

# Flow- and Context-Sensitive Points-to Analysis using Generalized Points-to Graphs

Pritam M. Gharat and Uday P. Khedker, Indian Institute of Technology Bombay  
 Alan Mycroft, University of Cambridge

Computing precise (fully flow-sensitive and context-sensitive) and exhaustive (as against demand driven) points-to information is known to be computationally expensive. Therefore many practical tools approximate the points-to information trading precision for efficiency. This often has adverse impact on computationally intensive analyses such as model checking. Past explorations in top-down approaches of fully flow- and context-sensitive points-to analysis (FCPA) have not scaled. We explore the alternative of bottom-up interprocedural approach which constructs summary flow functions for procedures to represent the effect of their calls. This approach has been effectively used for many analyses. However, this approach seems computationally expensive for FCPA which requires modelling unknown locations accessed indirectly through pointers. Such accesses are commonly handled by using placeholders to explicate unknown locations or by using multiple call-specific summary flow functions.

We generalize the concept of points-to relations by using the counts of indirection levels leaving the unknown locations implicit. This allows us to create summary flow functions in the form of *generalized points-to graphs* (GPGs) without the need of placeholders. By design, GPGs represent both memory (in terms of classical points-to facts) and memory transformers (in terms of generalized points-to facts). We perform FCPA by progressively reducing generalized points-to facts to classical points-to facts. GPGs distinguish between *may* and *must* pointer updates thereby facilitating strong updates within calling contexts.

The size of GPG for a procedure is linearly bounded by the number of variables and is independent of the number of statements in the procedure. Empirical measurements on SPEC benchmarks show that GPGs are indeed compact in spite of large procedure sizes. This allows us to scale FCPA to 158 kLoC using GPGs (compared to 35 kLoC reported by liveness-based FCPA). At a practical level, GPGs hold a promise of efficiency and scalability for FCPA without compromising precision. At a more general level, GPGs provide a convenient abstraction of memory in presence of pointers. Static analyses that are influenced by pointers may be able to use GPGs by combining them with their original abstractions.

**CCS Concepts:** •**Theory of computation** → **Program analysis**; •**Software and its engineering** → **Imperative languages; Compilers; Software verification and validation**;

## 1. INTRODUCTION

Points-to analysis discovers information about indirect accesses in a program and its precision influences the precision and scalability of other program analyses significantly. Computationally intensive analyses such as model checking are ineffective on programs containing pointers partly because of imprecision of points-to analyses [1; 2; 3; 8; 13; 14].

We focus on exhaustive as against demand-driven [6; 11; 30] points-to analysis. A demand-driven points-to analysis computes points-to information that is relevant to a query raised by a client analysis; for a different query, the analysis needs to be repeated. An exhaustive analysis, on the other hand, computes all points-to information which can be queried later by a client analysis; multiple queries do not require points-to analysis to be repeated. For precision of points-to information, we are interested in full flow- and context-sensitive points-to analysis. A flow-sensitive analysis respects the control flow and computes separate data flow information at each program point. It provides more precise results but could be inefficient at the interprocedural level. A context-sensitive analysis distinguishes between different calling contexts of procedures and restricts the analysis to interprocedurally valid control flow paths (i.e. control flow paths from program entry to program exit in which every return from a procedure is matched with a call to the procedure such that all call-return matchings are properly nested). A fully context-sensitive analysis does not approximate calling contexts by limiting the call chain lengths even in presence of recursion. Both flow- and context-sensitivity bring in precision and we aim to achieve it without compromising on efficiency.

```

int a, b, c, d;

01 g()
02 {
03     c = a*b;
04     f(); /* call 1 */
05     a = c*d;
06     f(); /* call 2 */
07 }

08 f()
09 {
10     a = b*c;
11 }

```

(a.1) Context independent representation of context-sensitive summary flow function of procedure  $f$

$$f(X) = X \cdot 011 + 010$$

(a.2) Context dependent representation of context-sensitive summary flow function of procedure  $f$

$$f = \{100 \mapsto 010, 011 \mapsto 011\}$$

(b) Context-insensitive data flow information as a procedure summary of procedure  $f$

$$f = 010$$

Fig. 1. Illustrating different kinds of procedure summaries for available expressions analysis. The set  $\{a*b, b*c, c*d\}$  is represented by the bit vector 111.

The top-down approach to context-sensitive analysis propagates the information from callers to callees [36] effectively traversing the call graph top down. In the process, it analyzes a procedure each time a new data flow value reaches a procedure from some call. Several popular approaches fall in this category: call strings method [29], its value-based variants [16; 24] and the tabulation based functional method [26; 29]. By contrast, the bottom-up approaches [7; 10; 19; 27; 29; 31; 32; 33; 34; 35; 36] avoid analyzing a procedure multiple times by constructing its *summary flow function* which is used to incorporate the effect of calls to the procedure. Effectively, this approach traverses the call graph bottom up.

It is prudent to distinguish between three kinds of summaries of a procedure that can be created for minimizing the number of times a procedure is re-analyzed:

- (a.1) a bottom-up parameterized summary flow function which is context independent (context dependence is captured in the parameters),
- (a.2) a top down enumeration of summary flow function in the form of input-output pairs for the input values reaching a procedure, and
- (b) a bottom-up parameterless (and hence context-insensitive) summary information.

**EXAMPLE 1.1.** Figure 1 illustrates the three different kinds of summaries for available expressions analysis. Procedure  $f$  kills the availability of expression  $a*b$ , generates the availability of  $b*c$ , and is transparent to the availability of  $c*d$ .

- Summary (a.1) is a parameterized flow function, summary (a.2) is an enumerated flow function, whereas summary (b) is a data flow value (i.e. it is a summary information as against a summary flow function) representing the effect of all calls of procedure  $f$ .
- Summaries (a.1) and (a.2) are context-sensitive (because they compute distinct values for different calling contexts of  $f$ ) whereas summary (b) is context-insensitive (because it represents the same value regardless of the calling context of  $f$ ).
- Summaries (a.1) and (b) are context independent (because they can be constructed without requiring any information from the calling contexts of  $f$ ) whereas summary (a.2) is context dependent (because it requires information from the calling contexts of  $f$ ).

□

Note that context independence (in (a.1) above), achieves context-sensitivity through parameterization and should not be confused with context-insensitivity (in (b) above).

```

int ***x, **y;
int *z, *a, *b;
int d, e, u, v, w;
void f();
void g();
01 void f()
02 { x = &a;
03 z = &w;
04 g();
05 *x = z;
06 }
07 void g()
08 { a = &e;
09 if (...) {
10     *x = z;
11     z = &u;
12 } else {
13     y = &b;
14     z = &v;
15 }
16 x = &b;
17 *y = &d;
18 }

```

Fig. 2. A motivating example which is used as a running example through the paper. Procedures  $g$  and  $f$  are used for illustrating intraprocedural and interprocedural GPG construction respectively. All variables are global.

We focus on summaries of the first kind (a.1) because we would like to avoid re-analysis and seek context-sensitivity. We formulate our analysis on a language modelled on C.

### Our Key Idea, Approach, and an Outline of the Paper

Section 2 describes our motivation and contributions by contextualizing our work in the perspective of the past work on bottom-up summary flow functions for points-to analysis. As explained in Section 3, we essentially generalize the concept of points-to relations by using the counts of indirection levels leaving the indirectly accessed unknown locations implicit. This allows us to create summary flow functions in the form of *generalized points-to graphs* (GPGs) whose size is linearly bounded by the number of variables. By design, GPGs can represent both memory (in terms of classical points-to facts) and memory transformers (in terms of generalized points-to facts).

**EXAMPLE 1.2.** Consider procedure  $g$  of Figure 2 whose GPG is shown in Figure 4(c). The edges in GPGs track indirection levels: indirection level 1 in the label “1,0” indicates that the source is assigned the address (indicated by indirection level 0) of the target. Edge  $a \xrightarrow{1,0} e$  is created for line 8. The indirection level 2 in edge  $x \xrightarrow{2,1} z$  for line 10 indicates that the pointees of  $x$  are being defined; since  $z$  is read, its indirection level is 1. The combined effect of lines 13 (edge  $y \xrightarrow{1,0} b$ ) and 17 (edge  $y \xrightarrow{2,0} d$ ) results in the edge  $b \xrightarrow{1,0} d$ . However edge  $y \xrightarrow{2,0} d$  is also retained because there is no information about the pointees of  $y$  along the other path reaching line 17.  $\square$

The generalized points-to facts are composed to create new generalized points-to facts with smaller indirection levels (Section 4) whenever possible thereby converting them progressively to classical points-to facts. This is performed in two phases: construction of GPGs, and use of GPGs to compute points-to information. GPGs are constructed flow-sensitively by processing pointer assignments along the control flow of a procedure and collecting generalized points-to facts. (Section 5).

Function calls are handled context-sensitively by incorporating the effect of the GPG of a callee into the GPG of the caller (Section 6). Loops and recursion are handled using a fixed point computation. GPGs also distinguish between *may* and *must* pointer updates thereby facilitating strong updates.

Section 7 shows how GPGs are used for computing classical points-to facts. Section 8 defines formal semantics of GPGs and provides a proof of soundness of the proposed points-to analysis using GPGs. Section 9 describes the handling of advanced features of the language such as function pointers, structures, unions, heap, arrays and pointer arithmetic. Section 10 presents the empirical measurements. Section 11 describes the related work. Section 12 concludes the paper.

| Pointer Statement |           | Flow Function $f \in \mathbf{F} = \{\text{ad}, \text{cp}, \text{st}, \text{ld}\}$ ,<br>$f : 2^{\text{PTG}} \mapsto 2^{\text{PTG}}$                    | Placeholders<br>in $X$ |
|-------------------|-----------|---|------------------------|
| Address           | $x = \&y$ | $\text{ad}_{xy}(X) = X - \{(x, l_1) \mid l_1 \in L\} \cup \{(x, y)\}$   | $\emptyset$            |
| Copy              | $x = y$   | $\text{cp}_{xy}(X) = X - \{(x, l_1) \mid l_1 \in L\} \cup \{(x, \phi_1) \mid (y, \phi_1) \in X\}$   | $\phi_1$               |
| Store             | $*x = y$  | $\text{st}_{xy}(X) = X - \{(\phi_1, l_1) \mid (x, \phi_1) \in X, l_1 \in L\} \cup \{(\phi_1, \phi_2) \mid \{(x, \phi_1), (y, \phi_2)\} \subseteq X\}$ | $\phi_1, \phi_2$       |
| Load              | $x = *y$  | $\text{ld}_{xy}(X) = X - \{(x, l_1) \mid (x, l_1) \in L\} \cup \{(x, \phi_2) \mid \{(y, \phi_1), (\phi_1, \phi_2)\} \subseteq X\}$                    | $\phi_1, \phi_2$       |

Fig. 3. Points-to analysis flow functions for basic pointer assignments.

The core ideas of this work were presented in [25]. Apart from providing better explanations of the ideas, this paper covers the following additional aspects of this work:

- Many more details of edge composition such as (a) descriptions of  $ST$  and  $TT$  compositions (Section 4), (b) derivations of usefulness criteria depending upon the type of compositions (Section 4.2), and (c) comparison of edge composition with matrix multiplication (Section 4.4) and dynamic transitive closure (Section 5.1.4).
- Soundness proofs for points-to analysis using GPGs by defining the concepts of a concrete memory (created along a single control flow path reaching a program point) and an abstract memory (created along all control flow paths reaching a program point) (Section 8).
- Handling of advanced features such as function pointers, structures, unions, heap memory, arrays, pointer arithmetic, etc. (Section 9).

## 2. MOTIVATION AND CONTRIBUTIONS

This section highlights the issues in constructing bottom-up summary flow functions for points-to analysis. We also provide a brief overview of the past approaches along with their limitations and describe our contributions by showing how our representation of summary flow functions for points-to analysis overcomes these limitations.

### 2.1. Issues in Constructing Summary Flow Functions for Points-to Analysis

Construction of bottom-up parameterized summary flow functions requires

- *composing* statement-level flow functions to summarize the effect of a sequence of statements appearing in a control flow path, and
- *merging* the composed flow functions to represent multiple control flow paths reaching a join point in the control flow graph.

An important requirement of such a summary flow function is that it should be compact and that its size should be independent of the size of the procedure it represents. This seems hard because the flow functions need to handle indirectly accessed unknown pointees. When these pointees are defined in caller procedures, their information is not available in a bottom-up construction; information reaching a procedure from its callees is available during bottom-up construction but not the information reaching from its callers. The presence of function pointers passed as parameters pose an additional challenge for bottom-up construction for a similar reason.

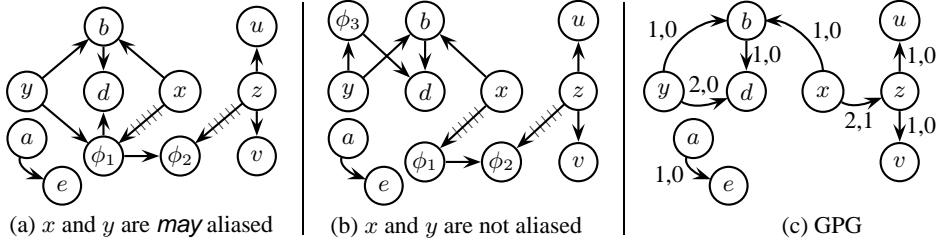


Fig. 4. PTFs/GPG for procedure  $g$  of Figure 2 for points-to analysis using placeholders  $\phi_i$ . Edges deleted due to flow-sensitivity are struck off. Our proposed representation GPG with no explicit placeholders.

## 2.2. Modelling Access of Unknown Pointees

The main difficulty in reducing meets (i.e. merges) and compositions of points-to analysis flow functions is modelling the accesses of pointees when they are not known. For the statement sequence  $x = *y; z = *x$  if the pointee information of  $y$  is not available, it is difficult to describe the effect of these statements on points-to relations symbolically. A common solution for this is to use *placeholders*<sup>1</sup> for indirect accesses. We explain the use of placeholders below and argue that they prohibit compact representation of summary flow functions because

- the resulting representation of flow functions is not closed under composition, and
- for flow-sensitive points-to analysis, a separate placeholder may be required for different occurrences of the same variable in different statements.

Let  $L$  and  $P \subseteq L$  denote the sets of locations and pointers in a program. Then, the points-to information is subset of  $PTG = P \times L$ . For a given statement, a flow function for points-to analysis computes points-to information after the statement by incorporating its effect on the points-to information that holds before the statement. It has the form  $f : 2^{PTG} \rightarrow 2^{PTG}$ . Figure 3 enumerates the space of flow functions for basic pointer assignments.<sup>2</sup> The flow functions are named in terms of the variables appearing in the assignment statement and are parameterized on the input points-to information  $X$  which may depend on the calling context. This is described in terms of placeholders in  $X$  denoted by  $\phi_1$  and  $\phi_2$  which are placeholders for the information in  $X$ . It is easy to see that the function space  $\mathbf{F} = \{\text{ad}, \text{cp}, \text{st}, \text{ld}\}$  is not closed under composition.

EXAMPLE 2.1. Let  $f$  represent the composition of flow functions for the statement sequence  $x = *y; z = *x$ . Then

$$\begin{aligned} f(X) = \text{Id}_{zx}(\text{Id}_{xy}(X)) &= (X - (\{(x, l_1) \mid (x, l_1) \in L\} \cup \{(z, l_1) \mid (z, l_1) \in L\})) \\ &\quad \cup \{(x, \phi_2) \mid \{(y, \phi_1), (\phi_1, \phi_2)\} \subseteq X\} \\ &\quad \cup \{(z, \phi_3) \mid \{(y, \phi_1), (\phi_1, \phi_2), (\phi_2, \phi_3)\} \subseteq X\} \end{aligned}$$

This has three placeholders and cannot be reduced to any of the four flow functions in the set.  $\square$

EXAMPLE 2.2. Consider the code snippet on the right for constructing a flow-sensitive summary flow function. Assume that we use  $\phi_1$  as the placeholder to denote the pointees of  $y$  and  $\phi_2$  as the placeholder to denote the pointees of pointees of  $y$ . We cannot guarantee that the pointees of  $y$  or pointees of pointees of  $y$  remains same in  $s_1$  and  $s_3$  because statement  $s_2$  could have a side effect of changing either one of them depending upon the aliases present in the calling context. Under the C model, only one of the first two combinations of aliases is possible. Assuming that  $\phi_3$  is the placeholder for  $q$ ,

|                 |
|-----------------|
| $s_1 : x = *y;$ |
| $s_2 : *z = q;$ |
| $s_3 : p = *y;$ |

<sup>1</sup>Placeholders are referred to as external variables in [19] and as extended parameters in [33]. They are parameters of the summary flow function (and not of the procedure for which the summary flow function is constructed).

<sup>2</sup>Other pointer assignments involving structures and heap are handled as described in Section 9.2.

- When  $*z$  is aliased to  $y$  before statement  $s_1$ ,  $y$  is redefined and hence, the placeholder for pointees of  $y$  in  $s_3$  will now be  $\phi_3$  otherwise it will be  $\phi_1$ .
- When  $z$  is aliased to  $y$  before statement  $s_1$ , pointees of  $y$  i.e.,  $\phi_1$  is redefined and hence, the placeholder for pointees of pointees of  $y$  in  $s_3$  will be represented by  $\phi_3$  otherwise it will be  $\phi_2$ .
- When  $z$  and  $y$  are not related, neither  $y$  nor pointees of  $y$  are redefined and hence, the placeholders for pointees of  $y$  and pointees of pointees of  $y$  for statement  $s_3$  will be same as that of statement  $s_1$ .

Thus the decision to reuse the placeholder for a flow-sensitive summary flow function depends on the aliases present in the calling contexts. It is important to observe that the combination of aliasing patterns involving other variables are ignored. Only the aliases that are likely to affect the accesses because of a redefinition need to be considered when summary flow functions are constructed.

This difficulty can be overcome by avoiding the kill due to  $s_2$  and using  $\phi_1$  for pointees of  $y$  and  $\phi_2$  for pointees of pointees of  $y$  in both  $s_1$  and  $s_3$ . If  $z$  is aliased to  $y$  or  $*z$  is aliased to  $y$  before statement  $s_1$  then both  $x$  and  $p$  will point to both  $\phi_2$  and  $\phi_3$  which is imprecise. Effectively, the summary flow function becomes flow-insensitive affecting the precision of the analysis.

Thus, introducing placeholders for the unknown pointees is not sufficient but the knowledge of aliases in the calling context is also equally important for introducing the placeholders.  $\square$

### 2.3. An Overview of Past Approaches

In this section, we explain two approaches that construct the summary flow functions for points-to analysis. Other related investigations have been reviewed in Section 11; the description in this section serves as a background to our contributions.

- *Using aliasing patterns to construct a collection of partial transfer functions (PTFs).*

This approach is “context-based” as the information about the aliases present in the calling contexts is used for summary flow function construction. A different summary flow function is constructed for every combination of aliases found in the calling contexts using the placeholders for representing the unknown pointees. This requires creation of multiple versions of a summary flow function which is represented by a collection of *partial transfer functions* (PTFs). A PTF is constructed for the aliases that could occur for a given list of parameters and global variables accessed in a procedure [33; 36].

**EXAMPLE 2.3.** For procedure  $g$  of the program in Figure 2, three placeholders  $\phi_1$ ,  $\phi_2$ , and  $\phi_3$  have been used in the PTFs shown in Figures 4(a) and (b). The possibility that  $x$  and  $y$  may or may not be aliased gives rise to two PTFs.  $\square$

The main limitation of this approach is that the number of PTFs could increase combinatorially with the number of dereferences of globals and parameters.

**EXAMPLE 2.4.** For four dereferences, we may need 15 PTFs. Consider four pointers  $a, b, c, d$ . Either none of them is aliased (1 possibility); only two of them are aliased:  $(a, b)$ ,  $(a, c)$ ,  $(a, d)$ ,  $(b, c)$ ,  $(b, d)$ , or  $(c, d)$  (6 possibilities); only three of them are aliased:  $(a, b, c)$ ,  $(a, b, d)$ ,  $(a, c, d)$ , or  $(b, c, d)$  (total 4 possibilities); all four of them are aliased:  $(a, b, c, d)$  (1 possibility); groups of aliases of two each:  $\{(a, b), (c, d)\}$ ,  $\{(a, c), (b, d)\}$ , or  $\{(a, d), (b, c)\}$  (3 possibilities). Thus the total number of PTFs is  $1 + 6 + 4 + 1 + 3 = 15$ .  $\square$

PTFs that do not correspond to actual aliasing patterns occurring in a program are irrelevant. They can be excluded by a preprocessing to discover the combination of aliases present in a program so that PTF construction can be restricted to the discovered combinations [33; 36]. The number of PTFs could still be large.

Although this approach does not introduce any imprecision, our measurements show that the number of aliasing patterns occurring in practical programs is very large which limits the usefulness of this approach.

- *Single summary flow function without using aliasing patterns.*

This approach does not make any assumption about aliases in the calling context and constructs a single summary flow function for a procedure. Hence, it is “context independent”. Owing to the absence of alias information in the calling contexts, this approach uses a new placeholder  $\phi_4$  for pointee of  $y$  and also  $\phi_5$  for pointee of pointee of  $y$  in  $s_3$  in Example 2.2. Thus, different placeholders for different accesses of the same variable are required thereby increasing the number of placeholders and hence the size of summary flow function. In a degenerate case, the size of summary flow function may be proportional to the number of statements represented by the summary flow function. This is undesirable because it may be better not to create summary flow functions and retain the original statements whose flow functions are applied one after the other. Separate placeholders for different occurrences of a variable can be avoided if points-to information is not killed by the summary flow functions [19; 31; 32]. Another alternative is to use flow-insensitive summary flow functions [7]. However, both these cases introduces imprecision.

#### 2.4. Our Contributions

A fundamental problem with placeholders is that they explicate unknown locations by naming them, resulting in either a large number of placeholders (e.g., a GPG edge  $\cdot \xrightarrow{i,j} \cdot$  would require  $i + j - 1$  placeholders) or multiple summary flow functions for different aliasing patterns that exist in the calling contexts. We overcome this difficulty by representing the summary flow function of a procedure in the form of a graph called *Generalized Points-to Graph* (GPG) and use it for flow-and context-sensitive points-to analysis.

GPGs leave pointees whose information is not available during summary construction, implicit. Our representation is characterized by the following:

- We do not need placeholders (unlike [19; 31; 32; 33; 36]). This is possible because we encode indirection levels as edge labels by replacing a sequence of indirection operators “\*” by a number.<sup>3</sup>
- We do not require any assumptions/information about aliasing patterns in the calling contexts (unlike [33; 36]) and construct a single summary flow function per procedure (unlike [33; 36]) without introducing the imprecision introduced by [19; 31; 32].
- The size of our summary flow function for a procedure does not depend on the number of statements in the procedure and is bounded by the number of global variables, formal parameters of the procedure, and its return value variable (unlike [19; 31; 32]).
- Updates can be performed in the calling contexts (unlike [7; 19; 31; 32]).

