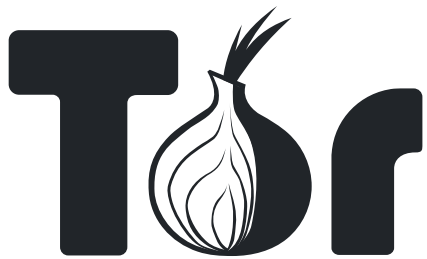
Safe measurement of live networks

draft-learmonth-pearg-safe-internet-measurement-01

Iain R. Learmonth
Tor Project

Monday 25th March
IETF 104, PEARG Session
Prague, Czech Republic



whoami

Iain R. Learmonth

irl@torproject.org
@irl@57n.org

Tor Metrics Team Member

Looking at Internet Measurement since 2013

Contributing to Tor Project since 2015

A8F7 BA50 41E1 3333 9CBA 1696 76D5 8093 F540 ABCD

What is safety?

- Safety $\neq$ Ethics
- Universities and research organizations do not currently have review boards equipped to evaluate Internet measurement research methods [5]
- Measurement of the technical specifics of censorship (what content the censor blocks, and technically how they impose the blocking) falls outside of human subjects research
- These measurements can still create risk for humans

What is safety?

When performing research on a platform shared with live traffic from other users, that research is considered safe if and only if other users are protected from or unlikely to experience danger, risk, or injury, now or in the future, due to the research.

Related Work

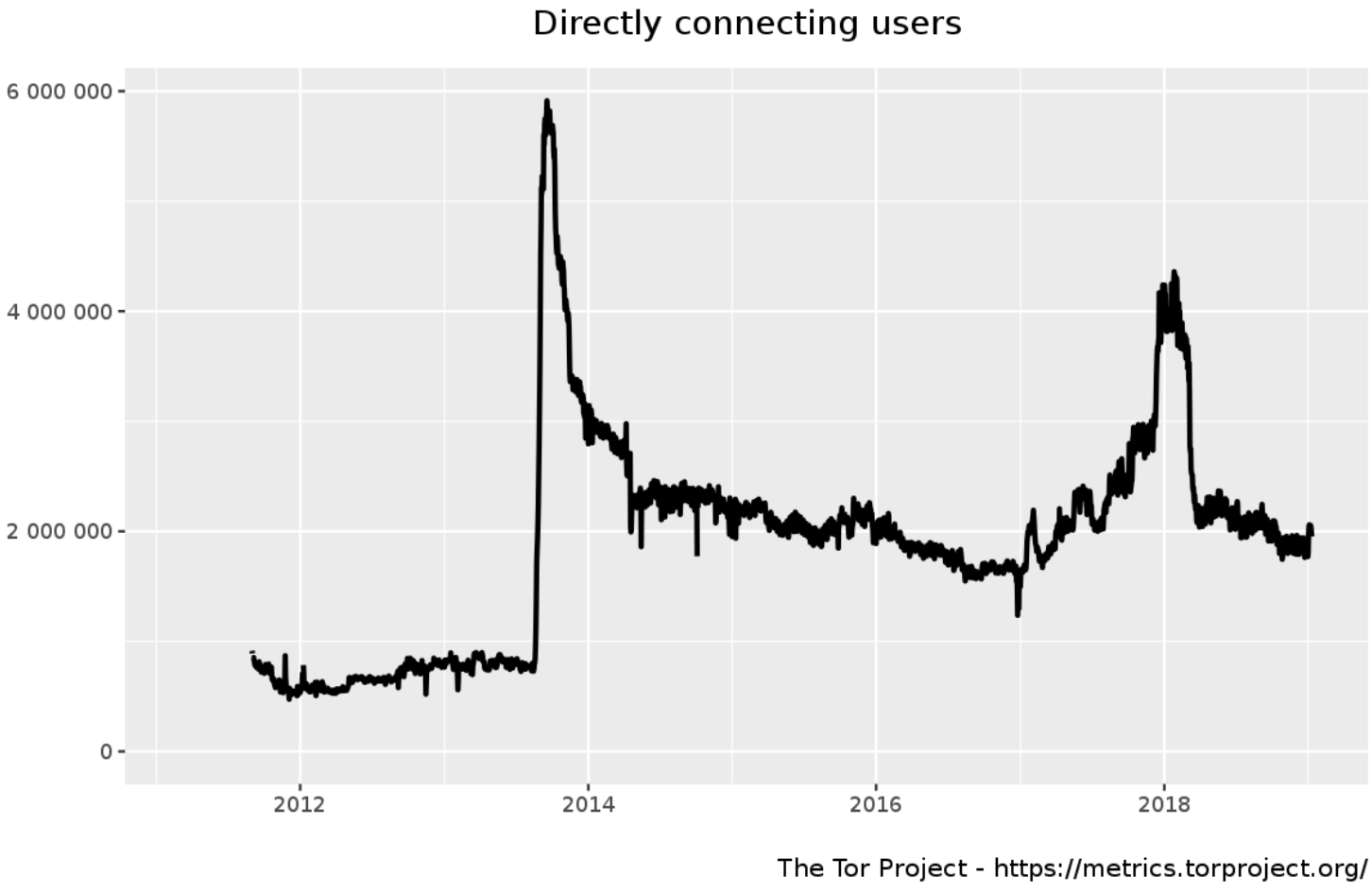
- Workshop on Ethics in Networked Systems Research
- CAIDA's Promotion of Data Sharing Webpage
- Menlo report and its companion
- EFF whitepaper: Unreliable Informants: IP Addresses, Digital Tips and Police Raids
- Tor Research Safety Board

