Ivan Ivanov

2672214

Bachelor’s Thesis

Vrije Universiteit Amsterdam

Supervisor: Ivano Malavolta

Second Reader: Kousar Aslam

CO2 and The Internet: a quantitative study on environmental impact

Contents

[Abbreviations 2](#_Toc106646663)

[Abstract 3](#_Toc106646664)

[1. Introduction 3](#_Toc106646665)

[Thesis Structure 4](#_Toc106646666)

[2. Related Studies 4](#_Toc106646667)

[3. Research Method 4](#_Toc106646668)

[Methodology 5](#_Toc106646669)

[Errors and Limitations 6](#_Toc106646670)

[Software 7](#_Toc106646671)

[4. API Calculations and scope 8](#_Toc106646672)

[Factors 8](#_Toc106646673)

[System Boundaries 8](#_Toc106646674)

[5. Data Analysis & Results 9](#_Toc106646675)

[General Analysis 10](#_Toc106646676)

[Overview 10](#_Toc106646677)

[Outliers 11](#_Toc106646678)

[Top and Bottom 1000 15](#_Toc106646679)

[Hosting Type 16](#_Toc106646680)

[Per Domain 17](#_Toc106646681)

[Original Domains 17](#_Toc106646682)

[Regional Domains 18](#_Toc106646683)

[6. Discussion & Results Analysis 19](#_Toc106646684)

[7. Reflection 19](#_Toc106646685)

[8. Conclusion 20](#_Toc106646686)

[References 20](#_Toc106646687)

[Appendices 22](#_Toc106646688)

# Abbreviations

|  |  |
| --- | --- |
| TWh | Terawatt-hour |
| MMT | Million Metric Tons |
| IEA | International Energy Agency |
| EPA | (US) Environmental Protection Agency |
| WRI | World Resources Institute |
| KWg | Kilowatt-hour per GB |
| TLD | Top-level Domain |

# Abstract

[abstract body]

# Introduction

Awareness towards the threat of climate change has increased in the in the last years (Leiserowitz, 2019. Sustainability, a practice concerned with the recognition of the finite nature of resources, and their conscious usage with regards to future generations has also experienced a similar growth in recognition (EPA, 2021). As a result of this, there has been a larger amount of CO2-related scientific publications published in the past years (Fabbrizzi et al., 2016).

Along with the surge in climate change related concerns, there has also been a growth in global Internet data traffic. In the year 2020 alone (Krisetya et al., n.d.) web traffic exchange has expanded with an average of 35% in size and has been previously predicted to experience further expansion with close to five and a half billion internet users expected by 2023 (Cisco Systems, Inc., n.d.). This explosion in traffic growth is driven partially by the SARS-CoV-2 pandemic’s effect, but also by the consistently upward developments in usage of Internet-connected devices from 2011 onward (Telefonaktiebolaget LM Ericsson, 2021, p. 3), and it is mostly caused by video streaming, conferencing, online gaming, and social networking (IEA, 2021). Device adoption rates vary widely per region, though it is estimated that in 2016 a United States citizen owned about 7.8 Internet connection-capable devices on average and consumes 97 gigabytes of data per month, several times more than the world average. Meanwhile a Chinese citizen had 2.5 and uses only 12 gigabytes. In the next two years, the statistics for the USA have increased to 10 devices and 140 gigabytes respectively (The Shift Project, 2019, p. 61).

As these devices become more and more of a central point in human lives the Internet usage grows rapidly as well. 92% of internet users nowadays access the web with mobile phones (Statista & Johnson, 2022). Because of this, the amount of site hits grows larger with each year and that, in turn, affects the overall data consumption as well (United Nations, 2019, p. 15).

Data centers are the backbone of the internet. They consume around 1-1.5% of global electricity use (Masanet et al., 2020) which translates to an estimated use of around 200-250 TWh of electricity per year (IEA, 2021) with some estimates going as high as 400 TWh. In general, the electricity usage has fallen “by a factor of four” since 2010 due to hardware improvements in processor efficiency and idle power usage (Masanet et al., 2020). A report by the International Energy Agency (IEA) from 2014 stated that the development of energy efficiency metrics was one of three key considerations needed for effective policy making and a reduction of the networks’ energy usage and in turn, the carbon dioxide footprint (IEA, 2014b).

CO2 emissions have risen rapidly for more than a hundred years. Greenhouse gases, especially carbon dioxide (CO2) emissions, are viewed as one of the main reasons of climate change, and it has slowly become one of the most important environmental problems for our world (Rehman et al., 2021). In the early 1900s there have been estimated to have been around 1000 MMT of carbon dioxide. In the 2010s these numbers reached almost 10,000 MMT (EPA, n.d.). This has been driven by a variety of factors: increased globalization, fossil fuel usage, population increase and more (Lindsey, 2020). According to the World Meteorological Organization’s State of the Global Climate report, the global average temperature in 2020 was about 1.2°C above preindustrial level (World Meteorological Organization, 2021). Numerous attempts have been made to decrease the emissions. To mitigate the threat, the Paris Agreement called to limit global warming to below 2°C but preferably to 1.5°C, a number comparable to the previously measured pre-industrial levels. Additionally, there have also been public awareness campaigns and government reforms (Department for Business, Energy & Industrial Strategy, 2021) (Ge & Ross, 2019). The IT sector alone amounts to 1.4% of those global emissions but it can be reduced to 20% less of that if a switch to renewables were to happen (Telefonaktiebolaget LM Ericsson, 2020).

There has been research made about the IT sector’s emission generation but despite the importance of ICT and the ubiquity of the Internet, but none of them have attempted to analyze the web’s CO2 footprint. In this thesis, an attempt will be made to create a clear overview of some of the web’s most popular websites and the amount of CO2 they generate.

To do so, the following research question will be addressed:

*What is the current state of energy consumption of Web sites?*

## Thesis Structure

In the next chapter a series of related papers are briefly examined. Chapter 3 talks about the research methods used and Chapter 4 focuses on the methodology and the formulas used to calculate the final numbers. Chapter 5 is an analysis on the received data and in the last two chapters an answer to the research questions will be provided, the results will be discussed and there will be suggestions for further research as well.

