



Bias in Internet Measurement Platforms

**Pavlos Sermpezis, Lars Prehn, Sofia Kostoglou,
Marcel Flores, Athena Vakali, Emile Aben**

Data & Web Science Laboratory (Datalab), <https://datalab.csd.auth.gr/>
Computer Science Dept., Aristotle University of Thessaloniki



RIPE NCC
RIPE NETWORK COORDINATION CENTRE

Edgio



Bias in data: a motivational example

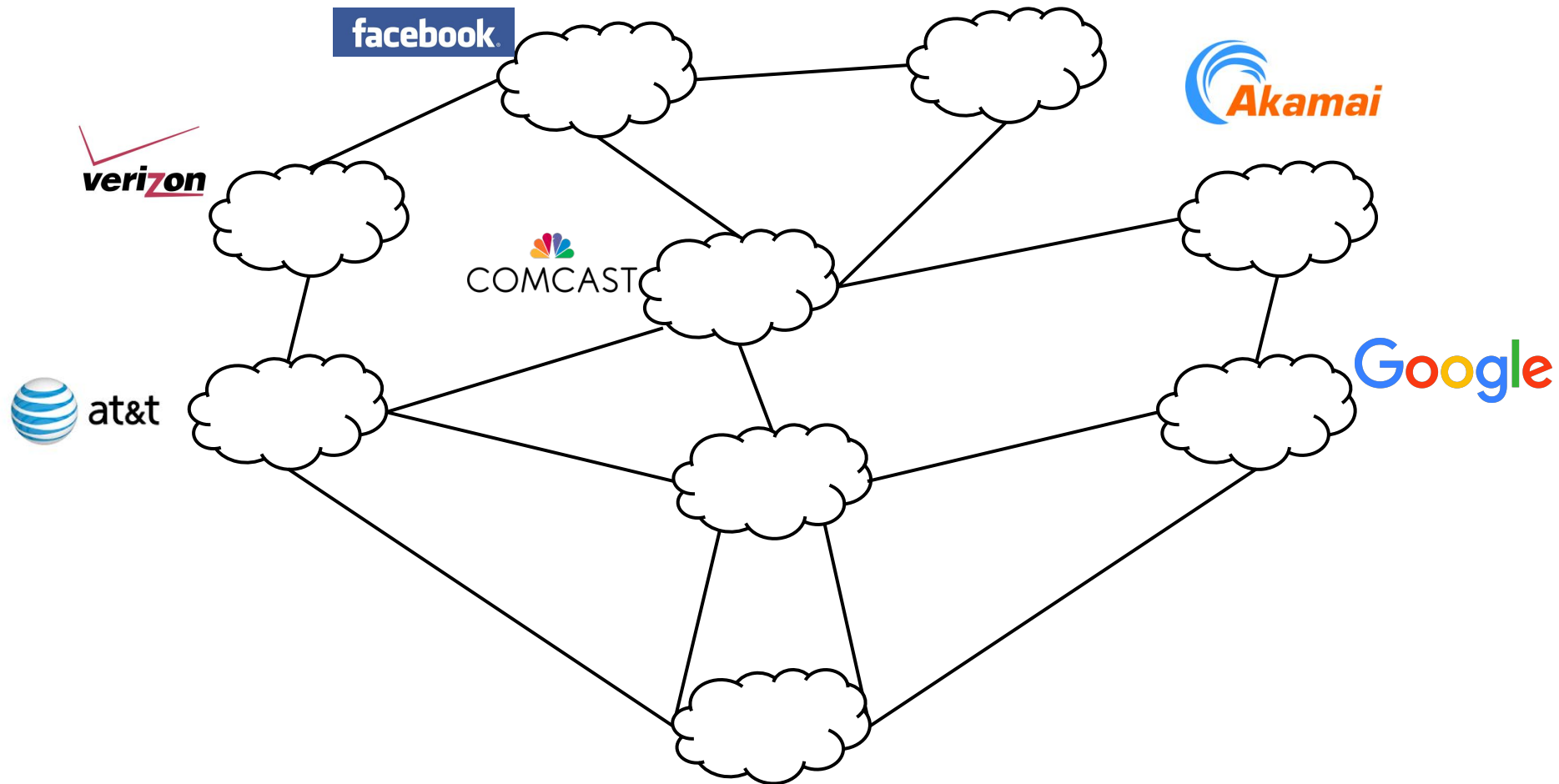
- Assume an entire population of 100 people
 - 50 men, 50 women
 - 70 from country A, 30 from country B
- We do a survey with 10 participants
 - 8 men, 2 women
 - 8 from country A, 2 from country B

	Men	Women	Country A	Country B
Entire population	50%	50%	70%	30%
Survey sample	80%	20%	80%	20%

- Is there bias? → Yes! difference in the gender/country distributions between population & sample
- Is bias the same along gender/country? → No! sample is more biased wrt. the gender dimension
- Is bias a problem? → It depends!
 - Goal: estimate the average population height (gender bias **is** a problem, country bias **may be** a problem)
 - Goal: calculate % of native spoken languages (gender bias **is not** a problem, country bias **is** a problem)

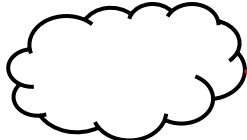


The Internet





The Internet



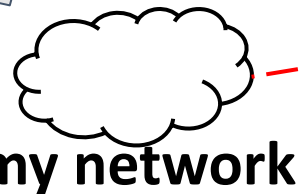
my network





The Internet

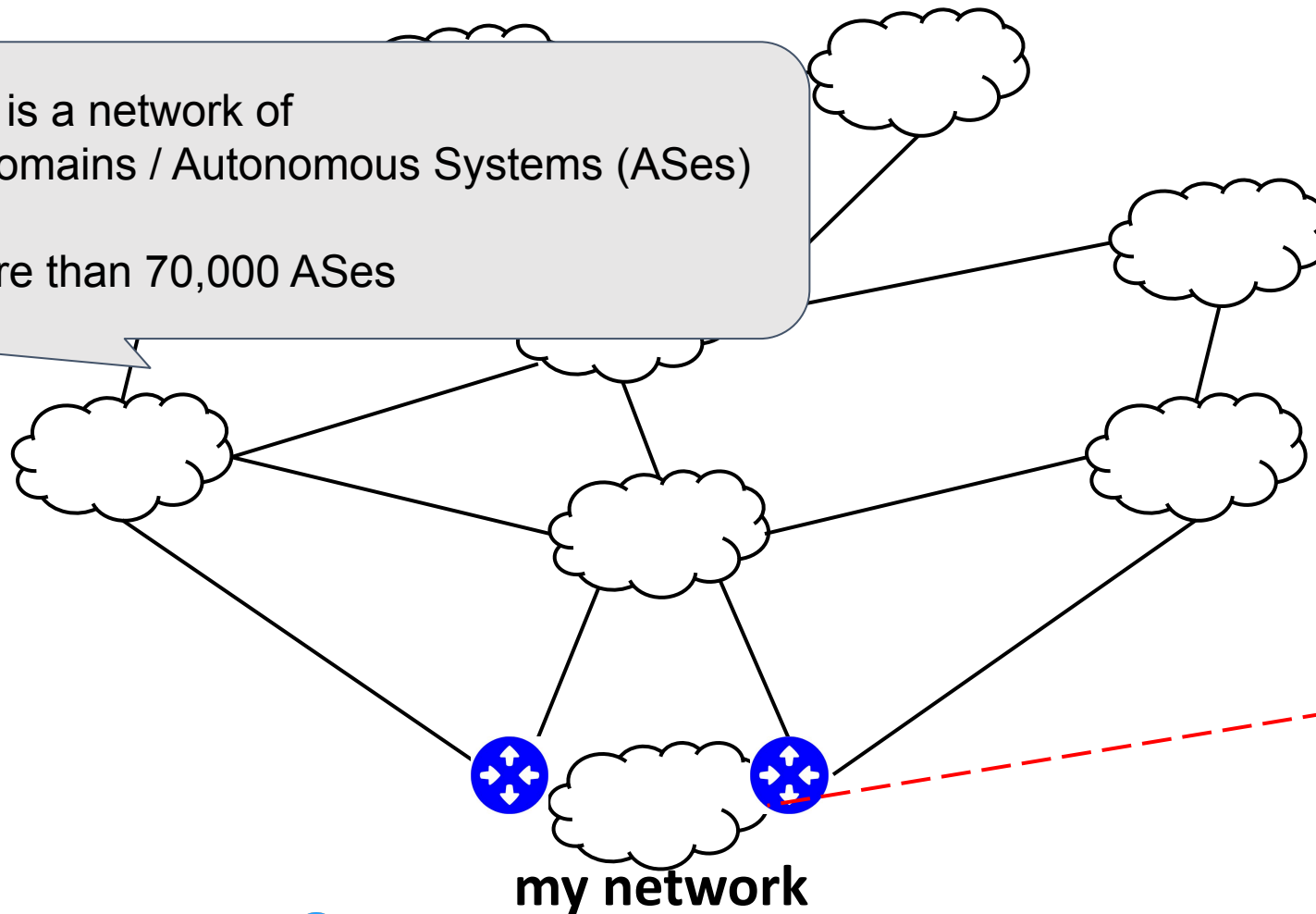
A “**domain**” or “**Autonomous System (AS)**” is a network or a collection of networks that are all managed, controlled and supervised by a single entity or organization





