A Case Study on Measuring Statistical Data in the Tor Anonymity Network

Karsten Loesing[1], Steven J. Murdoch[1,2], Roger Dingledine[1]

[1] The Tor Project
[2] Computer Laboratory, University of Cambridge, UK

Workshop on Ethics in Computer Security Research (WECSR 2010)

Motivation

* Largest anonymity network
* Lots of users - and diversity
* Who uses Tor? Why is it slow? Does it still work in China?

Research questions

* What do people do when they're private?
* Can we deanonymize users by content? By traffic?
* Can we give them a Java applet to unmask them?
* What application protocols do they use? What languages are the web pages?
* How much SSL? Do they check SSL fingerprints?

Funders want us to

* Track user growth in these 18 countries
* Show that Tor is getting faster
* React quickly to blocking events

Papers on Tor usage

* Colorado paper - and their data set
* Yale paper that never got written
* Bunch of industry people trying to drum up business
* Wiretapping? Pen register? Foreign laws? California's bilateral consent law

Our suggestion

* The only people writing the papers were the ones doing it wrong
* So we shifted from
  * "Don't do that! You might get it wrong" to
  * "Here are some guidelines for getting it right"

Outline

* Principles to choose from
* Background on Tor Anonymity Network
* Case study on measuring statistical data without hurting users' privacy
* Discussion of general guidelines

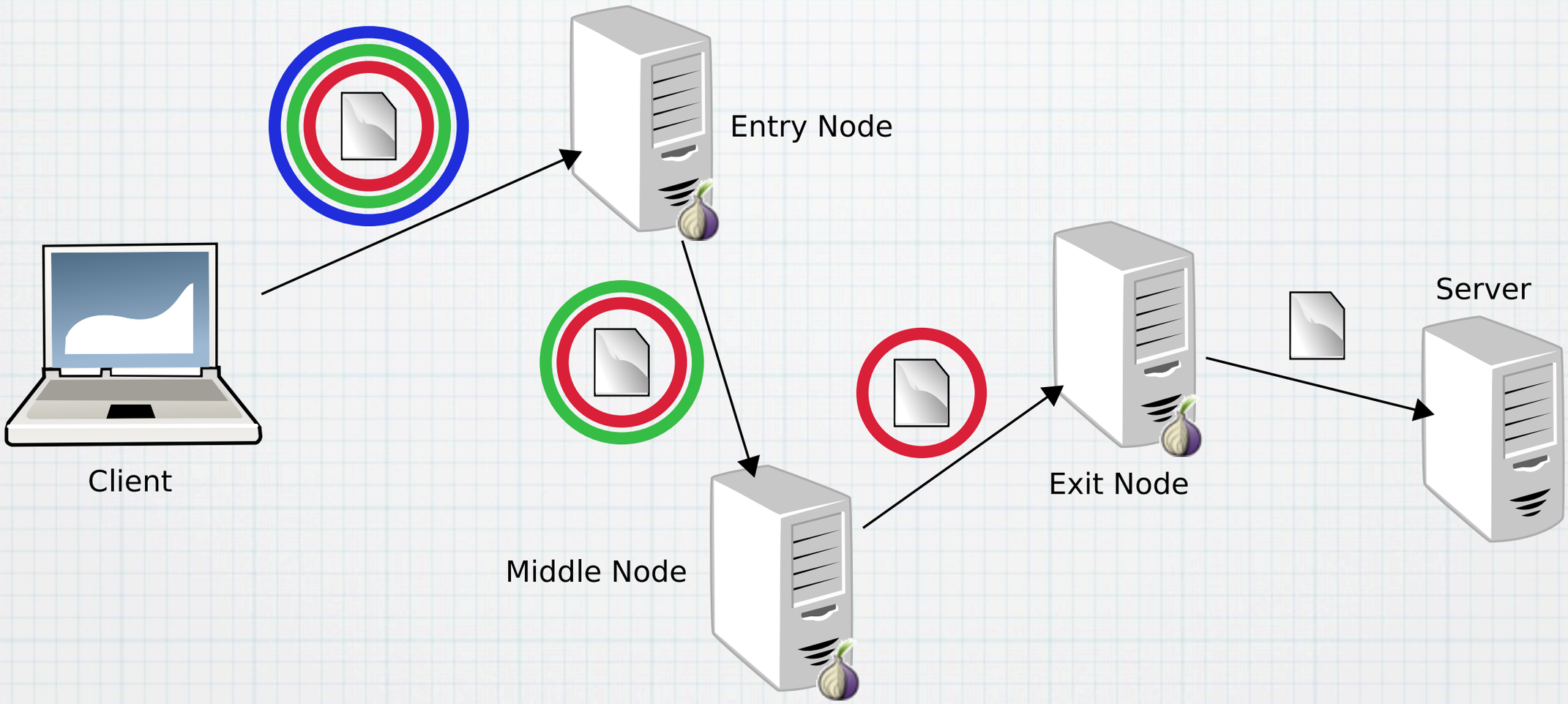
Principles to choose from

Principles to choose from

* Legal requirements
* User privacy
* Ethical approvement
* Informed consent
* Community acceptance

