}$, while Retypd is not better $\vdash \text{size32} <: y$ where $\vdash y <: \text{size32}$.

```
foo(signed32 x) {
    p = &x;
    q = an alias of p;
    y = *q;
    return y;
}
```

Listing 2. Data movement with pointer alias

Our compiler *uCc* disrupts the data-flow by: before an use (which needs to be protected) of a variable, *uCc* creates a stack slot to store it, generates an alias for the stack slot, then loads the variable from the alias.

3.3. Function signature masquerading

While types of local variables can be obfuscated by hiding the data-flow, obfuscating types of function parameters needs different techniques. For inter-procedural type analysis, the data-flow disrupting still works because the type of passed arguments are already hidden: they are local

variables of the calling function. But for intra-procedural type analysis, we can still hide the sign property (signed, unsigned) or whether they are pointers or not, but we cannot hide their size (recall that our goal is to obtain \top for all referred types). That is because of the size of a function parameter can be detected by looking at the register used to access this parameter inside the function. For example, if the first argument is of type `int` (i.e. `signed32`) then the register `edi` is used, not `rdi`. The parameter size is masqueraded by the following techniques:

Signature rewriting. For each function, *uCc* keeps an original copy of the signature but creates another signature where size of each parameter is extended to `size64`. The masqueraded signature is used for any call to the function in the translation unit, so for from an external (but in the translation unit) view, the function behaves as it has the masqueraded signature. The original signature is used inside the function.

Inner trampoline. In the function, *uCc* creates a trampoline as the pseudo-entry basic block, this trampoline uses the masqueraded signature to retrieve the parameters but then convert them to the real ones using the original signature. The conversion applies the data-flow disrupting to hide the source of the parameters. The real parameters then passed into the real entry basic block.

Semantics gap. The techniques used for function signature masquerading have actually exploited a *semantics gap* between the high-level and the machine language. In the high-level language, a function parameter has some high-level type and the program will not type check if this type violates the type system (e.g. passing an integer in a function whose the parameter type is pointer). But the ABI does not have such a constraint, it simply states that, for example, the first parameter is passed via `rdi`, no matter what its type is. Basically, the signature rewriting and inner trampoline have exploited the low-level polymorphism to wrap a type by its supertype.

3.4. Other techniques

uCc uses also some supplemental techniques to augment its resistance from type analysis algorithms and deobfuscation efforts in general.

Split and merge. To bypass the type constraints generated from `load/store` operations in TIE, as well as the derived types `.load/.store` in Retypd. *uCc* may split the loading of a value into several loads, each of them retrieves a part of the value, then merges these parts.

Code diversity. *uCc* is a probabilistic compiler, the generation of pointer aliases is not the same over generated binary codes: the coefficients and even the degree of the invertible polynomial are generated randomly for each run of *uCc*. The number of split parts in a value loading is randomized, as well as the size of each part.

4. Implementation

In implementing *uCc*, beside the principal goal is type flattening obfuscation, we try to avoid any other obfus-

$$\begin{aligned}
 P(x) &= 576460752303423488 \\
 &+ 8860833797731488269 \times x \\
 &+ 137438953472 \times x^2 \\
 &+ 288230376151711744 \times x^3 \\
 &+ 576460752303423488 \times x^4 \\
 &+ 68719476736 \times x^5 \\
 Q(x) &= -2882303761517117440 \\
 &+ 6993199436152966341 \times x \\
 &+ 8843495579821539328 \times x^2 \\
 &+ 4323455642275676160 \times x^3 \\
 &+ 6341068275337658368 \times x^4 \\
 &+ 2497183802250493952 \times x^5
 \end{aligned}$$

Figure 3. Invertible polynomial and its inverse

cation tricks that may bring unwanted effects on evaluating the obfuscation. Concretely, we want that the functions in binaries generated by *uCc* would be perfectly detectable, disassemblable and decompilable (but with flattened types). For these goals, we have to skip almost all popular obfuscation techniques [13, 26].