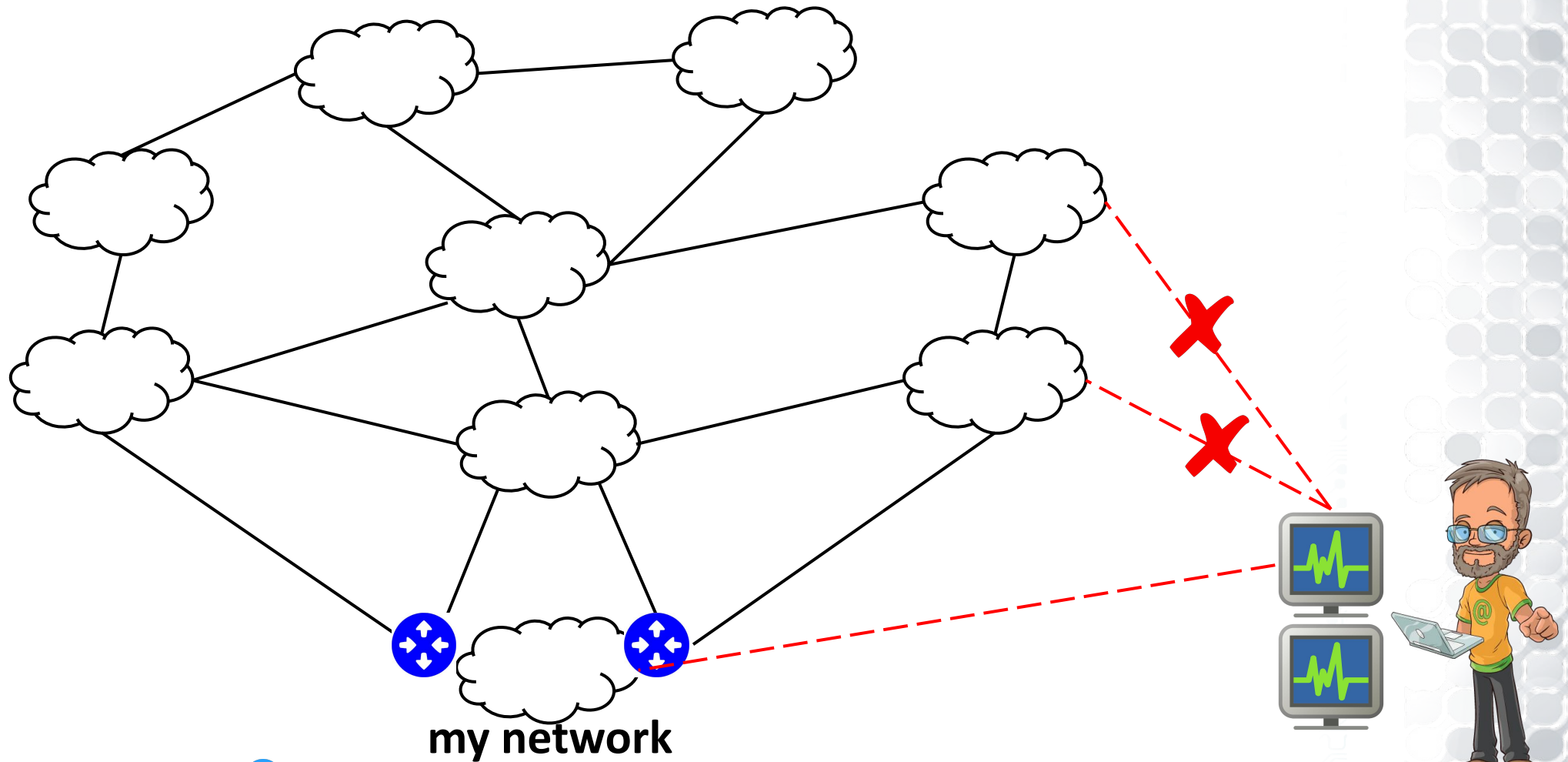
The Internet

- The Internet is a network of networks / domains / Autonomous Systems (ASes)
- today → more than 70,000 ASes



