Competition among Virtual Communities and User Valuation: The Case of Investor Communities[1]

Bin Gu[*]
Phone: 512-471-1582
E-mail: Bin.Gu@mccombs.utexas.edu

Prabhudev Konana[*]
Phone: 512-471-5219
E-mail: Prabhudev.Konana@mccombs.utexas.edu

Balaji Rajagopalan[**]
Ph:(248)-370-4958
E-mail: rajagopa@oakland.edu

Hsuan-Wei Michelle Chen[*]
E-mail: Hsuan-Wei.Chen@phd.mccombs.utexas.edu

* Department of Information, Risk, and Operations Management
Red McCombs School of Business
CBA 5.202 ; B6500
The University of Texas at Austin
Austin, TX 78712
Fax: (512) 471-0587

** Department of Decision and Information Sciences
School of Business Administration
Oakland University
Rochester, MI 48309

[1] Prabhudev Konana and Balaji Rajagopalan acknowledge support from National Science Foundation's Information Technology Research grant # IIS-0218988 and IIS-0219107, respectively. The authors acknowledge initial help from Chan-Gun Lee (researcher at Intel), Dharmaraj Karthekeyan, and Matt Wimble, and feedback from participants in AIS conference, 2004.

Competition among Virtual Communities and User Valuation: The Case of Investor Communities

ABSTRACT

Virtual communities are becoming a significant source of information sharing for consumers and businesses. This research examines how users value virtual communities and how virtual communities grow and compete with each other. In particular, the nature of trade-offs between network size and information quality, and the sources of positive and negative externalities are examined. We address these issues based on over 600,000 postings from three large virtual investing-related communities (VICs) for 14 different stocks. We developed an algorithmic methodology to process textual data and to categorize messages as *noise* or *signal* to evaluate information quality. The results provide interesting insights into competition among virtual communities. There is a trade-off between network size and information quality. We find support for the hypothesis that the value of VICs increases with useful postings – demonstrating positive network externalities – but the marginal contribution decreases with the size. On the contrary, the cost associated with using VICs increases with size, while the marginal cost increases with each additional posting indicating negative externalities. The negative externality due to consumer information processing thresholds leads to a bounded network size. Our analysis also suggests that the community network size depends on two important dimensions: the degree of integrated service offerings (e.g., email services, complementary information) and characteristics of the context (e.g., stock characteristics such as speculative or stable stock). The contributions of the study include extending our understanding of the virtual community evaluation by consumers, the exposition of role of network externalities in virtual community networks, and the development of an algorithmic methodology to evaluate the quality of textual data. The results provide useful guidance for practice on the design and control of VICs.

KEYWORDS: Network economics; Computer-mediated communication and collaboration; Virtual communities; IT diffusion and adoption

1. Introduction

Virtual communities provide unprecedented opportunities for consumers to interact with each other and to influence how businesses function. BusinessWeek magazine in its cover story (June 20, 2005) noted "*companies are using Internet-powered services [virtual communities] to tap into the collective intelligence of employees, customers, and outsiders, transforming their internal operations*." However, attracting large number of consumers into a network to tap into *collective intelligence* will also result in increased information processing costs, generating negative externalities for participants and the community owner. To understand the dynamics of virtual communities, this paper addresses three questions: (a) How do consumers value online community networks? (b) How do competing networks differ in their value propositions? and (c) How do users differ in their preferences for, and their choices of, virtual communities? We address the above questions in the context of virtual investing-related communities (VIC, henceforth). In addition, to aid in the analysis of large unstructured text data, the study describes an algorithmic methodology for processing the context-specific virtual community interactions (i.e., textual data) and to understand their value.

