A Network Coding Approach to IP Traceback

Pegah Sattari, Minas Gjokas, Athina Markopoulou

EECS, UC Irvine
Outline

- Background on Traceback
- Main idea PPM+NC
- Practical PPM+NC
- Simulation Results
- Conclusion and future work
Where is malicious traffic coming from?

Goal: traceback source and path of attack
Prior Work on Traceback

- Early ideas [Burch and Cheswick 1999]
- Send specialized (ICMP) packets [Bellovin et al. 2001]
- Routers keep logs of all packets [Snoeren et al. 2001] ...
- Packet Marking
  - routers mark packets with information about their ID, victim uses the marks of several packets to reconstruct path
  - [Savage et al. 2001]: probabilistically mark fragments of IP addresses
  - Authentication + hashing [Song et al. 2001], [Yaar et al. 05], adjusting marking probability, ...
- Algebraic Traceback
  - [Dean et al. 2002]: encodes the information of n routers on the attack path as coefficients of a polynomial of degree n-1.
  - [Das et al. 2010]: tracks changes in a single path, network coding
- Information theoretical [Adler 2002]
  - studied the tradeoff of #bits vs. #packets
Traceback
via Probabilistic Packet Marking (PPM)
Outline

- Background on Traceback
- Main idea
  - Problem statement
  - PPM+NC
- Practical PPM+NC
- Simulation Results
- Conclusion and future work
Probabilistic Packet Marking (PPM):

- Routers probabilistically mark packets with (partial) information about their address.
- The goal of PPM is to enable the victim to recover d router IDs after receiving a sufficient number of packets.
- PPM+NC tries to achieve the same goal with a smaller number of packets, by appropriately choosing the marking scheme at intermediate routers.
Main Idea

PPM+NC

- PPM is essentially a coupon collector’s problem
  - Collect all router ids \( \{R_d, R_{d-1}, ..., R_2, R_1\} \)
  - A coupon collector’s problem with unequal probabilities:
    * The further a router is from the victim, the less likely that its mark will not be overwritten as the packet moves along the path.

\[
E[X_{PPM}] = \int_0^\infty (1 - \prod_{i=1}^d (1 - e^{-p(1-p)^{i-1}x}))dx
\]

- NC helps the coupon collector problem:
  - NC increases the chance of getting an innovative coupon
  - equally likely coupons: \( E[X] \) reduces from \( \Theta(d \log d) \) to \( \Theta(d) \)
**Main Idea**

**PPM+NC cont’d**

- **Router i:**
  - instead of marking with its own id “$R_i$”, picks a random coefficient “$c_i$”, and adds $c_i \cdot R_i$ to the existing mark.

- **Victim:**
  - instead of ids themselves, it receives random linear combinations of router ids ($\sum c_i \cdot R_i$):
  - solves a system of equations and find the ids.
Main Idea
PPM+NC for a single path

Setup:
• path length $d=1...31$, field $F_4$, $p=1/25$, 500 realizations.
• Metric of interest: number of marks $X$ needed to reconstruct the attack path

Observations:
• $E[X_{PPM+NC}] < E[X_{PPM}]$
• Models perfectly agree with simulation
Main Idea
Multiple-path scenario as the union of multiple paths

- Typically DDoS attacks is distributed:

- The attack path from \{A_i\} is the ordered list of routers between \{A_i\} and V that the attack packet has gone through.
Outline

- DDoS and Traceback
- Main idea
- Practical PPM+NC
  - Practical constraints
  - Marking procedure
  - Reconstruction procedure
  - Processing costs
- Simulation results
- Conclusion and future work
Practical PPM+NC

Practical Constraints

- Limited number of bits (16 ID + 1 flag = 17)
  - Mark with Fragments of IP addresses
  - $f=4$ fragments (of 8 bits each), 2-bit fragment offset, $k=3$ coefficients, of $b=2$ bits each, distance=1 bit. Total: 17 bits.
  - 8 bits used for the linear combination, 2 bits for the coefficients.

\[
\left\lfloor \frac{32}{f} \right\rfloor + \lfloor \log_2 f \rfloor + k \cdot b + \text{distance} \leq \text{bit budget}
\]

- Spoofing by the attacker
  - Probabilistically overwrite the previous mark
  - Distance field (approximate traceback)

- Identifying nodes vs. reconstructing the attack graph
  - Distance field
  - Markings from consecutive routers
Practical PPM+NC
Marking Procedure

- Each router probabilistically decides whether to overwrite or not.
- If overwrite:
  - zero out the field+ mark with a fragment of the router ID.
- If not_overwrite & there is space:
  - add to the combination of the same fragment
  - increase distance field
Practical PPM+NC
Tradeoff in the packet header

- \( \sum c_i R_{i,j} \) 

- Linear combination of larger fragments.

