Preliminary Project Total = 20pts.

Answer the following questions for your final project.

1. Identify your project. Here state what do you want to do. [5pts]

2. Describe your project and why it is important. [5pts]

3. Three citations that are relevant to your project and set a timetable for tentative work: [3pts + 2pts]

It could be research papers/websites/books etc.

Professional write-up of the preliminary report. [5pts(Answered all questions with details), 2pts(Needs improvement or missed a answering a question), 0 (Inadequate)]

Title?: "Graph-based Analysis and Prediction of Malware Families using Divergence Measures and Graph Neural Networks."

In this project I will conduct the analysis and prediction of malware families, leveraging graph theory and machine learning. The main steps of the project are to use trace data to construct a similarity matrix using Kullback-Leibler Divergence (KLD) and Jensen-Shannon Divergence (JSD), and build a Minimum Spanning Tree (MST) from this matrix. We then apply community detection algorithms (Girvan Newman and Louvain) to identify potential malware families within the MST. We validate these communities against known malware families through a confusion matrix. Finally, we use graph neural networks to predict the family of new, unknown malware samples. This project is important because it can provide an effective way of classify and predicting malware families. This could be of great use in the cybersecurity field as a defensive measure for malware mitigation.

**Tentative Timeline:**

* Week 1-2: rewrite code from steps 1-3 of project outline.
* Week 2-4: steps 4-5 of project outline
* Week 4-7: Step 6 project outline

Project steps.

1. Get data and read data to construct the graphs of each trace data when compared with another trace data.
2. Construct a matrix from the trace data from step 1 with KLD (Kulbach leiber) values and JSD Jensen Shannon Divergence.
3. Run Minimum Spanning Tree (Prims or Kruskal) algorithm on the matrix that you computed. (Last Project Covered Steps 1-3)
4. Run Community detection- Girvan Newman algorithm and Louvain algorithm (Python packages) on MST.
5. Use the output of Step 4 and create confusion matrix of the families. (For each family of malware from the original data- compare how the constructed communities which we call as constructed families)
6. Pick the families and use graph neural nets to test the prediction.

Papers  
- <https://link.springer.com/chapter/10.1007/978-3-030-24907-6_35>

* <https://www.mdpi.com/2076-3417/12/21/10837>
* <https://ar5iv.labs.arxiv.org/html/0906.0612>