A study of denial of service attacks on the Internet

David J. Marchette

marchettedj@nswc.navy.mil

Naval Surface Warfare Center

Code B10

Outline

- Background
- Description of the Data
- Discussion of Results.
- Conclusions/Discussion

Computer Security

- Companies report hundreds of denial of service attacks each year.
- They report millions (billions?) of dollars lost.

< > - +

Computer Security

- Companies report hundreds of denial of service attacks each year.
- They report millions (billions?) of dollars lost.
- They lie.

< > - +

Computer Security

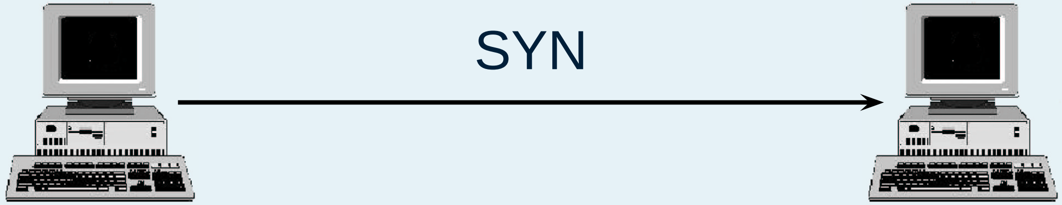
- Companies report hundreds of denial of service attacks each year.
- They report millions (billions?) of dollars lost.
- They lie.

We need a way to reliably estimate the number, type, and sizes of denial of service attacks on the Internet, without relying on self-reporting by victims. And it must be timely, not days (weeks) after the fact.

Introduction to Backscatter

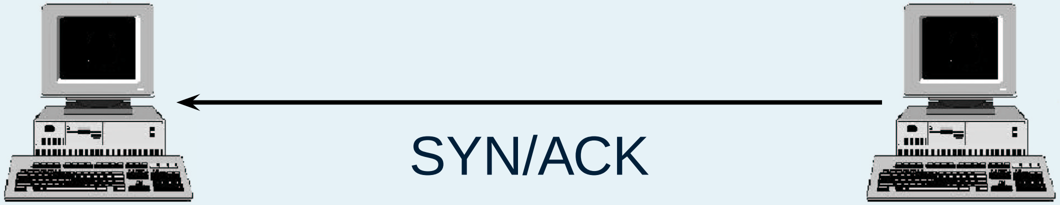
- This builds on work by David Moore et al, CAIDA, “Inferring Internet Denial-of-Service Activity”, Proceedings of the 10th USENIX Security Symposium, 2001.
- Many DOS attacks operate by sending packets to a victim with the source address spoofed.
- This results in response packets sent to the spoofed addresses.
- By monitoring the unsolicited packets sent to a network, one can estimate the level of attack, how many attacks there are, etc.

TCP 3-Way Handshake



Client sends a SYNchronize packet.

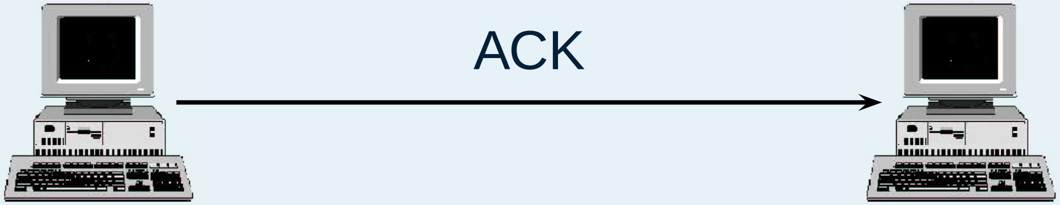
TCP 3-Way Handshake



Server ACKnowledges the SYNchronize.

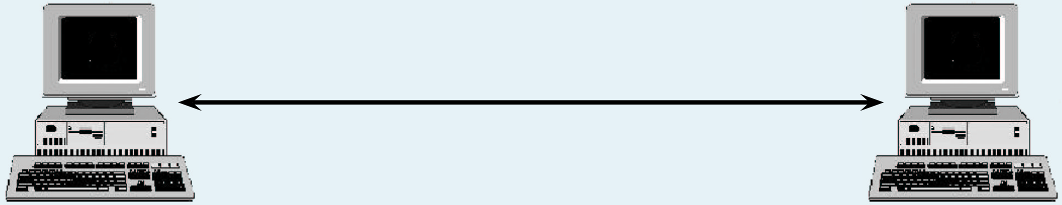
< > - +

TCP 3-Way Handshake



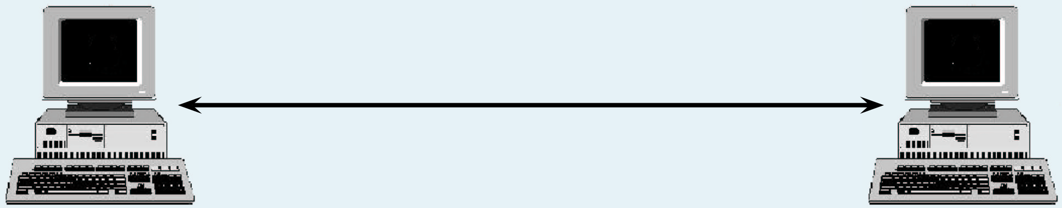
Client ACKnowledges the ACKnowledgment.

TCP 3-Way Handshake



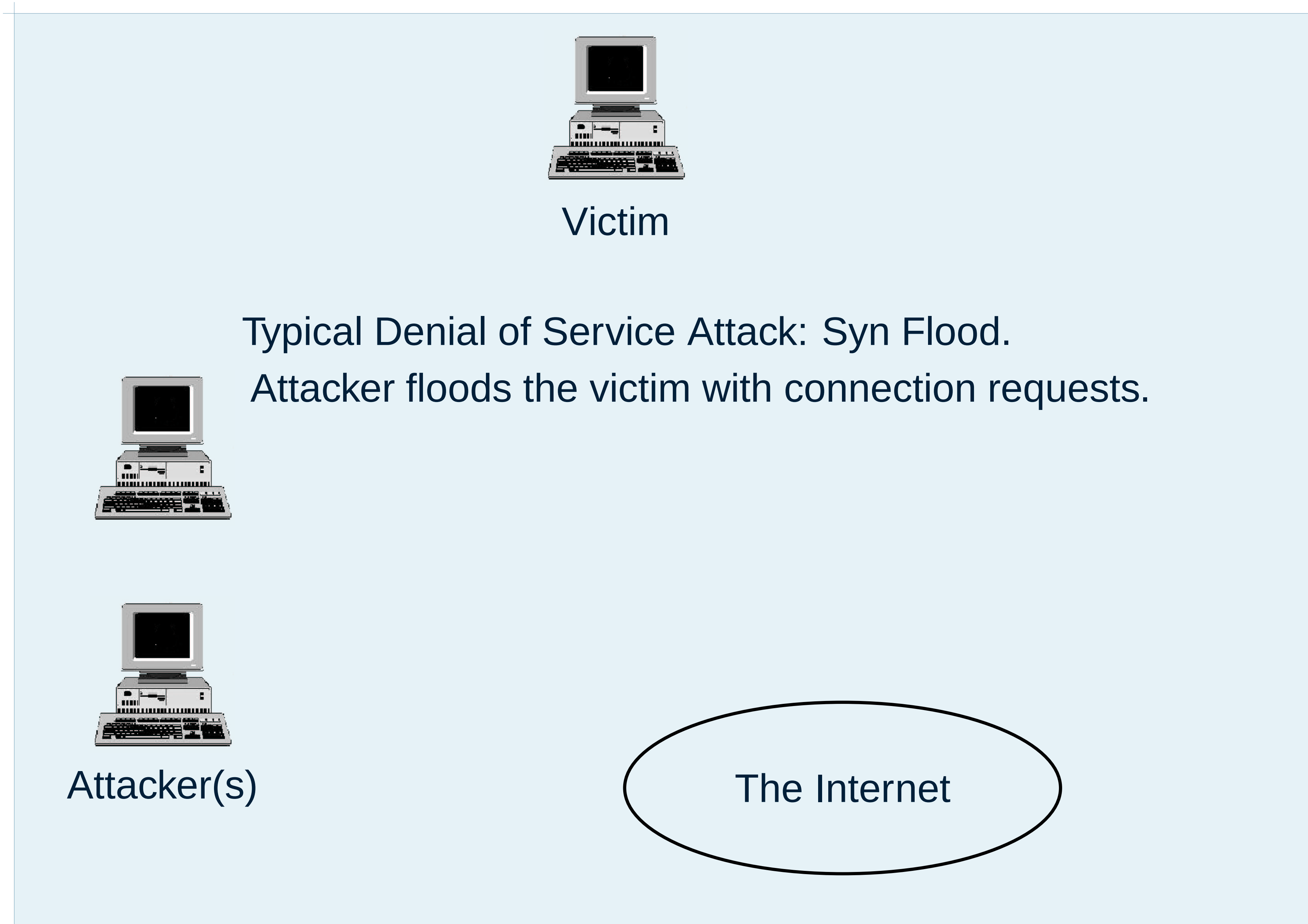
The communication channel is ready for use.