This facilitates the scalability of fully flow- and context-sensitive exhaustive points-to analysis. We construct context independent summary flow functions and context-sensitivity is achieved through parameterization in terms of indirection levels.

#### 2.5. Our Language and Scope

We have described our formulations for a language modelled on C and have organized the paper based on the features included in the language. For simplicity of exposition, we divide the language features into three levels. Our description of our analysis begins with Level 1 and is progressively extended to the Level 3.

---

<sup>3</sup>This is somewhat similar to choosing a decimal representation for integers over Peano’s representation or replacing a unary language by a binary or n-nary language [12].

| Feature   | Level |   |   | Sections |
|---|-------|---|---|----------|
|   | 1     | 2 | 3 |          |
| Pointers to scalars   | ✓     |   |   | 3, 4, 5  |
| Function Calls and Recursion                                    |       | ✓ |   | 6        |
| Function Pointers   |       |   | ✓ | 9        |
| Pointers to Structures, Unions, and Heap                        |       |   | ✓ | 9        |
| Pointer Arithmetic, Pointers to Arrays, Address Escaping Locals |       |   | ✓ | 9        |

For the first three features, the information flows from top to bottom in the call graph (i.e., from callers to callees) and hence are naturally handled by the top-down approaches of interprocedural analysis. However, a special attention is required for representing this information in the bottom-up approaches. In case of recursion, the presence of cycle in the call graph requires a fixed point computation regardless of the approach used.

Levels 1 and 2 handle the core features of the language whereas Level 3 handles the advanced features.<sup>4</sup> A preliminary version of this paper restricted to Levels 1 and 2 appeared as [25].

### 3. GENERALIZED POINTS-TO GRAPHS (GPGS)

This section defines generalized points-to graph (GPG) which represents memory manipulations without needing placeholders for unknown locations. We define the basic concepts assuming scalars and pointers in the stack and static memory; see Section 9 for extensions to handle structures, heap, function pointers, etc.

#### 3.1. Memory and Memory Transformer

We assume a control flow graph representation containing 3-address code statements. Program points  $t, u, v$  represent the points just before the execution of statements. The successors and predecessors of a program point are denoted by  $\text{succ}$  and  $\text{pred}$ ;  $\text{succ}^*$ ,  $\text{pred}^*$  denote their reflexive transitive closures. A *control flow path* is a finite sequence of (possibly repeating) program points  $q_0, q_1, \dots, q_m$  such that  $q_{i+1} \in \text{succ}(q_i)$ .

Recall that  $L$  and  $P \subseteq L$  denote the sets of locations and pointers respectively (Section 2.2). Every location has a content and an address. The *memory* at a program point is a relation  $M \subseteq P \times (L \cup \{?\})$  where “?” denotes an undefined location. We view  $M$  as a graph with  $L \cup \{?\}$  as the set of nodes. An edge  $x \rightarrow y$  in  $M$  indicates that  $x \in P$  contains the address of  $y \in L$ . The memory associated with a program point  $u$  is denoted by  $M_u$ ; since  $u$  could appear in multiple control flow paths and could also repeat in a given control flow path,  $M_u$  denotes the memory associated with all occurrences of  $u$ .

| <i>Definition 1: Memory Transformer <math>\Delta</math></i>  |  |
|--|--|
| $\Delta(u, v) := B(u, v) \sqcap \bigcap_{\substack{t \in \text{succ}^*(u) \\ v \in \text{succ}(t)}} \delta(t, v) \circ \Delta(u, t)$ |  |
| $B(u, v) := \begin{cases} \Delta_{id} & v = u \\ \delta(u, v) & v \in \text{succ}(u) \\ \emptyset & \text{otherwise} \end{cases}$    |  |

The pointees of a set of pointers  $X \subseteq P$  in  $M$  are computed by the relation application  $M[X] = \{y \mid (x, y) \in M, x \in X\}$ . Let  $M^i$  denote a composition of degree  $i$ . Then,  $M^i[x]$  discovers the  $i^{th}$  pointees of  $x$  which involves  $i$  transitive reads from  $x$ : first  $i - 1$  addresses are read followed by the content of the last address. For composability of  $M$ , we extend its domain to  $L \cup \{?\}$  by inclusion map. By definition,  $M^0[x] = \{x\}$ .

<sup>4</sup>Since our language is modelled after C, statements such as  $x = *x$  are prohibited by typing rules, and cycles in points-to graph exist only in the presence of structures.

| Pointer asgn. | Memory graph after the assignment   | Pointers defined | Pointees   | GPG edge                | Pointers over-written | Effect on $M$ after the assignment |
|---------------|---|------------------|------------|-------------------------|-----------------------|------------------------------------|
| $x = \&y$     | $x \bullet \rightarrow \bullet y$   | $M^0\{x\}$       | $M^0\{y\}$ | $x \xrightarrow{1,0} y$ | $M^0\{x\}$            | $M^1\{x\} = M^0\{y\}$              |
| $x = y$       | $x \bullet \rightarrow \bullet \leftarrow \bullet y$  | $M^0\{x\}$       | $M^1\{y\}$ | $x \xrightarrow{1,1} y$ | $M^0\{x\}$            | $M^1\{x\} = M^1\{y\}$              |
| $x = *y$      | $x \bullet \rightarrow \bullet \leftarrow \bullet \leftarrow \bullet y$                     | $M^0\{x\}$       | $M^2\{y\}$ | $x \xrightarrow{1,2} y$ | $M^0\{x\}$            | $M^1\{x\} = M^2\{y\}$              |
| $*x = y$      | $x \bullet \rightarrow \bullet \rightarrow \bullet \leftarrow \bullet \leftarrow \bullet y$ | $M^1\{x\}$       | $M^1\{y\}$ | $x \xrightarrow{2,1} y$ | $M^1\{x\}$ or none    | $M^2\{x\} \supseteq M^1\{y\}$      |

Fig. 5. GPG edges for basic pointer assignments in C. Figure 18 shows GPG edges for structures and heap. In the memory graph, a double circle indicates a shared location, a thick arrow shows the newly created edge in the memory and unnamed nodes may represent multiple pointees.

For adjacent program points  $u$  and  $v$ ,  $M_v$  is computed from  $M_u$  by incorporating the effect of the statement between  $u$  and  $v$ ,  $M_v = (\delta(u, v))(M_u)$  where  $\delta(u, v)$  is a *statement-level flow function* representing a *memory transformer*. For  $v \in \text{succ}^*(u)$ , the effect of the statements appearing in all control flow paths from  $u$  to  $v$  is computed by  $M_v = (\Delta(u, v))(M_u)$  where the memory transformer  $\Delta(u, v)$  is a *summary flow function* mapping the memory at  $u$  to the memory at  $v$ . Definition 1 provides an equation to compute  $\Delta$  without specifying a representation for it.

Since control flow paths may contain cycles,  $\Delta$  is the maximum fixed point of the equation where (a) the composition of  $\Delta$ s is denoted by  $\circ$  such that  $(g \circ f)(\cdot) = g(f(\cdot))$ , (b)  $\Delta$ s are merged using  $\sqcap$ , (c)  $B$  captures the base case, and (d)  $\Delta_{id}$  is the identity flow function. Henceforth, we use the term memory transformer for a summary flow function  $\Delta$ . The rest of the paper proposes GPG as a compact representation for  $\Delta$ . Section 3.2 defines GPG and Section 3.3 defines its lattice.

### 3.2. Generalized Points-to Graphs for Representing Memory Transformers

The classical memory transformers explicate the unknown locations using placeholders. This is a low level abstraction close to the memory, defined in terms of classical points-to facts: Given locations  $x, y \in L$ , a classical points-to fact  $x \rightarrow y$  in memory  $M$  asserts that  $x$  holds the address of  $y$ . We propose a higher level abstraction of the memory without explicating the unknown locations.

**Definition 2:** *Generalized Points-to Graph (GPG).* Given locations  $x, y \in L$ , a *generalized points-to fact*  $x \xrightarrow{i,j} y$  in a given memory  $M$  asserts that every location reached by  $i - 1$  dereferences from  $x$  can hold the address of every location reached by  $j$  dereferences from  $y$ . Thus,  $M^i\{x\} \supseteq M^j\{y\}$ . A *generalized points-to graph* (GPG) is a set of edges representing generalized points-to facts. For a GPG edge  $x \xrightarrow{i,j} y$ , the pair  $(i, j)$  represents indirection levels and is called the *indlev* of the edge ( $i$  is the *indlev* of  $x$ , and  $j$  is the *indlev* of  $y$ ).

Figure 5 illustrates the generalized points-to facts corresponding to the basic pointer assignments in C. Observe that a classical points-to fact  $x \rightarrow y$  is a special case of the generalized points-to fact  $x \xrightarrow{i,j} y$  with  $i = 1$  and  $j = 0$ ; the case  $i = 0$  does not arise.

The generalized points-to facts are more expressive than the classical points-to facts because they can be composed to create new facts as shown by the example below. Section 4 explains the process of composing the generalized points-to facts through *edge composition* along with the conditions when the facts can and ought to be composed.

**EXAMPLE 3.1.** Statements  $s_1$  and  $s_2$  to the right are represented by GPG edges  $x \xrightarrow{1,0} y$  and  $z \xrightarrow{1,1} x$  respectively. We can compose the two edges by creating a new edge  $\boxed{s_1 : x = \&y; s_2 : z = x;}$   $z \xrightarrow{1,0} y$  indicating that  $z$  points-to  $y$ . Effectively, this converts the generalized points-to fact for  $s_2$  into a classical points-to fact.  $\square$

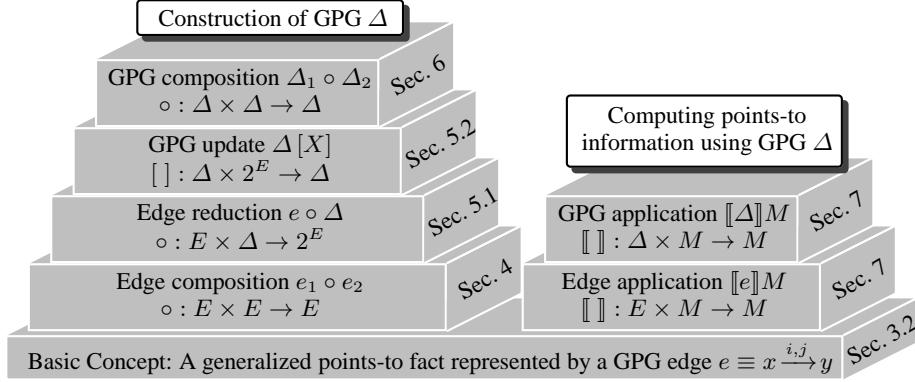


Fig. 6. A hierarchy of operations for points-to analysis using GPGs. Each operation is defined in terms of the layers below it.  $E$  denotes the set of GPG edges. By abuse of notation, we use  $M$  and  $\Delta$  also as types to indicate the signatures of the operations. The operators “ $\circ$ ” and “[ ]” are overloaded and can be disambiguated using the types of the operands.

Edges in a set are unordered. However, we want a GPG to represent a flow-sensitive memory transformer which requires the edges to be ordered. We impose this ordering externally which allows us to view the set of GPG edges as a sequence. A reverse post order traversal over the control flow graph of a procedure dictates this sequence. It is required only at the interprocedural level when the effect of a callee is incorporated in its caller. Since a sequence is totally ordered but control flow is partially ordered, the GPG operations (Section 6) internally relax the total order to ensure that the edges appearing on different control flow paths do not affect each other. While the visual presentation of GPGs as graphs is intuitively appealing, it loses the edge-ordering; we annotate edges with their ordering explicitly when it matters.

A GPG is a uniform representation for a memory transformer as well as (an abstraction of) memory. This is analogous to a matrix which can be seen both as a transformer (for a linear translation in space) and also as an absolute value. A points-to analysis using GPGs begins with generalized points-to facts  $\cdot \xrightarrow{i,j} \cdot$  representing memory transformers which are composed to create new generalized points-to facts with smaller *indlevs* thereby progressively reducing them to classical points-to facts  $\cdot \xrightarrow{1,0} \cdot$  representing memory.

### 3.3. The Lattice of GPGs

Definition 3 describes the meet semi-lattice of GPGs. For reasons described later in Section 6, we need to introduce an artificial  $\top$  element denoted  $\Delta_\top$  in the lattice. It is used as the initial value in the data flow equations for computing GPGs (Definition 5 which instantiates Definition 1 for GPGs).

| <i>Definition 3: Lattice of GPGs</i>  |
|---|
| $\Delta \in \{\Delta_\top\} \cup \{(\mathcal{N}, \mathcal{E}) \mid \mathcal{N} \subseteq N, \mathcal{E} \subseteq E\}$  |
| <b>where</b>  |
| $N := L \cup \{?\}$   |
| $E := \{x \xrightarrow{i,j} y \mid x \in P, y \in N, 0 < i \leq  N , 0 \leq j \leq  N \}$   |
| $\Delta_1 \sqsubseteq \Delta_2 \Leftrightarrow (\Delta_2 = \Delta_\top) \vee (\mathcal{N}_1 \supseteq \mathcal{N}_2 \wedge \mathcal{E}_1 \supseteq \mathcal{E}_2)$  |
| $\Delta_1 \sqcap \Delta_2 := \begin{cases} \Delta_1 & \Delta_2 = \Delta_\top \\ \Delta_2 & \Delta_1 = \Delta_\top \\ (\mathcal{N}_1 \cup \mathcal{N}_2, \mathcal{E}_1 \cup \mathcal{E}_2) & \text{otherwise} \end{cases}$ |

| Statement sequence     | GPG                |                   |
|------------------------|--------------------|-------------------|
|                        | Before composition | After composition |
| $x = \&y$<br>$z = x$   |                    |                   |
| $x = \&zy$<br>$*x = z$ |                    |                   |

Fig. 7. Examples of edge compositions for points-to analysis.

The sequencing of edges is maintained externally and is explicated where required. This allows us to treat a GPG (other than  $\Delta_{\top}$ ) as a pair of a set of nodes and a set of edges. The partial order is a point-wise superset relation applied to the pairs. Similarly, the meet operation is a point-wise union of the pairs. It is easy to see that the lattice is finite because the number of locations  $L$  is finite (being restricted to static and stack slots). When we extend GPGs to handle heap memory (Section 9.2), explicit summarization is required to ensure finiteness. The finiteness of the lattice and the monotonicity of GPG operations guarantee the convergence of GPG computations on a fixed point; starting from  $\Delta_{\top}$ , we compute the maximum fixed point.

For convenience, we treat a GPG as a set of edges leaving the set of nodes implicit; the GPG nodes can always be inferred from the GPG edges.

### 3.4. A Hierarchy of GPG Operations

Figure 6 lists the GPG operations based on the concept of the generalized points-to facts. They are presented in two separate columns according to the two phases of our analysis and each layer is defined in terms of the layers below it. The operations are defined in the sections listed against them in Figure 6.

**Constructing GPGs.** An *edge composition*  $e_1 \circ e_2$  computes a new edge  $e_3$  equivalent to  $e_1$  using the points-to information in  $e_2$  such that the *indlev* of  $e_3$  is smaller than that of  $e_1$ . An *edge reduction*  $e_1 \circ \Delta$  computes a set of edges  $X$  by composing  $e_1$  with the edges in  $\Delta$ . A *GPG update*  $\Delta_1 [X]$  incorporates the effect of the set of edges  $X$  in  $\Delta_1$  to compute a new GPG  $\Delta_2$ . A *GPG composition*  $\Delta_1 \circ \Delta_2$  composes a callee’s GPG  $\Delta_2$  with GPG  $\Delta_1$  at a call point to compute a new GPG  $\Delta_3$ .

**Using GPGs for computing points-to information.** An *edge application*  $\llbracket e \rrbracket M$  computes a new memory  $M'$  by incorporating the effect of the GPG edge  $e$  in memory  $M$ . A *GPG application*  $\llbracket \Delta \rrbracket M$  applies the GPG  $\Delta$  to  $M$  and computes a new memory  $M'$  using edge application iteratively.

These operations allow us to build the theme of a GPG being a uniform representation for both memory and memory transformers. This uniformity of representation leads to the following similarity in operations: (a) an edge application to a memory ( $\llbracket e \rrbracket M$ ) is similar to an edge reduction in GPG ( $e \circ \Delta$ ), and (b) GPG application to a memory ( $\llbracket \Delta \rrbracket M$ ) is similar to GPG composition ( $\Delta_1 \circ \Delta_2$ ).

## 4. EDGE COMPOSITION

This section defines edge composition as a fundamental operation which is used in Section 5 for constructing GPGs. We begin by introducing edge composition and then explore the concept in its full glory by describing the types of compositions and characterizing the properties of compositions such as *usefulness*, *relevance*, and *conclusiveness*. Some of these considerations are governed by the goal of including the resulting edges in a GPG  $\Delta$ ; the discussion on inclusion of edges in GPG  $\Delta$  is relegated to Section 5.2.

Regardless of the direction of an edge,  $i$  in *indlev* “ $i, j$ ” represents its source while  $j$  represents its target.  
Balancing the *indlevs* of  $x$  (the pivot of composition) in  $p$  and  $n$  allows us to join  $y$  and  $z$  to create a reduced edge  $r = n \circ p$  shown by dashed arrows.

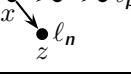
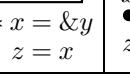
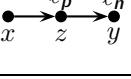
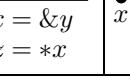
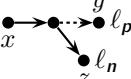
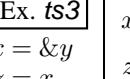
| Possible SS Compositions                           |   |  | Possible TS Compositions                           |   |  |
|--|---|--|--|---|--|
| Statement sequence                                 | Memory graph  | GPG edges  | Statement sequence                                 | Memory graph  | GPG edges  |
| $s_n^c < s_p^c$                                    |   |  | $T_n^c < T_p^c$                                    |   |  |
| Ex. ss1<br>$*x = \&y$<br>$x = \&z$                 |  | $p: x \xrightarrow{2,0} y$<br>$n: x \xrightarrow{1,0} z$<br>(irrelevant)               | Ex. ts1<br>$*x = \&y$<br>$z = x$                   |  | $p: x \xrightarrow{2,0} y$<br>$n: z \xrightarrow{1,1} x$<br>(not useful)               |
| $s_n^c > s_p^c$ (Additionally $T_p^c \leq s_p^c$ ) |   |  | $T_n^c > s_p^c$ (Additionally $T_p^c \leq s_p^c$ ) |   |  |
| Ex. ss2<br>$x = \&z$<br>$*x = \&y$                 |  | $p: x \xrightarrow{1,0} z$<br>$n: x \xrightarrow{2,0} y$<br>$r: z \xrightarrow{1,0} y$ | Ex. ts2<br>$x = \&y$<br>$z = *x$                   |  | $p: x \xrightarrow{1,0} y$<br>$n: z \xrightarrow{1,2} x$<br>$r: z \xrightarrow{1,1} y$ |
| $s_n^c = s_p^c$                                    |   |  | $T_n^c = s_p^c$ (Additionally $T_p^c \leq s_p^c$ ) |   |  |
| Ex. ss3<br>$*x = \&y$<br>$*x = \&z$                |  | $p: x \xrightarrow{2,0} y$<br>$n: x \xrightarrow{2,0} z$<br>(irrelevant)               | Ex. ts3<br>$x = \&y$<br>$z = x$                    |  | $p: x \xrightarrow{1,0} y$<br>$n: z \xrightarrow{1,1} x$<br>$r: z \xrightarrow{1,0} y$ |

Fig. 8. Illustrating all exhaustive possibilities of SS and TS compositions (the pivot is  $x$ ). Dashed edges are killed. Unmarked compositions are *relevant* and *useful* (Section 4.2); since the statements are consecutive, they are also *conclusive* (Section 4.3) and hence *desirable*.

Let a statement-level flow function  $\delta$  be represented by an edge  $n$  (“new” edge) and consider an existing edge  $p \in \Delta$  (“processed” edge). Edges  $n$  and  $p$  can be composed (denoted  $n \circ p$ ) provided they have a common node called the *pivot* of composition. The goal is to *reduce* (i.e., simplify)  $n$  by using the points-to information from  $p$ . This is achieved by using the pivot as a bridge to join the remaining two nodes resulting in a reduced edge  $r$ . This requires the *indlevs* of the pivot in both edges to be made the same. For example, given edges  $n \equiv z \xrightarrow{i,j} x$  and  $p \equiv x \xrightarrow{k,l} y$  with a pivot  $x$ , if  $j > k$ , then the difference  $j - k$  can be added to the *indlevs* of nodes in  $p$ , to view  $p$  as  $x \xrightarrow{j,(l+j-k)} y$ . This balances the *indlevs* of  $x$  in the two edges allowing us to create a reduced edge  $r \equiv z \xrightarrow{i,(l+j-k)} y$ . Although this computes the transitive effect of edges, in general, it cannot be modelled using multiplication of matrices representing graphs as explained in Section 4.4.

EXAMPLE 4.1. In the first example in Figure 7, the *indlevs* of pivot  $x$  in both  $p$  and  $n$  is the same allowing us to join  $z$  and  $y$  through an edge  $z \xrightarrow{1,0} y$ . In the second example, the difference  $(2-1)$  in the *indlevs* of  $x$  can be added to the *indlevs* of nodes in  $p$  viewing it as  $x \xrightarrow{2,1} y$ . This allows us to join  $y$  and  $z$  creating the edge  $y \xrightarrow{1,1} z$ .  $\square$

Let an edge  $n$  be represented by the triple  $(S_n, (s_n^c, T_n^c), T_n)$  where  $S_n$  and  $T_n$  are the source and the target of  $n$  and  $(s_n^c, T_n^c)$  is the *indlev*. Similarly,  $p$  is represented by  $(S_p, (s_p^c, T_p^c), T_p)$  and the reduced edge  $r = n \circ p$  by  $(S_r, (s_r^c, T_r^c), T_r)$ ;  $(s_r^c, T_r^c)$  is obtained by balancing the *indlev* of the pivot in  $p$  and  $n$ . The pivot of a composition, denoted  $\mathbb{P}$ , may be the source or the target of  $n$  and  $p$ . Thus, a composition  $n \circ p$  can be of the following four types: SS, TS, ST, and TT composition.

- TS composition. In this case,  $T_n = S_p$  i.e., the pivot is the target of  $n$  and the source of  $p$ . Node  $S_n$  becomes the source and  $T_p$  becomes the target of the reduced edge  $r$ .
- SS composition. In this case,  $S_n = S_p$  i.e., the pivot is the source of both  $n$  and  $p$ . Node  $T_p$  becomes the source and  $T_n$  becomes the target of the reduced edge  $r$ .

