Composition Inference for UML Class Diagrams

Ana Milanova
Department of Computer Science
Rensselaer Polytechnic Institute

milanova@cs.rpi.edu

ABSTRACT

Knowing which associations are compositions is important in a tool for the reverse engineering of UML class diagrams. Firstly, recovery of composition relationships bridges the gap between design and code. Secondly, since composition relationships explicitly state a requirement that certain representations cannot be exposed, it is important to determine if this requirement is met by component code. Verifying that compositions are implemented properly may prevent serious program flaws due to representation exposure.

We propose an implementation-level composition model based on ownership and a novel approach for identifying compositions in Java software. Our approach is based on a static ownership inference; it is parameterized by class analysis and is designed to work on incomplete programs. We present empirical results from one instantiation of our approach. In our experiments, on average 40% of the examined fields account for relationships that are identified as compositions. We also present a precision evaluation which shows that for our code base our analysis achieves almost perfect precision—that is, it almost never misses composition relationships. The results indicate that precise identification of interclass relationships can be done with a simple and inexpensive analysis, and thus can be easily incorporated in reverse engineering tools that support iterative model-driven development.

Categories and Subject Descriptors

D.2.7 [**Software Engineering**]: Distribution, Maintenance, and Enhancement—*Restructuring, reverse engineering, and reengineering*; F.3.2 [**Logics and Meanings of Programs**]: Semantics of Programming Languages—*Program Analysis*

General Terms

Algorithms

Keywords

UML, points-to analysis, reverse engineering, ownership

1. INTRODUCTION

In modern software development design recovery through reverse engineering is performed often; in a typical iterative development process reverse engineering is performed at the beginning of every iteration to recover the design from the previous iteration [23].

UML class diagrams describe the architecture of the program in terms of classes and interclass relationships; they are scalable, informative and widely-used design models. While the UML concepts of class and inheritance have corresponding first-class concepts in object-oriented programming languages, the UML concepts of *association, aggregation* and *composition* do not have corresponding language concepts. Thus, while the reverse engineering of classes and inheritance hierarchies is straightforward, the reverse engineering of associations presents various challenges.

UML associations model relatively permanent interclass relationships; conventionally, they are implemented using instance fields of reference type [23] (e.g., an association from class A to class B is implemented using a reference field of type B in class A). Thus, reverse engineering tools infer associations by examining instance fields of reference type; however, the inference is often non-trivial. One challenge is the recovery of one-to-many associations implemented using pseudo-generic containers (e.g., `Vector`). Another challenge is the recovery of compositions. Modern reverse engineering tools such as Rational Rose do not address these challenges and produce inconsistent class diagrams (see Guéhéneuc and Albin-Amiot [19] for detailed examples). Clearly, this leads to a gap between design class diagrams and reverse engineered class diagrams which hinders understanding, round-trip engineering and identification of design patterns.

Towards the goal of bridging this gap, this paper proposes a methodology for inference of binary associations for UML class diagrams. Our major focus is the inference of composition relationships, which we believe is challenging and inadequately addressed in previous work. While the UML concept of aggregation is "strictly meaningless" [13, Chapter 5] (i.e., it has no well-defined semantics to distinguishe it from association), the UML concept of composition has a well-defined semantics that emphasizes the notion of *ownership*: a "composition is a strong form of [whole-part] association with strong ownership of parts by the composite and coincident lifetime of parts with the composite. A part may belong to only one component at a time" [38, Chapter 14]. Therefore, a composition relationship at design level states the requirement for ownership and no *representation exposure* at implementation level (i.e., the owned component object cannot be exposed outside of its composite owner object); if composition is implemented correctly ownership

must be preserved.

It is important to investigate techniques for recovery of composition relationships. Firstly, it helps bridge the gap between the design class diagram and the reverse engineered diagram. Secondly, since composition relationships explicitly state a requirement that certain representations cannot be exposed, it is important to determine if this requirement is met by component code. Verifying that compositions are implemented properly may prevent serious program flaws due to representation exposure such as the well-known `Signers` bug in Java 1.1.[1]

Therefore, the goals of this work are (i) to define an implementation-level ownership model that captures the notion of composition in design and (ii) to design an analysis algorithm that infers ownership and composition using this model. Our definition of implementation-level composition is based on the *owners-as-dominators* ownership model [9, 32]; in this model the owner object (the composite) should dominate an owned object (a component)—that is, all access paths to the owned object should pass through its owner. The owners-as-dominators model defines an ownership boundary for each owner; intuitively, an owned object may be accessed by its owner as well as other objects within the boundary of the owner (e.g., an owned object stored in an instance field may be passed to an owned container). As pointed out by Clarke et al. [9, 32] and observed during our empirical study, the owners-as-dominators model captures well the notion of composition in modeling.

We propose a novel static analysis for ownership inference. If the ownership inference determines that all objects stored in a field are owned by their enclosing object, the analysis identifies a composition through that field. Our approach works on incomplete programs. This is an important feature because in the context of reverse engineering tools it is essential to be able to perform separate analysis of software components. For example, it is typical to have to analyze a component without having access to the clients of that component. Our ownership inference analysis is parameterized by *class analysis*, which determines the classes of the objects a reference variable or a reference object field may refer to. We use the class analysis solution to approximate the possible accesses between run-time objects. Our work defines a general framework for ownership and composition inference; it encompasses a wide range of analyses with different degrees of cost and precision.

We have implemented one instantiation of this framework based on the well-known Andersen-style points-to analysis for Java [35]. We present empirical results on several components. In our experiments, on average 40% of the examined fields account for relationships that are identified as compositions. We also present a precision evaluation which shows that for our code base, the analysis achieves almost perfect precision—that is, it almost never misses composition relationships identified in our model. The results indicate that precise identification of interclass relationships can be done with a simple and inexpensive analysis, and thus can be easily incorporated in reverse engineering tools that support iterative development.

This work has the following contributions:

- We propose an implementation-level ownership and composition model that captures well the notion of composition in modeling.
- We propose a general analysis framework for static inference of ownership and composition relationships in accordance with our model; the analysis framework works on incomplete programs.
- We present an empirical study that evaluates one instance of our framework. The results indicate that precise identification of composition relationships can be achieved with relatively inexpensive analysis.

2. PROBLEM STATEMENT

Reverse engineering tools typically infer associations by examining instance fields of reference type in the code. In our model, an association relationship through a field f is refined as composition if it can be proven that all objects referred by f are owned by their enclosing object. Thus, given a suitable definition of implementation-level ownership and composition, our goal is to design a static analysis that answers the question: given a set of Java classes (i.e, a component to be analyzed) for what instance fields we observe implementation-level composition throughout all possible executions of arbitrary client code built on top of these classes? The output is a set of fields for which the relationship is guaranteed to be a composition for arbitrary clients.

The input to the analysis contains a set *Cls* of interacting Java classes. We will use "classes" to denote both Java classes and interfaces as the difference is irrelevant for our purposes. A subset of *Cls* is designated as the set of *accessible classes*; these are classes that may be accessed by unknown client code from outside of *Cls*. Such client code can only access fields and methods from *Cls* that are declared in some accessible class; these accessible fields and methods are referred to as *boundary fields* and *boundary methods*.

Sections 2.1 and 2.2 describe the ownership model and the notion of implementation-level composition based on it. Section 2.3 discusses some constraints to the model that allow more precise detection of ownership and composition.

2.1 Ownership Model

The ownership model is based on the notion of *owners-as-dominators* [9, 8, 32]. It is essentially the model proposed by Potter et al. [32] with several modifications that allow more precise handling of popular object-oriented patterns such as iterators, composites and factories [14]. In this model, each execution is represented by an *object graph* which shows access relationships between run-time objects:

- Let f be a reference instance field in a run-time object o. There is an edge $o \xrightarrow{f} o'$ in the object graph if and only if field f in o refers to o' at some point of program execution.[2]

[1] In Java 1.1 the security system function `Class.getSigners` returned a pointer to an internal array allowing clients to modify the array and compromising the security of the system.

[2] We require that all newly created objects appear in the object graph explicitly [9]. That is, at the point of creation a new object is stored in a new local variable; this does not change program semantics.

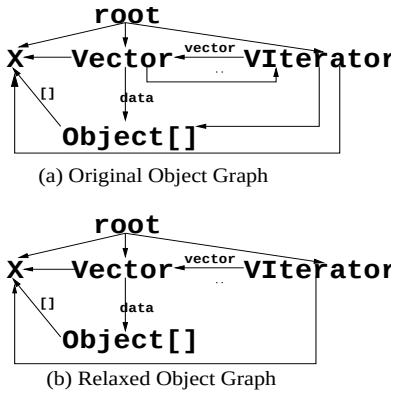
```
public class Vector {
 protected Object[] data;
 public Vector(int size) {
1 data = new Object[size]; }
 public void addElement(Object e,int at) {
2 data[at] = e; }
 public Object elementAt(int at) {
3 return data[at]; }
 public Enumeration elements() {
4 return new VIterator(this); }
}
final class VIterator implements Enumeration {
 Vector vector;
 int count;
 VIterator(Vector v) {
5  this.vector = v;
6  this.count = 0; }
 Object nextElement() {
7  Object[] data = vector.data;
8  int i = this.count;
9  this.count++;
10 return data[i]; }
}
main() {
11 Vector v = new Vector(100);
12 X x = new X();
13 v.addElement(x,0);
14 Enumeration e = v.elements();
15 x = (X) e.nextElement();
16 x.m();
}
```

Figure 1: Simplified vector and its iterator.

- There is an edge $o \xrightarrow{[]} o'$ if and only if some element of array o refers to o' at some point of program execution.
- There is an edge $o \rightarrow o'$ if and only if an instance method or constructor invoked on receiver o has local variable r that refers to o', or a static method called from an instance method or constructor invoked on o, has a local variable r that refers to o'. There is an edge of this kind only if there is no edge of the first kind from o to o'.

A run-time object o' is accessed in the *context* of o iff there is an edge from o to o' in the object graph. The start of program execution is expressed with a special node `root`. Context `root` represents the context for `main` and for objects referenced by static fields. For example, executing `main` in Figure 1 results in the object graph in Figure 2(a). Node `Vector` corresponds to the object created at the *new* site at line 11, node `Object[]` corresponds to the array created at the site at line 1, node `VIterator` corresponds to the iterator created at the site at line 4, and node `X` corresponds to the object created at the site at line 12.