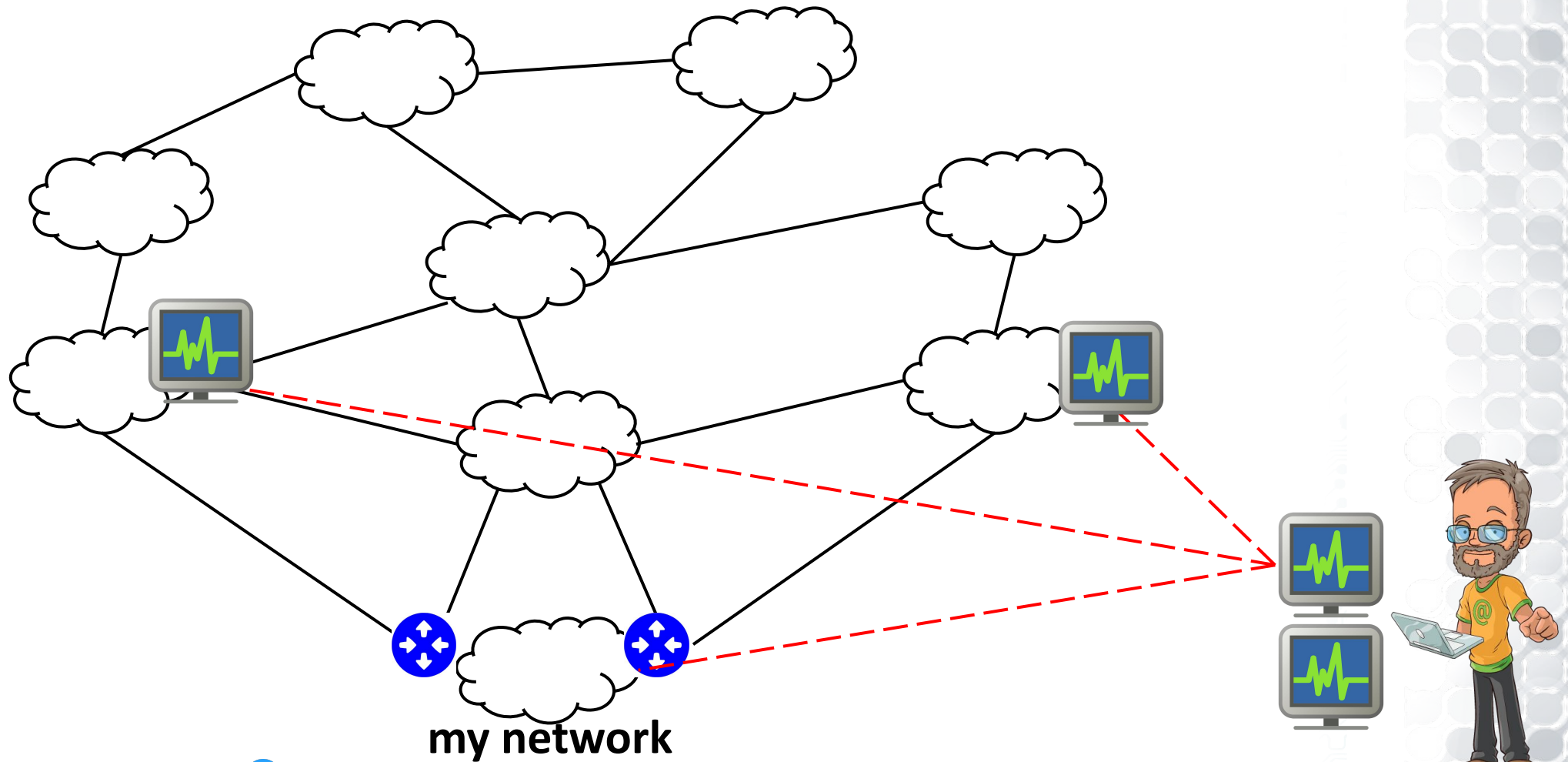


Inter-domain monitoring





Internet measurement platforms





Internet measurement platforms



RIPE NCC

RIPE Atlas



<https://atlas.ripe.net/>

- data plane measurements
- > 11,000 probes & anchors
- in > 3000 ASNs



<http://www.routeviews.org>

- BGP RIBs & updates
- 36 route collectors
- peering with > 300 ASNs



<https://ris-live.ripe.net/>

- BGP RIBs & updates
- 27 route collectors
- peering with > 500 ASNs

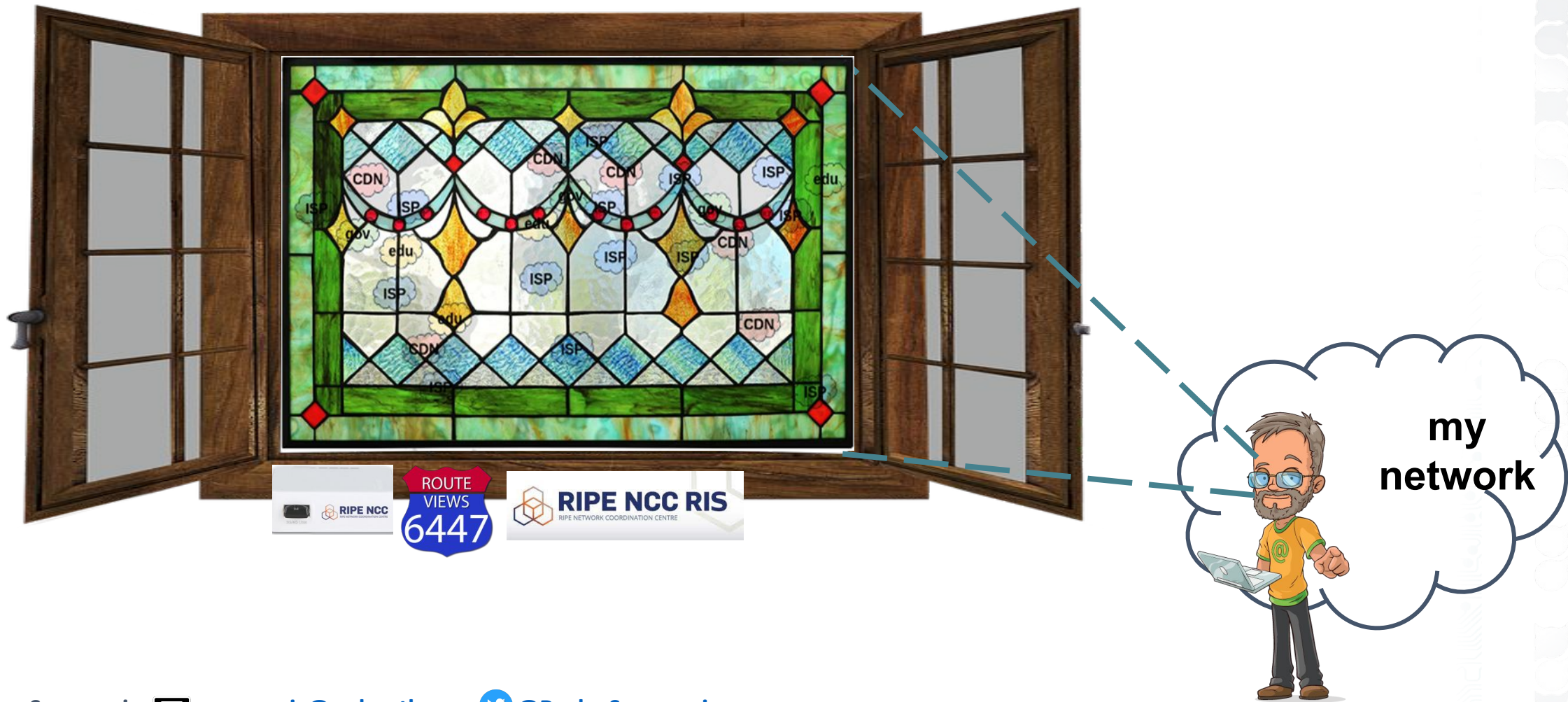


Measurement platforms: a window to the Internet





... but, in practice: a *stained glass* window





The “stained glass” view == Bias

not all network types can be equally
seen by the platforms

→ our view of the Internet is **biased**

Example 1 (location bias)

- RIPE Atlas & RIPE RIS have more probes/peers in Europe



RIPE Atlas probes

<https://atlas.ripe.net/results/maps/network-coverage/>

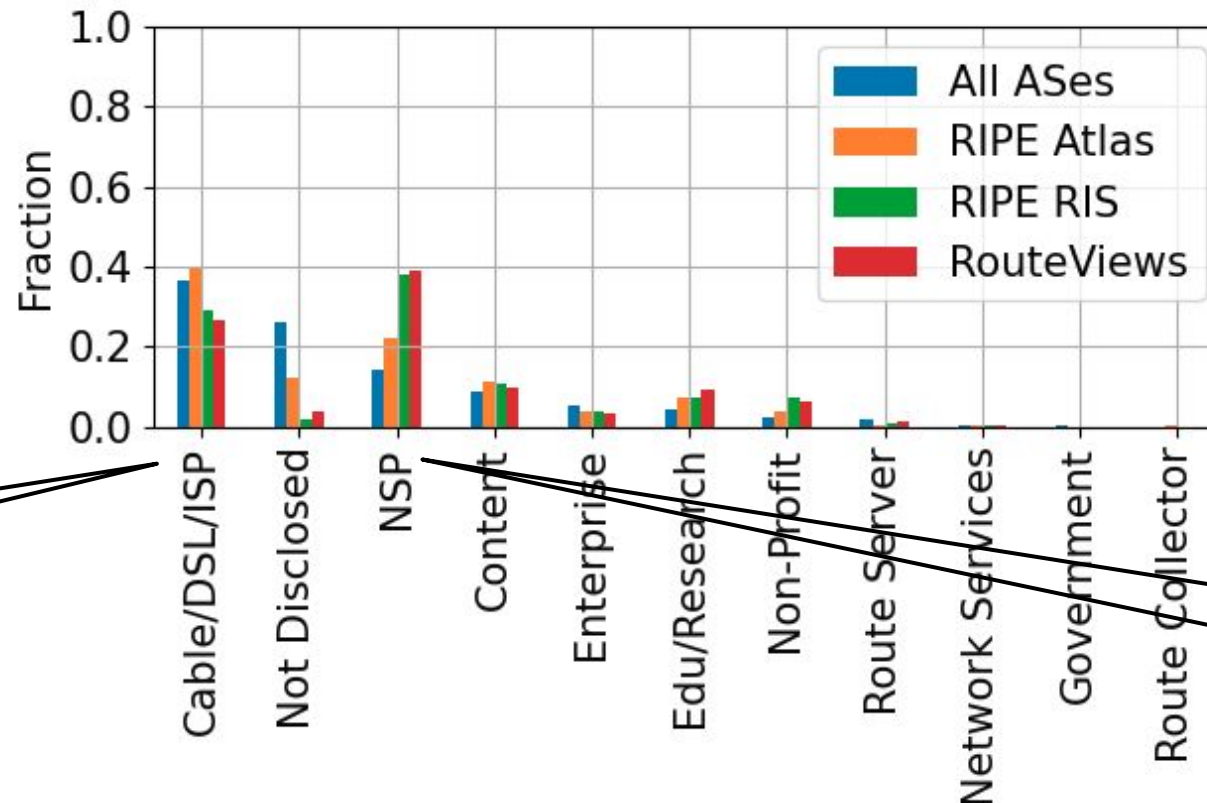


RIPE RIS route collectors

<https://observablehq.com/@emileaben/ris-route-collectors-and-peer-locations>

Example 2 (network-type bias)

- Peers of **RIPE RIS** and **RouteViews** do not equally represent all network types



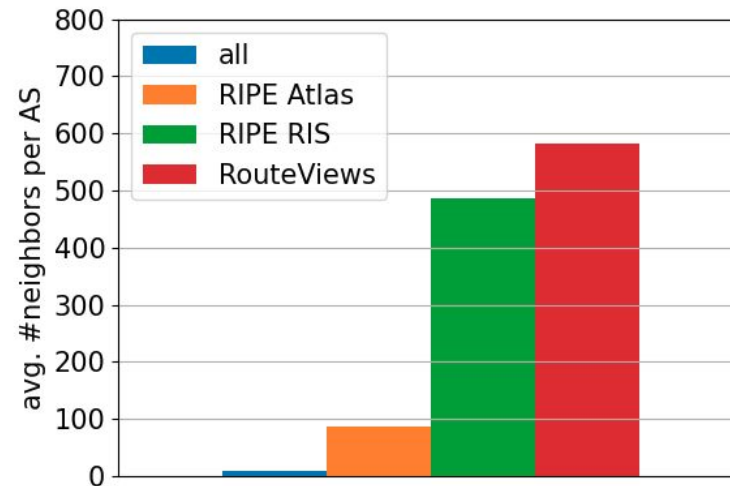
Cable/DSL/ISP are *under-represented*

NSPs are *over-represented*



Example 3 (topological bias)

- ASes that feed to **RIPE RIS/RouteViews** or host **RIPE Atlas** probes, are networks that typically peer with many other networks





Quantifying bias

- Many dimensions of bias
 - *location, network size, topology, IXP connectivity, network type, etc.*





Internet data sources

- [CAIDA AS-rank](#)
 - Network information: location, network size, topology, etc.
- [CAIDA AS-relationships](#)
 - Graph information: edgelist (i.e., peering links)
- [Peering DB](#)
 - Network information: connectivity, network type, traffic, etc.
- [AS hegemony](#)
 - Network information: size, topology
- [Country-level Transit Influence \(CTI\)](#)
 - Network information: size, topology
- [ASDB](#)
 - Network information: network types

**22 features/characteristics
per network
(i.e., the bias “dimensions”)**

The compiled dataset

← 22 network features/characteristics →

↑
74k networks
↓

ASN	Location-related information		Network-size related information			Topology-related information		IXP-related information		Network type-related information		
	RIR Region	Continent	Customer cone (in #ASNs)	AS hegemony	...	#neighbors (in #ASNs)	...	#IXPs connected to	...	Net. type (PeeringDB)	Net. type (ASDB)	...
174	ARIN	North America	32457	0.09	...	6614	...	0	...	NSP	ICT	...
1299	RIPE	Europe	37162	0.10		2328		0		NSP	ICT	
2497	APNIC	Asia	507	0.01		338		16		NSP	NaN	
3320	RIPE	Europe	3015	0.01		667		5		NSP	ICT	
3333	RIPE	Europe	3	0.00		320		1		Non-profit	ICT	
5470	RIPE	Europe	1	0.00		1		NaN		NaN	Education & Research	
15169	ARIN	North America	12	0.01		366		214		Content	ICT	
...

