Find Out: How Do Your Data Packets Travel?

Thomas Dreibholz
Simula Metropolitan Centre for Digital Engineering
OsloMet – storbyuniversitetet
Pilestredet 52, 0167 Oslo, Norway
dreibh@simula.no

Somnath Mazumdar
Department of Digitalization
Copenhagen Business School
Solbjerg Plads 3, 2000 Frederiksberg, Denmark
sma.digi@cbs.dk

***Abstract*—In today's communication-centric world, users generate and exchange a massive amount of data. The Internet helps user data to travel from one part of the world to another, via a complex set of network systems. These systems are intelligent, heterogeneous, and non-transparent to users. This paper presents an extensive, trace-driven study of user data traffic covering five years of observations, six large ISPs, 22 different autonomous systems, and a total of 12 countries. This work aims to make users aware of how their data travels in the Internet, as the interests of ISPs majorly influence the data traffic path. Although data traffic should prefer to travel through countries that share land borders, we found that the shortest land distance between the two countries does not impact data path selection.**[1]

I. Introduction

Today's Internet communication ecosystem is becoming more complex and ubiquitous, while leaving online users less informed or not informed at all. It is now understood that the main reason why online users are tracked by third parties, including Internet service providers (ISPs), is primarily for their financial profits, for better product developments or better service offerings [1]. Unfortunately, some websites provide misleading information related to collected data processing to visitors and also track them [2]. In general, data related to users' web browsing activities can be further exploited to link with users' social media profiles [3]. It has been found that some ISPs collect too much data, including sensitive personal data, than required from across the product lines and they use web browsing as well as app usage data for targeted advertisements. Some of the ISPs also share real-time location data with third-parties for financial gains [1]. However, countries have implemented laws to protect online users. There are multiple categories of online user privacy. They are identity, access, location, and temporal level identity. User data privacy should be preserved, because personal data can directly identify the user. From the data packets, it is also possible to trace back to their actual source host (individual Ethernet) without even depending on intervening ISPs [4]. Informed Internet users now want to know more about how their data is travelling from source to destination. Also, if there is exploitation of user data by a third entity without properly informing the online users.

In this short paper, our research question is *Is the path taken by data packets deterministic?* and *What are the primary factors affecting user data packet transmissions?* We answered these two research questions via an experimental campaign conducted on data collected for five years (from 2018 to 2022) spanning over 12 countries using six large ISPs including 18 different autonomous systems (ASs). We visualised results using an intuitive map, and our data offers AS-level mapping information. Unlike others, our work is different from existing literature, due to three reasons [5]: *i)* widely used BGP routing data or RIPE RIS data were not meant to be used to infer AS-level mappings, because by definition, BGP was not designed with AS-level topology discovery feature [5, Subsubsection I.A.4], *ii)* Traceroute data lacks AS-level mapping information [5, Subsubsection I.A.5] [6, Subsection II.A], and *iii)* existing AS-related study results are hard to interpret [5, Subsubsection I.A.3].

II. The Impact of ISPs on Data Packet Flow

Path inflation is the routing policy choice among ISPs. Generally, ISPs do not share topology and routing policies, causing vagueness in the data packet transmission. ISPs use the BGP as an inter-domain routing protocol to exchange path information. It is typically a combination of ISP-paths and AS-paths. Traditionally, ISPs use a combination of local policy that includes commercial relationships, length of AS-paths and resource constraints to select a path. When a sender composes a message using a client application (such as an email client), after pressing the "Send" button, the message is decomposed into multiple data packets using complex networking rules. These data packets leave the sender's device (e.g. computer or smartphone) using a router and reach the nearest ISP. Based on the network service providers' internal infrastructural arrangements[2], data packets reach ASs which are nothing but a collection of routers. Each AS is assigned a number, which is used by routing protocols[3] and is owned by ISPs. Such a network segment constitutes the core network part of an ISP. Again, based on the ISP's commercial arrangements, it can use multiple layers of ASs, which can spread across countries and continents. Three canonical contractual relationships between ASs are customer-to-provider, peer-to-peer, and sibling-to-sibling relationships [7]. It is understood that multiple factors can influence the route a data packet will take. Such factors

[1]The authors would like to thank Maria Normann for her friendly support, as well as the anonymous reviewers for their helpful comments.

