Week 5: Traffic measurements

1. Introduction
2. Packet capture
3. Interface counts
4. Flow capture
5. Traffic matrix
6. **Anonymization of packet traces**
7. Conclusion

Benefits of sharing data

- Good scientific practice
- Get others to work on relevant problems
- Learn from analysis of others
- Get broader view

But, packet traces contain lots of sensitive information

- Headers
  - Connection endpoints: who is talking to who; sites visited
  - Protocol, ports: applications used
- Payload
  - Visited content
  - Passwords, etc.

Solution: Anonymization

- Process to sanitize data to ensure anonymity
  - Absence of identity
  - Prevent others from linking identity to actions of an individual

Anonymizing payload

- Payload contains most sensitive information
  - Better if removed completely
  - If not possible, get minimum necessary
    - ✓ E.g., HTTP host better than full URL

Anonymizing packet headers

- Packet headers can be shared with care
  - MAC addresses
    - ✓ Potential to link records with the same MAC across datasets
  - IP addresses often need to be anonymized
  - IP addresses appear in other parts of the packet
    - ✓ IP options (e.g., record route)
    - ✓ ICMP/DNS packets

Anonymization of IP addresses

- Simply hashing IP addresses is too restrictive
  - IPs in the same prefix announced by the same AS
  - Many analysis require knowledge of IP prefix
- Prefix-preserving anonymization
  - All IP addresses that share a prefix in the raw data will also do so in the anonymized data
- Packet trace anonymization tools
  - tcpdpriv, ipsumdump, ip2anonip