The owners-as-dominators model states that the owner of an object o is the immediate dominator of o in the object graph [32].[3] Thus, according to this model `Object[]` is not owned by its enclosing `Vector` object for this execution due to the access relationship (although only temporary) between `VIterator` and `Object[]`. To make the model less restrictive, we introduce the *relaxed object graph* which omits edges due to certain temporary access relationships. We consider two kinds of temporary access relationships. The first kind arises when an object is created in one context and immediately passed to another context without being used; the relationship between the creating object and the new object is only temporary but if shown on the graph it is likely to restrict ownership. This notion captures the situations when an object is created and immediately returned (e.g., as in *return new VIterator*(*this*); in method `elements` in Figure 1) and when an object is created and immediately passed to another context (e.g., as in *new BufferedReader*(*new FileReader*(*fileName*))). This situation occurs in popular object-oriented design patterns such as factories, decorators and composites; in these cases the temporary relationship between the creating object and the newly created one is a matter of safety and flexibility of the implementation rather than an intention of the design. The second kind of temporary access relationships arises from field read statements $r = l.f$, where r is not assigned, passed as an implicit or explicit argument, or returned. This notion captures the situation that arises in iterators (consider statement *data* = *vector.data* in `nextElement` in Figure 1)—iterator objects have temporary references to the representation of their collections, which allows efficient access of collection elements; however, the collection object is always in scope. Therefore, if all accesses of o' in the context of o are due to such temporary access relationships, edge $o \rightarrow o'$ is not shown in the relaxed object graph.



(a) Original Object Graph

(b) Relaxed Object Graph

Figure 2: Object graphs for Figure 1.

The relaxed object graph for the execution of `main` in Figure 1 is shown in Figure 2(b). Edge `Vector`→`VIterator` is omitted because it is due to a temporary access relationship of the first kind; edge `VIterator`→`Object[]` is omitted as well because it is due to a temporary access relationship of

[3]Node m *dominates* node n if every path from the root of the graph that reaches node n has to pass through node m. The root dominates all nodes. Node m *immediately dominates* node n if m dominates n and there is no node p such that m dominates p and p dominates n.

the second kind. The owner of o is the immediate dominator of o in the relaxed object graph. Thus, `root` owns `X`, `Vector` and `Viterator` and `Vector` owns `Object[]`.

2.2 Implementation-level Composition

Let A be a class in *Cls*, and f be a field of type B declared in A where B is a reference type (class, interface or array type [15]). The ownership property holds for f if throughout all possible executions of arbitrary clients of *Cls*, every instance of A owns the instances of B that its f field refers to. Consider the case when f is a collection field—that is, all objects stored in the field are arrays or instances of one of the standard `java.util` collection classes (e.g., `java.util.Vector`). If every instance of A owns all corresponding instances stored in the collection, there is a *one-to-many composition* relationship between A and C, where C is the lowest common supertype of the instances stored in the collection[4]; otherwise, there is a one-to-many regular association. For collection fields for which the ownership property holds, there is an attribute of the association {*owned collection*} that indicates that the collection is owned by its enclosing object. Consider the case when f is not a collection field. If the ownership property holds for f, the association between A and B is a *one-to-one composition*; otherwise it is a regular one-to-one association.

Example. Consider the package in Figure 3. This example is based on classes from the standard Java library package `java.util.zip`, with some modifications made to simplify the presentation and better illustrate the problem and our approach. *Cls* contains the classes from Figure 3 plus class `ZipEntry`. The accessible classes are `ZipInputStream`, `ZipOutputStream` and `ZipEntry` and the boundary methods are all public methods declared in those classes (i.e., the component can be accessed from client code through the public methods declared in these classes).

Clearly, the `CRC32` objects are always owned by their enclosing streams. Thus, there is a one-to-one composition relationship between class `ZipInputStream` and class `CRC32` through field `crc`. Similarly, there is a one-to-one composition relationship between `ZipOutputStream` and `CRC32` through field `crc`. There is a regular one-to-one association through field `entry` in `ZipInputStream`; it is easy to construct client code on top of these classes such that the `ZipEntry` instances created in `ZipInputStream` objects are leaked to client code from `getNextEntry`. Similarly, there is a regular one-to-one association through `entry` in `ZipOutputStream` because the `ZipEntry` objects are passed from client code to `putNextEntry`. The associations through fields `names` and `entries` are both one-to-many regular associations between `ZipOutputStream` and `ZipEntry`; both have attribute {*owned collection*}. The `ZipOutputStream` instance trivially owns the `Hashtable` instance. It owns the `Vector` instance as well, although the `Vector` instance is referred to in the context of its iterator (recall the example in Figure 1); however, the iterator is a local object owned by the enclosing `ZipOutputStream` object which ensures that the `Vector` instance is dominated by the enclosing `ZipOutputStream` and may be accessed only within its own-

[4]Note that in Java, a unique non-trivial (i.e., non-`Object`) common supertype may not exist. A detailed discussion appears in [26].

```
package zip;

public class InflaterInputStream {
  protected Inflater inf;
  protected byte[] buf;
  public InflaterInputStream(Inflater inf,
   int size) {
    this.inf=inf;
    buf=new byte[size]; }
  public InflaterInputStream(Inflater inf) {
    this(inf, 512); }
  // methods read and fill contain instance calls on inf
}
public class ZipInputStream extends
 InflaterInputStream {
  private ZipEntry entry;
  private CRC32 crc=new CRC32();
  public ZipInputStream() {
    super(new Inflater(true), 512); }
  public ZipEntry getNextEntry() {
    crc.reset();
    inf.reset();
    if ((entry=readLOC())==null) return null;
    return entry; }
  private ZipEntry readLOC() {
    ZipEntry e=new ZipEntry();
    // code reads and writes fields of e
    return e; }
}
public class ZipOutputStream extends
 DeflaterOutputStream {
  private ZipEntry entry;
  private Vector entries=new Vector();
  private Hashtable names=new Hashtable();
  private CRC32 crc=new CRC32();
  public ZipOutputStream() {
    super(new Deflater(...)); }
  public void putNextEntry(ZipEntry e) {
    // code reads and writes fields of e
    if (names.put(e.name, e)!=null) { ...  }
    entries.addElement(e);
    entry=e; }
  public void closeEntry() {
    ZipEntry e=entry;
    // code reads and writes fields of e
    crc.reset();
    entry=null; }
  public void finish() {
    Enumeration enum=entries.elements();
    while (enum.hasMoreElements()) { ...  } }
}
```

Figure 3: Sample package `zip`.

ership boundary.

2.3 Discussion

In order to allow more precise detection of implementation-level composition, we employ the following constraint, standard for other problem definitions that require analysis of incomplete programs [36, 34]. We only consider executions in which the invocation of a boundary method does not leave *Cls*—that is, all of its transitive callees are also in *Cls*. In particular, if we consider the possibility of unknown subclasses, all instance calls from *Cls* could potentially be "redirected" to unknown external code that may affect the composition inference. For example, a field may be identified as composition in the current set of classes but an unknown subclass may override some method and the overriding method may leak the field (e.g., by passing it to a static field).

Thus, *Cls* is augmented to include the classes that provide component functionality as well as all other classes transitively referenced. In the experiments presented in Section 6 we included all classes that were transitively referenced by *Cls*. This approach restricts analysis information to the currently "known world"—that is, the information may be invalidated in the future when new subclasses are added to *Cls*. Another approach is to change the analysis to make worst case assumptions for calls that may enter some unknown overriding methods. However, in this case, the analysis will be overly conservative and likely report fewer compositions. Thus, we believe that it is more useful to restrict the analysis to the known world; of course, the analysis user must be aware that the information is valid only for the given set of known classes.

3. CLASS ANALYSIS FOR OBJECT GRAPH CONSTRUCTION

Class analysis determines the set of objects that a given reference variable or a reference object field may refer to. This information has a wide variety of uses in software tools and optimizing compilers. In this paper, class analysis information is needed to construct a graph that approximates all possible object graphs that can happen when arbitrary client code is built on top of *Cls*; subsequently the approximation of the object graph is used to infer ownership and composition relationships. There is a large body of work on class analysis with different trade-offs between cost and precision [30, 31, 1, 3, 18, 37, 41, 44, 43, 25, 35, 17, 27, 28, 47, 24, 5, 48]. In this paper, we propose a general framework for class analysis that encompasses a large number of analyses with varying degrees of cost and precision. The framework can be instantiated to relatively inexpensive and imprecise analyzes such as RTA [3] as well as to relatively expensive and precise analyses such as object-sensitive points-to analysis [27, 28] and call-string context-sensitive analysis. The object graph construction uses the output of class analysis—therefore, the precision and cost of constructing the object graph and subsequently inferring ownership and composition depend on the precision and cost of the underlying class analysis.

3.1 Generalized Class Analysis for Java

The two major dimensions of precision in class analysis are *flow sensitivity* and *context sensitivity*. Flow-insensitive analyses do not take into account the flow of control between program points and are less precise and less expensive than flow-sensitive analyses. Context-sensitive analyses distinguish between different calling contexts of a method and are more precise and more expensive than context-insensitive ones. Note that the term *context* was used in Section 2.1 when describing the notion of the object graph and it had a different meaning. Thus, it is warranted to make a distinction between the use of the term in Section 2.1 and the use of the term in the current section.[5] In the sense of the object graph, the context refers to the receiver of the method; the intuition is that the receiver controls the method and thus all objects that are referenced within that method are referenced by its receiver—that is, the objects are accessed in the *context of the receiver object*. In the sense of context-sensitive flow analysis, the context characterizes a *particular invocation of a method*; typically, the analysis keeps a copy of the method for each relevant context of invocation of the method. Popular context sensitivity schemes are the sequence of k enclosing call sites, the receiver object [27, 28] and the cartesian product of the set of classes at an invocation site [1]. Although certain context sensitivity schemes in flow analysis, such as using the receiver object as context as outlined below, are related to the meaning of the term in the sense of the object graph, context in flow analysis stands for a more general concept; for our purposes we regard the two uses of the term as independent. For the rest of the paper we carefully clarify each use of the term context.

Other dimensions of precision include *field sensitivity*, *directionality*, the *call graph construction scheme*, the *reference representation scheme* and the *object naming scheme*. Field-sensitive analyses are able to distinguish flow through different object fields while field-insensitive analyses merge flow through different fields of an object; thus field-sensitive analyses are more precise than field-insensitive ones. Directional analyses, also referred to as propagation-based analyses, process assignments in one direction, while bi-directional analyses, also referred to as unification-based analyses process assignments in both directions; for example a directional analysis processes assignment $l = r$ by propagating the set of classes for r into l; in contrast, a bi-directional analysis propagates the set of classes for r into l and the set of classes for l into r. With respect to call graph construction, the analysis may construct a call graph on-the-fly while propagating classes, or pre-compute the call graph using an inexpensive technique such as Class Hierarchy Analysis (CHA) [11].[6] Typically, the analyses that construct the call graph on-the-fly are significantly more precise than the ones that pre-compute the call graph.

