Lecture 12
Pointer Analysis

1. Motivation: security analysis
2. Datalog
3. Context-insensitive, flow-insensitive pointer analysis
4. Context sensitivity

Readings: Chapter 12

A Simple SQL Injection Pattern

o = *req*.getParameter ( );

stmt.executeQuery (*o*);

In Practice

ParameterParser.java:586
String session.ParameterParser.getRawParameter(String name)

```
public String getRawParameter(String name)
        throws ParameterNotFoundException {
        String[] values = request.getParameterValues(name);
        if (values == null) {
            throw new ParameterNotFoundException(name + " not found");
        } else if (values[0].length() == 0) {
            throw new ParameterNotFoundException(name + " was empty");
        }
        return (values[0]);
    }
```

ParameterParser.java:570
String session.ParameterParser.getRawParameter(String name, String def)

```
public String getRawParameter(String name, String def) {
   try {
    return getRawParameter(name);
   } catch (Exception e) {
    return def;
   }
}
```


In Practice (II)

ChallengeScreen.java:194
Element lessons.ChallengeScreen.doStage2(WebSession s)

```
String user = s.getParser().getRawParameter( USER, ""
);
StringBuffer tmp = new StringBuffer();
tmp.append("SELECT cc_type, cc_number from user_data
WHERE userid = '");
tmp.append(user);
tmp.append("'");
query = tmp.toString();
Vector v = new Vector();
try
{
  ResultSet results = statement3.executeQuery( query );
...
```

Vulnerabilities in Web Applications

Inject		Exploit
Parameters	X	SQL injection
Hidden fields		Cross-site scripting
Headers		HTTP splitting
Cookie poisoning		Path traversal

Key: Information Flow

PQL: Program Query Language

o = *req*.getParameter ( );

stmt.executeQuery (*o*);

- Query on the dynamic behavior based on object entities
- Abstracting away information flow

Dynamic vs. Static Pattern

Dynamically:

o = *req*.getParameter ( );
stmt.executeQuery (*o*);

Statically:

p_1 = *req*.getParameter ( );
stmt.executeQuery (p_2);

p_1 and p_2 point to same object?

Pointer alias analysis

Security Analysis

- Classifications
  - Conservative
    - All errors are reported
    - Include: false positives
  - Opportunistic
    - Only a subset of errors is reported
    - Include: false positives and false negatives
- Pointer alias analysis
  - No pointers
  - Flow-sensitive analysis?
  - Context-sensitive analysis?

Automatic Analysis Generation

Programmer: Security analysis in 10 lines → **PQL**

Compiler Writer: Flow-insensitive Context-sensitive Ptr analysis in 10 lines → **Datalog**

bddbddb (**BDD**-**b**ased **d**eductive **d**ata**b**ase) with Active Machine Learning

1000s of lines 1 year tuning → **BDD operations**

BDD (Binary Decision Diagrams): 10,000s-lines library

Goals of the Lecture

- Pointer analysis
  - Interprocedural, context-sensitive, flow-insensitive (Dataflow: intraprocedural, flow-sensitive)
- Power of languages and abstractions
- Elegant abstractions
  - Datalog: A deductive database (A database that can make deductions from stored data)
  - BDDs: Binary decision diagrams (Most cited CS papers for many years)

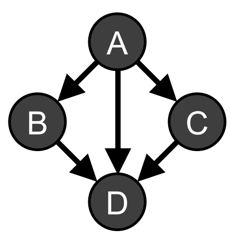
Outline
Pointer Analysis

1. Motivation: security analysis
2. Datalog
3. Context-insensitive, flow-insensitive pointer analysis
4. Context sensitivity

2. Datalog: a Deductive Database

- Relations as predicates
  - p(X_1, X_2, … X_n)
    - X_1, X_2, … X_n are variables or constants
- Database operations: logical rules
  - With recursion
- Unified syntax
  - Raw data: Extensional database
  - Deduced results: Intensional database

Example: Call graph edges Predicate vs. Relation



calls(A,B)
calls(A,C)
calls(A,D)
calls(B,D)
calls(C,D)

Predicates

- Calls (x,y): x calls y is true
- Ground atoms:
  predicates with constant arguments

Relations

- Calls (x,y) :
  x, y is in a “calls” relationship
- Extensional database:
  tuples representing facts

Datalog Programs: Set of Rules (Intensional DB)