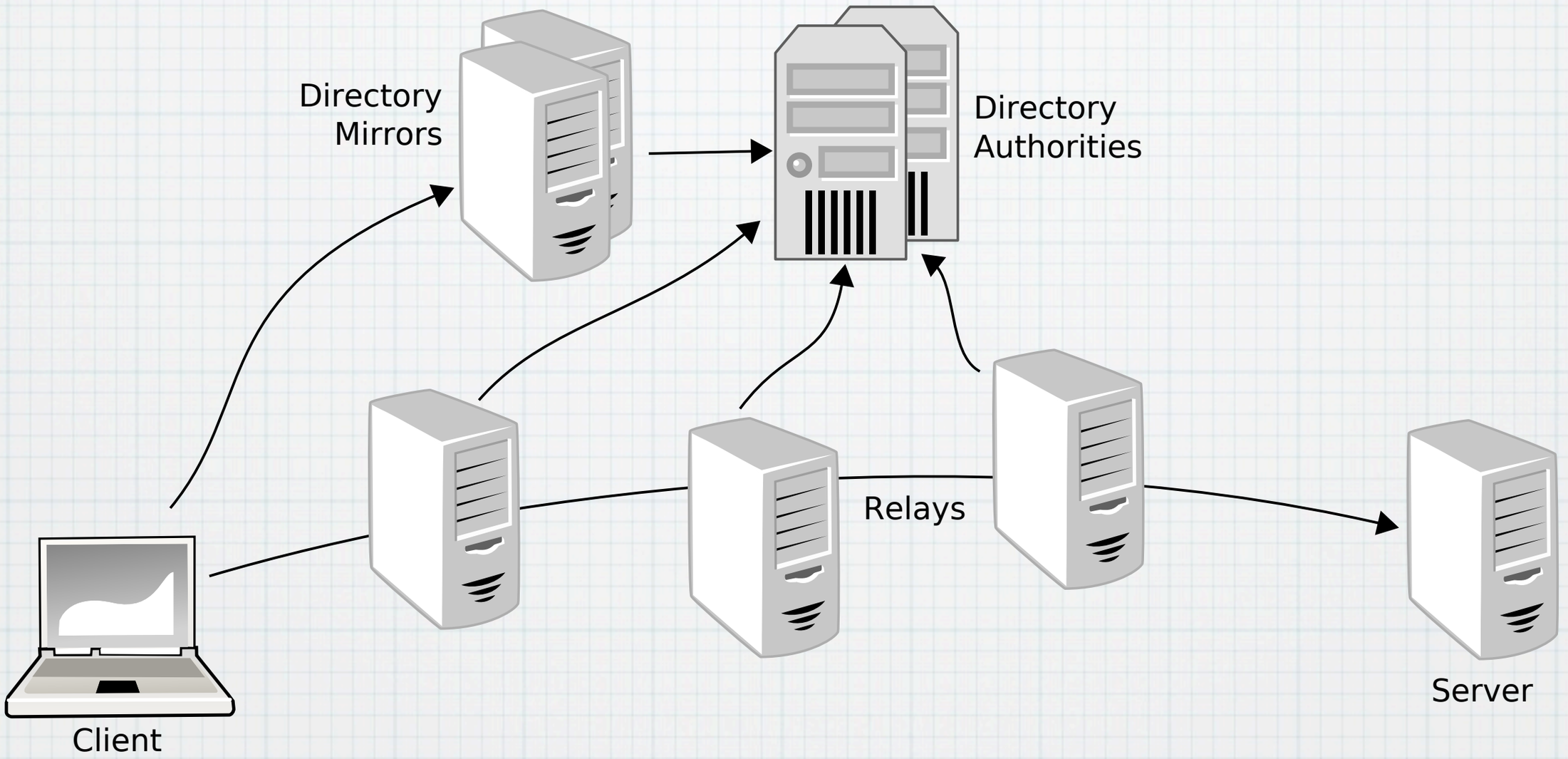
Background on Tor

Onion routing



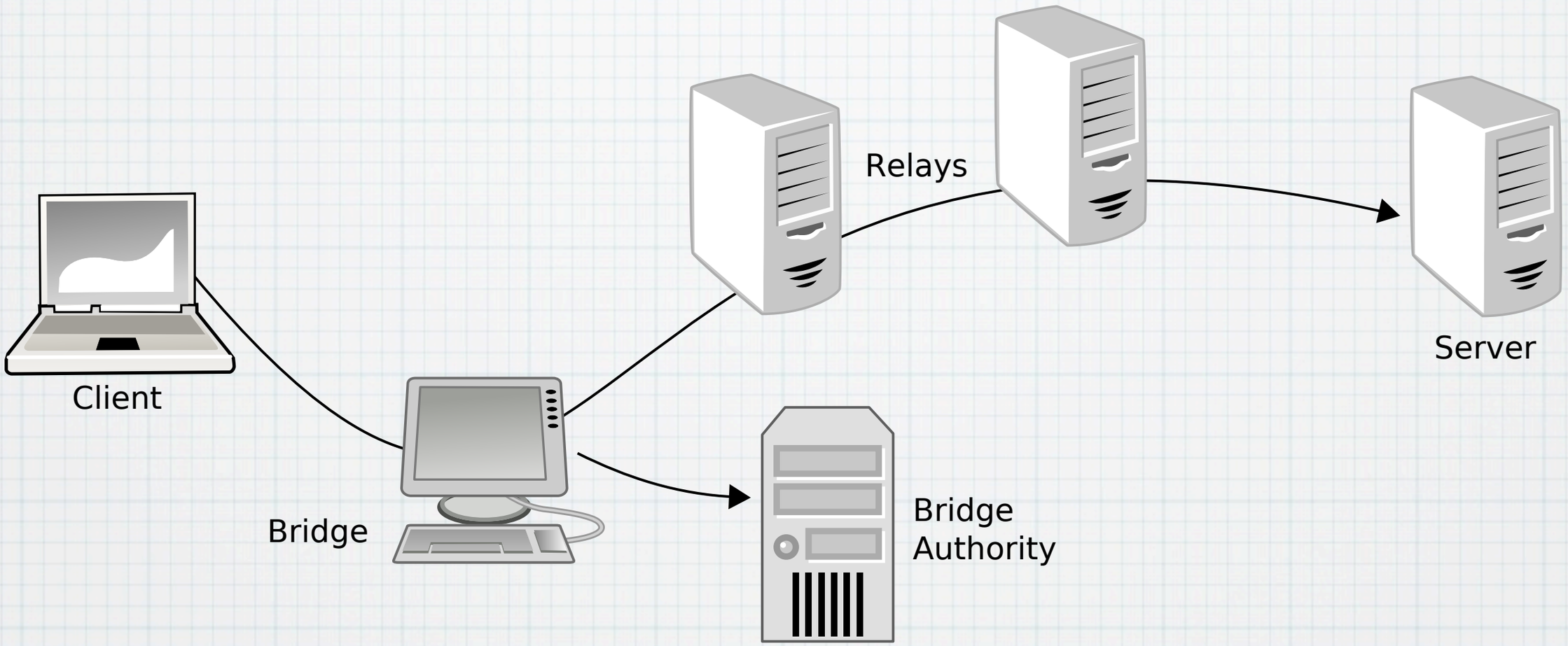
* User remains anonymous as long as not both entry and exit node hop are logging

Directory system



* Clients learn about relays from directories to select paths and build circuits

Bridge relays



* Bridge relays = non-public relays given out to blocked clients via email or website