TCP 3-Way Handshake



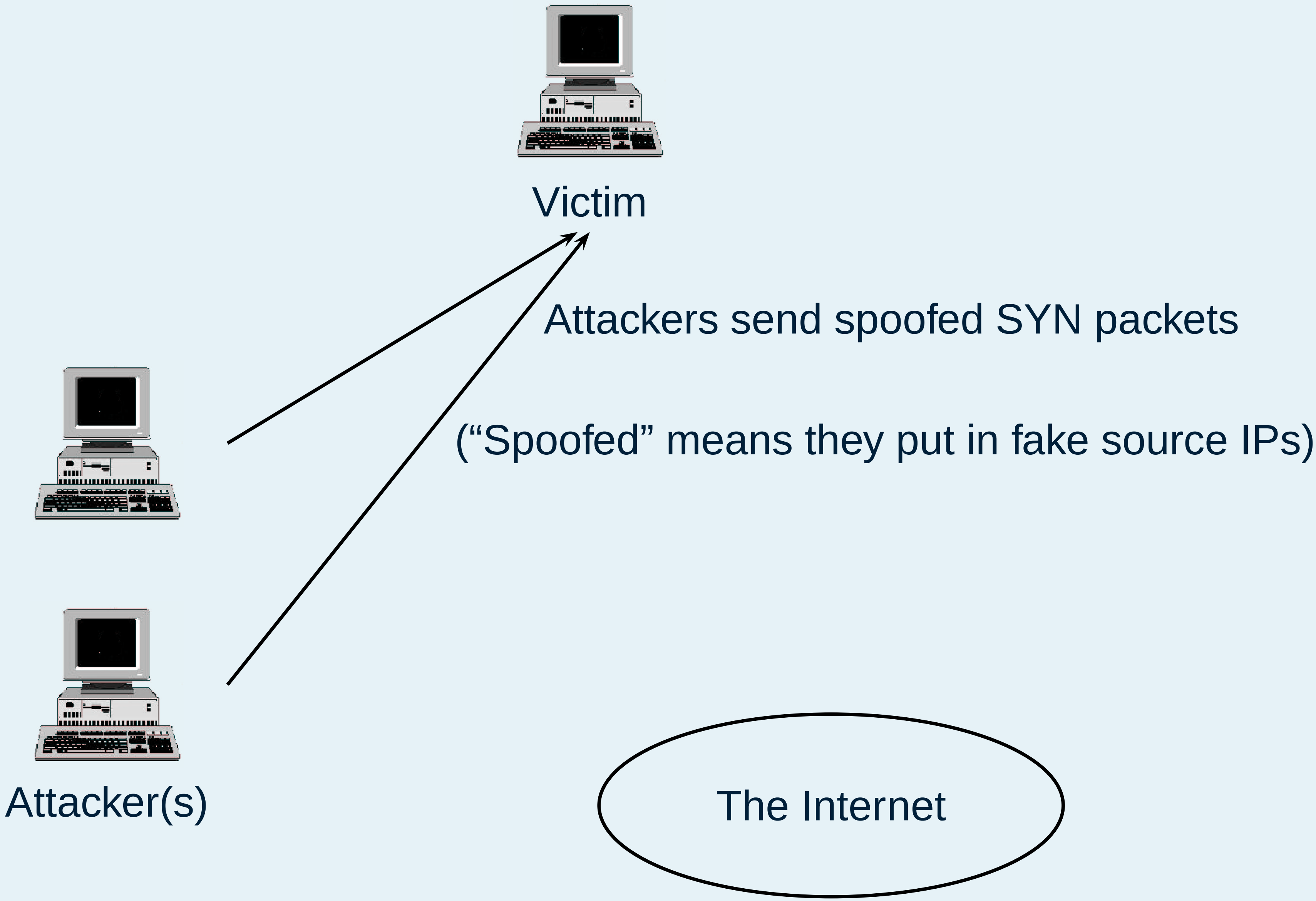
This all works because the machine’s IP addresses are in the packets, so all the routers know where to send the packets. If the client lies about this, you have a denial of service attack.

Backscatter Cartoon

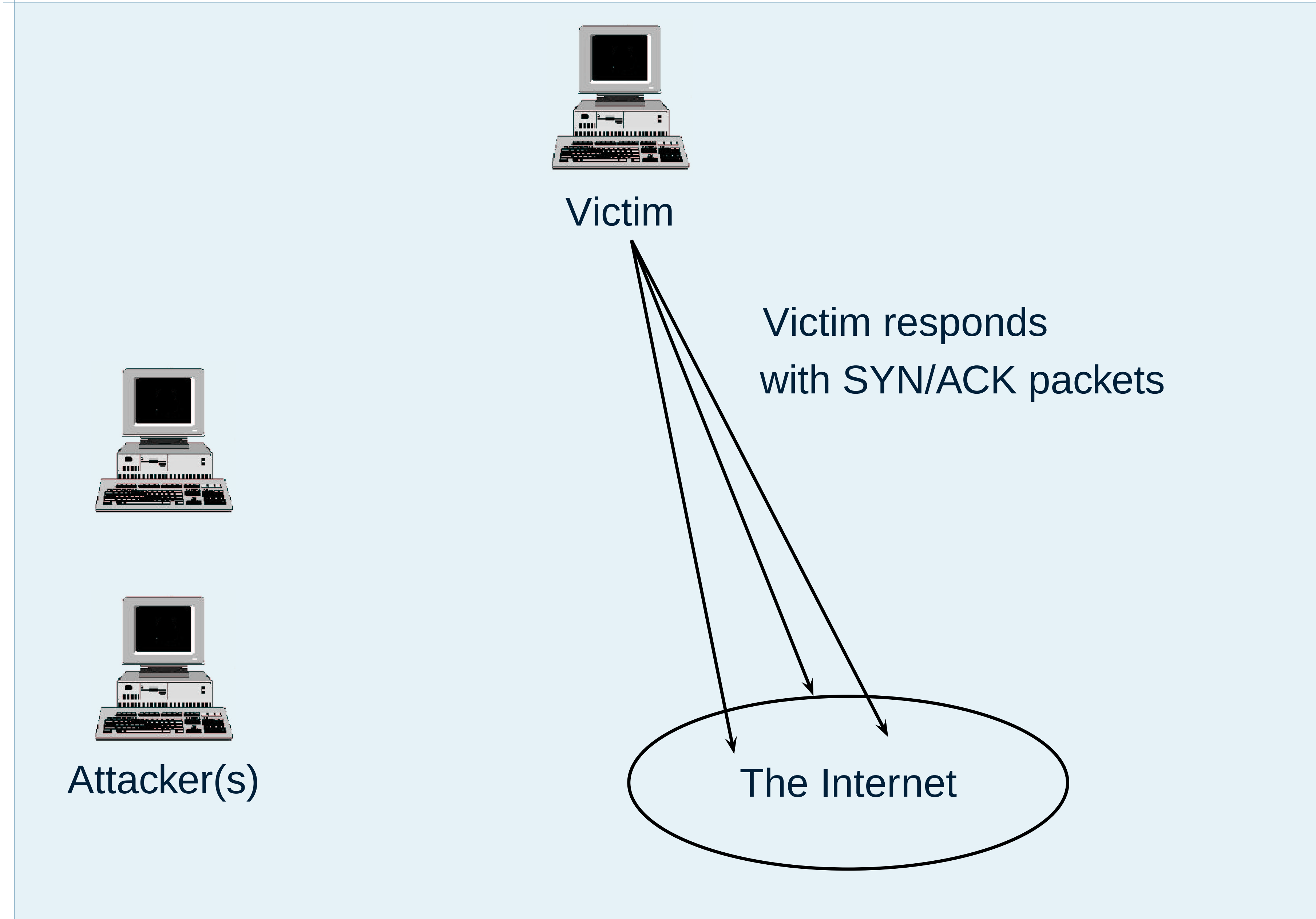


< > - +

Backscatter Cartoon

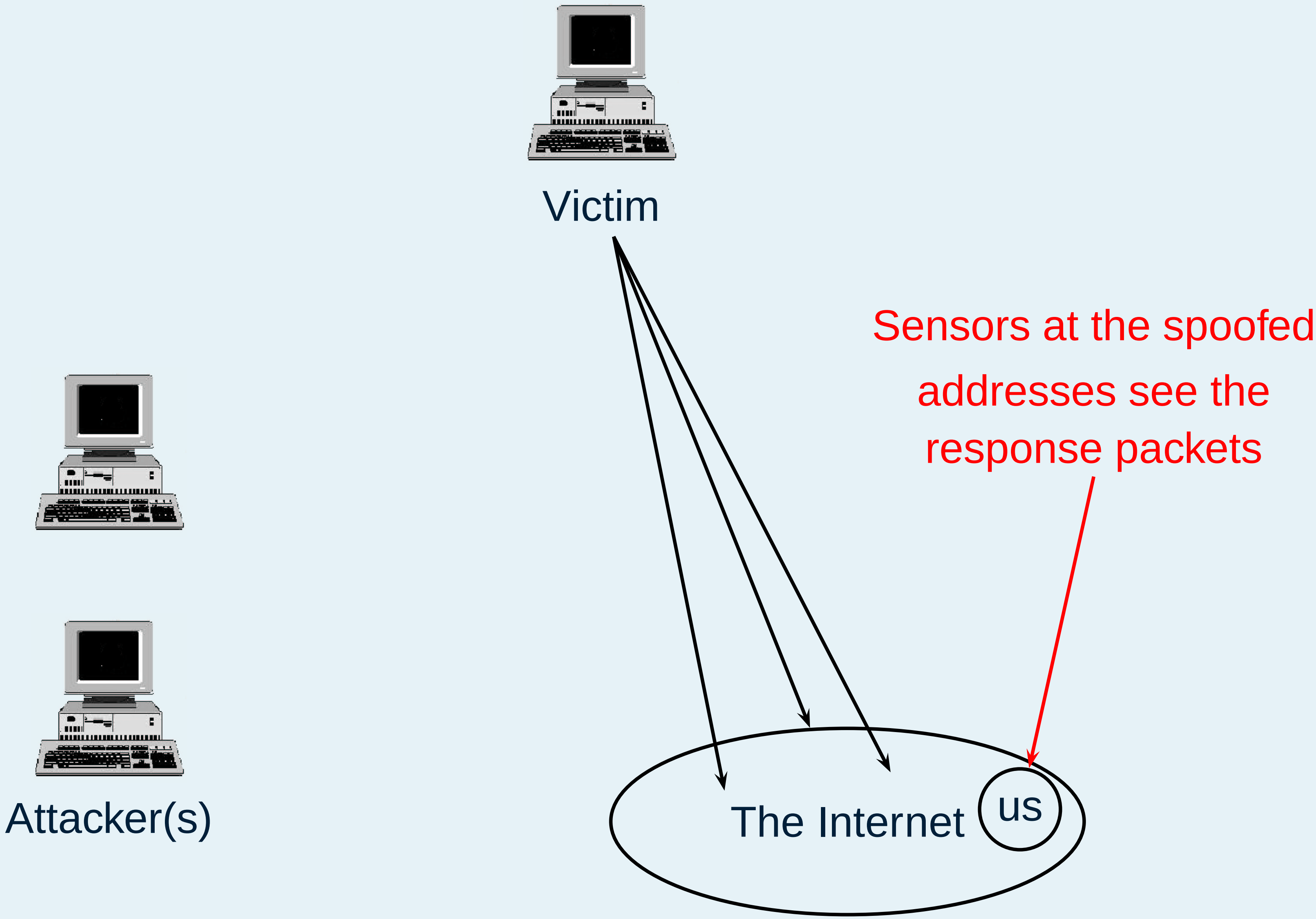


Backscatter Cartoon



< > - +

Backscatter Cartoon



< > - +

Probability of Detecting an Attack

- Assume the spoofed IPs are generated randomly, uniformly and independently. Assume m packets are sent in the attack.
- Assume we monitor n of the $N = 2^{32}$ possible IP addresses. Assume no packet loss.
- Then the probability of detecting an attack is:

$$P[\text{detect attack}] = 1 - \left(1 - \frac{n}{N}\right)^{m}.$$

- The expected number of backscatter packets we detect is:

$$\frac{nm}{N}.$$

< > - +

Estimating the Size of an Attack

- The probability of seeing exactly j packets is:

$$P[j\text{ packets}] = \binom{m}{j}\left(\frac{n}{N}\right)^{j}\left(1-\frac{n}{N}\right)^{m-j}.$$

- This allows us to estimate the size of the original attack:

$$\hat{m} = \left\lfloor \frac{jN}{n} \right\rfloor.$$

- Note that the attacker may choose to select from a subset of the 2^{32} possible IP addresses (many tools do this). Usually $N = 2^{32}, 2^{24}, 2^{16}$ or 2^{8}.
- We need to be able to determine N.

