DFRWS
DIGITAL FORENSIC RESEARCH CONFERENCE

LogExtractor: Extracting Digital Evidence from Android Log Messages via String and Taint Analysis

By:
Chris Chao-Chun Cheng (Iowa State University), Chen Shi (Iowa State University), Neil Zhenqiang Gong (Duke University), and Yong Guan (Iowa State University)

From the proceedings of

The Digital Forensic Research Conference

DFRWS USA 2021

July 12-15, 2021

LogExtractor: Extracting Digital Evidence from Android Log Messages via String and Taint Analysis

Chris Chao-Chun Cheng[1,3], Chen Shi[1,3], Neil Zhenqiang Gong[2], Yong Guan[1,3]

[1] Iowa State University
[2] Duke University
[3] NIST Center of Excellence in Forensic Science – CSAFE

Acknowledgement: Barbara Guttman and James Lyle (NIST)

DFRWS USA 2021, Virtual

IOWA STATE UNIVERSITY

csafe
Center for Statistics and Applications in Forensic Evidence

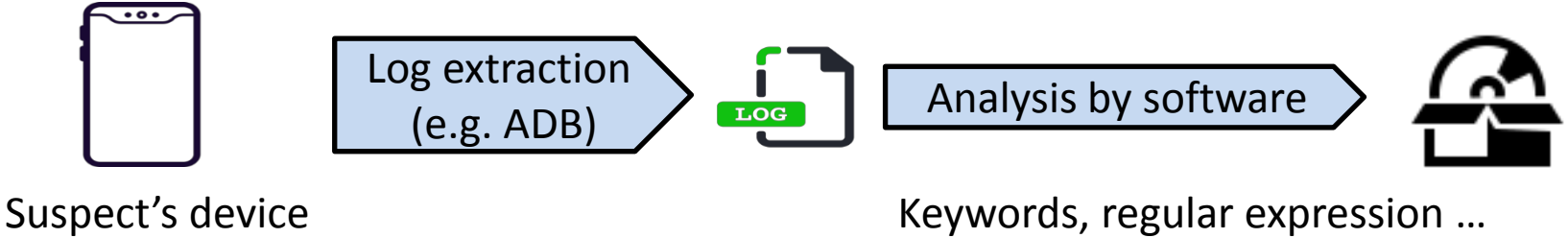
NIST
National Institute of Standards and Technology
Center of Excellence

Evidence from Android Logging System

Source-to-sink Info flows analyzed by R-Droid [AsiaCCS 16’] from 22.7 K apps.

Sink \ Source	Location	Network	Unique Identifier
File	9	37	19
Log	1,676 **(16.1%)**	4,401 **(26.5%)**	3,148 **(42.1%)**
Network	493	1,255 (7.6)	897 (12.2%)