- H :- B_1 & B_2 … & B_n
- LHS is true if RHS is true
  - Rules define the intensional database
- Example: Datalog program to compute call*
  - transitive closure of calls relation
  - calls*(x, y) if x calls y directly or indirectly
  - calls* (x, y) :- calls (x, y)
  - calls* (x, z) :- calls* (x, y) & calls* (y, z)
- Result:
  - set of ground atoms inferred by applying the rules until no new inferences can be made

Datalog vs. SQL

- SQL
  - Imperative programming:
    - join, union, projection, selection
  - Explicit iteration
- Datalog: logical database language
  - Declarative programming
  - Recursive definition: fixpoint computation
  - Negation can lead to oscillation
  - Stratified: separates rules into groups
    - Compute one group at a time
    - Can negate only the results from previous strata

Datalog vs. Prolog

- Syntactically a subset of Prolog
- No function variables e.g. b in a(b(x,y), c)
- Truly declarative:
  - Rule ordering does not affect program semantics
- Bottom-up evaluation
  - Stratified Datalog always terminates on a finite database

Why use a Deductive Database for Pointer Analysis?

- Pointer analysis produces "intermediate" results to be consumed in analysis.
- Allow query of specific subsets of results
- Analysis as queries
- Results of queries can be further queried in a uniform way

Outline
Pointer Analysis

1. Motivation: security analysis
2. Datalog
3. Context-insensitive, flow-insensitive pointer analysis
4. Context sensitivity

3. Flow-insensitive Points-to Analysis

- Alias analysis:
  - Can two pointers point to the same location?
  - *a, *(a+8)
- Points-to analysis:
  - What objects does each pointer points to?
  - Two pointers cannot be aliased
    if they must point to different objects

How to Name Objects?

- Objects are dynamically allocated
- Use finite names to refer to unbounded # objects
- 1 scheme: Name an object by its allocation site

```
main () {                f () {
    p = f();                A: a = new O ();
    q = f();                B: b = new O ();
}                                return a;
                         }
```


Points-To Analysis for Java

- Variables ($v \in V$):
  - local variables in the program
- Heap-allocated objects ($h \in H$)
  - has a set of fields ($f \in F$)
  - named by allocation site

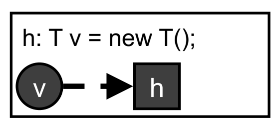
Program Abstraction

- Allocations   h: v = new c
- Store $v_1.f = v_2$
- Loads $v_2 = v_1.f$
- Moves, arguments: $v_1 = v_2$
- Assume: a (conservative) call graph is known a priori
  - Call: formal = actual
  - Return: actual = return value

Pointer Analysis Rules

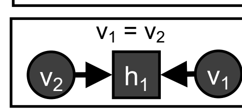
Object creation

pts(v, h) :- "h: T v = new T()".



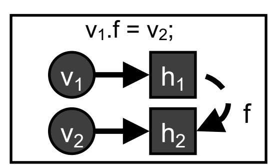
Assignment

pts(v_1, h_1) :- "$v_1 = v_2$" & pts(v_2, h_1).



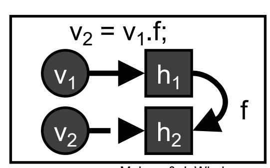
Store

hpts(h_1, f, h_2) :- "$v_1.f = v_2$" &
pts(v_1, h_1) & pts(v_2, h_2).



Load

pts(v_2, h_2) :- "$v_2 = v_1.f$" &
pts(v_1, h_1) & hpts(h_1, f, h_2).



Pointer Alias Analysis

- Specified by a few Datalog rules
  - Creation sites
  - Assignments
  - Stores
  - Loads
- Apply rules until they converge

Example program

```
void main() {
  x = new C();
  y = new C();
  z = new C();
  m(x,y);
  n(z,x);
  q = z.f;
}

void m(C a, C b) {
  n(a,b);
}

void n(C c, C d) {
  c.f = d;
}
```

Pointer Analysis in Datalog

Domains
- V = variables
- H = heap objects
- F = fields

EDB (input) relations
- vP_0 (v:V, h:H) : object allocation sites
- assign(v_1:V, v_2:V) : assignment instructions ($v_1 = v_2$;) and parameter passing
- store (v_1:V, f:F, v_2:V) : store instructions ($v_1.f = v_2$;)
- load (v_1:V, f:F, v_2:V) : load instructions ($v_2 = v_1.f$;)

IDB (computed) relations
- vP (v:V ,h:H) : variable points-to relation (v can point to object h)
- hP (h_1:H, f:F, h_2:H) : heap points-to relation (object h_1 field f can point to h_2)

Rules
- vP (v, h) :- vP_0 (v, h).
- vP (v_1, h) :- assign (v_1, v_2), vP (v_2, h).
- hP (h_1, f, h_2) :- store (v_1, f, v_2), vP (v_1, h_1), vP (v_2, h_2).
- vP (v_2, h_2) :- load (v_1, f, v_2), vP (v_1, h_1), hP(h_1, f, h_2).