Expected Time Between Observed Packets

- Assume the attacker sends a packet every t time units, and there is no delay effect on the network.
- The expected number of attack packets between two detected packets (assuming independence) is:

$$\sum_{s=1}^{N}\left(1-\frac{n}{N}\right)^{s-1}\frac{n}{N}s = \frac{(1-(n+1)(1-\frac{n}{N})^{N})N}{n}$$

$$\approx \frac{N(1-e^{-N})}{n}$$

$$\approx \frac{N}{n}$$

Time Between Observed Packets

- The variance of the number of packets between two detected packets is:

$$
\begin{aligned}
& \sum_{s=1}^{N}\left(1-\frac{n}{N}\right)^{s-1}\frac{n}{N}s^2-\left(\sum_{s=1}^{N}\left(1-\frac{n}{N}\right)^{s-1}\frac{n}{N}s\right)^2 \\
= \quad & \frac{N(N-n-N(1+n)^2(1-\frac{n}{N})^{2N}-n(1-\frac{n}{N})^N(nN-1))}{n^2} \\
\approx \quad & \frac{N(N-n)}{n^2}.
\end{aligned}
$$

< > - +

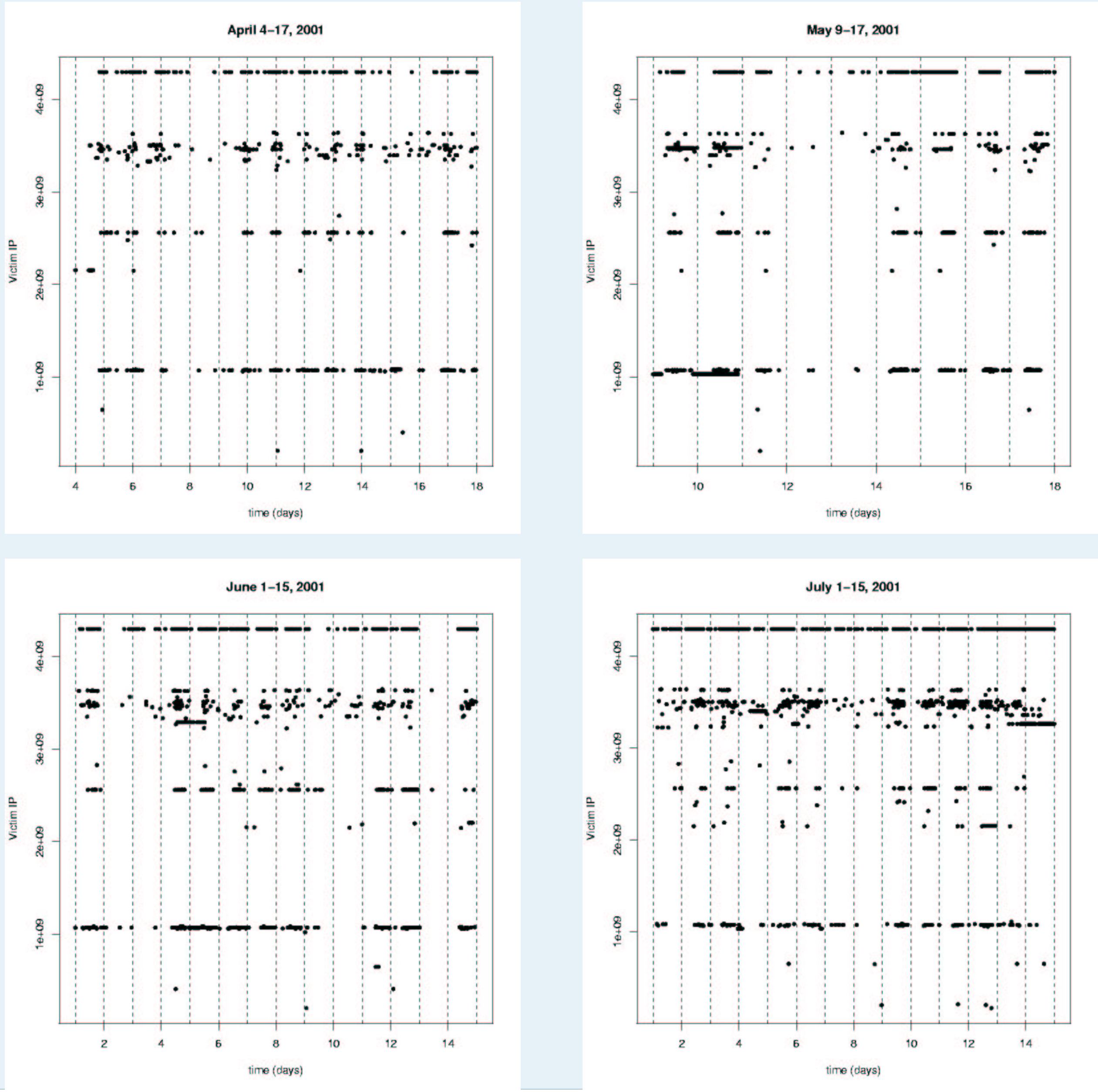
The Data

- A network of $n = 2^{16}$ IP addresses was monitored from April 2001 through January 2002.
- Only TCP packets considered in this study.
- Packets were assumed to be unsolicited if there had been no legitimate session between the source/destination pair (IPs and ports) for 20 minutes prior to the packet.
- In this study, only SYN/ACK packets were considered.
- SYN/ACKS are the response to a SYN flood, or a half-open scan.
- 8 datasets of contiguous data extracted, 7,672,597 unsolicited SYN packets during 193 days.

The Data Sets

Data Set Name	Duration	# days	# packets
April	April 4 – April 17	14	10,449
May	May 9 – May 17	9	23,264
June	June 1 – June 15	15	27,845
July	July 1 – July 15	15	59,666
Sept	Sept 1 – Sept 17	17	210,774
Oct	Sept 19 – Oct 15	26	1,253,714
Dec	Oct 28 – Dec 12	66	5,421,893
Jan	Jan 1 – Jan 31	31	665,392
	Total	193	7,672,597

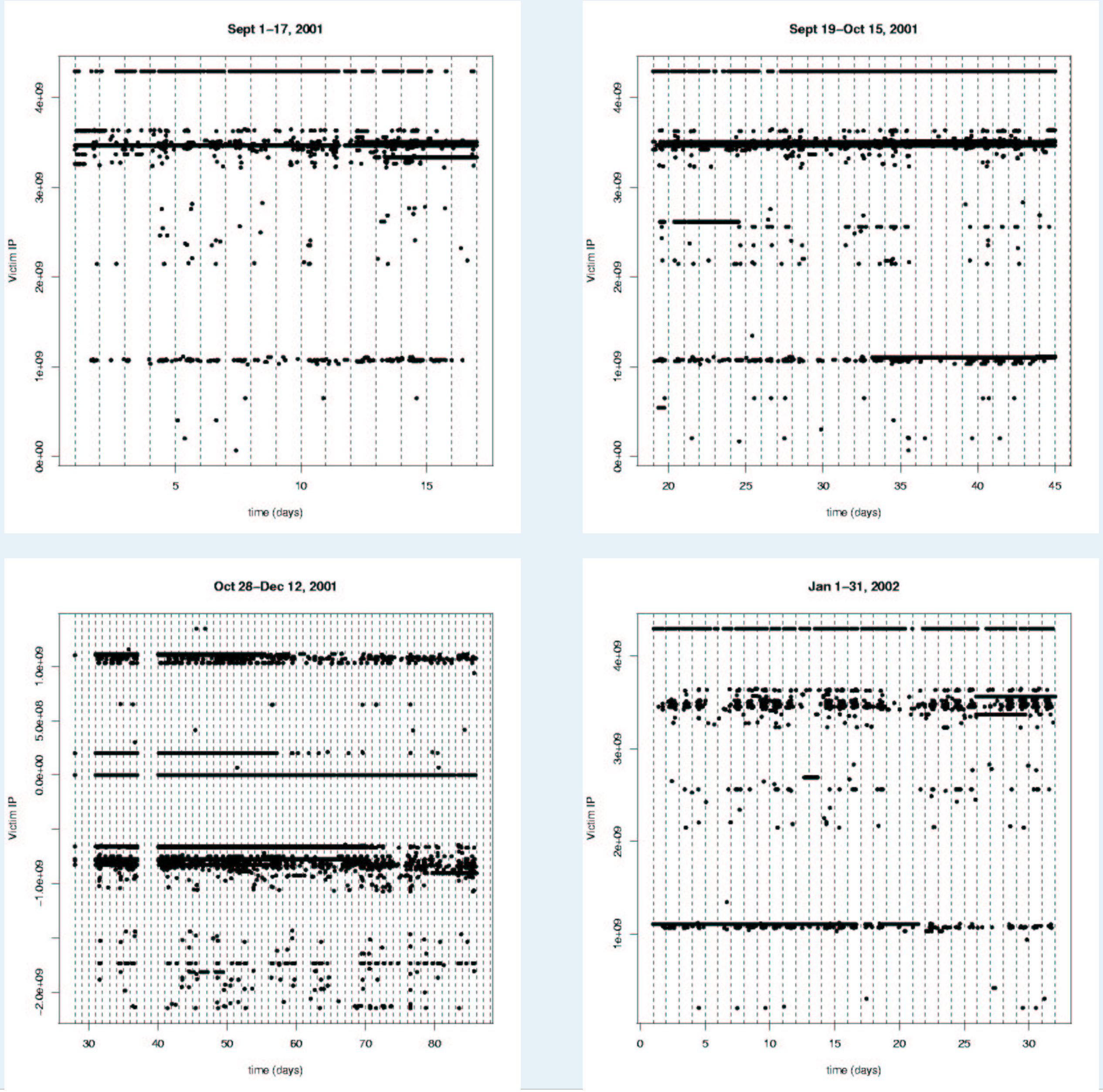
The Attacks



April 4–17, 2001
May 9–17, 2001
June 1–15, 2001
July 1–15, 2001
Victim IP
time (days)

< > - +

The Attacks



Sept 1–17, 2001
Victim IP
time (days)
Sept 19–Oct 15, 2001
Victim IP
time (days)
Oct 28–Dec 12, 2001
Victim IP
time (days)
Jan 1–31, 2002
Victim IP
time (days)

< > - +

Number of Attacks

Let T be the gap between attacks. Then the number of attacks is:

Data Set	T = 5 minutes	T = 1 hour
April	1,510	1,231
May	3,072	1,585
June	2,901	2,248
July	1,727	1,220
Sept	3,493	1,520
Sept/Oct	5,216	1,847
Oct/Dec	48,050	3,990
Jan	3,804	3,070
	69,773	16,831

< > - +

What's Going On?