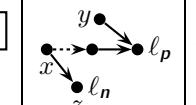
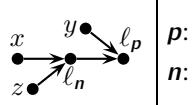
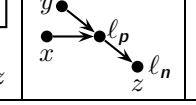
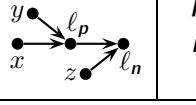
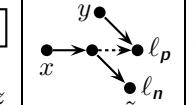
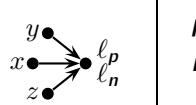
| Possible $ST$ Compositions                         |   |  | Possible $TT$ Compositions                         |  |  |
|--|---|--|--|--|--|
| Statement sequence                                 | Memory graph  | HRG edges  | Statement sequence                                 | Memory graph   | HRG edges  |
| $S_n^C < T_p^C$                                    |   |  | $T_n^C < T_p^C$                                    |  |  |
| Ex. <b>st1</b><br>$y = *x$<br>$x = \&z$            |  | $p: y \xrightarrow{1,2} x$<br>$n: x \xrightarrow{1,0} z$<br>(irrelevant)               | Ex. <b>tt1</b><br>$y = *x$<br>$z = x$              |  | $p: y \xrightarrow{1,2} x$<br>$n: z \xrightarrow{1,1} x$<br>(not useful)               |
| $S_n^C > T_p^C$ (Additionally $S_p^C \leq T_p^C$ ) |   |  | $T_n^C > T_p^C$ (Additionally $S_p^C \leq T_p^C$ ) |  |  |
| Ex. <b>st2</b><br>$y = x$<br>$*x = \&z$            |  | $p: y \xrightarrow{1,1} x$<br>$n: x \xrightarrow{2,0} z$<br>$r: y \xrightarrow{2,0} z$ | Ex. <b>tt2</b><br>$y = x$<br>$z = *x$              |  | $p: y \xrightarrow{1,1} x$<br>$n: z \xrightarrow{1,2} x$<br>$r: z \xrightarrow{1,2} y$ |
| $S_n^C = T_p^C$                                    |   |  | $T_n^C = T_p^C$ (Additionally $S_p^C \leq T_p^C$ ) |  |  |
| Ex. <b>st3</b><br>$y = *x$<br>$*x = \&z$           |  | $p: y \xrightarrow{1,2} x$<br>$n: x \xrightarrow{2,0} z$<br>(irrelevant)               | Ex. <b>tt3</b><br>$y = x$<br>$z = x$               |  | $p: y \xrightarrow{1,1} x$<br>$n: z \xrightarrow{1,1} x$<br>$r: z \xrightarrow{1,1} y$ |

Fig. 9. Illustrating all exhaustive possibilities of  $ST$  and  $TT$  compositions (the pivot is  $x$ ). See Figure 8 for illustrations of  $SS$  and  $TS$  compositions. In each case, the pivot of the composition is  $x$ .

- $ST$  composition. In this case,  $S_n = T_p$  i.e., the pivot is the source of  $n$  and the target of  $p$ . Node  $S_p$  becomes the source and  $T_n$  becomes the target of the reduced edge  $r$ .
- $TT$  composition. In this case,  $T_n = T_p$  i.e., the pivot is the target of both  $n$  and  $p$ . Node  $S_n$  becomes the source and  $S_p$  becomes the target of the reduced edge  $r$ .

Consider an edge composition  $r = n \circ p$ ,  $p \in \Delta$ . For constructing a new  $\Delta$ , we wish to include  $r$  rather than  $n$ : Including both of them is sound but may lead to imprecision; including only  $n$  is also sound but may lead to inefficiency because it forsakes summarization. An edge composition is *desirable* if and only if it is *relevant*, *useful*, and *conclusive*. We define these properties below and explain them in the rest of the section.

- A composition  $n \circ p$  is *relevant* only if it preserves flow-sensitivity.
- A composition  $n \circ p$  is *useful* only if the *indlev* of the resulting edge does not exceed the *indlev* of  $n$ .
- A composition  $n \circ p$  is *conclusive* only if the information supplied by  $p$  used for reducing  $n$  is not likely to be invalidated by the intervening statements.

When the edge composition is *desirable*, we include  $r$  in  $\Delta$  being constructed, otherwise we include  $n$ . In order to explain the *desirable* compositions, we use the following notation: Let  $\ell_p$  denote a  $(\mathbb{P}_p^C)^{th}$  pointee of pivot  $\mathbb{P}$  accessed by  $p$  and  $\ell_n$  denote a  $(\mathbb{P}_n^C)^{th}$  pointee of  $\mathbb{P}$  accessed by  $n$ .  $\mathbb{P}$  is never used as a subscript. Thus a  $p$  appearing in a subscript (e.g. in  $\ell_p$ ) refers to an edge  $p$ .

#### 4.1. Relevant Edge Composition.

An edge composition is *relevant* if it preserves flow-sensitivity. This requires the indirection levels in  $n$  to be reduced by using the points-to information in  $p$  (where  $p$  appears before  $n$  along a control flow path) but not vice-versa. The presence of a points-to path in memory (which is the transitive closure of the points-to edges) between  $\ell_p$  and  $\ell_n$  (denoted by  $\ell_p \rightarrow \ell_n$  or  $\ell_n \rightarrow \ell_p$ ) indicates that  $p$  can be used to resolve the indirection levels in  $n$ .

EXAMPLE 4.2. For  $s_n^c < s_p^c$  in Figure 8 (Ex. **ss1**), edge **p** updates the pointee of **x** and edge **n** redefines **x**. As shown in the memory graph, there is no path between  $\ell_p$  and  $\ell_n$  and hence **y** and **z** are unrelated rendering this composition *irrelevant*. Similarly, edge composition is *irrelevant* for  $s_n^c = s_p^c$  (Ex. **ss3**),  $s_n^c < \tau_p^c$  (Ex. **st1**), and  $s_n^c = \tau_p^c$  (Ex. **st3**).

For  $s_n^c > s_p^c$  (Ex. **ss2**),  $\ell_p \rightarrow \ell_n$  holds in the memory graph and hence this composition is *relevant*. For Ex. **ts1**,  $\ell_n \rightarrow \ell_p$  holds; for **ts2**,  $\ell_p \rightarrow \ell_n$  holds; for **ts3** both paths hold. Hence, all three compositions are *relevant*.  $\square$

Owing to flow-sensitivity, edge composition is not commutative although it is associative.

LEMMA 4.1. *Edge composition is associative.*

$$(e_1 \circ e_2) \circ e_3 = e_1 \circ (e_2 \circ e_3)$$

PROOF. Edge composition computes *indlevs* using arithmetic expressions involving binary plus (+) and binary minus (−). They can be made to associate by replacing binary minus (−) with binary plus (+) and unary minus (−), eg.  $a + b + (-c)$  instead of  $a + b - c$ .  $\square$

#### 4.2. Useful Edge Composition.

The *usefulness* of edge composition characterizes progress in conversion of the generalized points-to facts to the classical points-to facts. This requires the *indlev* ( $s_r^c, \tau_r^c$ ) of the reduced edge **r** to satisfy the following constraint:

$$s_r^c \leq s_n^c \wedge \tau_r^c \leq \tau_n^c \quad (1)$$

Intuitively, this ensures that the *indlev* of the new source and the new target does not exceed the corresponding *indlev* in the original edge **n**.

EXAMPLE 4.3. Consider Ex. **ts1** of Figure 8, in which  $\tau_n^c < s_p^c$ , and  $\ell_n \rightarrow \ell_p$  holds in the memory graph. Although this composition is *relevant*, it is not *useful* because the *indlev* of **r** exceeds the *indlev* of **n**. For this example, a *TS* composition will create an edge  $z \xrightarrow{2,0} y$  whose *indlev* is higher than that of **n** ( $z \xrightarrow{1,1} x$ ). Similarly, an edge composition is not *useful* when  $\tau_n^c < \tau_p^c$  (Ex. **tt1**).  $\square$

Thus, we need  $\ell_p \rightarrow \ell_n$ , and not  $\ell_n \rightarrow \ell_p$ , to hold in the memory graph for a *useful* edge composition. We can relate this with the *usefulness* criteria (Inequality 1). The presence of a path  $\ell_p \rightarrow \ell_n$  ensures that the *indlev* of edge **r** does not exceed that of **n**.

From Figure 8, we conclude that an edge composition is *relevant* and *useful* only if there exists a path  $\ell_p \rightarrow \ell_n$  rather than  $\ell_n \rightarrow \ell_p$ . *Intuitively, such a path guarantees that the updates made by **n** do not invalidate the generalized points-to fact represented by **p**.* Hence, the two generalized points-to facts can be composed by using the pivot as a bridge to create a new generalized points-to fact represented by **r**.

*Deriving the Composition Specific Conditions for Usefulness of Edge Compositions.* Constraint 1 can be further refined for a composition based on its type. We show the derivation of the *usefulness* criterion by examining the cases for *relevant* edge compositions. For simplicity, we consider only *TS* composition. There are three cases to be considered:  $\tau_n^c > \tau_p^c$ ,  $\tau_n^c < \tau_p^c$  and  $\tau_n^c = \tau_p^c$ . We have already seen that the case  $\tau_n^c < s_p^c$  is *irrelevant* in that it results in an imprecision in points-to information and hence we ignore this case. We derive a constraint for the case  $\tau_n^c > s_p^c$ . The *indlev* ( $s_r^c, \tau_r^c$ ) of the reduced edge **r** for the case  $\tau_n^c > s_p^c$ , by balancing the *indlev* of the pivot  $T_n/S_p$  in edges **n** and **p**, is given as

$$(s_r^c, \tau_r^c) = (s_n^c, \tau_p^c + \tau_n^c - s_p^c)$$

By imposing the *usefulness* constraint (Inequality 1) we get:

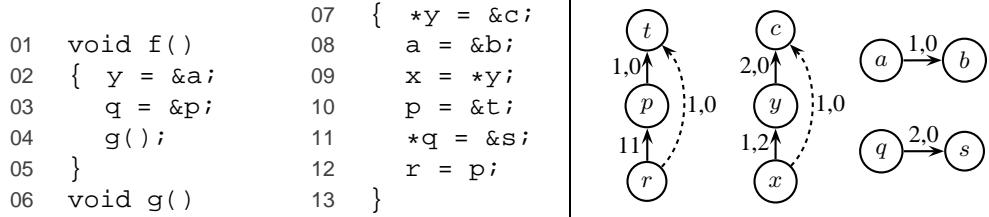


Fig. 10. Excluding inconclusive compositions (reduced edges shown by dashes are excluded).

$$\begin{aligned}
& (\tau_n^c > s_p^c) \wedge (s_r^c \leq s_n^c) \wedge (\tau_r^c \leq \tau_n^c) \\
\Rightarrow & (\tau_n^c > s_p^c) \wedge (s_n^c \leq s_p^c) \wedge (\tau_p^c + \tau_n^c - s_p^c \leq \tau_n^c) \\
\Rightarrow & (\tau_n^c > s_p^c) \wedge (\tau_p^c \leq s_p^c) \\
\Rightarrow & \tau_p^c \leq s_p^c < \tau_n^c
\end{aligned}$$

We can also derive a *usefulness* constraint for the case  $\tau_n^c = s_p^c$ . The final condition for a *useful* TS composition combined for both the cases is:

$$\tau_p^c \leq s_p^c \leq \tau_n^c \quad (\text{TS composition}) \quad (2)$$

Similarly, we can derive the criterion for other compositions by examining the *relevant* and *useful* cases for them which turn out to be:

$$\tau_p^c \leq s_p^c < s_n^c \quad (\text{SS composition}) \quad (3)$$

$$s_p^c \leq \tau_p^c < s_n^c \quad (\text{ST composition}) \quad (4)$$

$$s_p^c \leq \tau_p^c \leq \tau_n^c \quad (\text{TT composition}) \quad (5)$$

**EXAMPLE 4.4.** Consider a TS composition where  $n$  is  $z \xrightarrow{1,1} x$  and  $p$  is  $x \xrightarrow{2,1} y$  violating the constraint  $s_p^c < \tau_n^c$  (Inequality 2) because  $2 \not\leq 1$ . Edge  $n$  needs pointees of  $x$  whereas  $p$  provides information in terms of the pointees of pointees of  $x$ . Similarly, a TS composition of  $z \xrightarrow{1,2} x$  as  $n$  and  $x \xrightarrow{1,2} y$  as  $p$  violates the constraint  $\tau_p^c \leq s_p^c$  (Inequality 2). In this case,  $n$  needs pointees of pointees of  $x$  whereas  $p$  provides information in terms of pointees of pointees of  $y$ .  $\square$

In both these cases, the *indlev* of  $r$  exceeds the *indlev* of  $n$  and hence we do not perform such compositions. Similarly, we can reason about the *usefulness* constraint (Inequalities 3–5) for other types of compositions.

#### 4.3. Conclusive Edge Composition.

Recall that  $r = n \circ p$  is *relevant* and *useful* if we expect a path  $\ell_p \rightarrow \ell_n$  in the memory. This composition is *conclusive* when location  $\ell_p$  remains accessible from the pivot  $\mathbb{P}$  in  $p$  when  $n$  is composed with  $p$ . Location  $\ell_p$  may become inaccessible from  $\mathbb{P}$  because of a combined effect of the statements in a calling context and the statements in the procedure being processed. Hence, the composition is *undesirable* and may lead to unsoundness if  $r$  is included in  $\Delta$  instead of  $n$ .

**EXAMPLE 4.5.** Line 07 of procedure  $g$  in Figure 10 indirectly defines  $a$  (because  $y$  points to  $a$  as defined on line 02 of procedure  $f$ ) whereas line 08 directly defines  $a$  overwriting the value assigned on line 07. Thus,  $x$  points to  $b$  and not  $c$  after line 09. However, during GPG construction of procedure  $g$ , the relationship between  $y$  and  $a$  is not known. Thus, the composition of  $n \equiv x \xrightarrow{1,2} y$  with  $p \equiv y \xrightarrow{2,0} c$  results in  $r \equiv x \xrightarrow{1,0} c$ . In this case,  $\ell_p$  is  $c$ , however it is not reachable from  $y$

anymore as the pointee of  $y$  (which is  $a$ ) is redefined by line 08. Worse still, we do not have  $y \xrightarrow{2,0} b$  and hence edge  $x \xrightarrow{1,0} b$  cannot be created leading to unsoundness.

Similarly, line 10 defines  $p$  directly whereas line 11 defines  $p$  indirectly (because  $q$  points to  $p$  as defined on line 03 of procedure  $f$ ). The composition of  $n \equiv r \xrightarrow{1,1} p$  with  $p \equiv p \xrightarrow{1,0} t$  results in  $r \equiv r \xrightarrow{1,0} t$ . In this case,  $\ell_p$  is  $t$ , however it is not reachable from  $p$  anymore as the pointee of  $p$  is redefined indirectly by line 11. Thus we miss out the edge  $r \xrightarrow{1,0} s$  leading to unsoundness.  $\square$

Since the calling context is not available during GPG construction, we are forced to retain edge  $n$  in the GPG, thereby missing an opportunity of reducing the *indlev* of  $n$ . Hence we propose the following condition for *conclusiveness*:

- (a) The statements of  $p$  and  $n$  should be consecutive on every control flow path.
- (b) If the statements of  $p$  and  $n$  are not consecutive on some control flow path, we require that
  - (i) the intervening statements should not have an indirect assignment (e.g.,  $*x = \dots$ ), and
  - (ii) the pointee of pivot  $\mathbb{P}$  in edge  $p$  has been found i.e.  $\mathbb{P}_p^c = 1$ .

**EXAMPLE 4.6.** For the program in Figure 10, consider the composition of  $n \equiv x \xrightarrow{1,2} y$  with  $p \equiv y \xrightarrow{2,0} c$ . Since the pointee of  $y$  (which is  $a$ ) is redefined by line 08 violating the condition  $\mathbb{P}_p^c = 1$ , this composition is not conclusive and we add  $n \equiv x \xrightarrow{1,2} y$  instead of  $r \equiv x \xrightarrow{1,0} c$ . Similarly, for the composition of  $n \equiv r \xrightarrow{1,1} p$  with  $p \equiv p \xrightarrow{1,0} t$ , the pointee of  $p$  is redefined indirectly by line 11 violating the condition that  $p$  and  $n$  should not have an intervening indirect assignment. Thus this composition is inconclusive and we add  $n \equiv r \xrightarrow{1,1} p$  instead of  $r \equiv r \xrightarrow{1,0} t$ .  $\square$

This avoids a greedy reduction of  $n$  when the available information is *inconclusive*.

#### 4.4. Can Edge Composition be Modelled as Matrix Multiplication?

Edge composition  $n \circ p$  computes transitive effects of edges  $n$  and  $p$ . This is somewhat similar to the reachability computed in a graph: If there are edges  $x \rightarrow y$  and  $y \rightarrow z$  representing the facts that  $y$  is reachable from  $x$  and  $z$  is reachable from  $y$ , then it follows that  $z$  is reachable from  $x$  and an edge  $x \rightarrow z$  can be created. If the graph is represented by an adjacency matrix  $A$  in which the element  $(x, y)$  represents reachability of  $y$  from  $x$ , matrix multiplication  $A \times A$  can be used to compute the transitive effect.

It is difficult to model edge composition in this manner because of the following reasons:

- Edge labels are pairs of numbers representing indirection levels. Hence we will need to device an appropriate operator and the usual multiplication would not work.
- Edge composition has some additional constraints over reachability because of desirability; undesirable compositions are not performed. These restrictions are difficult to model in matrix multiplication.
- Transitive reachability considers only the edges of the kind  $x \rightarrow y$  and  $y \rightarrow z$ ; i.e. the pivot should be the target of the first edge and the source of the second edge. Edge composition considers pivot as both source as well as target in both the edges and hence considers all four compositions ( $SS$ ,  $TT$ ,  $TS$ , and  $ST$ ). For example, we compose  $x \xrightarrow{1,0} z$  and  $x \xrightarrow{2,0} y$  in an  $SS$  composition to create a new edge  $z \xrightarrow{1,0} y$ . Transitive reachability computed using matrix multiplication can consider only  $TS$  composition.

Thus, matrix multiplication does not model edge composition naturally.

## 5. CONSTRUCTING GPGS AT THE INTRAPROCEDURAL LEVEL

In this section we define edge reduction, and GPG update for computing a new GPG by incorporating the effect of an edge in the existing GPG. GPG composition is described in Section 6 which shows how procedure calls are handled.

### 5.1. Edge Reduction $n \circ \Delta$

In this section, we motivate the need for edge reduction and discuss the issues arising out of cascaded compositions across different types of compositions. Then, we define edge reduction which in turn is used for constructing  $\Delta$ .

**5.1.1. The Need of Edge Reduction.** Given an edge  $n$  and a GPG  $\Delta$ , edge reduction, denoted  $n \circ \Delta$ , reduces the *indlev* of  $n$  progressively by using the edges from  $\Delta$  through a series of edge compositions. For example, an edge  $x \xrightarrow{1,2} y$  requires two *TS* compositions: first one for identifying the pointees of  $y$  and second one for identifying the pointees of pointees of  $y$ . Similarly, for an edge  $x \xrightarrow{2,1} y$ , *SS* and *TS* edge compositions are required for identifying the pointees of  $x$  which are being defined and the pointees of  $y$  whose addresses are being assigned. Thus, the result of edge reduction is the fixed point computation of the cascaded edge compositions.

**5.1.2. Restrictions on Cascaded Edge Compositions.** An indiscriminate series of edge compositions may cause a reduced edge  $r = n \circ p$  to be composed again with  $p$ . In some cases, this may restore the original edge  $n$  i.e.,  $(n \circ p) \circ p = n$ , nullifying the effect of the earlier composition. To see why this happens, assume that the first composition eliminates the pivot  $x$  which is replaced by  $y$  in the reduced edge. The second composition may eliminate the node  $y$  as the pivot re-introducing node  $x$ . This is illustrated as follows:

EXAMPLE 5.1. Consider the example *st2* in Figure 9. An *ST* composition between  $p \equiv y \xrightarrow{1,1} x$  and  $n \equiv x \xrightarrow{2,0} z$  eliminates the pivot  $x$  and replaces it with  $y$  in the reduced edge  $r \equiv y \xrightarrow{2,0} z$ . This reduced edge can now be treated as edge  $n$  to further compose with  $p \equiv y \xrightarrow{1,1} x$  using  $y$  as the pivot resulting in a new reduced edge  $x \xrightarrow{2,0} z$  in which  $x$  has been re-introduced, thereby restoring the original edge.  $\square$

This nullification may happen with an *ST* composition followed by an *SS* composition or vice-versa. Similarly, a *TT* composition and a *TS* composition nullify the effect of each other. Since, edge reduction uses a fixed point computation, the computation may oscillate between the original and the reduced edges causing non-termination. In order to ensure termination, we restrict the combinations of edge compositions to the following four possibilities: *SS + TS*, *SS + TT*, *ST + TS*, and *ST + TT*. For our implementation, we have chosen the first combination i.e, *SS + TS*. We formalize the operation of edge reduction for this combination in the rest of this section.

**5.1.3. Edge Reduction using SS and TS Edge Compositions.** Edge reduction  $n \circ \Delta$  uses the edges in  $\Delta$  to compute a set of edges whose *indlevs* do not exceed that of  $n$  (Definition 4).

The results of *SS* and *TS* compositions are denoted by  $SS''_{\Delta}$  and  $TS''_{\Delta}$  which compute *relevant* and *useful* edge compositions; the *inconclusive* edge compositions are filtered out independently. The edge ordering is not required at the intraprocedural level; a reverse post order traversal over the control flow graph suffices.

