Tweets and peers: defining industry groups and strategic peers based on investor perceptions of stocks on Twitter

Timm O. Sprenger* and Isabell M. Welpe†

*Technische Universität München, TUM School of Management, 80804 Munich, Germany, phone: +49-(0)89-289-24820, fax: +49-(0)89-289-24805, E-mail: sprenger@tum.de
†Technische Universität München, TUM School of Management, 80804 Munich, Germany, phone: +49-(0)89-289-24800, fax: +49-(0)89-289-24805, E-mail: welpe@tum.de

Abstract. Delineating industry groups of related firms and identifying strategic peers is important for both financial practitioners and scholars. Our study explores whether the degree to which pairs of companies are associated with each other in an online stock forum is related to the comovement of their stocks. We find that our news-based measure of relatedness can explain stock returns with the same power as the established SIC classification scheme. We investigate, whether our method can serve to define strategic peer groups and conclude that news-based relatedness can help delineate meaningful industry groups.

Keywords: Twitter; microblogging; stock market; industry classification; strategic peer group; comovement.

1. Introduction

“Our ability to define industries is more art than science.”

Dranove et al. (1998)

Delineating industry groups of related firms and identifying strategic peers is important for both financial practitioners and scholars. Among financial professionals, industry-related stock market indices, such as Dow Jones’ Industrial Average or Internet Composite Index, are the most important indicators of market developments. Investment managers tend to take industry affiliations into account when they structure portfolios and corporate managers often need to identify the most relevant competitors for peer-group comparisons (Chan et al., 2007). In scholarly publications, industry classifications are used by hundreds of scientific studies and serve 4 primary purposes (Kahle and Walkling, 1996): first, to identify control (matched) firms (48% of studies using industry classifications), second, to describe the industrial composition of the sample (35%), third, to restrict the sample of interest (32%), and fourth, to categorize acquisitions and divestitures as conglomerate or nonconglomerate (9%).

However, identifying related companies and defining industry groups can be challenging. Fan and Lang (2000) have pointed out that “objectively measuring firm relatedness on a large sample is difficult” (p. 629) and “some practitioners even suggest that the selection of comparable firms is essentially an ‘art form’ that should be left to professionals” (Bhojraj and Lee, 2002, p. 408). There is much empirical evidence illustrating the limited power of established classification schemes. For example, in a cross-sectional analysis of 63 companies, King (1966) finds the industry to explain only about 10% of the variance in quarterly stock returns. Many recent studies have called into question the accuracy of popular methods, such as SIC codes, for industry classification (e.g., Bhojraj et al., 2003; Clarke, 1989; Fan and Lang, 2000). In particular, Bhojraj et al. (2003), who evaluate the usefulness of industry classifications based on their power to explain cross-sectional stock comovements within an industry, find the SIC to be of limited use.
Yet only a few studies offer alternative methods for industry classification (e.g., Fan and Lang, 2000; Ramnath, 2002). However, those methods are either limited to specific types of firms (Ramnath, 2002) or vulnerable to subjective judgments (Fan and Lang, 2000). All of these methods provide industry classifications that are fairly stagnant (i.e., they do not reflect changes in firms’ relatedness quickly) and provide us with a nominal categorization only (i.e., they do not allow us to interpret the relative strength of firm relatedness).

In this study, we propose an alternative approach to define industry groups based on investor perceptions of the impact of information on various groups of stocks. Ultimately, stock prices are driven by information and the degree to which new information affects two companies (e.g., news stories mentioning both firms) should thus allow us to derive a measure of their relatedness (i.e., the relative frequency with which two stocks are mentioned together). We use the degree to which pairs of companies are associated with each other in an online stock forum to develop a network of relationships and leverage methods from social network analysis in order to extract cohesive groups from an investor perspective. If market participants consider a set of companies closely related, their stocks should experience coincident movement (Chan et al., 2007). Thus, we explore return correlations in order to determine whether our method to delineate industry groups is economically meaningful.

Our study explores the following research questions. First, we investigate whether the degree to which pairs of companies are associated with each other in an online stock forum (i.e., relatedness) corresponds to the comovement of their stocks. Second, we explore whether our measure of relatedness can serve to define strategic peer groups1 for individual firms that are meaningful in terms of stock comovement and relative to SIC peers. Third, we analyze whether industry groups defined by our measure of relatedness are a viable alternative to established industry classification schemes.

We find that the degree to which companies are mentioned jointly in an internet stock forum can explain the comovement of their stocks. Our measure of relatedness can help identify a firm’s strategic peers from an investor perspective and delineate industry groups, which explain stock returns with the same power as established classification methods, but offers a number of promising advantages. Our approach is, for example, independent of the arbitrary assignment of companies by individual experts by leveraging the insights of hundreds of investors and, in contrast to fixed classification schemes, can quickly reflect changes in firm relatedness.

The main contributions of this study are as follows. First, our study provides empirical evidence supporting the theory that information associated with a set of firms is an indicator of relatedness and the comovement of their stocks (e.g., King, 1966). Second, leveraging methods from social network analysis, we present a novel news-based approach to determine industry classifications and define strategic peer groups from an investor perspective. Third, we demonstrate that our measure offers advantages over established classification schemes in that it provides a continuous measure of relatedness for every pair of two companies, which can be updated in real-time and at arbitrary intervals.

The remainder of the paper is structured as follows. First, we review related work and derive our research questions. Second, we describe our data set and methodology illustrating how online stock forums can be used to define industry groups. Third, we provide results of our analysis of the usefulness of our classification method with respect to the comovement of stocks, the identification of strategic peer groups and as a classification scheme vis-à-vis the established SIC classification system. We conclude that the user perception of strategic peers can be used to delineate meaningful industry groups. Finally, we discuss the implications of our findings and provide suggestions for further research.

2. Related work

2.1. Limitations of prevalent industry classification schemes

There are various industry classification schemes assigning individual companies to a particular industry group (for a comprehensive overview, see Bhojraj et al., 2003). The Standard Industrial Classification (SIC) is the predominant classification system in

Note that we use the term “strategic group” loosely referring to related firms from the perspective of an individual firm. It does not directly follow the definition traditionally used in strategy research referring to intraindustry subgroups of companies following a similar strategy (e.g., DeSarbo et al., 2009).
capital market research, with more than 90% of relevant studies making use of this classification scheme (Bhojraj et al., 2003). SIC codes are based on industry categories defined by the U.S. Census Bureau reflecting similarities between firms with respect to the products they produce or the manufacturing technologies they employ (Clarke, 1989). It has become the primary method for delineating industrial activity in the United States. However, Clarke (1989), who has examined whether firms in the same SIC category exhibit similar changes in sales, profit or stock prices, concludes that SIC codes are not particularly successful at identifying firms with such similar characteristics. Responding to changes in the U.S. economy, the North American Industry Classification System (NAICS) was developed, but has, in practice, not yet replaced the well established SIC system. In the meantime, financial practitioners have developed the Global Industry Classifications Standard (GICS) and financial scholars are making adjustments to the SIC categorization for research purposes (Fama and French, 1997). In sum, due to significant shortcomings of established industry classification systems a considerable amount of time and effort is being spent to divide firms into homogenous groups.

