Data Mining Project Proposal

Group- Jaspinder Singh(3770406), Nomaan Imran Saiyed(3769353), Syed Owais Haider Kazmi(3768888)

Working title: Mining and Explaining Darknet Traffic Patterns from Network Flow Data

1) Dataset + readiness date + size

Dataset: CIC-Darknet2020 (network flow dataset for classifying darknet-related traffic).

Source: Canadian Institute for Cybersecurity (UNB) dataset page.

What the dataset contains:
Each row is a network flow with many numeric traffic features (packet counts/length stats, timing/inter-arrival stats, active/idle time etc.) typically produced by CICFlowMeter-style flow feature extraction.
The CIC description notes darknet traffic categories like browsing, chat, streaming, email, P2P, transfer, VOIP and that the dataset combines Tor/VPN traffic sources into these categories.

Our local file: We already have the dataset as Darknet.CSV. In our copy we have 141,530 flows and 85 columns, including two label columns:

- **Traffic type label** (Tor / VPN / NonVPN / Non-Tor)
- **Application/behavior label** (e.g :- browsing, chat, streaming, P2P)
  This size is sufficient for effective mining, including clustering + multi-class classification and model comparison.

2) What information/insight we want to mine + proposed use

We want to mine patterns that distinguish:

1. **Traffic type** (Tor/VPN vs normal traffic)
2. **Application behavior** (browsing, chat, streaming, P2P etc.)

Planned mining approach:

- **Unsupervised discovery:** Use clustering (K-means + DBSCAN) to see whether flows form natural groups based on traffic behavior, and whether these clusters align with the labels.
- **Supervised detection:** Train multi-class models to predict traffic type and application type, compare models, and evaluate performance carefully under class imbalance.
- **Explainability/insight extraction:** Identify which traffic features most strongly separate Tor/VPN from normal traffic and which separate application behaviors (feature importance / permutation importance).

Proposed use of the insight:
This produces an explainable, reproducible analysis of whether privacy enhancing traffic shows detectable statistical patterns at the flow level (useful as a baseline for network monitoring, traffic characterization, and anomaly style reasoning in cybersecurity contexts).

3) Why it’s useful/interesting + what we hope to learn

Why interesting:

- It’s a real-world style problem: you’re classifying traffic that is often encrypted, using only flow statistics.
- The dataset supports both pattern mining (clustering) and prediction (classification), which matches course topics well.

What we hope to learn:

- Do Tor/VPN flows form distinct clusters from non-Tor/NonVPN flows without using labels?
- Which models perform best for multi-class classification and why (e.g:- Random Forest vs SVM vs KNN)?
- Which features consistently matter (timing stats, packet length stats, active/idle behavior, etc.) and what that suggests about traffic characteristics.

4) Preliminary timeline (built around March 10 interim report)

Date	What’s happening	Milestone
Tue Feb 17	Project proposal due	Dataset chosen + repo link + proposal submitted
Feb 18 – Feb 24	Work period	Data cleaned/usable + basic EDA started
Thu Feb 26 (Lab 5)	Lab checkpoint	Preprocessing + EDA + baseline model started in GitHub
Mar 3 – Mar 9 (Reading Week)	Work period	Baseline classification results + clustering (PCA + K-means/DBSCAN)
Tue Mar 10	Interim report due	Data summary + progress + results + next steps + member contributions
Thu Mar 12	Project Review in Lab	Show interim results + questions + updated plan
Thu Mar 19	Project Review in Lab	Show improved models/insights since interim
Tue Apr 7 / Thu Apr 9	Project presentations	Slides + final story ready (all members present)
After Apr 9 (final due TBA)	Final report	Final report + repo polished