Final Project Abstract

# Internet Access Disparity

Aditya Kumar Bej, Shivani Kalamadi
ECS 252 Computer Networks WQ 2024

## Research Question

We aim to answer the following research questions in our study:

1. To what extent do differences in socioeconomic factors such as income level, geographic type (rural/urban), population diversity impact internet access?
2. Does the presence of IXPs in certain regions differentially impact internet access in speed, reliability, and internet affordability across demographics?
3. How does the presence of political biases influence internet access?

## Hypothesis

**Hypothesis 1: Regions with lower socioeconomic status are served by lower-quality network providers, leading to significant disparities in internet access, speed, and reliability.**

We anticipate that variations in socioeconomic factors such as income level, geographic type (rural/urban), population diversity significantly influence internet access **[1]**. These factors compound to disparities, creating a complex web of challenges that contribute to unequal access to high-quality internet services across diverse demographics and regions.

**Hypothesis 2: The growth of Internet Exchange Points (IXPs) in specific regions over the last five years is likely to impact internet access disparities.**

We expect that increased IXP development will lead to improvements in speed, reliability, and affordability, thereby narrowing the gaps across demographic groups **[2]**. The enhanced local internet infrastructure is anticipated to contribute to a more equitable distribution of high-quality internet services.

**Hypothesis 3: The presence of political biases significantly influences internet access**

Our own hypothesis suggests that political biases—stemming from government ideologies, regulatory bodies, and policy-making processes—can lead to unequal distribution of internet access across different regions, communities, and socio-economic groups.

## Related work

[1] The article highlights the digital divide despite a significant increase in global internet users and penetration from 2021 to 2022 due to the pandemic. As of 2022, 53% of the world population lack high-speed broadband. **However**, the study doesn't delve into technical details like how such disparities were calculated or even specify which regions had better internet access than others.

[2] The paper surveys largely fragmented public information on Internet Exchange Points (IXPs), detailing their technical aspects. It underscores the crucial role IXPs play in the current Internet ecosystem and highlights how IXP-driven innovation in Europe is reshaping the global Internet marketplace. While the study offers substantial evidence highlighting the significance of Internet Exchange Points (IXPs), **it does not** delve into how IXPs in regions with suboptimal internet conditions might experience benefits from IXPs.

[3] The study shows a persistent political bias in internet access globally only among ethnic groups. The global expansion of the Internet is thought to enhance government transparency and democracy, but the study shows marginalized groups have significantly lower Internet access rates compared to those in power, highlighting a barrier to the liberating potential of technology. **Nevertheless**, the study is outdated, and due to its global scope, it lacks in-depth exploration of specific countries or regions.

## Theoretical contribution

This study aims to contribute in the following ways:

1. This project aims to contribute a novel dataset and analysis framework for evaluating the combined impact of socioeconomic factors and network provider quality on internet access disparities. By building upon existing tools like M-Lab Speed Test, Ookla's Open Data, and RIPE Atlas data, we will create a geographically weighted, multi-layered dataset encompassing speed, reliability, network provider quality, and relevant socioeconomic indicators for select regions. This innovative approach, focusing on specific neighborhoods within diverse regions, allows us to map internet access disparities with granular detail.
2. We will collect the current and past 5 years of information on the number of Internet Exchange Points (IXPs) in selected regions to explore their impact on internet access.
3. We aim to conduct a comprehensive literature review to understand the existing political climate for the selected regions. For a particular region, we will collect data through various social media and news, and compare findings with our technical data from Ookla, Mlab and RIPE Atlas datasets to draw conclusions about any potential biases.

We will make the findings publicly available through the GitHub repository - https://github.com/AdityaKumarBej/Internet-Access-Disparities

## Method

The research methodology is designed to address the complexities inherent in analyzing large-scale internet access disparities. By integrating **diverse data sources and analytical tools such as RIPE ATLAS**, our approach is well-suited to investigate the research questions and test our hypotheses effectively.