2.2. Alternative methods for industry classification

Few scholars so far have offered viable alternatives to prevalent industry classification schemes. Fan and Lang (2000) use commodity flow data from input-output (IO) tables to construct IO-based measures of interindustry and intersegment relatedness. However, commodity flows are available only for roughly 500 private-sector industries, and are thus not well-suited for firm-level analysis. In a study of analyst reactions to earnings announcements, Ramnath (2002) as well as Zuckerman and Rao (2004) have used an analyst-based definition of industry groups. The authors define an industry as a group of firms having a certain number of security analysts in common. This approach to industry definition is, of course, limited to large companies with a sufficient analyst following. In equity research and valuation, accounting-based multiples (e.g., price-to-earnings, price-to-book ratios) are often used to select comparable firms, but this approach is meaningful only within an industry previously specified by standard classification schemes (Bhojraj and Lee, 2002). Next to these objective approaches, there are a few studies, particularly related to the literature on firm and portfolio diversification, which use subjective criteria to classify companies into economic sectors (e.g., Rumelt, 1982) or broad categories characterized by growth, cyclical and stable return characteristics (Farrell, 1974) and are thus dependent on expert judgment. In addition, there are attempts to group companies outside the field of financial research that are related to our objective. Strategy researchers identify strategic groups by clustering firms based on firm-level dimensions that characterize strategy (e.g., cost structure, degree of product diversification, formal organization; DeSarbo et al., 2009; Dranove et al., 1998). However, due to the industry-specific definition of firm-level variables these methods are largely limited to defining subgroups within one particular industry and do not allow to group firms across the entire industry landscape. Finally, in econophysics, graph-theoretical methods have been applied to develop networks of financial markets using stocks as nodes and stock comovement as indicators of the strength of their ties (e.g., Bonanno et al., 2004; Mantegna, 1999; Omelka et al., 2003). Using similar methods, practitioners have identified strategic opportunities by drawing maps of competing firms based on semantic clusters of key phrases associated with individual companies in millions of corporate documents (Gourley, 2011). Exploratory and qualitative evidence from these studies suggest that a network perspective of company relationships can be leveraged for firm-level analysis, Fan and Lang (2000) use their own conversion table to link company SIC codes to IO codes and limit the use of their relatedness measure to the analysis of corporate diversification strategies among multi-segment firms.
to derive groups of stocks that are homogeneous with respect to traditional industry classifications and may be used to design stock indices (Tse et al., 2010). However, research in this field focuses on the analysis of graph-theoretical properties and network topologies without a rigorous, quantitative comparison with existing industry classification systems.

2.3. A news-based measure of firm-relatedness

We propose an alternative approach to define industry groups based on the perception of the impact of information on various groups of stocks. Ultimately, stock prices are driven by information. King (1966) has pointed out that “the stock market is subject to a steady inflow of information, much of which will […] fall into various classes according to the scope of its effect on the market” (p. 140). There are some news items, which will have a market-wide impact (e.g., changes in monetary policy), other information, which will affect only a subgroup of stocks (e.g., changes in defense policy affecting the aircraft industry), and a third class of information, which will be relevant only to a particular security (e.g., earnings announcements). The degree to which new information affects two companies should thus allow us to derive a measure of relatedness. While this line of argument is intriguing, it is usually difficult to observe such as “steady inflow of information” and link it to a particular set of stocks. However, the rise of online stock forums provides us with a unique data source documenting previously unavailable facets of information processing by online investors in real-time (for an example of the use of internet stock message boards in academic research, see Antweiler and Frank, 2004). A study investigating the online stock forum Yahoo!Finance indicates that stocks that are associated with each other on internet message boards, exhibit stronger comovement than other stocks (Das and Sisk, 2005). Even though these results are encouraging with respect to our hypothesis, they come with two limitations, which we address in our study. First, Das and Sisk (2005) define relatedness as a large share of common users on the message boards of two companies suggesting that “message boards belonging to the same community may result in similarity of opinion reflected in stock trades, ultimately impounded in stock returns” (p. 8). However, users often leave messages at different points in time and thus a common user base among message boards may not necessarily translate into “similarity of opinion reflected in stock trades”. Therefore, instead of studying online users as the carriers of static information sets, this study focuses on individual bits and pieces of information that are directly associated with a set of stocks in real-time. Second, Das and Sisk (2005) limit their analysis to the existence of stronger comovement among related stocks without comparing the strength of this effect to objective benchmarks. We leverage our measure of relatedness to delineate industry groups and compare these to existing classification schemes.

In line with previous research (Bhojraj et al., 2003; Chan et al., 2007), we focus the comparison of these classification schemes on their explanatory power for stock comovement as a benchmark of firms’ similarity. The literature distinguishes between fundamentals-based and industry-specific comovement (Barberis et al., 2005; Pindyck and Rotemberg, 1993). According to the fundamentals-based theory, comovement of stock returns can be linked to similar firm fundamentals (i.e., the assets owned by a firm). The theory of industry-specific comovement attributes comovement of stocks to market frictions and investor sentiment. Barberis et al. (2005) suggest that industry-specific comovement may result from investor preference for certain industries (habitat view) or an industry focus due to limited processing power (category view). In either case, investors direct their funds on the industry-level resulting in industry-specific comovement. Our approach to industry classification is based on investor perceptions of firms’ relatedness, which may reflect both sources of comovement. In other words, investors may mention two stocks jointly because of their perception of similar fundamentals or due to their preference for certain industries.
3. Data set and methodology

3.1. Data set and sample selection

In this section, we describe our data set and detail the methodology used to derive our measure of relatedness between firms. The advent of online stock forums has made observable many aspects of information processing by investors, which were previously unavailable. Internet stock forums allow investors to exchange stock-related information and trading ideas online. Das et al. (2005) have profiled users of these forums and suggest that the majority of them are individual investors and day traders. We chose the microblogging platform Twitter as our data source. Twitter allows users to publish short messages with up to 140 characters, so-called “tweets”. Users can subscribe to (i.e., “follow”) a selection of favorite authors or search for messages containing a specific key word (e.g., a stock symbol). The public timeline has turned into an extensive real-time information stream of millions of messages per day. Many of these messages discuss public companies, trading ideas and current news. Some commentators see in the conversations on this platform “the modern version of traders shouting in the pits” (BusinessWeek, 2009). The investor community has come to call Twitter and related third-party applications “a Bloomberg for the average guy” (BusinessWeek, 2009). Academic researchers were only recently drawn to Twitter as a field for capital market research. A few studies suggest that the information content of Twitter messages may help predict macroeconomic indicators such as the Index of Consumer Sentiment (O’Connor et al., 2010), stock market indices such as the Dow Jones Industrial Average (Bollen, et al., 2010) or the S&P 500 (Zhang et al., 2010) and even returns and trading volume of individual stocks (Sprenger and Welpe, 2010).

Traders have adopted the convention of tagging stock-related messages by a dollar sign followed by the relevant ticker symbol (e.g., “$AAPL”). We focus on this explicit subset of messages. This focus allows us to investigate the most relevant news items and avoid “noise” (i.e., messages that are not related to publicly traded companies). Messages are accessible via the website’s application programming interface (API). We study the 6 month period between January 1st and June 30th, 2010, to deal with stable developments on the U.S. financial markets and to avoid potentially distorting repercussions of the subprime mortgage crisis in 2009. During this period, we have collected 439,960 English-language, stock-related microblogging messages containing the dollar-tagged ticker symbol of an S&P 500 company. We focus on the S&P 500 to adequately reflect a wide spectrum of U.S. equities, which permits a cross-industry analysis, while limiting our study to well-known companies that trigger a substantial number of stock microblogs.