4.1. Pointer alias generation

The most "obfuscated" part of *uCc* is data-flow disrupting, thanks to results about invertible *permutation polynomials* [12], the generation of pointer aliases become straightforward. Basically, we generate randomly an invertible polynomial P over $\mathbb{Z}/2^n$ where n is the machine word bit size ($n = 64$), and its inverse $Q = P^{-1}$ (see fig. 3 for example), the aliases of some value v is $Q(P(v))$. The calculation of $Q(P(v))$ is embedded into the binary code.

5. Evaluation

6. Related work

Acknowledgment

References

- [1] J. R. Hindley. "The Principal Type-Scheme of an Object in Combinatory Logic". In: *Transactions of the American Mathematical Society* 146 (1969), pp. 29–60.
- [2] R. Milner. "A Theory of Type Polymorphism in Programming". In: *Journal of Computer and System Science* 17 (1978), pp. 348–375.
- [3] L. Damas and R. Milner. "Principal Type-Schemes for Functional Programs". In: *POPL*. 1982.
- [4] O. Shivers. "Data-Flow Analysis and Type Recovery in Scheme". In: *Topics in Advanced Language Implementation*. MIT, 1990.
- [5] C. Cifuentes. "Reverse Compilation Techniques". PhD thesis. 1994.
- [6] A. Mycroft. "Type-Based Decompilation (or Program Reconstruction via Type Reconstruction)". In: *ESOP*. 1999.

TABLE 1. FUNCTIONS AND ORIGINAL TYPES

Description	Signature
identity	<code>int id(int)</code>
division	<code>int div(short, char)</code>
modular	<code>char mod(short, char)</code>
increase $p \leftarrow p + 1$ then dereference	<code>int inc_deref(int *p)</code>
strlen	<code>int slen(char* s)</code>
sdbm hash	<code>int sdbm(char *s)</code>
djb2 hash	<code>int djb2(char *s)</code>
sum of array a	<code>long sum(short *a, int n)</code>
n -th fibonacci number (recursive impl.)	<code>int fibo(int n)</code>
num. of steps until $n \rightsquigarrow 1$ (Collatz's conj.)	<code>int collatz(int n)</code>

