Considerations and Pitfalls for Conducting Intrusion Detection Research

Vern Paxson

International Computer Science Institute and
Lawrence Berkeley National Laboratory
Berkeley, California USA

vern@icsi.berkeley.edu

July 12, 2007

Outline

- Perspectives & biases
- Nature of the research domain
- Pitfalls & considerations for problem selection
- Pitfalls & considerations for assessment
- Summary

Perspectives

- Worked in intrusion detection since 1994
  - Came into field by accident (from network meas.)
- 20+ security program committees
  - Chaired/co-chaired USENIX Security, IEEE S&P
  - 400+ reviews
    - (Many repeated mistakes!)
- Much work in the field lacks soundness or adequate generality
  - Some of the sharpest examples come from rejected submissions, so this talk light on "naming names"

Biases

- **Network** intrusion detection rather than **host**-based
  - This is simply a bias in emphasis
- Empiricism rather than theory
  - … But I’m going to argue this is correct!
- Primary author of the “Bro” network intrusion detection system
  - … But even if I weren’t, I’d still trash Snort!

Problematic Nature of the Research Domain

- Intrusion detection spans very wide range of activity, applications, semantics
- Much is **bolt-on** / **reactive**
  - Solutions often lack completeness / coherence
  - Greatly increases evasion opportunities
- Problem space is inherently adversarial
  - Rapid *evolution*
  - Increasingly complex *semantics*
  - *Commercialization* of malware is accelerating pace

The Research Process

1) Problem selection
2) Development of technique
3) Assessment
4) Iteration of these last two

The Research Process

1) Problem selection

2) Development of technique

3) Assessment

4) Iteration of these last two

Pitfalls for Problem Selection

- Research fundamental: understanding the **state-of-the-art**
- Pitfall: coming to intrusion detection from another domain, especially:
  - Machine learning
  - Hardware
  - Mathematical/statistical modeling …

⇒ Due to field's rapid innovation, very easy to underestimate evolution of the problem domain

Coming From Machine Learning:

- Pitfall:
  Showing that a new ML technique performs somewhat better than a previous one against a particular dataset = *Exceeding Slim Contribution* (**ESC**)
  - Proof: see below
- What's instead required:
  Develop a technique that
    - Exhibits broad applicability …
    - … and conveys insight into its power & limitations

Coming From Machine Learning, con't

- General problem (R. Sommer):
  Much of classical ML focuses on understanding
  - The **common** cases …
  - … for which classification errors **aren't costly**

- For intrusion detection, we generally want to find
  - Outliers ….
  - … for which classification errors cost us either in vulnerability or in wasted analyst time

Coming From Hardware:

- Pitfall:
  More quickly/efficiently matching sets of strings / regular expressions / ACLs = **ESC**
  - (Especially if done for Snort - see below)
- What's instead required:
  Hardware in support of *deep packet inspection*
  - Application-level analysis
    - **Not**: transport-level (byte stream w/o app. semantics)
    - **Certainly not**: network-level (per-packet)
  - Correlation across flows or activity

Coming From Modeling:

- Pitfall:

  Refining models for worm propagation = **ESC**

  - Particularly given published results on different, more efficient propagation schemes

- What's instead required:

  Modeling that *changes perception* of how to deal with particular threats

  - Operational relevance (see below)

  Modeling that provides insight into tuning, FP/FN tradeoffs, detection speed

Commercial Approaches vs. Research

- Legitimate concern for problem selection:
  Is it interesting research if commercial vendors already do it?
    - Not infrequent concern for field due to combination of (1) heavy commercialization + (2) heavy competition = diminished insight into vendor technology
- Response:
  Yes, there is significant value to exploring technology in open literature
- Valuable to also frame *apparent* state of commercial practice

Problem Selection: Snort is *not* State-of-the-art

- NIDS problem space long ago evolved beyond **per-packet analysis**
- NIDS problem space long ago evolved beyond **reassembled stream analysis**
- Key conceptual difference: syntax versus semantics
  - Analyzing semantics requires parsing & (lots of) state
  - … but is **crucial** for (1) much more powerful analysis and (2) resisting many forms of evasion
- Snort ≈ syntax
  - ⇒ Research built on it fundamentally limited

Problem Selection & Operational Relevance

- Whole point of intrusion detection: work in the Real World
- Vital to consider how security works in practice. E.g.:
- Threat model
  - Pitfall: worst-case attack scenarios with attacker resources / goals outside the threat model
- Available inputs
  - Pitfall: correlation schemes assuming ubiquitous sensors or perfect low-level detection
  - Pitfall: neglecting aliasing (DHCP/NAT) and churn
  - Pitfall: assuming a single-choke-point perimeter