3.2. Investor perceptions of strategic peer groups

Table 1 shows several random examples of messages from stock microblogs. In line with King (1966), we find news items that investors relate to one particular stock (e.g., “$TGT Target Q4 Profits Surge”) as well as others that are associated with multiple firms (e.g., “Energy doing well. SCHK SOXY”). Roughly 13.4% of all messages mention more than one stock. Reasons for these joint mentions include the impact of macroeconomic developments (e.g., “Big banks up or down with Bernanke’s re-nomination? $C $BAC $WFC”), the launch of new products affecting competitors (e.g., “Crazy Google now building super-high-speed fiber Internet network to scare Comcast and AT&T: http://bit.ly/dvWSzL $GOOG $T $CMCSA”) and legal actions of one firm against another (e.g., “Goldman Sachs $GS demands 4 billions from AIG $AIG to cover mortgage securities AIG insured, helped trigger crisis, forced gov to bail out AIG”). Irrespective of the content of individual messages, they all indicate that one company is associated with another and impacted by the same piece of news. These messages containing joint mentions are the focus of our analysis.

We investigate whether joint mentions can serve as a measure of relatedness between firms. To make the comparison more flexible and the interpretation more straightforward, we focus on the pairwise relationships between companies. Based on the overall probability that any one firm is mentioned in a message, a

---

7 We will refer to these tweets as messages or news items.

8 Twitter provides only a limited history of data at any point in time. We, therefore, developed a webcrawler, which made requests to and downloaded data from the Twitter API 24 hours a day. A load balancing feature ensured that messages associated with more frequently mentioned stock symbols were downloaded more often ultimately providing us with an uninterrupted stream of messages for the 6 months covered in this study.

9 Specifically, we focus on those companies that have been included in the S&P 500 as of January 1, 2010.
conditional probability that two firms are mentioned together can be computed. If all combinations were equally likely, this conditional probability should be equal to the observed share of messages mentioning these two firms. Due to different base rates, we divide the observed share of joint mentions by the conditional probability to derive a comparative measure. If $share(AAPL, GOOG)$ represents the share of observed joint mentions of these two stocks, the relative frequency $R$, is calculated as follows:

\[ R = \frac{\frac{share(AAPL, GOOG)}{P(AAPL|GOOG) + P(GOOG|AAPL)}}{2} \]  

(1)

The relative frequency is a measure of Relatedness and illustrates how often two stocks are mentioned together relative to the random probability based on the overall “share of voice” of the individual stocks. If $R$ equals 1.5 the share of observed joint mentions is 50% higher than pure chance would suggest (i.e., $R = 1.5$). This measure has been used successfully in the context of microblogs to derive a measure of the relative frequency of joint mentions of political parties, which was related to their ideological proximity (Tumasjan et al., 2010). We have limited our analysis to firms that were mentioned at least 100 times in our sample period, leaving us with 415 stocks from the S&P 500. We refer to any major financial website to match the ticker symbols used, for brevity, throughout the paper with the corresponding company names of these stocks.
Table 2
Most and least related firms

<table>
<thead>
<tr>
<th>Stock 1</th>
<th>Stock 2</th>
<th>Relatedness</th>
<th>Comovement</th>
<th>Same SIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AYE Allegheny Energy</td>
<td>FE FirstEnergy Corp</td>
<td>685.3</td>
<td>0.64</td>
<td>1</td>
</tr>
<tr>
<td>PGN Progress Energy Inc.</td>
<td>SO Southern Co.</td>
<td>549.9</td>
<td>0.81</td>
<td>1</td>
</tr>
<tr>
<td>AIV AIMCO</td>
<td>WMB Williams Cos.</td>
<td>456.8</td>
<td>0.71</td>
<td>0</td>
</tr>
<tr>
<td>FIS Fidelity National Information Services</td>
<td>FISV Fiserv Inc.</td>
<td>399.7</td>
<td>0.34</td>
<td>1</td>
</tr>
</tbody>
</table>

Top 10

<table>
<thead>
<tr>
<th>Stock 1</th>
<th>Stock 2</th>
<th>Relatedness</th>
<th>Comovement</th>
<th>Same SIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>FPL FPL Group</td>
<td>PGN Progress Energy Inc.</td>
<td>356.4</td>
<td>0.85</td>
<td>1</td>
</tr>
<tr>
<td>FPL FPL Group</td>
<td>SO Southern Co.</td>
<td>340.1</td>
<td>0.76</td>
<td>1</td>
</tr>
<tr>
<td>MDT Medtronic Inc.</td>
<td>STJ St Jude Medical</td>
<td>337.5</td>
<td>0.76</td>
<td>1</td>
</tr>
<tr>
<td>BXP Boston Properties</td>
<td>VNO Vornado Realty Trust</td>
<td>311.3</td>
<td>0.90</td>
<td>1</td>
</tr>
<tr>
<td>KR Kroger Co.</td>
<td>SWY Safeway Inc.</td>
<td>311.1</td>
<td>0.59</td>
<td>1</td>
</tr>
<tr>
<td>CTL CenturyTel Inc.</td>
<td>Q Qwest Communications Int.</td>
<td>299.1</td>
<td>0.24</td>
<td>1</td>
</tr>
</tbody>
</table>

Bottom 10

<table>
<thead>
<tr>
<th>Stock 1</th>
<th>Stock 2</th>
<th>Relatedness</th>
<th>Comovement</th>
<th>Same SIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADBE Adobe Systems</td>
<td>BAC Bank of America Corp.</td>
<td>0.1</td>
<td>0.52</td>
<td>0</td>
</tr>
<tr>
<td>C Citigroup Inc.</td>
<td>ORCL Oracle Corp.</td>
<td>0.1</td>
<td>0.57</td>
<td>0</td>
</tr>
<tr>
<td>GOOG Google Inc.</td>
<td>HAL Halliburton Co.</td>
<td>0.1</td>
<td>0.44</td>
<td>0</td>
</tr>
<tr>
<td>AAPL Apple Inc.</td>
<td>KFT Kraft Foods Inc.</td>
<td>0.1</td>
<td>0.12</td>
<td>1</td>
</tr>
<tr>
<td>MS Morgan Stanley</td>
<td>MSFT Microsoft Corp.</td>
<td>0.1</td>
<td>0.44</td>
<td>0</td>
</tr>
<tr>
<td>AAPL Apple Inc.</td>
<td>WFC Wells Fargo</td>
<td>0.1</td>
<td>0.56</td>
<td>0</td>
</tr>
<tr>
<td>GOOG Google Inc.</td>
<td>WFC Wells Fargo</td>
<td>0.1</td>
<td>0.51</td>
<td>0</td>
</tr>
<tr>
<td>GS Goldman Sachs Group</td>
<td>ORCL Oracle Corp.</td>
<td>0.0</td>
<td>0.31</td>
<td>0</td>
</tr>
<tr>
<td>AAPL Apple Inc.</td>
<td>HBAN Huntington Bancshares</td>
<td>0.0</td>
<td>0.62</td>
<td>0</td>
</tr>
<tr>
<td>GOOG Google Inc.</td>
<td>HBAN Huntington Bancshares</td>
<td>0.0</td>
<td>0.37</td>
<td>0</td>
</tr>
</tbody>
</table>

Notes: This table shows the firms that are most and least related in terms of our measure of relatedness. Results were limited to pairs of stocks with at least 10 joint mentions. The most related firms tend to come from the same SIC industry group and their stocks exhibit, on average, higher comovement.

that are associated with one another more closely tend to come from the same top-level SIC industry group. The comovement of their stocks is, on average, substantially higher than the comovement of stocks that are mentioned jointly less frequently (c = 0.66 vs. 0.45 with p < 0.01).

