III: Metrics and radar plots

Nick Merrill

November 6, 2019

I've collected metrics for all of 2019, for all of the data sources in our dataset. There may be an issue with our layer 3 (internet interference) data, but as we diagnose, I thought it would be a fair time to introduce the particulars of the metrics and share some preliminary radar charts (with the caveat that data may change).

We may be all on the same page with metrics, but often, the devil is in the details. It's worth going through each layer's metrics and discussing their particulars.

1 One metric per layer

1.1 Layer 2: IPv6 adoption

As a proxy for interlink layer (layer 2) fragmentation, we use Google's per-country IPv6 adoption statistics, which Google collects from all Google services and analytics users. I suspected that they would radically underestimate adoption in China, where Google is blocked, but surprisingly, Google's statistic roughly matches the one given by state-run Chinese media.

1.2 Layer 3: Network interference events

To measure transport layer (layer 3) fragmentation, we use data collected by the Open Observatory of Network Interference (OONI). OONI requires volunteers to install a plugin, which periodically performs tests to measure circumvention on the transport layer. These include mostly state-launched attacks (such as DNS manipulation or traffic filtering), but may also include private-sector manipulation, such as throttling streaming traffic.

Since different countries produce radically different numbers of reports per year, we compute a "rate" of anomalous observations per all reports generated.

1.3 Layer 4: Website ranking locality

We use the Alexa rankings to determine the 50 most popular websites in every country (by traffic). We then compare each country's most popular websites to the worldwide most popular websites, using Levenshtein distance to compute the edit distance between the two lists. This leaves us with a metric expressing how much each country's web browsing habits differs from the global rankings. [1]

1.4 Layer 5: Data locality laws

Right now, I count one type of data locality law: whether a given country restricts the cross-border flow of data (any type). We code this as a binary variable.

One clear issue with this metric: it's binary, which removes some subtlety from the radar chart. I have brainstormed different ways to make this metric a bit more (numerically) nuanced in the future (see 3.1).

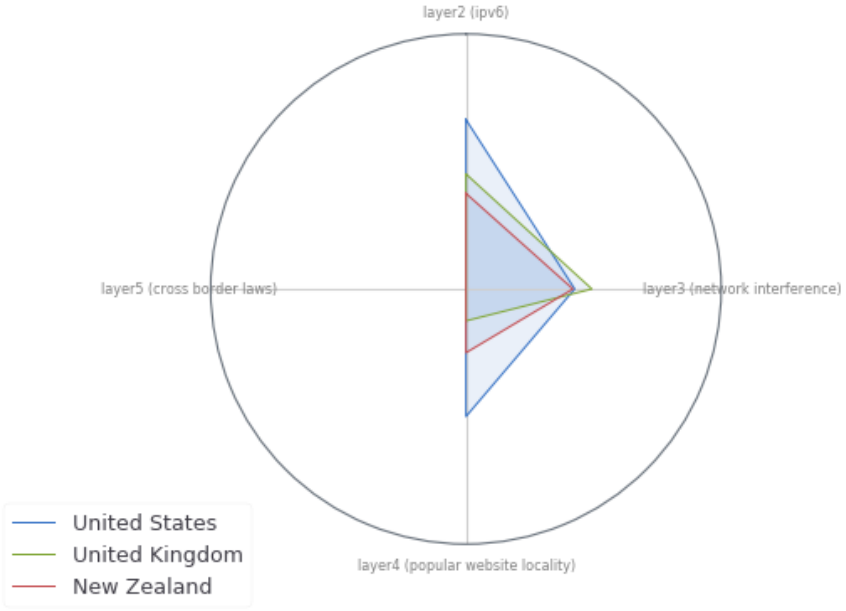
2 Visualizing the metrics: Radar charts

How do we visualize these metrics? Right now, our strategy is to produce a "radar" (or "spiderweb") plot. The radar plot describes a given country's spread across the four metrics we currently measure. By visually comparing these radar plots, we can establish rough "profiles" for countries. [2]

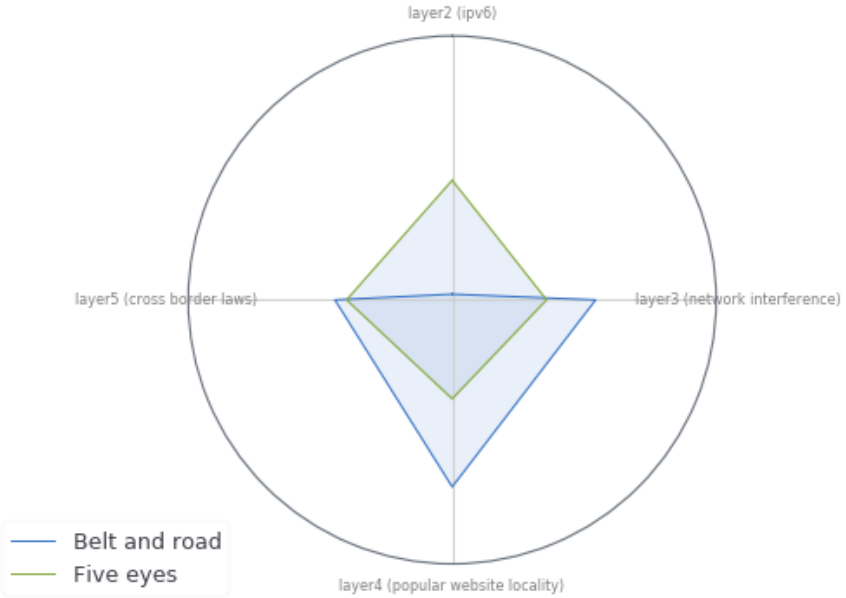
[1] To pump the intuition here... At the bottom of these rankings (most country-specific), we have China; at the top (most similar to the global rankings), we have Luxembourg.