Step 1: Assign numbers to elements in domain

```
void main() {
  x = new C();
  y = new C();
  z = new C();
  m(x,y);
  n(z,x);
  q = z.f;
}

void m(C a, C b) {
  n(a,b);
}

void n(C c, C d) {
  c.f = d;
}
```

Domains

V
- 'x' : 0
- 'y' : 1
- 'z' : 2
- 'a' : 3
- 'b' : 4
- 'c' : 5
- 'd' : 6

H
- 'main@1' : 0
- 'main@2' : 1
- 'main@3' : 2

F
- 'f' : 0

Step 2: Extract initial relations (EDB) from program

```
void main() {
  x = new C();
  y = new C();
  z = new C();
  m(x,y);
  n(z,x);
  q = z.f;
}

void m(C a, C b) {
  n(a,b);
}

void n(C c, C d) {
  c.f = d;
}
```

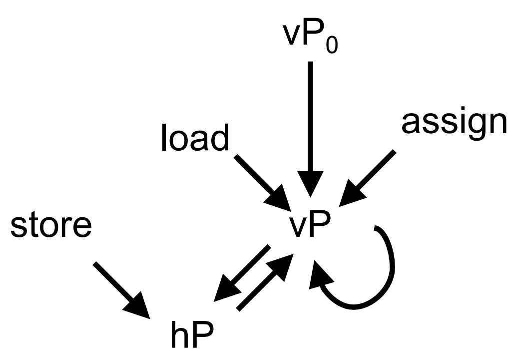
```
vP0('x', 'main@1').
vP0('y', 'main@2').
vP0('z', 'main@3').
assign('a','x').
assign('b','y').
assign('c','z').
assign('d','x').
load('z','f','q').
assign('c','a').
assign('d','b').
store('c','f','d').
```


M. Lam & J. Whaley

Step 3: Generate Predicate Dependency Graph

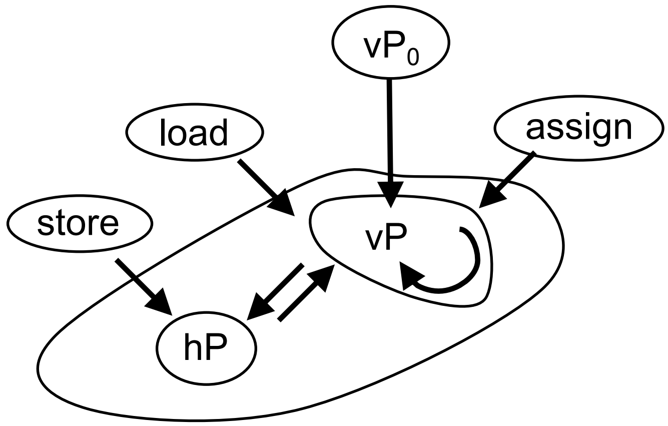
Rules

$vP(v,h)$:- $vP_0(v,h)$.
$vP(v_1,h)$:- $assign(v_1,v_2)$, $vP(v_2,h)$.
$hP(h_1,f,h_2)$:- $store(v_1,f,v_2)$, $vP(v_1,h_1)$, $vP(v_2,h_2)$.
$vP(v_2,h_2)$:- $load(v_1,f,v_2)$, $vP(v_1,h_1)$, $hP(h_1,f,h_2)$.



M. Lam & J. Whaley

Step 4: Determine Iteration Order



Step 5: Apply rules until convergence

Rules

$vP(v,h)$:- $vP_0(v,h)$.
$vP(v_1,h)$:- $assign(v_1,v_2)$, $vP(v_2,h)$.
$hP(h_1,f,h_2)$:- $store(v_1,f,v_2)$, $vP(v_1,h_1)$, $vP(v_2,h_2)$.
$vP(v_2,h_2)$:- $load(v_1,f,v_2)$, $vP(v_1,h_1)$, $hP(h_1,f,h_2)$.