- Even with the more strict definition of attack, this is over 80 attacks per day.
- Is this realistic?

< > - +

What’s Going On?

- Even with the more strict definition of attack, this is over 80 attacks per day.
- Is this realistic?
- If each attacker attacks once in this period, then there are about 1,600 active attackers.

What’s Going On?

- Even with the more strict definition of attack, this is over 80 attacks per day.
- Is this realistic?
- If each attacker attacks once in this period, then there are about 1,600 active attackers.
- This might be true.

< > - +

What's Going On?

- Even with the more strict definition of attack, this is over 80 attacks per day.
- Is this realistic?
- If each attacker attacks once in this period, then there are about 1,600 active attackers.
- This might be true.
- Some explanations:
  - dropped packets
  - scans against the monitored network
  - scans against the victim with a few spoofs
  - there really are 80 attacks per day

< > - +

What Do We Do?

- We can eliminate the dropped packets by considering only attacks with several packets.
- This biases our estimate of the number of attacks by eliminating “small” attacks.
- There are ways to detect some kinds of scans, and we can eliminate these.
- The best solution: better and more sensors.

Number of Attacks Revisited

Only consider “big” attacks, those of more than 10 packets:

Data Set	T = 5 minutes	T = 1 hour
April	54	42
May	62	60
June	97	80
July	149	107
Sept	375	192
Sept/Oct	1,324	177
Oct/Dec	6,551	414
Jan	263	206
	8,875	1,278
	46/day	7/day

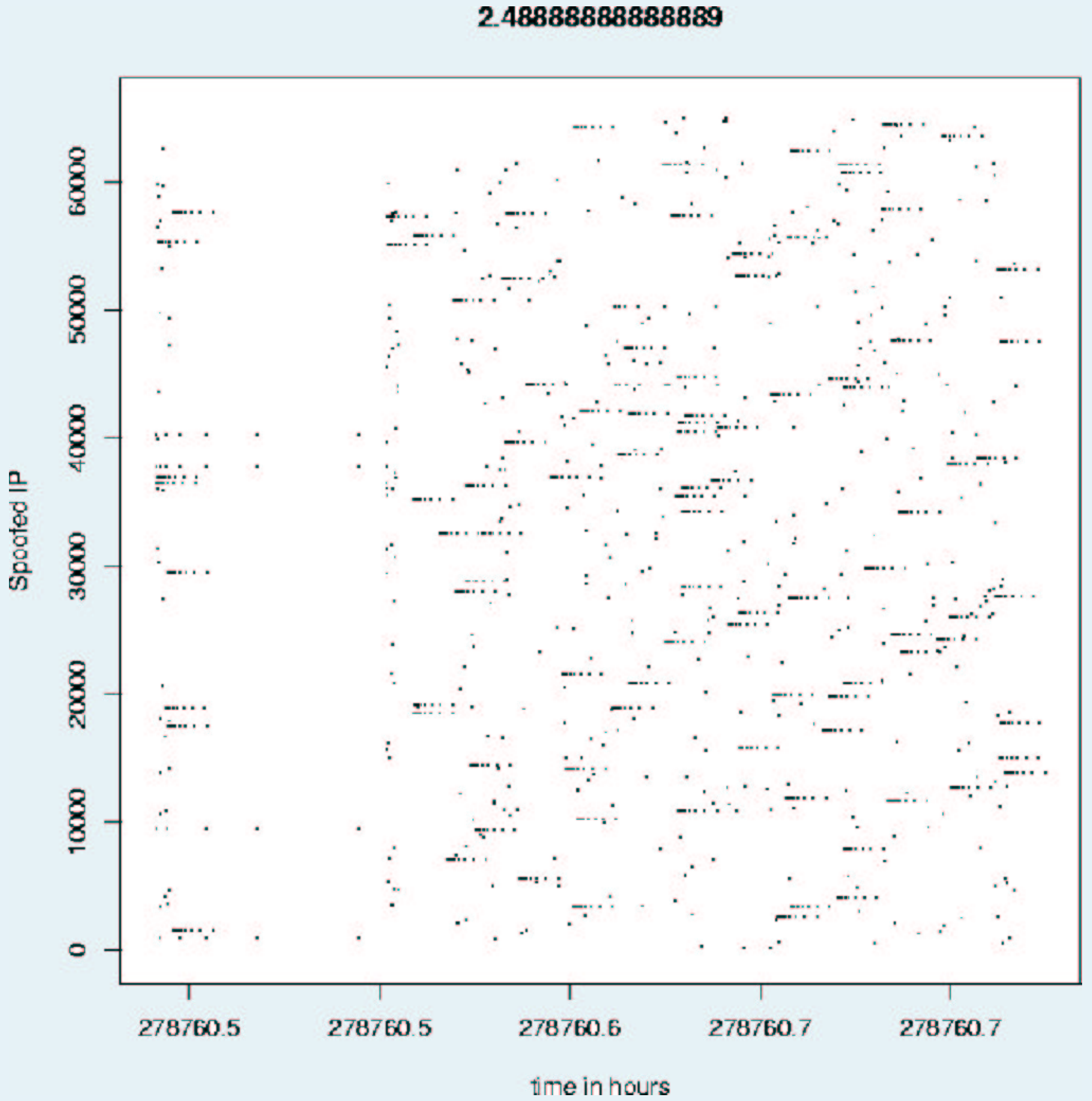
< > - +

Are the Random Assumptions Valid?

- Our models assume random, independent spoofed IP addresses.
- We will now consider some attacks to determine whether these assumptions are valid.
- We are also interested in determining (if possible):
  - the effect/success of the attack.
  - the number of attackers.
  - the attack tool used.

< > - +

Attack #1: 2,160 Packets

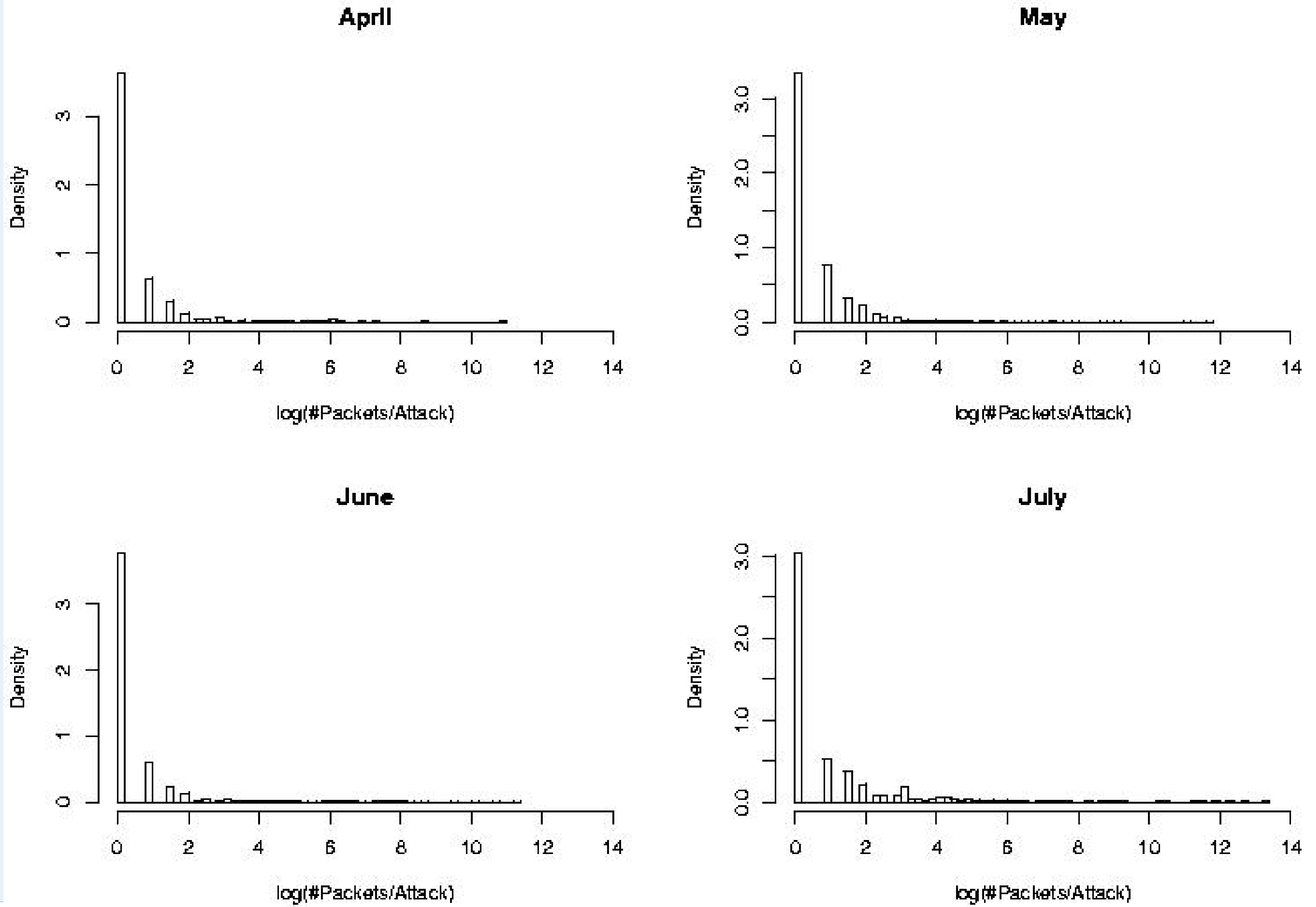


< > - +

What’s Going On?