# Related Studies

Previously, separate studies have analyzed the electricity consumption of the average Internet data transfer, for various devices (Thiagarajan et al., 2012; Zhu & Reddi, 2013)[!]. Direct estimate comparisons between these are difficult to do because of inconsistent usage of methodologies and data uncertainty. Some, are based on either estimates of regional or of worldwide energy consumption while on the web, combined with traffic estimates to compute the amount of energy consumed per some data amount. The distinction with the largest influence on the result is how the analysis boundaries have been set. Though some studies include the terminal equipment (e.g., personal computers and servers) within the system boundaries (Taylor and Koomey 2008; Weber et al. 2010), others do not (Hinton et al. 2011; Kilper et al. 2011). Most studies include the overhead for cooling and power distribution, but Lanzisera and colleagues (2012) do not. Because of these differences and of a lack of access to more modern and detailed data regarding the exact electricity usage habits of modern data centers, a comparison of such statistics will not be undertaken here. Instead, there is a focus on a data which is particularly well characterized and the system boundaries clear and consistent. The paper will be focused strictly on recently gathered data which reflects the state of the internet nowadays and will explicitly consider websites only. [!]

# Research Method

The purpose of this chapter is to explain the research methods used in this paper.

## Methodology

To write this paper, books, academic articles, journals, reports, and statistical analyses have been used. Additionally, the data used has been collected from several sources: Website Carbon’s API, [list other sites].

The data analysis will be strictly quantitative. The goal here is to analyze the impact of our internet surfing habits on a relatively large scale and to compare an innocuous habit with the real-world impact it has. Due to the nature of the service that the data is sourced from, the analysis will be done entirely on the homepage of a given website and because of that there will be no focus on analysis of common types of data transfer like streaming video from a particular streaming service, loading multiple pages from the same website or infinitely scrolling pages (e.g., Twitter, Facebook, NBCNews.com).

The main data has been collected by using the Website Carbon API, which is located at <https://www.websitecarbon.com/>. It is an online tool created by Wholegrain Digital that provides an estimate for the amount of CO2 a website generates. It awaits a query in the form of a URL address and returns a JSON file containing several statistics about the website.

The dataset where all the statistics have been collected is a .csv file generated from the data retrieved by the aforementioned API. It contains information about the 50,034 most popular websites and it has the following eight features:

1. Website URL
2. Type of hosting – Depends on the energy source used by the data center. Saved as either ‘True’ for websites hosted by a service provider using green energy or ‘unknown’ for those whose green status could not be determined. The status is determined by the Green Web Foundation’s own API. ([found here](https://www.thegreenwebfoundation.org/)) There, any websites that is hosted by a ‘Green’ data center is shown as using Green energy (Note: Not all centers mentioned on the GWF website use Green energy, in some cases they use standard grid and the emissions are offset afterwards. Additionally, the hosting status can sometimes not be detected dependent on whether or not the website is using a CDN. In such a case the IP address of the host cannot be identified and the CDN is deemed to be the host.). If the website is found to be using green energy, then the carbon emissions are reduced.
3. Number of bytes – The amount transferred upon the initial page load, provided that the website has not been visited before.
4. Number of bytes (adjusted) – An adjusted value for the second visit of a website which takes into consideration browser caching.
5. Percentage of sites it is cleaner than – A simple comparison between the amount of CO2 the currently tested website generates and the others in the database. Comparison is done at [ ].
6. Energy – The amount transferred upon a single page load, in Kilowatts per Gigabyte.
7. Grams and liters of CO2 transferred by a renewable grid
8. Grams and liters of CO2 by a standard national grid.

The website list we are using has been sourced from the [Tranco](https://tranco-list.eu/) list of 1 million most popular websites. This is a list which uses averaged data from four other ranking providers (Alexa, Cisco Umbrella, Majestic and Quantcast). The reason for using Tranco, and not either of the four other rankings is that they have been found to often disagree on which sites are actually the most popular. Those lists can change daily and are manipulatable by third parties. The data has been sourced from the original rankings and then averaged over a thirty-day period (le Pochat et al., 2019). The list used to write this project was retrieved on 03/05/2022.

The dataset that was used for analysis has been sourced between the period of 04/05/2022 and 21/5/2022.

### Errors and Limitations

The set has been cleaned up of any accidentally repeated data and several manual edits have been made for the following issues:

1. Google redirects: Some URLs, like several of Google’s regional domains and multiple other unrelated websites) redirect to either the main Google.com domain or one of the regional variations. This created several hundred duplicate “https://www.google.[region]/” entries. In this case the redirected entries have been removed and only the first one from each region has been kept with the presumption that it is the original one.
2. General regional redirects: The same issue appeared for several other websites and has been dealt with in the same way.

Additionally, the data has been formatted to account for visual clarity and ease of use (decimal sign placement, general formatting).

An occurring issue encountered during this data collection process was also a large number of websites which could not be analyzed for different reasons. These reasons are as follows:

1. The website was offline.
2. The website was hosted by Cloudflare – In this case the API returned a response in an HTML form and not JSON. Each time that happens the corresponding website is skipped.

A few other reasons, as mentioned by Wholegrain Digital on their website are:

1. The website can be accessed by the public through a standard web browser.
2. It does not require login.
3. It allows search engines.
4. Contains unique content aimed at human visitors – this excludes holding pages, error pages, server notification pages, demo pages or pages that are generally useless (this is highly subjective).
5. Is free from illegal or explicit content.

The original goal of the project was to parse the entire one million website list. Unfortunately, another limitation was encountered at the data collection process: the API used has a daily limit of 25,000 hits available to it and it is also shared with other users. Because of that the number of sites that could be parsed per day was no more than two to three thousand.

Overall, the first 65,600 websites from the Tranco list were parsed. 52,431 of those were actually processed (due to the issues mentioned above) and after the removal of any duplicates there were 50,034 usable websites left. The loss from parsed to parsable is 21.1% and from parsable to usable is an additional 4.6%.

## Software

The research question and the scope of the paper were defined in the previous sections. Here, the tools used for the data collection and manipulation processes are described.

Essential software used:

* Visual Studio Code 1.68.1 (for Windows)
* Jupyter v2022.5.1001601848
* Python 3.10.4

There were two main tools used for data collection:

* An HTTP Request/Response program (get\_requester.py) which main function is to send out asynchronous requests to the Website Carbon API and receives JSON formatted responses in return. It has been written with Python 3.10.4 64bit.