Relations

vP_0	assign	vP	hP
vP_0('x','main@1').	assign('a','x').		
vP_0('y','main@2').	assign('b','y').		
vP_0('z','main@3').	assign('c','z').		
	assign('d','x').		
store	assign('c','a').		
store('c','f','d').	assign('d','b').		
load			
load('z','f','q').			

Step 5: Apply rules until convergence

Rules

$vP(v,h)$:- $vP_0(v,h)$.
$vP(v_1,h)$:- $assign(v_1,v_2)$, $vP(v_2,h)$.
$hP(h_1,f,h_2)$:- $store(v_1,f,v_2)$, $vP(v_1,h_1)$, $vP(v_2,h_2)$.
$vP(v_2,h_2)$:- $load(v_1,f,v_2)$, $vP(v_1,h_1)$, $hP(h_1,f,h_2)$.

Relations

vP_0

vP_0('x','main@1').
vP_0('y','main@2').
vP_0('z','main@3').

store

store('c','f','d').

load

load('z','f','q').

assign

assign('a','x').
assign('b','y').
assign('c','z').
assign('d','x').
assign('c','a').
assign('d','b').

vP

vP('x','main@1').
vP('y','main@2').
vP('z','main@3').

hP

Step 5: Apply rules until convergence

Rules

$vP(v,h)$:- $vP_0(v,h)$.
$vP(v_1,h)$:- $assign(v_1,v_2)$, $vP(v_2,h)$.
$hP(h_1,f,h_2)$:- $store(v_1,f,v_2)$, $vP(v_1,h_1)$, $vP(v_2,h_2)$.
$vP(v_2,h_2)$:- $load(v_1,f,v_2)$, $vP(v_1,h_1)$, $hP(h_1,f,h_2)$.

Relations

vP_0

vP_0('x','main@1').
vP_0('y','main@2').
vP_0('z','main@3').

store

store('c','f','d').

load

load('z','f','q').

assign

assign('a','x').
assign('b','y').
assign('c','z').
assign('d','x').
assign('c','a').
assign('d','b').

vP

vP('x','main@1').
vP('y','main@2').
vP('z','main@3').
vP('a','main@1').
vP('d,'main@1').
vP('b','main@2').
vP('c','main@3').

hP

Step 5: Apply rules until convergence

Rules

$vP(v,h)$:- $vP_0(v,h)$.
$vP(v_1,h)$:- $assign(v_1,v_2)$, $vP(v_2,h)$.
$hP(h_1,f,h_2)$:- $store(v_1,f,v_2)$, $vP(v_1,h_1)$, $vP(v_2,h_2)$.
$vP(v_2,h_2)$:- $load(v_1,f,v_2)$, $vP(v_1,h_1)$, $hP(h_1,f,h_2)$.

Relations

vP_0

vP_0('x','main@1').
vP_0('y','main@2').
vP_0('z','main@3').

store

store('c','f','d').

load

load('z','f','q').

assign

assign('a','x').
assign('b','y').
assign('c','z').
assign('d','x').
assign('c','a').
assign('d','b').

vP

vP('x','main@1').
vP('y','main@2').
vP('z','main@3').
vP('a','main@1').
vP('d,'main@1').
vP('b','main@2').
vP('c','main@3').
vP('c','main@1').
vP('d','main@2').

hP

Step 5: Apply rules until convergence

Rules

$vP(v,h)$:- $vP_0(v,h)$.
$vP(v_1,h)$:- $assign(v_1,v_2)$, $vP(v_2,h)$.
$hP(h_1,f,h_2)$:- $store(v_1,f,v_2)$, $vP(v_1,h_1)$, $vP(v_2,h_2)$.
$vP(v_2,h_2)$:- $load(v_1,f,v_2)$, $vP(v_1,h_1)$, $hP(h_1,f,h_2)$.

Relations

vP_0

vP_0('x','main@1').
vP_0('y','main@2').
vP_0('z','main@3').

store

store('c','f','d').

load

load('z','f','q').

assign

assign('a','x').
assign('b','y').
assign('c','z').
assign('d','x').
assign('c','a').
assign('d','b').

vP

vP('x','main@1').
vP('y','main@2').
vP('z','main@3').
vP('a','main@1').
vP('d,'main@1').
vP('b','main@2').
vP('c','main@3').
vP('c','main@1').
vP('d','main@2').

hP

hP('main@1','f','main@1').
hP('main@1','f','main@2').
hP('main@3','f','main@1').
hP('main@3','f','main@2').