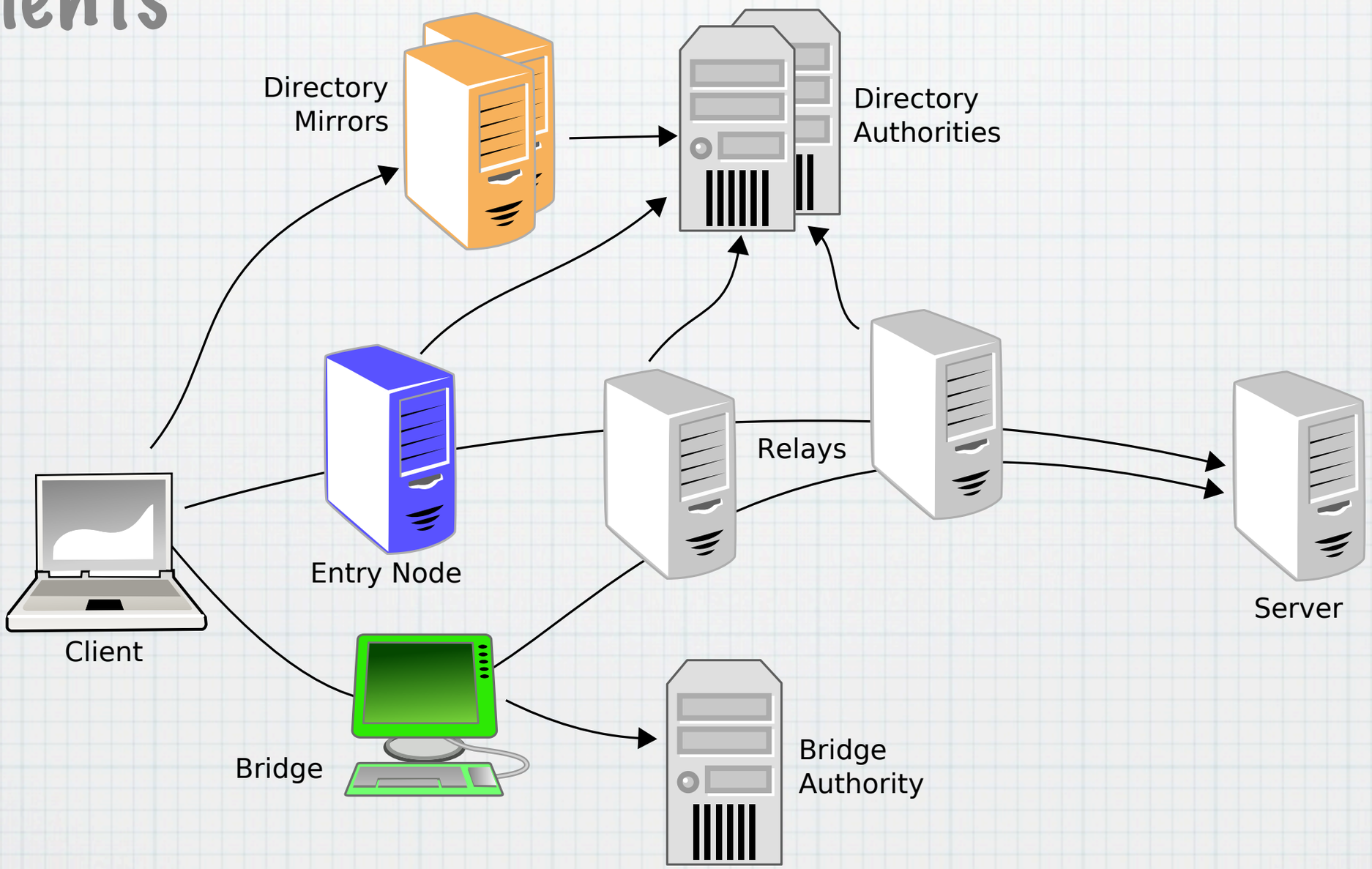
Case study

Statistics in Tor

* # relays, versions, dynamic IPs (HotPETS 2009)
* Bridge churn, # bridges for usable Tor
* Performance: torperf, circuit build times
* Usage: # users in total/by country,
  # bytes per port for improved load balancing

Who uses Tor?

* Learn IP addresses from directly connecting clients



Privacy problem

* IP addresses are highly sensitive data
* Don't leak IP addresses; even though using Tor is not secret/protected
* Don't allow adversary to correlate client IP addresses with exit traffic!

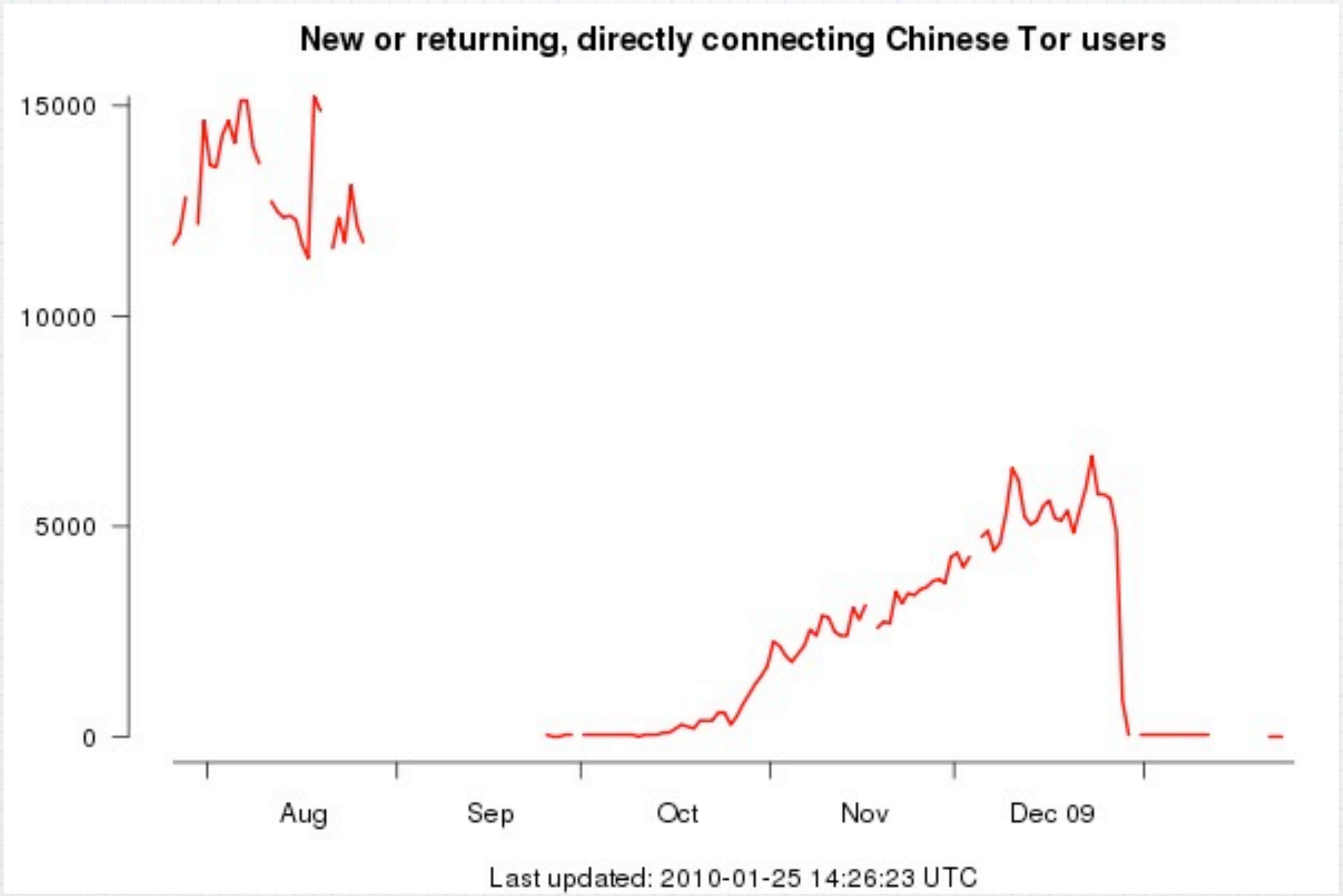
Data aggregation

* Resolve IP addresses to countries using GeoIP database (2.5 MiB) ASAP; never write to disk
* Only report data of 24h intervals
* Don't be too precise in numbers; round up