Yet another dimension of precision is the reference representation scheme. This dimension refers to the number of analysis variables used to represent reference variables;

[5] We use the term in the sense of the object graph because it is a standard term in the ownership type literature [9, 32]; similarly, it is a standard, widely used term in the flow analysis literature [40, 20, 39].

[6] Class Hierarchy Analysis (CHA) examines the declared type of the receiver variable and the class hierarchy and computes a set of possible run-time targets.

more precise analyses use more variables, while less precise ones use less variables and thus "merge" information for different reference variables. A typical reference representation scheme is to use an analysis variable for each reference variable in the program. The object naming scheme refers to the number of object names used to represent heap objects. More precise analyses use more object names, while less precise ones use less names and thus "merge" distinct run-time objects. One popular naming scheme is to represent each object by its class. Another popular naming scheme is to represent each object by its allocation site. Note that the reference representation and object naming schemes are related to context sensitivity and the degrees of context sensitivity—a context-sensitive analysis typically defines a representation for reference variables and objects; however, as demonstrated in this paper one may vary the reference and object naming schemes within context-insensitive analysis as well. A detailed examination of a wide variety of class analyses and their dimensions of precision is given in [39].

In this paper we consider analyses that are *flow-insensitive*, *field-sensitive*, *directional* and construct the call graph *on-the-fly*. The analysis designer may vary the context sensitivity scheme, the reference representation and the object naming scheme in order to achieve analysis of the desired precision and cost. The generalized class analysis is defined in terms of four sets. Set R is the set of locals, formals and static fields of reference type. Set O is the set of object allocation sites; the objects created at an allocation site s_i are represented by object name $o_i \in O$. Set F contains all instance fields in program classes and set Cl contains all program classes. There are three analysis parameters. Set C represents the set of all method contexts and $C_m \subseteq C$ is the set of contexts that are valid for method m; for example, to represent context-insensitive analysis, there is a single context ϵ for each method in the program. Function $v(r, c) : R \times C \rightarrow V$ where V is the set of reference variable representatives, defines the reference representation scheme; for example, the most typical context-insensitive reference representation scheme uses an analysis variable for each reference variable in the program. It is expressed as follows: $v(r, \epsilon) : R \rightarrow R$ (i.e., we have $v(r, \epsilon) = r$ for every r). Finally, function $h(o_i, c) : O \times C \rightarrow H$ where H is the set of heap object representatives, defines the object naming scheme used by the analysis; as an example, when objects are represented by their allocation site we have $h(o_i, c) : O \times C \rightarrow O$ (i.e., we have $h(o_i, c) = o_i$). The analysis solution is a *points-to graph* whose edges represent the following "may-refer-to" relationships:

- An edge $v \rightarrow h$ in the points-to graph (also denoted by (v, h)) means that at run-time some reference variable $r \in R$ represented by $v \in V$ may refer to some object represented by $h \in H$.
- Let $f \in F$ be a reference instance field in objects represented by some h. An edge $h \xrightarrow{f} h_2$ (also denoted by $(\langle h, f\rangle, h_2)$) means that at run time field f of some object represented by $h \in H$ may refer to some object represented by $h_2 \in H$.
- If h represents array objects, $h \xrightarrow{[]} h_2$ (also denoted by $(\langle h, []\rangle, h_2)$) shows that some element of some array represented by $h \in H$ may refer at run time to an object represented by $h_2 \in H$.

$\langle s_i$ in m: $l = new\ C, G\rangle \Rightarrow$ foreach context $c \in C_m$
add $\{(v(l, c), h(o_i, c))\}$ to G

$\langle l = r, G\rangle \Rightarrow$ foreach context $c \in C_m$
add $\{(v(l, c), h) \mid (v(r, c), h) \in G\}$ to G

$\langle l.f = r, G\rangle \Rightarrow$ foreach context $c \in C_m$
add $\{(\langle h, f\rangle, h_2) \mid (v(l, c), h) \in G \wedge (v(r, c), h_2) \in G\}$ to G

$\langle l = r.f, G\rangle \Rightarrow$ foreach context $c \in C_m$
add $\{(v(l, c), h) \mid (v(r, c), h_2) \in G \wedge (\langle h_2, f\rangle, h) \in G\}$ to G

$\langle s_i$: $l = r_0.m(r_1), G\rangle \Rightarrow$ foreach context c
$\{resolve(G, m, c, h, r_1, l) \mid (v(r_0, c), h) \in G\}$

$resolve(G, m, c, h, r_1, l)$ =
`let` $m_j(this, p_1, ret_j) = dispatch(h, m)$ `in`
c'=*findNewContext*(s_i, c, h); add c' to C_{m_j}
add $\{(v(this, c'), h)\} \cup$ to G
add $\{(v(p_1, c'), h_2) \mid (v(r_1, c), h_2) \in G\}$ to G
add $\{(v(l, c), h_3) \mid (v(ret_{m_j}, c'), h_3) \in G\}$ to G

Figure 4: Generalized Class Analysis.

The generalized class analysis algorithm is given in Figure 4; it propagates may-refer-to relationships by analyzing program statements. For the majority of statements the effects of the analysis are straight-forward. Consider statement "$l = new\ C$". The analysis processes this statement separately for each context c of the enclosing method m. It applies $v(l, c)$ and finds the representative of l for context c; similarly, it applies $h(o_i, c)$ and finds the representative of the objects allocated at that site when the enclosing method is invoked under context c. Subsequently, it creates a points-to edge $(v(l, c), h(o_i, c))$. Similarly, for statement $l = r$ the analysis infers that for each context c of the enclosing method, the representative of l in c may refer to the objects that the representative of r in c refers to.

At virtual calls the analysis performs resolution based on each object h in the current points-to set of the receiver variable (i.e., it constructs the call graph on-the-fly based on the current class analysis information). It finds the run-time target m_j based on the class of the receiver object h and the compile-time target m. Subsequently, the analysis finds a new context c' for the target method m_j according to the context sensitivity scheme; for example, for 1-CFA analysis, which distinguishes context by the last enclosing call site, the new context is the call site i, and for 2-CFA analysis which distinguishes context by the last two enclosing call sites, c' is formed by extracting the last call site j from c and attaching j to i (i.e., forming a string of the last two enclosing call sites). The new context c' is added to the set of contexts C_{m_j}. The analysis appropriately propagates values from actuals to formals and from the return variable to the left-hand side of the call, taking into account the context c of the caller and the context c' of the callee.

3.2 Instances of the Generalized Analysis

The generalized analysis outlined above can be instanti-

ated to many existing class analyses, ranging from the inexpensive RTA analysis to relatively expensive and precise context-sensitive analyses. Below, we present four representative instantiations in order of increasing precision and cost: RTA, 0-CFA, Andersen-style points-to analysis and object-sensitive points-to analysis. The class analysis information output by these analysis can be used to construct the object graph and subsequently infer ownership and compositions. The points-to graphs for the example in Figure 3 computed by 0-CFA, the Andersen-style points-to analysis and the object-sensitive points-to analysis are discussed in Section 3.3; they are shown in Figure 6.

3.2.1 Rapid Type Analysis (RTA)

RTA is a popular form of class analysis primarily used for call graph construction [3]. Intuitively, it starts from the main method of the program and keeps a set of currently instantiated classes and a set of currently reachable methods. RTA analyzes two kinds of program statements: call sites and allocation sites. When it encounters a call site in a currently reachable method it examines all potential edges according to CHA and for each edge records the classes that trigger the edge. If at least one class that triggers the edge is in the set of instantiated classes, the edge becomes valid and the target method reachable. When RTA encounters an allocation site which instantiates class A, it adds A to the set of instantiated classes and makes valid all previously visited edges that are triggered by A.

In order to instantiate our framework we need to define the set of contexts C, sets V and H and functions $v(r, c)$ and $h(o, c)$. Clearly, RTA is a context-insensitive analysis and thus there is a single context $C = \{\epsilon\}$. RTA can be regarded as keeping a single variable v that represents *all* reference variables [44]. Thus, function $v(r, \epsilon)$=v for every r and we have $V = \{v\}$. Finally, RTA represents objects by their class. Thus, function $h(o, \epsilon)$=A where A is the class of object o (i.e., the class instantiated at allocation site o) and we have $H = Cl$. It is easy to see that this instantiation of the generalized analysis from Figure 4 is equivalent to RTA.

3.2.2 0-CFA

0-CFA [30, 44] is another well-known class analysis at the low end of the cost/precision spectrum; it propagates sets of classes to reference variables and reference object fields. It is context-insensitive, maintains a set for each reference variable and represents heap objects by their class. For example, for statement $l = new\ A$ the analysis adds class A to the set for reference variable l. Similarly, for statement $l = r$, the analysis propagates the set for variable r to the set for variable l. In our framework 0-CFA can be achieved by instantiating C to $\{\epsilon\}$, $v(r, \epsilon)$ to $v(r, \epsilon) = r$ and $h(o, \epsilon) = A$ where A is the class of o. Thus, we have $V = R$ and $H = Cl$. Clearly, 0-CFA is more precise than RTA because it keeps a separate analysis variable for each reference variable in the program; in terms of the other dimensions, it uses the same representation as RTA.

We consider two other analyses at the low end of the cost/precision spectrum; these analyses are inspired by the XTA-style analyses from [44]. These analyses vary the reference representation scheme between the singleton representation of RTA and the R representation of 0-CFA while keeping C and h the same as in RTA and 0-CFA; thus, the precision of these analyses varies between RTA and 0-CFA. The first instance, referred to as mTA, maps each reference variable r to a representative valid for the enclosing method m of r—that is, we have $v(r, \epsilon)$=v_m and $V = M$ where M is the set of all program methods. The second instance, referred to as cTA maps each variable r to a representative valid for the enclosing class of r—that is, we have $v(r, \epsilon)$=v_c and $R = Cl$. Therefore, RTA is the least precise analysis, followed by cTA, mTA and 0-CFA.

3.2.3 Andersen-style Points-to Analysis

So far, we considered analyses that represent heap objects by their class. Another group of class analyses, typically referred to as points-to analyses, represents heap objects more precisely, usually by allocation site. The Andersen-style points-to analysis for Java is a well-known flow- and context-insensitive analysis [24, 5, 35, 41]. It uses an analysis variable for each reference variable and represents heap objects by their allocation site. In terms of our framework we have $C = \{\epsilon\}$, $v(r, \epsilon) = r$ and $h(o, \epsilon) = o$; thus, we have $V = R$ and $H = O$. This analysis is at the heart of our implementation of ownership and composition inference as we believe it is the most suitable for that purpose.