Step 5: Apply rules until convergence

Rules

$vP(v,h)$:- $vP_0(v,h)$.
$vP(v_1,h)$:- $assign(v_1,v_2)$, $vP(v_2,h)$.
$hP(h_1,f,h_2)$:- $store(v_1,f,v_2)$, $vP(v_1,h_1)$, $vP(v_2,h_2)$.
$vP(v_2,h_2)$:- $load(v_1,f,v_2)$, $vP(v_1,h_1)$, $hP(h_1,f,h_2)$.

Relations

vP_0

vP_0('x','main@1').
vP_0('y','main@2').
vP_0('z','main@3').

store

store('c','f','d').

load

load('z','f','q').

assign

assign('a','x').
assign('b','y').
assign('c','z').
assign('d','x').
assign('c','a').
assign('d','b').

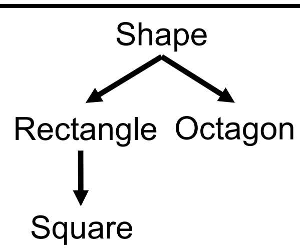
vP

vP('x','main@1').
vP('y','main@2').
vP('z','main@3').
vP('a','main@1').
vP('d,'main@1').
vP('b','main@2').
vP('c','main@3').
vP('c','main@1').
vP('d','main@2').
vP('q','main@1').
vP('q','main@2').

hP

hP('main@1','f','main@1').
hP('main@1','f','main@2').
hP('main@3','f','main@1').
hP('main@3','f','main@2').

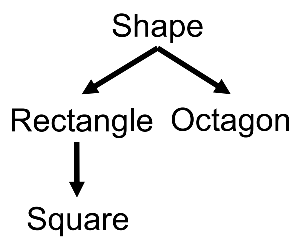
Virtual Method Invocation




```
void draw (shape s)  {
      int i = s.lines();
      …
  }
```

- Class hierarchy analysis cha (t, n, m)
  - Given an invocation v.n (…), if v points to object of type t, then m is the method invoked
  - t's first superclass that defines n

Virtual Method Invocation




```
void draw (shape s)  {
      int i = s.lines();
      …
  }
```

- Class hierarchy analysis cha (t, n, m)
  - Simple analysis: can determine the type if the program only allocates one type of objects.

Pointer Analysis Can Improve Call Graphs

Discover points-to results and methods invoked on the fly

invokes (s, m): statement s calls method m
hType (h, t): h has type t

invokes (s, m) :- "s: v.n (…)" & pts (v, h) &
hType (h, t) & cha (t, n, m)

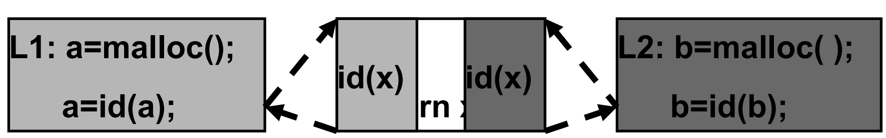
actual (s, i, v): v is the ith actual parameter in call site s.
formal (m, i, v): v is the ith formal parameter declared in method m.

pts(v, h) :- invokes (s, m) &
formal (m, i, v) & actual (s, i, w) &
pts (w, h)

Outline
Pointer Analysis

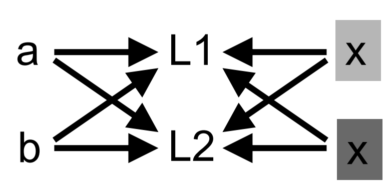
1. Motivation: security analysis
2. Datalog
3. Context-insensitive, flow-insensitive pointer analysis
4. Context sensitivity