A single-level composition (*s/c*) combines  $SS''_{\Delta}$  with  $TS''_{\Delta}$ . When both *TS* and *SS* compositions are possible (first case in *s/c*), the join operator  $\bowtie$  combines their effects by creating new edges by joining the sources from  $SS''_{\Delta}$  with the targets from  $TS''_{\Delta}$ . If neither of *TS* and *SS* compositions are possible (second case in *s/c*), edge  $n$  is considered as the reduced edge. If only one of them is possible, its result becomes the result of *s/c* (third case). Since the reduced edges computed by *s/c* may compose with other edges in  $\Delta$ , we extend *s/c* to multi-level composition (*m/c*) which recursively composes edges in  $X$  with edges in  $\Delta$  through function *s/ces* which extends *s/c* to a set of edges.

| <b>Definition 4: Edge reduction in <math>\Delta</math></b>   |  |
|--|--|
| $n \circ \Delta := mlc(\{n\}, \Delta)$   |  |
| <b>where</b>   |  |
| $mlc(X, \Delta) := \begin{cases} X & slices(X, \Delta) = X \\ mlc(slices(X, \Delta), \Delta) & \text{Otherwise} \end{cases}$   |  |
| $slices(X, \Delta) := \bigcup_{e \in X} slc(e, \Delta)$  |  |
| $slc(n, \Delta) := \begin{cases} SS_\Delta^n \bowtie TS_\Delta^n & SS_\Delta^n \neq \emptyset, TS_\Delta^n \neq \emptyset \\ \{n\} & SS_\Delta^n = TS_\Delta^n = \emptyset \\ SS_\Delta^n \cup TS_\Delta^n & \text{Otherwise} \end{cases}$ |  |
| $SS_\Delta^n := \{n \circ p \mid p \in \Delta, S_n = S_p, T_p^c \leq S_p^c < S_n^c\}$  |  |
| $TS_\Delta^n := \{n \circ p \mid p \in \Delta, T_n = S_p, T_p^c \leq S_p^c \leq T_n^c\}$   |  |
| $X \bowtie Y := \{(S_n, (S_n^c, T_p^c), T_p) \mid n \in X, p \in Y\}$  |  |

EXAMPLE 5.2. When  $n$  represents a statement  $x = *y$ , we need multi-level compositions: The first-level composition identifies pointees of  $y$  while the second-level composition identifies the pointees of pointees of  $y$ . This is facilitated by function  $mlc$ . Consider the code snippet on the right.  $\Delta = \{y \xrightarrow{1,0} a, a \xrightarrow{1,0} b\}$  for  $n \equiv x \xrightarrow{1,2} y$  (statement  $s_3$ ). This involves two consecutive  $TS$  compositions. The first composition involves  $y \xrightarrow{1,0} a$  as  $p$  resulting in  $TS_\Delta^n = \{x \xrightarrow{1,1} a\}$  and  $SS_\Delta^n = \emptyset$ .

This satisfies the third case of  $slc$ . Then,  $slices$  is called with  $X = \{x \xrightarrow{1,1} a\}$ . The second  $TS$  composition between  $x \xrightarrow{1,1} a$  (as a new  $n$ ) and  $a \xrightarrow{1,0} b$  (as  $p$ ) results in a reduced edge  $x \xrightarrow{1,0} b$ .  $slices$  is called again with  $X = \{x \xrightarrow{1,0} b\}$  which returns  $X$ , satisfying the base condition of  $mlc$ .  $\square$

EXAMPLE 5.3. Single-level compositions are combined using  $\bowtie$  when  $n$  represents  $*x = y$ .

For the code snippet on the right,  $SS_\Delta^n$  returns  $\{a \xrightarrow{1,1} y\}$  and  $TS_\Delta^n$  returns  $\{x \xrightarrow{2,0} b\}$  when  $n$  is  $x \xrightarrow{2,1} y$  (for statement  $s_3$ ). The join operator  $\bowtie$  combines the effect of  $TS$  and  $SS$  compositions by combining the sources from  $SS_\Delta^n$  and the targets from  $TS_\Delta^n$  resulting in a reduced edge  $r \equiv a \xrightarrow{1,0} b$ .  $\square$

**5.1.4. A Comparison with Dynamic Transitive Closure.** It is tempting to compare edge reduction  $n \circ \Delta$  with dynamic transitive closure [4; 5]: edge composition computes a new edge that captures the transitive effect and this is done repeatedly by  $mlc$ . However, the analogy stops at this abstract level. Apart from the reasons mentioned in Section 4.4, the following differences make it difficult to model edge reduction in terms of dynamic transitive closure.

- Edge reduction does not compute unrestricted transitive effects. Dynamic transitive closure computes unrestricted transitive effects.
- We do not perform closure. Either the final set computed by  $mlc$  is retained in  $\Delta$  or  $n$  is retained in  $\Delta$ . Dynamic transitive closure implies retaining all edges including the edges computed in the intermediate steps.

## 5.2. Constructing GPGs $\Delta(u, v)$

For simplicity, we consider  $\Delta$  only as a collection of edges, leaving the nodes implicit. Further, the edge ordering does not matter at the intraprocedural level and hence we treat  $\Delta$  as a set of edges. The construction of  $\Delta$  assigns sequence numbers in the order of inclusion of edges; these sequence numbers are maintained externally and are used during GPG composition (Section 6).

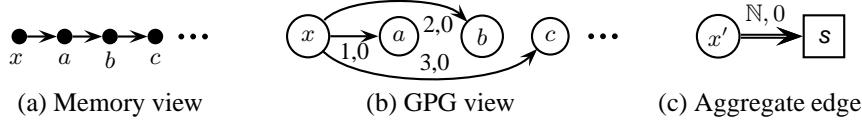


Fig. 11. Aggregate edge for handling strong and weak updates. For this example,  $s = \{a, b, c, \dots\}$ .

By default, the GPGs record the *may* information but a simple extension in the form of *boundary definitions* (described in the later part of this section) allows them to record the *must* information. This supports distinguishing between strong and weak updates and yet allows a simple set union to combine the information.

**Assumption**:  $n$  is  $\delta(t, v)$  and  $\Delta$  is a set of edges

$$\Delta(u, v) := B(u, v) \cup \bigcup_{\substack{t \in \text{succ}^+(u) \\ v \in \text{succ}(t)}} (\Delta(u, t)) [n \circ \Delta(u, t)]$$

$$B(u, v) := \begin{cases} n & v \in \text{succ}(u) \\ \emptyset & \text{otherwise} \end{cases}$$

where

$$\Delta[X] := (\Delta - \text{conskill}(X, \Delta)) \cup (X)$$

$$\text{conskill}(X, \Delta) := \{e_1 \mid e_1 \in \text{match}(e, \Delta), e \in X, |\text{def}(X)| = 1\}$$

$$\text{match}(e, \Delta) := \{e_1 \mid e_1 \in \Delta, S_e = S_{e_1}, S_e^c = S_{e_1}^c\}$$

$$\text{def}(X) := \{(S_e, S_e^c) \mid e \in X\}$$

Definition 5 is an adaptation of Definition 1 for GPGs. Since  $\Delta$  is viewed as a set of edges, the identity function  $\Delta_{id}$  is  $\emptyset$ , meet operation is  $\cup$ , and  $\Delta(u, v)$  is the least fixed point of the equation in Definition 5. The composition of a statement-level flow function ( $n$ ) with a summary flow function ( $\Delta(u, t)$ ) is performed by GPG update which includes all edges computed by edge reduction  $n \circ \Delta(u, t)$ ; the edges to be removed are under-approximated when a strong update cannot be performed (described in the rest of the section). When a strong update is performed, we exclude those edges of  $\Delta$  whose source and *indlev* match that of the shared source of the reduced edges (identified by *match*( $e, \Delta$ )). For a weak update,  $\text{conskill}(X, \Delta) = \emptyset$  and  $X$  contains reduced edges. For an *inconclusive* edge composition,  $\text{conskill}(X, \Delta) = \emptyset$  and  $X = \{n\}$ .

*Computing Edge Order for GPG Edges to Facilitate Flow-Sensitivity.* GPGs represent flow-sensitive memory transformers when their edges are viewed as a sequence matching the control flow order of statements represented by the GPGs. Hence we impose an ordering on GPG edges and maintain it externally explicating it whenever required. This allows us to treat GPGs as set of edges by default in all computations and bring in the ordering only when required.

The ordering of GPG edges for a procedure is governed by a reverse post order traversal over the control flow graph of the procedure. It is required only when the effect of a callee is incorporated in its caller because the control flow of the callee is not available. Since a sequence is totally ordered but control flow is partially ordered, the GPG operations (Section 6) internally relax the total order to ensure that the edges appearing on different control flow paths do not affect each other.

Let  $E$ ,  $Stmt$ , and  $order$  denote the set of edges in a GPG, set of statements, and a set of positive integers representing order numbers. Then, the edge order is maintained as a map  $Stmt \rightarrow 2^E \rightarrow order$ . A particular statement may cause inclusion of multiple edges in a GPG and all edges resulting from the same (non-call) statement should be assigned the same order number. We also maintain reverse map  $order \rightarrow 2^E$  for convenience.

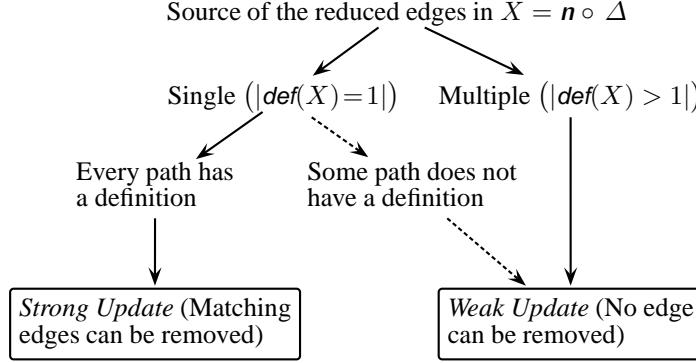


Fig. 12. Criteria for strong and weak updates in  $\Delta$ . Our formulations eliminate the dashed edge simplifying strong updates.

### 5.3. Extending $\Delta$ to Support Strong Updates.

Conventionally, points-to information is killed based on the following criteria: An assignment  $x = \dots$  removes all points-to facts  $x \rightarrow \cdot$  whereas an assignment  $*x = \dots$  removes all points-to facts  $y \rightarrow \cdot$  where  $x$  *must*-points-to  $y$ ; the latter represents a *strong update*. When  $x$  *may*-points-to  $y$ , no points-to facts can be removed representing a *weak update*.

Observe that the use of points-to information for strong updates is inherently captured by edge reduction. In particular, the use of edge reduction allows us to model the edge removal for both  $x = \dots$  and  $*x = \dots$  statements uniformly as follows: the reduced edges should define the same pointer (or the same pointee of a given pointer) along every control flow path reaching the statement represented by  $n$ . This is captured by the requirement  $|\text{def}(X)| = 1$  in *conskill* in Definition 5 where  $\text{def}(X)$  extracts the source nodes and their indirection levels of the edges in  $X$ .

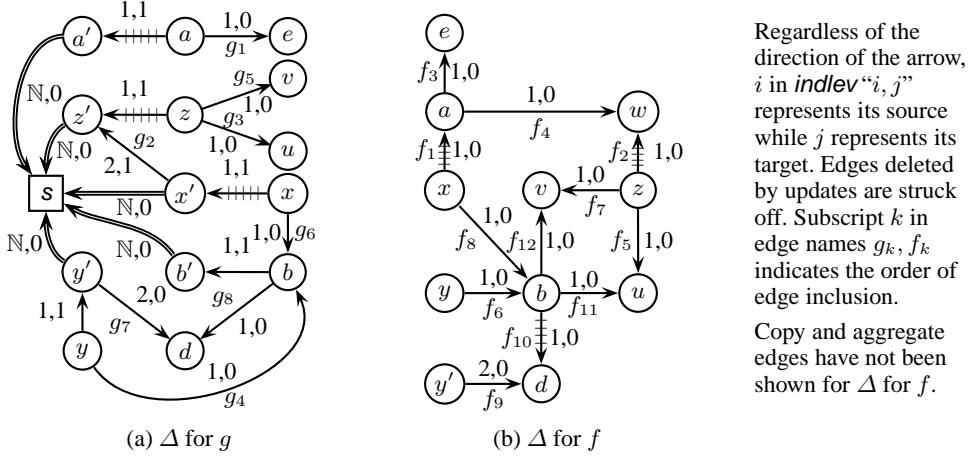
When  $|\text{def}(X)| > 1$ , the reduced edges define multiple pointers (or different pointees of the same pointer) leading to a weak update resulting in no removal of edges from  $\Delta$ . When  $|\text{def}(X)| = 1$ , all reduced edges define the same pointer (or the same pointee of a given pointer). However, this is necessary but not sufficient for a strong update because the pointer may not be defined along all the paths—there may be a path which does not contribute to  $\text{def}(X)$ . We refer to such paths as definition-free paths for that particular pointer (or some pointee of a pointer). The possibility of such a path makes it difficult to distinguish between strong and weak updates as illustrated in Figure 12.

Since a pointer  $x$  or its transitive pointees may be defined along some but not all control flow paths from  $u$  to  $v$ , we eliminate the possibility of definition-free paths from  $u$  to  $v$  by introducing *boundary definitions* of the following two kinds at  $u$ : (a) a pointer assignment  $x = x'$  where  $x'$  is a symbolic representation of the initial value of  $x$  at  $u$  (called the *upwards exposed* [15] version of  $x$ ), and (b) a set of assignments representing the relation between  $x'$  and its transitive pointees.

These boundary definitions are represented by special GPG edges—the first, by a *copy edge*  $x \xrightarrow{1,1} x'$  and the others, by an *aggregate edge*  $x' \xrightarrow{\mathbb{N},0} s$  where  $\mathbb{N}$  is the set of all possible *indevs* and  $s$  is the summary node representing all possible pointees. As illustrated in Figure 11,  $x' \xrightarrow{\mathbb{N},0} s$  is a collection of GPG edges (Figure 11(b)) representing the relation between  $x$  with its transitive pointees at  $u$  (Figure 11(a)).

A reduced edge  $x \xrightarrow{1,j} y$  along any path from  $u$  to  $v$  removes the copy edge  $x \xrightarrow{1,1} x'$  indicating that  $x$  is redefined. A reduced edge  $x \xrightarrow{i,j} y$ ,  $i > 1$  modifies the aggregate edge  $x' \xrightarrow{\mathbb{N},0} s$  to  $x' \xrightarrow{(\mathbb{N}-\{i\}),0} s$  indicating that  $(i-1)^{th}$  pointees of  $x$  are redefined.

The inclusion of aggregate and copy edges guarantees that  $|\text{def}(X)| = 1$  only when the source is defined along every path thereby eliminating the dashed path in Figure 12. This leads to a necessary

Fig. 13.  $\Delta$  for procedures  $f$  and  $g$  of Figure 2.

and sufficient condition for strong updates. Note that the copy and aggregate edges improve the precision of analysis by enabling strong updates and are not required for its soundness.

**EXAMPLE 5.4.** Consider the construction of  $\Delta_g$  as illustrated in Figure 13(c). Edge  $g_1$  created for line 8 of the program, kills edge  $a \xrightarrow{1,1} a'$  because  $|\text{def}\{g_1\}| = 1$ . For line 10, since the pointees of  $x$  and  $z$  are not available in  $g$ , edge  $g_2$  is created from  $x'$  to  $z'$ ; this involves composition of  $x \xrightarrow{2,1} z$  with the edges  $x \xrightarrow{1,1} x'$  and  $z \xrightarrow{1,1} z'$ . Edges  $g_3, g_4, g_5$  and  $g_6$  correspond to lines 11, 13, 14, and 16 respectively.

Edge  $z \xrightarrow{1,1} z'$  is killed along both paths (lines 11 and 14) and hence is struck off in  $\Delta_g$ , indicating  $z$  is *must*-defined. On the other hand,  $y \xrightarrow{1,1} y'$  is killed only along one of the two paths and hence is retained by the control flow merge just before line 16. Similarly  $x' \xrightarrow{2,0} s$  in the aggregate edge  $x' \xrightarrow{\mathbb{N},0} s$  is retained indicating that pointee of  $x$  is not defined along all paths. Edge  $g_6$  kills  $x \xrightarrow{1,1} x'$ . Line 17 creates edges  $g_7$  and  $g_8$ ; this is a weak update because  $y$  has multiple pointees ( $|\text{def}\{g_7, g_8\}| \neq 1$ ). Hence  $b \xrightarrow{1,1} b'$  is not removed. Similarly,  $y' \xrightarrow{2,0} s$  in the aggregate edge  $y' \xrightarrow{\mathbb{N},0} s$  is not removed.  $\square$

## 6. CONSTRUCTING GPGS AT THE INTERPROCEDURAL LEVEL

We have discussed the construction of intraprocedural GPGs in Section 5. We now extend GPG construction to Level 2 of our language which includes handling function calls and recursion.

### 6.1. Handling Function Calls

Definition 6 shows the construction of GPGs at the interprocedural level by handling procedure calls. Consider a procedure  $f$  containing a call to  $g$  between two consecutive program points  $u$  and  $v$ . Let  $\text{Start}_g$  and  $\text{End}_g$  denote the start and the end points of  $g$ .  $\Delta$  representing the control flow paths from  $\text{Start}_f$  to  $u$  (i.e., just before the call to  $g$ ) is  $\Delta(\text{Start}_f, u)$ ; we denote it by  $\Delta_f$  for brevity.  $\Delta$  for the body of procedure  $g$  is  $\Delta(\text{Start}_g, \text{End}_g)$ ; we denote it by  $\Delta_g$ . Then  $\Delta(\text{Start}_f, v)$  using  $\Delta_f$  and  $\Delta_g$  is computed as follows:

- Edges for actual-to-formal-parameter mapping are added to  $\Delta_f$ .
- $\Delta_f$  and  $\Delta_g$  are composed denoted  $\Delta_f \circ \Delta_g$ .

- An edge is created between the return variable of  $g$  and the receiver variable of the call in  $f$  and is added to  $\Delta_f$ .

The composition of a callee's GPG with the caller's GPG can be viewed as incorporating the effect of inlining the callee in the body of the caller. This intuition suggests the following steps for the composition of GPGs: we select an edge  $e$  from  $\Delta_g$  and perform an update  $\Delta_f[e \circ \Delta_f]$ . We then update the resulting  $\Delta$  with the next edge from  $\Delta_g$ . This is repeated until all edges of  $\Delta_g$  are exhausted. The update of  $\Delta_f$  with an edge  $e$  from  $\Delta_g$  involves the following:

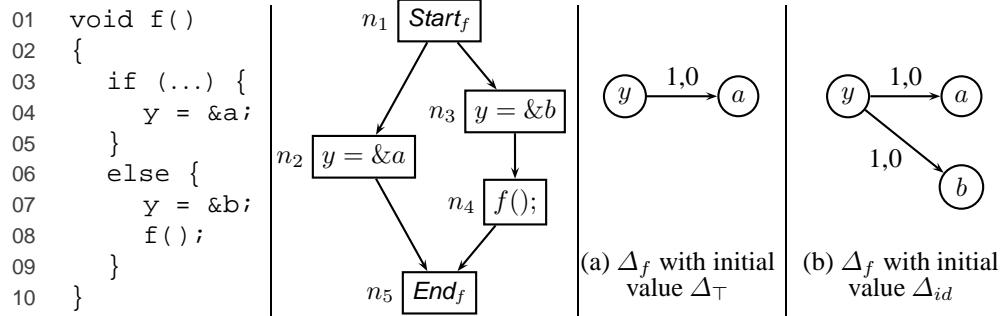
- Substituting the callee's upwards exposed variable  $x'$  occurring in  $\Delta_g$  by the caller's original variable  $x$  in  $\Delta_f$ ,
- Including the reduced edges  $e \circ \Delta_f$ , and
- Performing a strong or weak update.

| <b>Definition 6: <math>\Delta</math> for a call <math>g()</math> in procedure <math>f</math></b>  |
|---|
| $\text{/* let } \Delta_f \text{ denote } \Delta(\text{Start}_f, u) \text{ and } \Delta_g \text{ denote } \Delta(\text{Start}_g, \text{End}_g) */$   |
| $\Delta(\text{Start}_f, v) := \Delta_g \circ \Delta_f$  |
| $\Delta_g \circ \Delta_f := \Delta_f[\Delta_g]$   |
| <b>where</b> $\text{/* let } \Delta_g \text{ be } \{e_1, e_2, \dots, e_k\} */$  |
| $\Delta_f[\Delta_g] := \Delta_f[e_1, \Delta_g][e_2, \Delta_g] \dots [e_k, \Delta_g]$  |
| $\Delta_f[e, \Delta_g] := (\Delta_f - \text{callkill}(e, \Delta_f, \Delta_g)) \cup (e \circ \Delta_f)$  |
| $\text{callkill}(e, \Delta_f, \Delta_g) := \{e_2 \mid e_2 \in \text{match}(e_1, \Delta_f), e_1 \in e \circ \Delta_f, \text{callsup}(e, \Delta_f, \Delta_g)\}$   |
| $\text{callsup}(e, \Delta_f, \Delta_g) := ( \text{def}(e \circ \Delta_f)  = 1) \wedge \text{mustdef}(e, \Delta_g)$  |
| $\text{mustdef}(x \xrightarrow{i,j} y, \Delta) \Leftrightarrow (x \xrightarrow{i,k} z \in \Delta \Rightarrow k = j \wedge z = y) \wedge ((i > 1 \wedge x' \xrightarrow{i,0} s \notin \Delta) \vee (i = 1 \wedge x \xrightarrow{1,1} x' \notin \Delta))$ |

A strong update for summary flow function composition  $\Delta_f \circ \Delta_g$  i.e., when a call is processed, is identified by function  $\text{callsup}$  (Definition 6). Observe that a copy edge  $x \xrightarrow{1,1} x' \in \Delta$  implies that  $x$  has not been defined along some path. Similarly, an aggregate edge  $x' \xrightarrow{\mathbb{N},0} s \in \Delta$  implies that some  $(i-1)^{th}$  pointees of  $x$ ,  $i > 1$  have not been defined along some path. We use these to define  $\text{mustdef}(x \xrightarrow{i,j} y, \Delta)$  which asserts that the  $(i-1)^{th}$  pointees of  $x$ ,  $i > 1$  are defined along every control flow path. We combine it with  $\text{def}(x \xrightarrow{i,j} y \circ \Delta)$  to define  $\text{callsup}$  for identifying strong updates to be performed for a call. Note that we need  $\text{mustdef}$  only at the interprocedural level and not at the intraprocedural level. This is because, when we use  $\Delta_g$  to incorporate its effect in  $\Delta_f$ , performing a strong update requires knowing whether the source of an edge in  $\Delta_g$  has been defined along every control flow path in  $g$ . However, we do not have the control flow information of  $g$  when we incorporate its effect in  $\Delta_f$ . When a strong update is performed, we delete all edges in  $\Delta_f$  that match  $e \circ \Delta_f$ . These edges are discovered by taking a union of  $\text{match}(e_1, \Delta_f)$ , for all  $e_1 \in (e \circ \Delta_f)$ .

The total order imposed by the sequence of GPG edges is interpreted as a partial order as follows: If an edge to be added involves an upwards exposed variable  $x'$ , it should be composed with an original edge<sup>5</sup> in  $\Delta_f$  rather than a reduced edge included in  $\Delta_f$  created by  $e_1 \circ \Delta_f$  for some  $e_1 \in \Delta_g$ . Further, it is possible that an edge  $e_2$  may kill an edge  $e_1$  that was added to  $\Delta_f$  because it coexisted with  $e_2$  in  $\Delta_g$ . However, this should be prohibited because their coexistence in  $\Delta_g$  indi-

<sup>5</sup>By an original edge in  $\Delta_f$ , we mean an edge included in  $\Delta_f$  before processing the call to  $g$ . This edge could well be an edge because of a call in  $f$  processed before processing the current call.