3.3. Identification of strategic peers and delineation of industry groups

Our measure of relatedness informs us about the news-based proximity of any pair of stocks. In this section, we will introduce the methods used to meet our objective to leverage this information in order to derive a firm’s strategic peers and delineate relevant industry groups. The pairwise relationships that our measure of relatedness provides us with essentially create a network of all stocks in our sample. Figure 1 shows a graph for an extract from this network. The network contains information regarding the link (i.e., a line or, in network theory, the edge or tie) between two stocks as well as the strength of this link (i.e., the thickness, which represents Relatedness), the SIC industry classification (i.e., the shape of the symbol) and the absolute volume of company mentions (i.e., the size of the stock symbol). In line with the approach taken by Das and Sisk (2005) and researchers in the field of econophysics (e.g., Bonanno et al., 2004; Mantegna, 1999; Onnela et al., 2003), from this network we will extract strategic peers related to a

Note that this so-called network is not made up of archetypical network features (i.e., a set of nodes that are independently connected to other nodes), but derived from pairwise relationships among a set of stocks.
Fig. 1. Investor perceptions of the relationship of S&P 500 stocks

Notes: This figure shows the relationship of S&P 500 stocks in terms of joint mentions in stock microblogs. The size of the stock symbol represents the total number of mentions, the thickness of the lines is indicative of the relative frequency of joint mentions (i.e., Relatedness as defined in our methodology section). For better readability the figure was limited to stocks that were mentioned together at least 50 times.
particular stock as well as cohesive subgroups that define an industry.\textsuperscript{14}

Taking the perspective of a particular company, we can simply define the peer group as those firms that are most closely related to it, i.e., firms with the highest measure of relatedness.

The above mentioned approach cannot be employed to delineate clear-cut industry groups which clearly assign each company to exactly one category. Thus, to be consistent with established industry classification systems, we also want to partition the network into a pre-determined number of groups. In social network analysis, so-called faction analysis is used to achieve this objective. In short, this clustering algorithm optimizes a cost function, which measures the degree to which a partition forms a cohesive subgroup and takes into account both the ties as well as their strength (for details regarding this method, see Glover, 1990). Cohesiveness can be thought of as the average tie strength within each partition. We use the Ucinet implementation for the faction analysis of our dataset\textsuperscript{15} (Borgatti et al., 2002).

### 3.4. Similarity of stocks

In order to objectively evaluate the quality of our measure of relatedness as a proxy for firms’ similarity, we need to define external benchmarks for comparison. In this section, we will describe stock comovement and existing SIC industry classifications as input parameters to construct these benchmarks.

At the level of company pairs, we can compare Relatedness to the comovement of stocks. In line with related research (e.g., Bhojraj et al., 2003), we calculate comovement as the correlation\textsuperscript{16} of stock returns. To isolate the market-related component of this comovement, we start with a market model

\[
    r_i = \alpha_t + \beta_t r_m + \varepsilon_{i,t},
\]

in which the return of a stock \( r_i \) is explained by a firm-specific component \( \alpha_t \) and a market-related component, which depends on the stock’s sensitivity (\( \beta_t \)) to overall market returns \( r_m \). \( \varepsilon_{i,t} \) is a time- and company-specific error term. The S&P 500 serves as our proxy for the market return. The market-related comovement of two stocks \( i \) and \( j \) can then be isolated as

\[
    \sigma_{i,j} = \beta_i \beta_j \sigma_m^2,
\]

where \( \sigma_m \) is the standard deviation of the market return. We use our entire 6 month sample period to estimate the above mentioned parameters.

At the aggregate level, the most obvious comparison for our industry classification is the widely established SIC system. There are varying levels of granularity in the SIC classification scheme. For the purpose of our study, we use the most common two-digit level of analysis, which Clarke (1989) suggests to have the greatest explanatory power with respect to industry returns.\textsuperscript{17} In order to calculate the explanatory power of our classification for the stock returns of the companies in a peer or industry group, we model a firm’s stock returns, \( r_{i,t} \), with the simple OLS regression

\[
    r_{i,t} = \alpha_t + \beta r_{\text{ind},t} + \varepsilon_{i,t},
\]

where \( r_{\text{ind}} \) is the equally weighted industry return. This permits us to evaluate the explanatory power of alternative industry definitions by comparing adjusted \( R^2 \)’s of various industry definitions, following the methodology use by Bhojraj et al. (2003). All returns are calculated as the log difference of total return to shareholders, which reflects both price changes and dividend payments.

### 3.5. QAP methodology

When we compare our measure of relatedness to the comovement of stocks, we are dealing with sets of company pairs. The pairwise structure of so-called dyadic data requires special attention because every

\textsuperscript{14}Note that this approach is related to that in strategy research where strategic groups are identified by clustering firms based on firm-level variables such as R&D expenditures or distribution of sales (e.g., DeSarbo et al., 2009).

\textsuperscript{15}In Ucinet 6 the procedure can be found under Tools/Cluster Analysis/Optimization.

\textsuperscript{16}To ensure the robustness of our results, we have repeated the regression analysis with the covariance of stock returns. The results are in line with those using the correlation of stock returns as dependent variable (i.e., Relatedness and SIC group have about the same explanatory power for covariance). To save space, we do not report the results tables separately.

\textsuperscript{17}Others, such as Fertuck (1975), suggest that three-digit SIC codes are the most useful in predicting return variation. Thus, we have replicated our analyses at the three-digit level and find our results to be largely in line with those that we report in this paper. However, we prefer the less granular two-digit level of analysis for our relatively small sample of 500 firms to avoid the exclusion of too many industry groups with too few sample companies.
company is included in multiple sets. We can think of the data as a matrix with every stock listed in one row and one column and the attributes of the resulting pairs (e.g., relatedness, comovement) occupying the cells. Thus, the observations are not independent preventing the use of ordinary regression models if the dependency between the observations is not controlled for. This is a common phenomenon in social network analysis, where the resampling-based nonparametric Quadratic Assignment Procedure (QAP) is used to deal with dyadic data sets (for a comprehensive description of the QAP, see Krackhardt, 1988). Similar to bootstrapping methods, the QAP permutes the rows and columns of the above-mentioned matrix, but maintains rows and columns for individual companies. As a result, the permuted datasets comply with the null hypothesis. Resampling multiple times and running OLS regressions on the “scrambled datasets” permits us to determine the percentile of the original coefficients relative to the empirical distribution of permuted datasets. In sum, we employ the QAP to account for the pairwise nature of our dataset.

4. Results

4.1. Overview and anecdotal evidence

The results section is structured as follows. We will first highlight noteworthy aspects of the network graph of all stocks introduced in our methodology section in order to provide first descriptive evidence of identifying company groups with our proposed methodology. Then, we will turn to our three research questions and investigate, first, whether relatedness of companies in the eyes of online investors can explain the comovement of their stocks, second, whether relatedness can serve to identify a company’s strategic peers and, third, if we can extract meaningful industry groups from the network of relationships.

Figure 1 shows the relationship of S&P 500 stocks in terms of joint mentions in stock microblogs. For a rough comparison with the SIC coding, the shape and color of each stock symbol represents the one-digit SIC industry group. The layout of the network graph is not the product of a random process, but was automatically derived in order to maintain roughly an equal distance between nodes while ensuring readability through node repulsion. As a result, apart from tracing direct links, we can interpret proximity of a group of stocks as the degree to which they are related. We find that many subgroups of stocks are consistent with our intuition of classic industry delineations. Financial firms, such as Goldman Sachs (GS), JP Morgen (JPM), and Bank of America (BAC) are closely interconnected. There are also tight-knit smaller groups of stocks, such as the media companies Disney (DIS), CBS Broadcasting (CBS) and Time Warner (TWX). In addition, some subgroups exist that are not connected to the rest of the network, for example the insurance companies WellPoint (WLP) and United Health (UNH), the logistics firms United Parcel Service (UPS) and Fedex (FDX), and the hardware stores Home Depot (HD) and Lowe’s (LOW). Their isolated position suggests that these companies form micro industries that are often subject to the same news items, but not frequently associated with other firms.