4. Context-Sensitive Pointer Analysis

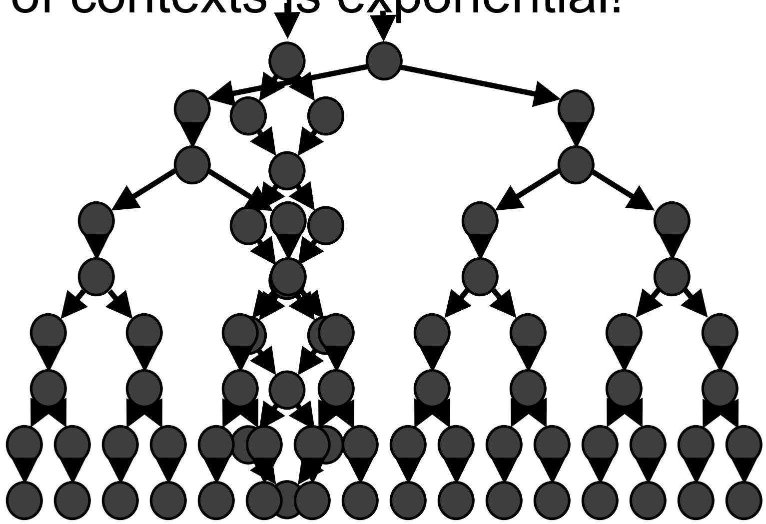


context-sensitive

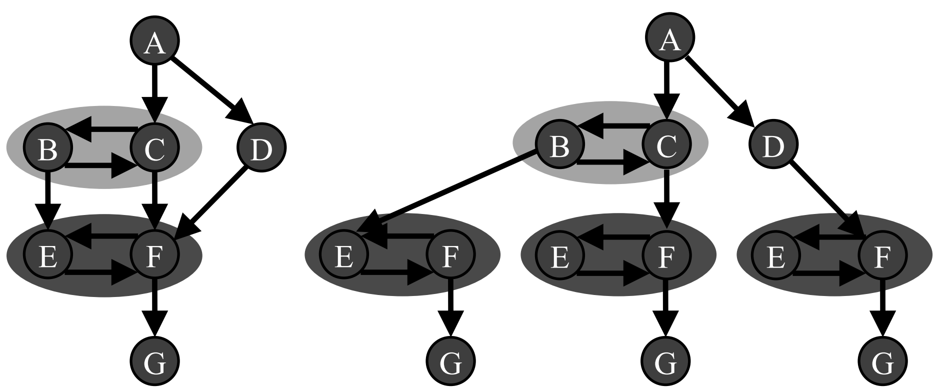
context-insensitive



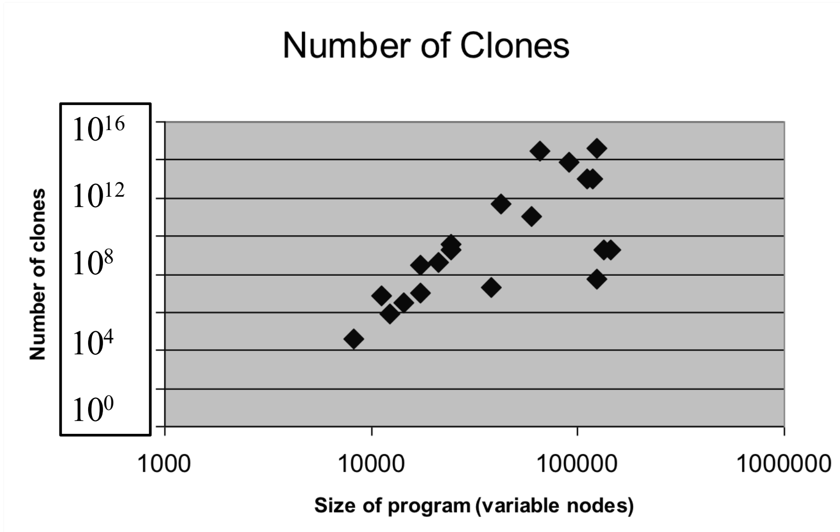
Even without recursion, # of contexts is exponential!



Recursion



Top 20 Sourceforge Java Apps



Cloning-Based Algorithm

- Apply the context-insensitive algorithm to the program to discover the call graph
- Find strongly connected components
- Create a “clone” for every context
- Apply the context-insensitive algorithm to cloned call graph
- Lots of redundancy in result
- Exploit redundancy by clever use of BDDs (binary decision diagrams)

Whaley&Lam, PLDI 2004 (best paper award)

Automatic Analysis Generation

Programmer:
Security analysis
in 10 lines

PQL

Compiler Writer:
Flow-insensitive
Context-sensitive
Ptr analysis in 10 lines

Datalog

bddbddb
(**BDD**-**b**ased
deductive **d**ata**b**ase)
with
Active Machine Learning

1000s of lines
1 year tuning

BDD operations

BDD: 10,000s-lines library

Advanced Compilers

M. Lam & J. Whaley

Goals of the Lecture

- Pointer analysis
  - Interprocedural, context-sensitive, flow-insensitive
    (Dataflow: intraprocedural, flow-sensitive)
- Power of languages and abstractions
- Elegant abstractions
  - Datalog: A deductive database
    (A database that can make deductions from stored data)
  - BDDs: Binary decision diagrams
    (Most cited CS papers for many years)

Advanced Compilers

M. Lam & J. Whaley