Additionally, these libraries were used:

1. ASYNCIO ([Documentation](https://docs.python.org/3/library/asyncio.html)). Part of the Python Standard Library. Utilizes concurrent programming concepts to make asynchronous programming possible. It is used because asynchronous requests are usually multiple times faster than sequentially programmed ones, depending on the implementation and the receiving server’s limitations. Concurrency allows for different parts (e.g., functions) of a program to be executed out of order. It means that executing functions can be done “in parallel” which significantly boosts the overall speed of execution of the program. Note that ASYNCIO does not truly support parallelism, but the execution of functions is so quick so that to the purposes of this experiment and the bandwidth of the data processed it works essentially the same.
2. AIOHTTP ([Documentation](https://docs.aiohttp.org/en/stable/)). Written by Nikolay Kim and Andrew Svetlov. A client/server library that utilizes ASYNCIO for making asynchronous HTTP GET requests targeted towards a public endpoint from the Website Carbon API.
3. Throttler 1.2.1 ([Documentation](https://pypi.org/project/throttler/)). Used for throttling the amount of outgoing GET requests, as to not overload the receiving server.

* The Website Carbon API which provides the data to the Python program.

And the following tool was used for the data manipulation and analysis process:

* An .ipynb notebook file (notebook.ipynb) which was used in VSCode to combine the collected .csv files into one file (called main.csv), format it, clean any potential issues with it and generate data for the thesis.

Additionally, these libraries were used:

1. Pandas 1.4.2 ([Documentation](https://pandas.pydata.org/docs/reference/)). A Python written software library used for data science purposes. This is the library used the most in this project for reading files, plotting graphs, storing data and more.
2. NumPy 1.22.3 ([Documentation](https://numpy.org/doc/stable/reference/index.html)). Similar to Pandas, but mostly focused on array and math function handling. Used for a few functions in the notebook file.
3. Matplotlib 3.5.2 ([Documentation](https://matplotlib.org/stable/api/index.html)). A plotting library, written for Python as well and used for a few of the plots present in the thesis.
4. Tld 0.12.6 ([Documentation](https://tld.readthedocs.io/en/latest/)). A small package which main function is to extract the top-level domain of the URL’s present in the main dataset.

Python’s pathlib, warnings, re and os were used as well for miscellaneous purposes.

# API Calculations and scope

### Factors

The amount of energy and emissions generated by a webpage are calculated with the following factors:

* Data transfer over the wire – the amount of data that is transferred from the server to the user upon a page load.
* Energy intensity of web data – The amount of energy used by data centers, networks, and the user’s device to load a page. An estimated average is used here due to the many different variations possible when all devices and computers are considered.
* Energy source used by the data center – as described in “Type of hosting” previously.
* Carbon intensity of electricity – Carbon intensity is the amount of carbon dioxide generated to create a unit of electricity (nationalgridESO, n.d.). Usually measured in grams of CO2 per kilowatt-hour. Here it is based on the international average for grid electricity.
* Website traffic – The number of page views on a website. Multiplying the amount of carbon generated per page view by the number of expected views for a website gives an overview of the actual impact a website has.

### System Boundaries

System boundaries are defined in research by Klaus Büchel (1996) as what “define the processes to be analyzed with regard to material and energy flows and emissions.”. As such they act as “scope limiters” of sorts on the material that is being analyzed and contextualize the goals of the API.

To bring an accurate estimate to the energy usage of the network, the system boundaries have to be defined first. Defining them to be smaller in scope would lead to a misrepresentation of the energy output and usage of the hardware involved here(again, data centers, networks, and end devices). If the opposite were to happen, the broadening would overestimate the amount of elements/hardware with any influence that need to be looked at and add unnecessary complexity.

The boundaries set here for the different system segments are based on Anders Andrae’s (2020) ‘New perspectives on internet electricity use in 2030’ study.

* Consumer device use: 52% End-users, 25% of those are returning visitors.
* Network use: The data that is transferred through the network. 14% of the system.
* Data center use: The energy used by the centers for operation. 15% of the system.
* Hardware production: An estimate for the energy used to create all of the devices taking part in the data transfer process. 19% of the system.

Additional estimates for the average energy usage of all devices have been derived for the purposes of this API, both from the Andrae study and from other sources as well. As far as carbon intensity goes, there is an average of 442g/kWh used here, which is sourced from [Ember’s Data Explorer](https://ember-climate.org/data/data-explorer/).

The key metric used in the calculations is kWh/GB, or “kilowatt-hour per gigabyte” – the kilowatts per hour for each gigabyte of data transferred.

With all of that said, the API uses a number of formulas to make the exact calculations. They work as follows [(Sustainable Web Design 2022](https://sustainablewebdesign.org/calculating-digital-emissions/)):

#### Energy per visit in kWh (E):

E = [Data Transfer per Visit (new visitors) in GB x 0.81 kWh/GB x 0.75] + [Data Transfer per Visit (returning visitors) in GB x 0.81 kWh/GB x 0.25 x 0.02]

#### Emissions per visit in grams CO2e (C):

C = E x 442 g/kWh (or alternative/region-specific carbon factor)

#### Annual energy in kWh (AE):

AE = E x Monthly Visitors x 12

#### Annual emissions in grams CO2e (AC):

AC = C x Monthly Visitors x 12

#### Annual Segment Energy:

Consumer device energy = AE x 0.52

Network energy = AE x 0.14

Data center energy = AE x 0.15

Production energy = AE x 0.19

#### Annual Segment Emissions:

Consumer device emissions = AC x 0.52

Network emissions = AC x 0.14

Data center emission = AC x 0.15

Production emission = AC x 0.19

# Data Analysis & Results

It can start with tables showing the amount of data I’m working with.

Then there will be a structured analysis, on a few major categories and subcategories.

Show:

Mode / Median / average co2 generation

bytes / adjusted bytes proportions

min / max of co2 / bytes

frequency of co2 / bytes in groups

## General Analysis

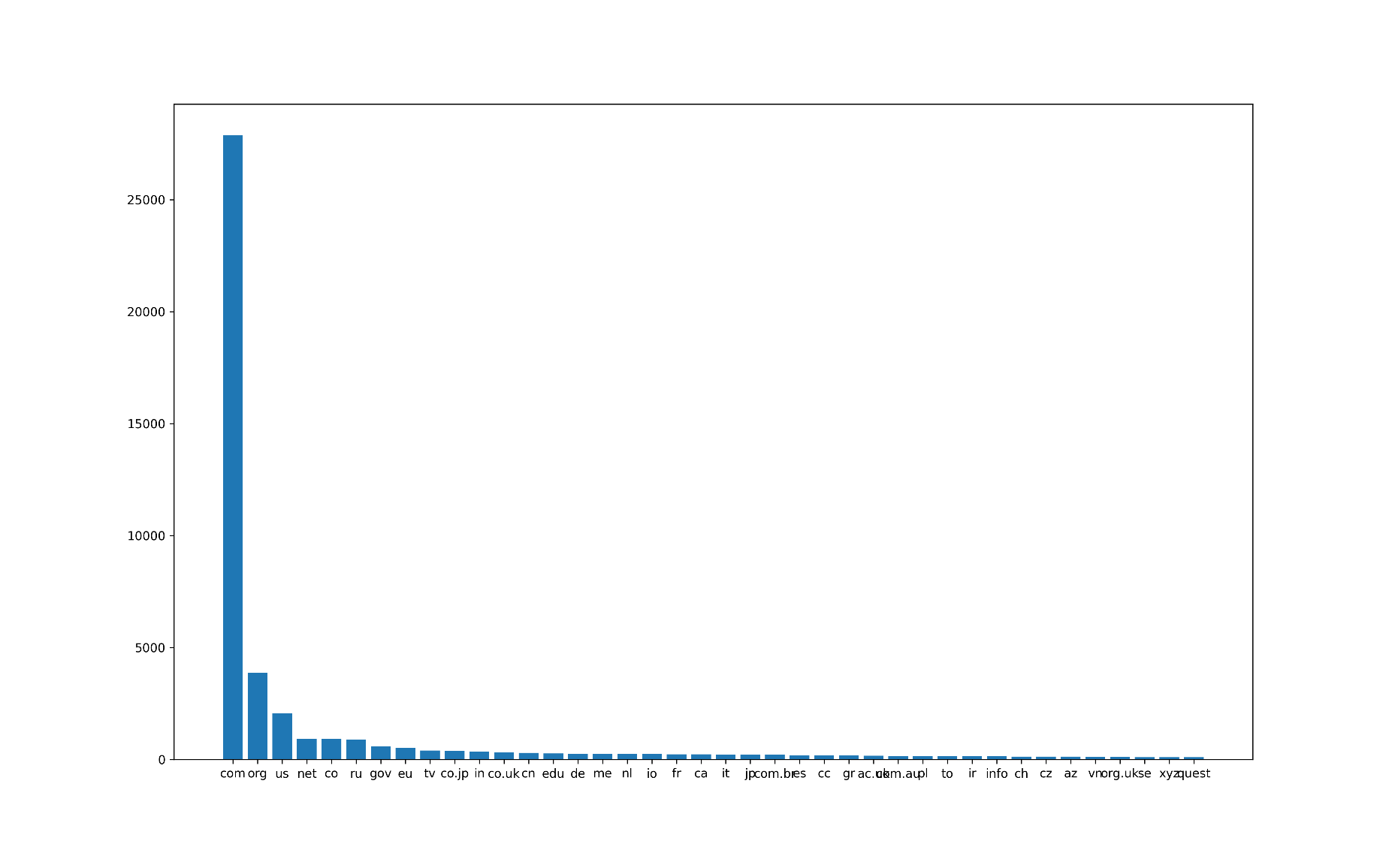
### Overview

#### Dataset Distribution

To provide a sound picture of websites and their CO2 emission impact, more than 65 thousand websites were overall analyzed. All websites were sourced from the original Tranco list and there were no distinctions done to the website’s importance during the data collection process, they were all collected in the original order of popularity.

Top-Level Domains: There are 50,034 usable websites in the final dataset. From those, 27,873 have the “.com” top-level domain, 3868 have “.org “, and 2065 have “.net “ as the domain. Full distribution follows in the table snippet below. Overall, there are 716 different top-level domains [full table appendix]. Keeping only those which occur more than 100 times leaves us with exactly 40 domains. The distribution is heavily skewed towards the first 3 TLD’s which account for 67.5% of the entire dataset.

Figure : TLD Distribution



Website Sizes: A website’s size affects the time it will take to load it and the amount of CO2 it will generate. The collected websites vary greatly in range. Looking at the size on an initial load, the smallest one is 168 bytes and the largest one is <https://www.lematin.ma/> at 304MB. These are only outliers though as the mean size is only 3.69MB and the less affected by outliers median stands at 2.23MB. Overall, all the websites take 184.93GB of space.

#### CO2 Distribution

In total, the sites generate 46554 grams/25893 liters of CO2 and on average, a single site generates 0.562 grams of CO2 per first page load. The minimum amount registered is 0.000042 but the largest one is much bigger, at 76 grams per load. The mean and median numbers here are 0.9304637 grams/0.5175239 liters and 0.5629283 grams/0.3131007 liters.

#### Energy

All sites use up 105 KWG of electricity, the biggest one uses 0.17 KWG and the smallest one 0.00000009568445 KWG. The mean and median figures are 0.002105122 and 0.001273593 respectively. That is a 99 and 99.993% difference when compared to the one using up the most.

#### Green Hosting

Here we see that there are less websites classified as using green hosting than regular grid. The green websites take up 80.8GB of space against 104.06GB for the standard ones and the other three statistics follow a similar pattern: 20 kilograms of CO2 are generated for all energy efficient sites against 26 kilograms for the ones labelled “unknown”.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Green Hosting | Bytes | Statistics: Energy (KWG) | Statistics: Grams of CO2 (Grid) | Statistics: Liters of CO2 (Grid) |
| Unknown | 104068160248 | 59.272054 | 26198.247878 | 14571.46547 |
| True | 80863090299 | 46.055599 | 20356.574757 | 11322.32688 |

## Outliers

One of the first interesting things noticed after the inspection of the dataset in both by histograms and manually was the presence of outliers in the data. Outliers are defined in differing ways in statistical literature. Hawkins (1980 [ref](Hawkins,%20%20D.M.%20%20(1980),%20)) defines an outlier as “an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism”. On the other hand, Grubbs (1969 [ref](Grubbs,%20%20F.E.%20%20(1969),%20)) states them as “An outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs”. The meaning behind the different definitions is that an outlier is a point of data located much farther from the average than most of the dataset. [!]

Outliers can have a negative impact on a quantitative analysis for a variety of reasons, one of them being that they can influence important estimates in a very negative way. In a general sense, there are two major reasons for the existence of errors in a dataset: human error (wrongly inputted data) or technical error (miscalculation by the system).

Chart, bar chart

Description automatically generatedOutliers come in different types. In data mining, they can be global, collective, contextual and in general statistics they can also be univariate and multivariate. Global outliers are those that “all outside the normal range for an entire dataset” (Alghushairy et al., 2020) and univariate outliers are defined by Tabachinck & Fidell as “a case with an extreme value that falls outside the expected population values for a single variable” ([Tabachinck & Fidell 2013](Tabachnick,%20B.%20G.,%20&%20Fidell,%20L.%20S.%20(2013).Using%20multivariatestatistics(6th%20ed.).%20Boston,%20MA:%20Pearson.)). All of the outliers which will be discussed in this section fall in those two categories.

