Big Data: Measuring Information Diffusion about Stocks on the Web

Researchers/Advisors:

Johan Walden
University of California at Berkeley
Associate Professor of Finance

Deniz Yavuz
Purdue University
Assistant Professor of Finance

Background:

The proposed project is part of an ongoing research initiative that explores the relationship between stock returns, diffusion of information, and investor networks in the market, and which has received attention both in academia and among practitioners (e.g., in CFA Digest).

How is information incorporated into financial markets and asset prices? Network theory offers a novel approach to this important question. In a market where investors share private information through a network, specific predictions arise about the distribution of trading profits, the correlation of trades and the trading volume.

Theoretical models were developed in Ozsoylev and Walden (2011) and Ozsoylev et al (2014). Specifically, the models introduce slow information diffusion through an *information network*. In the models, some information is neither completely public, nor completely private. The network provides an additional channel through which information is incorporated into asset prices. This view, that there is "semipublic" channel through which information reaches asset prices in addition to the traditional channels, is quite different that the standard approach to asset pricing with heterogeneous information, since the standard approach assumes completely private signals. Anecdotal evidence suggests that the semipublic channel is important.

The theory is consistent with several stylized facts in real-world markets:

1. Most large price movements are not related to public signals.
2. Price volatility is highly time-varying.
3. Trading volume is high and heterogeneous – in a predictable way.
4. Portfolio investments are heterogeneous – in a predictable way

The theory is explored in Ozsoylev et al. (2014), who use account level data of trades on the Istanbul Stock Exchange to infer the market's information network.

The current project takes a different approach, by directly measuring information diffusion among market participants in Internet discussion boards. Specifically, by crawling these webpages for relevant discussions, one can measure whether information tends to spread from one forum to another in a sequential fashion, and how this process relates to the dynamics of stock returns. This is an exciting new area of research, which potentially can deliver academically important insights about what determines stock valuations, as well as insights that are relevant to practitioners about whether such data can be used to generate abnormal returns.

A student at Berkeley, Jun-Mok Kim, has carried out an initial analysis, writing a web-crawler that collected data for 20 stocks on 3 discussion boards (Yahoo, Investorhub, Lion), providing a proof of concept of the approach. Specifically, Jun-Mok collected data about how many times specific words (from a pre-specified list of about 500 words) were mentioned within a specific time period (measured hourly) for a specific stock in a specific discussion board, over a 9-month period. He then used the data to infer whether discussions in some boards of specific stocks seemed to precede those in the others.

The project builds upon Jun-Mok's initial analysis.

Project description:

Jun-Mok's code and data are available as a starting point for a more extensive analysis. The project has three phases:

1. Extending the analysis to a larger stock universe, a longer time period, and more discussion boards. This will require efficient implementation of algorithms to handle big datasets, a couple of order of magnitudes larger than what the current code does.
2. Implementing and running sophisticated tools to analyze the data. So far, the analysis has basically been restricted to word counts, where information diffusion from webpage A to B is assumed have occurred if there is a sufficiently high match between word scores on webpage A at time *t* with the word scores of webpage B at time *t+1*. There are, however, much more sophisticated, linear algebra based, algorithms for quantitative/computerized text analysis. Designing, implementing and running such an algorithm, to create a good measure of information diffusion in specific stocks, constitutes the second phase of the project.
3. The third phase of the project is to relate the measure(s) developed in phase 2 to actual stock performance. For example, the novel measures may be related to existing factors (size, value and momentum, etc.) that have been known to explain stock returns.

Ideally, the bulk of the project will be spent on phase 2, and to some extent on phase 3. However, the exact time division is uncertain and will depend on how the project evolves, as is usually the case in real-world situations.

Deliverables:

1. The program and collected data
2. The data analysis and results
3. A report that summarizes the results

The specific details of later parts of the project will depend on the initial results. Students will interact with the Professors (who are in Berkeley, CA, and Lafayette, IN) on a regular basis. Johan may also visit Uppsala for a short period during the project.

Prerequisites:

- Experience of programming in an appropriate language, e.g., Java
- A good understanding of numerical methods in linear algebra (e.g., eigenvector decomposition/spectral analysis of matrices)
- Some experience with network/graph theory is a plus
- Some knowledge of finance is also a plus

Contact:

Students who are interested and/or have questions can contact:

Johan Walden
University of California at Berkeley
e-mail: walden@haas.berkeley.edu

Please provide some information about yourself in the e-mail (e.g., about relevant courses and experience).

Further Reading:

Students can get further information from Ozsoylev et al. (2014) "Investor Networks in the Stock Market", downloadable at: http://faculty.haas.berkeley.edu/walden/HaasWebpage/ein.pdf