- A linear combination of as many fragments of IP addresses as possible (random coefficients).

- Always an optimal \( k \) minimizes \#packets. For bit budget 17, it is \( k = 3 \) (our selection).

- \( R_{i,j} \): The \( j^{th} \) fragment of \( R_i \).

- We want both parts to be as large as possible:
  - A linear combination of larger fragments.
  - A linear combination of as many fragments of IP addresses as possible (random coefficients).
Best choice: 8 bits for fragments (f=4), 2 bits for fragment offset, 3 coefficients (k=3), of 2 bits each (b=2), 1 bit for distance.

17 bits in total, within the bit-budget.
Once the victim receives the packet P, it forms:
\[ c_L \cdot R_L^j + c_{L-1} \cdot R_{L-1}^j + c_{L-2} \cdot R_{L-2}^j = P \cdot \text{linearCombination} \]

The unknowns are the fragments of the IP addresses:
\[ R_i^j, \ i=1\ldots d, \ j=1\ldots f \]

The victim can solve the system of linear equations after receiving \( d \cdot f \) innovative packets.

Use fragment offset to order fragments of same router ID (same distance).

Path consists of router IDs ordered by distance.
**Practical PPM+NC**

**Reconstruction Procedure, cont’d**

- **Multiple-paths:**
  - Multiple routers at the same distance from the victim.
  - Need to distinguish equations coming from different paths.

- E.g., victim receives 2 packets from distance=4
- One from $R_8, R_4, R_2$, the other from $R_{15}, R_7, R_3$
- Do they belong to the same triplet or not?!
Practical PPM+NC
Reconstruction Procedure, cont’d

- Two solutions:
  1. Use 8 bits (TOS field) to store a checksum that helps identify a triplet of marking routers
     - E.g., each router pre-computes a hash of its IP address
     - The less bits we use, the larger the probability of collision
  2. Assume the victim has knowledge of the map of its upstream routers [Song et al., Yaar et al.]
     - Given the distance value, fragment offset, and random coefficients, the victim tries all possible triplets in the map and picks the one that matches.
     - Does not even solve a system of linear equations
Practical PPM+NC

Cost

- Benefit of the PPM+NC approach
  - Reconstruct the paths after receiving a smaller number of marked packets

- Cost of PM+NC approach:
  - increased computational complexity and processing time.

- Need to generate more random numbers,
  - both for the marking decision and for the random coefficients:
    - only when there is space
    - can be pre-computed and used for all packets

- Routers need to compute linear combinations in $F_{256}$
  - can be done quickly using a transition (log) table

- Victim needs to solve a system of linear equations or to try addresses against a given linear combination
Outline

- DDoS and Traceback
- Main idea
- Practical PPM+NC
- Simulation Results
- Conclusion and future work
Simulation Results
paths vs. trees

- Single path, d=1...31
- Binary tree, 3...127 nodes

- Fair comparison against modified FMS [Savage et al. 2001], such that it uses 17bits + TTL-based distance.
- p=1/25, 500 realizations
Simulation Results

power-law graphs

Setup:
- BRITE topology generator
- Router-only mode, GLP model, preferential connectivity, incremental growth, random node placement.
- #links added per new node = 2
- generated a 150 node graph, extracted a tree out of it, and tried different #attackers.
- p = 1/25, 500 realizations.
Outline

- DDoS and Traceback
- Main idea
  - Problem statement
  - PPM+NC
- Practical PPM+NC
  - Practical constraints
  - Marking procedure
  - Reconstruction procedure
  - Processing costs
- Simulation results
- Conclusion and Future work
Conclusion

- A network coding-based approach to PPM: marking packets with random linear combinations of router IDs, instead of individual IDs.

- Implemented the idea in practice, taking into account the bit limitations and other constraints.

- Simulated several attack scenarios. Showed it significantly reduces number of required packets.
NC + other PPM Schemes

- NC-based marking is orthogonal to and can be combined with:
  - hashing-based PPM
  - authentication schemes
  - adjusted probabilities
Future Work
inter-path coding for multipath traceback

- When network coding is deployed in the network
  - use one mark $f(R_1, R_2, R_3)$
  - instead of two $g(R_1, R_3), h(R_2, R_3)$

- Potential Benefits
  - Can signal coding point
  - Can distinguish among paths
  - Can signal the distance

- Connections with the work on topology inference + network coding
Thank you!

{psattari, athina} @uci.edu