3.2.4 Object-sensitive Points-to Analysis

Finally, we consider a context-sensitive points-to analysis, which is referred to as *object-sensitive* analysis [27, 28]. With object sensitivity, each instance method and each constructor is analyzed separately for each object on which this method/constructor may be invoked. More precisely, if a method/constructor may be invoked on run-time objects represented by object name o, the object-sensitive analysis maintains a separate contextual version of that method/constructor that corresponds to invocation context o. For static methods, the analysis uses a special context ϵ.

In terms of our framework, the analysis is defined as follows. The set of contexts C equals O—that is, the set of contexts is the set of object allocation sites. Map $v(r, c) = r^c$ where r^c is the context copy of r that corresponds to the invocation of the enclosing method of r in context c (in this case, with receiver object represented by c). Thus, we have $V = R \times C = R \times O$. Similarly, map $h(o, c) = o^c$ where o^c is the set of context-sensitive object names; intuitively, o^c represents the objects that are allocated at the site of o when the enclosing method is invoked with context c. Thus, we have $H = O \times O = O^2$.

At virtual calls, the analysis resolves the call based on the heap object $h = o^{c''}$ and compile-time target m. The new context c' for run-time target m_j is set to o and points-to edge $(this^o, o^{c''})$ is added to the points-to graph. Also points-to edges (p_1^o, h_2) are added for every h_2 in the points-to set of r_1^c in order to account for flow from actuals to formals. Similarly, edges (l^c, h_3) are added for every h_3 in the points-to set of $ret_{m_j}^o$. Note that this analysis maintains a context of depth one, that is, one allocation site for both reference variables and heap objects. It would be trivial to define C, v and h to maintain contexts of higher depths.

3.3 Fragment Class Analysis

The class analyses are typically designed as *whole-program*

```
void main() {
    ZipEntry ph_ZE;
    ZipInputStream ph_ZIS;
    ZipOutputStream ph_ZOS;
    ph_ZE = new ZipEntry();
    ph_ZIS = new ZipInputStream();
    ph_ZOS = new ZipOutputStream();
    ph_ZE.setCRC(0);
    ph_ZE = ph_ZIS.getNextEntry();
    ph_ZOS.putNextEntry(ph_ZE);
    ph_ZOS.closeEntry();
    ph_ZOS.finish();
}
```

Figure 5: Placeholder `main` method for `zip`.

analyses; they take as input a complete program and produce points-to graphs that reflect relationships in the entire program. However, the problem considered in this paper requires class analysis information for a partial program. The input is a set of classes *Cls* and the analysis needs to construct an approximate object graph that is valid across all possible executions of arbitrary client code built on top of *Cls*. To address this problem we make use of a general technique called *fragment analysis* due to Nasko Rountev [33, 36, 34]. Fragment analysis works on a program fragment rather than on a complete program; in our case the fragment is the set of classes *Cls*.

Initially, the fragment analysis produces an artificial `main` method that serves as a placeholder for client code written on top of *Cls*. Intuitively, the artificial `main` simulates the possible flow of objects between *Cls* and the client code. Subsequently, the fragment analysis attaches `main` to *Cls* and uses some whole-program class analysis engine to compute a points-to graph which summarizes the possible effects of arbitrary client code. The fragment analysis approach can be used with a wide variety of class analyses; for the purposes of this paper we consider fragment analysis used with three of the class analyses described in the previous section: 0-CFA, Andersen-style points-to analysis and object-sensitive points-to analysis.

The placeholder `main` method for the classes from Figure 3 is shown in Figure 5. The method contains variables for types from *Cls* that can be accessed by client code. The statements represent different possible interactions involving *Cls*; their order is irrelevant because the whole-program analysis is flow-insensitive. Method `main` invokes all public methods from the classes in *Cls* designated as accessible.

The details of the fragment analysis will not be discussed here; they can be found in [36]. For the purposes of our analysis we discuss the *object reachability* [34] property of the results computed by the fragment analysis. Consider some client program built on top of *Cls* and an execution of this program (the program must satisfy the constraints discussed in Section 2.3). Let $r \in R$ be a variable declared in *Cls* and at some point during execution r is the start of a chain of object references that leads to some heap object. In the fragment analysis solution, there will be a chain of points-to edges that starts at the representative of r, $v \in V$ and leads to some object name $h \in H$ that represents the run-time object. A similar property holds if r is declared outside of *Cls*. In this case, in the fragment analysis solution, the starting point of the chain is the representative of the variable from `main` that has the same type as r. This property is relevant for the ownership and composition analysis described in Section 5 as the points-to graph is used to approximate all possible object graphs and thus all possible accesses must be taken into account.

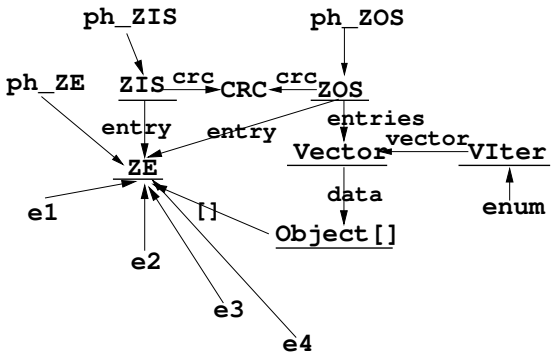
We illustrate the analyses described in Section 3.1. Consider the example from Figures 3 and 5. There are three allocation sites in the `main` method; they are denoted by names `ZE1`, `ZIS1` and `ZOS1`. Name `byte[]` corresponds to the allocation site in class `InflaterInputStream`. There are three allocation sites in class `ZipInputStream`; they are denoted by names `CRC1`, `Inflater1` and `ZE2`. There are four allocation sites in class `ZipOutputStream`; they are denoted by `Vector1`, `Hashtable1`, `Deflater1` and `CRC2`. In addition, we consider the allocation sites in `Vector` (recall Figure 1), which are transitively reachable; they are denoted by `Object[]` and `VIter1`. For brevity, the relevant classes are further denoted in a similar abbreviated fashion; for example we use `ZE` to denote class `ZipEntry`, `ZIS` to denote `ZipInputStream` and `ZOS` to denote `ZipOutputStream`.

The points-to graphs computed from the code in Figures 5, 3 and 1 when the generalized class analysis algorithm is instantiated to 0-CFA, the Andersen-style points-to analysis and to the object-sensitive points-to analysis are shown in Figure 6. Variable `e1` denotes variable `readLOC.e` (i.e., local variable `e` in method `readLOC` in `ZipInputStream`); similarly, `e2` stands for `putNextEntry.e`, `e3` stands for `closeEntry.e` and `e4` stands for `Vector.addElement.e`. Variable `enum` denotes `finish.enum`. Heap object names are underlined in Figure 6. For simplicity, implicit parameters `this` and objects `Inflater`, `byte[]`, `Hashtable` and `Deflater` are not shown.

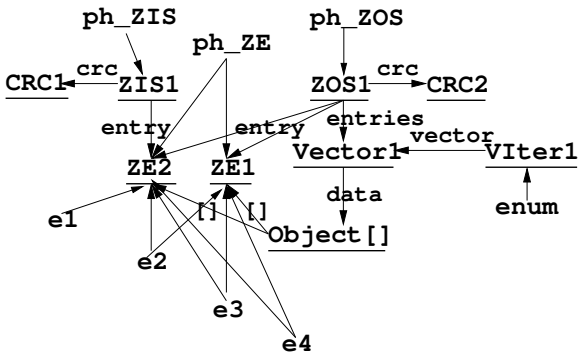
The points-to graph computed when the generalized class analysis algorithm is instantiated to 0-CFA is shown in Figure 6(a). In this case, the objects are represented by their corresponding classes. As a result, there is a single name for the two `CRC` objects even though there are two allocation sites that allocate `CRC` objects; similarly, there is a single name for the two `ZipEntry` objects.

The points-to graph computed when the generalized algorithm is instantiated to the Andersen-style points-to analysis is shown in Figure 6(b). In this case, there are separate object names, `CRC1` and `CRC2` for the two allocation sites that instantiate `CRC` objects. Similarly, there are two object names, `ZE1` and `ZE2` for the two allocation sites that instantiate `ZipEntry` objects.

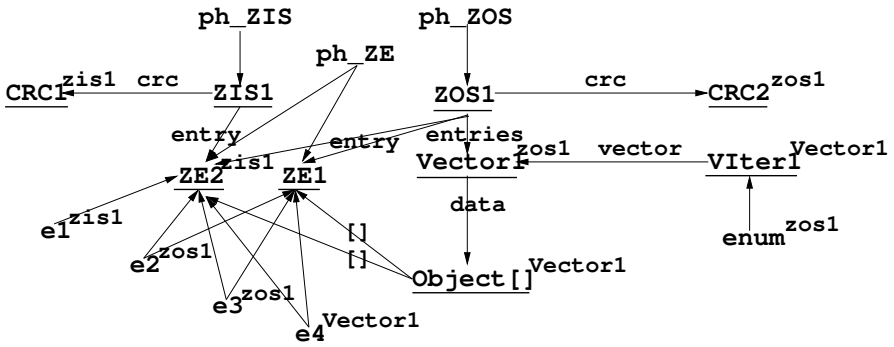
Finally, the points-to graph computed when the generalized algorithm is instantiated to the object-sensitive points-to analysis is shown in Figure 6(c); the superscripts on reference variables and object names denote object contexts. The placeholder `main` method is analyzed in the special static context ϵ and there is a single set variable for each placeholder variable and a single objet name for each object allocated in `main`—that is, object names `ZIS1`, `ZOS1` and `ZE1` as well as reference variables `ph_ZOS`, `ph_ZIS` and `ph_ZE` have no superscripts. For the remaining object names, $\texttt{CRC1}^{\texttt{zis1}}$ and $\texttt{ZE2}^{\texttt{zis1}}$ denote objects that are created in the context of `ZIS1` and thus are annotated with subscript `zis1`. Simi-



(a) Points−to graph for 0−CFA analysis



(b) Points−to graph for Andersen−style points−to analysis



(c) Points−to graph for object−sensitive points−to analysis

Figure 6: Points-to graphs computed by the fragment points-to analysis.

larly, `CRC2`$^{\texttt{zos1}}$ and `Vector1`$^{\texttt{zos1}}$ denote objects created in the context of `ZOS1`. `VIter`$^{\texttt{Vector1}}$ denotes the iterator created by `Vector1` and `Object[]`$^{\texttt{Vector1}}$ denotes the data array created by `Vector1` (in methods `elements` and `Vector` in Figure 1 respectively). For the reference variables, `e1`$^{\texttt{zis1}}$ denotes the context copy of `e1` when the enclosing method `readLOC` is invoked with receiver `ZIS1`. Similarly, `e2`$^{\texttt{zos1}}$, `e3`$^{\texttt{zos1}}$ and `enum`$^{\texttt{zos1}}$ denote the context copies of `e2`, `e3` and `enum` when their enclosing methods are invoked with receiver `ZOS1`. Finally, `e4`$^{\texttt{Vector1}}$ denotes the context copy of `e4` when its enclosing method `addElement` is invoked on receiver `Vector1`.

