A Technique for Counting NATted Hosts

`smb@research.att.com`

`http://www.research.att.com/~smb`

973-360-8656

AT&T Labs Research

Florham Park, NJ 07932

Why is this Interesting?

- Because of the shortage of IPv4 addresses, many people use Network Address Translators (NATs).
- Internet censuses can't easily count NATted hosts.
- How many machines are out there?

Basic Technique

- Observation: the `IPid` is usually implemented as a counter.
- By detecting *approximate sequences* of `IPid`, we can detect distinct hosts.
- Packets with the same IP address but belonging to different `IPid` sequences come from different hosts.

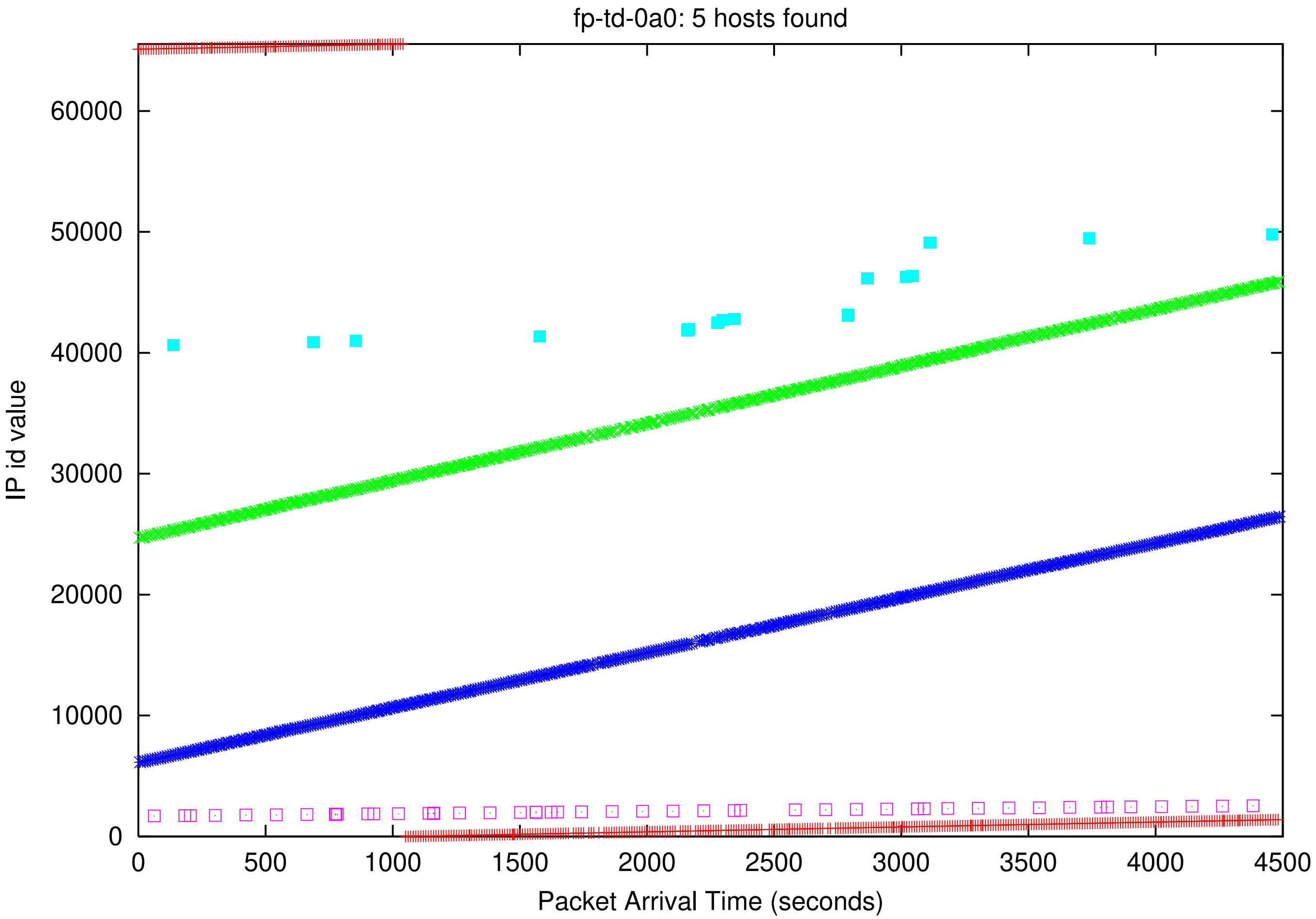
Methodology

- To permit proper control analysis, used “synthetic NAT”.
- Used packet header traces from AT&T Florham Park lab. (To preserve privacy, destination addresses and port numbers were omitted.)
- Packets from each /28 were treated as having the same source address.
- After the analysis run, comparison was made to the real data.