[2]Some network operators rely on other network operators by subscribing their resources.

[3]See https://www.iana.org/assignments/as-numbers/as-numbers.xhtml.

978-3-903176-51-5 © 2022 IFIP

Table I
THE TEST SITES USED FOR OUR EXPERIMENT (NB: ADSL: CONSUMER-GRADE ADSL, CF: CONSUMER-GRADE FIBRE AND RF: RESEARCH NETWORK FIBRE.)

Site (Country)	ISP 1	ISP 2
NTNU Trondheim (NO)	Uninett^{RF}	PowerTech^{ADSL}
Karlstads Universitet (SE)	SUNET^{RF}	–
Universität Duisburg-Essen (DE)	DFN^{RF}	–
Hainan University (CN)	CERNET^{RF}	China Unicom^{CF}

are primarily financial. Data packets leave the ISP premise and land in the transit network. Generally, transit networks typically consist of long-distance fibres between large cities, countries and continents. After the journey via the transit network, user data packets land in their respective ISP infrastructure. A reverse directional process follows, and the user message reaches the receiver.

III. RESULTS

A. Experimental Setup

We have used the NORNET research testbed infrastructure [8], consisting of distributed sites in Norway (NO), Sweden (SE), Germany (DE), and China (CN). We collected data packet traces (covering 12 countries, six ISPs and 22 ASs) from the year 2018 to mid of year 2022 between the sites shown in Table I.

HIPERCONTRACER [6] has been used to perform high-volume, long-term traceroute measurements over IPv4 and IPv6. Traceroute uses the Time-to-Live header field in IPv4 and Hop Count header field in IPv6 to trigger ICMP Time Exceeded errors from routers to extract IP addresses and associated IP packet routes. HIPERCONTRACER uses parallelisation and checksum adjustments to cope with load balancing. Traceroute has been performed approximately every five minutes with three rounds, over almost five years, for each relationship between the chosen sites (i.e. over 1,600,000 measurements per relation). While IP addresses correspond to locations in the network, they do not provide geo-location-related information. Simple geo-location using the free GEOLITE2 databases[4] does not provide very accurate information about routers. To approximate a router's geo-location, we used HLOC [9] for location approximation by using RIPE ATLAS Ping measurements to obtain the round-trip time (RTT) to a router's address from known vantage points. In addition, we use AS information from CIDR REPORT[5] and the AS number lookup from the free GEOLITE2 database. We selected three scenarios for further analysis focusing on AS-level and link-level. They are *i)* between neighbouring countries (Karlstad, SE to Trondheim, NO), *ii)* intra-continental (Essen, DE to Trondheim, NO), *iii)* inter-continental (Haikou, CN to Trondheim, NO).

[4]GEOLITE2: https://dev.maxmind.com/geoip/geoip2/geolite2/.
[5]CIDR REPORT: https://www.cidr-report.org/as2.0/autnums.html.

B. AS-level Findings

We can verify from the AS-level HIPERCONTRACER Traceroute results that: *i)* even simple scenarios can involve third-party countries, *ii)* traffic may take unexpected routes, and *iii)* routes change over the time between two fixed points.

1) Between Neighbouring Countries Scenario: The site in Karlstad has IPv4 connectivity only. It can be seen from Figure 1 that 100 % of the traffic must leave the site via only one router via SUNET (Swedish research network; AS 1653 in orange colour). The Trondheim site is connected to two ISPs: Uninett (Norwegian research network; AS 224 in red colour) and PowerTech (part of GlobalConnect, AS 2116 in green). Therefore, the incoming traffic is split between Uninett and GlobalConnect. While the research networks (Uninett in Norway and SUNET in Sweden) mainly connect within their corresponding country, possible connections between Karlstad and Trondheim may involve inter-country links from GlobalConnect, NorduNet (AS 2603; in blue colour) and IPO-EU (AS 12552, in purple colour). Between Karlstad and Trondheim, multiple routes are possible (e.g. via Copenhagen or Stockholm). Traffic can also take significant detours compared to the geographically shortest route. NorduNet connects Sweden and Norway via Denmark, i.e. traffic between the neighbouring countries Sweden and Norway may be routed via the third-party country Denmark (DK). In this scenario, all nations belong to the European Economic Area (EEA) and have similar privacy rules. It is also worth noting that data packet movement is influenced by many factors implemented by ISPs, including financial. We examine such facts in more detail in Subsubsection III-C1.