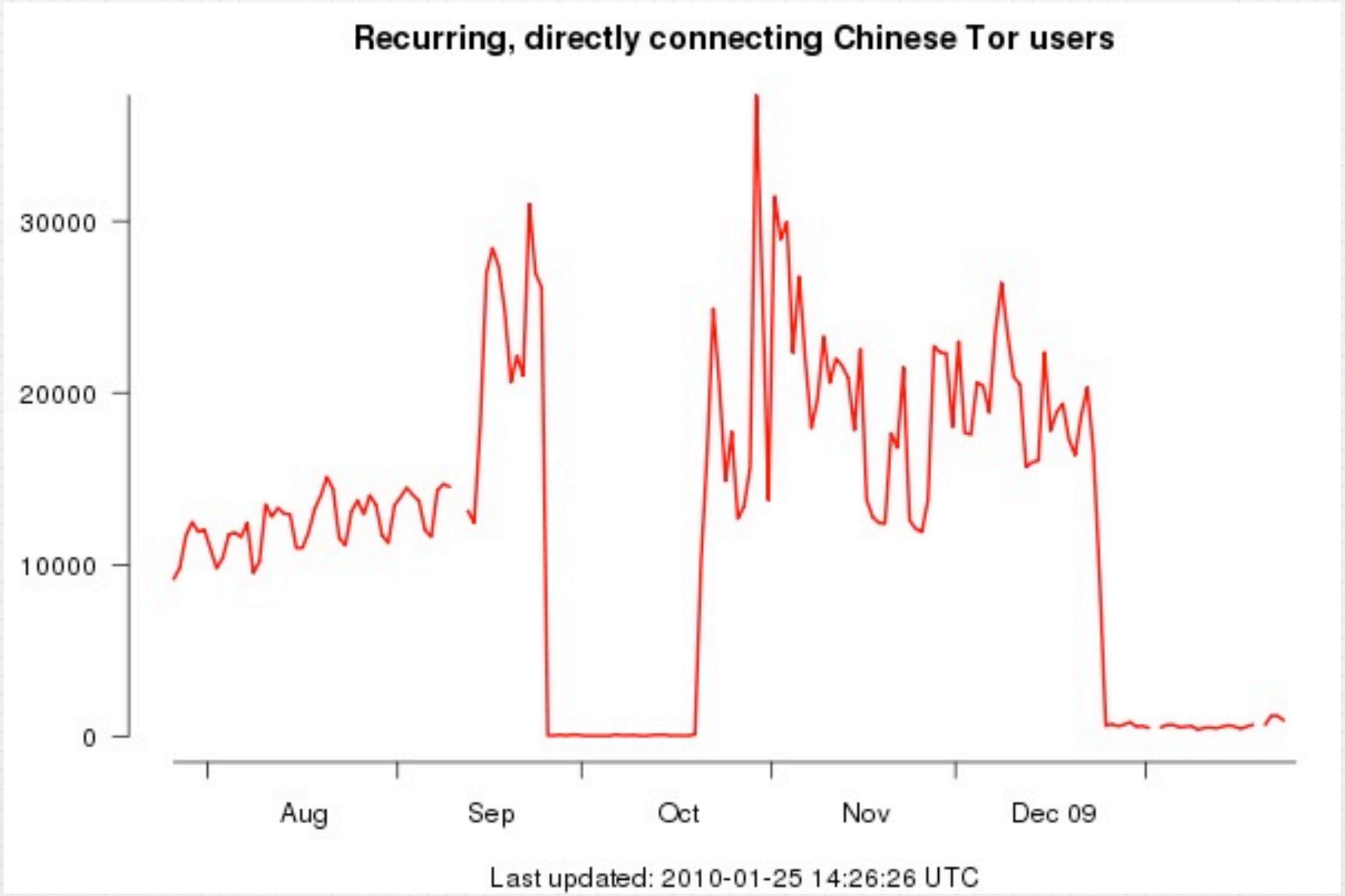
Example data

```
dirreq-stats-end 2009-08-20 17:16:35 (86400 s)
dirreq-v2-ips us=4136,de=3744,cn=3552,gb=1120,ir=1024,[...]
dirreq-v3-ips us=6024,de=5176,cn=3384,fr=2208,kr=1328,[...]
dirreq-v2-reqs us=7136,cn=5608,de=4728,kr=3816,gb=1568,[...]
dirreq-v3-reqs us=7800,de=5944,kr=4368,cn=4208,fr=2632,[...]
```