We commence our **large-scale data analysis** by harnessing data from premier network metrics providers such as MLAB, OOKLA, and RIPE Atlas. This multilateral data acquisition enables a robust foundation for our subsequent analysis. To ensure the accuracy and relevance of our data, we employ a custom interface for data extraction, followed by a rigorous data ingestion process that includes data correlation and cleaning. The ingested data is then stored in our local file servers. We leverage Git for version control to manage our codebase and datasets efficiently.

For **data visualization**, we utilize Tableau and Python's rich libraries for plotting, thus enabling us to transform complex data sets into comprehensible visual representations. These visualizations are not only instrumental for exploratory data analysis but also serve as a medium to communicate our findings effectively.

In parallel, our **low-level data analysis** targets specific regions for a granular analysis. This involves a careful selection process informed by a cross-reference of network data against **socio-economic and political climate data** from various authoritative sources, including ITU and census websites. By integrating **dynamic traceroute information** from RIPE Atlas, we gain real-time insights into network topology and performance. An added parameter is the IXP data from **CAID Dataset** **[4]** which will allow us to get the information on the geographical locations. The affordability of ISPs can be understood by a quick search of existing ISPs and their plans of that particular region.

The final phase of our research encapsulates the synthesis of our findings into comprehensive reports. These reports will feature analyses and hypotheses on the identified disparities, backed by potential results graphs that are both plausible and illustrative of our research outcomes. The graphs will underscore patterns and trends in internet access quality across demographics, providing a clear answer to our hypothesis.

In summary, our method stands out for its validity, coherence, and careful consideration of various data dimensions. It is tailored to address our research question meaningfully and is poised to yield insights that could influence policy and practice in the realm of internet access equality.

One example of the on-going work (in GitHub) is visualized in Figure 1 which was plotted from data collected through OOKLA Dataset. We will be doing a similar visualization for both the large scale and small scale analysis and have more network metrics (which is mentioned in Figure 2 Data Parameters table)

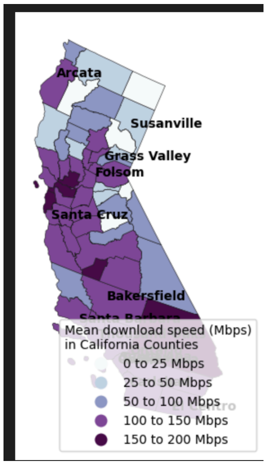

**Figure 1: California Counties Internet Metrics**

## Analysis and/or systems building plan

### Large Scale Data Analysis

**1. Data Sources:** Our methodology utilizes multiple data sources for network metric data:

- **MLAB** (Measurement Lab): Provides open-source internet performance data.
- **OOKLA**: Offers data from speed tests that measure internet connection speeds.
- **RIPE Atlas**: Maintains a global network of probes that measure internet connectivity and reachability.

**2. Data Pull:** Data is extracted from these sources, through a custom interface designed to query and collect the necessary information from the above stated data sources.

**3. Data Ingestion:** The collected data may undergo an ingestion process to a data storage system where it can be processed. Data ingestion will involve:

- **Data Correlation:** Associating different data points with each other based on common identifiers.
- **Data Cleaning:** Removing or correcting erroneous or incomplete data entries.