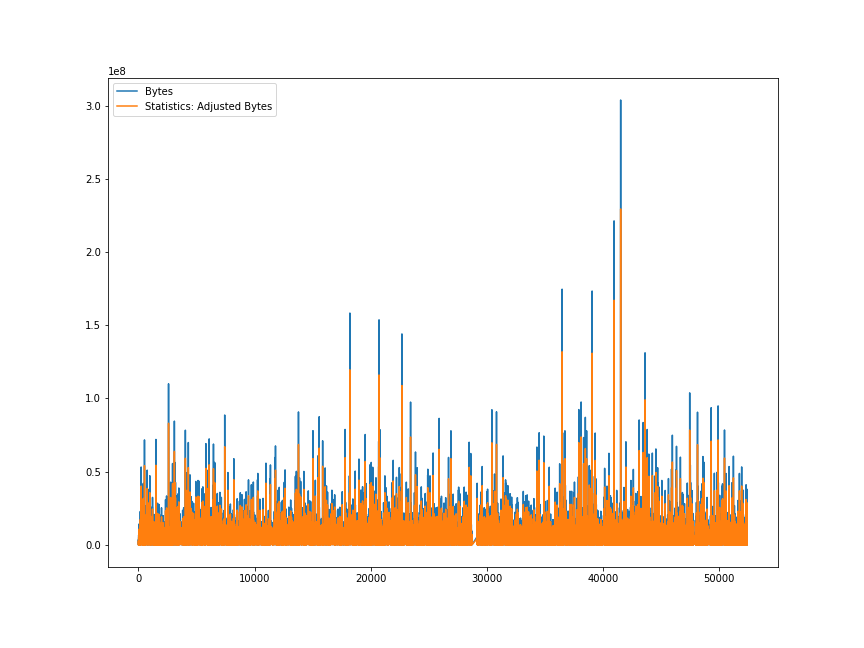
Figure 2: Average, Mean and Standard Deviation for "Bytes"

Understanding the nature of our outliers is important to the nature and validity of the data. What made the outliers occur? Was it human error or a technical one, and what do they say about the websites?

At first look at the dataset, we notice a large difference in website sizes. Looking at the “Bytes” and “Adjusted Bytes” columns we see that although the average website size in “Bytes” stands at 3.69MB, the mean is only 2.23MB, a clear sign of the distribution being off. Looking at the top 10 and bottom 10 values in “Bytes” shows us just how large the difference is.

|  |  |
| --- | --- |
| Top 10 Bytes | Bottom 10 Bytes |
| 304084175 | 168 |
| 221431538 | 170 |
| 174746287 | 185 |
| 173488420 | 200 |
| 158429770 | 207 |
| 153772730 | 209 |
| 144149539 | 209 |
| 131234648 | 218 |
| 110049081 | 227 |
| 103846179 | 230 |

The largest recorded website is 304084175 bytes (304MB) and the smallest only 168 bytes. From the top 10 we see that each website drops off in size from the previous one by anywhere between 5 and more than 80 megabytes and the 10th is “only” 110MB. The bottom ten looks very different with all the websites weighing at less than 1/5th of a megabyte. All of this can also be seen in the following plot, few entries truly stand out from the rest:



So, what is the reason then? After performing tests on a random selection of the outliers a few different patterns are easily observed, but none of them can be called the sole reason for the difference in estimates. In some cases, the reason was simple: the website was offline; thus, the API did not record anything beyond a generic browser response. Others were a blank page, sometimes with a few lines of text, or were not indexable due to the several reasons listed by Website Carbon which have been mentioned earlier.

All of these relate to the bottom ten. For the top ten though, the reason was much more surprising and interesting. It had to do mostly with the page contents not being optimized at all.

Before anything, it bears remembering that some websites are more demanding by nature. Streaming services of course often download videos locally, but news sites on the other hand have a lot of dynamic content which gets updated daily.

The largest website mentioned previously, and one that will serve as a general example is <https://lematin.ma/>, the online version of Le Matin, a Moroccan daily newspaper. This site, tested initially on 15/05/2022 was estimated to be 304MB, as previously stated, but on further testing it returned vastly differing sizes, ranging anywhere from the original estimate, to 11.4MB, as of 10/06/2022. Examining the website with Google Chrome’s Developer ToolA picture containing graphical user interface

Description automatically generateds, caching disabled, no advertisement blocking, and screen resolution forced to 1080p within the browser (to prevent the mobile version from appearing) was done on several different days and it showed that the website is filled with heavily unoptimized images and videos, most of them coming from article previews, automatically scrolling sections and ads. For example, on 27/05/2022 there were two identical ad videos hosted on the website each of them being 84.4MB. On 06/10 again, minutes after the previously mentioned test, the website ballooned from 11.4 to 169 megabytes after an automatically triggered refresh and multiple videos ranging from 5 to 73mb were loaded. Upon further refreshes, none of which were triggered from my side, the size decreased to a “mere” 117MB and then finally at 54MB. Due to the dynamically changing nature of the website, some of the biggest files previously noted were only loaded on some occasions which is the main reason the amount of data transferred differed so much.

Figure : An example of the largest files



Figure : 06/10/2022, 16:53, LeMatin.ma



Figure : 06/10/2022, 16:59, LeMatin.ma

The same essential pattern shows up on many of the large websites in the dataset. Ittefaq.com.bd, a Bangladeshi newspaper exhibited similar size changes, it was initially recorded as being 158MB in both Website Carbon and Chrome Developer Tools, but on 10/06 it only transferred 11.9MB.

The other sites exhibiting those patterns were either sites with heavy graphics like <https://warnerbrosgames.com/> where the heavy content was in the shape of game advertisement videos or adult content streaming services some of which were loading the videos automatically and one particular example even hosted an entire visual novel game on the home page. Going back to the Warner Bros Games example, that website was initially measured at 174mb by Website Carbon. On 18/06 though it stood at 52.1mb as measured by Chrome, with 49.4 of them being all video files. That is 94% of the site’s weight contributed to a few files each of which took seconds to load.