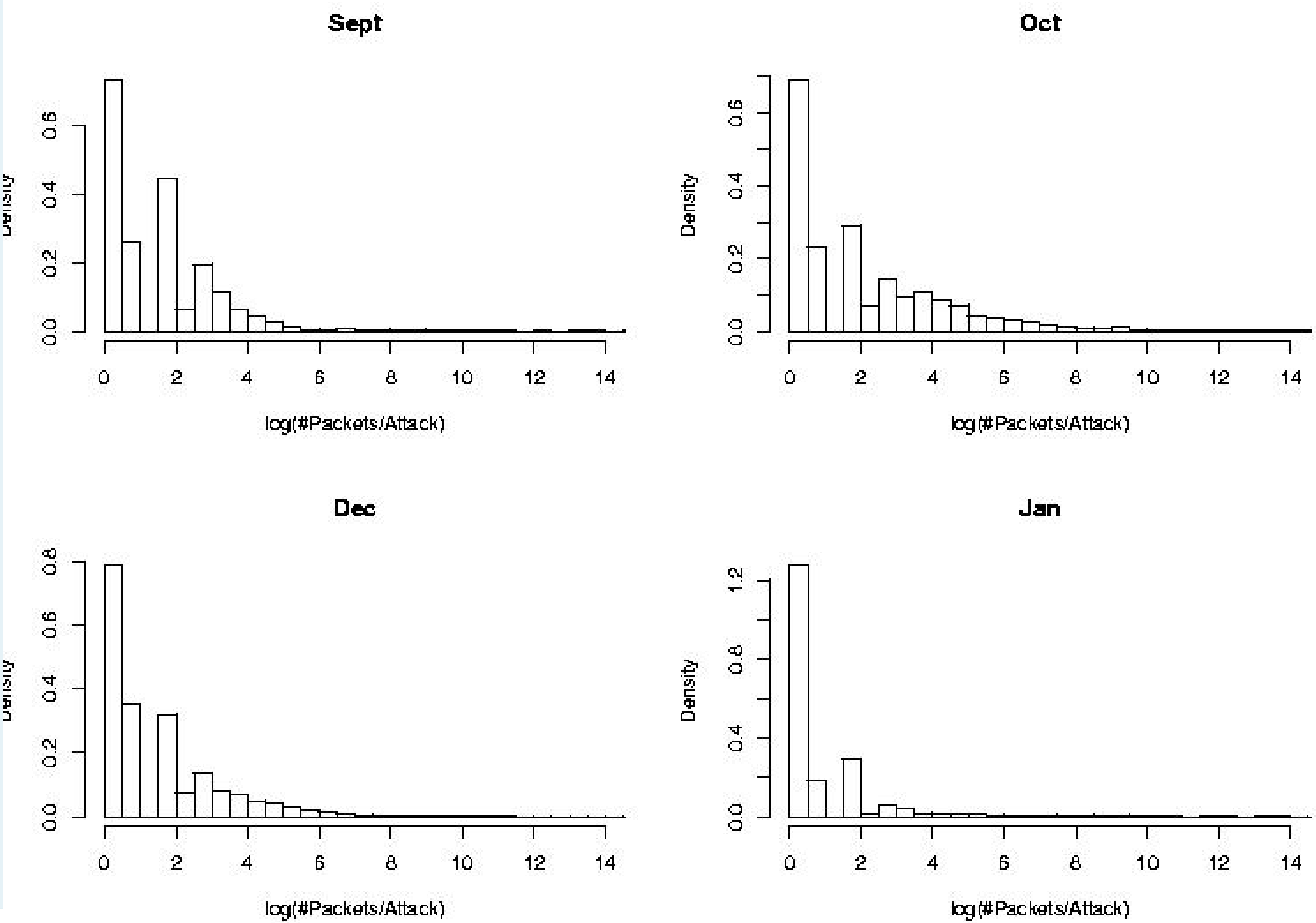
- The “streaks” are caused by resends:
  - When no response is forthcoming the victim waits, then resends the packet.
  - The victim waits twice as long, then resends.
  - The victim waits twice as long, then resends.
  - Three or four resends, then the victim gives up.
- Resends can be detected by looking at the IP/port pairing and the sequence number, as well as the time between packets.
- From here on out we eliminate these resends.