The network graph not only confirms classical industry groupings, but also reveals interesting connections between these industries. For example, while Exxon Mobile (XOM) is obviously closely related to other major energy firms such as Chevron (CVX) and ConocoPhillips (COP), it is also linked to subcontractors such as the exploration firms Anadarko Petroleum (APC) and Halliburton (HAL), which in turn are associated with their equipment suppliers Schlumberger (SLB) and Cameron (CAM). Note that these relationships cut across traditional SIC categories. Online retailer Amazon (AMZN) is another interesting example in this respect, because the company appears to be a hub that is associated with traditional “brick and mortar” retailers, such as Walmart (WMT), Target (TGT) and Costco (COST), online retailers, such as Priceline (PCLN) and Ebay (EBAY), computer soft- and hardware firms, such as Apple (AAPL), Microsoft (MSFT) and Intel (INTC), and communication providers, such as AT&T (T). The network also highlights particularly competitive relationships, such as those between Coke (KO) and Pepsi (PEP) or Visa (V) and Mastercard (MA), as well as issues related to corporate control and ownership, such as the strong link between Disney (DIS) and its parent company General Electric (GE).

We conclude, that the news-based investor perception of strategic peer groups offers unique and rich descriptive insights into the relationship between companies. In the following, we will investigate whether we can leverage this insight to identify

---

18 The null hypothesis is the hypothesis that there is no relationship in the matrix.
Explaining stock comovement through Relatedness

<table>
<thead>
<tr>
<th>Comovement interval</th>
<th>Model 1</th>
<th>Model 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Daily</td>
<td>Weekly</td>
</tr>
<tr>
<td>Estimation method</td>
<td>QAP</td>
<td>Clustered</td>
</tr>
<tr>
<td>Investor perception</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relatedness</td>
<td>0.006***</td>
<td>0.006***</td>
</tr>
<tr>
<td>SIC industry component</td>
<td></td>
<td></td>
</tr>
<tr>
<td>same SIC dummy</td>
<td>0.0666***</td>
<td>0.0666***</td>
</tr>
<tr>
<td>Market component</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Market-related comovement</td>
<td>0.0118***</td>
<td>0.0118***</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.234</td>
<td>0.234</td>
</tr>
<tr>
<td>F-value</td>
<td>8838.35***</td>
<td>169.1***</td>
</tr>
</tbody>
</table>

Notes: This table shows the explanatory power of our measure of relatedness for stock comovements. Stock comovement, defined as the correlation of stock returns, serves as the dependent variable. The independent variables were standardized to compare relative effect strengths. We can see that relatedness adds to the explanation of comovement beyond the control variable market-related comovement and the industry component due to traditional SIC classifications. The sample is based on all pairs of 415 firms \( (N = 171,810) \).

Regarding the estimation methods: We have used the QAP with 500 iterations of OLS estimations. In addition, we provide the results for clustered regressions, where robust standard errors are adjusted for intragroup correlation (with one company in each pair defining the group), as a robustness check of our results.

*** indicates significance at the 1% level.

meaningful peer groups and establish a link to stock market returns.

4.2. Comovement

In this section, we will explore whether our news-based measure of relatedness is indicative of stock comovement. In order to deal with the dyadic nature of our dataset we employ the QAP introduced in our methodology section. Following related studies that have used a SIC-based variable of relatedness (which equals one if two companies belong to the same SIC code and zero otherwise) and market-related comovement as control variables (e.g., Fan and Lang, 2000), we likewise control for these two measures in our model.

Table 3 shows the results of the QAP regression. Model 1 uses daily comovement as the dependent variable (left panel), whereas model 2 is based on the comovement of weekly returns (right panel). Next to the results for the QAP regressions and as a robustness check, we provide the results for clustered regressions, where robust standard errors are adjusted for intragroup correlation (with one company in each pair defining the group). The independent variables were standardized to compare relative effect strengths.

We can see that the same-industry-dummy has the strongest explanatory power for stock comovement \( (c = 0.066) \). It is roughly 6 times as strong as general market-related comovement \( (c = 0.012) \). However, beyond these traditional measures of proximity, our measure of relatedness has statistically significant explanatory power for stock comovement \( (c = 0.006) \). These results, even though based on short-lived news items, are not limited to daily stock comovement. The analysis of weekly comovement shows a very similar pattern (right panel of Table 3). We conclude that relatedness can help explain the comovement of stock returns over and above considering the presence of both companies in the same SIC industry.19

The two models in Table 3 were based on relatedness and comovement over our entire sample

19Note that the same-industry dummy is a binary variable (either 1 or 0, depending on whether two stocks are in the same industry or not) whereas Relatedness is discrete. As a result, we have to be cautious when we compare the coefficients in the regression models and cannot directly gauge the variables relative power in explaining comovement.
period. However, our news-based measure of relatedness may have a crucial advantage over static SIC classifications. News-based relatedness could be updated frequently and reflect changes over time. The results from Table 3 do not show whether our measure of relatedness can adjust to changes in firm proximity quickly. Therefore we have conducted OLS-based QAP panel regressions with a time series of Relatedness and stock comovement of all company pairs at a monthly interval. In this case, the permutations described in our methodology section ensure that the time series of the observation corresponding to a row or column in the matrix is kept together. Table 4 shows the results of these regressions. Model 2 illustrates that changes in Relatedness over time are positively associated with changes in the comovement of daily stock returns. As the change in $R^2$ shows, Relatedness can partially explain comovement even when market-related comovement and SIC-industry dummies are included. Thus, at least at monthly intervals, an increase in Relatedness reflects an increase in stock comovement.

### 4.3. Strategic peer groups

In this section, we will investigate whether relatedness can help identify a firm’s strategic peers. For selected companies we compare the peers according to the SIC and our measure of relatedness as well as their explanatory power for industry returns.

Table 5 shows the 10 two-digit SIC groups with more than 20 sample companies. We have selected the company with the highest message volume in our sample for each of these industry groups and show all sample companies with the same SIC code. For comparison, we have chosen the same number of most closely related companies in terms of Relatedness. Peers found in both groups are bolded. We can see that the overlap of peers derived by the two methods is around one quarter (23%). Adjusted $R^2$'s are shown for the OLS regressions that try to explain stocks returns through industry returns, where the industry is defined as the specified peer group. While the explanatory power of SIC groups for industry returns is generally higher than that of news-based relatedness (average adjusted $R^2$ of 0.52 vs. 0.47), this difference is not significant at the 5% level. Thus SIC-defined peer groups are not a significantly better representation of stock-related proximity than peer groups defined by our measure of relatedness.

Next to the return-based quantitative support for Relatedness as a meaningful instrument to identify strategic peer groups, the measure also makes sense from a qualitative viewpoint. News-based peer groups reflect many intuitive competitive relationships. Google (GOOG), for instance, is associated with Yahoo (YHOO), eBay (EBAY) and Microsoft (MSFT) according to both methods. However, beyond a mere reflection of standard classifications, Relatedness reveals meaningful relationships between companies, which the SIC scheme does not capture. One of the

---

**Notes:**

This table shows the results for an OLS-based QAP panel regressions with a timeseries of Relatedness and stock comovement of all company pairs at a monthly interval. The QAP permutations described in our methodology section ensure that the time series of an observation corresponding to a row or column in the matrix is kept together. The time series consists of monthly observations of the dependent variable stock comovement at daily intervals and the independent variables Relatedness etc. for the same 4 week period. The independent variables were standardized to compare relative effect strengths. The coefficients illustrate the explanatory power of our measure of relatedness for stock comovement over time. We can see that changes in Relatedness are positively associated with changes in stock comovement. Thus, this measure reflects changes in firm relatedness. The sample is based on all pairs of 415 firms ($N = 171,810$ company pairs).