© 2018-2023, Pavlos Sermpezis. All rights reserved.



Quantifying bias

- Many dimensions of bias
 - *location, network size, topology, IXP connectivity, network type, etc.*
- Bias score per dimension
 - Bias == Difference between two distributions (**all networks** vs. **networks with vantage points**)
 - **Bias score**: Kullback-Leibler divergence metric
 - i.e, a value between 0 (low bias) and 1 (high bias)

ASN	Location-related information		Network-size related information		
	RIR Region	Continent	Customer cone (in #ASNs)	AS hegemony	...
174	ARIN	North America	32457	0.09	...
1299	RIPE	Europe	37162	0.10	
2497	APNIC	Asia	507	0.01	
3320	RIPE	Europe	3015	0.01	
3333	RIPE	Europe	3	0.00	
5470	RIPE	Europe	1	0.00	
15169	ARIN	North America	12	0.01	
...	



Quantifying bias

- Many dimensions of bias
 - *location, network size, topology, IXP connectivity, network type, etc.*
- Bias score per dimension
 - Bias == Difference between two distributions (**all networks** vs. **networks with vantage points**)
 - **Bias score**: Kullback-Leibler divergence metric
 - i.e, a value between 0 (low bias) and 1 (high bias)

ASN	Location-related information		Network-size related information		
	RIR Region	Continent	Customer cone (in #ASNs)	AS hegemony	...
174	ARIN	North America	32457	0.09	...
1299	RIPE	Europe	37162	0.10	
2497	APNIC	Asia	507	0.01	
3320	RIPE	Europe	3015	0.01	
3333	RIPE	Europe	3	0.00	
5470	RIPE	Europe	1	0.00	
15169	ARIN	North America	12	0.01	
...	



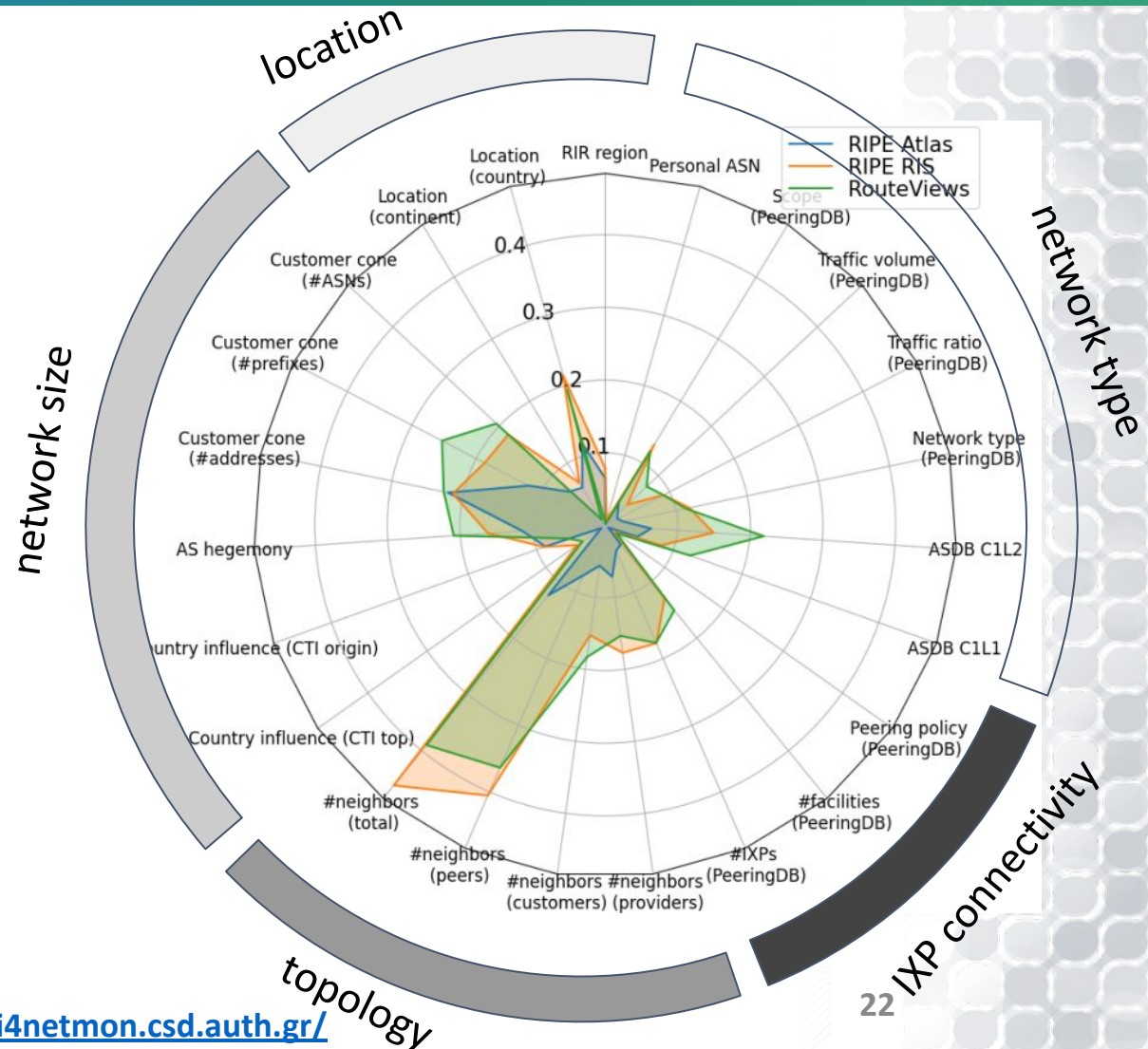
Quantifying bias

- Many dimensions of bias
 - *location, network size, topology, IXP connectivity, network type, etc.*
- Bias score per dimension
 - Bias == Difference between two distributions (**all networks** vs. **networks with vantage points**)
 - **Bias score**: Kullback-Leibler divergence metric
 - i.e, a value between 0 (low bias) and 1 (high bias)

ASN	Location-related information		Network-size related information		
	RIR Region	Continent	Customer cone (in #ASNs)	AS hegemony	...
174	ARIN	North America	32457	0.09	...
1299	RIPE	Europe	37162	0.10	
2497	APNIC	Asia	507	0.01	
3320	RIPE	Europe	3015	0.01	
3333	RIPE	Europe	3	0.00	
5470	RIPE	Europe	1	0.00	
15169	ARIN	North America	12	0.01	
...	