[2] For ease of display in a radar plot, we transform all data to be between 0 and 1, and assure all continuous variables are roughly normally distributed. Setting aside our binary variable (layer 5), both Layer 2 and Layer 4's metrics

are (rougly) normally distributed. Layer 3's metric, on the other hand, is roughly a power law distribution (i.e., a few countries have 10-100x the rate of network inteference as the majority). We "view" these data on a logarithmic scale in the radar plot, (I confirmed that the log-scale data are roughly normal).



Here's an example. It compares the US, UK and New Zealand. This radar chart shows us that the UK and New Zealand are similar, but the United States shows more fragmentation on layers 2 and 3. However, all three countries have no restrictions on cross-border data flow, and they all have about the same amount of network interference. **NOTE**: /Data may change slightly here as we sanity check our preprocessing—layer 3 data in particularespecially /.



We can also average multiple countries' metrics together to create *blocks*. Here's an example comparing Five Eyes countries with Belt and Road countries. Overall, Belt & Road countries have higher website locality (perhaps as a result of censorship), higher rates of network interference, and very low adoption of IPv6. (They have slightly higher rates of having cross-border data transfer laws, but it's not clear yet whether the difference is significant).

3 Reflections

3.1 Introducing variation into Layer 5

This binary variable doesn't give us enough variation. As much as I'm loathe to "generate quantification," one possibility is to bucket data-related laws by domain (cross-border data flows, privacy, data sovereignty), code them as binaries per category, and sum them. This will produce an aggregate score in which all data-related laws are equally weighted. What do folks think about this approach? See Appendix 4.1 for the types of data laws we coded over the summer.

Cons:

- This may get us in trouble with sticklers, who will argue that not all data laws are equally consequential.

Pros:

- It will certainly introduce variation into our metric, and hopefully will produce a normal (rather than bimodal) distribution among countries.

3.2 Dealing with the European Union

For our layer 5 metric, the European Union (GDPR) is tricky here. GDPR does restrict certain types of cross-border data flows. For now, I've counted all EU member states as "yes," as the GDPR laws apply there. However, this law restricts data flows *out of the EU*, not (e.g.) between Belgium and Netherlands. So having each country "inherit" the EU law may be misleading.

In the future, we may want to include EU **and** member states as separate, if geographically overlapping entities for our analysis. Any thoughts on this would be much appreciated.

3.3 Layer 5: Are laws enforced?

Are the laws described in Layer 5 enforced? Our metric does not currently capture this question. It could, but may require some bespoke input from people "on the ground," which we are trying to avoid right now. I'm inclined to ignore this for now, and add it to the "think about later" pile (a pile I am, for the record, diligently collecting).

4 Appendix

4.1 Data laws by type

- Online sales
- Domain name (DNS) registration requirements
- Export restrictions
- Bandwidth, net neutrality
- Lack of safe harbor for intermediary liability
- Sanctions for non-compliance
- Administrative requirements on data privacy
- Data retention

- Restrictions on cross-border data flows
- Other restrictive practices related to business mobility
- Quotas, Labour Market Tests, Limits of Stay
- Other restrictive practices related to competition policy
- Competition
- Copyright
- Patents
- Screening of investment and acquisitions
- Restrictions on ownership
- Technology mandate
- Preferential purchase schemes covering digital products and services
- Discriminatory tax regime on online services
- Discriminatory tax regime on digital goods and products
- Antidumping, CVD & Safeguards
- Applied tariffs on digital goods
- Barriers to fulfillment
- Product screening and testing requirements
- Product safety certification (EMC/EMI, radio transmission)
- Import restrictions
- Censorship and filtering of web content
- Notice and takedown requirement
- Other
- Personal rights to data privacy
- Trade secrets

- Other restrictive practices related to foreign investment
- Restrictions on board of directors and managers
- Taxation on data usage
- Other restrictive practices related to IPR
- Telecom network and base standards
- Local Content Requeriments for commercial market
- Requirement to surrender patents, source codes, trade secrets
- Subsidies and favourable tax regime
- Other restrictive practices related to content access
- Encryption
- Other restrictive practices related to standards
- Discriminatory / disproportionate consumer protection
- Other restrictive practices related to intermediary liability