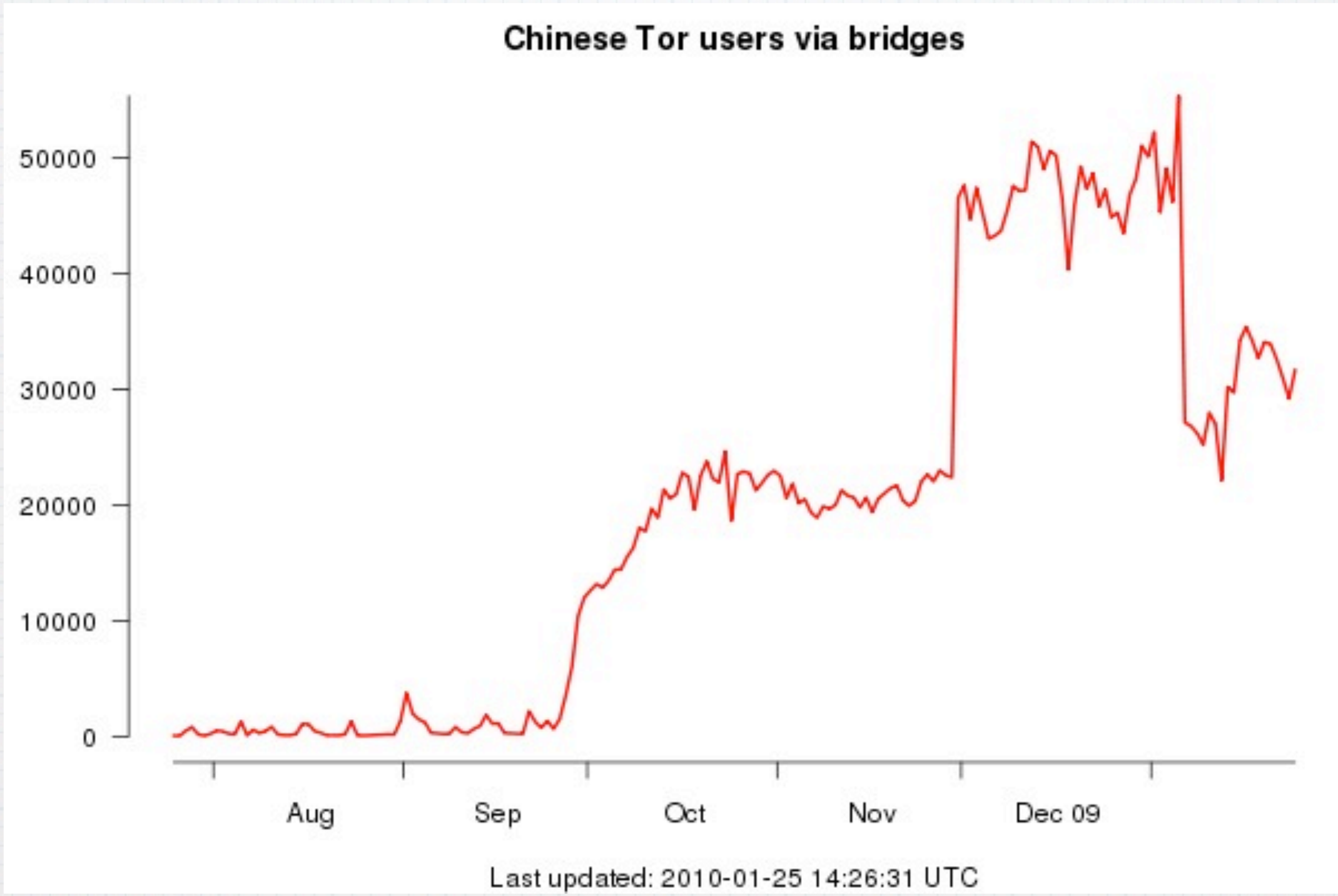
China 1/3


New or returning, directly connecting Chinese Tor users
15000
10000
5000
0
Aug
Sep
Oct
Nov
Dec 09
Last updated: 2010-01-25 14:26:23 UTC

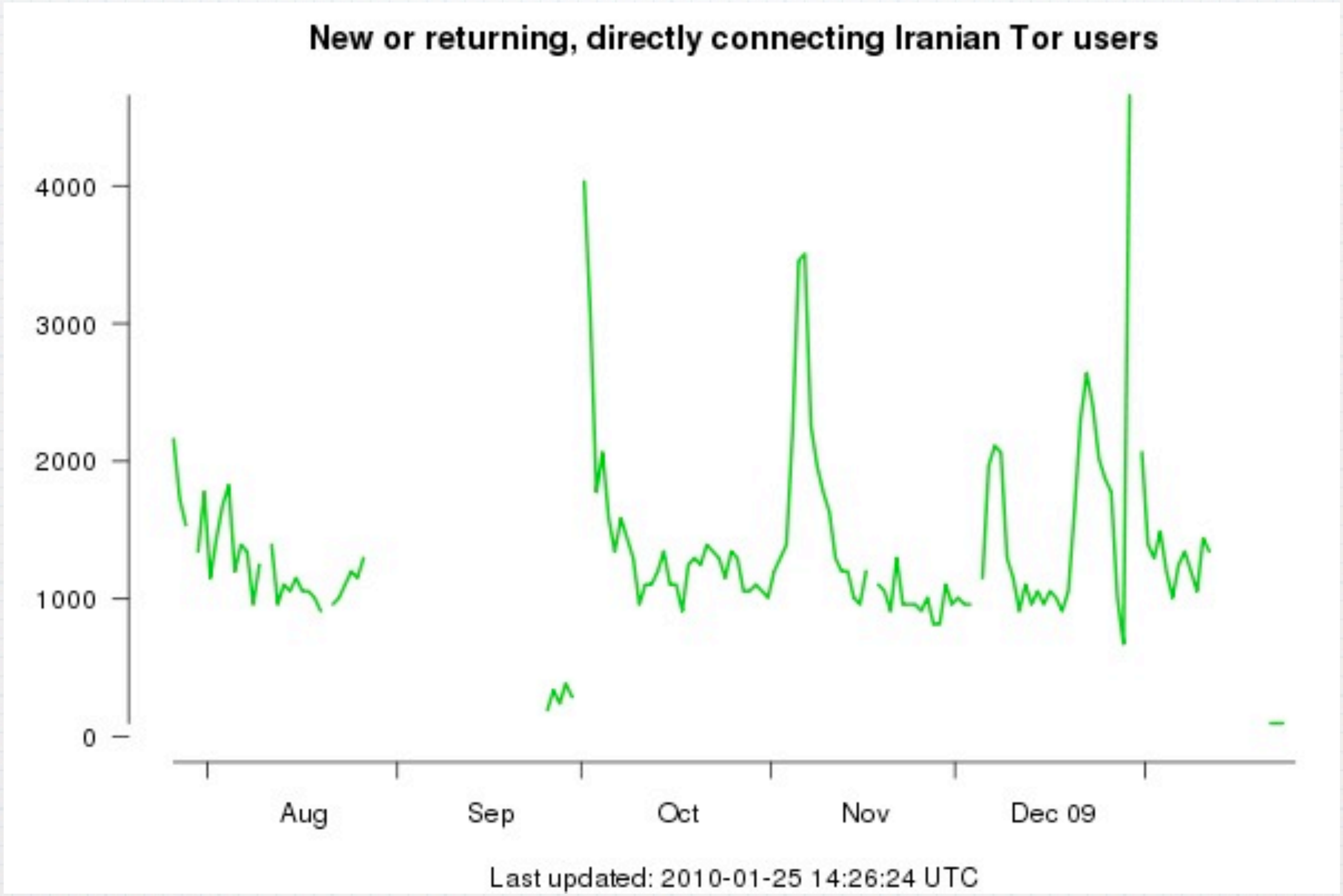
China 2/3


Recurring, directly connecting Chinese Tor users
30000
20000
10000
0
Aug
Sep
Oct
Nov
Dec 09
Last updated: 2010-01-25 14:26:26 UTC