Operational Relevance, con't

- The need for actionable decisions:
  - False positives ⇒ *collateral damage*
- Analyst burden:
  - E.g., honeypot activity stimulates alarms elsewhere; FPs
- Management considerations:
  - E.g., endpoint deployment is expensive
  - E.g., navigating logs, investigating alarms is expensive

Operational Relevance, con’t

- Legal & business concerns:
  - E.g., data sharing
- Granularity of operational procedures:
  - E.g., disk wipe for rooted boxes vs. scheme to enumerate altered files, but w/ some errors

- These concerns aren’t necessarily “*deal breakers*” …
  - … but can significantly affect research “**heft**”

The Research Process

1) Problem selection

2) Development of technique

3) Assessment

4) Iteration of these last two

Development of Technique

- Pitfall: failing to separate data used for development/analysis/training from data for assessment
  - Important to keep in mind the process is iterative
- Pitfall: failing to separate out the contribution of different components
- Pitfall: failing to understand range/relevance of parameter space

- Note: all of these are <u>standard</u> for research in general
  - Not intrusion-detection specific

The Research Process

1) Problem selection
2) Development of technique
3) Assessment
4) Iteration of these last two

Assessment Considerations

- Experimental design
  - Pitfall: user studies
- Acquiring & dealing with data
- Tuning / training
- False positives & negatives (also **true** +/-'s!)
- Resource requirements
- Decision speed
  - Fast enough for intrusion prevention?
- … Evasion & evolution

Assessment - The Difficulties of Data

- Arguably most significant challenge field faces
  - Very few public resources ….
  - …. due to issues of legality/privacy/security
- Problem #1: lack of **diversity** / **scale**
  - Pitfall: using data measured in own CS lab
    - Nothing tells you this isn’t sufficently diverse!
  - Pitfall: using simulation
    - See *Difficulties in Simulating the Internet*, Floyd/Paxson, IEEE/ACM Transactions on Networking, 9(4), 2001
  - Hurdle: the problem of “crud” …

1 day of “crud” seen at ICSI (155K times)

active-connection-reuse	DNS-label-len-gt-pkt	HTTP-chunked-multipart	possible-split-routing
bad-Ident-reply	DNS-label-too-long	HTTP-version-mismatch	SYN-after-close
bad-RPC	DNS-RR-length-mismatch	illegal-%-at-end-of-URI	SYN-after-reset
bad-SYN-ack	DNS-RR-unknown-type	inappropriate-FIN	SYN-inside-connection
bad-TCP-header-len	DNS-truncated-answer	IRC-invalid-line	SYN-seq-jump
base64-illegal-encoding	DNS-len-lt-hdr-len	line-terminated-with-single-CR	truncated-NTP
connection-originator-SYN-ack	DNS-truncated-RR-rdlength	malformed-SSH-identification	unescaped-%-in-URI
data-after-reset	double-%-in-URI	no-login-prompt	unescaped-special-URI-char
data-before-established	excess-RPC	NUL-in-line	unmatched-HTTP-reply
too-many-DNS-queries	FIN-advanced-last-seq	POP3-server-sending-client-commands	window-recision
DNS-label-forward-compress-offset	fragment-with-DF		

The Difficulties of Data, con't

- Problem #2: stale data
  - Today's attacks often greatly differ from 5 years ago
  - Pitfall: Lincoln Labs / KDD Cup datasets (as we'll see)
- Problem #3: failing to tell us about the data
  - Quality of data? Ground truth? Meta-data?
  - Measurement errors & artifacts?
    - *How do you know*? (calibration)
  - Presence of noise
    - Internal scanners, honeypots, infections
    - "*Background radiation*"
  - **Frame the limitations**

The KDD Cup Pitfall / *Vortex*

- Lincoln Labs DARPA datasets (1998, 1999)
  - Traces of activity, including attacks, on hypothetical air force base
  - Virtually the **only** public, labeled intrusion datasets
- Major caveats
  - Synthetic
    - Unrelated artifacts, little “crud”
  - Old!
  - Overstudied! (answers known in advance)
- Fundamental: ***Testing Intrusion detection systems: A critique of the 1998 and 1999 DARPA intrusion detection system evaluations as performed by Lincoln Laboratory***, John McHugh, ACM Transactions on Information and System Security 3(4), 2000