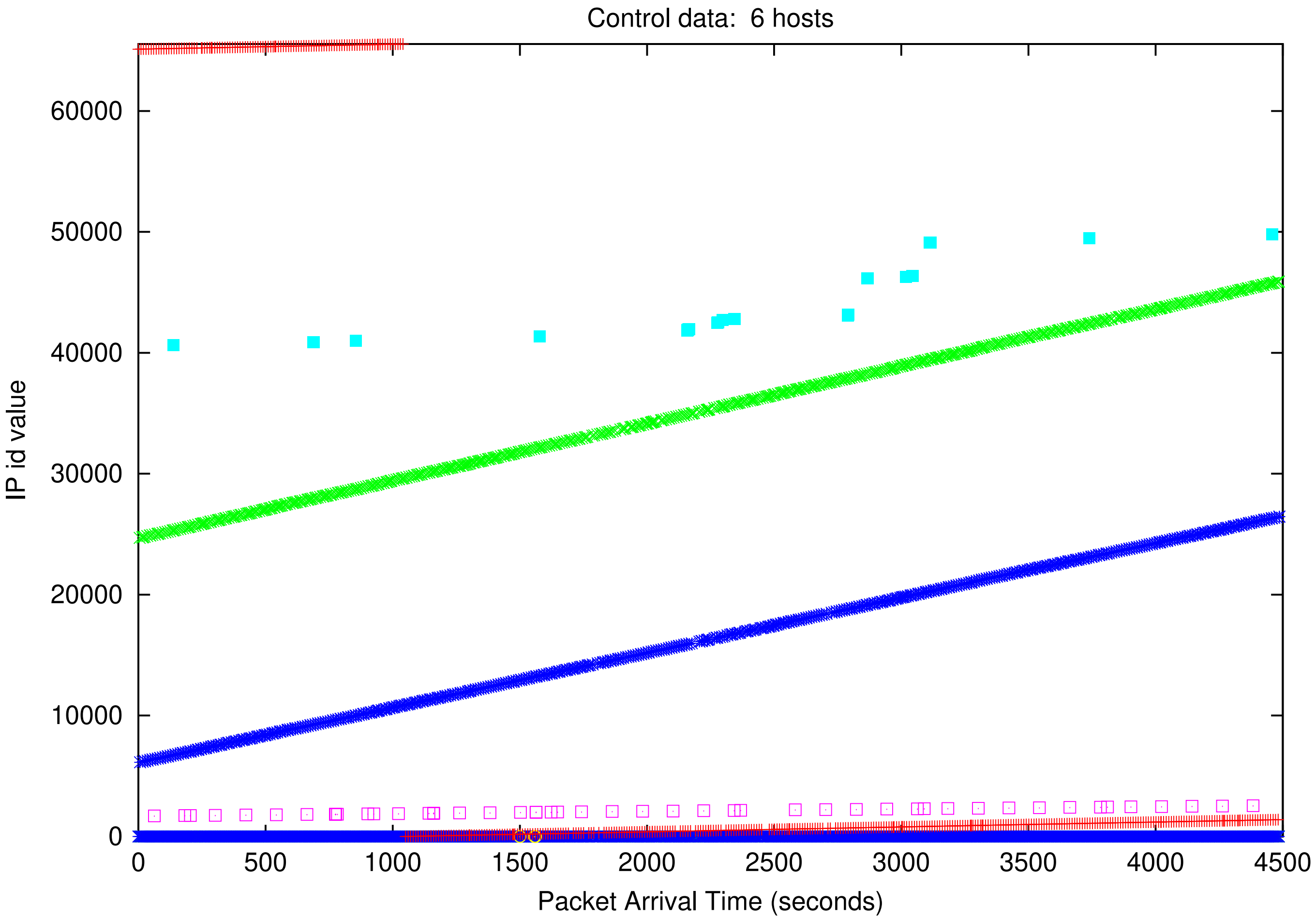
Sequence Identification Rules

- Drop `IPid` of 0; try `IPid` normal and byte-swapped.
- Packets must be “close enough” together in time.
- Bias towards exact `IPid` matches.
- `IPid` values must be “close enough”.
- After collection, “close-enough” adjacent sequences are coalesced.

Analytic Graph



Control Graph



Limitations

- Collisions can cause miscounts.
- Large gaps in `IPid` space, caused by intranet traffic, confuse the program.
- Much more suited for counting SOHO hosts than corporate NATs. (Better algorithms may change this.)
- Some operating systems (Linux, OpenBSD, FreeBSD, Solaris) sometimes use different algorithms for `IPid` assignment.
- With Path MTU enabled, `IPid` doesn't matter, and may be constant.

Privacy Issues

- Properly-designed NATs can rewrite `IPid` field.
- In fact, they *must*, to avoid fragment collisions.
- Scheme related to passive OS fingerprinting.

Future Directions

- Obvious: use technique on real trace data.
- Use other header data (TCP/UDP connection 4-tuple, TCP timestamp option) to improve packet grouping).
- Use signal processing algorithms to pick out lines.

Related Work

- Armitage counted non-default port numbers from Quake III clients.*
- Wendland uses `IPid`s to identify the identical host for Netcraft's Web server surveys. (Similar technique used by Burch and Cheswick; Mahajan et al.; probably others.)

*http://www.caia.swin.edu.au/reports/020712A/CAIA-TR-020712A.pdf