Fig. 14. A recursive example demonstrating the need for  $\Delta_{\top}$ .

cates that they are *may* edges. This is ensured by checking the presence of multiple edges with the same source in  $\Delta_g$ . For example, edge  $f_7$  of Figure 13(d) does not kill  $f_5$  as they coexist in  $\Delta_g$ .

EXAMPLE 6.1. Consider the construction of  $\Delta_f$  as illustrated in Figure 13(d). Edges  $f_1$  and  $f_2$  correspond to lines 2 and 3. The call on line 4 causes the composition of  $\Delta_f = \{f_1, f_2\}$  with  $\Delta_g$  selecting edges in the order  $g_1, g_2, \dots, g_8$ . The edges from  $\Delta_g$  with their corresponding names in  $\Delta_f$  (denoted name-in- $g$ /name-in- $f$ ) are:  $g_1/f_3, g_3/f_5, g_4/f_6, g_5/f_7, g_6/f_8, g_7/f_9$ , and  $g_8/f_{10}$ . Edge  $f_4$  is created by SS and TS compositions of  $g_2$  with  $f_1$  and  $f_2$ . Although  $x$  has a single pointee (along edge  $f_1$ ), the resulting update is a weak update because the source of  $g_2$  is *may*-defined indicated by the presence of  $x' \xrightarrow{2,0} s$  in the aggregate edge  $x' \xrightarrow{\mathbb{N},0} s$ .

Edges  $g_3/f_5$  and  $g_5/f_7$  together kill  $f_2$ . Note that the inclusion of  $f_7$  does not kill  $f_5$  because they both are from  $\Delta_g$ . Finally, the edge for line 5 ( $x \xrightarrow{2,1} z$ ) undergoes an SS composition (with  $f_8$ ) and TS compositions (with  $f_5$  and  $f_7$ ). This creates edges  $f_{11}$  and  $f_{12}$ . Since  $x \xrightarrow{2,1} z$  is accompanied by the aggregate edge  $x' \xrightarrow{\mathbb{N}-\{2\},0} s$  indicating that the pointee of  $x$  is *must*-defined, and  $x$  has a single pointee (edge  $f_8$ ), this is a strong update killing edge  $f_{10}$ . Observe that all edges in  $\Delta_f$  represent classical points-to facts except  $f_9$ . We need the pointees of  $y$  from the callers of  $f$  to reduce  $f_9$ .  $\square$

## 6.2. Handling Recursion

The summary flow function  $\Delta$  of a procedure is complete only when it incorporates the effect of all its callees. Hence  $\Delta$  of callee procedures are constructed first to incorporate its effect in their callers resulting in a postorder traversal over the call graph. However, in case of recursion,  $\Delta$  of a callee procedure may not have been constructed yet because of the presence of a cycle in the call graph. This requires us to begin with an approximate version of  $\Delta$  which is then refined to incorporate the effect of recursive calls. When the callee's  $\Delta$  is computed, its call statements will have to be reprocessed needing a fixed point computation. This is handled in the usual manner [15; 29] by over-approximating initial  $\Delta$  that computes  $\top$  for *may* points-to analysis (which is  $\emptyset$ ). Using any other function would be sound but imprecise. Such a GPG, denoted  $\Delta_{\top}$ , kills all points-to relations and generates none. Clearly,  $\Delta_{\top}$  is not expressible as a GPG and is not a natural  $\top$  element of the meet semi-lattice [15] of GPGs. It has the following properties related to the meet and composition:

- *Meet Operation.* Since we wish to retain the the meet operation  $\sqcap$  as  $\cup$ , we extend  $\cup$  to define  $\Delta \cup \Delta_{\top} = \Delta$  for any GPG  $\Delta$ . This property is also satisfied by a GPG  $\Delta = \emptyset$  denoted  $\Delta_{id}$ , however, it is an identify flow function and not a function computing  $\top$  because it does not kill points-to information.
- *Composition.* Since  $\Delta_{\top}$  is a constant function returning  $\top$  value of the lattice of *may* points-to analysis, it follows that  $\forall \Delta, \Delta \circ \Delta_{\top} = \Delta_{\top}$ . Similarly,

$$\forall X, \forall \Delta, \Delta \circ \Delta_{\top}(X) = \Delta(\Delta_{\top}(X)) = \Delta(\top) = \emptyset[\Delta]$$

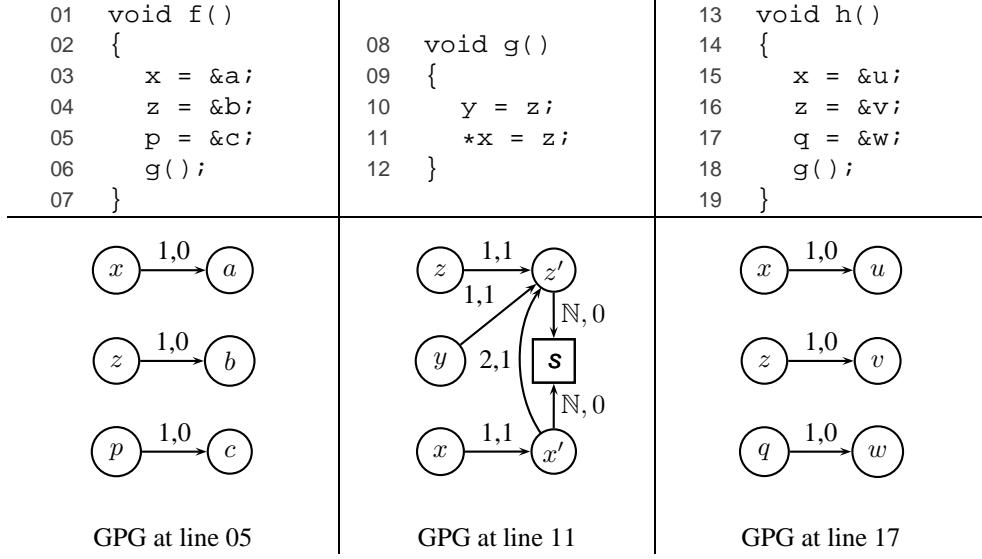


Fig. 15. An example demonstrating the bypassing performed.

which implies that  $\Delta \circ \Delta_{\top} = \Delta$ . This is because  $\top$  for *may* points-to analysis is  $\emptyset$  and empty memory updated with  $\Delta$  returns  $\Delta$ . Although,  $\Delta \circ \Delta_{\top} = \Delta$ , it is only an intermediate function because the fixed point computation induced by recursion will eventually replace  $\Delta_{\top}$  by an appropriate summary flow function.

**EXAMPLE 6.2.** Consider the example of Figure 14 to understand the difference between using  $\Delta_{id}$  and  $\Delta_{\top}$  as the initial value. If we use the initial  $\Delta$  for procedure  $f$  at  $n_4$  as  $\Delta_{id}$ , a GPG with no edges, then the  $\Delta$  at the *Out* of  $n_4$  has a GPG with one edge  $y \xrightarrow{1,0} b$ . Thus, the summary flow function of procedure  $f$  ( $\Delta_f$ ) computed at  $n_5$  after the meet is as shown in Figure 14(b). After reprocessing the call at  $n_4$ , we still get the same GPG. However, if we consider  $\Delta_{\top}$  as the initial value for procedure  $f$ , the GPG at *Out* of  $n_4$  is an empty GPG as  $\Delta_{\top}$  kills all points-to relations and generates none. Thus,  $\Delta_f$  at  $n_5$  is as shown in Figure 14(a) which remains the same even after re-processing. The resulting summary flow function is more precise than the summary flow function computed using  $\Delta_{id}$  as the initial value because it excludes  $y \xrightarrow{1,0} b$  from the GPG of procedure  $f$ . It is easy to see that after a call to  $f$  ends,  $y$  cannot point to  $b$ ; it must point to  $a$ .  $\square$

## 7. COMPUTING POINTS-TO INFORMATION USING GPGS

Recall that the points-to information is represented by a memory  $M$ . We define two operations to compute a new memory  $M'$  using a GPG or a GPG edge from a given memory  $M$ .

- An *edge application*  $\llbracket e \rrbracket M$  computes memory  $M'$  by incorporating the effect of GPG edge  $e \equiv x \xrightarrow{i,j} y$  in memory  $M$ . This involves inclusion of edges described by the set  $\{w \xrightarrow{1,0} z \mid w \in M^{i-1}\{x\}, z \in M^j\{y\}\}$  in  $M'$  and removal of edges by distinguishing between a strong and a weak update. The edges to be removed are characterized much along the lines of *call/kill* (Definition 6).
- A *GPG application*  $\llbracket \Delta \rrbracket M$  applies the GPG  $\Delta$  to  $M$  and computes the resulting memory  $M'$  using edge application iteratively.

We now describe the computation of points-to information using these two operations. Let  $PT_v$  denote the points-to information at program point  $v$  in procedure  $f$ . Then,  $PT_v$  can be computed by (a) computing *boundary information* of  $f$  (denoted  $Bl_f$ ) associated with the program point  $Start_f$ , and (b) computing the points-to information at  $v$  from  $Bl_f$  by incorporating the effect of all paths from  $Start_f$  to  $v$ .

$Bl_f$  is computed as the union of the points-to information reaching  $f$  from all of its call points. For the main function,  $Bl$  is computed from static initializations. In the presence of recursion, a fixed point computation is required for computing  $Bl$ .

EXAMPLE 7.1. For the program in Figure 15, the  $Bl$  of procedure  $g$  (denoted  $Bl_g$ ) is the points-to information reaching  $g$  from its callers  $f$  and  $h$  and hence a union of GPG at the  $Out$  of line numbers 05 and 17. Let  $\Delta_{10}$  represent the GPG that includes the effect of line 10. Then the points-to information after line number 10 is  $(\Delta_{10} \circ Bl_g)$  as discussed in Section 6. Similarly, the points-to information at line number 11 can be computed by  $(\Delta_{11} \circ Bl_g)$ .  $\square$

If  $v$  is  $Start_f$ , then  $PT_v = Bl_f$ . For other program points,  $PT_v$  can be computed from  $Bl_f$  in the following ways; both of them compute identical  $PT_v$ .

- (a) *Using statement-level flow function (Stmt-ff):* Let  $stmt(u, v)$  denote the statement between  $u$  and  $v$ . If it is a non-call statement, let its flow function  $\delta(u, v)$  be represented by the GPG edge  $n$ . Then  $PT_v$  is computed as the least fixed point of the following data flow equations where  $In_{u,v}$  denotes the points-to information reaching program point  $u$  from its predecessor  $v$ .

$$In_{u,v} = \begin{cases} \llbracket \Delta(Start_q, End_q) \rrbracket PT_u & stmt(u, v) = call\ q \\ \llbracket n \rrbracket PT_u & otherwise \end{cases}$$

$$PT_u = \bigcup_{u \in pred(v)} In_{u,v}$$

- (b) *Using GPGs:*  $PT_v$  is computed using GPG application  $\llbracket \Delta(Start_f, v) \rrbracket Bl_f$ . This approach of  $PT_v$  computation is oblivious to intraprocedural control flow and does not involve fixed point computation for loops because  $\Delta(Start_f, v)$  incorporates the effect of loops.

Our measurements show that the *Stmt-ff* approach takes much less time than using GPGs for  $PT_v$  computation. This may appear surprising because the *Stmt-ff* approach requires an additional fixed point computation for handling loops which is not required in case of GPGs. However, using GPGs for all statements including the non-call statements requires more time because the GPG at  $v$  represents a cumulative effect of the statement-level flow functions from  $Start_f$  to  $v$ . Hence the GPGs tend to become larger with the length of a control flow path. Thus computing  $PT_v$  using GPGs for multiple consecutive statements involves redundant computations.

EXAMPLE 7.2. In our example in Figure 15,  $\Delta_{10}$  has only one edge  $y \xrightarrow{1,1} z'$  (ignoring the aggregate and copy edges) whereas  $\Delta_{11}$  consists of two edges  $y \xrightarrow{1,1} z'$  and  $x' \xrightarrow{1,2} z'$  incorporating the effect of all control flow paths from start of procedure  $g$  to line number 11 which also includes the effect of line number 10.

As an alternative, we can compute points-to information using statement level flow functions using the points-to information computed for the  $In$  of the statement (instead of  $Bl$ ) thereby avoiding redundant computations. Thus at line number 10, we have  $y \xrightarrow{1,1} z$  and at line number 11 we have only  $x \xrightarrow{2,1} z$ . For a call statement, we can use the GPG representing the summary flow function of the callee instead of propagating the values through the body of the callee. This reduces the computation of points-to information to an intraprocedural analysis.  $\square$

### Bypassing of $Bl$

Our measurements show that using the entire  $Bl$  of a procedure may be expensive because many points-to pairs reaching a call may not be accessed by the callee procedure. Thus the efficiency of computing points-to information can be enhanced significantly by filtering out the points-to information which is irrelevant to a procedure but merely passes through it unchanged. This concept of *bypassing* has been successfully used for data flow values of scalars [22; 23]. GPGs support this naturally for pointers with the help of upwards exposed versions of variables. An upwards exposed version of a variable in a GPG indicates that there is a use of the variable in the procedure requiring pointee information from the callers. Thus, the points-to information of such a variable is relevant and should be a part of  $Bl$ . For variables that do not have their corresponding upwards exposed versions in a GPG, their points-to information is irrelevant and can be discarded from the  $Bl$  of the procedure, effectively bypassing the calls to the procedure.

**EXAMPLE 7.3.** In our example of Figure 15, the GPG at the *Out* of line number 11 (which represents the summary flow function of procedure  $g$ ) contains upwards exposed versions of variables  $x$  and  $z$  indicating that some pointees of  $x$  and  $z$  from the calling context are accessed in the procedure  $g$ . Since the *indlev* of  $x'$  is 2 which is the source of one of the GPG edge, its pointee is being defined by  $g$ . Thus, pointee of  $x$  needs to be propagated to the procedure  $g$ . Similarly, the *indlev* of  $z'$  is 1 which is the target of an GPG edge specifying that pointee of  $z$  is being assigned to some pointer in procedure  $g$ . Thus, pointees of  $x$  and  $z$  are accessed in procedure  $g$  but are defined in the calling context and hence should be part of the  $Bl$  of procedure  $g$ . Note that points-to information of  $p$  or  $q$  is neither accessed nor defined by procedure  $g$  and hence can be bypassed. Thus,  $Bl_g$  is not the union of GPGs at the *Out* of line numbers 05 and 17. It excludes edges such as  $p \xrightarrow{1,0} c$  and  $q \xrightarrow{1,0} w$  as they are irrelevant to procedure  $g$  and hence are bypassed.  $\square$

## 8. SEMANTICS AND SOUNDNESS OF GPGS

We prove the soundness of points-to analysis using GPGs by establishing the semantics of GPGs in terms of their effect on memory and then showing that GPGs compute a conservative approximation of the memory. For this purpose, we make a distinction between a *concrete memory* at a program point computed along a single control flow path and an *abstract memory* computed along all control flow paths reaching the program point. The soundness of the abstract memory computed using GPGs is shown by arguing that it is an over-approximation of concrete memories.

The memory that we have used so far is an abstract memory. This section defines concrete memory and also the semantics of both concrete and abstract memories.

### 8.1. Notations for Concrete and Abstract Memory

We have already defined a *control flow path*  $\pi$  as a sequence of (possibly repeating) program points  $q_0, q_1, \dots, q_m$ . When we talk about a particular control flow path  $\pi$ , we use  $psucc$  and  $ppred$  to denote successors and predecessors of a program point along  $\pi$ . Thus,  $q_{i+1} = psucc(\pi, q_i)$  and  $q_i = ppred(\pi, q_{i+1})$ ;  $psucc^*$ ,  $ppred^*$  denote their reflexive transitive closures. In presence of cycles, program points could repeat; however, we do not explicate their distinct occurrences for notational convenience; the context is sufficient to make the distinction.

The *concrete memory* at a program point along a control flow path  $\pi$  is an association between pointers and the locations whose addresses they hold and is represented by a function  $M : P \rightarrow (L \cup \{?\})$ . For static analysis, when the effects of multiple control flow paths reaching a program point are incorporated in the memory, the resulting memory is a relation  $M \subseteq P \times (L \cup \{?\})$  as we have already seen in Section 3.1. We call it an *abstract memory* because it is an over-approximation of the union of concrete memories along all paths reaching the program point.

When concrete and abstract memories need to be distinguished, we denote them by  $\overline{M}$  and  $M$ , respectively.  $\overline{M}_{u,\pi}$  denotes the concrete memory associated with a particular occurrence of  $u$  in a given  $\pi$  whereas  $M_u$  denotes the memory associated with all occurrences of  $u$  in all possible  $\pi$ s.

Definition 7 provides an equation to construct a concrete summary flow function  $\overline{\Delta}(\pi, u, v)$  by composing the flow functions  $\delta$  of the statements appearing on a control flow path  $\pi$  from  $u$  to  $v$ .

| <b>Definition 7: Memory Transformer <math>\overline{\Delta}</math></b>  |
|---|
| $\overline{\Delta}(\pi, u, v) := B(\pi, u, v) \sqcap \delta(t, v) \circ \overline{\Delta}(\pi, u, t)$                                       |
| $B(\pi, u, v) := \begin{cases} \overline{\Delta}_{id} & v = u \\ \delta(u, v) & v \in psucc(u) \\ \emptyset & \text{otherwise} \end{cases}$ |

The summary flow function  $\overline{\Delta}(\pi, u, v)$  is used to compute  $\overline{M}_{v,\pi}$  as follows:

$$\overline{M}_{v,\pi} = [\overline{\Delta}(u, v)] (\overline{M}_{u,\pi})$$

**8.1.1. Difference between  $\overline{M}$  and  $M$ : An Overview.** The operations listed in Figure 6 were defined for abstract memory. An overview of how they differ for the two memories is as follows:

- *Edge composition  $n \circ p$ .* This is same for both memories.
- *Edge reduction.* For a concrete GPG  $\overline{\Delta}$ , the reduction  $n \circ \overline{\Delta}$  creates a single edge whereas for an abstract GPG  $\Delta$ , the reduction  $n \circ \Delta$  could create multiple edges because  $\Delta(u, v)$  needs to cover all paths from  $u$  to  $v$  unlike  $\overline{\Delta}(\pi, u, v)$  which covers only a single control flow path  $\pi$  from  $u$  to  $v$ .
- *GPG application.* A concrete memory  $\overline{M}$  is a function and the update  $\llbracket e \rrbracket \overline{M}$  reorients the out edge of the source of  $e$ . An abstract memory  $M$  is a relation and the source of  $e$  may have multiple edges. This may require under-approximating deletion.
- *GPG update.* Like GPG application,  $\overline{\Delta}$  update is exact whereas  $\Delta$  update may have to be approximated.

$\Delta(u, v)$  should be an over-approximation of  $\overline{\Delta}(\pi, u, v)$  for every path  $\pi$  from  $u$  to  $v$ . Hence, the inclusion of pointees of a pointer is over-approximated while their removal is under-approximated; the latter requires distinguishing between strong and weak updates.

## 8.2. Computing Points-to GPGs $\overline{\Delta}(\pi, u, v)$ for a Single Control Flow Path

This section formalizes the concept of GPG for points-to analysis over concrete memory  $\overline{M}$  created by a program along a single execution path.

In the base case, the program points  $u$  and  $v$  are consecutive and  $\overline{\Delta}(\pi, u, v)$  is  $\delta(t, v)$ . When they are farther apart on  $\pi$ , consider a program point  $t \in psucc^+(\pi, u) \cap ppred(\pi, v)$ . We define  $\overline{\Delta}(\pi, u, v)$  recursively by extending  $\overline{\Delta}(\pi, u, t)$  to incorporate the effect of  $\delta(t, v)$  for computing the concrete memory  $\overline{M}$  at program point  $v$  from  $\overline{M}$  at  $u$  (Definition 8).

Extending  $\overline{\Delta}(\pi, u, t)$  (denoted  $\overline{\Delta}$  for simplicity) to incorporate the effect of  $\delta(t, v)$  (denoted by the edge  $n$ ) involves two steps:

- Reducing  $n$  by composing it with edges in  $\overline{\Delta}$ . This operation is denoted by  $n \circ \overline{\Delta}$  (i.e. reduce *idle* of  $n$  using points-to information in  $\overline{\Delta}$ ). This is explained in Definition 4 which is applicable to both  $\Delta$  and  $\overline{\Delta}$  uniformly.
- Updating  $\overline{\Delta}$  with the reduced edge. This operation is denoted by  $\overline{\Delta}[n \circ \overline{\Delta}]$ .

The first step is same for both  $\overline{\Delta}$  and  $\Delta$ . However, the second step differs and is formulated in Definition 8 for  $\overline{\Delta}$ . Unlike Definition 5, the GPG update for  $\overline{\Delta}$  is defined in terms of an update using a single edge. Hence, we define  $\overline{\Delta}[X] = \overline{\Delta}[r]$  where  $X = \{r\}$ . This allows us to define GPG update

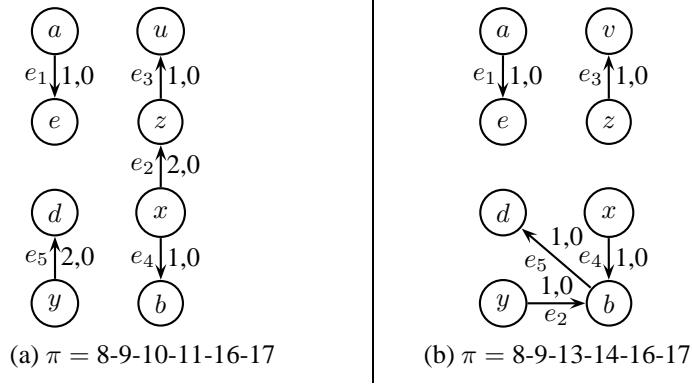


Fig. 16.  $\overline{\Delta}$  for the two control flow paths in procedure  $g$  of Figure 2. The control flow paths are described in terms of line numbers. For  $\text{indlev } mn$ , regardless of the direction of the edge,  $m$  is for the source while  $n$  is for the target. The numbers in the subscripts of edge names (e.g.,  $e_i$ ) indicate the order of their inclusion.

generically in terms of a set of edges  $X$ . Given a reduced edge  $r$ , the update  $\overline{\Delta}[r]$ , reorients the out edge of the source whose  $\text{indlev}$  matches that in  $r$ ; if no such edge exists in  $\overline{\Delta}$ ,  $r$  is added to it. For this purpose, we view  $\overline{\Delta}$  as a mapping  $L \times I \rightarrow L \times I$  and an edge  $x \xrightarrow{i,j} y$  as a pair  $(x, i) \mapsto (y, j)$  in  $\overline{\Delta}$  where  $I$  is the set of integers. Then, the update  $\overline{\Delta}[x \xrightarrow{i,k} z]$  changes the mapping of  $(x, i)$  in  $\overline{\Delta}$  to  $(z, k)$ .

| <i>Definition 8: Construction of <math>\overline{\Delta}</math></i>   | <i>/* <math>n</math> is <math>\delta(t, v)</math> */</i>     |
|---|--|
| $\overline{\Delta}(\pi, u, v) := \begin{cases} \delta(u, v) & v \in \text{psucc}(\pi, u) \\ (\overline{\Delta}(\pi, u, t)) [ n \circ \overline{\Delta}(\pi, u, t) ] & t \in \text{psucc}^+(\pi, u) \cap \text{ppred}(\pi, v) \end{cases}$ |  |
| <b>where</b>  |  |
| $\overline{\Delta}[X] := \overline{\Delta}[r]$  | <i>/* let <math>X</math> be <math>\{r\}</math> */</i>        |
| $\overline{\Delta}[e] := \overline{\Delta}[(x, i) \mapsto (y, j)]$  | <i>/* let <math>e \equiv x \xrightarrow{i,j} y</math> */</i> |

EXAMPLE 8.1. Figure 16 shows the summary flow function along two paths in procedure  $g$  of our motivating example in Figure 2. The edges are numbered in the order of their inclusion.  $\square$

### 8.3. The Semantics of the Application of GPG to Concrete and Abstract Memory

We first define the semantics of  $\overline{\Delta}$  to  $\overline{M}$  and then extend it to the application of  $\Delta$  to  $M$ .