KDD Cup Pitfall / *Vortex*, con't

- KDD Cup dataset (1999)
  - Distillation of Lincoln Labs 1998 dataset into features for machine learning
  - Used in competition for evaluating ML approaches
- Fundamental problem #1
- Fundamental problem #2
  - There is nothing "holy" about the features
    - And in fact some things unholy ("tells")
  - *Even more over-studied than Lincoln Labs*
  - See ***An Analysis of the 1999 DARPA/Lincoln Laboratory Evaluation Data for Network Anomaly Detection***, Mahoney & Chan, Proc. RAID 2003

KDD Cup Pitfall / *Vortex*, con't

- Data remains a magnet for ML assessment

- All that the datasets are good for:
  - Test for "showstopper" flaws in your approach
  - **Cannot** provide insight into utility, correctness

Assessment - Tuning & Training

- Many schemes require “fitting” of parameters (tuning) or profiles (training) to operational environment
- Assessing significance requires multiple datasets
  - Both for initial development/testing …
  - … and to see behavior under range of conditions
  - Can often sub-divide datasets towards this end
    - But do so **in advance** to avoid bias
- Longitudinal assessment:
  - If you tune/train, for how long does it remain effective?

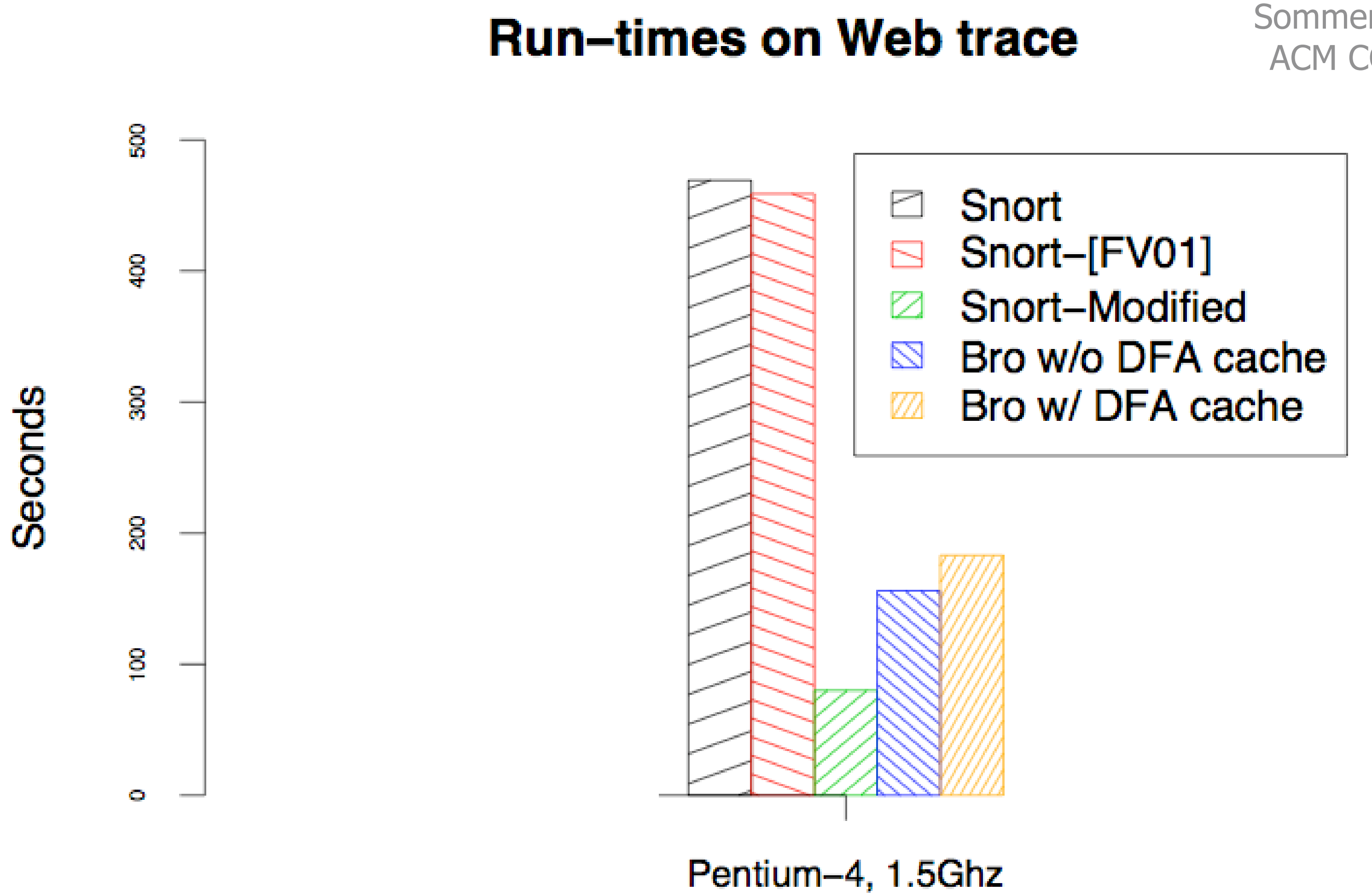
General Tuning/Training Considerations

- Very large benefit to *minimizing parameters*
  - In addition, if training required then tolerating noisy data
- When comparing against other schemes, crucial to assess whether you fairly tuned them too
- General technique: assess **range** of parameters / training rather than a single instance

- Even so, comparisons can exhibit striking variability …

Performance Comparison Pitfall …

Sommer/Paxson, ACM CCS 2003



Snort gets worse on P4, Bro gets better - *which is "correct"*?

If we hadn't tried two different systems, we never would have known …

Assessment - False Positives & Negatives

- FP/FN tradeoff is of **fundamental** interest
- FPs can often be assessed via manual inspection
  - For large numbers of detections, can employ random sampling
- FNs more problematic
  - Inject some and look for them
  - Find them by some other means
    - e.g., simple brute-force algorithm
  - Somehow acquire labeled data
- Common pitfall (esp. for machine learning):
  - For both, need to analyze **why** they occurred

False Positives & Negatives, con't

- For "opaque" algorithms (e.g., ML) need to also assess why **true** positives & negatives occur!
  - What does it mean that a feature exhibits power?
- Key operational concern: is detection actionable?