What is Tor?

- Community of researchers, developers, users and relay operators
- Open Source
- Open Network
    - Security, Privacy, Anonymity, Robust, Authenticated, Integrity

https://torproject.org/

What is Tor?



Estimated average 2,000,000+ concurrent Tor users [7]

Use Cases

Data and analysis can be used to:

- detect possible censorship events
- detect attacks against the network
- evaluate effects on performance of software changes
- evaluate how the network is scaling

Tor Metrics Philosophy

We only handle public, non-sensitive data. Each analysis goes through a rigorous review and discussion process before publication.

Tor Research Safety Board

The goals of a **privacy and anonymity network** like Tor are not easily combined with *extensive data gathering*, but at the same time data is needed for monitoring, understanding, and improving the network.

Safety and privacy concerns regarding data collection by Tor Metrics are guided by the *Tor Research Safety Board's guidelines*.

`https://research.torproject.org/safetyboard.html`

Key Safety Principles

- Data Minimalisation
- Source Aggregation
- Transparency

Data Minimalisation

The first and most important guideline is that only the **minimum amount** of statistical data should be gathered to solve a given problem. The level of detail of measured data should be as **small as possible**.

“Over time, the probability that any entity holding a large store of sensitive private data will remain both competent enough to protect it adequately and honest enough to want to goes to zero.” —@mattblaze

Source Aggregation

Possibly sensitive data should exist for **as short a time as possible**. Data should be aggregated at its source, including categorizing single events and memorizing category counts only, summing up event counts over large time frames, and being imprecise regarding exact event counts.

“For almost every person on earth, there is at least one fact about them stored in a computer database that an adversary could use to blackmail, discriminate against, harass, or steal the identity of him or her. I mean more than mere embarrassment or inconvenience; I mean legally cognizable harm.” —Paul Ohm [8]

Transparency

All algorithms to gather statistical data need to be **discussed publicly** before deploying them. All measured statistical data should be made **publicly available** as a safeguard to *not gather data that is too sensitive*.

