**Assignment 1**

*This assignment accounts for 15% of the grade. There are two options. You can choose either one. The first option is completing a simple, practical Internet measurement exercise where you will analyze a dataset to answer specific questions. The second option is to summarize and highlight the salient points from an Internet measurement research article.*

**Option1**

We discussed the workload characterization of the WWW 2007 Web server in class. We discussed the usage behaviour, popularity characteristics, client errors, and geographic distributions of users, among other things. The Web site for the WWW2007 conference (http://www2007.org) is hosted at the University of Calgary. The site remains online to this date. Your task is to analyze the content hosted on the Web site. This exercise will help you attain skills in data analytics, statistical analysis, graphing, and interpretation of networking data. This exercise is courtesy of Prof. Carey Williamson from the University of Calgary.

The file www2007data.txt contains the output of the Unix command "ls -lR" in the home directory of the WWW2007 Web site (/home/projects/www2007). The output shows information such as the name of each file and directory, the file permissions, the file size, the file modification date, and so on.

Answer the following questions:

1. How many different regular files (not directories) are stored on the site? What is the aggregate size of these files (in bytes)?
2. What is the largest file on the site? How big is it? How many empty files (0 bytes) are there? What is the smallest non-empty file on the site? How big is it?
3. What is the mean file size on the site? What is the standard deviation of file size? What is the median file size (50-th percentile value)? What is the mode (most frequently occurring value) of the file size distribution?
4. Plot a graph showing the file size distribution. Make one graph for the empirical probability density function (pdf), and a separate one for the cumulative distribution function (CDF). Use a graph style (e.g., lines, boxes, histogram, scatter plot) and axis scaling (e.g., linear, logarithmic, log-linear, log-log) to convey the distribution effectively. Comment on your observations.
5. Analyze the file type distribution: File types can be determined heuristically based on the (optional) suffix in the file name (e.g., foo.html, paper127.pdf, painful.doc). Produce a table showing the site’s top 10 known file types, in sorted order from most prevalent to least prevalent. Within this table, show the number of files of each type, the percentage of files of each type, the number of bytes for each file type, and the percentage of bytes for each file type. If necessary, use a category "Unknown" for any file types that are not easily discernible from the file name suffix. In the table, add a category "Other" for those files not accounted for among the top 10 file types so that the percentages in the table sum correctly to 100%. Comment on your observations.
6. Plot a graph showing the file size distribution for the PDF versions of the papers and posters in the conference proceedings (i.e., from the subdirectories ./papers and ./posters). Plot a CDF graph with two lines (one for papers, one for posters). Use a graph style and axis scaling to convey the distributions effectively. Comment on your observations.
7. Calculate (or estimate) the age of each file on the Web site (i.e., the number of days since it was last modified). What is the oldest file on the Web site? How old is it? What is the newest file on the Web site? How old is it? What are the mean, median, and mode for the file age distribution?
8. Plot a CDF graph showing the file age distribution. Use a graph style and axis scaling to convey the distribution effectively. Comment on your observations.

**Option 2**

The work-from-home phenomenon introduced as a result of the pandemic has increased the use of video conferencing and collaboration tools such as Zoom, MS Team, Skype, etc. Several papers have analyzed the impact of this shift in work patterns on Internet traffic. Two such articles published in a top-rated Internet measurements conference called PAM 2022 have reported results on the use of Zoom in a university network environment. Choose one paper and write a 1–2-page commentary. Note this is not a review of the paper, which would require critical analysis and a good understanding of the literature to determine the merits and drawbacks of the work. The commentary should be organized into the following sections:

Introduction: Briefly summarize the paper, the problem, the context, and the motivation

Methodology: Briefly summarize the methodology used to conduct the research. You are encouraged to use visual methods to describe the methodology. When you use visual methods, you should still use text to describe the figure or diagram.

Results Highlights: List 3 key results that you found fascinating. Justify why you selected those results and what their implications are.

Contributions: Briefly describe how the paper advances knowledge in the field of networking and measurements. It would be best if you justified whether it reports new results or employs a unique methodology.

Takeaways: Taking a high-level perspective, briefly describe two crucial points that you should remember from this paper. These points should be lasting impressions of the work.

Future Work: Briefly describe how you will extend this work and what challenges/opportunities you may face.

This option requires good writing skills and synthesizing ideas from the paper in your own words. Do not simply take sentences from the article. Instead, read the paper, take notes, and understand the crucial aspects of the work.

Paper 1: Zoom Session Quality: A Network-Level View (PDF, 680 KB) by Albert Choi et al. Proceedings of Passive and Active Measurement (PAM) Conference, virtual event, March 2022. (Springer LNCS, Vol. 13210, pp. 555-572, March 2022)

Paper 2: Zoomiversity: A Case Study of Pandemic Effects on Post-Secondary Teaching and Learning (PDF, 1.5 MB) by Mehdi Karamollahi et al. Proceedings of Passive and Active Measurement (PAM) Conference, virtual event, March 2022. (Springer LNCS, Vol. 13210, pp. 573-599, March 2022)