Quantifying bias

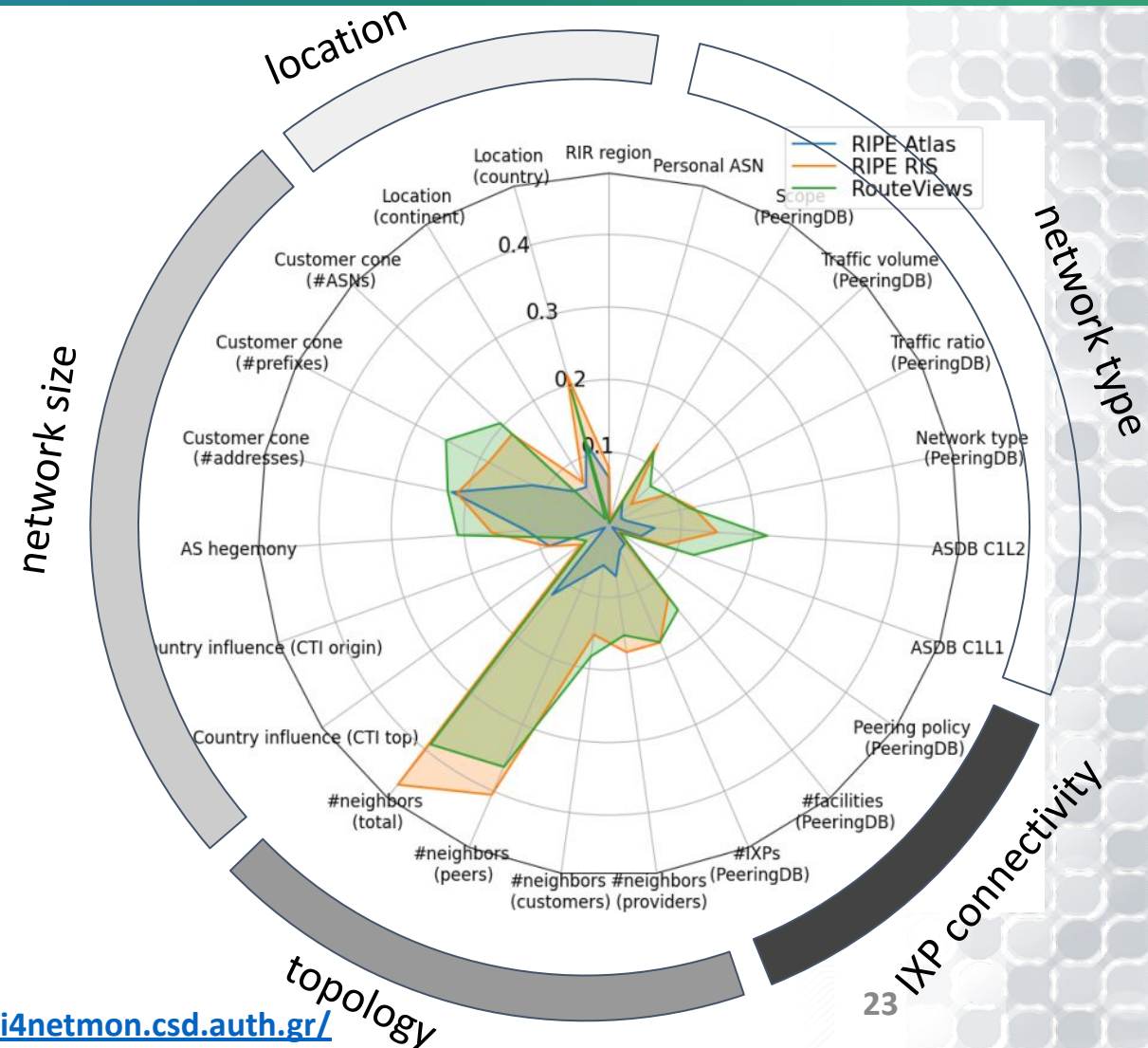
- Many dimensions of bias
 - *location, network size, topology, IXP connectivity, network type, etc.*
- Bias score per dimension
 - Bias == Difference between two distributions (**all networks** vs. **networks with vantage points**)
 - **Bias score**: Kullback-Leibler divergence metric
 - i.e, a value between 0 (low bias) and 1 (high bias)
- Radar plot of bias
 - each radius → a bias dimension
 - colored lines/areas → bias score
 - high bias → far from center





Bias in Internet Measurement Platforms

RIPE Atlas is significantly less biased than **RIPE RIS** and **RouteViews** in almost all dimensions

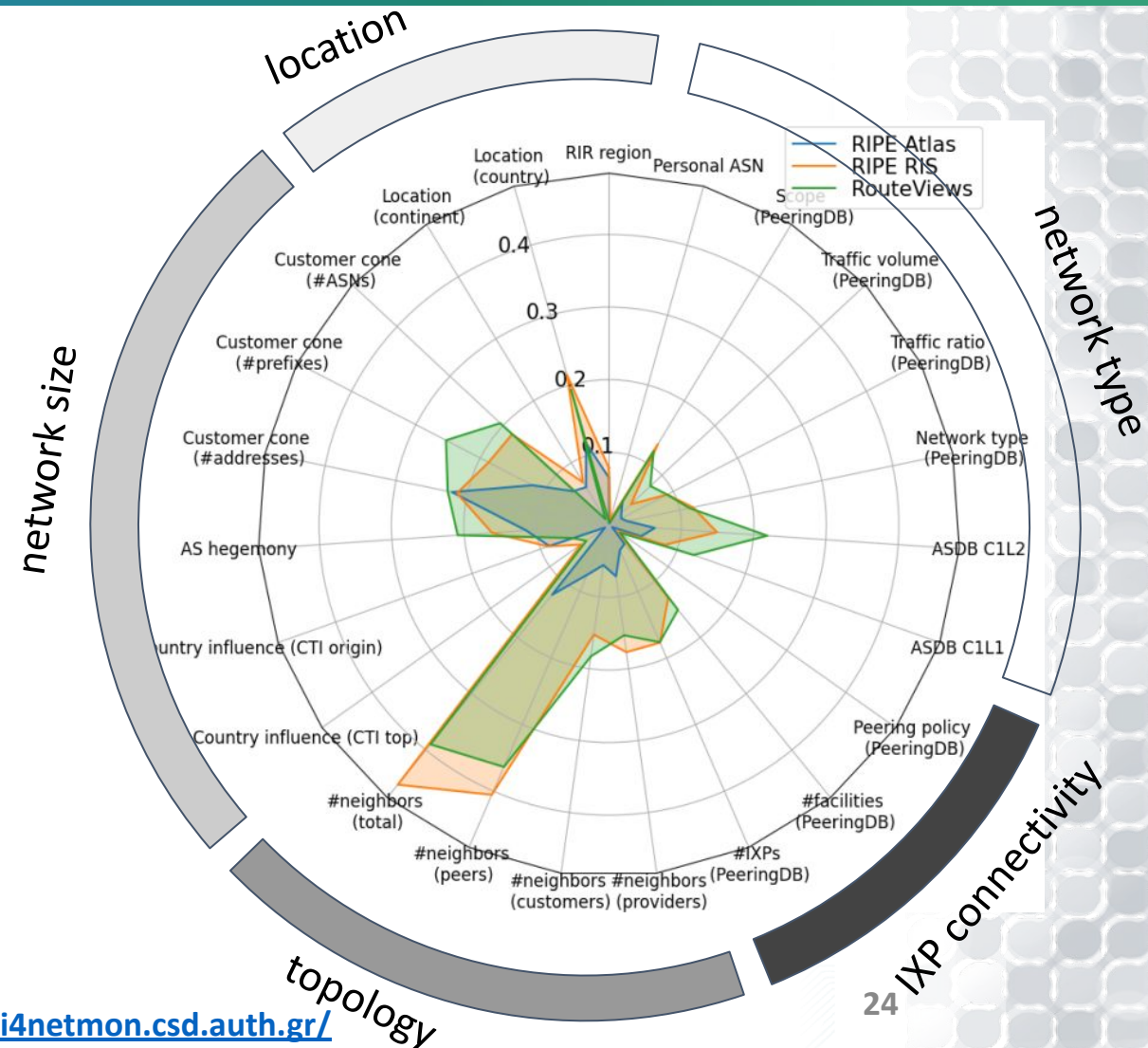




Bias in Internet Measurement Platforms

RIPE Atlas is significantly less biased than **RIPE RIS** and **RouteViews** in almost all dimensions

RIPE RIS has high topology bias (due to route collectors at IXPs) and high network size bias (peers are large networks)



