**BI Capstone Project**

**Mapping the Internet**

**Problem Definition**

Link - <https://www.kaggle.com/c/facebook-ii>

For this project we will be exploring the map of the entire internet. Unlike the map of a city, where best routes are relatively fixed except for the occasional construction or parade detour, the paths that information travels over the web are constantly changing. There is no centralized system of stop-lights or traffic cops. Instead, there are tens of thousands of autonomous systems using a common protocol to advertise the next available hops, updated depending on service-agreements, capacity, and load. This will be a test of both our engineering know-how and ability to statistically learn on complex, dynamic graph structures.

The Task: We have data that gives a path which, at one point in the training time period, *was* an optimal path from node A to B. The question is then to make a probabilistic prediction, for each of the 5 test graphs, whether the given path is still an optimal path. This is a much more difficult task than link prediction alone. The global structure of the graph may affect many optimal routes, paths can have varying lengths, and there may be multiple optimal routes for a given source & destination.



**BI Use Case**

The topology of internet changes a lot everyday because of various factors like service agreements, capacity and load. Using the probabilistic prediction from the models that we work on, we can predict the usage of links in the near future given the history. This will help the ISP's or the link providers actually predict the usage of links which may be useful in many ways. If the link is predicted to be used extensively than they can allocate more resources.

**Data Sets**

Link:<https://www.kaggle.com/c/facebook-ii/data>

There are 15 graphs in the training set and 5 in the test set. These are directed, dynamic, real-world Internet topology graphs with nodes that correspond to [Autonomous Systems](http://en.wikipedia.org/wiki/Autonomous_System_(Internet)) (AS). Each graph is sampled at a fixed time interval, delta, apart.

Edges may appear and disappear as routing patterns change over time. Some nodes are peers (A->B means A can exchange traffic through B for free) and some are connected by regular edges (A->B means A pays B to route their traffic). Peers have weights of zero in the training data and regular edges have weights of one.

For the prediction task, we are given a path which, at one point in the training time period, was an optimal path from node A to B. Here, “optimal” accounts for the peer relationships with zero-weight edges (i.e. it is the weighted shortest path). Many paths can have the same best optimal weight and are all considered optimal, regardless of the path length.

**Demo**

We are planning to give a demo about how different methods stand in solving the problem at hand. Simple comparison of how the methods perform in finding the optimal paths. It may change over the course of the project.

**Method/Algorithms**

We are planning to implement variants of logistic regression and/or Random walk forest for getting the probabilistic predictions. We are planning to implement at least 3 for now. Any other suggestions are welcome.

**Who will Contribute to What**

Data cleaning: Alok and Dhyey

Implementation of methods: Karthik, Dhara and Shubham

Getting the results, report and ppt: Ankit and Aditya

This is a rough estimate and the roles may change depending on the needs, once we dive deep.