*** indicates significance at the 1% level.

---

20 We would like to point out that our approach does not rely on any particular seed company. The idea of this subsection is to identify peers for any given firm. The selected companies were chosen with respect to the SIC classification scheme only to ensure that a representative sample of industries are covered in the analysis.

21 To check the robustness of our results, we have repeated this analysis using weekly stock returns. As in the case of daily returns, SIC industries are only slightly better at explaining weekly industry returns (average adjusted $R^2 = 0.497$ vs. 0.457), but this difference is not significant at the 10% level.
<table>
<thead>
<tr>
<th>Sample firm</th>
<th>SIC name</th>
<th>Same SIC</th>
<th>SIC peers</th>
<th>Adj. $R^2$</th>
<th>Relatedness</th>
</tr>
</thead>
<tbody>
<tr>
<td>GOOG (Google)</td>
<td>Business services</td>
<td>36</td>
<td>ADBE, ADP, ADSK, AKAM, BMC, CA, CPWR, CRM, CSC, CTSH, CTXS, EBAY, EFZ, ERTS, FIS, FISV, INTU, IPG, IRM, JNPR, MA, MCO, MFE, MSFT, MWL, NOVL, OMC, ORCL, RHI, RHT, SYMC, TSS, V, VRSN, WU, YHOO</td>
<td>0.49</td>
<td>Same SIC: AAPL, ADBE, AMZN, BBY, CCMCA, CRM, CSC, CTSH, DELL, EBAY, EMC, F, FCX, FHN, GE, GS, HPQ, IBM, IGT, INTC, ISRG, MFE, MOT, MSFT, NVDA, PCLN, PPG, QCOM, STZ, VZ, WPO, XRX, YHOO, YUM</td>
</tr>
<tr>
<td>AIG (American Int'l)</td>
<td>Insurance carriers</td>
<td>23</td>
<td>AET, AFL, AIZ, ALL, CB, CI, CINF, CVH, GNW, HIG, HUM, L, LNC, LUK, MET, PFG, PGR, PRU, TMK, TRV, UNH, UMN, WLP</td>
<td>0.56</td>
<td>Same SIC: AEE, AEP, AES, AFL, ALL, APD, AYE, BDK, C, CB, CNP, DOW, ECL, FRX, GS, HAK, HIG, IPG, MAS, MET, MIL, ODP, PRU</td>
</tr>
<tr>
<td>C (Citigroup)</td>
<td>Depository institutions</td>
<td>22</td>
<td>BAC, BBT, BK, CMA, COF, FHN, FITB, HBAN, HCBK, JPM, KEY, MI, MTB, NTRS, PBCT, PNC, RF, STI, TRV, UNH, WFC, ZION</td>
<td>0.64</td>
<td>Same SIC: A, AIG, AXP, BAC, BK, CMA, COF, DFS, DNR, DOW, ETFC, F, GS, JPM, MI, MO, MS, NTRS, TH, UNP, USB, WFC</td>
</tr>
<tr>
<td>D (Dominion Resources)</td>
<td>Electricity, gas, and sanitary services</td>
<td>39</td>
<td>AEE, AEP, AES, AYE, CEG, CMS, CNP, DTE, DUK, ED, EIX, EP, EQT, ET, EXC, FE, FPL, GAS, NI, NU, PCG, PEG, PGN, PNN, POM, PPL, RSG, SGC, SE, SO, SRCI, SRE, STR, TE, TEG, WEC, WM, WMB, XEL</td>
<td>0.63</td>
<td>Same SIC: A, ADM, AEE, AEP, AFL, APA, AVP, BMY, CA, CAG, CHK, CLX, CNX, DUK, EMR, EP, FE, FPL, GT, HES, K, KEY, L, LLY, MMM, MO, NVDA, NWL, PGN, PXD, SLE, SO, SYY, TSN, TSO, UTX, WM, WY, XTO</td>
</tr>
<tr>
<td>A (Agilent Technologies)</td>
<td>Instruments and related products</td>
<td>24</td>
<td>AGN, BAX, BCR, BDX, BSX, CFN, COL, DHR, EK, FLIR, ISRG, KLAC, MDT, MLI, PKI, RTN, STJ, SYK, TER, TMO, WAT, XRAY, XRX, ZMH</td>
<td>0.40</td>
<td>Same SIC: AEP, APD, AIG, BCR, BDX, BEN, CNP, D, DHR, DLR, EXPD, FIS, JBL, K, L, M, MFE, NWL, PPG, R, STR, TER, VAR, WDC</td>
</tr>
</tbody>
</table>