Size of Attacks, Histograms



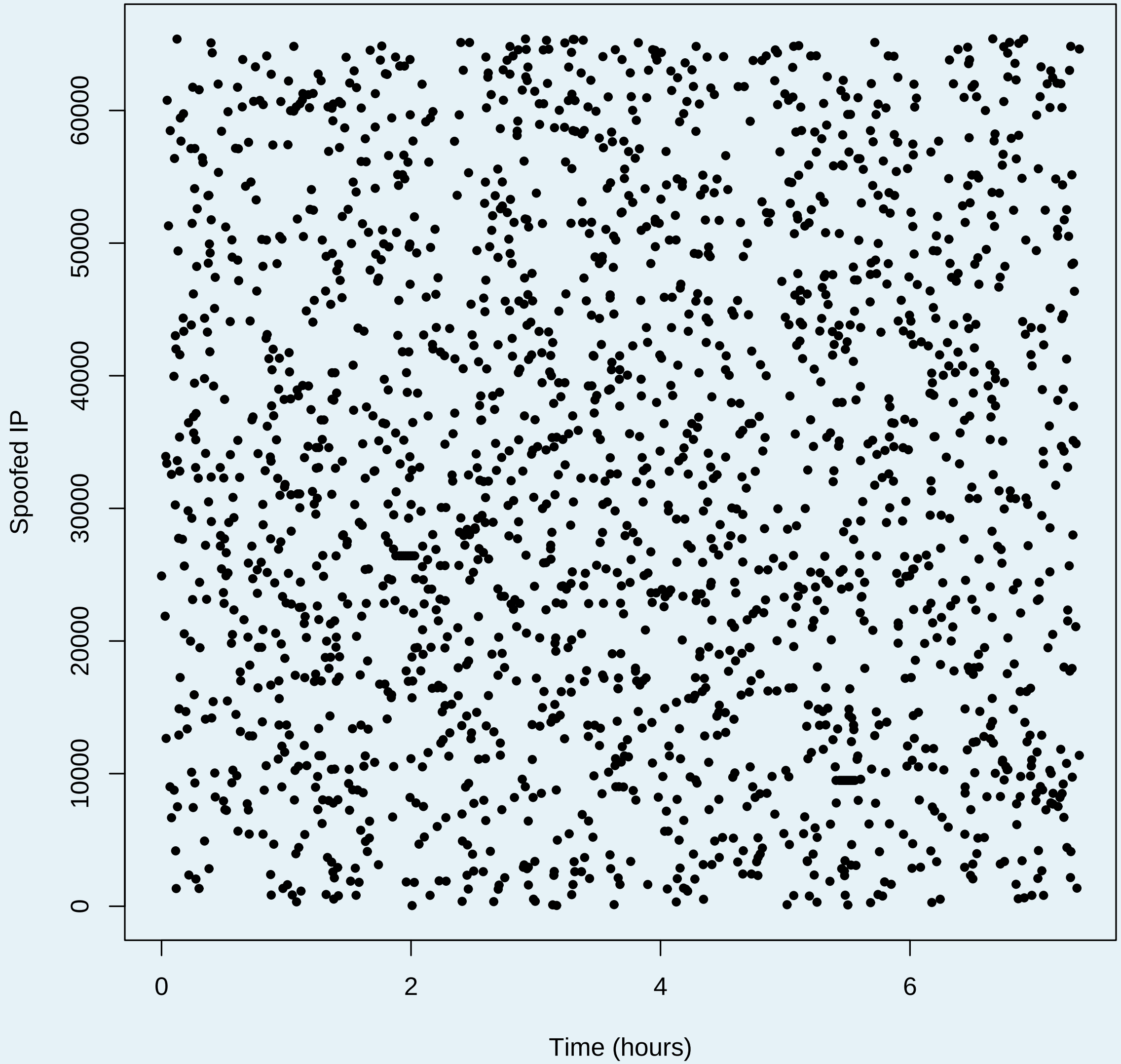
< > - +

Size of Attacks, Histograms



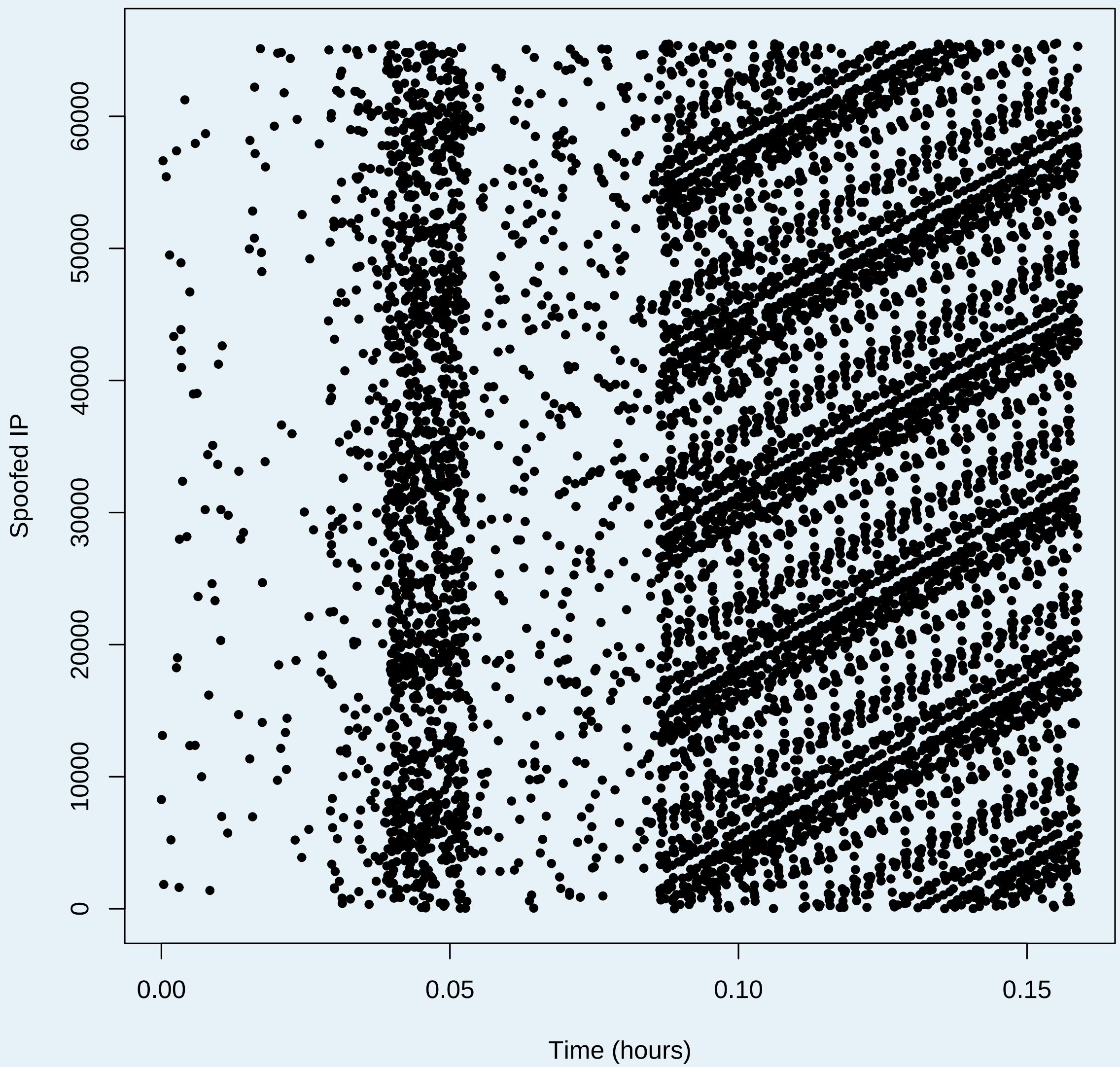
< > - +

Attack #2; 1,997 Packets



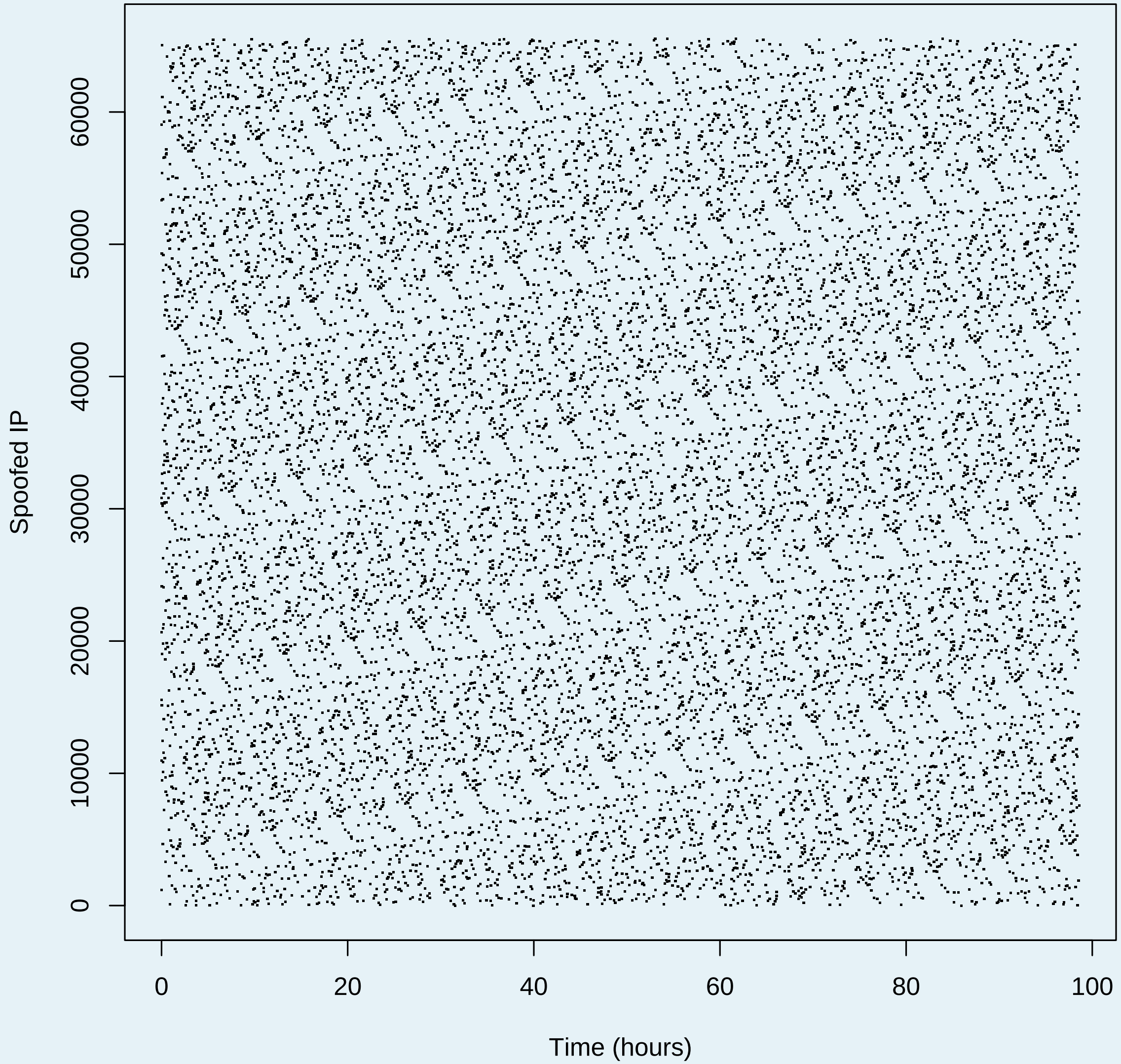
< > - +

Attack #3; 7,137 Packets



< > - +

Attack #5


Spoofed IP
0
10000
20000
30000
40000
50000
60000
0
20
40
60
80
100
Time (hours)

< > - +

From Whence the Patterns?

Three possibilities:

- It is caused by the attacker code (non-random spoofed IP selection).
- It is caused by something to do with the way packets are routed, possibly with multiple attackers.
- It is caused by the victim (load balancing?).

< > - +

Hypothesis: Supreme Random Leetness

From the code to stacheldrahtV4:

```
srandom ((time (0) + random () % getpid ()));
/* supreme random leetness */
```

Notes:

- time(0) returns seconds.
- This code is only executed when the attacker chooses not to select over all 2^{32} addresses, but instead only (a subset of) the last three octets.
- If the code is executed once, it is executed for every spoofed IP address.
- Does calling random() in the seed introduce structure?.
- This does not appear to produce the observed patterns.

Hypothesis: Routing

- Assume multiple attackers, different distances away.
- Packets from each take different length routes.
- These are interleaved at the sensor.
- Can this cause the dependence that is observed?

Hypothesis: Routing

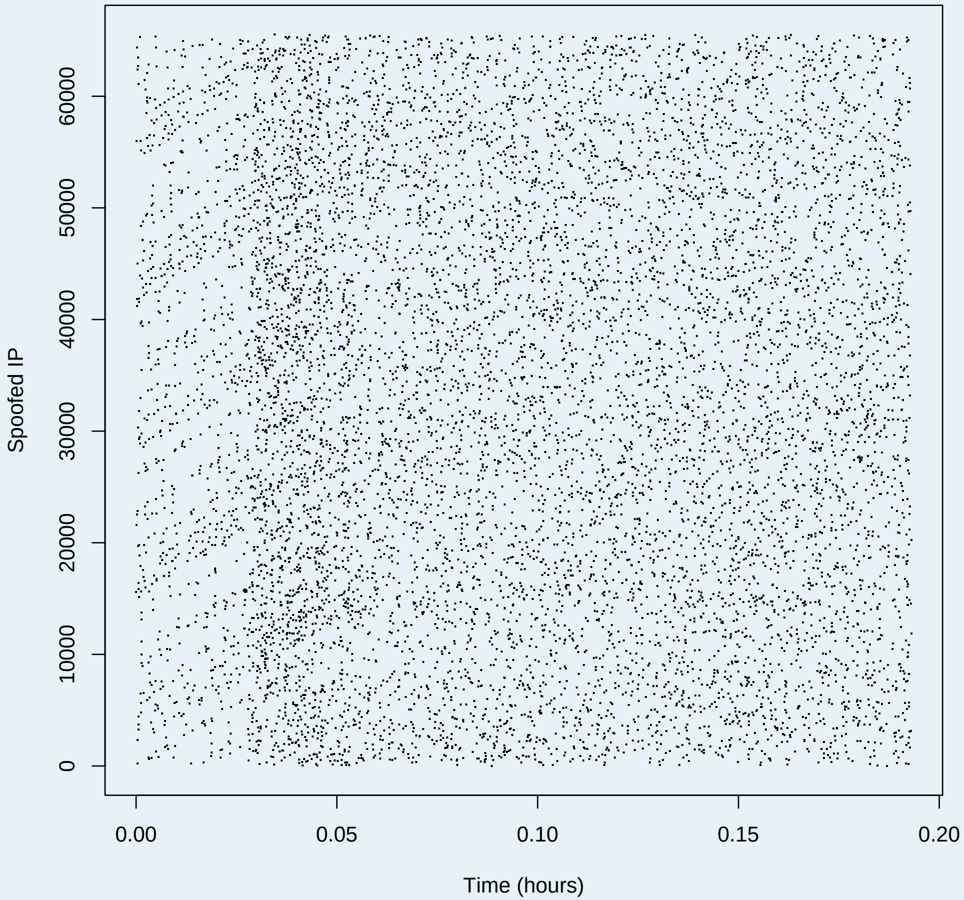
- Assume multiple attackers, different distances away.
- Packets from each take different length routes.
- These are interleaved at the sensor.
- Can this cause the dependence that is observed?
- Routes must depend on spoofed IP address.
- Only if the attackers split up the spoofed addresses.

< > - +

Hypothesis: Routing

- Assume multiple attackers, different distances away.
- Packets from each take different length routes.
- These are interleaved at the sensor.
- Can this cause the dependence that is observed?
- Routes must depend on spoofed IP address.
- Only if the attackers split up the spoofed addresses.
- This does not seem to explain the structure.

Hypothesis: Victim Actions



- Does this contradict the hypothesis?