Evidence Identification From Android Logging System

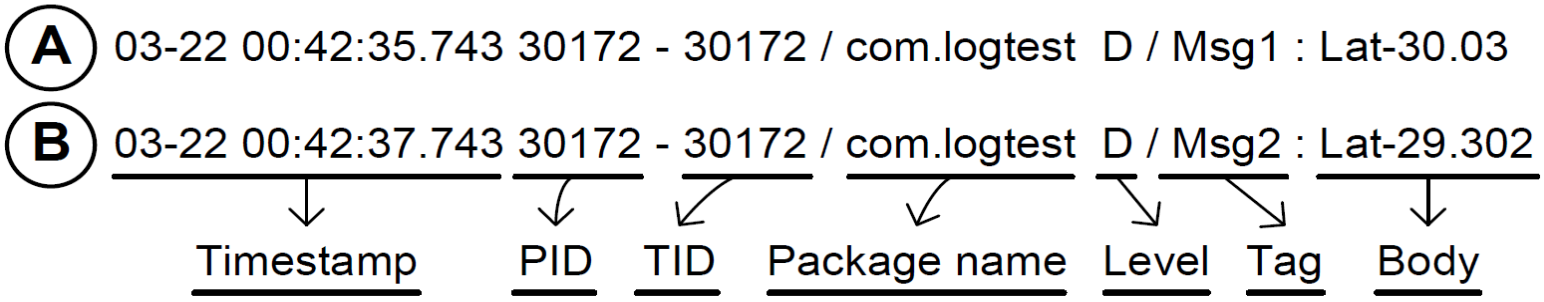


Real-world Challenge of Identifying Evidence from Log

```
04-15 12:54:53.651  1105 11682 I ActivityManager: START u0 {act=android.intent.action.MAIN
cat=[android.intent.category.LAUNCHER] flg=0x10200000 cmp=de.ecspride/.LocationLeak3
bnds=[641,559][843,821]} from uid 10060
04-15 12:54:53.651   740   740 D QCOM PowerHAL: LAUNCH HINT: ON
04-15 12:54:53.654   740   740 D QCOM PowerHAL: Activity launch hint handled
04-15 12:54:53.672 13514 13514 D Location: Location: -50.23456, 42.0283713
04-15 12:54:53.702   740   740 D QCOM PowerHAL: LAUNCH HINT: OFF
04-15 12:54:54.055  1514  1514 I WallpaperService: engine paused
```

- How fast can a forensic analyst find this in 10K+ messages?
- And how to parse these information?
  - Latitude, Longitude
  - Longitude, Latitude

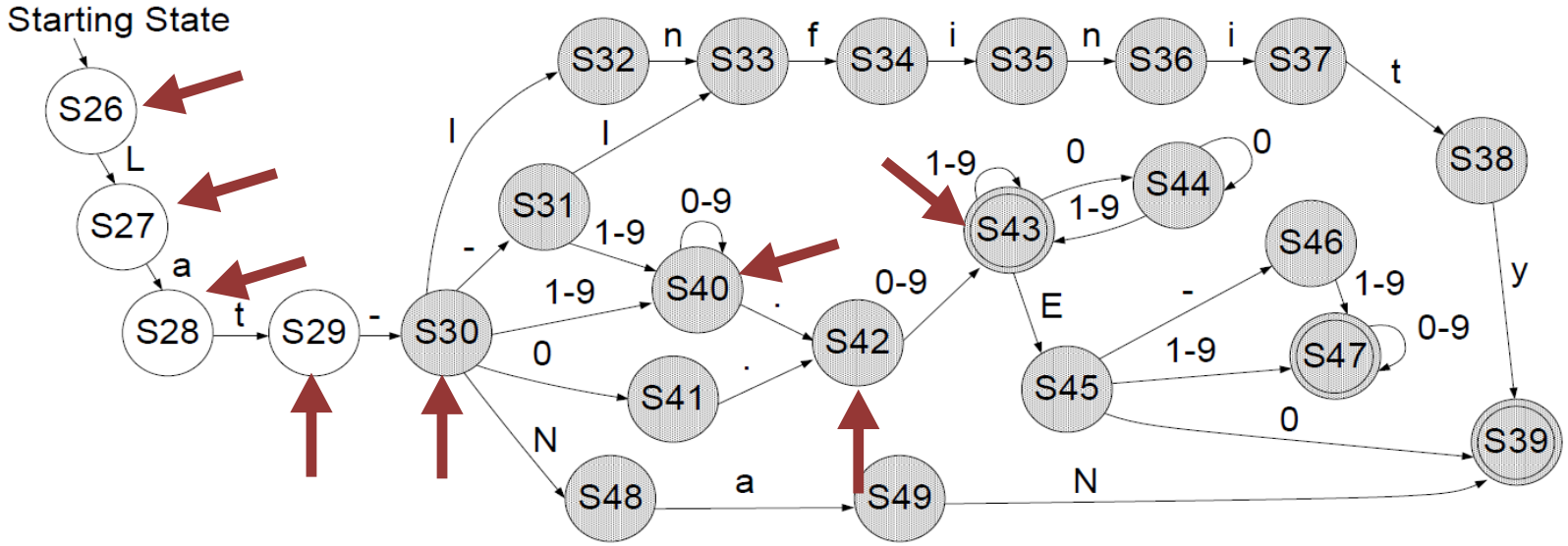
Evidence Identification and Extraction Problem



- Which message contains evidence? What types?
- If the first one contains GPS latitude, is the exact value “30.03” or “-30.03”?