2) Intra-Continental Scenario: Furthermore, we take a look at intra-continental communication between two non-neighbouring countries. Figure 2 presents the data traffic from Essen, DE to Trondheim, NO. The site in Essen is connected to DFN only (German research network; AS 680 in light-green colour). Similar to the previous scenario, routes may involve third-party countries. A significantly larger number of third-party countries can be seen (e.g. BE, DK, FR, GB, NL, SE, and even the US)[6] in this case, although the geographic distance is around 1,400 km. That is, traffic may leave the EEA and run via GB or even trans-Atlantic via the US. Further analysis reveals that the route over the US is for DFN to PowerTech (consumer ISP) over IPv4 only, while IPv6 routing to PowerTech remains within EEA countries. However, from DFN to Uninett (both research networks), routing is via GB (non-EEA country) in many cases. The accuracy of geo-location becomes important here. HLOC [9] based its geo-location on RIPE ATLAS Ping measurements. The routes via the US are likely *not* false positives here. Instead, the trans-Atlantic connectivity of some ISPs sometimes makes routing a more cost-effective choice. However, from the perspective of privacy, this may not be an expected behaviour an ordinary Internet user would "naïvely" expect.

[6]Belgium (BE), France (FR), Great Britain (GB), Netherlands (NL), United States (US).

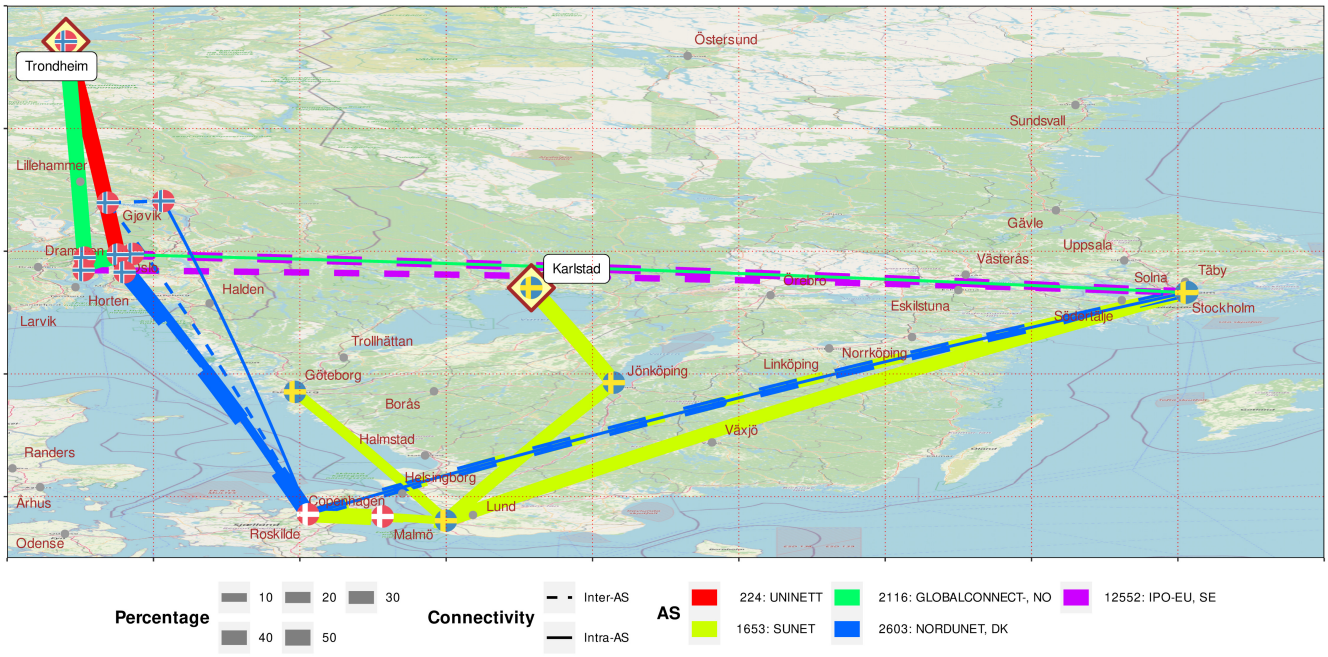

Figure 1. AS Map for IPv4 Traffic from Karlstad (SE) to Trondheim (NO).

