Smart Unpacking Research:
Using Mathematics To Unpack More

Benjamin Long

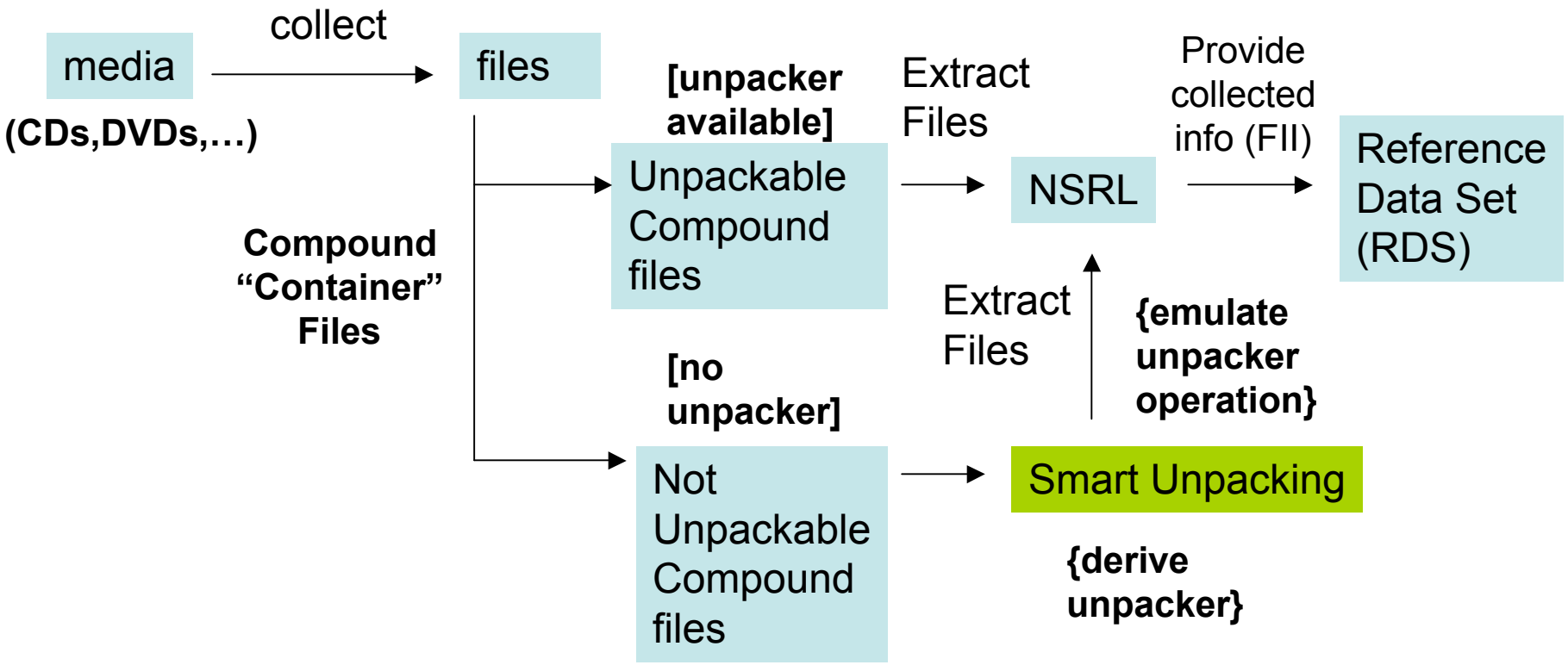
NIST United States Department of Commerce
National Institute of Standards and Technology

Disclaimer

Disclaimer
Trade names and company products are mentioned in the text or identified. In no case does such identification imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the products are necessarily the best available for the purpose.

Statement of Disclosure
This research was funded by the National Institute of Standards and Technology Office of Law Enforcement Standards, the Department of Justice National Institute of Justice, the Federal Bureau of Investigation and the National Archives and Records Administration.

Context – Smart Unpacking and NSRL Operations



Smart Unpacking

- NSRL collects files by unpacking
- New file types appear faster than new unpackers
- NSRL goal: unpack and provide FII for as many files as possible
- SU – attempts to facilitate this goal by
  - Unpacking files when have no unpacker
  - Forming measurements (completeness,accuracy) of unpacking ops

Nature of Unpacking

- Unpacking: locate, identify, and extract files completely and accurately from compound structures
- SU studies and formalizes
  - How to use mathematical methods to find structure in content and nonstandard packaging
  - How to use this to detect, extract, and measure files

SU Research – Early Stages

- To date has –
  - Established basis for theoretical and practical implementation
  - Not yet formalized unpacking systems
  - Implemented specific prototypes and experiments
- Experiments have shown capability to
  - Do basic inline extractions
  - Do similarity and grammatical pattern correlation across many file types (binary and non-binary)

Experiment – Binary Similarity – MSI File Type

```
$ ./cmpB.exe 1.msi.pt 2.msi.pt | more
+ STATUS: analyzing f1([1.msi.pt], size=[2233724]), f2([2.msi.pt], size=[3609563])
0x00000000

a   b       d       f                   k           n
012 067     003     002                 002         002
a   b       d       f                   k
021 080     017     002                 010
 - != same = 201, diff = 181
match = 76.425856, mismatch = 23.574144
        f1: b+z+kzbzbfwzbz+ z+bz+bz+bz+ z+bf+z+ z+bz+bz+bz+          bz+  z+  z+b+z+
+z+b+z+b+z+b+z+  z+  z+  z+  z+  z+  z+  z+b+z+  z+b+z+  z+  z+  z+  z+  z+  z+  z+
+z+b+z+b+z+b+z+b+z+b+z+  z+  z+b+z+
        f2: b+z+kzbzbfwzbz+ z+bz+bz+bz+ z+bf+z+ z+bz+bz+bz+          bz+  z+  z+b+z+
+z+b+z+b+z+b+z+  z+  z+  z+  z+  z+  z+  z+b+z+  z+b+z+  z+  z+  z+  z+  z+  z+  z+
+z+b+z+b+z+b+z+b+z+b+z+  z+  z+b+z+
     match: ............... ........... ....... ...........          ...  ..  ......
..............  ..  ..  ..  ..  ..  ..  ......  ......  ..  ..  ..  ..  ..  ..  ..
......................  ..  ......
```