“Given enough eyeballs, all bugs are shallow” —Linus’ Law

Shortcut to Safety

- Use simulations
- Use a testbed

Case Study: Counting Unique Users of Tor

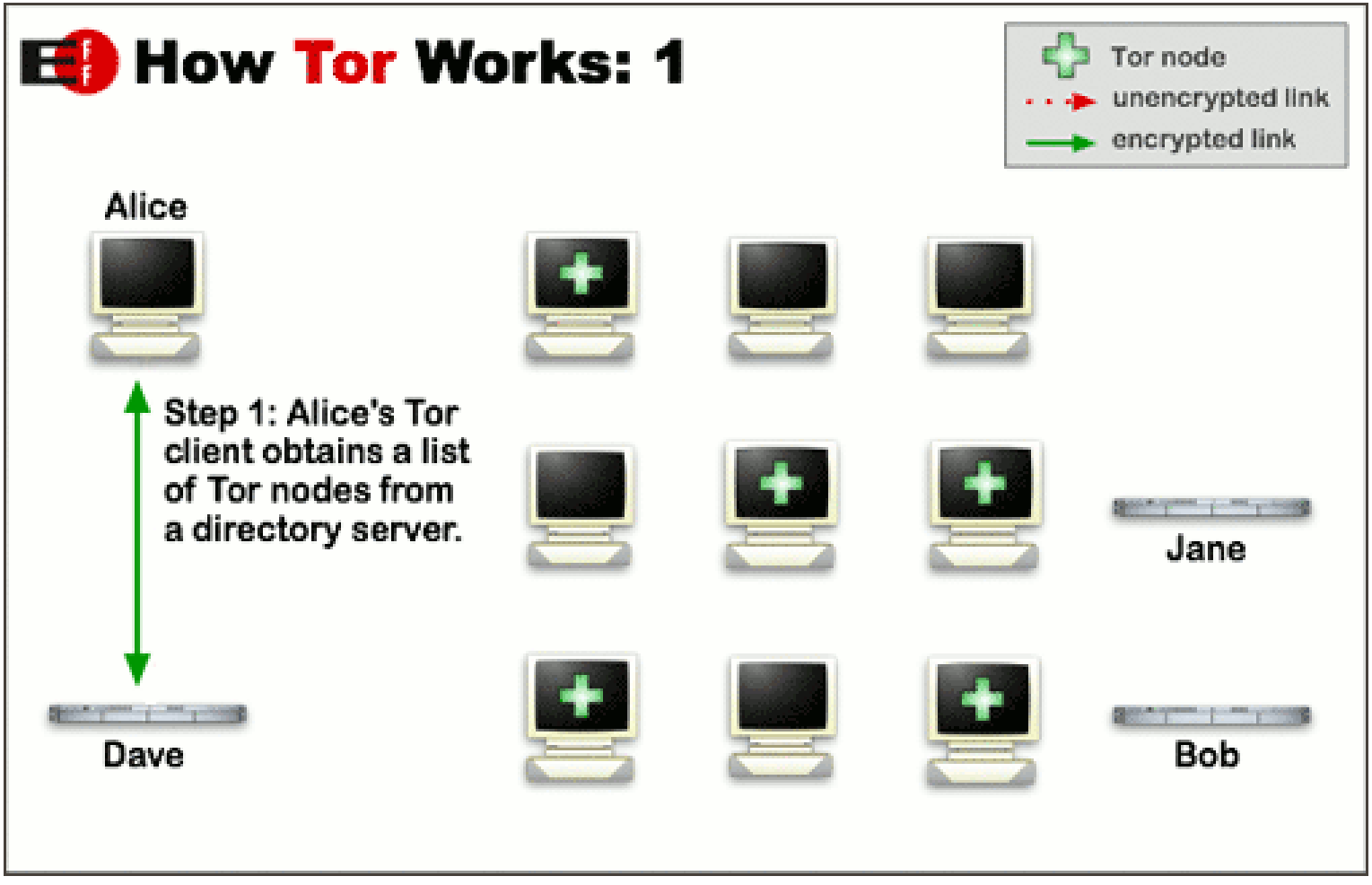
The Easy Way:

- Each relay keeps track of all the IP addresses it has seen
- These all get uploaded to a central location
- Unique IP addresses are counted