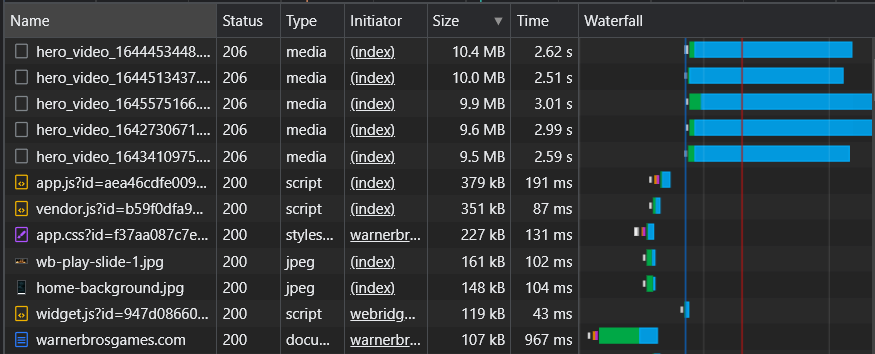


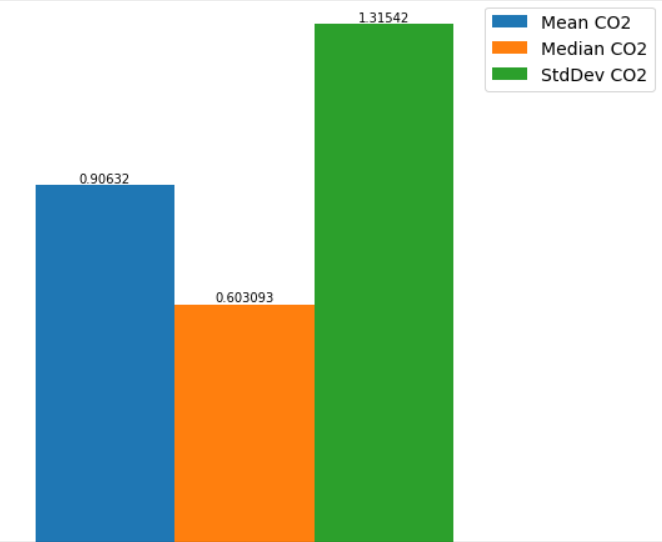
Figure : 18/06/2022, https://warnerbrosgames.com/

These tests were generally performed at random intervals, and for one reason: to determine whether retesting would be needed to verify the correctness of the data. One important fact about the nature of the outliers has been verified by this, the data is not a product of human or technical error, it truly represents a website’s state at the time of testing and gives clear examples to the importance of proper web development done with respect to standards and quality. [change]

## Top and Bottom 1000

Figure 8: Bottom 1000, Mean, Median, St Dev in 'Bytes'

Figure 7: Top 1000, Mean, Median, St Dev in 'Bytes'

Here we observe similar numbers to the entire dataset. The top 1000’s average sizes are only slightly lower than the entire dataset’s and the same goes for the bottom. Overall, the differences are negligible with the top 1000 being only slightly lower (0.09mb) in terms of size than the 3.69 mb originally measured and the bottom is 0.14mb less. Those differences can be explained as small deviations caused by the presence or lack thereof of one of the previously detected outliers in this slice of the data, and that can be further seen in the bigger difference in the median sizes where there is a 0.4mb disparity. As far as the CO2 output goes, we also see a similar picture, as the averages stand at 0.90 and 0.86. These numbers are very similar to the ones reported by [agency, average size from another paper] in 2019 and the small difference with the 2022 data is caused by either natural increases in file sizes year-by-year or by the outliers again.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Top, Bytes | Bottom, Bytes | Top, CO2 | Bottom, CO2 |  |  |  |
| Mean | 3.6002 | 3.43557 | 0.90632 | 0.86487 |  |  |  |
| Median | 2.3957 | 1.99612 | 0.60309 | 0.50250 |  |  |  |
| St. Dev. | 5.2253 | 4.94506 | 1.31542 | 1.22449 |  |  |  |

## Hosting Type

|  |  |
| --- | --- |
| Unknown | 25708 |
| True | 24326 |

As previously noted, we have used two different definitions for the hosting type. “True” signifies that the site is powered by a data center using renewable energy and “unknown” means that the energy type could not be confirmed. It is worth mentioning again that the hosting status is strictly determined by whether the domain is registered with the Green Web Foundation, and it is entirely possible that some of the “unknown” labeled sites could also be renewable. That is only confirmable on a case-by-case basis though and is beyond the scope of this paper. The main takeaway here should be that the “unknown” values might be slightly skewed. In this case though we will take them at face value.

For the entire set we have approximately the same number of sites for both types, 25708 unknowns and 24326 greens, the difference is only 6%. What is interesting though is that the difference is much larger in the other statistics. Referring back to table 1 we notice a much larger 23% difference in website size, energy usage and also CO2 generation. On average, a site using renewables uses 3.32mb of space and those that do not use 4.04mb which means that either there simply are more outliers in the second group or that developers who have built overly large sites are not particularly concerned with whether they are carbon neutral or not.

To verify that, we first split the dataset in two, one half only containing True and the other only Unknown. The distribution between green and regular usage in the top websites is roughly equal, with many often-used websites contained in both the first and second groups. Some examples include Google, YouTube, Facebook, Netflix, and Instagram being hosted renewable and Microsoft, Twitter, LinkedIn, and Wikipedia using regular energy.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Type | >= 100mb | >= 50mb | >= 25mb | >= 15mb | >= 10mb |
| True Bytes | 897840743 | 4082672748 | 9812265588 | 18259742599 | 27146551273 |
| Unknown B. | 777391624 | 4024409192 | 11584901699 | 22420267851 | 35320806777 |
| True Count | 5 | 53 | 222 | 672 | 1407 |
| Unknown C. | 5 | 53 | 285 | 867 | 1935 |
| True, Per site | 179,568,148.6 | 77,031,561.28 | 44,199,394.54 | 27,172,236.01 | 19,293,924.14 |
| Unknown, P.S. | 155,478,324.8 | 75,932,248.90 | 40,648,777.89 | 25,859,593.83 | 18,253,646.91 |

What is interesting though is that there are indeed many more large sites in “unknown”, but on average they are actually slightly smaller in size than the renewable ones. That can be seen by further modifying the two split datasets to be arranged by descending and seeing an overview of the sites in both statistics. If we define a large site to be one over 100mb then we see that True and unknown both contain only 5 sites, and the difference mostly comes from the first two sites which are much larger than the others in size. By filtering down to those larger than 50mb, the picture starts to change, the sizes are almost the same in all rows. But once we filter the sets to only include anything over 25, 15 and 10mb we start seeing some large differences. The 25mb column already has a large difference in size, caused by the larger number of sites in “unknown”, even if those are generally smaller than the trues and this continues into the 15 and 10mb columns too. Although the average size stays slightly lower in weight in “unknown”, there are 28% more of them and that bloats the overall weight.