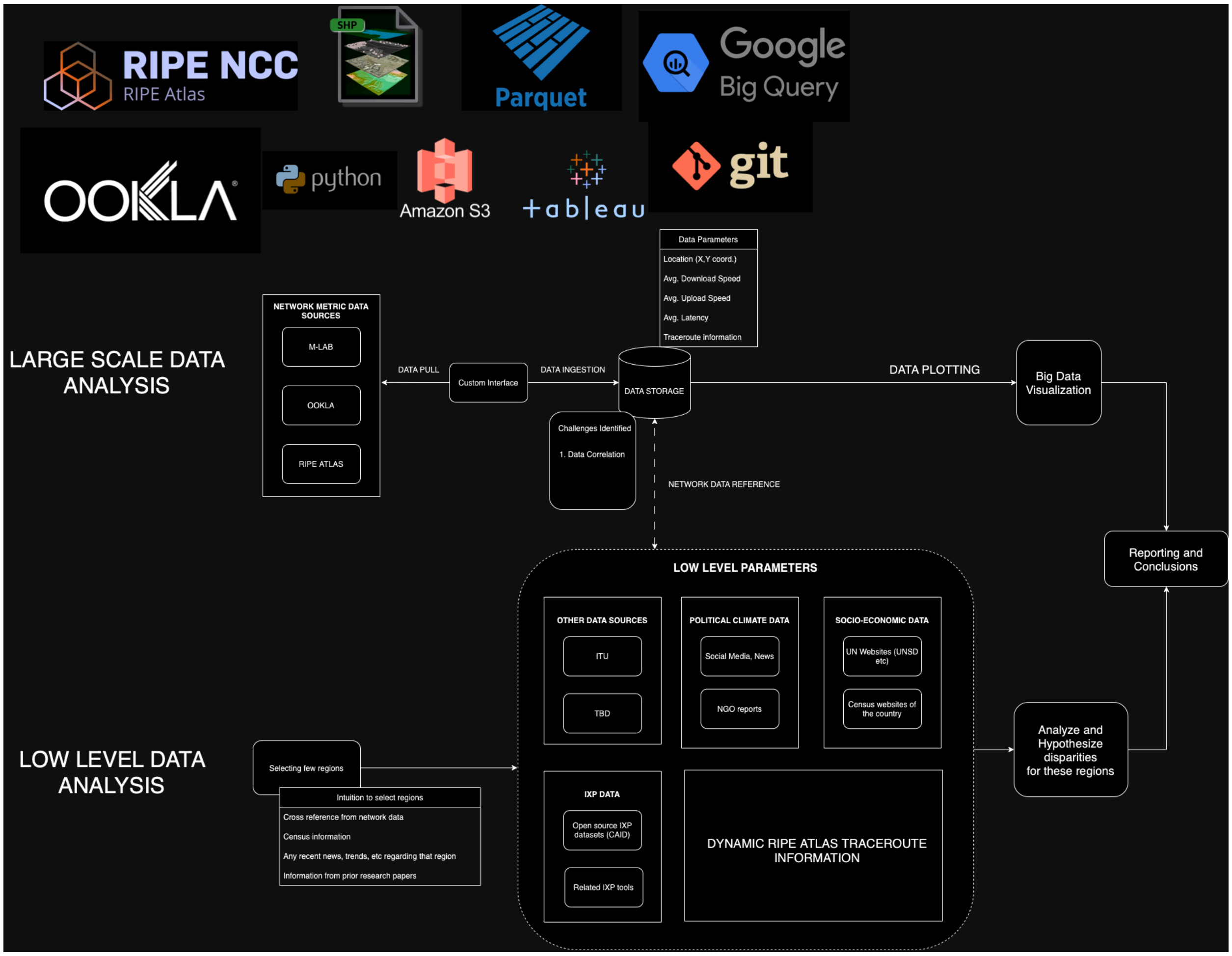

**Figure 2. Proposed System Design**

**4. Data Storage:** The ingested data is stored in a format and system that facilitates analysis. We will be storing this on our local file server systems since we will be dealing with large volumes of data which will amount to nearly 50 GB.

**5. Data Plotting:** The stored data is then plotted for visualization. This could involve:

- **Tableau:** For creating interactive data visualizations.
- **Python:** Using libraries like Matplotlib, seaborn, or Plotly for creating static or interactive plots.

### Low-Level Data Analysis

*Note on the selection of specific regions*: A few states from the US and all the metropolitan cities in the following countries/continent: India, Africa, Middle East and SouthEast Asia. This selection of countries will be refined once we have the large scale dataset ready.