LogExtractor – Automated Evidence Extraction

A 03-22 00:42:35.743 30172 - 30172 / com.logtest D / Msg1 : Lat-30.03



Input String: L a t - 3 0 . 0 3

Latitude: 3 0 . 0 3

Prior Studies

- Taint analysis.
  - FlowDroid [PLDI'14]
  - DroidSafe [NDSS '15]

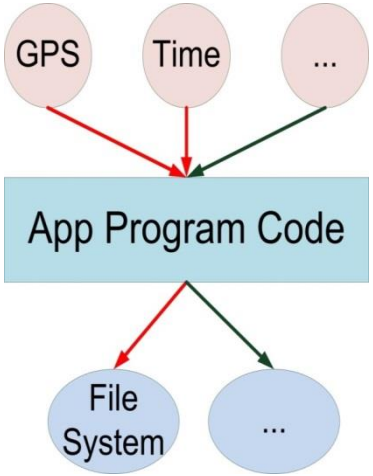
❌ Doesn't support log string patterns.

- String analysis.
  - Java String Analyzer (JSA) [SAS'03]
  - Violist [ESEC/FSE'15]

❌ Doesn't support evidence types tracking.

- Log parsers.
  - Spell [ICDM'16]
  - LKE [ICDM'09]

❌ Doesn't support evidence types tracking.



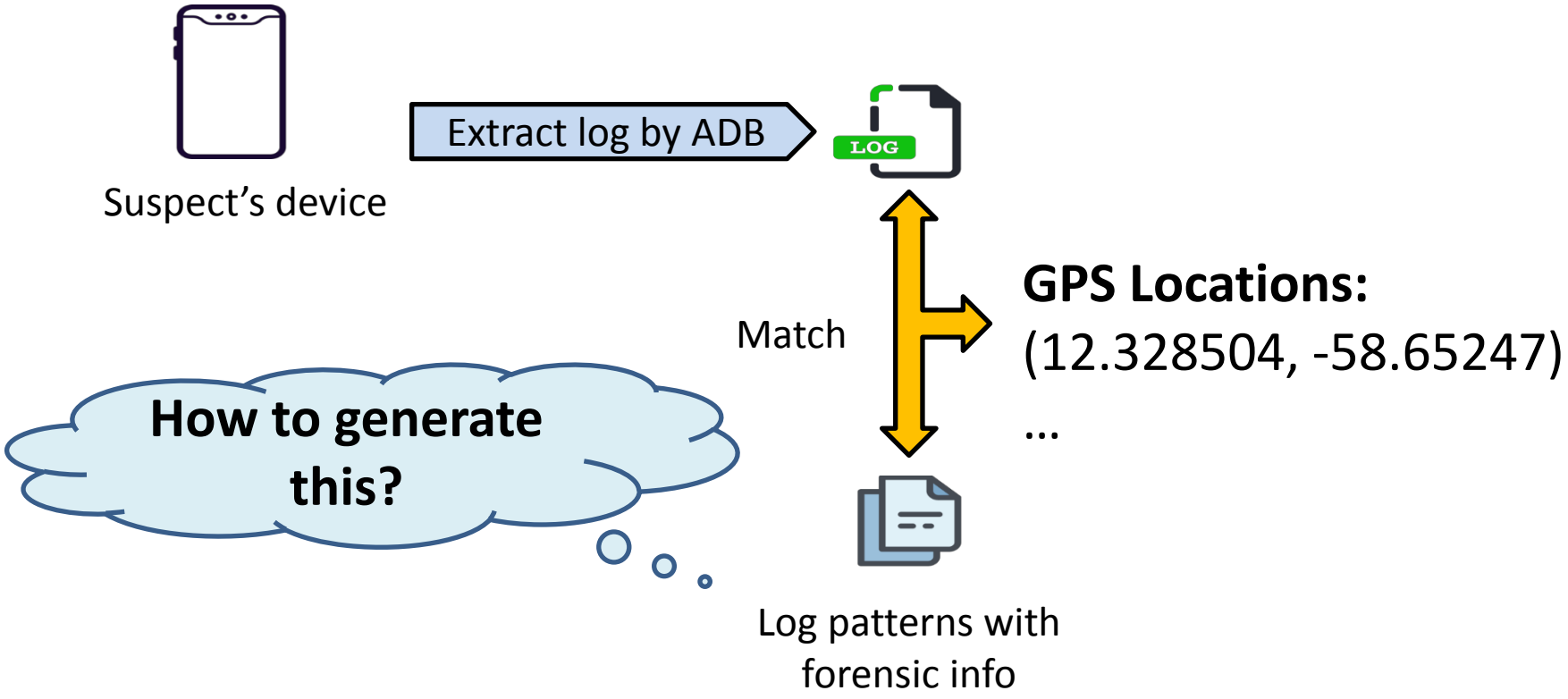
Background – Android Log

- 04-12 05:26:16.143 31735 - 31735 / V / KF Engine: 12.328504x-58.65247

Time | PID-TID | Log lvl | Log tag | Log message

- Time, PID, and, TID are runtime information.
- By *Log.v*(v_1, v_2), the app writes log message with log level V (verbose), where
  - *Log tag* = v_1
  - *Log message* = v_2