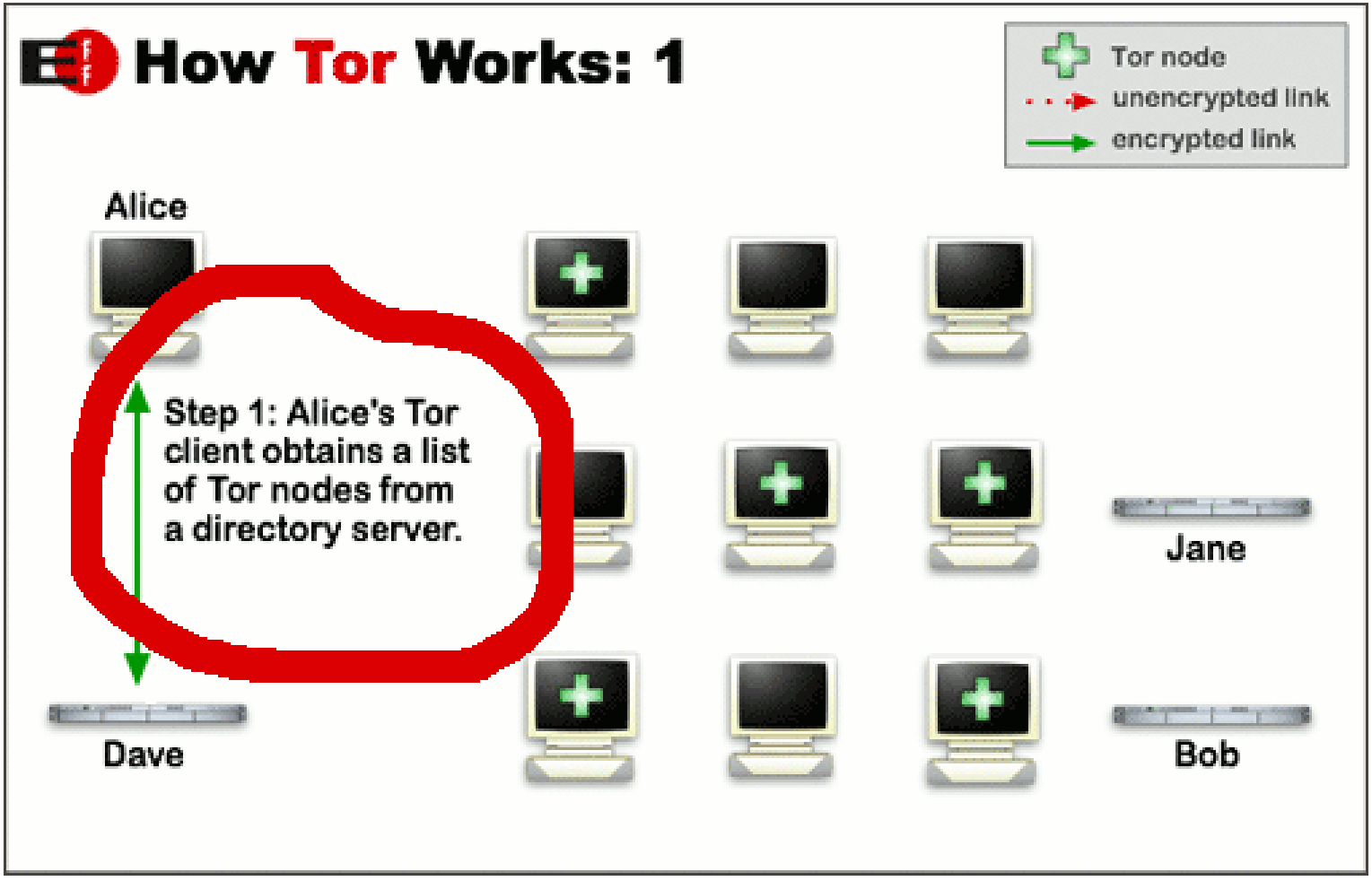
Indirect Measurement

In 2010, Tor Metrics set out to develop a safe method of counting users [3].