4. APPROXIMATE OBJECT GRAPH

The output of fragment class analysis is needed to construct the *approximate object graph* Ag which approximates all possible run-time object graphs that can happen when client code is built on top of *Cls*. Subsequently, Ag is used for ownership inference. The nodes in Ag are taken from the set of object names H and the edges represent "may-access" relationships. Figure 7 outlines the construction of Ag given a points-to graph Pt—that is, the algorithm is parameterized by a class analysis. Recall that set C represents the contexts of invocation of a method for the purposes of class analysis; although the generalized class analysis algorithm may be instantiated with different context sensitivity schemes, for clarity in this paper we consider only empty contexts (i.e., $C = \{\epsilon\}$) and receiver object contexts (i.e., $C = O$). The analysis in Figure 9 processes each statement that may contribute object graph edges in each appropriate context. In other words, for analyses that are context-insensitive (RTA, 0-CFA and the Andersen-style points-to analysis) statement s is processed once. For analyses that are context-sensitive (the object-sensitive points-to analysis in our example) statement s in method m is processed for each context in C_m.

Set $\mathcal{C}_{m,c}$ denotes the set of object names that represent the contexts of invocation of method m in context c. Note that $\mathcal{C}_{m,c}$ are the contexts in the sense of the object graph as described in Section 2.1—that is, the *receivers* of m when m is invoked in context c. Clearly, each $\mathcal{C}_{m,c}$ can be approximated using class analysis information. If m is an instance method or constructor, $\mathcal{C}_{m,c}$ is the points-to set of `this`^{c} (i.e., the points-to set of the context copy of implicit parameter `this` for context c). If m is a static method $\mathcal{C}_{m,\epsilon}$ includes the union of the points-to sets of `this` for all instance methods or constructors that may call m (directly or through a sequence of static calls); it includes `root` if m is `main` or may be called from `main`.

Lines 1-2 account for edges due to flow from the contexts of the callee to the contexts of the caller (using the term context in the sense of the object graph in Figure 2.1).

input *Stmt*: set of statements Pt: $V \cup H \rightarrow \mathcal{P}(H)$
output $Ag : H \rightarrow \mathcal{P}(H)$

[1] **foreach** statement s in method m
s: $l = new\ C(...)$ s.t. l not immediately passed or immediately returned to another context,
s: $l = r.m(...)$ s.t. $r \neq$ `this`,
s: $l = r.f$ s.t. $r \neq$ `this` and l assigned to a variable
foreach context $c \in C_m$
[2] add $\{h \rightarrow h_j \mid h \in \mathcal{C}_{m,c} \wedge (v(l,c), h_j) \in Pt\}$ to Ag
// add access edges due to flow from callees to callers
[3] **foreach** statement s in method m
s: $l = new\ C(r)$,
s: $l.m(r)$ s.t. $l \neq$ `this`,
s: $l.f = r$ s.t. $l \neq$ `this`
foreach context $c \in C_m$
[4] add $\{h_i \rightarrow h_j \mid (v(l,c), h_i) \in Pt \wedge (v(r,c), h_j) \in Pt\}$ to Ag
// add access edges due to flow from callers into callees
[5] **foreach** $h_i \xrightarrow{f} h_j \in Pt$ label $h_i \rightarrow h_j \in Ag$ with f
[6] **foreach** s in method m
s: $l = new\ C($`this`$)$ s.t. l not immediately passed or immediately returned to another context,
s: $r.m($`this`$)$,
s: $... =$ `this`
foreach context $c \in C_m$
[7] add $\{h_i \rightarrow h_i \mid (v($`this`$, c), h_i) \in Pt\ \}$ to Ag
// add self-loop edges due to `this` access

Figure 7: Construction of Ag. $\mathcal{P}(X)$ denotes the power set of X. Ag is initially empty.

For example, at a constructor call new edges are added to Ag from each context of the call to the name representing the newly created object. Similarly, at an instance call not through `this` new edges are added from each context of the call to each returned object. Note that when the newly constructed object is immediately passed to another context (e.g., as in $new\ A(new\ B(...))$), or immediately returned to another context (e.g., as in $return\ new\ VIterator(this)$), no new edges are added to that object from the contexts enclosing the constructor call. Also, at indirect read statements, no edges are added when variable l is not assigned or passed as an explicit or implicit argument later (e.g., it is used only to access instance or array fields such as in x=$l[i]$). This is consistent with the definition of the relaxed object graph in Section 2.1. Lines 3-4 account for edges due to flow from the contexts of the caller to the contexts of the callee. For example, at instance calls edges are added to each object in the points-to set of a reference argument, from each object in the points-to set of the receiver. Line 5 labels the edges with the appropriate field identifier and line 6 creates a self-loop that is due to a reference through implicit parameter `this`. For clarity, we omit detailed discussion of static fields. The actual implementation creates edges from `root` to each object in the points-to set of a static field; the case is handled correctly by this algorithm and by the algorithm in Section 5.1.

We discuss the *reachability* property of the approximate object graph. Consider some client program built on top of *Cls* and an execution of this program (the program must satisfy the constraints discussed in Section 2.3). Let c be a context in the sense of the object graph (i.e., `root` or a heap object) and at some point during execution c is the start of a chain in the relaxed object graph that leads to some heap object o^r. In Ag, there will be a chain of edges that starts at the representative of c and leads to the representative of o^r.

Figure 8 shows the approximate object graphs computed from the code on Figures 3, 5 and 1, and the points-to graphs in Figure 6 (only object names from Figure 6 are shown and fields are omitted for clarity). We consider in detail the object graph in Figure 8(b) resulting from the points-to graph computed by Andersen's points-to analysis in Figure 6(b). The other two object graphs are computed analogously and the reachability property holds for all. The analysis is context-insensitive and thus each statement is processed by the algorithm in Figure 7 once in the empty context ϵ; also we have $v(r, \epsilon) = r$ and $H = O$. For the majority of edges inference is straight-forward. For example, edges `root`→`ZIS1`, `root`→`ZIS2` and `root`→`ZE1` are due to the constructor calls in `main` and edges `ZIS1`→`CRC1` and `ZIS1`→`ZE2` are due to the constructor calls in class `ZipInputStream`. Edge `ZOS1`→`VIter1` is due to call *enum=entries.elements*() in method `finish`. Edge `VIter1`→`Vector1` is due to statement *return new VIterator*(*this*) in method `elements`; note that there is no edge `Vector1`→`VIter1` due to this statement. Edge `root`→`ZE2` is due to statement $ph\_ZE = ph\_ZIS.getNextEntry()$ in `main`, and edges `ZOS1`→`ZE2` and `ZOS1`→`ZE1` are due to statement $ph\_ZOS.putNextEntry(ph\_ZE)$ in `main`. Edges `Object[]`→`ZE2` and `Object[]`→`ZE1` are due to flow at statement $data[at] = e$ in `addElement`.

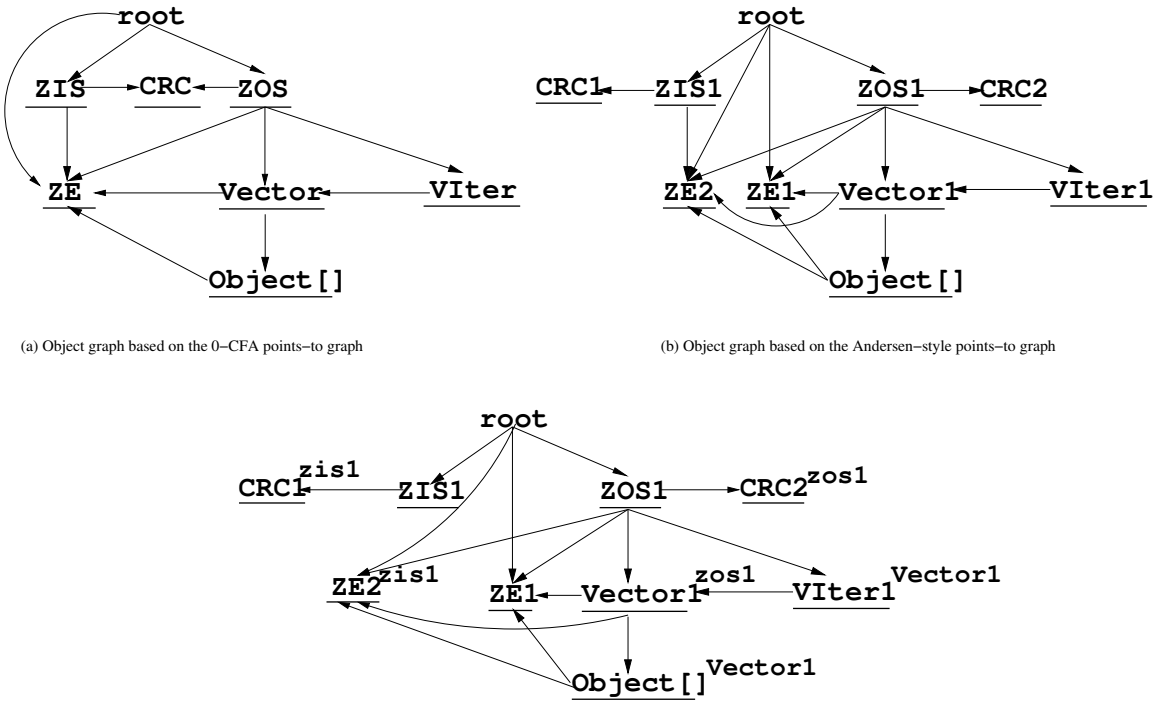
5. IDENTIFYING COMPOSITION RELATIONSHIPS

We propose a novel analysis for ownership inference. The analysis uses Ag to identify a *boundary* subgraph rooted at h for each object name h; the subgraph contains paths that are guaranteed to represent flow within the ownership boundary of h. Whenever the edge appears in the boundary of its source for *all* edges labeled with f, the relationship through f is identified as composition.

5.1 Ownership Boundary

Procedure **computeBoundary** in Figure 9 takes Ag and object name h_i as input and outputs subgraph $Bndry(h_i)$. Subgraph $Bndry(h_i)$ contains paths that are guaranteed to represent flow within the ownership boundary of an instance represented by h_i. From now on we will denote run-time time objects by o^r, o^r_i, o^r_j, etc; their corresponding analysis representatives will be denoted by h, h_i, h_j, etc. More precisely, we have the following *lemma*. Let o^r_i be a heap object represented by h_i. For every edge e: $h \rightarrow h_j \in Bndry(h_i)$ we have that if o^r_i dominates some o^r (represented by h) then o^r_i dominates the o^r_j (represented by h_j) that o^r refers to. Therefore, for every o^r_i and run-time path p: $o^r_i \rightarrow ...o^r \rightarrow o^r_j$, whose representative is in $Bndry(h_i)$, we have that o^r_i dominates o^r_j.