LogExtractor- Proposed Approach



App Log Evidence Database

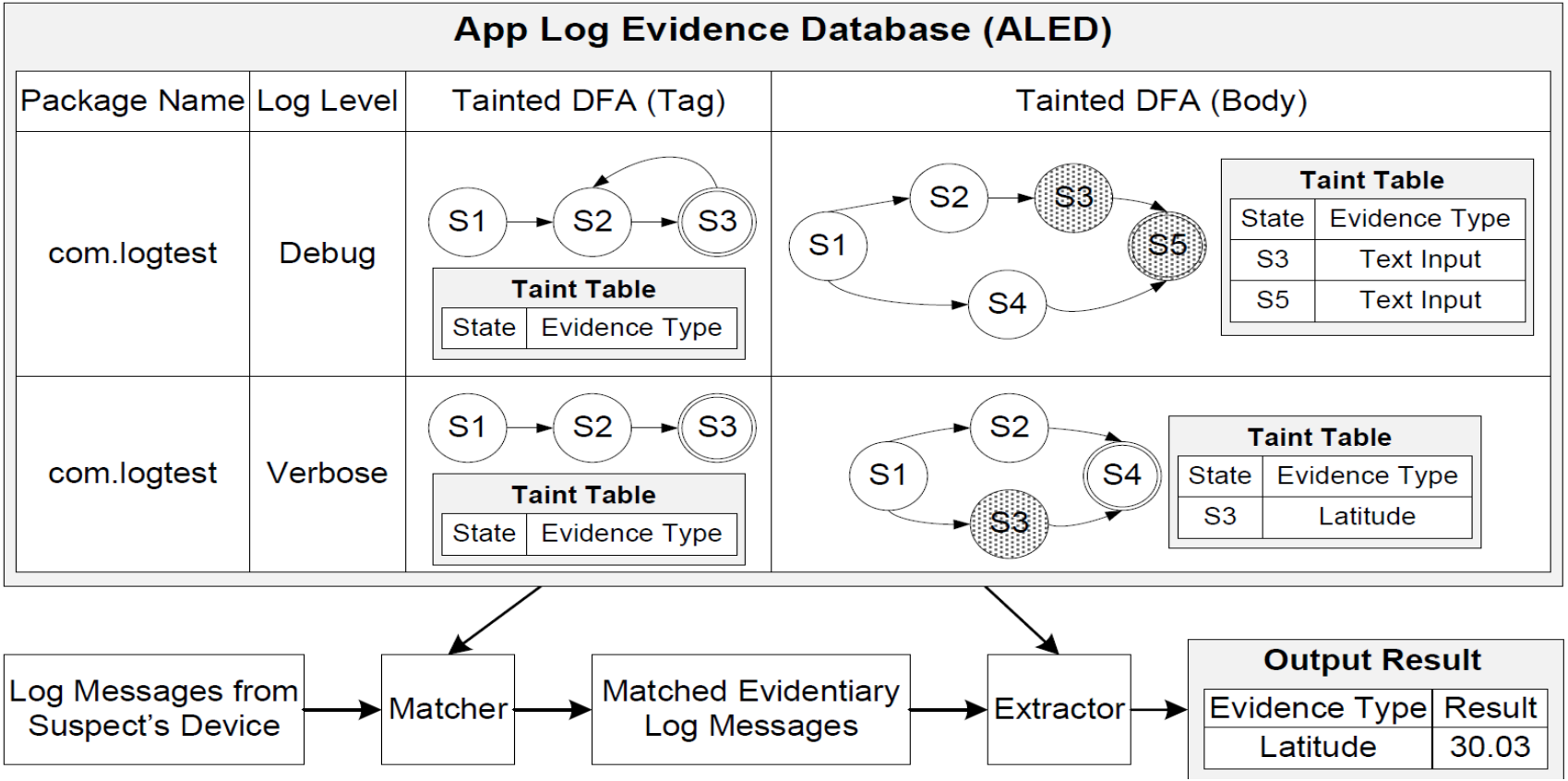
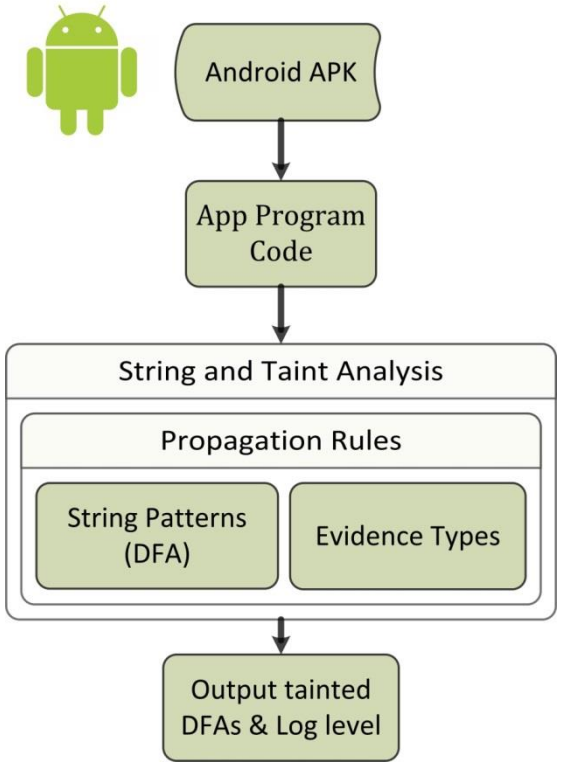