**1. Low-Level Parameters:** More detailed analysis is performed with a focus on specific regions, informed by:

- **Selection of Few Regions:** Targeting specific areas for in-depth analysis.
- **Cross-Reference:** Using network data, census information, recent news, trends, and research papers to understand the context of the selected regions.

**2. Other Data Sources:** Additional data to support the analysis might come from:

- **ITU (International Telecommunication Union):** Provides standardized global telecom data.
- **IXP Data:** Data from internet exchange points from open source IXPs datasets like CAIDA.

**3. Dynamic RIPE Atlas Traceroute Information:** Utilizing traceroute data from RIPE Atlas for real-time path analysis which can provide insight into network performance and topology.

**4. Political Climate and Socio-Economic Data:** Incorporating contextual data that could impact internet access and performance, such as:

- **Social Media and News:** For up-to-date information on the political climate.
- **NGO Reports:** Reports from Non-Governmental Organizations on regional developments.
- **UN Websites (UNSD):** United Nations Statistical Division for global statistical data or any other relevant UN body
- **Census Websites of the Country:** For demographic and socio-economic data.

### Conclusion and Reporting

After conducting both large-scale and low-level data analysis, the findings are compiled and synthesized into reports. These reports likely include:

- **Analysis and Hypothesis for Disparities:** Examining the reasons behind disparities in internet access and quality among different regions.
- **Reporting and Conclusions:** Presenting the analysis results, drawing conclusions, and possibly making policy recommendations.

This methodology is comprehensive, integrating both macro and micro-level data analyzes to understand internet access disparities and network performance. It leverages a wide array of tools and data sources to provide a multi-faceted view of internet metrics and contextual factors. The end goal is to report findings that are well-informed by a variety of qualitative and quantitative data points.

## Biggest risk

Although we are using real time data and existing databases from RIPE Atlas, the biggest risk of this project is the potential inconsistency in crowd-sourced data quality and availability across different regions from M-Lab and Ookla. This may impact the accuracy of correlating internet access quality with socioeconomic factors.

Another risk is data correlation and absence of data from various multiple sources. The data headers between OOKLA, M-LAB and RIPE Atlas might not be the same and can cause issues. The best way is to standardize the dataset when collecting information. There are a few headers which are common among the datasets which should be enough for us to conduct our analysis.

Another identified risk is the biases on our political influence research through news and social media. This can be mitigated/reduced through correlating the research from various trusted sources such as recognized NGO reports or news channels such as Reuters.

## Tech stack/Tools

1. Python (for various data fetching, ingestion and analysis scripts)
2. Git
3. RIPE ATLAS
4. Google Big Query (where M-LAB dataset can be queried from)
5. Amazon S3 (where ookla dataset is hosted)
6. Tableau or Python's graph related libraries (matplotlib)

## Contribution

**Aditya Kumar Bej** : Technical Methodology, Systems Building Plan, IXP related research (CAID Dataset analysis), OOKLA dataset analysis and code.

**Shivani Kalamadi** : Research on existing works, M-LAB Dataset analysis, Literature Review to analyze Research Questions and come up with hypotheses. Modified Technical Methodology for Low-Level Analysis.

## References

[1] Fixing the global digital divide and digital access gap
https://www.brookings.edu/articles/fixing-the-global-digital-divide-and-digital-access-gap/

[2] Nikolaos Chatzis, Georgios Smaragdakis, Anja Feldmann, and Walter Willinger. 2013. There is more to IXPs than meets the eye. SIGCOMM Comput. Commun. Rev. 43, 5 (October 2013), 19–28. https://doi.org/10.1145/2541468.2541473

[3] Nils B. Weidmann et al.x, Digital discrimination: Political bias in Internet service provision across ethnic groups. Science 353,1151-1155(2016). DOI:10.1126/science.aaf5062

[4] https://publicdata.caida.org/datasets/ixps/