Workshop on Internet Economics (WIE 2020) Final Report

kc claffy
UCSD/CAIDA
kc@caida.org

David Clark
MIT/CSAIL
ddc@csail.mit.edu

This article is an editorial note submitted to CCR. It has NOT been peer reviewed.
The authors take full responsibility for this article's technical content. Comments can be posted through CCR Online.

ABSTRACT

On 16-17 December 2020, CAIDA hosted the 11th interdisciplinary Workshop on Internet Economics (WIE) in a virtual Zoom conference. This year our goal was to gather feedback from researchers on their experiences using CAIDA's data for economics or policy research. We invited all researchers who reported use of CAIDA data in these disciplines. We discussed their successes and challenges of using the data, and how CAIDA could help these fields via Internet measurement and data curation. To avoid Zoom fatigue, we had a conversation-focused rather than presentation-focused workshop. Research topics we discussed included: Internet data for macroeconomics; connectivity and its effect on economic interdependence; effects of the EU's new GDPR on internet interconnection; measuring corporate cyber risk; measuring work-from-home trends; measuring the economic value of open source software; and more generally how to best support evidence-based policymaking.

CCS CONCEPTS

• **Networks → Public Internet**; • **Social and professional topics → Economic impact**; **Governmental regulations**.

KEYWORDS

Economics, Internet, Measurement, Security

1 INTERNET DATA FOR MACROECONOMICS

CAIDA's topology data sets include information about interconnection of IP interfaces, routers, autonomous networks, which networks own which routers, hostnames associated with observed router interfaces, IXP locations, etc. Although the data and annotations are rich, they do not provide sufficient data for some macroeconomics questions. The data may not cleanly match country-level statistics, or the time horizon may be too short.

For example, in the field of international macroeconomics, researchers might ask whether trade and financial links between countries align with interconnection links. Low latency promotes trade, just as in the physical world. Indeed, the structure of the Internet may reflect trade agreements between countries, as well as a given nation's power to influence other countries since countries may examine, disrupt, or otherwise control data flowing through them. But quantifying these relationships is elusive.

Many macroeconomic studies require adjacency matrices across countries. Relevant variables include: bandwidth capacity between countries, bandwidth usage over a given period, bilateral latency (round trip time); cost of transit, and number/type of users. Few data sources reveal bandwidth between countries. An open challenge is validation of whether any proxy measurement, e.g., observable paths, would suffice. Furthermore, spatial regression studies require *precise* geolocation data for nodes: IXPs, PoPs, core routers, and data centers. Some data is available from commercial providers such as Telegeography and Infrapedia, although its accuracy is unknown. The RIPE NCC's Atlas[1] data could provide latency data necessary to build an adjacency matrix.

Economists also seek empirical data to determine the effect of internet access on market outcomes. Internet access positively impacts economic activity, but there are disagreements about how to measure (or define) Internet access. Network measurement necessarily covers existing, not missing, access.

2 GDPR'S EFFECTS ON INTERCONNECTION

Economic data has revealed that the European Union (EU)âĂŹs General Data Protection Regulation (GDPR) legislation has imposed significant costs to application firms. Recent press has also asserted, without substantiation, that GDPR could have massive impact on the economy as a whole and the

interconnection market in particular.[1] Harvard researchers used CAIDA data to study whether the GDPR legislation had measurable effects on internet interconnection investment [2]. This study used a large sample of interconnection agreements as reflected in CAIDA's AS Relationships data [3], and several other CAIDA data sets: IP Prefix Probing Traceroutes[4], AS-to-Organization Mapping [5]; prefix-to-AS mapping data [6]. Given the perceived reduction of demand at the application layer (estimated at 10%), some thought that reduction would result in less need for infrastructure given fewer [streaming] ads. The study used data from before (2015) and after (2019) GDPR went into effect, and found no statistical change at any measurable margin in comparison to controls. This result provided rare empirical underpinning for a tense policy debate. Further discussion at the workshop entertained a thought experiment. What if vetted parties could query, compare, and analyze country-to-country interconnection agreements? Could meta-data about such agreements (number, port count, capacities, IP addresses announced per interconnection agreement) serve as proxy indicators of growth?

3 MEASURING CORPORATE CYBER RISK

Malicious events such as DDoS attacks, cyber break-ins, and phishing campaigns continue to make daily news. Corporations struggle to determine how to measure their own cyber risk. Cyber insurance is expensive partly because we do not have the data to do effective actuarial analysis in the cyber domain. Further, regulators require insurance companies to retain large reserves to offset the risk. Often insurers will only cover business interruption.

Researchers need improved methods of mapping global indexes of publicly traded companies to the network and service information we can measure. But challenges remain even with such linkages, e.g, subsidiary entities with shared or similar strings in the names. Risk assessment is again reduced to proxy information. For example, measuring how often version upgrades happen can provide a window into the cyber hygiene of a given company. Companies that upgrade more often, will likely have better records on managing cyber risk.