< > - +

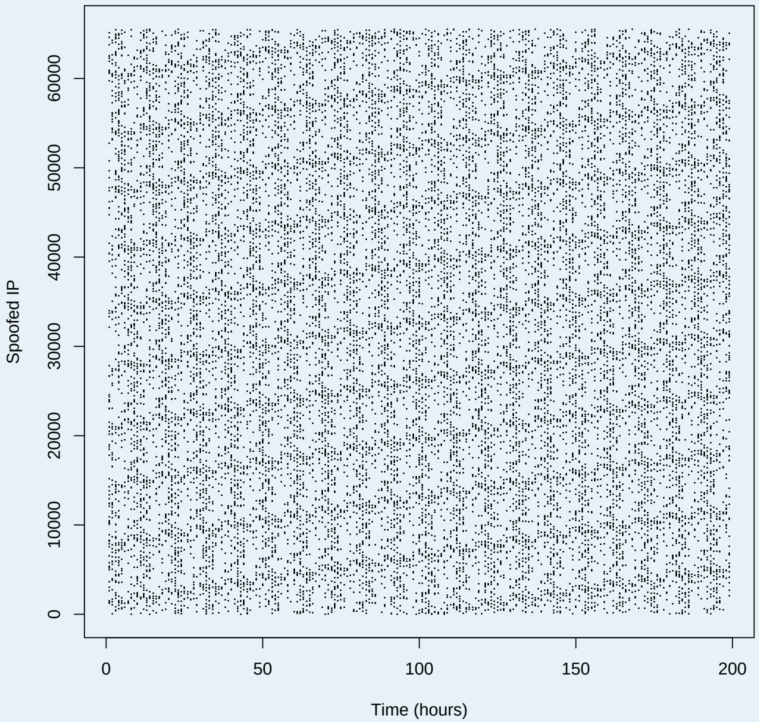
Deterministic Algorithm

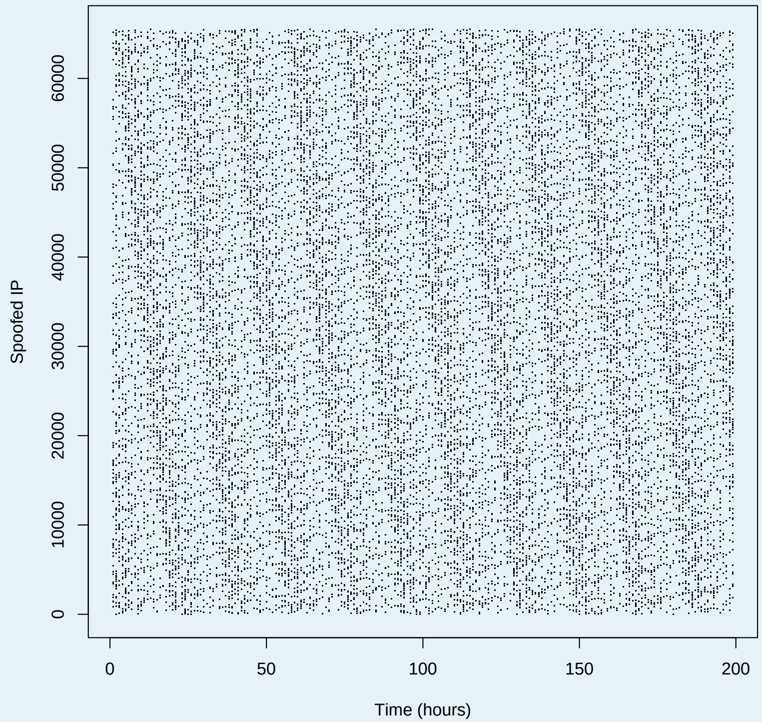
- Assume m attackers each pick a different starting IP address.
- Each attacker increments the IP address by a fixed amount.
- Packets arrive at a random time, with random interleaving.
- This should give a “linear” pattern like we see.
- Let’s look at this.

< > - +

Hypothesis: Deterministic Algorithm

100 attackers, each starting at a random IP, then incrementing by a fixed amount:

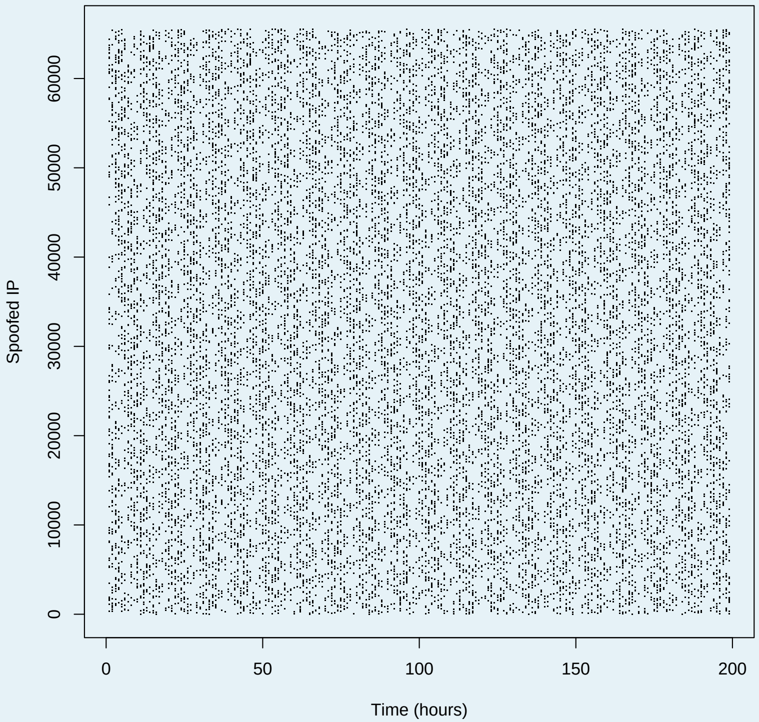


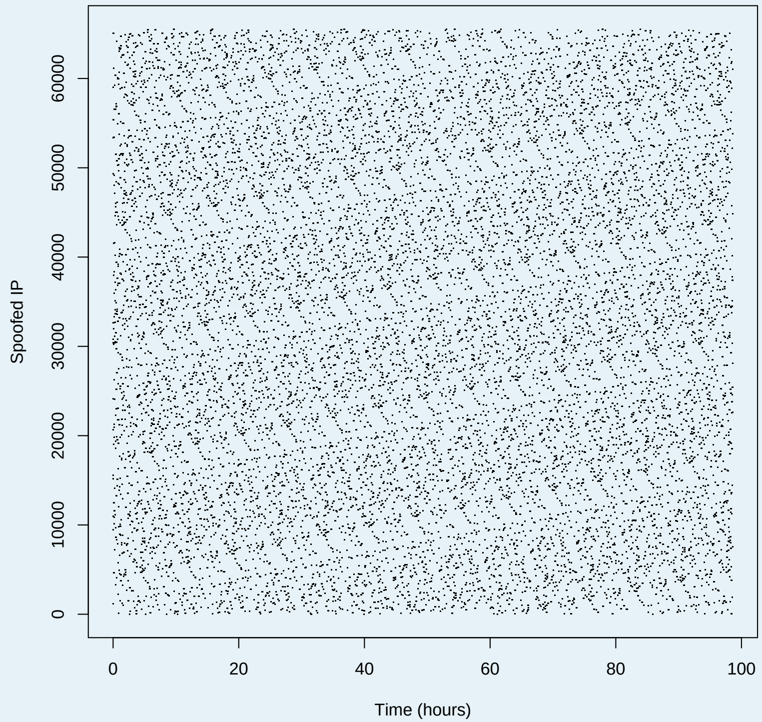


< > - +

Hypothesis: Deterministic Algorithm

100 attackers, each starting at a random IP, then incrementing by a fixed amount:





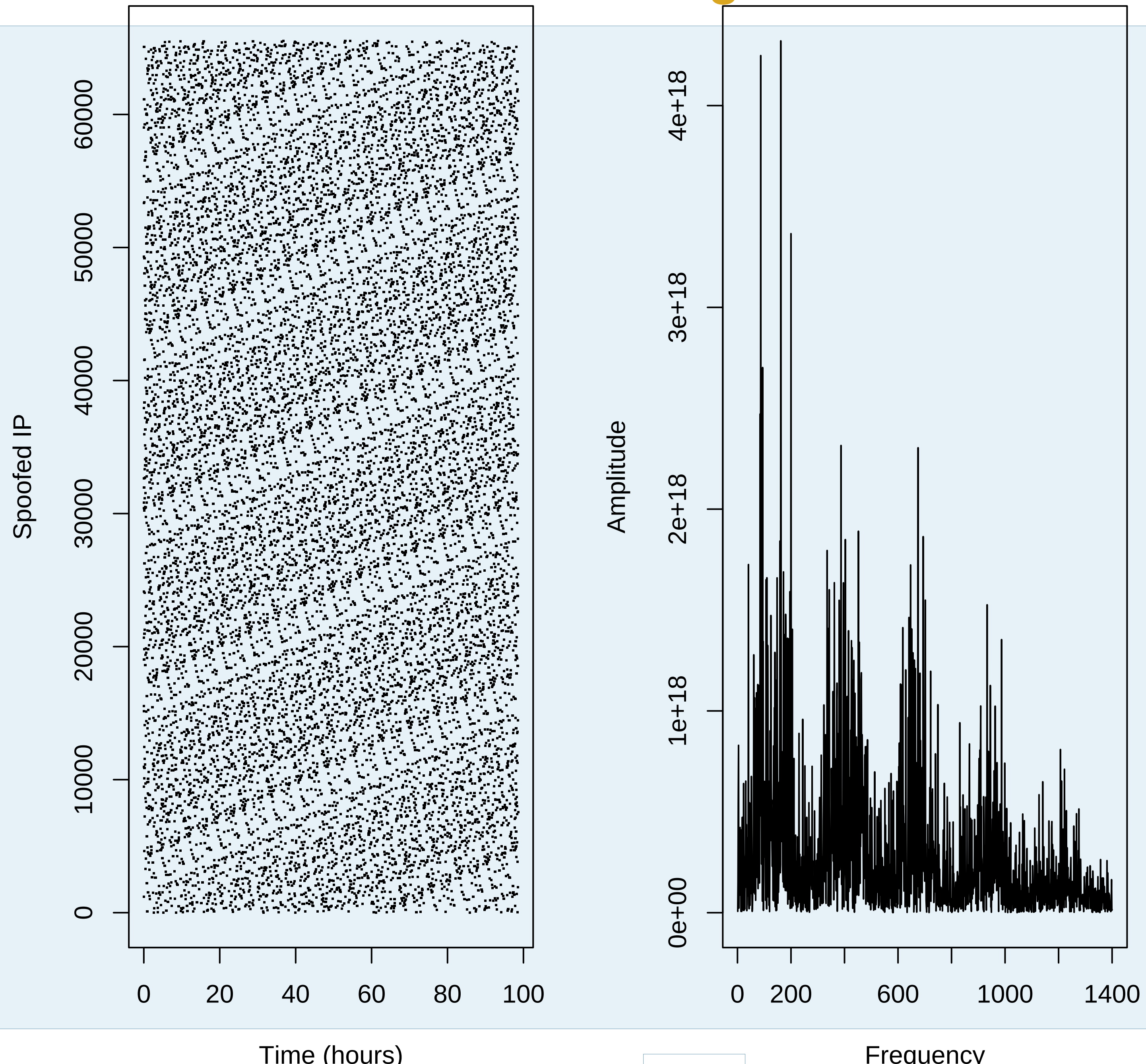
Which is the real attack?