Indirect Measurement



Indirect Measurement



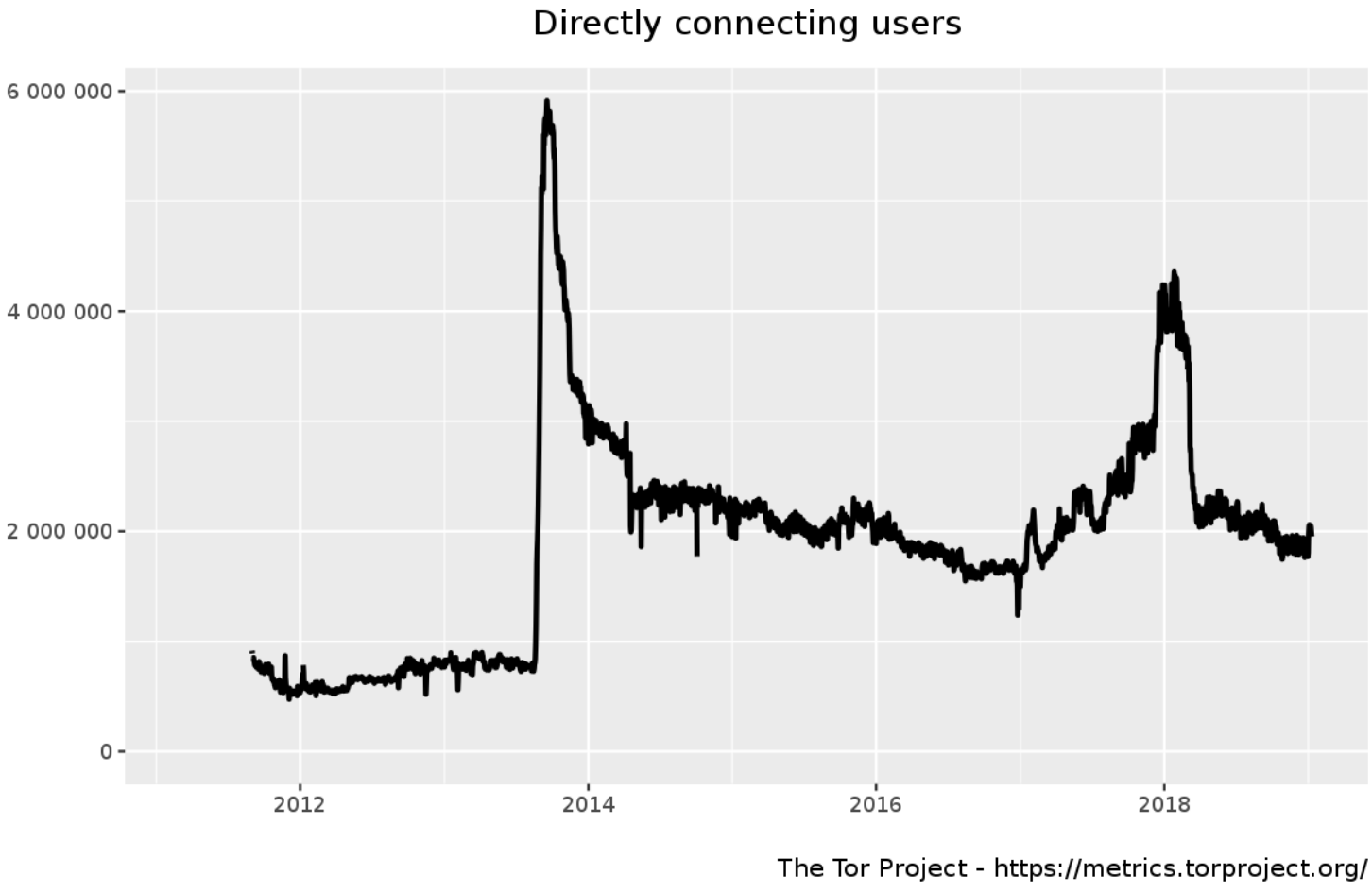
Indirect Measurement

The Safer Way:

- Relays don't store IP addresses at all
- Relays count number of directory requests
- Relays report numbers to a central location
- We have to guess how long an average session lasts
- We do not have the same detail in the data
- We still get the general ballpark figure and also see trends

https://metrics.torproject.org/reproducible-metrics.html

Indirect Measurement



Estimated average 2,000,000+ concurrent Tor users [7]

Count-Distinct Problem

Count-distinct problem

From Wikipedia, the free encyclopedia

In computer science, the **count-distinct problem**[1] (also known in applied mathematics as the **cardinality estimation problem**) is the problem of finding the number of distinct elements in a data stream with repeated elements. This is a well-known problem with numerous applications. The elements might represent IP addresses of packets passing through a router, unique visitors to a web site, elements in a large database, motifs in a DNA sequence, or elements of RFID/sensor networks.

HyperLogLog

Let $h : \mathcal{D} \to [0,1] \equiv \{0,1\}^\infty$ *hash data from domain* $\mathcal{D}$ *to the binary domain.*
Let $\rho(s)$*, for* $s \in \{0,1\}^\infty$*, be the position of the leftmost 1-bit* ($\rho(0001\cdots) = 4$).

Algorithm HYPERLOGLOG (**input** $\mathcal{M}$: multiset of items from domain $\mathcal{D}$).
assume $m = 2^b$ with $b \in \mathbb{Z}_{>0}$;
initialize a collection of m registers, $M[1], \ldots, M[m]$, to $-\infty$;

for $v \in \mathcal{M}$ **do**
    set $x := h(v)$;
    set $j = 1 + \langle x_1 x_2 \cdots x_b \rangle_2$; {*the binary address determined by the first* b *bits of* x}
    set $w := x_{b+1} x_{b+2} \cdots$; **set** $M[j] := \max(M[j], \rho(w))$;

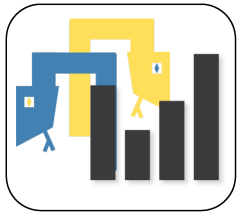
compute $Z := \left(\sum_{j=1}^{m} 2^{-M[j]} \right)^{-1}$; {*the "indicator" function*}
return $E := \alpha_m m^2 Z$

Algorithm designed for very large data sets [2] where you don't want to keep all the unique items around.