Diagram of LogExtractor



1. Obtain Android Package (APK) file.
2. Extract app code.
3. Perform string and taint analysis.
4. Output when reaching a sink method (logging system).

Example: Log Pattern **“Lat-30.03”**

```
1 void foo(Location loc){
2   String prefix = "Lat-";
3   StringBuffer buff = new StringBuffer(prefix);
4   double lat = loc.getLatitude();
5   buff.append(Double.toString(lat));
6   String toStr = buff.toString();
7   Log.d("Msg1", toStr);
8   Log.d("Msg2", "Lat-29.302");
9 }
```

Lat-

Lat-

<Latitude>

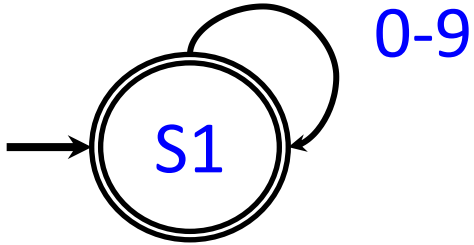
Lat-<Latitude>

Lat-<Latitude>

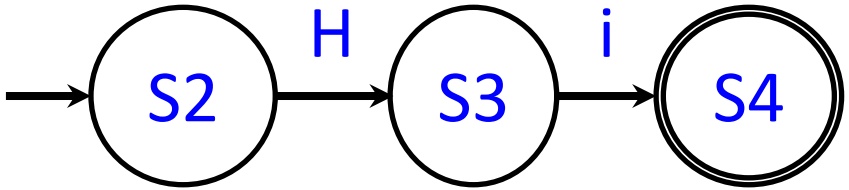
Lat-<Latitude>

Deterministic Finite-State Automaton (DFA)