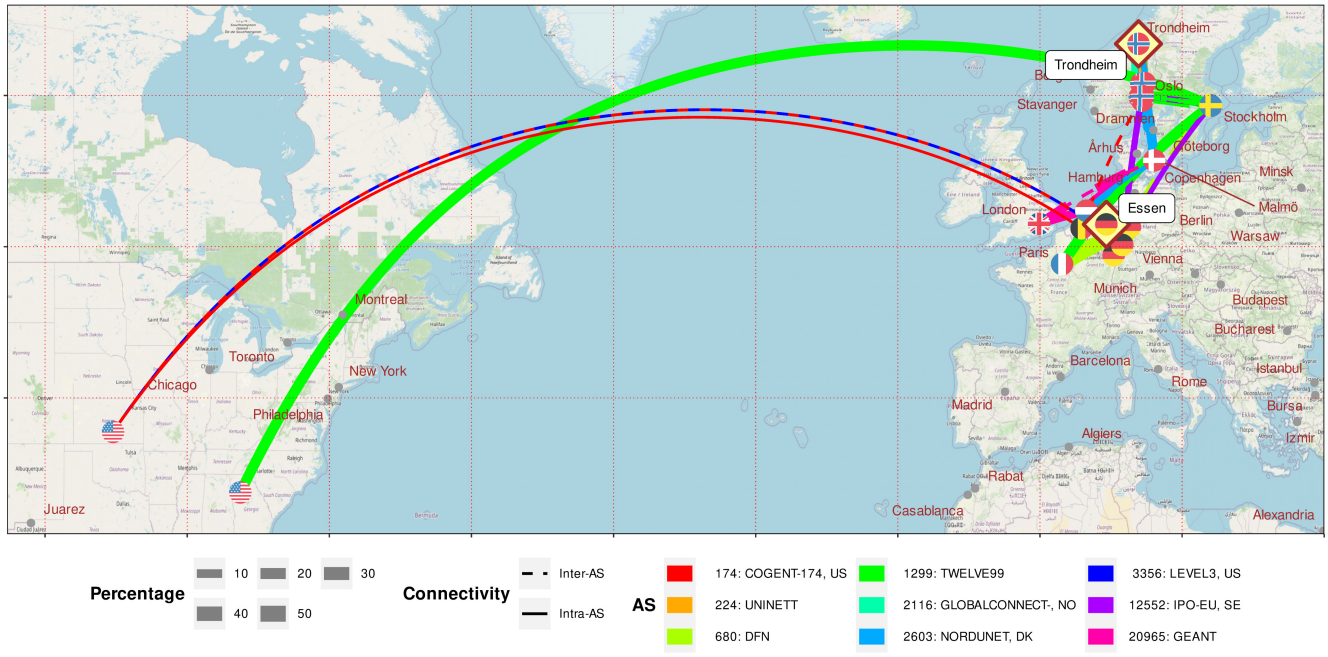

Figure 2. AS Map for IPv4 and IPv6 Traffic from Essen (DE) to Trondheim (NO).

3) Inter-Continental Scenario: Figure 3 shows an inter-continental setup: the traffic from Haikou, CN to Trondheim, NO. Similar to the previous two scenarios, data communication involves various regions and third-party countries (such as CA, DK, FR, GB, JP, NL, SE, US)[7], while the geographically shortest distance is just around 8,500 km (Asia to Europe via Russia). We can see that a significant fraction of the observed routes takes the eastern direction (from China) via different trans-Pacific and -Atlantic cables. During nearly five years of observations, various paths between the two sites have been seen, involving 19 different ASs and 8 third-party countries, including significant detours. The observed routes result from economic decisions made by the involved ISPs. It has an apparent effect on the RTT (see [6] for details), but it also significantly impacts the assumptions to be made about privacy for inter-continental data traffic.