< > - +

Deterministic Notes

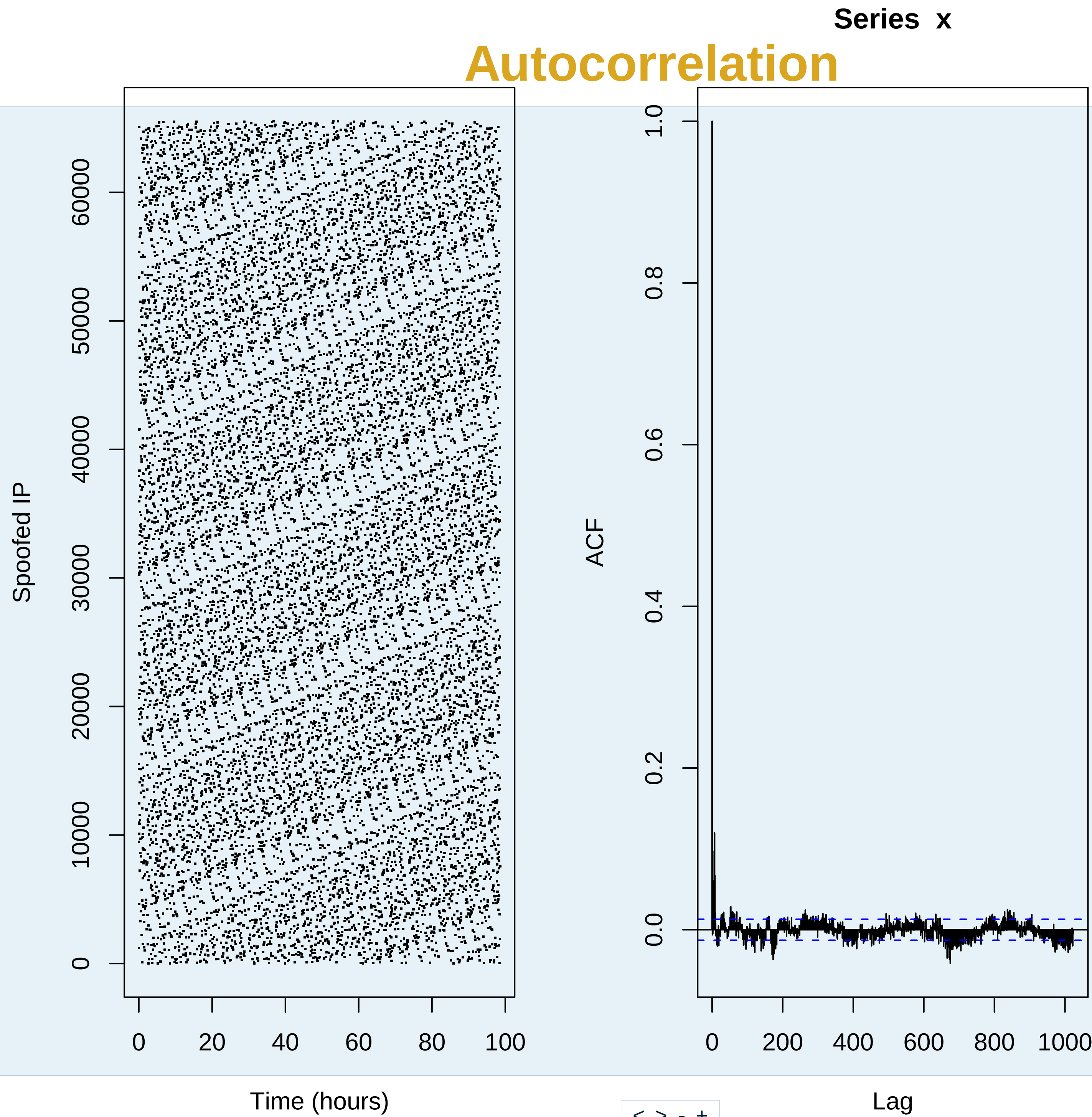
- The simulations are similar to the attack patterns.
- The attackers do not all seem to use the same increment.
- Adding multiple increments changes the slopes of the lines.
- There may be packet losses that are not present in the simulations.
- The simulation's packet interleaving is probably not quite right.

Periodogram



< > - +

Autocorrelation



< > - +

Thinking about Models

Even deterministic attacks have random aspects to them:

- Random start times of the attacks from multiple attackers.
- Random initial IP address.
- The path length to the victim differs.
- Different random delays on the path.
- Packet loss.

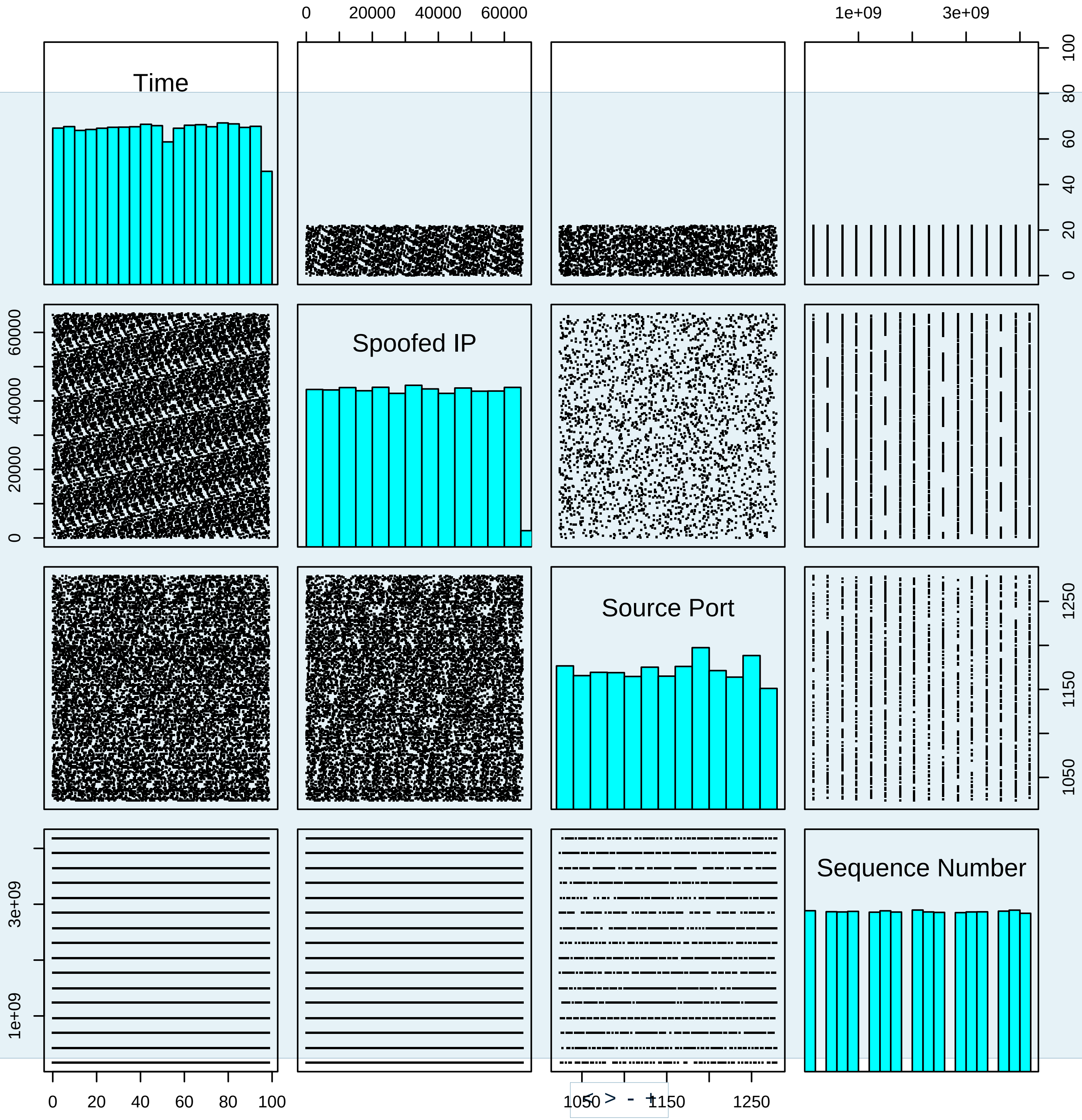
This results in a (possibly random) interleaving of the packets from different attackers, as well as random arrival times.

Some attacks may mix in some random IP selections.

Some attacks are purely random.

Discussion

- A single attack packet can generate multiple responses. This means that our estimates must take this into account.
- Some attack tools use purely random spoofed IP addresses.
- Some attack tools appear to use a deterministic algorithm.
  - This effects our estimates.
  - Pattern might allow a signature as to the tool used.
  - Pattern might allow for a determination of number of attackers.
  - Attack may not be purely deterministic.
- Attacks can overlap, making the definition of “attack” tricky.
- Other header features should be investigated:
  - Destination port.
  - Sequence number.


Time
Spoofed IP
Source Port
Sequence Number
0
20000
40000
60000
1e+09
3e+09
0
20
40
60
80
100
1050
1150
1250

Future Work

- Stochastic/deterministic model for attacks.
- Expand the investigation to other header parameters.
- Look at other attacks besides SYN floods.
- Explain the bumps in the “size of attack” histograms (are they really there?).
- Test network for running attack tools.
- See if we can determine the attack tool from the pattern of the attack.
- More sensors.