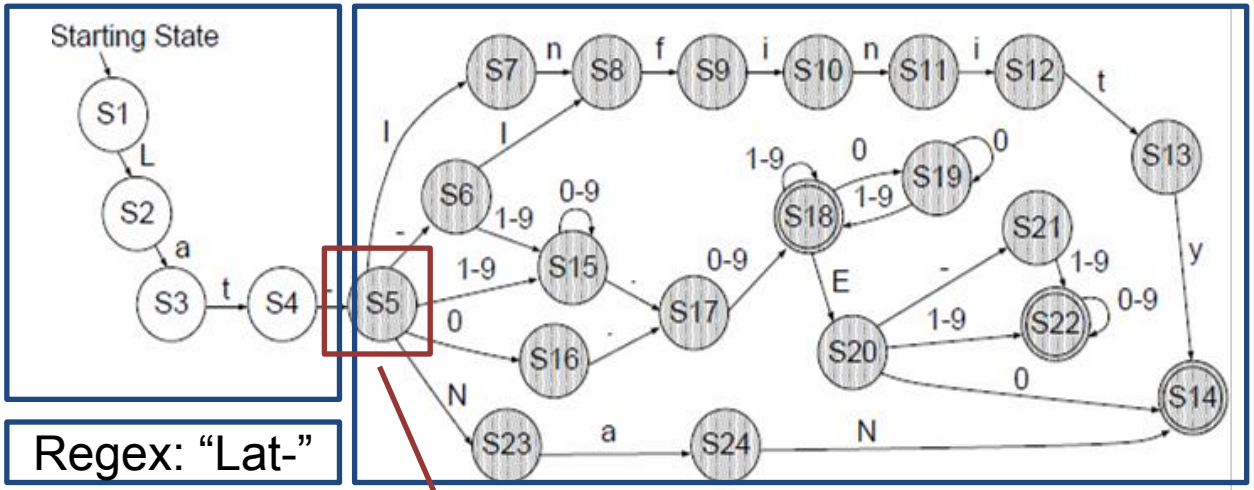
Regular expression: ^[0-9]*$



Regular expression: ^Hi$



LogExtractor – Tainted DFA “Lat-<Latitude>”



Regex: “Lat-”

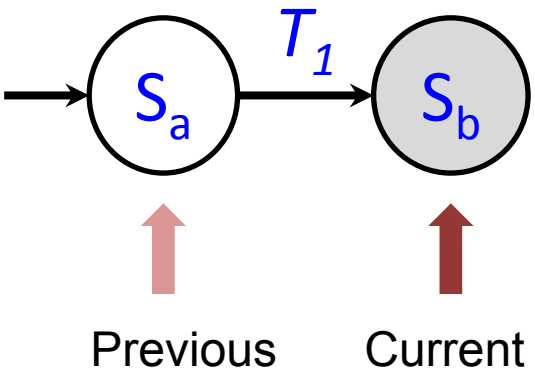
Regex <float num>

Taint Table

State	Evidence Type
S5	{Latitude}
S6	{Latitude}
• • •	
S24	{Latitude}

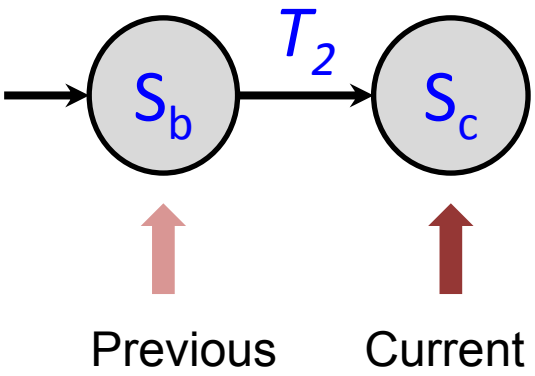
S5-S24

Use Tainted DFA for Evidence Extraction (1)



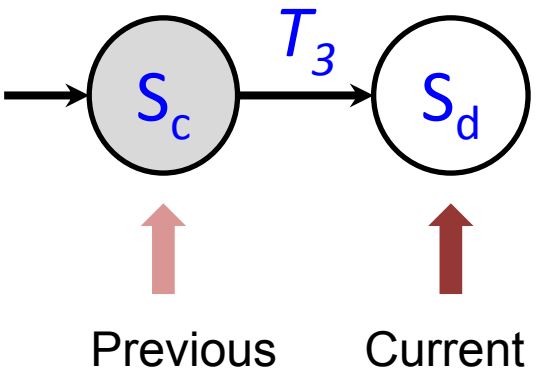
Initialize the output buffer for the evidence type E_{S_b}

Use Tainted DFA for Evidence Extraction (2)