Consider the object graph in Figure 8(b). The boundary of `ZOS1` includes nodes `ZOS1,CRC2,Vector1,Object[]` and `VIter1` and the edges between them. There are paths `ZOS1`→`CRC2`, `ZOS1`→`VIter1`, `ZOS1`→`Vector1`, `ZOS1`→`VIter1`



(a) Object graph based on the 0–CFA points–to graph

(b) Object graph based on the Andersen–style points–to graph

(c) Object graph based on the object–sensitive points–to graph

Figure 8: Approximate object graphs computed by the algorithm in Figure 7.

→`Vector1`, `ZOS1`→`Vector1`→ `Object[]` and `ZOS1`→`VIter1` →`Vector1`→`Object[]`. It is easy to see that for example for every run-time `ZOS1`^{r}→`Vector1`^{r}, `ZOS1`^{r} dominates `Vector1`^{r}.

Below we briefly outline the algorithm and the correctness argument. The algorithm uses the fact that o_j^r flows from object o_i^r to some object o_k^r only if one of the following is true: (1) o_k^r has a handle to both o_i^r and o_j^r (and due to the reachability property Ag contains edges $h_k \rightarrow h_i$, $h_k \rightarrow h_j$, $h_i \rightarrow h_j$), or (2) o_i^r has a handle to both o_k^r and o_j^r (and Ag contains edges $h_i \rightarrow h_k$, $h_i \rightarrow h_j$, $h_k \rightarrow h_j$). This observation helps identify encapsulation more precisely. Suppose that our running example has another input stream object, created by `root` and denoted by name `ZIS2`. The relationship between `ZIS2` and its crc object would be represented by edge `ZIS2`→`CRC1` in Figure 8(b). A naive algorithm may identify `root` as the dominator of the crc objects, and fail to identify the composition relationship between `ZipInputStream` and `CRC32`. In fact, the `CRC1` object is created and dominated by its enclosing `ZIS1` object because there is no h_k such that either h_k has handles to both `ZIS1` and `CRC1`, or `ZIS1` has handles to both h_k and `CRC1`; thus, the `CRC1` object created by the `ZIS1` object does not flow to or from any other context.

The algorithm builds the boundary of an object name h_i by adding edges. First, **computeBoundary** partitions the edges reachable from h_i into appropriate closure sets using auxiliary procedure **findClosureSet**. Intuitively, the closure set of edge $h \rightarrow h_j$ contains all edges $h_k \rightarrow h_j$ in the transitive closure of h_i, such that some o_k^r and o^r refer to the same o_j^r. For example, the closure set of `ZOS1`→`Vector1` is {`ZOS1`→`Vector1`,`VIter1`→ `Vector1`}, and the closure set of `ZOS1`→`ZE1` is {`ZOS1`→`ZE1`,`Vector1`→`ZE1`,`Object[]`→`ZE1`}. The role of the parent set *Prt* (discussed later) is to ensure that the relevant paths to $h \rightarrow h_j$ stay in boundary. $Bndry(h_i)$ grows from zero to one edge, $h_i \rightarrow h_j$, when (i) there is no h_k that has handles to both h_i and h_j and (ii) there is no h_k such that h_i has handles to both h_k and h_j, and h_k has a handle to h_j. The first condition is guaranteed by the check that the *Closure* set of $h_i \rightarrow h_j$ is not forbidden, and the second condition is guaranteed by the check that the *Prt* set of $h_i \rightarrow h_j$ is empty; both checks are performed at line 5. Thus, an edge $h_i \rightarrow h_j$ is added to the empty boundary of h_i only when it is guaranteed that the h_i object accesses the h_j object exclusively (i.e., no other object has a handle to it). Examples of such edges are `ZIS1`→`CRC1` and `ZOS1`→`CRC2`. Clearly, the lemma holds in this case.

Consider an edge $h \rightarrow h_j$ that is added to $Bndry(h_i)$ at line 6. Consider some client program built on top of *Cls* and an execution of this program (the program must satisfy the constraints discussed in Section 2.3). Let o_i^r be any run-time object represented by h_i and o^r be an object dominated by o_i^r. We need to examine all h_k such that some o_j^r referred by o^r may flow to or from o_k^r (i.e., there is an edge $o_k^r \rightarrow o_j^r$ in the relaxed object graph). If all these o_k^r are dominated by o_i^r then o_j^r is dominated by o_i^r.

Object o_j^r flows from o^r into some o_k^r when one of the

procedure **findClosureSet** // of $h \rightarrow h_j$ w.r.t. h_i
input Ag: $H \rightarrow \mathcal{P}(H)$ $h \rightarrow h_j$: $H \times H$ h_i: H n: Int
output $Closure(h_i, n)$: $\mathcal{P}(H \times H)$ $Prt(h_i, n)$: $\mathcal{P}(H \times H)$
initialize Wl={}, $Closure(h_i, n)$={}, $Prt(h_i, n)$={}
[1] mark $h \rightarrow h_j$, add it to Wl and to $Closure(h_i, n)$
[2] **while** Wl not empty
[3] remove $h \rightarrow h_j$ from Wl
[4] **foreach** $h_k \rightarrow h_j$ s.t. $h_k \rightarrow h$ and h_k reachable from h_i
[5] **if** $h_k \rightarrow h_j$ is unmarked
[6] mark $h_k \rightarrow h_j$, add it to Wl and $Closure(h_i, n)$
[7] add $h_k \rightarrow h$ to $Prt(h_i, n)$
[8] **foreach** $h_k \rightarrow h_j$ s.t. $h \rightarrow h_k$
[9] **if** $h_k \rightarrow h_j$ is unmarked
[10] mark $h_k \rightarrow h_j$, add it to Wl and $Closure(h_i, n)$
[11] add $h \rightarrow h_k$ to $Prt(h_i, n)$

procedure **computeBoundary** // of h_i
input Ag: $H \rightarrow \mathcal{P}(H)$ h_i: H
output $Bndry(h_i)$: $\mathcal{P}(H \times H)$
initialize n=0
[1] **foreach** unmarked edge $h \rightarrow h_j$ reachable from h_i
[2] findClosureSet($h \rightarrow h_j$,o_i,n++)
[3] **foreach** $h_i \rightarrow h_j$ s.t. $\exists h_k$ s.t. $h_k \rightarrow h_i$ and $h_k \rightarrow h_j$
[4] mark the $Closure$ set of $h_i \rightarrow h_j$ as forbidden
[5] **while** empty $Prt(h_i, k)$ and $Closure(h_i, k)$ not forbidden
[6] add $Closure(h_i, k)$ to $Bndry(h_i)$
[7] **foreach** $e \in Closure(h_i, k)$ remove e from each Prt
[8] remove $Prt(o_i, k)$ and $Closure(h_i, k)$

Figure 9: Ownership analysis.

following conditions is true. First, o_k^r has handles to both o^r and o_j^r (e.g., o_j^r may be returned to o_k^r from a method invoked on o^r, or it may be passed as an argument from o_k^r to a method invoked on o^r). Since o_i^r dominates o^r we have that o_i^r dominates o_k^r. This case is examined at lines 4-7 in **findClosureSet** and $h_k \rightarrow h_j$ is added to the worklist; it is examined in a subsequent iteration of the while loop in **findClosureSet** in order to find the representatives of the objects that o_j^r may flow to from o_k^r. In addition, $h_k \rightarrow h_j$ is added to $Closure(h_i, n)$, the closure set of $h \rightarrow h_j$. Second, o_j^r may flow from o^r into some o_k^r such that o^r has handles to both o_k^r and o_j^r. Clearly, in this case we have that $h \rightarrow h_k \in Bndry(h_i)$ because $h \rightarrow h_k$ is in the Prt set of $h \rightarrow h_j$; recall that an edge is removed from a Prt set only when it is added to the boundary at lines 6-7 in **computeBoundary**. We may assume that the lemma holds for $h \rightarrow h_k \in Bndry(h_i)$—that is, if o_i^r dominates o^r then o_i^r dominates the o_k^r referred to by o^r. Thus, we have that o_i^r dominates o_k^r. This case is examined at lines 8-11 in **findClosureSet** and appropriate $h_k \rightarrow h_j$ are added to the worklist and to the closure set.

We briefly illustrate the algorithm on our running example. Consider the boundary of `ZIS1`. There is a single closure set that is not forbidden, $Closure$(`ZIS1`, 0)={`ZIS1`→`CRC1`} with corresponding parent set Prt(`ZIS1`, 0)={} and edge `ZIS1`→`CRC1` is added to $Bndry$(`ZIS1`) at line 6. Consider the boundary of `ZOS1`. As a result of **findClosureSet** in lines 1-2 there are four closure sets that are not forbidden: $Closure$(`ZOS1`, 0)={`ZOS1`→`CRC2`}, $Closure$(`ZOS1`, 1) ={`ZOS1`→`VIter1`}, $Closure$(`ZOS1`, 2)={`ZOS1`→`Vector1`,`VIter1`→`Vector1`} and $Closure$(`ZOS1`, 3)={`Vector1`→`Object[]`}. Their corresponding parent sets are Prt(`ZOS1`, 0)={}, Prt(`ZOS1`, 1) ={}, Prt(`ZOS1`, 2)={`ZOS1`→`VIter1`}, and Prt(`ZOS1`, 3)={}. The algorithm processes the first closure set and adds edge `ZOS1`→`CRC2` to $Bndry$(`ZOS1`). Then it adds the second closure set—that is, edge `ZOS1`→`VIter1` to the boundary and deletes the edge from the third parent set. The third parent set becomes empty and `ZOS1`→`Vector1` and `VIter1`→`Vector1` are added to the boundary. Finally, edge `Vector1`→`Object[]` is added to the boundary. Thus we have the following boundary graphs: $Bndry$(`ZIS1`) = {`ZIS1`→`CRC1`}, $Bndry$(`Vector1`) = {`Vector1`→`Object[]`} and $Bndry$(`ZOS1`) = {`ZOS1`→`CRC2`, `ZOS1`→`Vector1`, `ZOS1`→`VIter1`, `Vector1`→`Object[]`,`VIter1` →`Vector1`}.

A corollary of the lemma is that whenever we have an edge $h_i \rightarrow h_j \in Bndry(h_i)$ each o_i^r owns the o_j^r instances that it may refer to. If for every edge labeled with f we have $h \xrightarrow{f} h' \in Bndry(h)$ the analysis identifies one-to-one implementation-level composition or collection ownership. As pointed earlier, the major focus of this paper is the inference of ownership and compositions. Our methodology handles inference of one-to-many relationships as well but in order to keep the focus of the paper it is not discussed here; it is addressed in [26].