U.S. NAVAL RESEARCH LABORATORY

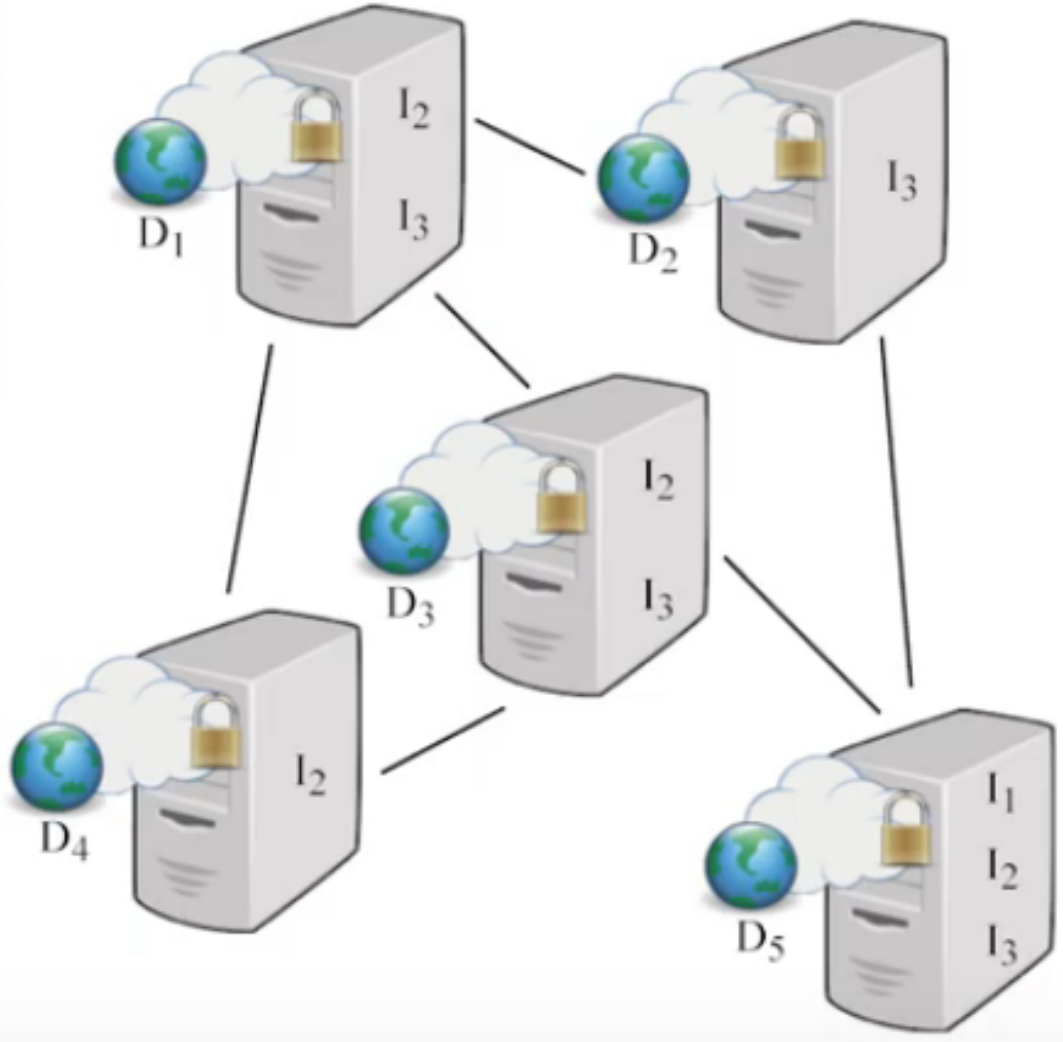
PrivCount: Overview

Distributed measurement system



- "Privacy-preserving counting" system
  - Tracks various types of Tor events, computes statistics from those events
  - Based on PrivEx-S2 by Elahi et al. (CCS 2014)
- Distributes trust using secret sharing across many operators
- Achieves forward privacy during measurement
  - the adversary cannot learn the state of the measurement before time of compromise
- Provides differential privacy of the results
  - prevents confirmation of the actions of a specific user given the output