4 MEASURING WORKING FROM HOME

With the global pandemic starting in Q2 2020, the world shifted. Much of the U.S. labor force began working from home full-time, accounting for more than 65% of U.S. economic activity [7]. Measurement of this shift is a challenge. Analysts can create digital footprints to track which companies employ workforces at home, patterns of content flows, and economic impact on companies and employees. They can use this data to compare performance and other outcomes by geographic and network region.

[1] See Section 3 and especially footnote 46 of [2].

5 REGIONALIZATION OF TRAFFIC

The evolution in Internet interconnection and usage inspires other questions regarding its impact on Internet topology and peering relationships. Has the pandemic sped the shortening of paths on the internet? When looking at interconnection and paths on the internet, policymakers would like to know, what fraction of traffic stays local, and how local does it stay? Are paths becoming so short that jurisdiction over it might be intrastate rather than interstate, with implication for state vs. federal regulatory oversight?

6 VALUE OF OPEN SOURCE SOFTWARE

As an interesting aside, the workshop discussed recent efforts to assess the challenge of quantifying the economic value of open source software, specifically the software used to support web servers. Some digital activity, maybe most digital creation, goes unmeasured. Harvard researchers have harvested public information from the Internet Archive and the data provided by the service software that identifies the software (Apache/IIS/nginx) to see which firms run which software and how that aligns with industry, region, and other characteristics [8]. They found that usage of server software was a reasonable proxy for behaviors that contribute to firm performance. Researchers have only scratched the surface of what can be done with the Internet Archive URLs, which are amenable to correlation with other financial data.

7 FUTURE DIRECTIONS

The researchers had useful feedback about CAIDA's data resources and resource discovery mechanisms that fell into two categories: 1) there is a steep learning curve when using new data for the first time, and in many cases one needs to have inside knowledge to figure out which software to use to process/query a given dataset; and 2) it is hard to cross dataset domains when doing simple lookups in bulk – for example, getting from IP addresses to ASNs is relatively easy, but one must consult a different database (file) to get the names of these ASNs. To address the first issue, we have built a new rich context data catalog, with a user interface that makes it easier to identify datasets, but also ties datasets to software libraries, tools, and sample code (notebooks) used to process the data sets, and to papers published with the data [9]. To address the second issue, we have designed and implemented backend microservice APIs that are designed both for simple interactive use as well as bulk queries. We explored (and implemented) both RESTful

and GraphQL APIs, and found that the GraphQL's complexity interfered with utility for many use cases, but it also provided more fine-grained control etc.

This workshop was part of an NSF-funded overhaul of CAIDA's data resources to lower the barrier to their usage by disciplines outside of computer science. The ultimate metric of success of this project is its ability to enable new empirical studies in the four targeted disciplines, promising innovations in: Internet mapping; detection of route hijacking and other disruptive events; cybersecurity preparedness; economic studies of correlations between ISP characteristics, market power, performance degradations, security practices, and regional economic growth; and regulatory discourse that has thus far occurred largely without data.

8 WORKSHOP PARTICIPANTS

Co-Hosts: kc claffy (CAIDA/UCSD) and Dave Clark (MIT/CSAIL). Participants: Klaus Ackermann (Monash University); Bob Cannon (FCC); Sarah George (U Penn); Shane Greenstein, Harry Oppenheimer, and Ran Zhuo (Harvard); Scott Jordan and Ali Nikkhah (UCI); Alan Kwan (U. of Hong Kong); Thomas Pellet (Northwestern); Ben Du, Roderick Fanou, Alexander Marder, Ricky Mok, Joshua Polterock, and Elena Yulaeva (CAIDA/UCSD).

ACKNOWLEDGMENTS

We thank participants for contributing their insights, and for feedback on this report. This workshop was supported by NSF OAC-1724853. Opinions expressed do not necessarily reflect views of the NSF. Any errors are the responsibility of the authors.

REFERENCES

[1] RIPE-NCC, "RIPE Atlas." https://atlas.ripe.net/.
[2] Ran Zhuo and Bradley Huffake and kc claffy and Shane Greenstein, "The Impact of the General Data Protection Regulation on Internet Interconnection," *Telecommunications Policy*, vol. 45, no. 2, 2021.
[3] "CAIDA AS Relationships Data." https://www.caida.org/data/as-relationships/.
[4] "CAIDA's IPv4 prefix probing data collection," 2016. http://www.caida.org/data/active/ipv4_prefix_probing_dataset.xml.
[5] CAIDA, "Inferred AS to Organization Mapping Dataset." https://www.caida.org/data/as-organizations/, 2017.
[6] CAIDA, "CAIDA Prefix to AS mappings Dataset (pfx2as) for IPv4 and IPv6 using RouteViews data." http://www.caida.org/data/routing/routeviews-prefix2as.xml.
[7] May Wong, "Stanford research snapshot of a new working-from-home economy." https://news.stanford.edu/2020/06/29/snapshot-new-working-home-economy/.
[8] Shane Greenstein and Klaus Ackermann, "The State of Open Source Server Software," tech. rep., Harvard University, 2018. https://www.hbs.edu/faculty/Pages/item.aspx?num=55073.
[9] "CAIDA Resource Catalog," 2020. https://catalog.caida.org/.