**8.3.1. The Semantics of the Application of  $\overline{\Delta}$  to  $\overline{M}$ .** The initial value of the memory at the start of a control flow path  $\pi$  is  $\overline{M}_0 = \{(x, ?) \mid x \in L\}$ . Since  $\overline{M}_0$  is a total function, any  $\overline{M}$  computed by updating it is also a total function. Hence,  $\overline{M}$  is defined for all variables at all program points. Let  $\overline{M}\{a\} = \{b\}$  implying that  $a$  points-to  $b$  in  $\overline{M}$ . Suppose that, as a consequence of execution of a statement,  $a$  ceases to point to  $b$  and instead points to  $c$ . The memory resulting from this change is denoted by  $\overline{M}[a \mapsto c]$ .

Definition 9 provides the semantics of the application of  $\overline{\Delta}(\pi, u, v)$  to  $\overline{M}_{u, \pi}$  to compute  $\overline{M}_{v, \pi}$  for the control flow path  $\pi$  from  $u$  to  $v$ .

| <b>Definition 9: Semantics of <math>\overline{\Delta}</math></b>   |
|--|
| $\overline{M}_{v, \pi} := \llbracket \overline{\Delta}(\pi, u, v) \rrbracket \overline{M}$   |
| <b>where</b> /* let $\overline{\Delta}$ be $\{e_1, e_2, \dots, e_k\}$ */   |
| $\llbracket \overline{\Delta} \rrbracket \overline{M} := (\llbracket e_k \rrbracket \dots (\llbracket e_2 \rrbracket (\llbracket e_1 \rrbracket \overline{M})) \dots) := \llbracket e_k \rrbracket \dots \llbracket e_2 \rrbracket \llbracket e_1 \rrbracket \overline{M}$ |
| $\text{eval}(x \xrightarrow{i,j} y, \overline{M}) := w \xrightarrow{1,0} z \text{ where } w = \overline{M}^{i-1}\{x\}, z = \overline{M}^j\{y\}$  |
| $\llbracket e \rrbracket \overline{M} := \overline{M}[\text{eval}(e, \overline{M})]$   |

The *application* of a GPG edge  $e \equiv x \xrightarrow{i,j} y$  to memory  $\overline{M}$  denoted  $\llbracket e \rrbracket \overline{M}$ , creates a points-to edge by discovering the locations reached from  $x$  and  $y$  through a series of indirections and updates  $\overline{M}$  by reorienting the existing edges. We define edge application as a two step process:

- *Edge evaluation* denoted  $\text{eval}(e, \overline{M})$  returns a points-to edge by discovering the locations reached indirectly from  $x$  and  $y$  where  $e \equiv x \xrightarrow{i,j} y$ . This operation is similar to edge reduction (Section 5.1) with minor differences such as, the edges used for reduction are from memory  $\overline{M}$  where each edge represents a classical points-to fact and not a generalized points-to fact. The reduced edge  $r$  also represents a points-to fact.
- *Memory update* denoted  $\overline{M}[e]$  re-orientation the existing edges. It is similar to the GPG update  $\overline{\Delta}[e]$ .

Suppose the evaluation  $\text{eval}(x \xrightarrow{i,j} y, \overline{M})$  creates an edge  $w \xrightarrow{1,0} z$  where  $w = \overline{M}^{i-1}\{x\}$  and  $z = \overline{M}^j\{y\}$ , then the memory update  $\overline{M}[x \xrightarrow{i,j} y]$  results in  $\overline{M}[w \mapsto z]$ . Although the two notations  $\overline{M}[x \xrightarrow{i,j} y]$  and  $\overline{M}[w \mapsto z]$  look similar, the arrow  $\rightarrow$  in the first indicates that it is a GPG edge whereas the arrow  $\mapsto$  in the second indicates that a mapping is being changed. Effectively we change  $\overline{M}_{v, \pi}$  such that  $\overline{M}_{v, \pi}^i\{x\} = \overline{M}_{v, \pi}^j\{y\}$ .

EXAMPLE 8.2. For our motivating example, let  $\overline{M}$  before the call to  $g$  be  $\{(a, ?), (b, ?), (x, a), (y, ?), (z, w)\}$ . The resulting memory after applying  $\overline{\Delta}$  of Figure 16(a) is  $\{(a, w), (b, ?), (x, b), (y, ?), (z, u)\}$ . When we apply  $\overline{\Delta}$  of Figure 16(b) representing the other control flow path to the same  $\overline{M}$  before the call to  $g$ , the resulting  $\overline{M}$  is  $\{(a, e), (b, d), (x, b), (y, b), (z, v)\}$ .  $\square$

8.3.2. *The Semantics of the Application of  $\Delta$  to  $M$ .* Definition 10 provides the semantics of  $\Delta(u, v)$  by showing how  $M_v$  is computed from  $M_u$ .

| <b>Definition 10: Semantics of <math>\Delta</math></b>  |
|---|
| $M_v := \llbracket \Delta(u, v) \rrbracket M_u$   |
| <b>where</b> /* let $\Delta$ be $\{e_1, e_2, \dots, e_k\}$ */   |
| $\llbracket \Delta \rrbracket M := (\llbracket e_k \rrbracket, \Delta) \dots (\llbracket e_2 \rrbracket, \Delta)(\llbracket e_1 \rrbracket, \Delta M) \dots)$ |
| $\text{eval}(x \xrightarrow{i,j} y, M) := \{w \xrightarrow{1,0} z \mid w \in M^{i-1}\{x\}, z \in M^j\{y\}\}$  |
| $\llbracket e, \Delta \rrbracket M := \text{eval}(e, M) \cup (M - \text{memkill}(e, M, \Delta))$  |
| $\text{memkill}(e, M, \Delta) := \{e_2 \mid e_2 \in \text{match}(e_1, M), e_1 \in \text{eval}(e, M), \text{memsup}(e, M, \Delta)\}$                           |
| $\text{memsup}(e, M, \Delta) \Leftrightarrow \text{singledef}(e, M) \wedge \text{mustdef}(e, \Delta)$   |
| $\text{singledef}(x \xrightarrow{i,j} y, M) \Leftrightarrow M^{i-1}\{x\} = \{z\} \wedge z \neq ?$   |

We assume that the pair  $(x, ?)$  is included in  $M$  for all variables at the start of the program.  $M\{a\}$  represents the set of pointees of  $a$ . The *application* of a GPG edge  $e \equiv x \xrightarrow{i,j} y$  to memory  $M$  denoted  $\llbracket e, \Delta \rrbracket M$ , first evaluates the edge  $e$  and then updates the memory  $M$  as follows:

- *Edge evaluation* returns a set of edges that are included to compute memory  $M_v$ . These are points-to edges obtained by discovering the locations reached from  $x$  and  $y$  through a series of indirections.
- *Memory update* An update of  $M$  with  $e$  is a strong update when  $e$  defines a single pointer and its source is *must* defined in  $\Delta$  (i.e., it is defined along all paths from  $u$  to  $v$ ).

Unlike edge application to  $\overline{M}$  (Definition 9), edge application to  $M$  requires two arguments (Definition 10). The second argument  $\Delta$  is required to identify that the source of the edge  $e$  is *must* defined which is not required for computing  $\overline{M}$  because  $\overline{M}$  considers only one control flow path  $\pi$  at a time.

The predicate  $\text{singledef}(x \xrightarrow{i,j} y, M)$  asserts that an edge  $x \xrightarrow{i,j} y$  in  $\Delta$  defines a single pointer. Contrast this with  $\text{def}(X)$  in Definition 5 which collects the pointers being defined. Observe that  $\text{singledef}(x \xrightarrow{i,j} y, M)$  trivially holds for  $i = 1$ . We discover that the source of a GPG edge is *must* defined with the provision of edges  $x \xrightarrow{1,1} x'$  and  $x' \xrightarrow{\mathbb{N},0} s$  (Section 5.3).  $\text{singledef}(x \xrightarrow{i,j} y, M)$  and  $\text{mustdef}(x \xrightarrow{i,j} y, \Delta)$  are combined to define  $\text{memsup}(e, M, \Delta)$  which asserts that an edge  $e$  in  $\Delta$  can perform strong update of  $M$ .

When a strong update is performed using  $\text{memsup}(e, M, \Delta)$ , we delete all edges in  $M$  that match  $e$  which is a reduced form of edge  $n$ . These edges are discovered by  $\text{match}(e, M)$ . The edges to be removed ( $\text{memkill}$ ) are characterized much along the lines of  $\text{callkill}$  (Section 6) with a couple of minor differences:

- The edge  $e_1$  now is a result of an evaluation of  $e$  in  $M$  rather than reduction  $e \circ \Delta_f$ , and
- matching edges  $e_2$  are from  $M$  instead of from  $\Delta_f$ .

#### 8.4. Soundness of GPGs

We first show the soundness of  $\overline{\Delta}(\pi, u, v)$  for a path  $\pi$  from  $u$  to  $v$  in terms of memory  $\overline{M}_{v,\pi}$  computed from memory  $\overline{M}_{u,\pi}$ . Then we show the soundness of  $\Delta(u, v)$  by arguing that it is an over-approximation of  $\overline{\Delta}(\pi, u, v)$  for every path  $\pi$  from  $u$  to  $v$ .

| <b>Definition 11: Soundness of <math>\overline{\Delta}</math> and <math>\Delta</math></b>   |
|---|
| <b>Soundness of Concrete Summary Flow Function <math>\overline{\Delta}</math></b> $\text{eval}(n, \llbracket p \rrbracket \overline{M}_{u,\pi}) := \text{eval}(n \circ p, \overline{M}_{u,\pi})$ $\text{eval}(n, \llbracket \overline{\Delta} \rrbracket \overline{M}_{u,\pi}) := \text{eval}(n \circ \overline{\Delta}, \overline{M}_{u,\pi})$ $\overline{M}_{v,\pi} = \llbracket \overline{\Delta}(\pi, u, v) \rrbracket \overline{M}_{u,\pi}$  |
| <b>Soundness of Abstract Summary Flow Function <math>\Delta</math></b> $\text{kill}(\pi, n, \overline{M}_{u,\pi}) := \{ e_1 \mid e_1 \in \text{match}(e, \overline{M}_{u,\pi}), e \in \llbracket n \rrbracket \overline{M}_{u,\pi} \}$ $\text{memkill}(n, M_u, \Delta(u, v)) \subseteq \bigcap_{\pi \in \text{Paths}(u, v)} \text{kill}(\pi, n, \overline{M}_{u,\pi})$ $n \circ \Delta(u, v) \supseteq \bigcup_{\pi \in \text{Paths}(u, v)} n \circ \overline{\Delta}(\pi, u, v)$ $\llbracket \Delta(u, v) \rrbracket M_u \supseteq \bigcup_{\pi \in \text{Paths}(u, v)} \llbracket \overline{\Delta}(\pi, u, v) \rrbracket \overline{M}_{u,\pi}$ |

Definition 11 articulates the formal proof obligations for showing soundness. We assume that we perform *TS* and *SS* compositions only. Further, the *conclusiveness* of edge compositions is checked independently prohibiting *inconclusive* edge compositions.

**8.4.1. Soundness of GPGs for Concrete Memory.** We argue below that  $\overline{\Delta}(\pi, \mathbf{u}, \mathbf{v})$  is sound because the effect of the reduced edge is identical to the effect of the original edge on  $\overline{M}_{\mathbf{u}, \pi}$ ; hence the evaluation of an edge  $\mathbf{n}$  in the resulting memory computed after application of  $\mathbf{p}$  to  $\overline{M}_{\mathbf{u}, \pi}$ , is same as the evaluation of the reduced edge  $\mathbf{n} \circ \mathbf{p}$  in  $\overline{M}_{\mathbf{u}, \pi}$ .

**LEMMA 8.1.** *Consider a memory  $\overline{M}'$  resulting from application of a GPG edge  $\mathbf{p}$  to  $\overline{M}$ . The evaluation of an edge  $\mathbf{n}$  in  $\overline{M}'$  is identical to the evaluation of the reduced edge  $\mathbf{r}$  in  $\overline{M}$  where  $\mathbf{r} = \mathbf{n} \circ \mathbf{p}$ .*

$$\text{eval}(\mathbf{n}, [\mathbf{p}] \overline{M}_{\mathbf{u}, \pi}) = \text{eval}(\mathbf{n} \circ \mathbf{p}, \overline{M}_{\mathbf{u}, \pi}) \quad (8.1.a)$$

**PROOF.** The lemma trivially follows when  $\mathbf{n}$  and  $\mathbf{p}$  do not compose because they have independent effects on  $\overline{M}_{\mathbf{u}, \pi}$  provided the order of execution is followed.

Consider *TS* composition for  $\mathbf{n} \circ \mathbf{p}$ . Let edge  $\mathbf{n} \equiv x \xrightarrow{i,j} y$  and edge  $\mathbf{p} \equiv y \xrightarrow{k,l} z$ . From Section 4.2,  $\mathbf{n} \circ \mathbf{p} = x \xrightarrow{i,(l+j-k)} z$  for a *useful* edge composition.

- For the RHS of (8.1.a), the evaluation of  $\mathbf{n} \circ \mathbf{p}$  in  $\overline{M}_{\mathbf{u}, \pi}$  results in  $[\mathbf{n} \circ \mathbf{p}] \overline{M}_{\mathbf{u}, \pi} = s_1 \xrightarrow{1,0} t_1$  where  $s_1 = \overline{M}_{\mathbf{u}, \pi}^{i-1}\{x\}$  and  $t_1 = \overline{M}_{\mathbf{u}, \pi}^{l+j-k}\{z\}$ . Thus edge  $s_1 \xrightarrow{1,0} t_1$  imposes the constraint

$$\overline{M}_{\mathbf{u}, \pi}^i\{x\} = \overline{M}_{\mathbf{u}, \pi}^{l+j-k}\{z\} \quad (8.1.b)$$

- For the LHS of (8.1.a), the evaluation of edge  $\mathbf{p}$  updates  $\overline{M}_{\mathbf{u}, \pi}$  as follows  $\overline{M}_{\mathbf{u}, \pi}[\mathbf{p}] = \overline{M}_{\mathbf{u}, \pi}[s_2 \mapsto t_2]$  where the pointer  $s_2 = \overline{M}_{\mathbf{u}, \pi}^{k-1}\{y\}$  and the pointee  $t_2 = \overline{M}_{\mathbf{u}, \pi}^l\{z\}$ .  $\overline{M}'$  is defined in terms of  $\overline{M}_{\mathbf{u}, \pi}$  by the following constraint resulting from the inclusion of the edge  $s_2 \xrightarrow{1,0} t_2$ .

$$\overline{M}_{\mathbf{u}, \pi}^k\{y\} = \overline{M}_{\mathbf{u}, \pi}^l\{z\} \quad (8.1.c)$$

The evaluation of  $\mathbf{n}$  in the updated memory  $\overline{M}' = [\mathbf{p}] \overline{M}_{\mathbf{u}, \pi}$  results in  $\text{eval}(\mathbf{n}, \overline{M}') = s_3 \xrightarrow{1,0} t_3$  where  $s_3 = (\overline{M}')^{i-1}\{x\}$  and  $t_3 = (\overline{M}')^j\{y\}$ . Edge  $s_3 \xrightarrow{1,0} t_3$  imposes the following constraint on  $\overline{M}'$ .

$$(\overline{M}')^i\{x\} = (\overline{M}')^j\{y\}$$

In order to map this constraint to  $\overline{M}_{\mathbf{u}, \pi}$ , we need to combine it with constraint (8.1.c), replace  $\overline{M}'$  by  $\overline{M}_{\mathbf{u}, \pi}$  and solve them together.

$$\begin{aligned} \overline{M}_{\mathbf{u}, \pi}^i\{x\} &= \overline{M}_{\mathbf{u}, \pi}^j\{y\} \wedge \overline{M}_{\mathbf{u}, \pi}^k\{y\} = \overline{M}_{\mathbf{u}, \pi}^l\{z\} \\ \Rightarrow \overline{M}_{\mathbf{u}, \pi}^i\{x\} &= \overline{M}_{\mathbf{u}, \pi}^j\{y\} \wedge \overline{M}_{\mathbf{u}, \pi}^{k+(j-k)}\{y\} = \overline{M}_{\mathbf{u}, \pi}^{l+(j-k)}\{z\} \\ \Rightarrow \overline{M}_{\mathbf{u}, \pi}^i\{x\} &= \overline{M}_{\mathbf{u}, \pi}^{l+j-k}\{z\} \end{aligned} \quad (8.1.d)$$

Constraint (8.1.d) is identical to constraint (8.1.b). Since the effect on the memory is identical, the two evaluations are identical.

The equivalence of evaluations for *SS* composition between  $\mathbf{n}$  and  $\mathbf{p}$  can be proved in a similar manner.  $\square$

LEMMA 8.2. Consider a memory  $\bar{M}'$  resulting from application of a GPG  $\bar{\Delta}$  to  $\bar{M}$ . The evaluation of an edge  $n$  in  $\bar{M}'$  is identical to the evaluation of the reduced edges  $n \circ \bar{\Delta}$  in  $\bar{M}$ .

$$\mathbf{eval}(n, [\bar{\Delta}] \bar{M}_{u,\pi}) = \mathbf{eval}(n \circ \bar{\Delta}, \bar{M}_{u,\pi})$$

PROOF. Let  $\bar{\Delta}_m$  denote  $\bar{\Delta}(\pi, u, v)$ , where the subpath of  $\pi$  from  $u$  to  $v$  contains  $m$  pointer assignment statements. We prove the lemma by induction on  $m$ . From Definition 8,

$$\bar{\Delta}_m = \bar{\Delta}_{m-1} [e_m \circ \bar{\Delta}_{m-1}] \quad (8.2.a)$$

$$= \bar{\Delta}_{m-1} [e] \quad \text{where } e = e_m \circ \bar{\Delta}_{m-1} \quad (8.2.b)$$

For basis  $m = 1$ ,  $\bar{\Delta}_1$  contains a single edge and  $\bar{\Delta}_0 = \emptyset$ . Hence the basis holds from Lemma 8.1. For the inductive hypothesis, assume

$$\mathbf{eval}(n, [\bar{\Delta}_m] \bar{M}_{u,\pi}) = \mathbf{eval}(n \circ \bar{\Delta}_m, \bar{M}_{u,\pi}) \quad (8.2.c)$$

To prove,

$$\mathbf{eval}(n, [\bar{\Delta}_{m+1}] \bar{M}_{u,\pi}) = \mathbf{eval}(n \circ \bar{\Delta}_{m+1}, \bar{M}_{u,\pi})$$

For  $m + 1$ , the RHS of (8.2.c) becomes

$$\begin{aligned} & \mathbf{eval}(n \circ \bar{\Delta}_{m+1}, \bar{M}_{u,\pi}) \\ \Rightarrow & \mathbf{eval}(n \circ (\bar{\Delta}_m[e_{m+1} \circ \bar{\Delta}_m]), \bar{M}_{u,\pi}) \quad (\text{using (8.2.a) for } \bar{\Delta}_{m+1}) \\ \Rightarrow & \mathbf{eval}(n \circ (\bar{\Delta}_m[e]), \bar{M}_{u,\pi}) \quad (\text{let } e_{m+1} \circ \bar{\Delta}_m = e) \\ \Rightarrow & \mathbf{eval}(n, [\bar{\Delta}_m[e]] \bar{M}_{u,\pi}) \quad (\text{from (8.2.d) and (8.2.c)}) \\ \Rightarrow & \mathbf{eval}(n, [\bar{\Delta}_{m+1}] \bar{M}_{u,\pi}) \quad (\text{from (8.2.e) and (8.2.b)}) \end{aligned} \quad (8.2.d)$$

Hence the lemma.  $\square$

THEOREM 8.3. (Soundness of  $\bar{\Delta}$ ). Let a control flow path  $\pi$  from  $u$  to  $v$  contain  $k$  statements. Then,  $\bar{M}_{v,\pi} = [\bar{\Delta}(\pi, u, v)] \bar{M}_{u,\pi}$

PROOF. From Lemma 8.2, the effect of the reduced form  $e \circ \bar{\Delta}$  of an edge  $e$  on memory  $\bar{M}_{u,\pi}$  is identical to the effect of  $e$  on the resulting memory obtained after GPG application of  $\bar{\Delta}$  to  $\bar{M}_{u,\pi}$ . This holds for every edge in  $\bar{\Delta}$  and the theorem follows from induction on the number of statements covered by  $\bar{\Delta}$ .  $\square$

**8.4.2. Soundness of GPGs for Abstract Memory.** We argue below that  $\Delta(u, v)$  is sound because it under-approximates the removal of GPG edges and over-approximates the inclusion of GPG edges compared to  $\bar{\Delta}(\pi, u, v)$  for any  $\pi$  from  $u$  to  $v$ .