- [7] M. Siff, S. Chandra, T. Ball, K. Kunchithapadam, and T. Reps. "Coping with Type Casts in C". In: *FSE*. 1999.
- [8] B. Schwarz, S. Debray, and G. Andrews. "Disassembly of Executable Code Revisited". In: *WCRE*. 2002.
- [9] M. Dalla Preda and R. Giacobazzi. "Semantic-Based Code Obfuscation by Abstract Interpretation". In: *ICALP*. 2005.
- [10] G. Balakrishnan and T. Reps. "DIVINE: DIScovering Variables IN Executables". In: *VMCAI*. 2007.
- [11] M. Van Emmerik. "Static Single Assignment for Decompilation". PhD thesis. 2007.
- [12] Y. Zhou, A. Main, Y. X. Gu, and H. Johnson. "Information Hiding in Software with Mixed Boolean-Arithmetic Transforms". In: *WISA*. 2007.
- [13] C. Collberg and J. Nagra. *Surreptitious Software: Obfuscation, Watermarking, and Tamperproofing for Software Protection*. 1st. Addison-Wesley Professional, 2009.
- [14] A. R. Bernat and B. P. Miller. "Anywhere, Any-Time Binary Instrumentation". In: *PASTE*. 2011.
- [15] J. Lee, T. Avgerinos, and D. Brumley. "TIE: Principled Reverse Engineering of Types in Binary Programs". In: *NDSS*. 2011.
- [16] L. Durfina, J. Kroustek, P. Zemek, and B. Kabele. "Detection and Recovery of Functions and their Arguments in a Retargetable Decompiler". In: *WCRE*. IEEE, 2012, pp. 51–60.
- [17] K. Anand et al. "A Compiler-level Intermediate Representation based Binary Analysis and Rewriting System". In: *EuroSys*. 2013.
- [18] K. ElWazeer, K. Anand, A. Kotha, M. Smithson, and R. Barua. "Scalable Variable and Data Type Detection in a Binary Rewriter". In: *PLDI*. 2013.
- [19] J. Caballero and Z. Lin. "Type Inference on Executables". In: *ACM Computing Surveys* 48.4 (2016).
- [20] O. Katz, R. El-Yaniv, and E. Yahav. "Estimating Types in Binaries using Predictive Modeling". In: *POPL* (2016).
- [21] M. Noonan, A. Loginov, and D. Cok. "Polymorphic Type Inference for Machine Code". In: *PLDI*. 2016.
- [22] E. Robbins, A. King, and T. Schrijvers. "From MinX to MinC: Semantics-Driven Decompilation of Recursive Datatypes". In: *POPL*. 2016.
- [23] T. Blazytko, M. Contag, C. Aschermann, and T. Holz. "Syntia: Synthesizing the Semantics of Obfuscated Code". In: *USENIX Security*. 2017.
- [24] N. Eyrolles. "Obfuscation with Mixed Boolean-Arithmetic Expressions: reconstruction, analysis and simplification tools". PhD Thesis. Université Paris-Saclay, 2017.
- [25] J. Křoustek, P. Matula, and P. Zemek. "RetDec: An Open-Source Machine-Code Decompiler". In: *Botconf*. 2017.
- [26] S. Banescu and A. Pretschner. "A Tutorial on Software Obfuscation". In: *Advances in Computers*. Vol. 108. Elsevier, Jan. 2018, pp. 283–353.
- [27] P. Goodman and A. Kumar. "Lifting program binaries with McSema". In: *ISSISP*. 2018.
- [28] A. Maier, H. Gascon, C. Wressnegger, and K. Rieck. "TypeMiner: Recovering Types in Binary Programs Using Machine Learning". In: *DIMVA*. 2019.
- [29] S. B. Yadavalli and A. Smith. "Raising Binaries to LLVM IR with MCTOLL (WIP Paper)". In: *LCTES*. 2019.
- [30] R. David, L. Coniglio, and M. Ceccato. "QSynth - A Program Synthesis based approach for Binary Code Deobfuscation". In: *BAR*. 2020.
- [31] *Ghidra*. URL: <https://ghidra-sre.org/>.
- [32] *Hex-Rays Decompiler*. URL: <https://www.hex-rays.com/>.
- [33] *JEB Decompiler*. URL: <https://www.pnfsoftware.com/>.
- [34] *Snowman decompiler*. URL: <http://derevenets.com/>.
- [35] T. D. Ta. *uCc: an untyped C compiler*. URL: <https://github.com/tathanhdinh/uCc>.
- [36] *Tigress: A Source-to-Source-ish Obfuscation Tool*. URL: <https://tigress.wtf>.

TABLE 2. TYPE FLATTENING EFFECT

Original type	Hex-Rays	JEB	Ghidra	Snowman	RetDec
int id(int)	int64 id(int64)	ulong id(ulong)	ulong id(void)	int64 id(int64)	int64 id(int64)
short div(short, char)	int64 div(uint64, uint64)	div_t div(int, int)	div_t div(int, int)	int64 div(int64, int64)	int64 div(int64, int64)
char mod(short, char)	int64 mod(int64, int64)	ulong mod(ulong, ulong)	ulong mod(void)	int64 mod(int64, int64)	int64 mod(int64, int64)
int inc_deref(int*)	int64 inc_deref(int64)	ulong inc_deref(ulong)	ulong inc_deref(undefined8)	int64 inc_deref(uint64)	int64 int_deref(int64)
int slen(char*)	int64 slen(int64)	ulong slen(ulong)	ulong slen(undefined8)	int64 slen(uint64)	int64 slen(int64)
int sdbm(char*)	int64 sdbm(uint64)	ulong sdbm(ulong)	ulong sdbm(undefined8)	int64 sdbm(uint64)	int64 sdbm(int64)
int djb2(char*)	int64 djb2(uint64)	ulong djb2(ulong)	ulong djb2(undefined8)	int64 djb2(uint64)	int64 djb2(int64)
long sum(short*, int)	int64 sum(int64, uint64)	ulong sum(ulong, ulong)	undefined8 sum(undefined8)	int64 sum(uint64, int64)	int64 sum(int64, int64)
int fibo(int)	int64 fibo(uint64)	ulong fibo(ulong)	ulong fibo(undefined8)	int64 fibo(int64)	int64 fibo(int64)
int collatz(int)	int64 collatz(int64)	ulong collatz(ulong)	ulong collatz(undefined8)	int64 collatz(int64)	int64 collatz(int64)