The takeaway here is that websites using regular energy are quite a bit more likely (28%) to be demanding and to have a larger emission footprint.

## Per Domain

This next section will look at the trends across the different domain types on the internet and will see if there is a difference between them as far as our original statistics go.

### Original Domains

The seven “original top-level domains” were created in the 1980s to cover the needs of the first websites on the internet. They are: .com, .edu, .gov, .int, .mil, .net, and .org. ([ICANN.org](http://archive.icann.org/en/tlds/#:~:text=In%20the%201980s%2C%20seven%20gTLDs%20(.com%2C%20.edu%2C%20.gov%2C%20.int%2C%20.mil%2C%20.net%2C%20and%20.org)%20were%20created.%20Domain%20names%20may%20be%20registered%20in%20three%20of%20these%20(.com%2C%20.net%2C%20and%20.org)%20without%20restriction%3B%20the%20other%20four%20have%20limited%20purposes.)).

|  |  |
| --- | --- |
| TLD | Amount |
| .com | 27884 |
| .org | 3873 |
| .net | 2066 |
| .edu | 921 |
| .gov | 387 |
| .int | 24 |
| .mil | 5 |

Between these seven we have 35161 websites for a total of 130604115266 bytes/130gb. The distribution is shown in the table on the left.

Applying the same type of analysis as before we see that the averages and medians are not too different than the entire dataset as they stand at 3.71/2.25mb.

When comparing the domains directly we see a slightly different picture.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Type | .com | .org | .net | .edu | .gov | .int | .mil |
| MB’s | 103289.11 | 13521.25 | 5951.96 | 6342.36 | 1351.43 | 111.69 | 36.30 |
| Mean MB. | 3.70 | 3.49 | 2.88 | 6.89 | 3.49 | 4.65 | 7.26 |
| Median MB. | 2.26 | 2.13 | 1.44 | 4.84 | 2.40 | 3.29 | 3.63 |
| Energy | 58.83 | 7.70 | 3.38 | 3.61 | 0.77 | 0.06 | 0.02 |
| Mean E. | 0.002110 | 0.001988 | 0.001641 | 0.003922 | 0.001989 | 0.002651 | 0.004135 |
| Median E. | 0.001290 | 0.001216 | 0.000825 | 0.002757 | 0.001366 | 0.001873 | 0.002071 |
| Grams | 26002.13 | 3403.85 | 1498.35 | 1596.63 | 340.21 | 28.12 | 9.14 |
| Mean G. | 0.93 | 0.88 | 0.72 | 1.73 | 0.88 | 1.17 | 1.83 |
| Median G. | 0.57 | 0.54 | 0.36 | 1.22 | 0.60 | 0.83 | 0.92 |

What this tells us is that most TLD’s follow the set averages, with the exceptions being .net, .edu, .int and .mil. Those four are smaller than the entirety of .com but they do contain a significant number of weighty sites.

And when we look at the energy consumption values, we see that

### Regional Domains

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Region | EU + UK | North America | South America | Asia | Middle East | Oceania | Africa |
| MBs | 14441.05 | 2452.13 | 3473.42 | 17369.23 | 2234.84 | 1473.18 | 1072.57 |
| Mean MB | 3.27 | 3.64 | 4.11 | 4.70 | 4.56 | 3.66 | 5.01 |
| Median MB | 2.23 | 2.41 | 2.69 | 2.98 | 2.83 | 2.98 | 1.78 |
| Energy (KWG) | 8.22 | 1.40 | 1.98 | 9.89 | 1.27 | 0.83 | 0.61 |
| Mean E. | 0.001862 | 0.002072 | 0.002338 | 0.002680 | 0.002597 | 0.002087 | 0.002854 |
| Median E. | 0.001268 | 0.001 | 0.001531 | 0.001699 | 0.001612 | 0.001701 | 0.001016 |
| Grams | 3635.40 | 617.30 | 874.40 | 4372.55 | 562.60 | 370.86 | 270.01 |
| Mean G. | 0.82 | 0.92 | 1.03 | 1.18 | 1.15 | 0.92 | 1.26 |
| Median G. | 0.56 | 0.61 | 0.68 | 0.75 | 0.71 | 0.75 | 0.44 |

|  |  |
| --- | --- |
| Region | Amount |
| EU + UK | 4416 |
| North America | 674 |
| South America | 846 |
| Asia | 3138 |
| Middle East | 490 |
| Oceania | 402 |
| Africa | 214 |

# Discussion & Results Analysis

In order to truly understand this data though and see the impact a website has on the environment it is necessary to know how it compares to real world examples. To do so the data will be now compared against measurements sourced from [ ]. Then we will look at a sample from the dataset and determine the number of images and videos in the websites.

First this will be a comparison with CO2 generation by people, cars, etc.

The second part will be an analysis on 1000 downloaded websites and seeing how many of them contain large images.

# Reflection

The process of writing this thesis is described below:

During the writing of the thesis there were several roadblocks encountered which led to limitations in the scope of the final deliverable. They are:

#### Lack of temporal data

The API used for gathering the data only provided measurements for the state of a website at the exact time the test was being established. Access to data older than what was gathered would have allowed for an overview of the progression of website sizes through the years. Ideally, such data would have existed for at least the last couple of years if not more. They would’ve at minimum included website sizes which would have helped calculate a very rough estimate for previous energy usage (ignoring technology efficiency improvements), but the API is too new for this and nobody before has done large scale measurements on a regular basis.

#### Lack of bandwidth

The initial suggestion was for the entire Tranco list to be processed. Unfortunately, that was not possible due to the earlier mentioned limitations with the API. The 2000 possible website measurements per day were too limiting to process the entire dataset as it would have taken at minimum 500 days to collect all of the data, excluding any other possible issues like site outages. From early on a relatively arbitrary amount of 50,000 was decided as it was neither too little to give a decent overview and neither too large to take up too much time to collect.