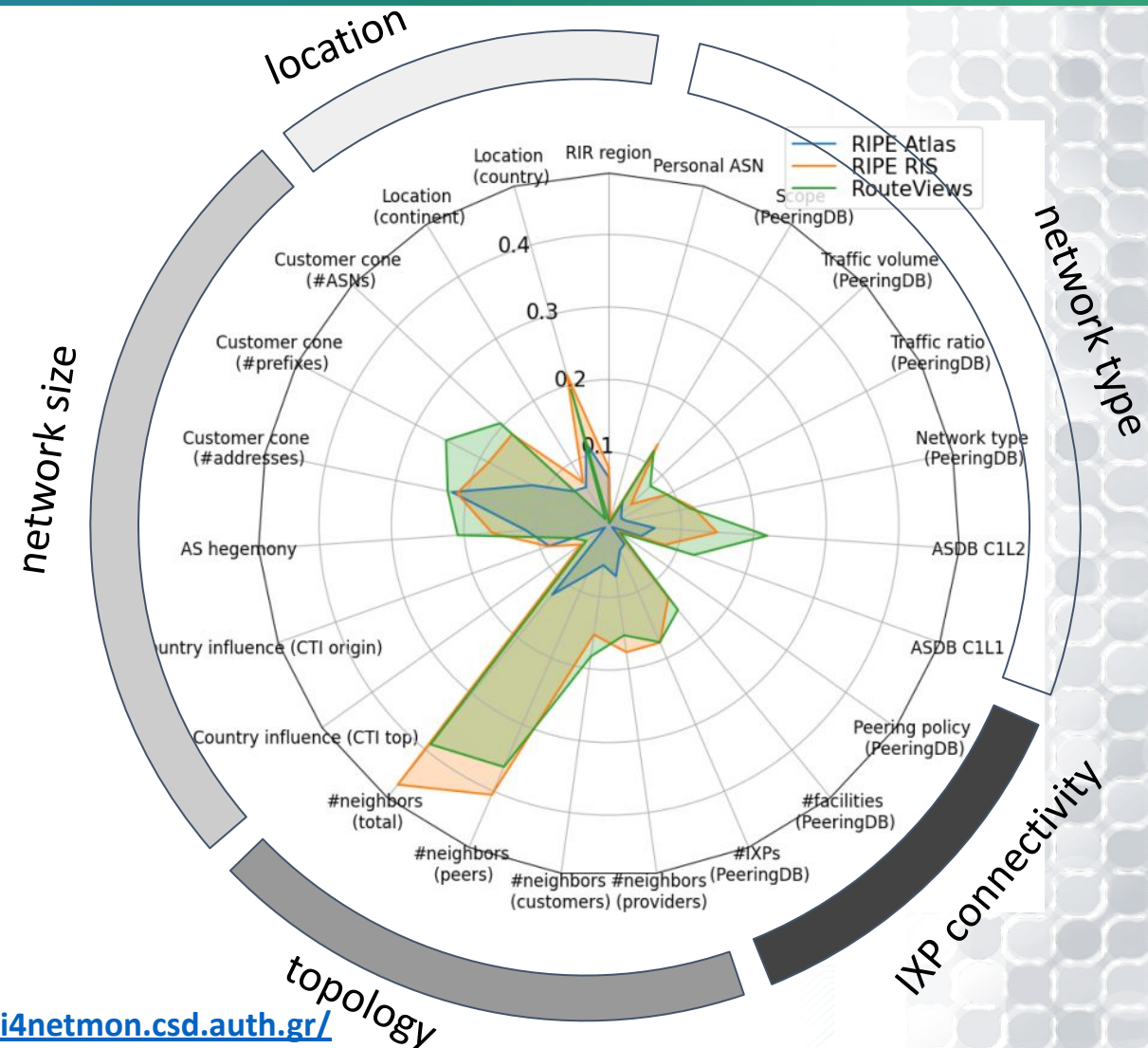


Bias in Internet Measurement Platforms

RIPE Atlas is significantly less biased than **RIPE RIS** and **RouteViews** in almost all dimensions

RIPE RIS has high topology bias (due to route collectors at IXPs) and high network size bias (peers are large networks)

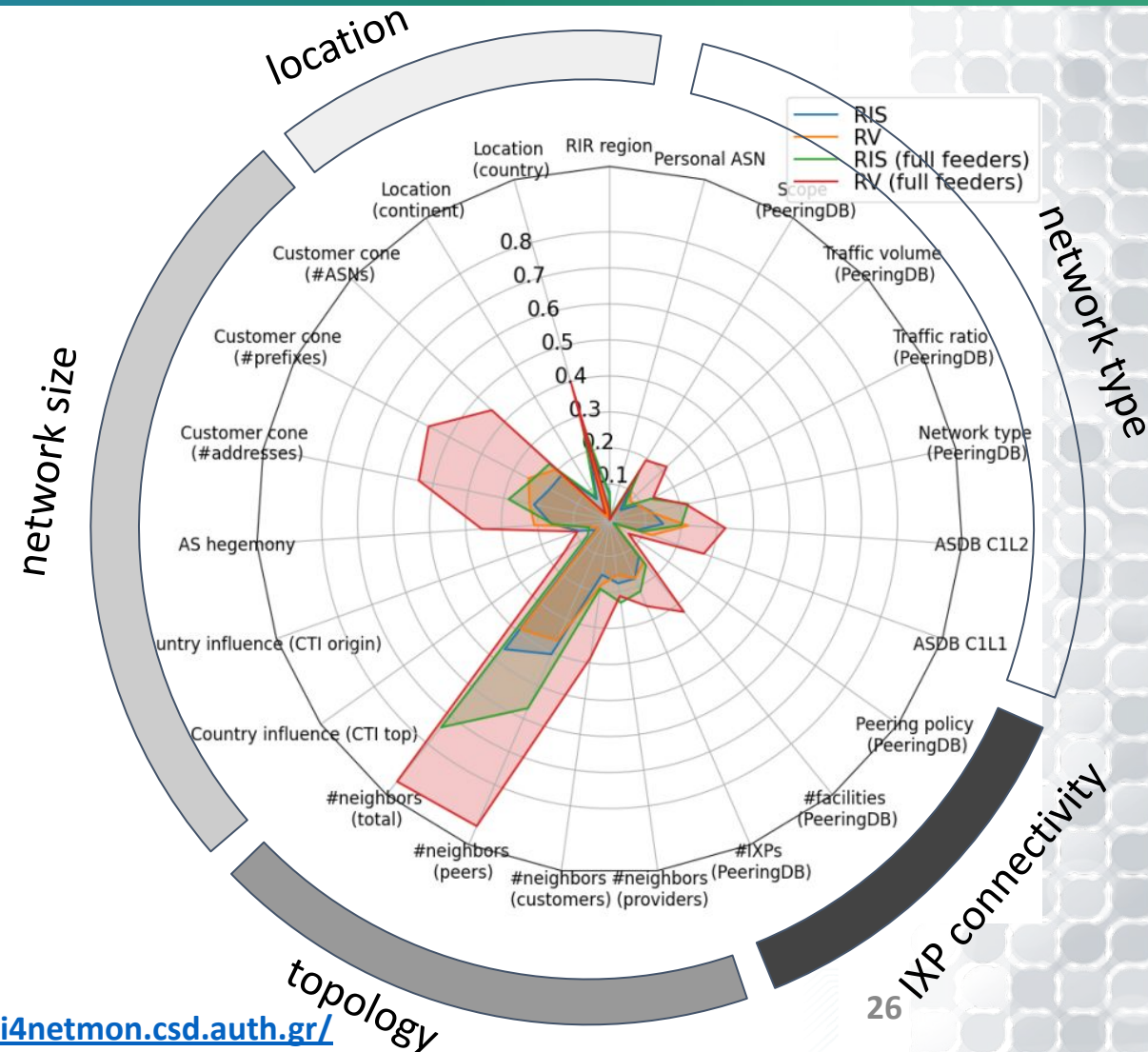
RIPE Atlas, **RIPE RIS** and **RouteViews** have relatively low network-type bias (PeeringDB vs ASDB)