(Continued)
<table>
<thead>
<tr>
<th>Sample firm</th>
<th>SIC name</th>
<th>Same SIC</th>
<th>SIC peers</th>
<th>Adj. $R^2$</th>
<th>Relatedness</th>
<th>Same SIC</th>
<th>Peers (most closely related)</th>
<th>Adj. $R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSCO (Cisco)</td>
<td>Electrical and</td>
<td>30</td>
<td>ADI, ALTR, AMD, APH, BRCM, EMR, FSLR, GE, HAR, HRS, JNTO, JDSU, LLL, LLTC, LSI, MCHP, MOLX, MOT, MU, NSM, NTAP, NVDA, NVLS, QCOM, QLGC, TLAB, TXN, WFR, WHR, XLLNX</td>
<td>0.56</td>
<td>4 13% AIZ, AKAM, BDK, BRCM, CLX, CMCSA, CSC, CTXS, DD, EMC, FO, GLW, HPQ, IBM, INTC, IP, JBL, JDSU, JNJ, JNPR, KSS, LLL, MBBM, MU, NOVL, NWSA, PBCT, PLD, WFRM, XLLNX</td>
<td>0.49</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AAPL (Apple)</td>
<td>Industrial</td>
<td>30</td>
<td>AMAT, BDK, BHI, CAM, CAT, CMI, DE, DELL, DOV, EMC, ETN, FLS, FTI, HPQ, IBM, JGT, ITT, JBL, LXK, MMM, NOV, PBI, PH, PLL, ROK, SII, SNKD, TDC, VAR, WDC</td>
<td>0.57</td>
<td>4 13% ADBE, AMZN, AVY, BBY, BRCM, CPWR, CRM, CSCO, DELL, DIS, EK, EMC, F, GOOG, HPQ, IBM, INTC, ISRG, MOT, MSFT, NYT, PCP, QCOM, T, TAP, TXN, VZ, WIN, WMT, WYNN</td>
<td>0.44</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PFE (Pfizer)</td>
<td>Chemicals and</td>
<td>30</td>
<td>ABT, AMGN, APD, AVP, BBIB, BMY, CELG, CEPF, CF, CL, CLX, DD, DOW, ECL, EL, EMN, FMC, FRX, GENZ, GILD, HSB, IFF, JNJ, KG, LIFE, LLY, MON, MKR, MYL, PG, PPG, PX, SIAL, WPI</td>
<td>0.44</td>
<td>7 21% ABT, ADM, AET, AMGN, BAX, BDK, BMS, BMY, BSX, CI, DUK, FRX, GILD, GIS, JNJ, KG, LLY, LM, LO, MCK, MDT, MKR, PPL, R, STJ, SYK, SYX, PVC, TPN, TVX, UTX, VMC, VAT, WIN, WLP</td>
<td>0.38</td>
<td></td>
<td></td>
</tr>
<tr>
<td>K (Kellogg)</td>
<td>Food and kindred</td>
<td>19</td>
<td>ADM, CAG, CCE, CPB, DPS, GIS, HNZ, HRL, HSY, KFT, KO, MJN, MKC, PEP, SJM, SLE, STZ, TAP, TSN</td>
<td>0.41</td>
<td>6 32% AVP, CAG, CL, CLX, CPB, DNB, FAST, GIS, HNZ, IP, KFT, NOV, RAI, SLM, SLE, TSN, VAR, VAT, WY</td>
<td>0.48</td>
<td></td>
<td></td>
</tr>
<tr>
<td>APC (Ampco-Pittsburgh)</td>
<td>Oil and gas</td>
<td>19</td>
<td>APA, BJS, CHK, COG, DNR, DO, DNV, EOG, ESV, HAL, NBL, NBR, OXY, PXD, RDC, RRC, SLB, SWN, XTO</td>
<td>0.64</td>
<td>7 37% APA, BHI, BJS, CAM, CHK, CNP, COP, DO, DNV, EOG, FTI, HAL, MIB, NBL, PAC, RLC, RRC, SLY, WAT, WFR, XOM</td>
<td>0.56</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td></td>
<td><strong>27.6</strong></td>
<td></td>
<td><strong>0.53</strong></td>
<td><strong>5.8 23%</strong></td>
<td><strong>0.47</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes: This table shows sample companies (selected by highest message volume for SIC industry groups with at least 20 companies) and their SIC peers as well as the most closely related companies in terms of Relatedness. Peers found in both groups are bolded. Adjusted $R^2$s are shown for the OLS regression $r_{i,t} = \alpha_t + \beta r_{i,ind,t} + \varepsilon_{i,t}$. The explanatory power of SIC groups for industry returns is generally higher than that of news-based relatedness, but the difference is not significant at the 5% level. We can see that there is an overlap of 23% between the two methods to define peer groups.
most strident examples is the frequently referenced rivalry in the high tech industry between Google and Apple. The SIC assigns Google to Business Services and classifies Apple as Technology, Hardware and Equipment. According to the news, Apple is one of Google’s closest competitors along with other technology firms that are assigned to different SIC codes, such as Dell (DELL), Hewlett-Packard (HPQ), and IBM (IBM). Amazon, which is traditionally classified as a retailer, is often mentioned jointly in the news with both Google (GOOG) and Apple (AAPL). While the previous examples are fairly timeless, our news-based measure of relatedness also reflects temporary relationships. In our sample period, insurance firm American International Group, for example, is associated closely with investment bank Goldman Sachs (GS). This reflects suspicions of potential conflicts of interest due to an AIG payment of $12.9 billion to Goldman Sachs, where then-Treasury Secretary Henry Paulson had previously worked as CEO, in the months after AIG was rescued by the government in September 2009. Ongoing legal actions between the two financial institutions followed (see the 6th message from Table 1 for an example). While we can only provide a few examples of this type of anecdotal evidence to illustrate the insights contained in news-based peer groups and cannot explain each and every relationship, there is enough evidence to support the notion that our measure of relatedness can be used to define meaningful strategic peer groups. In sum, we find that news-based relatedness can define meaningful strategic peer groups, which exhibit a substantial level of stock comovement.

4.4. Industry classification

In this section, we will explore whether our news-based measure of relatedness can serve to delineate meaningful industry groups. In the previous section we defined strategic peer groups as the most closely related companies from the perspective of one sample company and permitted firms to belong to multiple strategic peer groups. In the following, we will leverage social network analysis to enable the direct comparison of SIC- versus news-based industry groups.

We limit our analysis to the sample companies assigned to the 10 largest two-digit SIC groups by number of companies. This sample, along with the official SIC names, is shown in the left panel of Table 6. As in the case of traditional industry classification, we have reassigned each and every one of these firms to exactly one news-based industry group. For a direct comparison with the 10 SIC groups, we employ social network cluster analysis to delineate exactly 10 clusters of stocks. The right panel of Table 6 shows the resulting news-based industry groups.22 While there are some similarities to traditional SIC codings, the news-based classification shows distinct emphases. The cluster analysis, for instance, isolated biotech companies from the more broadly defined SIC group Chemicals and allied products, but combined pharmaceutical companies (e.g., Abbott (ABT), Pfizer (PFE), and Merck (MRK)) and food companies (e.g., Kraft (KFT), PepsiCo (PEP), and ConAgra Foods (CAG)). Internet-related firms, such as Google (GOOG) and Apple (AAPL) were assigned to one group as well as computer soft- and hardware companies, which the SIC system splits into Industrial machinery and equipment (e.g., Hewlett-Packard (HPQ), Cisco (CSCO) and Dell (DELL)) and Business services (e.g., Novell (NOVL), Juniper Networks (JNPR)). Given ever faster changes in the industrial landscape over the past decades, our methodology may be better suited to identify current industry lines than the increasingly outdated SIC scheme. One needs to keep in mind that our approach slightly favors the SIC classification, since we have used a clearly defined SIC sample as the starting point for news-based rearrangement. Even so, the explanatory power of news-based industry groups for stock returns is almost identical to SIC industry groups (average adjusted $R^2$ of 0.546 vs. 0.540).23 We conclude that news-based industry groups can be a plausible alternative to traditional industry classifications.

5. Conclusion

5.1. Discussion of results

Given the limitations of popular methods for industry classification (e.g., Bhojraj et al., 2003; Clarke, 1989; Fan and Lang, 2000), this study set out

22The names for these industry groups were assigned by the authors and seek to loosely describe the constituents.