#### The tools available were not functional

Originally this paper would have included a wider array of measurement tools. Many of them were either not functional, paid or too slow to use in this limited timeframe. Below follows an overview of each and the reason why they were not used:

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

# Conclusion

# References

Alghushairy, O., Alsini, R., Soule, T., & Ma, X. (2020). A Review of Local Outlier Factor Algorithms for Outlier Detection in Big Data Streams. *Big Data and Cognitive Computing*, *5*(1). <https://doi.org/10.3390/bdcc5010001>

Andrae, A. (2020). New perspectives on internet electricity use in 2030. *Engineering and Applied Science Letter*, *3*(2), 19–31. <https://doi.org/10.30538/psrp-easl2020.0038>

Büchel, K. (1996). System Boundaries. *Life Cycle Assessment (LCA) — Quo Vadis?*, 11–25. <https://doi.org/10.1007/978-3-0348-9022-9_2>

Cisco Systems, Inc. (n.d.). *Cisco Annual Internet Report (2018–2023) White Paper*. Retrieved June 19, 2022, from <https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html>

EPA. (n.d.). *Global Greenhouse Gas Emissions Data*. US EPA. Retrieved June 19, 2022, from [https://www.epa.gov/ghgemissions/global-greenhouse-gas-emissions-data#:%7E:text=Global%20carbon%20emissions%20from%20fossil,increase%20from%201970%20to%202011](https://www.epa.gov/ghgemissions/global-greenhouse-gas-emissions-data%23:~:text=Global%20carbon%20emissions%20from%20fossil,increase%20from%201970%20to%202011)

Fabbrizzi, S., Maggino, F., Marinelli, N., Menghini, S., Ricci, C., & Sacchelli, S. (2016). Sustainability and Food: A Text Analysis of the Scientific Literature. *Agriculture and Agricultural Science Procedia*, *8*, 670–679. <https://doi.org/10.1016/j.aaspro.2016.02.077>

Ge, M., & Ross, K. (2019, September 17). *Which Countries Have Long-term Strategies to Reduce Emissions?* World Resources Institute. Retrieved June 19, 2022, from <https://www.wri.org/insights/which-countries-have-long-term-strategies-reduce-emissions>

IEA. (2014b, July). *More Data, Less Energy*. <https://www.iea.org/reports/more-data-less-energy>

IEA. (2021, November). *Data Centers and Data Transmission Networks*. <https://www.iea.org/reports/data-centres-and-data-transmission-networks/>

Krisetya, M., Lairson, L., & Mauldin, A. (n.d.). *Global Internet Map 2021* [Graph]. Global Internet Map 2021. <https://global-internet-map-2021.telegeography.com/>

*Learn About Sustainability*. (2021, December 2). US EPA. Retrieved June 19, 2022, from <https://www.epa.gov/sustainability/learn-about-sustainability>

Leiserowitz, A. (2019, February 15). *Climate change in the American mind: December 2018*. Yale Program on Climate Change Communication. Retrieved June 19, 2022, from [https://climatecommunication.yale.edu/publications/climate-change-in-the-american-mind-december-2018/](https://climatecommunication.yale.edu/publications/climate-change-in-the-american-mind-december-2018/%20)

le Pochat, V., van Goethem, T., Tajalizadehkhoob, S., Korczynski, M., & Joosen, W. (2019). Tranco: A Research-Oriented Top Sites Ranking Hardened Against Manipulation. *Proceedings 2019 Network and Distributed System Security Symposium*. <https://doi.org/10.14722/ndss.2019.23386>

Lindsey, R. (2020, August 14). *Climate Change: Atmospheric Carbon Dioxide | NOAA Climate.gov*. Climate.Gov. Retrieved June 19, 2022, from [https://www.climate.gov/news-features/understanding-climate/climate-change-atmospheric-carbon-dioxide#:%7E:text=Carbon%20dioxide%20concentrations%20are%20rising,people%20are%20burning%20for%20energy](https://www.climate.gov/news-features/understanding-climate/climate-change-atmospheric-carbon-dioxide%23:~:text=Carbon%20dioxide%20concentrations%20are%20rising,people%20are%20burning%20for%20energy)

Masanet, E., Shehabi, A., Lei, N., Smith, S., & Koomey, J. (2020). Recalibrating global data center energy-use estimates. *Science*, *367*(6481), 984–986. <https://doi.org/10.1126/science.aba3758>

nationalgridESO. (n.d.). *What is carbon intensity? | National Grid ESO*. Nationalgrideso.Com. Retrieved June 20, 2022, from <https://www.nationalgrideso.com/future-energy/net-zero-explained/what-carbon-intensity>

Rehman, A., Ma, H., Ozturk, I., Murshed, M., & Dagar, V. (2021). The dynamic impacts of CO2 emissions from different sources on Pakistan’s economic progress: a roadmap to sustainable development. *Environment, Development and Sustainability*, *23*(12), 17857–17880. <https://doi.org/10.1007/s10668-021-01418-9>

Statista, & Johnson, J. (2022, May). *Share of users worldwide accessing the internet in 4th quarter 2021, by device*. Statista. <https://www.statista.com/statistics/1289755/internet-access-by-device-worldwide/>

Telefonaktiebolaget LM Ericsson. (2020, February). *A quick guide to your digital carbon footprint*. <https://www.ericsson.com/4907a4/assets/local/reports-papers/consumerlab/reports/2020/ericsson-true-or-false-report-screen.pdf>

Telefonaktiebolaget LM Ericsson. (2021, November). *Ericsson Mobility Report*. <https://www.ericsson.com/4ad7e9/assets/local/reports-papers/mobility-report/documents/2021/ericsson-mobility-report-november-2021.pdf/>

The Shift Project. (2019, March). *-LEAN ICT- TOWARDS DIGITAL SOBRIETY*. <https://theshiftproject.org/wp-content/uploads/2019/03/Lean-ICT-Report_The-Shift-Project_2019.pdf/>

Thiagarajan, N., Aggarwal, G., Nicoara, A., Boneh, D., & Singh, J. P. (2012). Who killed my battery?: analyzing mobile browser energy consumption. *WWW ’12: Proceedings of the 21st International Conference on World Wide Web*, 41–50. <https://doi.org/10.1145/2187836.2187843>

United Nations. (2019, July). *Digital Economy Report 2019*. <https://unctad.org/system/files/official-document/der2019_en.pdf/>

World Meteorological Organization. (2021). *State of the Global Climate 2020*. <https://library.wmo.int/doc_num.php?explnum_id=10618>

Zhu, Y., & Reddi, V. J. (2013). High-performance and energy-efficient mobile web browsing on big/little systems. *2013 IEEE 19th International Symposium on High Performance Computer Architecture (HPCA)*. <https://doi.org/10.1109/hpca.2013.6522303>

Adams, C. et al. (2022) Calculating Digital Emissions, In *Sustainable Web Design,* (Accessed: April, 12 20221) <https://sustainablewebdesign.org/calculating-digital-emissions/>

# Appendices

[notes]

Mention any issues, topics, and things to look into for future research