The state of user privacy: Guidelines for performing safe measurement on the internet

https://datatracker.ietf.org/doc/draft-irtf-pearg-safe-internet-measurement

User privacy in measurement: A principled approach

- Guidelines for ensuring any measurement, in any environment, can be carried out safely,
- From a user perspective,
- With a special focus on data minimisation,
- Since detection and measurement proliferates data, by and of itself,
- Where especially in encrypted environments enhancing metadata should not be used to compensate for lack of content.

Internet-draft in PEARG

- Link to the active internet-draft:

  https://datatracker.ietf.org/doc/html/draft-irtf-pearg-safe-internet-measurement
- Discussion of draft-safe-internet-measurement happens on the PEARG list:

  pearg@irtf.org
- Issues and pull requests are accepted here:

  https://github.com/IRTF-PEARG/draft-safe-internet-measurement.

Authors

- Iain Learmonth, Tor Project
- Gurshabad Grover, Centre for Internet and Society India
- Mallory Knodel, Center for Democracy and Technology

Goal

For industry and academia using measurements to research the functioning and usage of the Internet, this document describes guidelines for ensuring that such measurements can be carried out without violating user privacy.

Scope

- Not a substitute for any institutional ethics review process.
- Not legal advice.
- Restricted to guidelines for measurement of: the network, its constituent hosts and links, or its users traffic.
- An Internet user is an individual or organisation whose data is used in communications over the Internet, most broadly, and those who use the Internet to communicate or maintain Internet infrastructure.

Structure

- Consent: informed, proxy, implied
- Safety considerations
  - Isolate risk with a dedicated testbed
  - Be respectful of others' infrastructure
  - Maintain a "do not scan" list
  - Minimize Data
- Risk analysis

Consent

- Informed consent

  A researcher uses volunteer owned mobile devices to collect information about local Internet censorship. Connections will be made from the volunteer's device towards known or suspected blocked webpages.

- Proxy consent

  A researcher performs a packet capture to determine the TCP options and their values used by all client devices on an corporate wireless network.

- Implied consent

  A researcher performs A/B testing for protocol feature on web performance. The two software versions report telemetry. These updates are pushed to users at random by auto-update. The telemetry excludes PII or location data.

Safety: Isolate risk with a dedicated testbed

Wherever possible, use a testbed. An isolated network means that there are no other users sharing the infrastructure you are using for your experiments.

When measuring performance, competing traffic can have negative effects on the performance of your test traffic and so the testbed approach can also produce more accurate and repeatable results than experiments using the public Internet.

WAN link conditions can be emulated through artificial delays and/or packet loss using a tool like [netem]. Competing traffic can also be emulated using traffic generators.

Safety: Be respectful of others' infrastructure

If your experiment is designed to trigger a response from infrastructure that is not your own, consider what the negative consequences of that may be. At the very least your experiment will consume bandwidth that may have to be paid for.

In more extreme circumstances, you could cause traffic to be generated that causes legal trouble for the owner of that infrastructure. The Internet is a global network crossing many legal jurisdictions and so what may be legal for you is not necessarily legal for everyone.

If you are sending a lot of traffic quickly, or otherwise generally deviate from typical client behaviour, a network may identify this as an attack which means that you will not be collecting results that are representative of what a typical client would see.

Safety: Minimize data

When collecting, using, disclosing, and storing data from a measurement, use only the minimal data necessary to perform a task. Reducing the amount of data reduces the amount of data that can be misused or leaked.

When deciding on the data to collect, assume that any data collected might be disclosed… See section 6.1 of RFC6973 [RFC6973] for data minimalization considerations specific to this use case.

- Discard data
- Mask data
- Reduce accuracy
- Aggregate data.

Risk

The benefits should outweigh the risks. Consider auxiliary data (e.g. third-party data sets) when assessing the risks.

Future of the draft (open issues)

- Include responsible disclosure
- Consider availability, not just disclosure
- Consider IP addresses
- Discuss future computing capabilities
- Cite CAIDA's Promotion of Data Sharing Webpage
- Cite Workshop on Ethics in Networked Systems Research
- Data minimization section is largely unwritten
- Risk assessment could be extended or removed.

User privacy in measurement: A principled approach

- *Guidelines for ensuring any measurement, in any environment, can be carried out safely,*
- *From a user perspective,*
- *With a special focus on data minimisation,*
- *Since detection and measurement proliferates data, by and of itself,*
- *Where especially in encrypted environments enhancing metadata should not be used to compensate for lack of content.*

Open questions

- Is it thorough?
- Is there support for this draft?
- Should data minimization be elaborated here or elsewhere?