5.2 Analysis Complexity

We discuss the complexity of the analysis in terms of sets H, V and C and we emphasize the framework instance that is based on the Andersen-style points-to analysis which we believe is the most suitable for our purposes (i.e., it provides the best trade-off between analysis cost and analysis precision for the purposes of composition inference).

Let N be the size of the program being analyzed (i.e., Cls and the placeholder main)—that is, the number of statements, the number of object allocation sites and the number of reference variables is of order N. To reason about the complexity of the generalized class analysis in Figure 4 we consider a standard set-constraint-based solution procedure [12, 42, 44, 35]. In set-constraint-based analyses the solution is divided into *constraint generation* and *constraint resolution* where complexity is clearly dominated by constraint resolution. For example, when the algorithm in Figure 4 is instantiated into Andersen-style points-to analysis constraint generation processes each statement once and generates constraints of the form $v_r \subseteq v_l$ (i.e., this constraint denotes that the points-to set of r flows to the points-to set of l). Solving for Andersen's analysis requires propagating object names o_i to each v_r which clearly dominates the linear generation. To reason about constraint resolution in terms of sets H and V consider that H object names need to be propagated towards V reference variables through constraints of the form $h \subseteq v_i$ and $v_i \subseteq v_j$. Therefore, the complexity of propagation for the generalized analysis is $O(H * V^2)$—clearly, each constraint $o_i \subseteq v_j$ may be discovered though $O(V)$ intermediate variables. As a result for 0-CFA we have $O(Cl * R^2)$=$O(N^3)$, for the Andersen-style points-to analysis we have $O(O * R^2)$=$O(N^3)$ and for the object-sensitive points-to analysis we have $O(O * O * (R * O)^2)$=$O(N^6)$ (note that although the number of classes Cl is of order N, the number of classes Cl is substantially smaller

than the number of allocation sites O in practice).

The complexity of the construction of the approximate object graph in Figure 7 is $O(N*C*H^2)$: there are $O(N)$ statements, each statement is processed in at most $O(C)$ contexts and for each statement the algorithm performs at most $O(H^2)$ work (due to lines 2 and 4). Thus, for 0-CFA we have $O(N*Cl^2)$=$O(N^3)$, for the Andersen-style points-to analysis we have $O(N*O^2)$=$O(N^3)$ and for the object-sensitive points-to analysis we have $O(N*O*(O*O)^2)$=$O(N^6)$.

Finally, consider procedure **computeBoundary** in Figure 9. The code for partitioning the edges in the transitive closure of h_i into closure sets (lines 1-2) examines each edge and for each edge performs at most $O(H)$ work: for edge $h \rightarrow h_j$ there may be at most $O(H)$ nodes h_k such that $h_k \rightarrow h$ and $h_k \rightarrow h_j$ (examined at lines 4-7 in **findClosureSet**); similarly, there may be at most $O(H)$ nodes h_k such that $h \rightarrow h_k$ and $h_k \rightarrow h_j$ (examined at lines 8-11 in **findClosureSet**). Therefore, the complexity of lines 1-2 is $O(H^3)$. The while loop that adds edges to the boundary (lines 5-8) examines each edge at most once, and each edge is removed from at most $O(H)$ parent sets. Therefore, the complexity of lines 5-8 is $O(H^3)$ as well. To conclude, the complexity of our analysis is dominated by the computation of the boundary sets which is worst-case $O(H^4)$. For the Andersen-style points-to analysis, the overall complexity is $O(N^4)$.

6. EXPERIMENTAL STUDY

We implemented one instance of our framework—in particular, we considered class analysis based on the Andersen-style points-to analysis, and the object graph construction and ownership inference based on it. We believe that this instance of the framework is the most suitable for the purposes of inference of compositions in program fragments. The goal of the empirical study is to address two questions. First, how often does our analysis discover implementation-level composition? Second, how *imprecise* the analysis is—that is, how often it misses implementation-level composition?

We performed experiments on the 7 Java components listed in Table 1. The analysis implementation is based on the Soot framework [46]. The components are from the standard library packages `java.text` and `java.util.zip`, also used in [34]. The components are described briefly in the first two columns of Table 1. Each component contains the set of classes in *Cls* (i.e., the classes that provide component functionality plus all other classes that are directly or transitively referenced by the functionality classes). The number of classes in *Cls* and the number of classes that implement the component functionality is shown in column (3). We considered all reference instance fields in the classes that implement the component functionality; this number is given in column (4).

6.1 Results

We applied Andersen-style points-to analysis, object graph construction based on the Andersen-style points-to analysis and composition inference as described earlier in order to determine which fields accounted for composition relationships. Column (5) in Table 1 shows how many of the fields from column (4) are identified as one-to-one compositions and column (3) shows how many are identified as owned collections (i.e., arrays and standard `java.util` collections). On average, the analysis reported 30% one-to-one compositions and 10% owned collections—that is, 40% of the reference instance fields account for representation that is not being exposed outside of its enclosing object.

6.2 Analysis Precision

The issue of analysis precision is of crucial importance for software tools. If an analysis is imprecise, it may report that the relationship between two classes is not a composition while in reality it is, or that a collection is not owned while in reality it is owned (i.e., the analysis reports that certain representation may be exposed while in fact it is not). Such information is not useful and may confuse the user and even render the tool unusable. For example, if a user attempts to ensure the consistency between the code and the composition relationships in UML design class diagrams, imprecision will mean that a large chunk of code will have to be examined manually. Since imprecision results in waste of human time, analysis designers must carefully and precisely identify and evaluate any sources of imprecision.

In our experiments, we examined the fields that were not identified as compositions or owned collections. We attempted to prove that it was possible to write client code s.t. an object stored in such a field would be exposed (i.e., it would not be owned by its enclosing object in accordance with the ownership model in Section 2.1). In all cases, except one, we were able to prove exposure. Thus, the analysis achieves almost perfect precision.

Field `defaultCenturyStart` in component `date` accounted for the one case of imprecision. The object stored in the field comes from a call to a method `getTime` which creates and immediately returns a `Date` object. Although the `Date` object stored in `defaultCenturyStart` does not flow out of its enclosing `SimpleDateFormat` object, other `Date` objects created by `getTime` in `SimpleDateFormat` are being returned (i.e., there are edges in *Ag* to the `SimpleDateFormat` object and the only representative of `Date`). This imprecision may be resolved by using an instance of our framework that employs more precise object naming. In the case of `getTime` it may distinguish the `Date` objects for different call sites of `getTime`—for example, 1-CFA analysis which distinguishes context by the last enclosing call site would produce precise results; in this case, the target of the `defaultCenturyStart` edge would be a separate `Date` object that does not flow out and the ownership inference algorithm will correctly identify that there is a composition relationship through this field. However, it remains to be seen whether a more precise context-sensitive points-to analysis will result in substantial benefits for the ownership and composition analyses.

6.3 Conclusions

Our results indicate that the ownership model captures conceptual composition relationships appropriately—we encountered several cases when values of private fields were stored in other parts of the object representation. Thus, a model based on exclusive ownership (i.e., a model which requires that an owned object is referenced only by its owner) would not have been sufficient. The results also show that composition relationships occur often. Therefore, the analysis can provide useful information for reverse engineering

(1)Component	(2)Functionality	(3)#Classes *Cls*/Functionality	(4)#Fields	Compositions			
				(5)#One-to-one		(6)#Owned collections	
				Analysis	Perfect	Analysis	Perfect
`gzip`	GZIP IO streams	199/6	7	4(57%)	4(57%)	0(0%)	0(0%)
`zip`	ZIP IO streams	194/6	10	3(30%)	3(30%)	2(20%)	2(20%)
`checked`	IO streams with checksums	189/4	2	0(0%)	0(0%)	0(0%)	0(0%)
`collator`	text collation	203/15	24	10(42%)	10(42%)	6(25%)	6(25%)
`date`	date formatting	205/17	20	3(15%)	**4(20%)**	5(25%)	5(25%)
`number`	number formatting	198/10	3	2(67%)	2(67%)	0(0%)	0(0%)
`boundary`	iter. over boundaries in text	199/13	7	0(0%)	0(0%)	0(0%)	0(0%)
`Average`				30%	31%	10%	10%

Table 1: Java components and implementation-level compositions.

tools. It is important that precise results can be obtained with practical analysis—the combined running time of the points-to, object graph construction and composition inference analyses does not exceed 10 seconds on any component (executed on a 900MHz Sun Fire 380R).

Clearly, a threat to the validity of our results is the relatively small code base used in our experiments. Although the components used in this study are representative program fragments, the results need to be confirmed on more components. In the future we plan to investigate the impact of other framework instances, especially context-sensitive ones, on more components.

7. RELATED WORK

Work by Kollmann and Gogolla [22] and more recently by Guéhéneuc and Albin-Amiot [19] presents definitions and identification algorithms for implementation-level association, composition and aggregation relationships. Our work focuses on compositions and differs from [22] and [19] in both the definition of implementation-level composition and in the identification algorithm. The definition of composition in [22] and [19] is based on exclusive ownership. This may not be sufficient to model commonly used patterns such as iterators, decorators, and factories [14], as well as the common situation when instance fields refer to owned objects that are temporarily accessed by other parts of the representation of the owner. Our definition is based on the owners-as-dominators model which does not require exclusive relationship with the owner; as observed by us and other researchers [9, 32], this model captures well the notion of composition in modeling [38].

We present an identification algorithm that may be more appropriate. Guéhéneuc and Albin-Amiot propose the use of dynamic analysis, but point out serious disadvantages. First, dynamic analysis is slow, second, it requires a complete program, and third, the results that are obtained may be incomplete because they are based on particular runs of particular clients of the component. Kollmann and Gogolla use dynamic analysis as well. Our detection algorithm is based on practical static analysis that works on incomplete programs and produces a solution that is valid over all unknown clients of the component.

Work in [21] and [45] addresses the issue of recovering one-to-many associations through containers, since reverse engineering tools typically loose the association between the enclosing class and the class whose instances are stored in the container field (recall the `entries` field of `Vector` type in Figure 3). Identification of composition is not addressed in these papers.