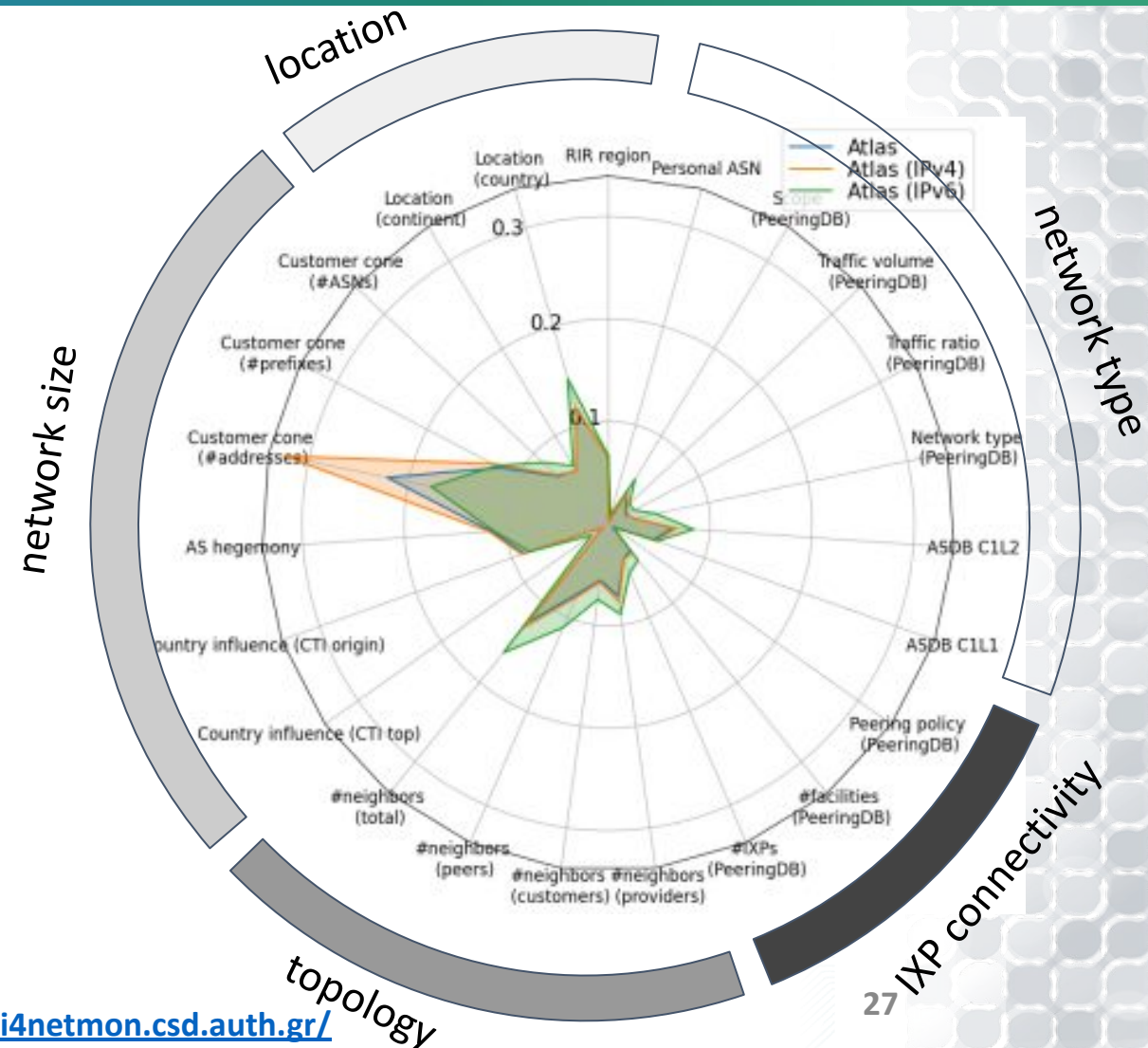
Bias in Internet Measurement Platforms

Full feeds are more biased





Bias in Internet Measurement Platforms

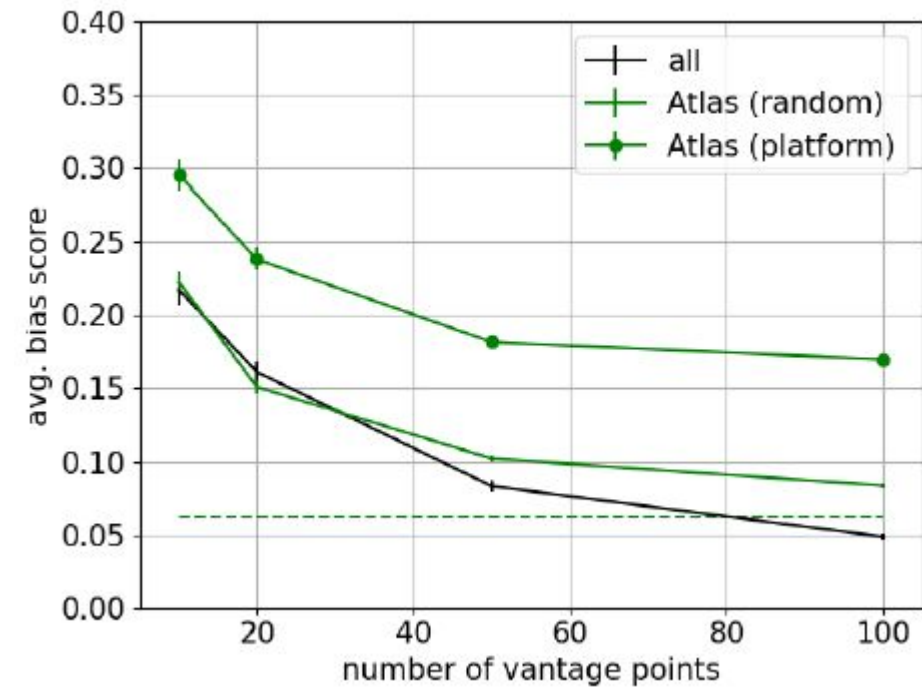


IPv6 Atlas probes are only *slightly more biased* than IPv4



Bias in RIPE Atlas measurements with few probes

- Bias vs. number of probes
- Less probes → higher bias
- Automatic selection by Atlas (“Atlas platform”) is *more biased* than randomly selecting probes (“Atlas random”)!

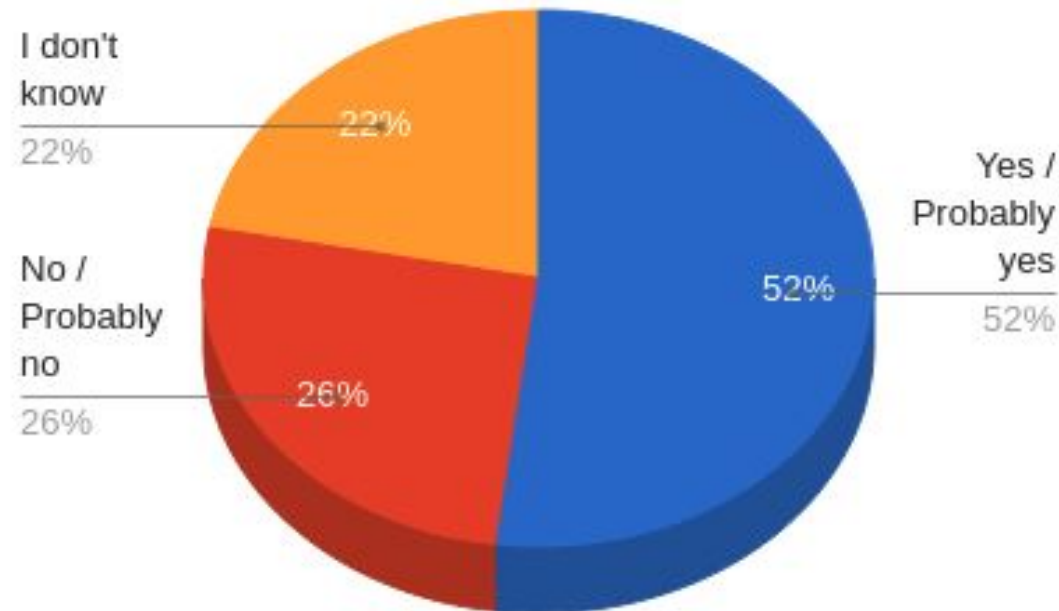




Do people know?

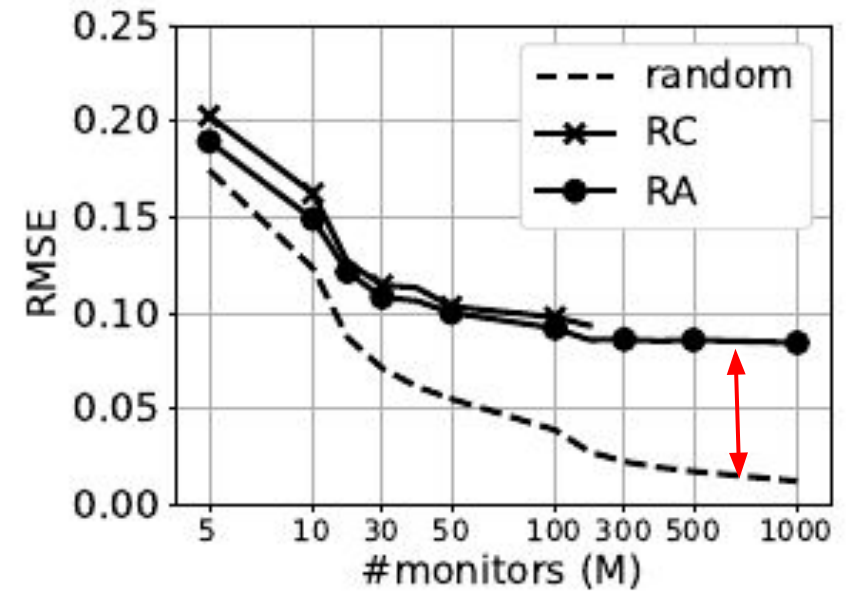
- Not all people know! → our main goal: raise awareness & deepen our understanding

Do you believe there is bias in Internet measurements?



So... what?

- Should I care? → Yes! Bias may affect the insights you get from your measurements
 - e.g., “Estimating the Impact of BGP Prefix Hijacking”, *IFIP Networking, 2021* [[link](#)]
- Be aware of bias! Carefully interpret your results
 - “Which dimensions affect my measurements?”
 - “Is there bias in my dimensions?”



**bias of public
infrastructure**



Dataset, code, API, Web app

- AI4NetMon project <https://ai4netmon.csd.auth.gr/>
 - You can find all the information about the project!
- Code & Data @ GitHub <https://github.com/sermpezis/ai4netmon/>
- API <https://ai4netmon.csd.auth.gr/api/>
 - Documentation @ GitHub
- Web app <https://app-ai4netmon.csd.auth.gr/>



Web app "Show me the bias"

- Available at <https://app-ai4netmon.csd.auth.gr/>



Select a custom set of vantage points. In the boxes below, add a list of ASNs/Probe IDs (only numbers, separated with commas, no spaces; e.g., 174,1299,3333)

Select the type of list of numbers

- ASNs probe IDs

Custom Set #1 (ASNs)

Custom Set #2 (ASNs)

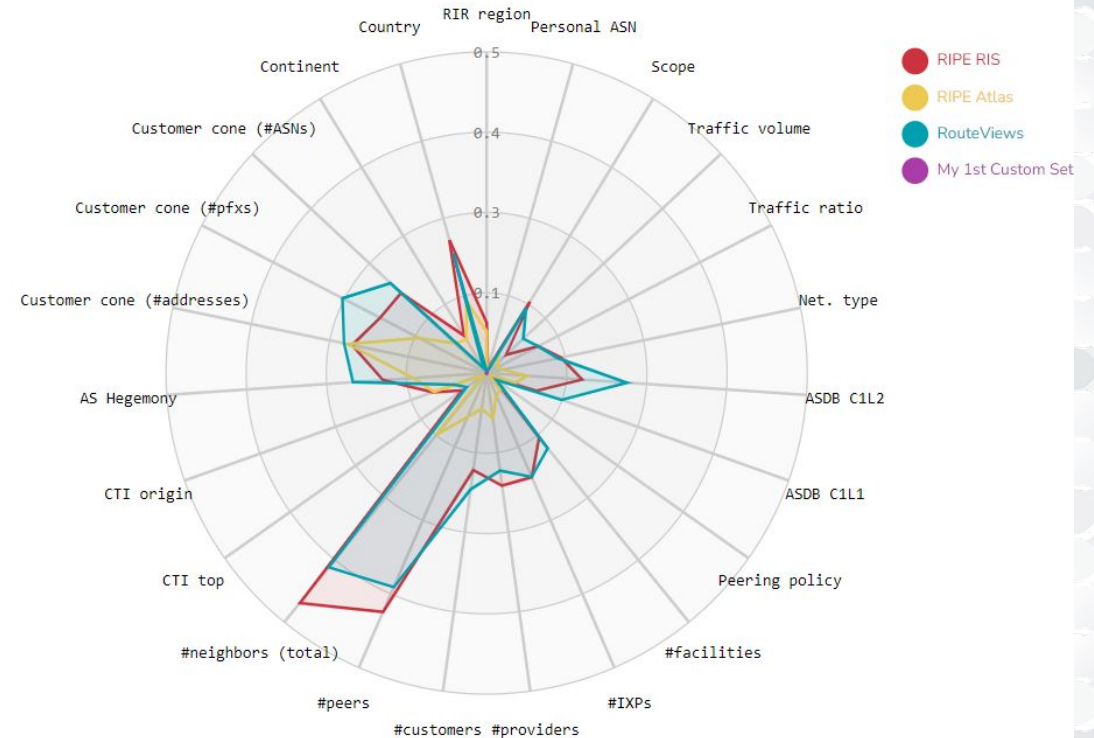
Custom Set #3 (ASNs)