Private Set Union Cardinality



- How many **unique** items are there, across a set of distributed private datasets?

$$|D_1 \cup D_2 \cup \ldots D_5| = |\{I_1, I_2, I_3\}| = 3$$

- Requirements
  - Input must stay private
  - Only output should be revealed

Other Privacy-Preserving Telemetry Schemes

- RAPPOR
  https://security.googleblog.com/2014/10/learning-statistics-with-privacy-aided.html
- PROCHLO
  https://ai.google/research/pubs/pub46411
- Prio
  https://hacks.mozilla.org/2018/10/testing-privacy-preserving-telemetry-with-prio/

draft-learmonth-pearg-safe-internet-measurement

[Docs] [txt|pdf|xml|html] [Tracker] [Email] [Diff1] [Diff2] [Nits]

Versions: 00 01

Network Working Group — I. Learmonth
Internet-Draft — Tor Project
Intended status: Informational — December 12, 2018
Expires: June 15, 2019

Guidelines for Performing Safe Measurement on the Internet
draft-learmonth-pearg-safe-internet-measurement-01

Abstract

Researchers from industry and academia will often use Internet measurements as a part of their work. While these measurements can give insight into the functioning and usage of the Internet, they can come at the cost of user privacy. This document describes guidelines for ensuring that such measurements can be carried out safely.

Work-in-progress in the IRTF [6]
(Discussion in the proposed Privacy Enhancements and Assessments Research Group (PEARG))

Next Steps

- Comprehensive general considerations checklist for any measurement
- Introduction to literature on secret sharing/multi-party computation telemetry systems (this may fall out of scope)
- Thinking about future computing power available
- Discussion on consent and proxy consent
- Ensure all types of harm are considered, e.g. unavailability

https://github.com/irl/draft-safe-internet-measurement/issues

References I

[1] Ellis Fenske, Akshaya Mani, Aaron Johnson, and Micah Sherr.
Distributed measurement with private set-union cardinality.
In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, CCS '17, pages 2295–2312, New York, NY, USA, 2017. ACM.

[2] Philippe Flajolet, Éric Fusy, Olivier Gandouet, and Frédéric Meunier.
HyperLogLog: the analysis of a near-optimal cardinality estimation algorithm.
In Philippe Jacquet, editor, *AofA: Analysis of Algorithms*, volume DMTCS Proceedings vol. AH, 2007 Conference on Analysis of Algorithms (AofA 07) of *DMTCS Proceedings*, pages 137–156, Juan les Pins, France, June 2007. Discrete Mathematics and Theoretical Computer Science.

[3] Sebastian Hahn and Karsten Loesing.
Privacy-preserving ways to estimate the number of Tor users.
Technical Report 2010-11-001, The Tor Project, November 2010.

References II

[4] Rob Jansen and Aaron Johnson.
Safely measuring tor.
In *Proceedings of the 23rd ACM Conference on Computer and Communications Security (CCS '16)*, October 2016.

[5] Ben Jones, Roya Ensafi, Nick Feamster, Vern Paxson, and Nick Weaver.
Ethical concerns for censorship measurement.
In *Proceedings of the 2015 ACM SIGCOMM Workshop on Ethics in Networked Systems Research*, NS Ethics '15, pages 17–19, New York, NY, USA, 2015. ACM.

[6] Iain Learmonth.
Guidelines for performing safe measurement on the internet.
Internet-Draft draft-learmonth-pearg-safe-internet-measurement-01, IETF Secretariat, December 2018.
http://www.ietf.org/internet-drafts/draft-learmonth-pearg-safe-internet-measurement-01.txt.

References III

[7] Karsten Loesing, Steven J. Murdoch, and Roger Dingledine.
A case study on measuring statistical data in the Tor anonymity network.
In *Proceedings of the Workshop on Ethics in Computer Security Research (WECSR 2010)*, LNCS. Springer, January 2010.

[8] Paul Ohm.
Broken promises of privacy: Responding to the surprising failure of anonymization.
UCLA Law Review, 57:1701, 2009.