Ownership type systems disallow certain accesses of object representation [29, 9, 8, 2, 6]. These systems require type annotations and typically do not include automatic inference algorithms or empirical investigations. In contrast, we infer ownership automatically and present an empirical study of the effectiveness of our approach; we believe that our analysis can be usefully incorporated in software tools for reverse engineering of class diagrams from Java code. The only type annotation inference analysis that we are aware of is given by Aldrich et al. [2] for the purposes of alias understanding. Similarly to [19], the `owned` annotation is used only when the analysis is able to prove *exclusive* ownership; in the majority of cases it infers alias parameters. Our work focuses on a different problem, composition inference, and infers ownership using a model that captures better the notion of composition in modeling. Grothoff et al. [16] and Clarke et al. [10] present tools for checking of confinement within a package and within a class respectively. They define confinement rules and the tools check if code conforms to these rules. Our work focuses on a different problem, composition inference, and takes a different approach, the use of semantic analysis that is based on points-to analysis. We believe that such analysis may be more appropriate than confinement rules for the purposes of the identification of object ownership and composition; for example, the rules in [16] and [10] do not handle pseudo-generic containers well.

Bruel et al. [7] and Barbier et al. [4] formalize UML interclass relationships by defining sets of characteristics for association, aggregation and composition; they do not address implementation-level relationships and the problem of reverse engineering. In contrast, we consider implementation-level relationships and propose a methodology for their reverse engineering with an empirical investigation.

8. CONCLUSIONS AND FUTURE WORK

We present an analysis that identifies composition relationships in Java components. We define an ownership-based implementation-level composition model and a static analysis framework for inference composition relationships in incomplete programs. Our experimental study indicates that (i) the ownership-based model captures well the notion of composition in modeling and (ii) implementation-level compositions occur often and almost *all* such compositions

can be identified using a relatively simple and inexpensive analysis. Clearly, no definitive conclusions can be drawn from these limited experiments. In the future, we plan to focus on further empirical investigation placing special emphasis on framework instances based on context-sensitive class analyses.

9. ACKNOWLEDGEMENTS

The author is very grateful to Nasko Rountev for providing the component data, and to the ASE'05 reviewers whose valuable comments improved this paper enormously.

10. REFERENCES

[1] O. Agesen. The cartesian product algorithm: Simple and precise type inference of parametric polymorphism. In *European Conference on Object-Oriented Programming*, pages 2–26, 1995.

[2] J. Aldrich, V. Kostadinov, and C. Chambers. Alias annotations for program understanding. In *Conference on Object-Oriented Programming Systems, Languages, and Applications*, pages 311–330, 2002.

[3] D. Bacon and P. Sweeney. Fast static analysis of C++ virtual function calls. In *Conference on Object-Oriented Programming Systems, Languages, and Applications*, pages 324–341, 1996.

[4] F. Barbier, B. Henderson-Sellers, A. L. Parc-Lacayrelle, and J.-M. Bruel. Formalization of the whole-part relationship in the unified modeling language. *IEEE Transaction Software Engineering*, 29(5):459–470, 2003.

[5] M. Berndl, O. Lhotak, F. Qian, L. Hendren, and N. Umanee. Points-to analysis using BDD's. In *Conference on Programming Language Design and Implementation*, pages 103–114, 2003.

[6] C. Boyapati, B. Liskov, and L. Shrira. Ownership types for object encapsulation. In *Symposium on Principles of Programming Languages*, pages 213–223, 2003.

[7] J.-M. Bruel, B. Henderson-Sellers, F. Barbier, A. L. Parc, and R. B. France. Improving the UML metamodel to rigorously specify aggregation and composition. In *International Conference on Object-Oriented Information Systems*, pages 5–14, 2001.

[8] D. Clarke and S. Drossopoulou. Ownership, encapsulation and the disjointness of type and effect. In *Conference on Object-Oriented Programming Systems, Languages, and Applications*, pages 292–310, 2002.

[9] D. Clarke, J. Potter, and J. Noble. Ownership types for flexible alias protection. In *Conference on Object-Oriented Programming Systems, Languages, and Applications*, pages 48–64, 1998.

[10] D. Clarke, M. Richmond, and J. Noble. Saving the world from bad beans: Deployment time confinment checing. In *Conference on Object-Oriented Programming Systems, Languages, and Applications*, pages 374–387, 2003.

[11] J. Dean, D. Grove, and C. Chambers. Optimizations of object-oriented programs using static class hierarchy analysis. In *European Conference on Object-Oriented Programming*, pages 77–101, 1995.

[12] M. Fähndrich, J. Foster, Z. Su, and A. Aiken. Partial online cycle elimination in inclusion constraint graphs. In *Conference on Programming Language Design and Implementation*, pages 85–96, 1998.

[13] M. Fowler. *UML Distilled Third Edition*. Addison-Wesley, 2004.

[14] E. Gamma, R. Helm, R. Johnson, and J. Vlissides. *Design Patterns: Elements of Reusable Object-Oriented Software*. Addison-Wesley, 1995.

[15] J. Gosling, B. Joy, G. Steele, and G. Bracha. *The Java Language Specification*. Addison-Wesley, 2nd edition, 2000.

[16] C. Grothoff, J. Palsberg, and J. Vitek. Encapsulating objects with confined types. In *Conference on Object-Oriented Programming Systems, Languages, and Applications*, pages 241–253, 2001.

[17] D. Grove and C. Chambers. Call graph construction in object-oriented languages. *ACM Transactions on Programming Languages and Systems*, 23(6):685–746, Nov. 2001.

[18] D. Grove, G. DeFouw, J. Dean, and C. Chambers. Call graph construction in object-oriented languages. In *Conference on Object-Oriented Programming Systems, Languages, and Applications*, pages 108–124, 1997.

[19] Y.-G. Guéhéneuc and H. Albin-Amiot. Recovering binary class relationships: Putting icing on the UML cake. In *Conference on Object-Oriented Programming Systems, Languages, and Applications*, pages 301–314, 2004.

[20] M. Hind. Pointer analysis: Haven't we solved this problem yet? In *Workshop on Program Analysis for Software Tools and Engineering*, pages 54–61, 2001.

[21] D. Jackson and A. Waingold. Lightweight extraction of object models from bytecode. *IEEE Transaction Software Engineering*, 27(2):156–169, 2001.

[22] R. Kollmann and M. Gogolla. Application of UML associations and their adornments in design recovery. In *Working Conference on Reverse Engineering*, pages 81–91, 2001.

[23] C. Larman. *Applying UML and Patterns*. Prentice Hall, 2nd edition, 2002.

[24] O. Lhotak and L. Hendren. Scaling Java points-to analysis using Spark. In *International Conference on Compiler Construction*, pages 153–169, 2003.

[25] D. Liang, M. Pennings, and M. J. Harrold. Extending and evaluating flow-insensitive and context-insensitive points-to analyses for Java. In *Workshop on Program Analysis for Software Tools and Engineering*, pages 73–79, 2001.

[26] A. Milanova. Precise identification of composition relationships for UML class diagrams. Technical Report RPI/DCS-05-10, Rensselaer Polytechnic Institute, 2005.

[27] A. Milanova, A. Rountev, and B. Ryder. Parameterized object-sensitivity for points-to and

side-effect analyses for Java. In *International Symposium on Software Testing and Analysis*, pages 1–12, 2002.

[28] A. Milanova, A. Rountev, and B. G. Ryder. Parameterized object sensitivity for points-to analysis for Java. *ACM Transactions on Software Engineering and Methodology*, 14(1):1–42, 2005.

[29] J. Noble, J. Vitek, and J. Potter. Flexible alias protection. In *European Conference on Object-Oriented Programming*, pages 158–185, 1998.

[30] J. Palsberg and M. Schwartzbach. Object-oriented type inference. In *Conference on Object-Oriented Programming Systems, Languages, and Applications*, pages 146–161, 1991.

[31] J. Plevyak and A. Chien. Precise concrete type inference for object-oriented languages. In *Conference on Object-Oriented Programming Systems, Languages, and Applications*, pages 324–340, 1994.

[32] J. Potter, J. Noble, and D. Clarke. The ins and outs of objects. In *Australian Software Engineering Conference*, pages 80–89, 1998.

[33] A. Rountev. *Dataflow Analysis of Software Fragments*. PhD thesis, Rutgers University, 2002.

[34] A. Rountev. Precise identification of side-effect free methods. In *International Conference on Software Maintenance*, pages 82–91, 2004.

[35] A. Rountev, A. Milanova, and B. G. Ryder. Points-to analysis for Java using annotated constraints. In *Conference on Object-Oriented Programming Systems, Languages, and Applications*, pages 43–55, 2001.

[36] A. Rountev, A. Milanova, and B. G. Ryder. Fragment class analysis for testing of polymorphism in Java softwhare. *IEEE Transaction Software Engineering*, 30(6):372–386, June 2004.

[37] E. Ruf. Effective synchronization removal for Java. In *Conference on Programming Language Design and Implementation*, pages 208–218, 2000.

[38] J. Rumbaugh, I. Jacobson, and G. Booch. *The Unified Modeling Language Reference Manual*. Addison-Wesley, 2nd edition, 2004.

[39] B. G. Ryder. Dimensions of precision in reference analysis of object-oriented languages. In *International Conference on Compiler Construction*, pages 126–137, 2003.

[40] M. Sharir and A. Pnueli. Two approaches to interprocedural data flow analysis. In S. Muchnick and N. Jones, editors, *Program Flow Analysis: Theory and Applications*, pages 189–234. Prentice Hall, 1981.

[41] M. Streckenbach and G. Snelting. Points-to for Java: A general framework and an emprirical comparison. Technical report, U. Passau, Sept. 2000.

[42] Z. Su, M. Fähndrich, and A. Aiken. Projection merging: Reducing redundancies in inclusion constraint graphs. In *Symposium on Principles of Programming Languages*, pages 81–95, 2000.

[43] V. Sundaresan, L. Hendren, C. Razafimahefa, R. Vallee-Rai, P. Lam, E. Gagnon, and C. Godin. Practical virtual method call resolution for Java. In *Conference on Object-Oriented Programming Systems, Languages, and Applications*, pages 264–280, 2000.

[44] F. Tip and J. Palsberg. Scalable propagation-based call graph construction algorithms. In *Conference on Object-Oriented Programming Systems, Languages, and Applications*, pages 281–293, 2000.

[45] P. Tonella and A. Potrich. Reverse engineering of the UML class diagram from C++ code in presence of weakly typed containers. In *International Conference on Software Maintenance*, pages 376–385, 2001.

[46] R. Vallée-Rai, E. Gagnon, L. Hendren, P. Lam, P. Pominville, and V. Sundaresan. Optimizing Java bytecode using the Soot framework: Is it feasible? In *International Conference on Compiler Construction*, LNCS 1781, pages 18–34, 2000.

[47] J. Whaley and M. Lam. An efficient inclusion-based points-to analysis for strictly-typed languages. In *Static Analysis Symposium*, pages 180–195, 2002.

[48] J. Whaley and M. Lam. Cloning-based context-sensitive pointer alias analysis using binary decision diagrams. In *Conference on Programming Language Design and Implementation*, pages 131–144, 2004.