The Social Internetwork

Abstract

We develop a comprehensive dynamic network of the economy that captures the interconnections of all firms. To do so, we utilize a rich dataset of over 50 million financial tweets. Initially, we use the Twitter dataset to extract the connections between every pair of firms, and subsequently, scale this process to map out all the interconnections among all firms in the economy. We then identify firms that are most ***central*** to the economy as well as those that are ***peripheral***. Finally, we study the impact of economic shocks applied at various points in the network and track how such shocks propagate. The computational complexity of generating these networks is high and requires the computational resources of XSEDE.

Objective and Project Overview:

The goal of this project is to construct a novel network of all firms in the economy based on social media discussions. We utilize a large dataset of 50 million financial tweets, originating from millions of Twitter users to first identify peer firms for each firm in the economy, and subsequently aggregate these connections to construct a network of the whole economy. This network representation of the economy can be viewed as a map of all firms and all their interconnections – including the strength of connections. This network view will also identify the formation of clusters – or economically linked firms – on the network. This is a novel approach to both identifying peer firms, as well as constructing a network of the economy based entirely on social media discussions. We expect this network to be more current, relevant, and timely than existing methods of classifying firms in that it uses the *wisdom of the crowd* revealed through social media discussions. Our network will be dynamic and time-varying in that it will constantly evolve as new information and connections are revealed via social media. We propose the term "*social internetwork"* to describe the network of the economy implied by the financial tweeting activities of all Twitter users. After constructing the *social internetwork*, we will use network theory to identify firms that are ***central*** as well are those that are ***peripheral***. We then examine the effect of economic shocks to central and peripheral firms with emphasis on the "ripple effect" of how shocks propagate within the economy.

One of our goals is to advance the literature that identifies peer firms. Particularly, instead of relying on static definitions of identifying peers, we propose the novel approach of relying on social media to discover peer or related firms. Our preliminary tests show that ties between firms – as identified by millions of Twitter users – describe deep, non-obvious links. Such links are not explained by industry or supply-chain ties. For example, Tesla Inc. is most closely connected to Apple, Netflix, Amazon, and Facebook! This list of peers would not be identified using standard classification methods. Most importantly, our tests show that firms' returns exhibit high correlations with their socially-connected peers – more so than they do with firms from their own industry. Figure 1 below shows Tesla Inc's closest peers:

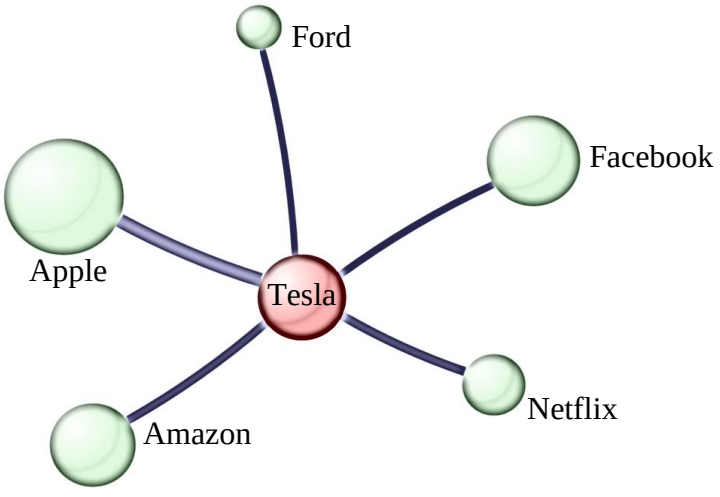


Figure 1: Tesla's closest peers based on the *social internetwork* connections.

Computing Languages and Constructing the *Social Internetwork*

The initial phase of this project utilizes the Python programming language to collect the corpus of tweets using the public Twitter API. We also use Python to process the social internetwork once generated. The tweeting dataset is housed on a MySQL Server database. The *Social Internetwork* is built by examining all tweets about each firm, and textually identifying the presence of all other firms within the text of the tweets; thus, the problem is of exponential complexity – a situation that is exacerbated by the sheer size of the dataset. In preliminary tests, it took approximately 2 weeks to generate one network using standard computing resources. Further, we plan to investigate the idea of *degrees of separation*, in which we study indirect links between firms – which is a much more demanding computational problem. Having access to the XSEDE resources will allow us to handle the computational complexity effectively.

Empirical Methodology

To study the flow of information and noise throughout the economy, we will use a version of the Tookes (2008) model to test the relationship between trading and returns in central firms, and trading and returns of firms related to central firms. We extend the vector autoregression (VAR) approach of Hasbrouck (1991) to include information arrivals as follows:

$$r_t = \sum_{i=1}^{I} \alpha_i \, r_{t-i} + \sum_{i=0}^{I} \beta_i \, x_{t-i} + \sum_{i=0}^{I} \delta_i \, information_{t-i} + \mu_t$$

Where r_t is the return and t is a time index.The variable x_t represents trades. The variable $information_t$ represents information arrivals. For ease of exposition we suppress the equations for trading and information above and in the following.

The VAR provides a compact framework to understand the relationship between price changes, trades, and information arrivals. The VAR can be further extended to include two-related firms as in Chan, Chung, and Fong (2002) and to include related firms and information events (earnings announcements) as in Tookes (2008).

$$r_t^c = \sum_{i=1}^{I} \alpha_i \, r_{t-i}^c + \sum_{i=0}^{I} \beta_i \, x_{t-i}^c + \sum_{i=0}^{I} \delta_i \, information_{t-i}^c$$
$$+ \sum_{i=0}^{I} \alpha_i \, r_{t-i}^r + \sum_{i=0}^{I} \beta_i \, x_{t-i}^r + \sum_{i=0}^{I} \delta_i \, information_{t-i}^r + \mu_t$$

Where c represents the central firm in the network and r represents the related firm in the network. The system can be made more general and incorporate *n* related firms. The methodology is flexible and allows for bidirectional information flow from central to related firms. Central firms may also be related firms in other clusters (or smaller networks). The methodology can be generalized to the whole economy as:

$$R_t = \sum_{i=1}^{I} \alpha_i \, R_{t-i} + \sum_{i=0}^{I} \beta_i \, X_{t-i} + \sum_{i=0}^{I} \delta_i \, Information_{t-i} + \mu_t$$

Where R, X, and Information are N * 1 vectors of returns, trading, and information arrivals. At the limit we can test how the information contained in returns, trading, and information releases flows through the entire economy. The methodology allows for firm specific, industry specific, and economy-wide shocks to impact all firms.