High Fidelity, High Risk, High Reward: Using High-Fidelity Networking Data in Ethically Sound Research

Mohammad Taha Khan & Chris Kanich
The University of Illinois at Chicago
Introduction

Networking Data

- Recent focus on Internet users
- Infrastructure to collect large data volumes

Motivation

- Experience with networking tap data.
- High Fidelity and High Risk
In This Report...

- Networking data sources
- Measurement techniques
- Example use cases
- Ethical guidelines for research
Data Sources

- HTTP and DNS Logs
- Middleware Data
- Data From ISPs
- Public Sniffing Probes
- Crawler Data
- Botnet/Honeypot Data
General Measurements

- **Macro Analysis**
  - Generating statistics from previous data
  - Using ML for actionable intelligence

  **Use Case:** Social network trends

- **Micro Analysis**
  - Focus on infrastructure quality

  **Use Case:** Page load times in an enterprise
Measuring Human Involvement

- Studying Behavior
  - Focus on activities of the general user space
    - **ISP Censorship in Pakistan** *(IMC 14)*

- Studying Misbehavior
  - Activities pertaining to a specific user subset
    - **PharmaLeaks** *(Usenix Sec 12)*

- Direct User Interaction
A Case Study of the “Tap” Data Set

- **Respect for persons**
  - IRB Approvals, Informed consent

- **Beneficence**
  - Giving back to the community

- **Justice**
  - Grounds for discrimination

- **Respect of the law**
  - Data used for allowed purposes
Learning Experiences

- Singling out users
- Disjoint relation of researchers and user subjects
- Disclosure of identity
- Operational feedback to volunteer organization
- Systems allowing individuals to opt out
- Data anonymized by employees
Thank You & Questions?