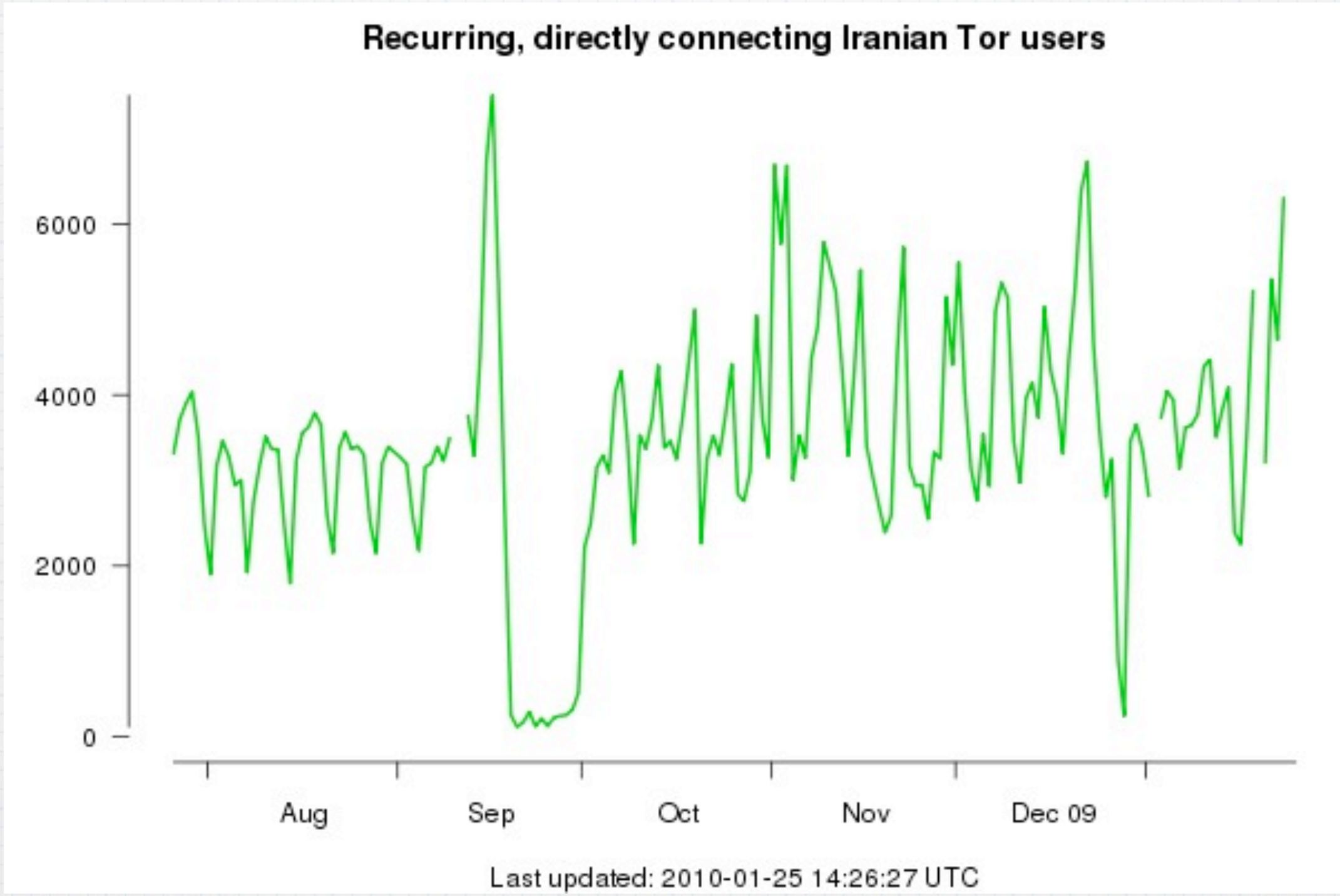
China 3/3


Chinese Tor users via bridges
50000
40000
30000
20000
10000
0
Aug
Sep
Oct
Nov
Dec 09
Last updated: 2010-01-25 14:26:31 UTC

Iran 1/4


New or returning, directly connecting Iranian Tor users
4000
3000
2000
1000
0
Aug
Sep
Oct
Nov
Dec 09
Last updated: 2010-01-25 14:26:24 UTC