23Similar to our analysis of strategic peer groups, we have repeated this analysis using weekly stock returns to check for robustness. As in the case of daily returns, SIC industries are only insignificantly better at explaining weekly industry returns (adjusted $R^2 = 0.516$ vs. 0.506).
<table>
<thead>
<tr>
<th>Industry Group</th>
<th>Constituents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oil and gas extraction</td>
<td>APA, APC, BJS, CHK, COG, DNR, DO, DVN, EOG, ESV, HAL, NBL, NBR, OXY, PXD, RDC, RRC, SLB, SWN, XTO</td>
</tr>
<tr>
<td>Food and kindred products</td>
<td>ADM, CAG, CCE, CPB, DPS, GIS, HNZ, HRL, HSY, K, KFT, KO, MJN, MKC, PBG, PEP, SJM, SLE, STZ, TAP, TSN</td>
</tr>
<tr>
<td>Chemicals and allied products</td>
<td>ABT, AMGN, APD, AVP, BIIB, BMY, CELG, CEPH, CF, CL, CLX, DD, DOW, ECL, EL, EMN, FMC, FRX, GENZ, GILD, HSP, IFF, JNJ, KG, LIFE, LLY, MON, MRK, MYL, PFE, PG, PPG, PX, SIAL, WPI</td>
</tr>
<tr>
<td>Industrial machinery &amp; equipment</td>
<td>AAPL, AMAT, BDK, BHI, CAM, CAT, CMI, DE, DELL, DOV, EMC, ETN, FLS, FTI, HPQ, IBM, IGT, ITT, JBL, LXX, MMN, NOV, PBI, PH, PLL, ROK, SII, SNDK, SWK, TDC, VAR, WDC</td>
</tr>
<tr>
<td>Electrical and electronic equipment</td>
<td>ADI, ALTR, AMD, APH, BRCM, CSCO, EMR, FSLR, GE, HAR, HRS, INTC, JDSU, LLI, LLTC, LSI, MCHP, MOLX, MOT, MU, NSM, NTAP, NVDA, NVL, QCOM, QLGC, TLAB, TXN, WFR, WHR, XLNX</td>
</tr>
<tr>
<td>Energy</td>
<td>APA, APC, BHI, CAM, CHK, CMI, COG, DNR, DO, DVN, EOG, EP, EQT, FTI, HAL, NBL, NBR, NOV, OXY, PXD, RDC, RRC, SII, SLB, STR, SWN, WMB, XTO</td>
</tr>
<tr>
<td>Pharmaceutical and food products</td>
<td>ABT, ADM, ADP, AET, AFL, AGN, AMGN, BAX, BDX, BIIB, BMY, CAG, CAT, CELG, CEPH, CI, CL, CLX, CPB, CVH, DD, DOW, EMR, GILD, GIS, HNZ, HSY, HUM, INJ, KG, KFT, KO, LLY, MMM, MRK, PFE, PGE, PG, SJM, SLE, SRCL, SYK, UNH, WHR, WLP, WM</td>
</tr>
<tr>
<td>Biotechnology</td>
<td>AVP, CCE, CF, CRM, CTSH, DHR, DPS, EL, FLS, FSLR, HSP, ISRG, LIFE, MIL, MJN, PLL, TMO, WAT</td>
</tr>
<tr>
<td>Internet</td>
<td>AAPL, ADBE, AKAM, CPWR, GOOG, IPG, JBL, MCO, MOT, MSFT, MWW, ORCL, YHOO</td>
</tr>
<tr>
<td>Computer software and hardware</td>
<td>ADI, ADSK, ALTR, AMAT, AMD, BCR, BRCM, CA, CSCO, CTXS, DELL, EBAY, EMC, HPQ, IBM, INTC, INTU, JNPR, KLAC, LLTC, LSI, MCHP, MU, NOVL, NSM, NTAP, NVDA, QCOM, RHT, SNDK, SYMC, TER, TXN, VRSN, XLNX</td>
</tr>
<tr>
<td>Instruments &amp; related products</td>
<td>Medical technology</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>--------------------</td>
</tr>
<tr>
<td>A, AGN, BAX, BCR, BDX, BSX, CFN, COL, DHR, EK, FLIR, ISRG, KLAC, MDT, MIL, PKI, RTN, STJ, SYK, TER, TMO, WAT, XRAY, XRX, ZMH</td>
<td>BSX, DE, FRX, GE, GENZ, KG, MDT, MON, MYL, PPG, STJ, STZ, TAP, TSN, WPI</td>
</tr>
</tbody>
</table>

| Adj. $R^2$ | 0.546 | Adj. $R^2$ | 0.540 |

Notes: This table shows the 10 largest SIC groups by number of sample companies and the 10 groups resulting from a cluster analysis of these firms based on our measure of relatedness. We can see that the network of joint mentions generates plausible industry groups, which differ from traditional classification schemes. These groups have the same explanatory power for industry returns as the well-established SIC.
to investigate in how far the investor perception of strategic peer groups can help determine the relatedness of firms and be used to identify homogenous groups of firms. We have employed a unique dataset and leveraged methods from social network analysis to compute the frequency with which pairs of companies are mentioned together in news-related messages in an online stock community. We explored the following three research questions: first, whether our measure of relatedness can help explain the comovement of stocks, second, whether the relationships can help identify a firm’s strategic peers, and, third, whether meaningful industry groups can be extracted from the network of relationships.

With respect to our first research question, we conclude that news-based relatedness can help explain the comovement of stock returns and adds explanatory power for comovement to traditional measures of proximity such as SIC classifications. An increase in Relatedness over time quickly reflects an increase in stock comovement. Our results support the theory that information associated with a set of firms is an indicator of relatedness and the comovement of their stocks (e.g., King, 1966). As far as the identification of strategic peers is concerned, we find that relatedness can define meaningful strategic peer groups, which exhibit a substantial level of stock comovement. This approach provides an alternative to clustering firms based on a multitude of firm-level dimensions frequently employed by strategy researchers (e.g., DeSarbo et al., 2009; Dranove et al., 1998). Finally, our results suggest that news-based industry groups can be a plausible alternative to traditional industry classifications. Overall, our results support the view that a news-based measure of firm relatedness can be used to reliably define strategic peer and industry groups. These findings confirm previous exploratory evidence that a network perspective of the stock market can be used to derive groups of stocks that are homogeneous with respect to traditional industry classifications (e.g., Bonanno et al., 2004; Mantegna, 1999; Omnela et al., 2003) and may be used to design stock indices (Tse et al., 2010).

Our innovative approach to measure firm relatedness has multiple advantages over traditional methods of industry classification: First, our approach is transparent and does not depend on the arbitrary assignment of companies by experts or the census bureau. Our method leverages the insights of hundreds of investors who associate one firm with another. Our study focuses on the context of stock microblogs which offer a large number of news items that users associate with one or several firms. However, our approach is not limited to this particular data source and could easily be adapted to other media such as newspaper articles or financial news wires. Second, our definition of relatedness provides a quantitative measure of the proximity of firms. In contrast to the nominal categorization of standard industry classifications, it offers a continuous measure of relatedness for all pairs of companies. Third, news-based relatedness can be calculated for various time horizons and may thus reflect changes in firm relatedness quickly compared to fairly stagnant manual classification schemes such as the SIC.

There are various promising applications for the accurate measurement of firm relatedness. As laid out in our section covering related research, both academics and practitioners spent considerable time and effort on the delineation of industry groups (e.g., Fama and French, 1997; Fan and Lang, 2000; Rammath, 2002). Research and practical applications include the design of meaningful stock indices that are composed of a coherent group of companies and the calculation of industry-specific costs of capital. In addition, financial professionals, such as analysts and investors, may find multiple applications ranging from the identification of relevant competitors to the selection of diversified portfolios (Chan et al., 2007).

5.2. Limitations and further research

Our study does not come without limitations. First, our quantitative analysis of firm relatedness was mainly limited to the similarity of stock returns. Stock returns are largely driven by new information and thus lend themselves in particular to our news-based definition of relatedness. However, there are many other dimensions of relatedness and future research should investigate in how far a news-based industry classification can delineate homogenous groups with respect to these dimensions (e.g., accounting figures, overlap in customer base, coverage by the same analyst, ownership by the same investors, etc.). Second, our method to delineate industry groups may prove to be helpful to the strategic groups literature. Dranove et al. (1998) have pointed to a shortcoming in much of the existing research, which groups firms based on firm-level factors that do not capture strategic interactions. Our approach to delineate industry groups through a network of relationships may address this shortcoming and serve to capture the patterns of interactions within industries. Third, we drew on basic
methods from social network analysis. However, there is a rich repository of analytical resources in the realm of social network analysis that may help refine the delineation of industry groups. For instance, while we have defined strategic peers to be the companies that are most closely related to a particular company directly (i.e., a direct link or joint mention had to be present) to be included in the investigation of Relatedness and comovement on the level of company pairs, other methods (such as so-called cliques or k-plexes) may be better suited to the analysis of larger networks. Exploring these methods may help identify even more relevant subgroups of strategic peers. Finally, our analysis focused on the mere joint mentioning of two companies and it is easy to imagine that additional firm-specific attributes may allow for more granular analysis. However, as one of the first studies in this area, our method sought to establish a general link between company associations in the news and industry classifications. Yet, it opens the door to more nuanced analysis of relatedness. Leveraging information regarding the context in which two stocks are associated with each other (e.g., a news item referring to a joint venture vs. a legal action) may enable us to define the type of relatedness and the competitive relationship in further detail (see Gourley, 2011).

References