[7]Canada (CA), Japan (JP).

C. Link-level Findings

In the following, we take a look at cross-border data traffic. We, therefore, computed the percentages of cross-border link observations for the three scenarios mentioned in Subsection III-B. We showed the observed percentage

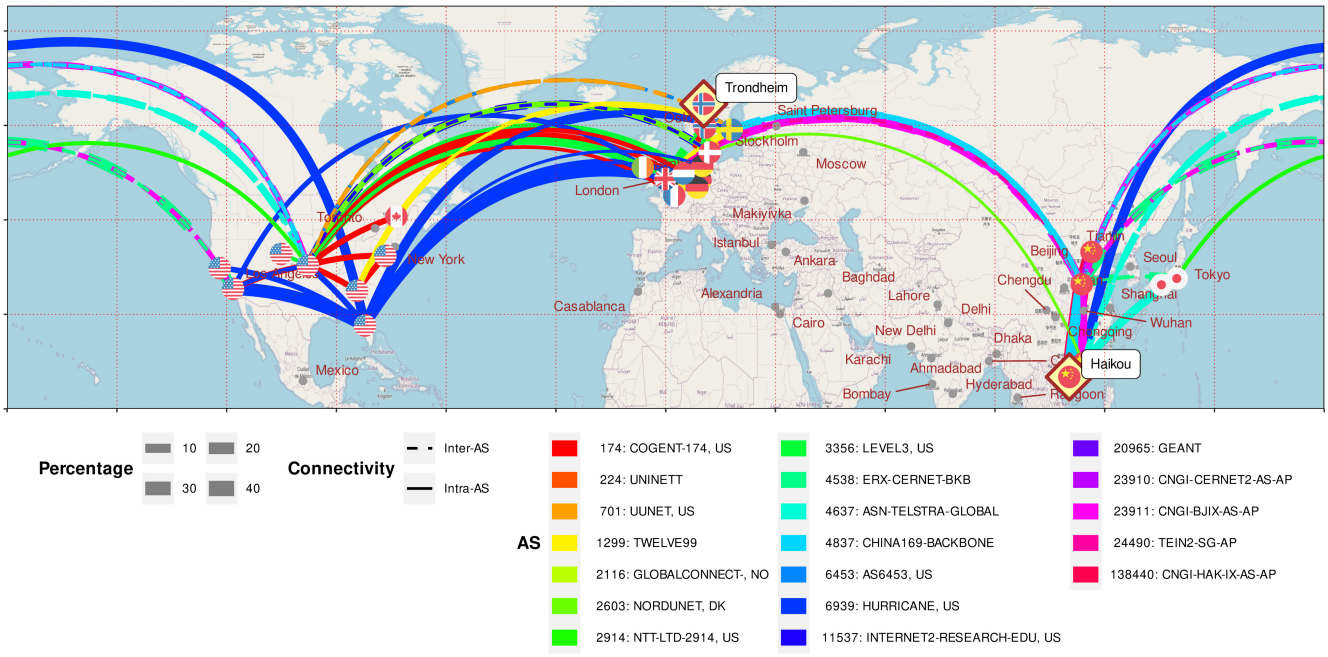

Figure 3. AS Map for IPv4 and IPv6 Traffic from Haikou (CN) to Trondheim (NO).

Table II
DATA TRAFFIC PERCENTAGES FROM KARLSTAD, SE TO TRONDHEIM, NO

From	To	Percent IPv4
DK	NO	67.65
DK	SE	14.67
SE	DK	99.69
SE	NO	14.63

for each cross-border relationship in all HIPERCONTRACER Traceroute results.

1) Between Neighbouring Countries Scenario: Table II shows the cross-border data (only IPv4) traffic percentages for the Karlstad, SE to Trondheim, NO scenario (mentioned in Subsubsection III-B1). Overall, we can see significant traffic detours through a third-party country (here: Denmark). In 99.69 % of the observations, a cross-border link SE-DK is observed. That is, concerning Figure 1, in many cases, traffic between ISPs is exchanged in Denmark and from there forwarded to Norway. 67.65 % of the observations contain a cross-border link DK-NO. So, traffic from DK may re-enter SE (likely in another AS), with 14.67 % of the observations showing a cross-border link DK-NO. Note that cross-border links may be hidden due to non-responding routers. Also, for security reasons, the ICMP response rate of routers is limited, i.e. not all routers respond to all Traceroute runs.

