**Constructing a Tracking Portfolio Using PCA and Graph Techniques**

### Supervisor: Chris Romano

### Team Members: Group members: Yitian Duan, Shengbo Lang, Yan Sang, Daoming Zhang

Tracking portfolio construction has always been a fundamental topic for exchange-traded fund (ETF) issuers. The objective is to track a portfolio's performance using only a subset of securities, with a desirable small tracking error, defined as the standard deviation of returns' difference. This is important for ETF issuers in order to provide market participants such as market makers with necessary information to hedge their risks correspondingly, meanwhile avoid disclosing too many details of their actual holdings.

In this work, we aim to track the stock market index Standard and Poor's 500 (S&P 500), which is one of the most commonly followed equity indices, indicating the stock performance of 500 large listed US companies.

We have approached the problem via [1] principal component analysis (PCA) on S&P 500 stocks, in order to generate mutual relationship information between any two stocks; [2] graph filtering methods, including planar maximally filtered graph (PMFG), minimal spanning tree (MST), and winner-take-all combined with spread sampling (SS), to construct a filtered graph with stocks being nodes for further analysis; [3] appropriate centrality measures and selection methods to rank and select stocks to construct the tracking portfolio; [4] different weighting strategies including equal weights, low volatility weights, Markowitz weights, price-adjusted weights and return-adjusted weights. The iShares S&P 100 ETF (OEF) is selected to be our benchmark model, as a reference to large-capitalization US equities.

We obtain current S&P 500 components and the components update history from Wikipedia, use Yahoo Finance to retrieve historical daily data, and generate numerical features by a Python package ‘Financial Technical Analysis’ (FinTa) to conduct PCA on a daily basis given a lookback period. Following that, the difference between normalized PCA coefficients of any two stocks is taken as a measure of similarity. Up to this point, we have developed a relevance measure to replace correlation, which is our key innovation. PCA window, length of the lookback period, is experimented to be 3 months. Using such PCA-derived information, we apply different graph filtering techniques to construct an incomplete graph, where only ‘strong relevance’ edges are left. Followed by centrality sort and selection (most central stocks, most peripheral stocks, random selected stocks or uniformly selected stocks), we can get 10, 30, 50 and 100 stocks used to construct the tracking portfolio. Such PCA, graph construction and stock selection are done every 1 month (at the beginning of each month). Weights are assigned to them by ways mentioned above, and the holding period is also 1 month to avoid excessive transaction costs.

Resultwise, we measure performance by annualized tracking error on portfolio returns against S&P 500 index returns. The key challenge is from the long running time and too many parameters, making it difficult to traverse all combinations. Thus, we selectively do local tests to select through parameters. Another justification for our approach is that annualized tracking errors are at similar levels when changing some conditions.

PCA window length (3/6/12 months) and graph filtering methods (PMFG, MST) yield highly similar results when having other parameters controlled to be constant. Most central chosen stocks from centrality rank gives better results than peripheral ones or uniformly ones, and significantly better than randomly chosen stocks, which proves our graph approach's functionality in constructing a tracking portfolio. Equal or low volatility weights generate small and stable results. Based on such choices of parameters, the long run performance gives less than 1% annualized tracking error.