In order to relate  $\bar{M}_{u,\pi}$  and  $M_u$  (i.e., concrete and abstract memory), we rewrite the update operation for concrete memory (Definition 9) which reorients the edges without explicitly defining the edges being removed. We explicate the edges being removed by rewriting the equation as:

$$[\bar{n}] \bar{M}_{u,\pi} = (\bar{M}_{u,\pi} - \mathbf{kill}(\pi, n, \bar{M}_{u,\pi})) \cup \mathbf{eval}(n, \bar{M}_{u,\pi}) \quad (6)$$

$$\mathbf{kill}(\pi, n, \bar{M}_{u,\pi}) = \{ e_1 \mid e_1 \in \mathbf{match}(e, \bar{M}_{u,\pi}), e \in \mathbf{eval}(n, \bar{M}_{u,\pi}) \} \quad (7)$$

Let  $\mathbf{Paths}(u, v)$  denote the set of all control flow paths from  $u$  to  $v$ .

LEMMA 8.4. Abstract summary flow function under-approximates the removal of information.

$$\mathbf{memkill}(n, M_u, \Delta(u, v)) \subseteq \bigcap_{\pi \in \mathbf{Paths}(u, v)} \mathbf{kill}(\pi, n, \bar{M}_{u,\pi}) \quad (8.4.a)$$

```

01 void f()
02 {
03     fp = p;
04     x = &a;
05     g(fp);
06     fp = q;
07     z = &b;
08     g(fp);
09     z = &c;
10     g(fp);
11 }
12 void g(fp)
13 {
14     fp();
15 }
16 void p()
17 {
18     y = x;
19 }
20 void q()
21 {
22     y = z;
23 }

```

Fig. 17. An example demonstrating the top-down traversal of call graph for handling function pointers.

PROOF. Observe that *memkill* (Definition 10) is more conservative than *kill* (Equation 7) because it additionally requires that  $n$  should cause a strong update. From Definition 10, for causing a strong update,  $n$  must be defined along every path and the removable edges must define the same source along every path. Hence 8.4.a follows.  $\square$

LEMMA 8.5. *Abstract summary flow function over-approximates the inclusion of information.*

PROOF. Since the rules of composition are same for both  $\Delta$  and  $\overline{\Delta}$ , it follows from Definition 4 that,

$$n \circ \Delta(u, v) \supseteq \bigcup_{\pi \in \text{Paths}(u, v)} n \circ \overline{\Delta}(\pi, u, v)$$

$\square$

THEOREM 8.6. (*Soundness of  $\Delta$* ). *Abstract summary flow function  $\Delta(u, v)$  is a sound approximation of all concrete summary flow functions  $\overline{\Delta}(\pi, u, v)$ .*

$$\llbracket \Delta(u, v) \rrbracket M_u \supseteq \bigcup_{\pi \in \text{Paths}(u, v)} \llbracket \overline{\Delta}(\pi, u, v) \rrbracket \overline{M}_{u, \pi}$$

PROOF. It follows because killing of points-to information is under-approximated (Lemma 8.4) and generation of points-to information is over-approximated (Lemma 8.5).  $\square$

## 9. HANDLING ADVANCED FEATURES FOR POINTS-TO ANALYSIS USING GPGS

So far we have created the concept of GPGs for (possibly transitive) pointers to scalars allocated on stack or in the static area. This section extends the concepts to data structures created using C style *struct* or *union* and possibly allocated on heap too, apart from stack and static area.

We also show how to handle arrays, pointer arithmetic, and interprocedural analysis in the presence of function pointers.

### 9.1. Handling Function Pointers

In the presence of indirect calls (eg. a call through a function pointer in C), the callee procedure is not known at compile time. In our case, construction of the GPG of a procedure requires incorporating

| Pointer assignment     | GPG edge                      | Alternative GPG edge      |
|------------------------|-------------------------------|---------------------------|
| $x = \text{new} \dots$ | $x \xrightarrow{[*,*]} h_0$   | —                         |
| $x = y.n$              | $x \xrightarrow{[*,*]} y.n$   | $x \xrightarrow{[*,n]} y$ |
| $x.n = y$              | $x.n \xrightarrow{[*,*]} y$   | $x \xrightarrow{[n,*]} y$ |
| $x = y \rightarrow n$  | $x \xrightarrow{[*,*,n]} y$   | —                         |
| $x \rightarrow n = y$  | $x \xrightarrow{[*,n],[*]} y$ | —                         |

Fig. 18. GPG edges with indirection lists (*indlist*) for basic pointer assignments in C for structures and heap.  $h_i$  is the heap location at the allocation site  $i$ .  $*$  is the dereference operator.

the effect of the GPGs of all its callees. In the presence of indirect calls, we would not know the callees whose GPGs should be used at an indirect call site.

If the function pointers are defined locally, their effect can be handled easily because the pointees of function pointers would be available during GPG construction. Consider the function pointers that are passed as parameters or global function pointers that are defined in the callers. A top-down interprocedural points-to analysis would be able to handle such function pointers naturally because the information flows from callers to callees and hence the pointees of function pointers would be known at the call sites. However, a bottom-up interprocedural analysis such as ours, works in two phases and the information flows from

- the callees to callers when GPGs are constructed, and from
- the callers to callees when GPGs are used for computing the points-to information.

By default, the function pointer values are available only in the second phase whereas they are actually required in the first phase.

A bottom-up approach requires that the summary of a callee procedure should be constructed before its calls in caller procedures are processed. If a procedure  $f$  calls procedure  $g$ , this requirement can be satisfied by beginning to construct the GPG of  $f$  before that of  $g$ ; when a call to  $g$  is encountered, the GPG construction of  $f$  can be suspended and  $g$  can be processed completely by constructing its GPG before resuming the GPG construction of  $f$ . Thus, we can traverse the call graph top-down and yet construct bottom-up context independent summary flow functions. We start the GPG construction with the *main* procedure and suspend the construction of its GPG  $\Delta_{\text{main}}$  when a call is encountered and then analyze the callee first. After the completion of construction of GPG of the callee, then the construction of  $\Delta_{\text{main}}$  is resumed. Thus, the construction of GPG of callees is completed before the construction of GPG of their caller. In the process, we collect the pointees of function pointers along the way during the top-down traversal. These values (i.e., only the function pointer values) from the calling contexts are used to build the GPGs.

Observe that a GPG so constructed is context independent for the rest of the pointers but is customized for a specific value of a function pointer that is passed as a parameter or is defined globally. In other words, a procedure with an indirect call should have different GPGs for distinct values of function pointer for context-sensitivity. This is important because the call chains starting at a call through a function pointer in that procedure could be different.

EXAMPLE 9.1. In the example of Figure 17, we first analyze procedure  $f$  as we traverse the call graph top-down and suspend the construction of its GPG at the call site at line number 05 to analyze its callee which is procedure  $g$ . We construct a customized GPG for procedure  $g$  with  $fp = p$ . The pointee information of  $x$  is not used for GPG construction of  $g$ . In procedure  $g$ , there is a call through function pointer whose value is  $p$  as extracted from the calling context, we now suspend the GPG construction of  $g$  and the GPG of  $p$  is constructed first and its effect is incorporated in  $g$  with  $\Delta = \{y \xrightarrow{1,1} x\}$ . We then resume with the GPG construction of procedure  $f$  by incorporating the

```

struct node *x, *y;
struct node z;

01 struct node{
02 {
03     struct node *m, *n;
04 };
05 void f()
06 {
07     x = malloc(...);
08     y = x;
09     w = y->n;
10     g();
11 }
12 void g()
13 {
14     while(...) {
15         y = x->m;
16         x = y->n;
17     }
18     z.m = x;
19 }

```

Fig. 19. An example for modelling structures and heap.

effect of procedure  $g$  at line number 05 which results in a reduced edge  $y \xrightarrow{1,0} a$  by performing the required edge compositions.

At the call site at line number 07, procedure  $g$  is analyzed again with a different value of  $fp$  and this time procedure  $q$  is the callee which is analyzed and whose effect is incorporated to construct GPG for procedure  $g$  with  $\Delta = \{y \xrightarrow{1,1} z\}$  for  $fp = q$ . Note that procedure  $g$  has two GPGs constructed for different values of function pointer  $fp$  so far encountered. However, procedure  $p$  and  $q$  has only one GPG as they do not have any calls through function pointers. At line number 07,  $y$  now points to  $b$  as  $z$  points to  $b$  (because  $\Delta_g = \{y \xrightarrow{1,1} z\}$  for  $fp = q$ ).

The third call to  $g$  at line number 10 does not require re-analysis of procedure  $g$  as GPG is already constructed because value of  $fp$  is not changed. So GPG of procedure  $g$   $\Delta_g = \{y \xrightarrow{1,1} z\}$  for  $fp = q$  is reused at line number 10. The pointee of  $y$  however is now  $c$  as the pointee of  $z$  has changed.  $\square$

## 9.2. Handling Structures, Unions, and Heap Data

In this section, we describe the construction of GPGs for pointers to structures, unions, and heap allocated data. We use allocation site based abstraction for heap in which all locations allocated at a particular allocation site are over-approximated and are treated alike. This approximation allows us to handle the unbounded nature of heap as if it were bounded. However, since the allocation site might not be available during GPG construction phase (because it could occur in the callers), the heap accesses within a loop may remain unbounded and we need a summarization technique to bound them. This section first introduces the concept of indirection lists (*indlist*) for handling structures and heap accesses which is then followed by an explanation of the summarization technique we have used.

Figure 18 illustrates the GPG edges corresponding to the basic pointer assignments in C for structures and heap. The *indlev* “ $i, j$ ” of an edge  $x \xrightarrow{i,j} y$  represents  $i$  dereferences of  $x$  and  $j$  dereferences of  $y$ . We can also view the *indlev* “ $i, j$ ” as lists (also referred to as indirection list or *indlist*) containing the dereference operator (\*) of length  $i$  and  $j$ . This representation naturally allows handling structures and heap field-sensitively by using indirection lists containing field dereferences. With this view, we can represent the two statements at line numbers 08 and 09 in the example of Figure 19 by GPG edges in the following two ways:

- *Field-Sensitively.*  $y \xrightarrow{[*,*]} x$  and  $w \xrightarrow{[*,*,n]} y$ ; field-sensitivity is achieved by enumerating the field dereferences.
- *Field-Insensitively.*  $y \xrightarrow{1,1} x$  and  $w \xrightarrow{1,2} y$ ; no distinction made between any field dereference.<sup>6</sup>

The dereference in the pointer expression  $y \rightarrow n$  on line 09 is represented by an  $indlist$   $[*, n]$  associated with the pointer variable  $y$ . On the other hand, the access  $z.m$  on line 18 can be mapped to location by adding the offset of  $m$  to the virtual address of  $z$  at compile time. Hence, it can be treated as a separate variable which is represented by a node  $z.m$  with an  $indlist$   $[*]$  in the GPG. We can also represent  $z.m$  with a node  $z$  and an  $indlist$   $[m]$ . For our implementation, we chose the former representation for  $z.m$ . For structures and heap, we ensure field-sensitivity by maintaining  $indlist$  in terms of field names. Unions are handled similarly to structures.

Recall that an edge composition  $n \circ p$  involves balancing the  $indlev$  of the pivot in  $n$  and  $p$ . With  $indlist$  replacing  $indlev$ , the operations remain similar in spirit although now they become operations on lists rather than operations on numbers. To motivate the operations on  $indlist$ , let us recall the operations on  $indlev$  as illustrated in the following example.

**EXAMPLE 9.2.** Consider the example in Figure 19. Edge composition  $n \circ p$  requires balancing  $indlevs$  of the pivot (Section 4) which involves computing the difference between the  $indlev$  of the pivot in  $n$  and  $p$ . This difference is then added to the  $indlev$  of the non-pivot node in  $n$  or  $p$ . Recall that an edge composition is useful (Section 4.2) only when the  $indlev$  of the pivot in  $n$  is greater than or equal to the  $indlev$  of the pivot in  $p$ . Thus, in our example with  $p \equiv y \xrightarrow{1,1} x$  and  $n \equiv w \xrightarrow{1,2} y$  with  $y$  as pivot, an edge composition is useful because  $indlev$  of  $y$  in  $n$  (which is 2) is greater than  $indlev$  of  $y$  in  $p$  (which is 1). The difference (2-1) is added to the  $indlev$  of  $x$  (which is 1) resulting in a reduced edge  $r \equiv w \xrightarrow{1,(2-1+1)} x$ .  $\square$

Analogously we can define similar operations for  $indlist$ . An edge composition is useful if the  $indlist$  of the pivot in edge  $p$  is a prefix of the  $indlist$  of the pivot in edge  $n$ . In our example, the  $indlist$  of  $y$  in  $p$  (which is  $[*]$ ) is a prefix of the  $indlist$  of  $y$  in  $n$  (which is  $[*, n]$ ) and hence the edge composition is useful. The addition of the difference in the  $indlevs$  of the pivot to the  $indlev$  of one of the other two nodes is replaced by an append operation denoted by  $\#$ .

The operation of computing the difference in the  $indlev$  of the pivot is replaced by the remainder operation  $remainder : indlist \times indlist \rightarrow indlist$  which takes two  $indlist$ s as its arguments where first is a prefix of the second and returns the suffix of the second  $indlist$  that remains after removing the first  $indlist$  from it. Given  $il_2 = il_1 \# il_3$ ,  $remainder(il_1, il_2) = il_3$ . Note that  $il_3$  is  $\epsilon$  when  $il_1 = il_2$ . Further  $remainder(il_1, il_2)$  is not computed when  $il_1$  is not a prefix of  $il_2$ .

**EXAMPLE 9.3.** In our example,  $remainder([*], [*, n])$  returns  $[n]$  and this  $indlist$  is appended to the  $indlist$  of node  $x$  (which is  $[*]$ ) resulting in a new  $indlist$   $[*] \# [n] = [*, n]$  and a reduced edge  $w \xrightarrow{[*,*,n]} x$ .  $\square$

Under the allocation site based abstraction for heap, line number 07 of procedure  $f$  can be viewed as a GPG edge  $x \xrightarrow{1,0} \text{heap}_{07}$  where  $\text{heap}_{07}$  is the heap location created at this allocation site. We expect the heap to be bounded by this abstraction but the allocation site may not be available during the GPG construction as is the case in our example where heap is accessed through pointers  $x$  and  $y$  in a loop in procedure  $g$  whereas allocation site is available in procedure  $f$  at line 07.

**EXAMPLE 9.4.** The fixed point computation for the loop in procedure  $g$  will never terminate as the length of the indirection list keeps on increasing. In the first iteration of the loop, at its exit, the edge composition results into a reduced edge  $x \xrightarrow{[*,*,m,n]} y$ . In the next iteration, the reduced edge is now  $x \xrightarrow{[*,*,m,n,m,n]} y$  indicating the access pattern of heap. This continues as the length of the indirection list keeps on increasing leading to a non-terminating sequence of computations. Heap

---

<sup>6</sup>This does not matter for the first edge but matters for the second edge.

access where the allocation site is locally available does not face this problem of non-termination.

This indicates the need of a summarization technique. We bound the indirection lists by  $k$ -limiting technique which limits the length of indirection lists upto  $k$  dereferences. All dereferences beyond  $k$  are treated as an unbounded number of field insensitive dereferences.

Note that an explicit summarization is required only for heap locations and not for stack locations because the *indlists* can grow without bound only for heap locations.

### 9.3. Using SSA Form for Compact GPGs

Although the Static Single Assignment (SSA) form is not a language feature, it is ubiquitous in any real IR of practical programs. In this section we show how we have used the SSA productively to make our analysis more efficient and construct compact GPGs.

SSA form makes use-def chains explicit in the IR because every use has exactly one definition reaching it and every definition dominates all its uses. Thus for every local non-address taken variable, we traverse the SSA chains transitively until we reach a statement whose right hand side has an address taken variable, a global variable, or a formal parameter. In the process, all definitions involving SSA variables on the left hand side are skipped.

**EXAMPLE 9.5.** Consider the code snippet in its SSA form on the right. The GPG edge  $x\_1 \xrightarrow{1,0} a$  corresponding to statement  $s_1$  is not added to the GPG. Statement  $s_2$  defines a global pointer  $y$  which is assigned the pointee of  $x\_1$  (use of  $x\_1$ ). The explicit use of use-def chain helps to identify the pointee of  $x\_1$  even though there is no corresponding edge in the GPG. SSA resolution leads to an edge  $y \xrightarrow{1,0} a$  which is the desired result, also indicating the fact that SSA resolution is similar to edge composition.  $\square$

The use of SSA has the following two advantages:

- The GPG size is small because local variables are eliminated.
- No special filtering required for eliminating local variables from the summary flow function of a procedure. These local variables are not in the scope of the callers and hence should be eliminated before a summary flow function is used at its call sites.

Both of them aid efficiency.

### 9.4. Handling Arrays, Pointer Arithmetic, and Address Escaping Locals

An array is treated as a single variable in the following sense: Accessing a particular element is seen as accessing every possible element and updates are treated as weak updates. This applies to both the situations: when arrays of pointers are manipulated, as well as when arrays are accessed through pointers. Since there is no kill, arrays are maintained flow-insensitively by our analysis.

For pointer arithmetic, we approximate the pointer being defined to point to every possible location. All address taken local variables in a procedure are treated as global variables because they can escape the scope of the procedure. However, these variables are not strongly updated because they could represent multiple locations.

| Program    | kLoC            | # of pointer stmts | Time for GPG based approach (in seconds)                        |                          |         |               |   |                    | Avg. # of pointees per pointer |                  |  |       |      |       | Avg. # of pointees per dereference                                     |      |      |     |  |  |
|------------|-----------------|--------------------|---|--------------------------|---------|---------------|---|--------------------|--------------------------------|------------------|--|-------|------|-------|--|------|------|-----|--|--|
|            |                 |                    | GPG Constr.   | computing points-to info |         |               | GPG   |                    |                                | GCC              |  | LFCPA |      |       |  |      |      |     |  |  |
|            |                 |                    |   | GPG NoByp                | GPG Byp | Stmt-ff NoByp | Stmt-ff Byp   | G/NoByp (per stmt) | G/Byp (per stmt)               | L+Arr (per proc) | G+L+Arr (per proc)   | GPG   | GCC  | LFCPA |  |      |      |     |  |  |
|            |                 |                    | A   | B                        | C       | D             | E   | F                  | G                              | H                | I  | J     | K    | L     | M  | N    | O    |     |  |  |
| lbm        | 0.9             | 370                | 0.10  | 0.22                     | 0.21    | 0.26          | 0.28  | 1.31               | 1.42                           | 2.21             | 17.74  | 0.05  | 1.09 | 2.25  | 1.50   |      |      |     |  |  |
| mcf        | 1.6             | 480                | 75.29   | 33.73                    | 30.05   | 1.25          | 0.91  | 18.73              | 6.10                           | 10.48            | 34.74  | 1.22  | 4.25 | 2.57  | 0.62   |      |      |     |  |  |
| libquantum | 2.6             | 340                | 6.47  | 10.23                    | 1.95    | 8.21          | 1.85  | 139.50             | 22.50                          | 1.11             | 4.49   | 3.34  | 1.50 | 2.93  | 0.83   |      |      |     |  |  |
| bzip2      | 5.7             | 1650               | 3.17  | 11.11                    | 8.71    | 4.73          | 3.30  | 43.39              | 8.38                           | 1.89             | 31.46  | 0.94  | 1.72 | 2.94  | 0.33   |      |      |     |  |  |
| milc       | 9.5             | 2540               | 7.36  | 6.08                     | 5.89    | 4.29          | 5.61  | 21.15              | 16.32                          | 4.52             | 14.06  | 31.73 | 1.18 | 2.58  | 1.61   |      |      |     |  |  |
| sjeng      | 10.5            | 700                | 9.36  | 39.66                    | 25.75   | 14.75         | 7.56  | 445.22             | 64.81                          | 3.07             | 2.68   | -     | 0.98 | 2.71  | -  |      |      |     |  |  |
| hmmer      | 20.6            | 6790               | 38.23   | 51.73                    | 14.86   | 31.32         | 13.50   | 43.49              | 5.85                           | 6.05             | 59.35  | 1.56  | 1.04 | 3.62  | 0.91   |      |      |     |  |  |
| h264ref    | 36.1            | 17770              | 208.47  | 1262.07                  | 199.34  | 457.26        | 74.62   | 219.71             | 9.24                           | 16.29            | 98.84  | -     | 0.98 | 3.97  | -  |      |      |     |  |  |
| gobmk      | 158.0           | 212830             | 652.78  | 3652.99                  | 1624.46 | 1582.62       | 1373.88   | 11.98              | 1.73                           | 6.34             | 4.08   | -     | 0.65 | 3.71  | -  |      |      |     |  |  |
| Program    | # of call sites | # of procs.        | Proc. count for different buckets of # of calls (reuse of GPGs) |                          |         |               | # of procs. requiring different no. of PTFs based on the no. of aliasing patterns |                    |                                |                  | # of procs. for different sizes of GPG in terms of the number of edges |       |      |       | # of procs. for different % of context ind. info. (for non-empty GPGs) |      |      |     |  |  |
|            |                 |                    |   |                          |         |               | Actually observed   |                    | Predicted                      |                  |  |       |      |       |  |      |      |     |  |  |
|            |                 |                    | 2-5   | 5-10                     | 10-20   | 20+           | 2-5   | 6-10               | 11-15                          | 15+              | 2-5  | 15+   | 0    | 1-2   | 3-4  | 5-8  | 9-50 | 50+ |  |  |
|            |                 |                    | P   | Q                        | R       |               | S   | T                  |                                |                  | U  |       | V    |       | W  |      |      |     |  |  |
| lbm        | 30              | 19                 | 5   | 0                        | 0       | 0             | 8   | 0                  | 0                              | 0                | 13   | 0     | 13   | 4     | 2  | 0    | 0    | 0   |  |  |
| mcf        | 29              | 23                 | 11  | 0                        | 0       | 0             | 0   | 0                  | 0                              | 4                | 0  | 10    | 5    | 2     | 3  | 2    | 1    | 5   |  |  |
| libquantum | 277             | 80                 | 24  | 11                       | 4       | 3             | 7   | 3                  | 1                              | 0                | 14   | 4     | 42   | 10    | 7  | 12   | 9    | 0   |  |  |
| bzip2      | 288             | 89                 | 35  | 7                        | 2       | 1             | 22  | 0                  | 0                              | 0                | 28   | 2     | 62   | 13    | 4  | 5    | 5    | 0   |  |  |
| milc       | 782             | 190                | 60  | 15                       | 9       | 1             | 37  | 8                  | 0                              | 1                | 35   | 25    | 157  | 11    | 19   | 2    | 7    | 0   |  |  |
| sjeng      | 726             | 133                | 46  | 20                       | 5       | 6             | 14  | 3                  | 1                              | 3                | 10   | 14    | 99   | 20    | 6  | 3    | 5    | 0   |  |  |
| hmmer      | 1328            | 275                | 93  | 33                       | 22      | 11            | 62  | 5                  | 3                              | 4                | 88   | 32    | 167  | 56    | 20   | 15   | 15   | 2   |  |  |
| h264ref    | 2393            | 566                | 171   | 60                       | 22      | 16            | 85  | 17                 | 5                              | 3                | 102  | 46    | 419  | 76    | 23   | 15   | 30   | 3   |  |  |
| gobmk      | 9379            | 2697               | 317   | 110                      | 99      | 134           | 206   | 30                 | 9                              | 10               | 210  | 121   | 1374 | 93    | 8  | 1083 | 97   | 42  |  |  |

Fig. 20. Time, precision, size, and effectiveness measurements for GPG Based Points-to Analysis. Byp (Bypassing), NoByp (No Bypassing), Stmt-ff (Statement-level flow functions), G (Global pointers), L (Local pointers), Arr (Array pointers).