  - Fundamental: *The Base-Rate Fallacy and its Implications for the Difficulty of Intrusion Detection*, S. Axelsson, Proc. ACM CCS 1999
    - E.g., FP rate of 10^{-6} with 50M events/day ⇒ 50 FPs/day
  - Particularly problematic for anomaly detection
- If not actionable, can still aim to:
  - Provide *high-quality information* to analyst
  - *Aggregate* multiple signals into something actionable

Assessment - Evasion

- One form of evasion: *incompleteness*
  - E.g., your HTTP analyzer doesn't understand Unicode
    - There are a zillion of these, so a pain for research
    - But important for operation …
- Another (thorny) form: *fundamental ambiguity*
  - Consider the following attack URL:

    http://…./c/winnt/system32/cmd.exe?/c+dir
  - Easy to scan for (e.g., "cmd.exe"), right?

Fundamental Ambiguity, con’t

- But what about

  http://…./c/winnt/system32/cm%64.exe?/c+dir

- Okay, we need to handle % escapes.

  (%64=‘d’)

- But what about

  http://…./c/winnt/system32/cm%25%54%52.exe?/c+dir

- Oops. Will server double-expand escapes … or not?
  - %25=‘%’   %54=‘6’ %52=‘4’

Assessment - Evasion, con't

- Reviewers generally recognize that a spectrum of evasions exists …
- … rather than ignoring these, you are better off identifying possible evasions and reasoning about:
  - Difficulty for attacker to exploit them
  - Difficulty for defender to fix them
  - *Likely evolution*
- Operational experience: there's a lot of utility in "*raising the bar*"
- <u>However</u>: if your scheme allows for easy evasion, or plausible threat model indicates attackers will undermine ….
  - …. then you may be in trouble

Assessment - General Considerations

- Fundamental question: what **insight** does the assessment illuminate for the approach?
  - Pitfall: this is especially often neglected for ML and anomaly detection studies …
  - Note: often the features that work well for these approaches can then be directly coded for, rather than indirectly
    - I.e., consider ML as a *tool* for developing an approach, rather than a final scheme

- Fundamental question: where do things break?
  - **And why**?

Summary of Pitfalls / Considerations

- Select an **apt** problem
  - State-of-the-art
  - Aligned with operational practices
  - Avoid ESCs! (Exceedingly Slim Contributions)
- **Beware** KDD Cup! ……. **Beware** Snort!
- Obtain *realistic*, *diverse* data
  - And tell us its properties
- What's the *range of operation*?
  - And accompanying trade-offs?
- How do the false positives **scale**?
  - How do you have confidence in the false negatives?
- What's the **insight** we draw from the assessment?