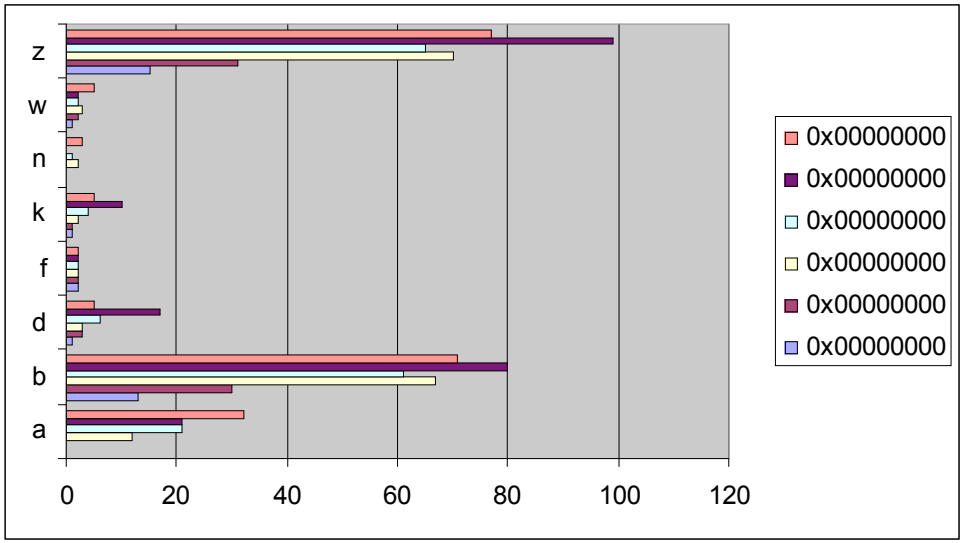
legend	
a	alpha
b	binary
d	delimiter
f	0xff
k	brackets
n	numeric
w	white-space
z	0x00

-Grammatical/location-specific similarity in binary data

-Graph of location-specific content-class frequencies

- basis for preliminary unpacker development

- **metadata types:** file size, location, data size, unpacker version, user/group ids, storage volume used



Technical and Scientific Basis and Methods

Methods used to Discern Structure and Metadata
-Modeling
-Uniqueness features in – syntax, content, abstract pattern representations

Pattern Theory

$$G = \{g_0, g_1, ..., g_i, g_i^0\} = generator\_space$$
$$S : G \longleftrightarrow G, s \in S = similarity\_group$$
$$G = \bigcup_{\alpha \in A} G^{\alpha} = partition\_of\_generators$$
$$b_1, b_2, ..., b_{\omega} = bond\_values$$
$$\mathrm{B}_s(g) = \{b_j; j = 1,2,3,..., w(g)\} = bond\_structures$$
$$B_v(g) = \{\beta j = 1,2,..., w(g)\}$$
$$...$$

Formal Language Theory

Context-free grammars

$$G = (V, \sum, R, S) = grammar$$
$$V = nonter \min als$$
$$\sum = ter \min als$$
$$S = start\_rule$$
$$R = rules$$
$$...$$

Mathematical Modeling based on:

- Information Theory
- Statistics
- Probability
- Universal Algebra
- Topology
- Graph Theory, and more …

Parser Theory

- Variable lookahead mechanisms
- Multi-channel token processors
- Augmentation with syntactic and semantic predicates for context-sensitivity

Measurements and Measurability

- Derived from mathematical models

Spectrum of Unpacking Scenarios

- Simplest – very little knowledge req'd, inline metadata and content; just need an unpacker
- More complex – more complicated structure, nesting, intermingled file, content, and encoding types, alignment considerations (MSI experiment)
- Harder – req's knowledge of metadata and structure relationships; mathematical + experimental models
- Even Harder – at least 1 transform – compressed, encrypted, etc – but still can get metadata and op properties
- Hardest – 2 or more transforms, little to no metadata – will often still get some metadata and specific properties

Unpacking Operation Measurement

- Measuring (at least) – accuracy and completeness of unpacking operation
  - Per file – from a given compound file
  - Across all files/content from a given compound file
- Beginning – focus on simple, coarse-grained ratios – total found/total extracted (files, content)
- Ongoing Objective –
  - develop content and format-specific measures – entropy-, probabilistic-, algebraic-based properties of content and structure
  - Simplify into classes of overall properties per file type

Work From Here

- Integrate prototype tools and experiments into automated unpacking framework
- Research and integrate
  - methods for harder cases
  - Relevant measurements
- Track ability to increase overall coverage of the NSRL

Contact

Benjamin Long
www.nsrl.nist.gov
nsrl@nist.gov

Barbara Guttman
Software Diagnostics & Conformance Testing Division
barbara.guttman@nist.gov

Sue Ballou, Office of Law Enforcement Standards
Rep. For State/Local Law Enforcement
susan.ballou@nist.gov