2) Intra-Continental Scenario: Table III presents the observation of both IPv4 and IPv6 data traffic percentages for the Essen, DE to Trondheim, NO scenario (refer to Subsubsection III-B2). There are many routes between DE and NO, but routes change over time, leading to smaller percentages (e.g. for IPv4, 45.25% of the observations have a cross-border link DE-FR, 50.15% a link DE-NL, 3.62% a direct link DE-NO) of traffic flow on different cross-border links. Due to the smaller deployment of IPv6 compared to IPv4, there is less variation

Table III
DATA TRAFFIC PERCENTAGES FROM ESSEN, DE TO TRONDHEIM, NO

From	To	Percent IPv4	Percent IPv6
BE	NO	0.07	–
DE	FR	45.25	–
DE	NL	50.15	50.63
DE	NO	3.62	–
DE	US	0.06	–
DK	NO	50.11	34.92
FR	SE	44.89	–
GB	DK	9.64	9.89
GB	NL	0.14	–
NL	DK	40.47	40.69
NL	GB	9.76	9.86
NO	SE	3.62	–
SE	NO	50.09	43.12
SE	US	16.91	–
US	BE	0.06	–
US	DE	0.06	–
US	SE	16.91	–

for IPv6. For example, the only cross-border link seen from DE via IPv6 is DE-NL with 50.63 %. Again, note that there may be hidden links. There is no long trans-Atlantic detour via the US for IPv6, but there is routing via GB for 9.86 % of the observed IPv6 routes. So, despite the routing differences between IPv4 and IPv6, the privacy issues remain the same for both protocols.

3) Inter-Continental Scenario: Finally, Table IV shows the observed IPv4 and IPv6 data traffic percentages for the Haikou, CN to Trondheim, NO scenario (refer to Subsubsection III-B3). The table reflects the results expected from Figure 3, with many different link possibilities. While a significant fraction of the traffic takes a direct route from China

Table IV
DATA TRAFFIC PERCENTAGES FROM HAIKOU, CN TO TRONDHEIM, NO

From	To	Percent IPv4	Percent IPv6
BE	NO	2.56	–
CA	US	2.48	–
CN	DK	0.05	–
CN	GB	18.74	8.09
CN	JP	9.61	–
CN	SE	42.47	–
CN	US	27.59	29.33
DK	NO	49.35	35.74
FR	NL	0.22	–
GB	DK	19.44	10.69
GB	NL	1.37	–
GB	NO	21.34	–
JP	CN	9.05	–
JP	US	0.51	–
NL	DK	1.37	–
NL	US	2.52	0.48
SE	DK	22.82	17.12
SE	NO	25.35	43.25
SE	US	3.34	–
US	BE	2.52	–
US	CA	2.48	–
US	DK	3.99	9.45
US	FR	0.22	–
US	GB	21.26	11.73
US	NL	2.30	11.59
US	SE	3.52	28.17

to Europe over IPv4 (via SE: 42.47 %, via GB: 18.74 %), the other part takes the opposite direction, via trans-Pacific and trans-Atlantic cables (e.g. 27.59 % directly CN-US, or 9.61 % via CN-JP). For IPv6, there is less variation. That is, the connections between China and Norway contain many border crossings, also between different political regions. It can motivate further research on the implications, particularly concerning privacy.

IV. RELATED WORK

Existing articles focus primarily on the routing policies [7], [10]. Gao et al. measured AS path lengths and found that AS paths are inflated due to inter-domain routing policies, where a shortest AS path routing policy is not used in most cases [10]. It is worth noting that most of the literature mainly focuses on the path of inflation. Golkar et al. compared IPv4 and IPv6 routing, showing that IPv6 routes may change more frequently, while the number of hops is similar to IPv4 routing [11]. Mahajan et al. present a negotiation framework to share information for data traffic flow based on a specific relationship [7].