Monitor sets

- RIPE RIS RIPE Atlas RouteViews My 1st Custom Set
 My 2nd Custom Set My 3rd Custom Set

Bias dimensions

- RIR region Country Continent Customer cone (#ASNs)
 Customer cone (#pfxs) Customer cone (#addresses)
 AS Hegemony CTI origin CTI top #neighbors (total)
 #peers #customers #providers #IXPs #facilities
 Peering policy ASDB C1L1 ASDB C1L2 Net. type
 Traffic ratio Traffic volume Scope Personal ASN



Web app "Show me the bias"

- Available at <https://app-ai4netmon.csd.auth.gr/>



Select a custom set of vantage points. In the boxes below, add a list of ASNs/Probe IDs (only numbers, separated with commas, no spaces; e.g., 174,1299,3333)

Select the type of list of numbers ASNs probe IDs

Custom Set #1 (ASNs)

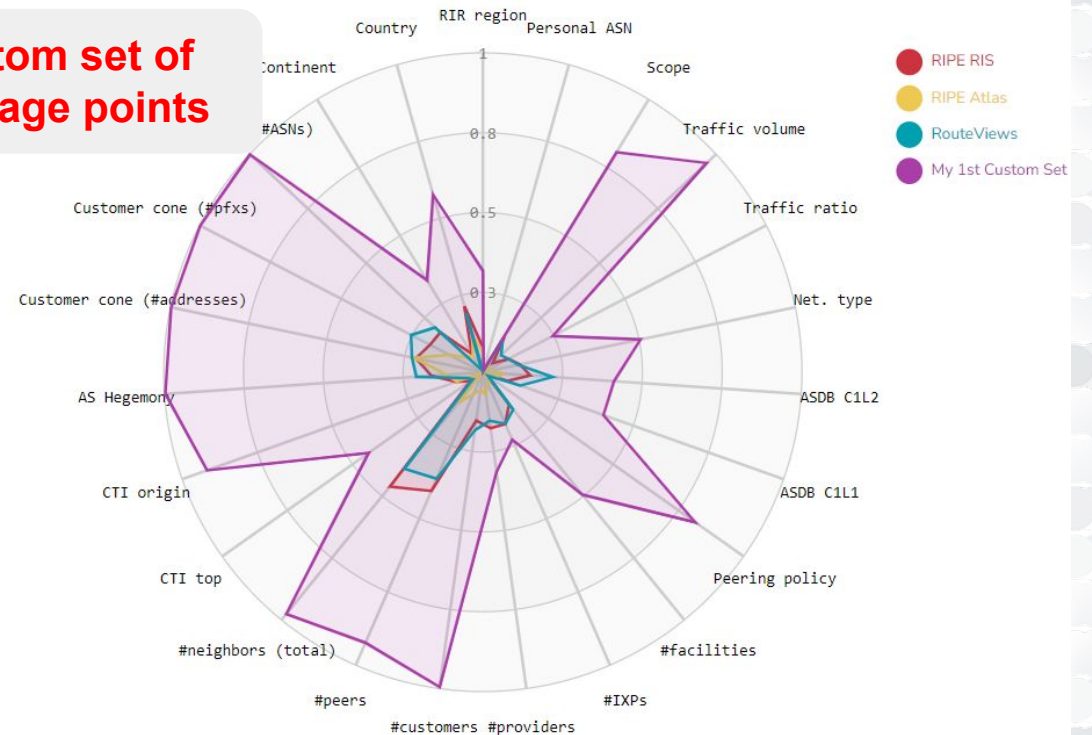
Custom Set #2 (ASNs)

Custom Set #3 (ASNs)

Monitor sets RIPE RIS RIPE Atlas RouteViews My 1st Custom Set My 2nd Custom Set My 3rd Custom Set

Bias dimensions RIR region Country Continent Customer cone (#ASNs) Customer cone (#pfxs) Customer cone (#addresses) AS Hegemony CTI origin CTI top #neighbors (total) #peers #customers #providers #IXPs #facilities Peering policy ASDB C1L1 ASDB C1L2 Net. type Traffic ratio Traffic volume Scope Personal ASN

Custom set of vantage points





Web app "Show me the bias"

- Available at <https://app-ai4netmon.csd.auth.gr/>



Select a custom set of vantage points. In the boxes below, add a list of ASNs/Probe IDs (only numbers, separated with commas, no spaces; e.g., 174,1299,3333)

Select the type of list of numbers ASNs probe IDs

Custom Set #1 (ASNs)

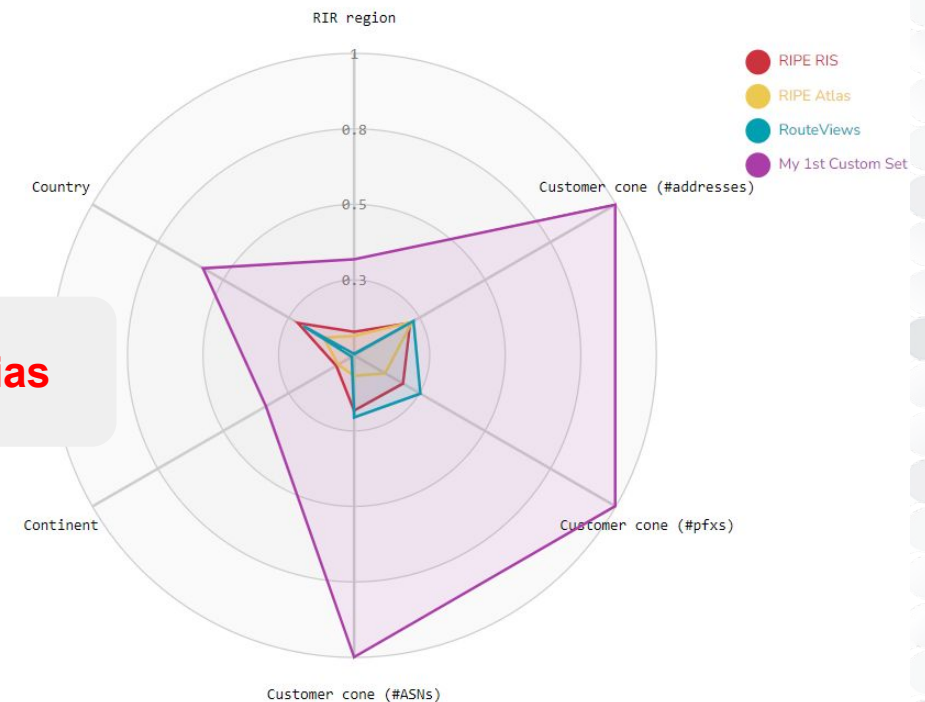
Custom Set #2 (ASNs)

Custom Set #3 (ASNs)

Monitor sets RIPE RIS RIPE Atlas RouteViews My 1st Custom Set My 2nd Custom Set My 3rd Custom Set

Bias dimensions RIR region Country Continent Customer cone (#ASNs) Customer cone (#pfxs) Customer cone (#addresses) AS Hegemony CTI origin CTI top #neighbors (total) #peers #customers #providers #IXPs #facilities Peering policy ASDB C1L1 ASDB C1L2 Net. type Traffic ratio Traffic volume Scope Personal ASN

keep only subset of bias dimensions





Summarizing...

- Our contributions
 - A framework (data, definitions, etc.) to quantify bias
 - Analysis of bias in Internet measurement platforms
 - Code & tools
 - Website <https://ai4netmon.csd.auth.gr/>
 - Web app <https://app-ai4netmon.csd.auth.gr/>
- Next steps
 - **Unbias** Internet measurements [ongoing work] :
 - (a) extend platforms (add extra vantage points)
 - (b) carefully select vantage points (subsampling)
 - **Use cases**: When the bias really hurts our findings?
 - **Bias in ML models** based on data from measurements