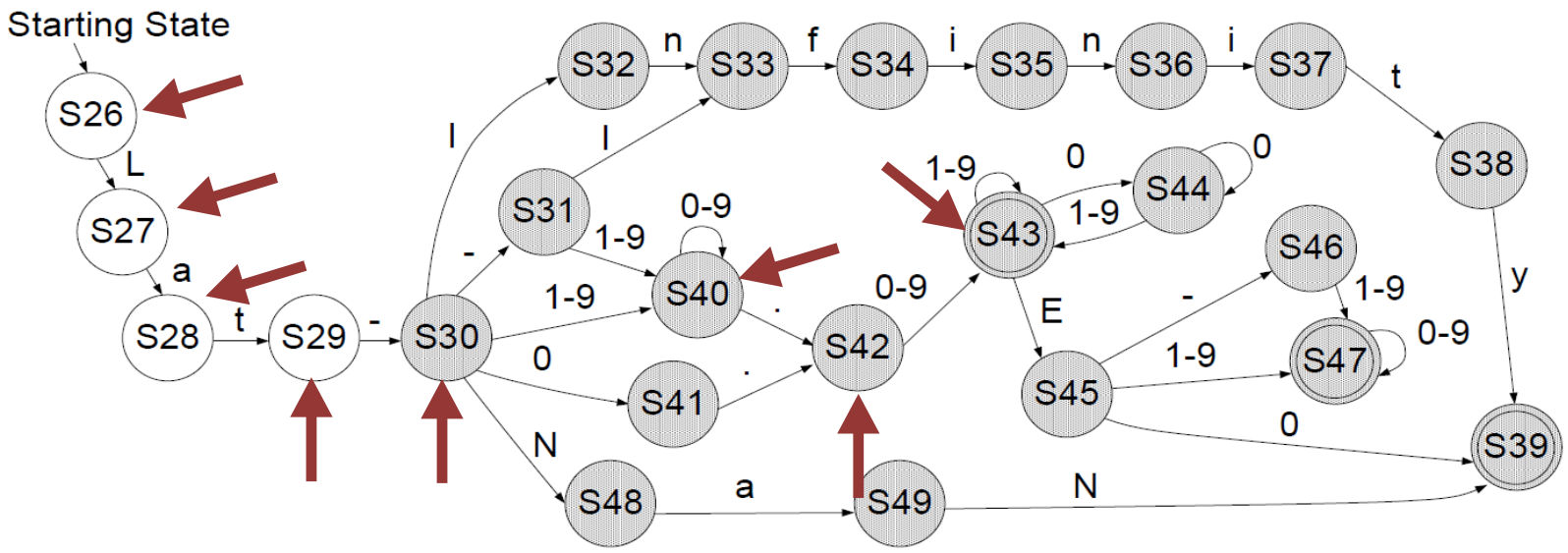
Append T_2 to the output buffer for the evidence type E_{S_b}

Use Tainted DFA for Evidence Extraction (3)



Stop appending text and output string for evidence type E_{S_b}

Extract Evidence by Tainted DFA



Input String: L a t - 3 0 . 0 3

Latitude: 3 0 . 0 3

Evaluation – Benchmark Apps

- DroidBench 65 apps.
  - Identification: precision = 97.7%, recall = 79.2%.
  - Extraction: 100% whenever tainted DFA is correctly built.

- Limitations:
  - Taint analysis: 7 FNs (ICC flows), 1 FN (file I/O), 1 FN (implicit flow), 1 FP (Android model)
  - String analysis: 2 FNs (Formatter, Matcher)

Evaluation – Real-world Apps

- Manual verification on 91 apps:
  - Identification: 86.5% precision and 91.3% recall.
  - Extraction: 100% whenever tainted DFA is correctly built.

- Limitations:
  - Two-dimensional data structure. (JSON object, HashMap).
  - Taint when handling the any string pattern.

Evaluation – GPS Locations in Real-world App Log

When there're **keywords**:

05-12 08:25:12.855 28918 - 28918 / D / Utils : Got Location - lat: 42.0284026, lon: -93.6504905

Latitude Longitude

No **keyword**:

05-12 05:26:16.143 31735 - 31760 / V / KF Engine: calculateTimeTable (valid) @ 42.028404x-93.65049: 71 357 791 1225

Latitude Longitude

No **keyword & reverse sequence**:

05-10 12:45:40.818 10920 – 10920 / E / 測試: hongyan:2=com.new5tv.zc000058,,logextractortest@gmail.com ,Pixel 2,27,8.1.0,null,357537086585379,null,非漫遊,,,,02:00:00:00:00:00,,2019-05-10-12:45:40,-93.65047184,42.02840526,

Longitude Latitude

Conclusion

- Propose ALED to identify evidence on Android logging system.
- Combine Android string and taint analysis to build ALED.
- Future works:
  - Improve time- and space-efficiency of Android program analysis.
  - More precise string models on complex data structures.