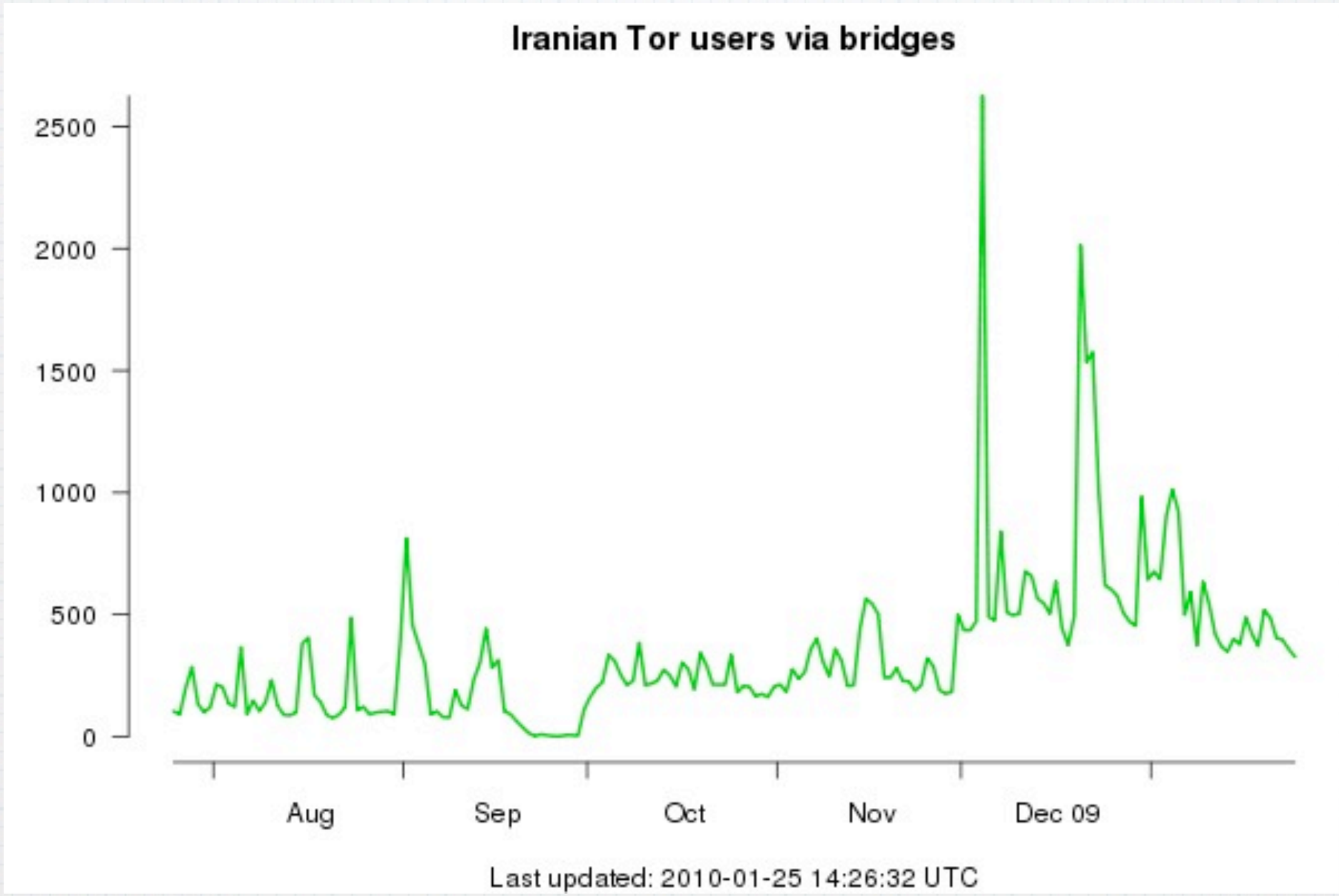
Iran 2/4


Recurring, directly connecting Iranian Tor users
6000
4000
2000
0
Aug
Sep
Oct
Nov
Dec 09
Last updated: 2010-01-25 14:26:27 UTC

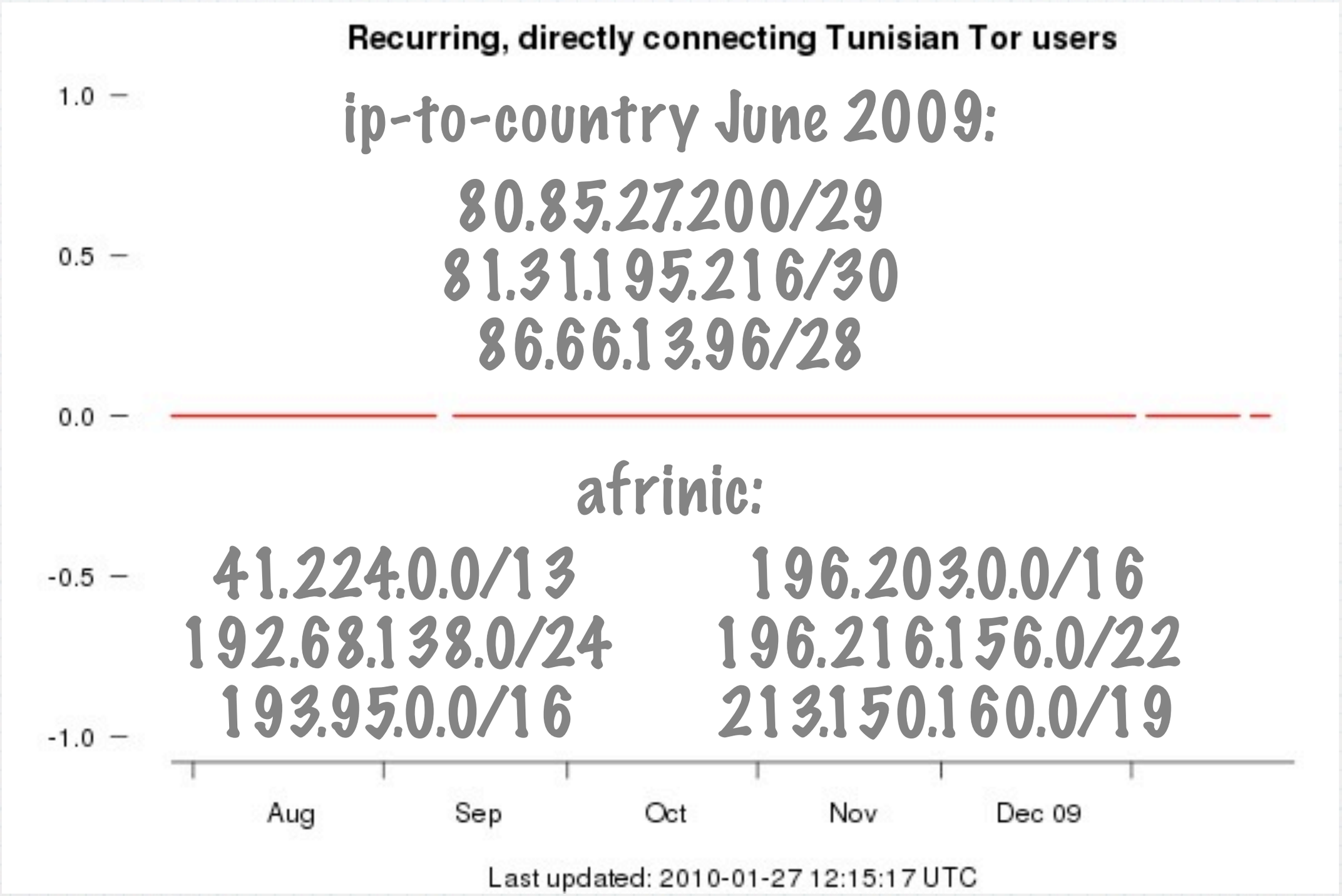
Iran 3/4

Iranian bridge users relative to June 1, 2009

Iran 4/4


Iranian Tor users via bridges
2500
2000
1500
1000
500
0
Aug
Sep
Oct
Nov
Dec 09
Last updated: 2010-01-25 14:26:32 UTC

Tunisia



ip-to-country June 2009:

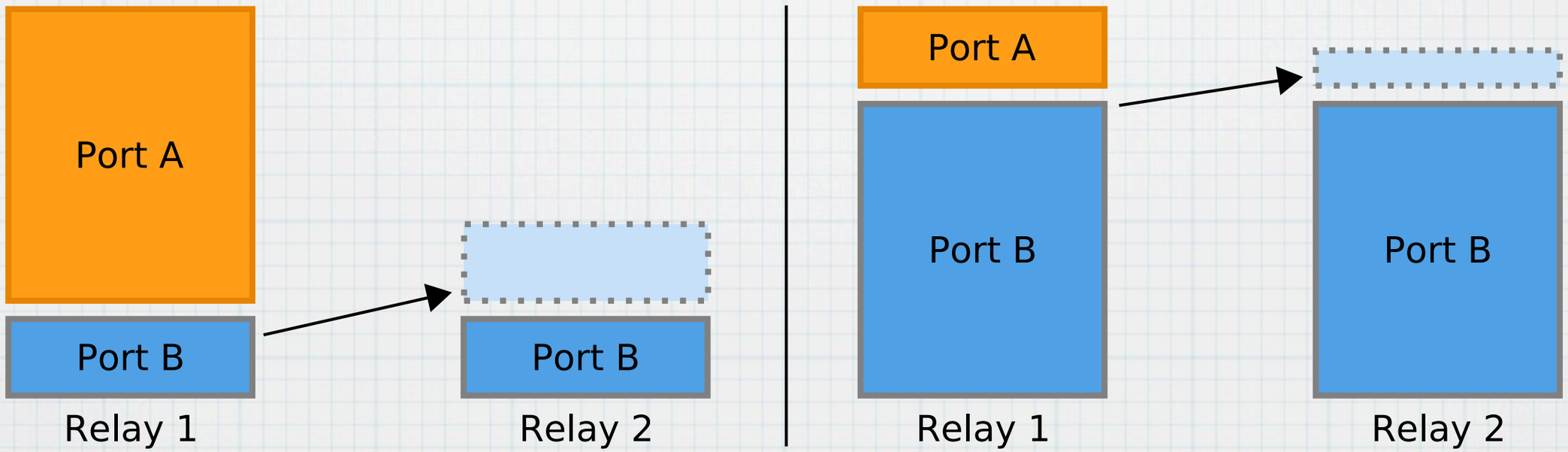
80.85.27.200/29
81.31.195.216/30
86.66.13.96/28

afrinic:

41.224.0.0/13
192.68.138.0/24
193.95.0.0/16
196.203.0.0/16
196.216.156.0/22
213.150.160.0/19

What is Tor used for?

* How much traffic exits Tor network by port?
* Improve load balancing among relays with different exit policies
* Shift traffic towards more restrictive exit nodes



Privacy problem

* Exit traffic is highly sensitive data
* Contents (possibly unencrypted) and server addresses must not be disclosed; even without knowing clients
* Don't allow adversary to correlate exit traffic with client IP addresses!

Data aggregation

* Only remember ports and written/read bytes
* Only report data of 24h intervals
* Discard data for ports below threshold
* Don't be too precise in numbers; round up

Example data

```
exit-stats-end 2009-07-24 20:40:35 (86400 s)
exit-kibibytes-written 17=58902,23=9616,25=262579,
  40=9546,76=5789,80=681732,[...],other=15332199
exit-kibibytes-read 17=15,23=79,25=13221,40=7,76=2,
  80=1841879,85=926,143=1038,222=85,[...],other=3035782
exit-streams-opened 17=12,23=88,25=141240,40=12,76=16,
  80=867896,85=2704,143=168,222=32,[...],other=3165052
```

metrics.torproject.org

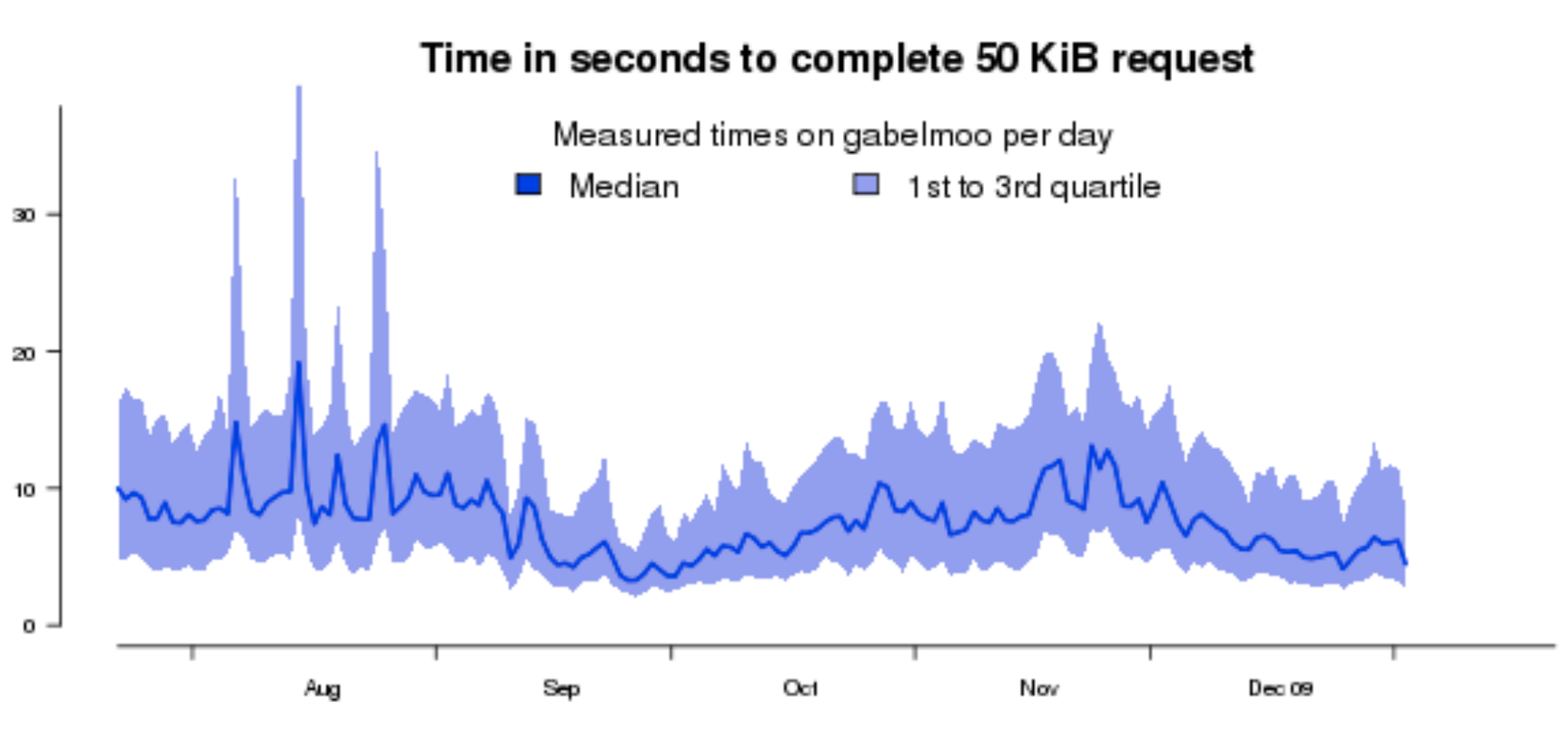
Tor Metrics Portal: Graphs

http://metrics.torproject.org/graphs.html

Google

Time to complete requests

The following graphs show the performance of the Tor network as experienced by its users. The graphs contain the average (median) time to request files of three different sizes over the network as well as first and third quartile of request times.



Discussion

Guidelines

* Data minimalism: Do we really need data?
* Source aggregation: How to measure safely?
* Transparency: Publish process and data