V. CONCLUSIONS AND OUTLOOK

This work aims to extend the state-of-the-art by not relying on public BGP routing data and standard Traceroute data, while making the results easier to understand via cartography-level visualisation. For this study, we had to solve the challenge of setting up an infrastructure to obtain long-term measurements (over 5 years) at a high frequency, at multiple sites (requiring negotiations with various entities), in different countries (involving different regulations), with different types of ISPs (i.e. particularly not only research networks, but also consumer-grade ADSL lines and business-grade commercial fibres). Our study found that the travel paths of data packets are not deterministic (shown via all three scenarios). We also found that the primary factor that affects data traffic path selection is the economic incentives received by the involved ISPs. Due to that, a significant detour from the shortest path distance can be seen in our study, where the shortest land distance does not count (valid for neighbouring countries and intra-continental scenarios). Moreover, data packets can be routed via the countries that do not share a land border. To conclude, this work shows that data traffic path selection is neither transparent nor deterministic. Such situations can impact user data privacy. It is recommended that the end-users and application developers are aware of these findings and get the option to decide which path the data should take and where it should not travel. As part of our ongoing work, we will analyse the HIPERCONTRACER Traceroute data in more detail and combine it with fine-granular HIPERCONTRACER Ping data to obtain information about connectivity outages and RTT changes.

REFERENCES

[1] F. Staff, "A Look At What ISPs Know About You: Examining the Privacy Practices of Six Major Internet Service Providers," Federal Trade Commission, Tech. Rep., 2021.

[2] I. Sánchez-Rola, M. Dell'Amico, P. Kotzias, D. Balzarotti, L. Bilge, P.-A. Vervier, and I. Santos, "Can I Opt Out Yet? GDPR and the Global Illusion of Cookie Control," in *Proceedings of the ACM Asia Conference on Computer and Communications Security*, 2019, pp. 340–351.

[3] J. Su, A. Shukla, S. Goel, and A. Narayanan, "De-Anonymizing Web Browsing Data with Social Networks," in *Proceedings of the 26th International Conference on World Wide Web*, 2017, pp. 1261–1269.

[4] H. Burch and B. Cheswick, "Tracing Anonymous Packets to their Approximate Source," in *Proceedings of the 14th Systems Administration Conference*, New Orleans, Louisiana/U.S.A., Dec. 2000, pp. 319–327.

[5] M. Roughan, W. Willinger, O. Maennel, D. Perouli, and R. Bush, "10 Lessons from 10 Years of Measuring and Modeling the Internet's Autonomous Systems," *IEEE Journal on Selected Areas in Communications*, pp. 1810–1821, 2011.

[6] T. Dreibholz, "HiPerConTracer - A Versatile Tool for IP Connectivity Tracing in Multi-Path Setups," in *Proceedings of the 28th IEEE International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*, Hvar, Dalmacija/Croatia, 2020.

[7] R. Mahajan, D. Wetherall, and T. Anderson, "Negotiation-based Routing between Neighboring ISPs," in *Proceedings of the 2nd conference on Symposium on Networked Systems Design & Implementation*, 2005, pp. 29–42.

[8] T. Dreibholz, "NorNet at the University of Sydney: From Simulations to Real-World Internet Measurements for Multi-Path Transport Research," Invited Talk at University of Sydney, Sydney, New South Wales/Australia, 2019.

[9] Q. Scheitle, O. Gasser, P. Sattler, and G. Carle, "HLOC: Hints-Based Geolocation Leveraging Multiple Measurement Frameworks," in *Proceedings the of Network Traffic Measurement and Analysis Conference (TMA)*, Dublin/Ireland, 2017.

[10] Q. Gao, F. Wang, and L. Gao, "Quantifying as Path Inflation by Routing Policies," *International Journal of Future Generation Communication and Networking*, pp. 167–186, 2016.

[11] F. Golkar, T. Dreibholz, and A. Kvalbein, "Measuring and Comparing Internet Path Stability in IPv4 and IPv6," in *Proceedings of the 5th IEEE International Conference on the Network of the Future (NoF)*, Paris/France, 2014.