## 10. IMPLEMENTATION AND MEASUREMENTS

We have implemented GPG based points-to analysis in GCC 4.7.2 using the LTO framework and have carried out measurements on SPEC CPU2006 benchmarks on a machine with 16 GB RAM with 8 64-bit Intel i7-4770 CPUs running at 3.40GHz. Figure 20 provides the empirical data.

Our method eliminates local variables using the SSA form and GPGs are computed only for global variables. Eventually, the points-to information for local variables is computed from that of global variables and parameters. Heap memory is approximated by maintaining indirection lists of field dereferences of length 2 (see Section 9.2). Unlike the conventional approaches [33; 35; 36], our summary flow functions do not depend on aliasing at the call points. The actually observed number of aliasing patterns (column *S* in Figure 20) suggests that it is undesirable to indiscriminately construct multiple PTFs for a procedure.

Columns *A*, *B*, *P*, and *Q* in Figure 20 present the details of the benchmarks. Column *C* provides the time required for the first phase of our analysis i.e., computing GPGs. The computation of points-to information at each program point has four variants (using GPGs or *Stmt-ff* with or without bypassing). Their time measurements are provided in columns *D*, *E*, *F*, and *G*. Our data indicates that the most efficient method for computing points-to information is to use statement-level flow functions and bypassing (column *G*).

Our analysis computes points-to information flow-sensitively for globals. The following points-to information is stored flow-insensitively: locals (because they are in the SSA form) and arrays (because their updates are conservative). Hence, we have separate columns for globals (columns *H* and *I*) and locals+arrays (column *J*) for GPGs. GCC-PTA computes points-to information flow-insensitively (column *K*) whereas LFCPA computes it flow-sensitively (column *L*).

The second table provides measurements about the effectiveness of summary flow functions in terms of (*a*) compactness of GPGs, (*b*) percentage of context independent information, and (*c*) reusability. Column *U* shows that GPGs for a large number of procedures have 0 edges because they do not manipulate global pointers. Besides, in six out of nine benchmarks, most procedures with non-empty GPGs have a significantly high percentage of context independent information (column *V*). Thus a top-down approach may involve redundant computations on multiple visits to a procedure whereas a bottom-up approach may not need much work for incorporating the effect of a callee’s GPG into that of its callers. Further, many procedures are called multiple times indicating a high reuse of GPGs (column *R*).

Interestingly, computing points-to information using summary flow functions seems to take much more time than constructing summary flow functions. As discussed in Section 7, computing points-to information at every program point within a procedure using the *Bl* of the procedure and the summary flow function ( $\Delta$ ) is expensive because of the cumulative effect of the  $\Delta$ . The time measurements (columns *F* and *G*) confirm the observation that the application of statement-level flow functions is much more efficient than the application of GPGs for computing points-to information at every program point. These measurements also highlight the gain in efficiency achieved because of the bypassing technique [22; 23]. Bypassing technique helps to reduce the size of the *Bl* of a procedure by propagating only the relevant information.

The effectiveness of bypassing is evident from the time measurements (columns *E* and *G*) as well as a reduction in the average number of points-to pairs (column *I*). We have applied the bypassing technique only to the flow-sensitive points-to information.

We have compared our analysis with GCC-PTA and LFCPA [17]. The number of points-to pairs per function for GCC-PTA (column *K*) is large because it is partially flow-sensitive (because of the SSA form) and context-insensitive. The number of points-to pairs per statements is much smaller for LFCPA (column *L*) because it is liveness-based. However LFCPA which in our opinion represents the state of the art in fully flow- and context-sensitive exhaustive points-to analysis, does not seem to scale beyond 35 kLoC. We have computed the average number of pointees of dereferenced variables which is maximum for GCC-PTA (column *N*) and minimum for LFCPA (column *O*) because it is liveness driven. The points-to information computed by these methods is incomparable because they

employ radically dissimilar features of points-to information such as flow- and context-sensitivity, liveness, and bypassing.

## 11. RELATED WORK

In this section, we briefly review the literature related to flow- and context-sensitive analyses. As described earlier in Section 1, a context-sensitive interprocedural analysis may visit the procedures in a program by traversing its call graph top-down or bottom-up. A top-down approach propagates the information from callers to callees [36]. In the process, it analyzes a procedure each time a new data flow value reaches a procedure from some call. Since the information is propagated from callers to callees, all information that may be required for analyzing a procedure is readily available. A bottom-up approach, on the other hand, avoids analyzing procedures multiple times by constructing *summary flow functions* which are used in the calling contexts to incorporate the effect of procedure calls. Since the callers' information is not available, analyzing a procedure requires a convenient encoding of accesses of variables which are defined in the caller procedures. The effectiveness of a bottom-up approach crucially depends on the choice of representation of procedure summaries. For some analyses, the choice of representation is not obvious. In the absence of pointers, procedure summaries for bit-vector frameworks can be easily represented by *Gen* and *Kill* sets whose computation does not require any information from the calling context [15]. In the presence of pointers, the representation needs to model unknown locations indirectly accessed through pointers that may have been defined in the callers.

Section 2 introduced two broad categories of constructing summary flow functions for points-to analysis. Some methods using placeholders require aliasing information in the calling contexts and construct multiple summary flow functions per procedure [33; 36]. Other methods do not make any assumptions about the calling contexts [18; 19; 28; 31; 32] but they construct larger summary flow functions causing inefficiency in fixed point computation at the intraprocedural level thereby prohibiting flow-sensitivity for scalability. Also, these methods cannot perform strong updates thereby losing precision.

Among the general frameworks for constructing procedure summaries, the formalism proposed by Sharir and Pnueli [29] is limited to finite lattices of data flow values. It was implemented using graph reachability in [20; 26; 27]. A general technique for constructing procedure summaries [9] has been applied to unary uninterpreted functions and linear arithmetic. However, the program model does not include pointers.

Symbolic procedure summaries [33; 35] involve computing preconditions and corresponding postconditions (in terms of aliases). A calling context is matched against a precondition and the corresponding postcondition gives the result. However, the number of calling contexts in a program could be unbounded hence constructing summaries for all calling contexts could lose scalability. This method requires statement-level transformers to be closed under composition; a requirement which is not satisfied by points-to analysis (as mentioned in Section 2). We overcome this problem using generalized points-to facts. Saturn [10] also creates summaries that are sound but may not be precise across applications because they depend on context information.

Some approaches use customized summaries and combine the top-down and bottom-up analyses to construct summaries for only those calling contexts that occur in a given program [36]. This choice is controlled by the number of times a procedure is called. If this number exceeds a fixed threshold, a summary is constructed using the information of the calling contexts that have been recorded for that procedure. A new calling context may lead to generating a new precondition and hence a new summary.

GPGs handle function pointers efficiently and precisely by traversing the call graph top-down and yet construct bottom-up summary flow functions (see Section 9.1). The conventional approaches [19; 31; 32] perform type analysis for identifying the callee procedures for indirect calls through function pointers. All functions matching the type of a given function pointer are conservatively considered as potential callees thereby over-approximating the call graph significantly.

The PTF approach [33] suspends the summary construction when it encounters an indirect call and traverse the call graph bottom-up until all pointees of the function pointer are discovered.

Although GPGs use allocation-site based heap abstraction, they additionally need  $k$ -limiting summarization as explained in Section 9.2. The approaches [19; 31; 32; 33] use allocation-site based heap abstraction. Since they use as many placeholders as required explicating each location in a pointee chain, they do not require  $k$ -limiting summarization.

### On the use of graphs for representing summary flow functions

Observe that all approaches that we have seen so far use graphs to represent summary flow functions of a procedure. The interprocedural analysis via graph reachability [26] also represents a flow function using a graph. Let  $D$  denote the set of data flow values. Then a flow function  $2^D \rightarrow 2^D$  is modelled as a set of functions  $D \rightarrow D$  and is represented using a graph containing  $2 \cdot |D|$  nodes and at most  $(|D| + 1)^2$  edges. Each edge maps a value in  $D$  to a value in  $D$ ; this is very convenient because the function composition simply reduces to traversing a path created by adjacent edges, therefore the term reachability. Also, the meet operation on data flow values now reduces to the meet on the edges of the graph.

A graph representation is appropriate for a summary flow function only if each edge in the graph has its independent effect irrespective of the other edges in the graph. Graph reachability ensures this by requiring the flow functions to be distributive: If a function  $2^D \rightarrow 2^D$  distributes over a meet operator  $\sqcap$  then it can be modelled as a set of unary functions  $D \rightarrow D$ . However, graph reachability can also model some non-distributive flow functions. Consider a flow function for a statement  $y = x \% 4$  for constant propagation framework. This function does not distribute over  $\sqcap$  defined for the usual constant propagation lattice [15] because  $f(10 \sqcap 6) = f(\perp) = \perp$  whereas  $f(10) \sqcap f(6) = 2 \sqcap 2 = 2$ . However, this function can be represented by an edge in a graph from  $x$  to  $y$ . Thus distributivity is a sufficient requirement for graph reachability but is not necessary. The necessary condition for graph reachability is that a flow function should be representable in terms of a collection of unary flow functions.

Representing a pointer assignment  $*x = y$  requires modelling the pointees of  $x$  as well as  $y$ . With the classical points-to relations, this function does not remain a unary function. Similarly, a statement  $x = *y$  requires modelling the pointees of pointees of  $y$  and this function too does not remain a unary function with classical points-to relations. It is for this reason that the state of the art uses placeholders to represent unknown locations, such as pointees of  $x$  and  $y$  in this case. Use of place holders allows modelling the functions for statements  $*x = y$  or  $x = *y$  in terms of a collection of unary flow functions facilitating the use of graphs in which edge can have its own well defined independent effect. GPGs uses *indlev* with the edges to represent pointer indirections and hence, model the effect of pointer assignments in terms of unary flow functions.

Graph reachability fails to represent indirect accesses through pointers.

## 12. CONCLUSIONS AND FUTURE WORK

Constructing bounded summary flow functions for flow- and context-sensitive points-to analysis seems hard because it requires modelling unknown locations accessed indirectly through pointers—a callee procedure’s summary flow function is created without looking at the statements in the caller procedures. Conventionally, they have been modelled using placeholders. However, a fundamental problem with the placeholders is that they explicate the unknown locations by naming them. This results in either (a) a large number of placeholders, or (b) multiple summary flow functions for different aliasing patterns in the calling contexts. We propose the concept of generalized points-to graph (GPG) whose edges track indirection levels and represent generalized points-to facts. A simple arithmetic on indirection levels allows composing generalized points-to facts to create new generalized points-to facts with smaller indirection levels; this reduces them progressively to classical points-to facts. Since unknown locations are left implicit, no information about aliasing patterns in the calling contexts is required allowing us to construct a single GPG per procedure. GPGs are

linearly bounded by the number of variables, are flow-sensitive, and are able to perform strong updates within calling contexts. Further, GPGs inherently support bypassing of irrelevant points-to information thereby aiding scalability significantly.

Our measurements on SPEC benchmarks show that GPGs are small enough to scale fully flow- and context-sensitive exhaustive points-to analysis to programs as large as 158 kLoC (as compared to 35 kLoC of LFCPA [17]). We expect to scale the method to still larger programs by (a) using memoisation, and (b) constructing and applying GPGs incrementally thereby eliminating redundancies within fixed point computations.

Observe that a GPG edge  $x \xrightarrow{i,j} y$  in  $M$  also asserts an alias relation between  $M^i\{x\}$  and  $M^j\{y\}$  and hence GPGs generalize both points-to and alias relations.

The concept of GPG provides a useful abstraction of memory involving pointers. The way matrices represent values as well as transformations, GPGs represent memory as well as memory transformers defined in terms of loading, storing, and copying memory addresses. Any analysis that is influenced by these operations may be able to use GPGs by combining them with the original abstractions of the analysis. We plan to explore this direction in the future.

In presence of pointers, current analyses use externally supplied points-to information. Even if this information is computed context-sensitively, its use by other analyses is context-insensitive because at the end of the points-to analysis, the points-to information is conflated across all contexts at a given program point. GPGs on the other hand, allows other analyses to use points-to information that is valid for each context separately by performing joint analyses. Observe that joint context-sensitive analyses may be more precise than two separately context-sensitive cascaded analyses. We also plan to explore this direction of work in future.

### Acknowledgments.

The paper has benefited from the feedback of many people; in particular, Supratik Chakraborty and Sriram Srinivasan gave excellent suggestions for improving the accessibility of the paper. Our ideas have also benefited from discussions with Amitabha Sanyal, Supratim Biswas, and Venkatesh Chopella. The seeds of GPGs were explored in a very different form in the Master’s thesis of Shubhangi Agrawal in 2010.

## REFERENCES

- Thomas Ball and Sriram K. Rajamani. The slam project: Debugging system software via static analysis. In *Proceedings of the 29th ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages*, POPL ’02, pages 1–3, New York, NY, USA, 2002. ACM.
- Dirk Beyer, Thomas A. Henzinger, and Grégory Théoduloz. Configurable software verification: Concretizing the convergence of model checking and program analysis. In *Proceedings of the 19th International Conference on Computer Aided Verification*, CAV’07, pages 504–518, Berlin, Heidelberg, 2007. Springer-Verlag.
- Edmund Clarke, Daniel Kroening, and Flavio Lerda. A tool for checking ANSI-C programs. In *In Tools and Algorithms for the Construction and Analysis of Systems*, pages 168–176. Springer, 2004.
- Camil Demetrescu, David Eppstein, Zvi Galil, and Giuseppe F. Italiano. Dynamic graph algorithms. In Mikhail J. Atallah and Marina Blanton, editors, *Algorithms and Theory of Computation Handbook*, pages 9.1–9.14. Chapman & Hall/CRC, 2010.
- Camil Demetrescu and Giuseppe F. Italiano. Maintaining dynamic matrices for fully dynamic transitive closure. *Algorithmica*, 51(4):387–427, May 2008.
- Isil Dillig, Thomas Dillig, and Alex Aiken. Sound, complete and scalable path-sensitive analysis. In *Proceedings of the 29th ACM SIGPLAN Conference on Programming Language Design and Implementation*, PLDI ’08, New York, NY, USA, 2008. ACM.
- Yu Feng, Xinyu Wang, Isil Dillig, and Thomas Dillig. Bottom-up context-sensitive pointer analysis for Java. In *Programming Languages and Systems - 13th Asian Symposium, APLAS 2015, Pohang, South Korea, November 30 - December 2, 2015, Proceedings*, 2015.
- Jeffrey Fischer, Ranjit Jhala, and Rupak Majumdar. Joining dataflow with predicates. In *Proceedings of the 10th European Software Engineering Conference Held Jointly with 13th ACM SIGSOFT International Symposium on Foundations of Software Engineering*, ESEC/FSE-13, pages 227–236, New York, NY, USA, 2005. ACM.

- S. Gulwani and A. Tiwari. Computing procedure summaries for interprocedural analysis. In R. De Nicola, editor, *European Symp. on Programming, ESOP 2007*, volume 4421 of *LNCS*, 2007.
- Brian Hackett and Alex Aiken. How is aliasing used in systems software? In *Proceedings of the 14th ACM SIGSOFT International Symposium on Foundations of Software Engineering*, SIGSOFT '06/FSE-14, New York, NY, USA, 2006. ACM.
- Nevin Heintze and Olivier Tardieu. Demand-driven pointer analysis. In *Proceedings of the ACM SIGPLAN 2001 Conference on Programming Language Design and Implementation*, PLDI '01, New York, NY, USA, 2001. ACM.
- John E. Hopcroft, Rajeev Motwani, and Jeffrey D. Ullman. Introduction to automata theory, languages, and computation, 2nd edition. *SIGACT News*, 32(1), March 2001.
- Franjo Ivančić, Ilya Shlyakhter, Aarti Gupta, Malay K Ganai, Vineet Kahlon, Chao Wang, and Zijiang Yang. Model checking C programs using F-Soft. In *IEEE International Conference on Computer Design*, pages 297–308. IEEE, 2005.
- Ranjit Jhala and Rupak Majumdar. Software model checking. *ACM Comput. Surv.*, 41(4):21:1–21:54, October 2009.
- U. P. Khedker, A. Sanyal, and B. Sathe. *Data Flow Analysis: Theory and Practice*. Taylor & Francis (CRC Press, Inc.), Boca Raton, FL, USA, 2009.
- Uday P. Khedker and Bageshri Karkare. Efficiency, precision, simplicity, and generality in interprocedural data flow analysis: resurrecting the classical call strings method. In *Proceedings of the Joint European Conferences on Theory and Practice of Software 17th international conference on Compiler construction*, CC'08/ETAPS'08, 2008.
- Uday P. Khedker, Alan Mycroft, and Prashant Singh Rawat. Liveness-based pointer analysis. In *Proceedings of the 19th International Static Analysis Symposium*, SAS'12, Berlin, Heidelberg, 2012. Springer-Verlag.
- Lian Li, Cristina Cifuentes, and Nathan Keynes. Precise and scalable context-sensitive pointer analysis via value flow graph. In *Proceedings of the 2013 International Symposium on Memory Management*, ISMM '13, New York, NY, USA, 2013. ACM.
- Ravichandran Madhavan, G. Ramalingam, and Kapil Vaswani. Modular heap analysis for higher-order programs. In *Proceedings of the 19th International Conference on Static Analysis*, SAS'12, Berlin, Heidelberg, 2012. Springer-Verlag.
- Nomair A. Naeem, Ondřej Lhoták, and Jonathan Rodriguez. Practical extensions to the ifds algorithm. In *Proceedings of the 19th Joint European Conference on Theory and Practice of Software, International Conference on Compiler Construction*, CC'10/ETAPS'10, Berlin, Heidelberg, 2010. Springer-Verlag.
- Erik M. Nystrom, Hong-Seok Kim, and Wen-mei W. Hwu. Bottom-up and top-down context-sensitive summary-based pointer analysis. In *Static Analysis, 11th International Symposium*, SAS 2004, Verona, Italy, August 26-28, 2004, Proceedings, 2004.
- Hakjoo Oh, Kihong Heo, Wonchan Lee, Woosuk Lee, and Kwangkeun Yi. Design and implementation of sparse global analyses for c-like languages. In *ACM SIGPLAN Conference on Programming Language Design and Implementation, PLDI '12, Beijing, China - June 11 - 16, 2012*, 2012.
- Hakjoo Oh, Wonchan Lee, Kihong Heo, Hongseok Yang, and Kwangkeun Yi. Selective context-sensitivity guided by impact pre-analysis. In *ACM SIGPLAN Conference on Programming Language Design and Implementation, PLDI '14, Edinburgh, United Kingdom - June 09 - 11, 2014*, 2014.
- Rohan Padhye and Uday P. Khedker. Interprocedural data flow analysis in SOOT using value contexts. In *Proceedings of the 2Nd ACM SIGPLAN International Workshop on State Of the Art in Java Program Analysis*, SOAP '13, New York, NY, USA, 2013. ACM.
- Uday P. Khedker Pritam M. Gharat and Alan Mycroft. Flow and context sensitive points-to analysis using generalized points-to graphs. In *Proceedings of the 23rd Static Analysis Symposium*, SAS'16, Berlin, Heidelberg, 2016. Springer-Verlag.
- Thomas Reps, Susan Horwitz, and Mooly Sagiv. Precise interprocedural dataflow analysis via graph reachability. In *Proceedings of the 22Nd ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages*, POPL '95, New York, NY, USA, 1995. ACM.
- Mooly Sagiv, Thomas Reps, and Susan Horwitz. Precise interprocedural dataflow analysis with applications to constant propagation. In *Selected Papers from the 6th International Joint Conference on Theory and Practice of Software Development*, TAPSOFT '95, Amsterdam, The Netherlands, The Netherlands, 1996. Elsevier Science Publishers B. V.
- Lei Shang, Xinwei Xie, and Jingling Xue. On-demand dynamic summary-based points-to analysis. In *Proceedings of the Tenth International Symposium on Code Generation and Optimization*, CGO '12, New York, NY, USA, 2012. ACM.
- A. Sharir M., Pnueli. Two approaches to interprocedural data flow analysis. S.S., Jones, N.D. (eds.) *Program Flow Analysis: Theory and Applications*, (ch. 7), 1981.
- Manu Sridharan, Denis Gopan, Lexin Shan, and Rastislav Bodík. Demand-driven points-to analysis for Java. In *Proceedings of the 20th Annual ACM SIGPLAN Conference on Object-oriented Programming, Systems, Languages, and Applications*, OOPSLA '05, New York, NY, USA, 2005. ACM.
- Alexandru Sălcianu and Martin Rinard. Purity and side effect analysis for Java programs. In *Proceedings of the 6th International Conference on Verification, Model Checking, and Abstract Interpretation*, VMCAI'05, Berlin, Heidelberg, 2005. Springer-Verlag.

- John Whaley and Martin Rinard. Compositional pointer and escape analysis for Java programs. In *Proceedings of the 14th ACM SIGPLAN Conference on Object-oriented Programming, Systems, Languages, and Applications*, OOPSLA '99, New York, NY, USA, 1999. ACM.
- R. P. Wilson and M. S. Lam. Efficient context-sensitive pointer analysis for C programs. In *Proceedings of the ACM SIGPLAN Conference on Programming Language Design and Implementation*, PLDI '95, 1995.
- Dacong Yan, Guoqing Xu, and Atanas Rountev. Rethinking SOOT for summary-based whole-program analysis. In *Proceedings of the ACM SIGPLAN International Workshop on State of the Art in Java Program Analysis*, SOAP '12, New York, NY, USA, 2012. ACM.
- Greta Yorsh, Eran Yahav, and Satish Chandra. Generating precise and concise procedure summaries. In *Proceedings of the 35th Annual ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages*, POPL '08, New York, NY, USA, 2008. ACM.
- Xin Zhang, Ravi Mangal, Mayur Naik, and Hongseok Yang. Hybrid top-down and bottom-up interprocedural analysis. In *Proceedings of the 35th ACM SIGPLAN Conference on Programming Language Design and Implementation*, PLDI '14, New York, NY, USA, 2014. ACM.