The value of a virtual community depends largely on the contributions from users in terms of time, resource and knowledge (Butler 2001). Virtual communities may exhibit positive network externality. That is, the more members a virtual community has, the higher the value for all the participants, and the more users it is likely to attract in the future; a phenomenon extensively studied in both academic research and business practice (e.g., Katz and Shapiro, 1985; Asvanund et al. 2004). An increasing number of businesses are leveraging this power of virtual communities. The success of eBay, for example, is attributed to such network effects. Software businesses are facilitating the creation of virtual user communities to support after sale services. The open source development model is built on the sustainable strength of online communities. Thus, competition among firms is

also a competition to attract individuals to their communities.

One of the most widely used and known communities over the Internet are the VICs for individual investors to exchange information (Das and Chen, 2001, Antweiler and Frank, 2004, Wysocki, 1999; Konana and Balasubramanian, 2005). VICs offer a unique opportunity to observe the evolution of, and competition among, multiple communities (e.g., Yahoo!, Raging Bull, Silicon Investor). The investors' participation in various VICs depends on the community characteristics such as message quality, community size and volume, and their information processing cost thresholds[2]. VICs, thus, offer an excellent context to examine the research questions of interest in this study.

First, we examine how users value VICs. Members of a virtual community receive value primarily from two sources: *Integrated* services provided by the community, and *information sharing* with community members. Examples of integrated services include free email accounts, business news, and stock charts offered by investment communities (e.g., Yahoo!). The value from these services is fixed for each member regardless of the community size. The second source of value is the opportunity to seek information actively (i.e., by positing messages), or passively (i.e., by only reading messages, but not posting) with fellow community members. This value is influenced by both positive and negative network externalities. A user's value of a network is an increasing function of its size as more postings lead to more information and higher valuation of the network. This represents positive network externality. However, the network effect is limited by the *quality* of messages – measured by noise and signal levels described later – which results in negative externalities due to higher information processing costs to the members. Such costs increase exponentially with community size. Thus, incorporating quality measures into consumer choice model enables us to assess network value and information processing costs simultaneously.

[2] Users with high information processing cost thresholds imply that they have higher tolerance for noise and low overall costs for processing information, and vice-versa.

Second, we explore how the multiple sources of value for VICs drive different competitive strategies. For example, Yahoo!Finance postings are skewed toward higher levels of noise (Das and Chen, 2001), but the portal offers a rich array of integrated services such as historic stock and financial performance data that bring value to its users. In contrast, other smaller forums like the Raging Bull have lower levels of integrated services, but offer a high quality information exchange network for a limited number of stocks. Clearly, the size of these networks makes it easier to monitor postings and reduce noise levels.

Third, we study how the differences in the positioning of different VICs create self-selection among consumers. Consumers with higher valuation for quality information and lower information processing cost threshold are more likely to be attracted to smaller communities with higher quality postings. Likewise, consumers with lower valuation for high quality information and higher information processing cost thresholds are more likely to join larger communities with lower quality postings. Indeed, this latter group appears to value the quantity of information more than the quality.

Based on the differentiation of VICs and consumer self-selection, we argue that the competition between VICs is not a simple competition for size. Instead, VICs make trade-offs between size and quality and so are the consumers, resulting in the coexistence of multiple VICs. The traditional network externality literature has largely focused on competition through penetration pricing and other means to increase network size (e.g. see Katz and Shapiro 1985). We extend this literature by showing empirical support for our hypotheses that such competition is a remarkably differentiated one. VICs not only compete for size, they are differentiated based on noise levels, network size, integrated services, and characteristics of consumers attracted to the network.

In examining consumers' valuation of virtual investment communities, this study also contributes to advancing techniques to analyze posting quality by developing an algorithmic

methodology. The algorithmic methodology provides a feasible method to analyze the large volume of message postings. Online message posts are notoriously noisy and their quality is difficult to quantify. For example, community users flame other members, use abbreviations or jargons. Building upon prior work in the automated extraction of message board posting sentiment (e.g., Das and Chen 2001) we develop generic classifiers and a framework that can be applied to different domains to classify messages according to relevance of the content (e.g., signal, noise and neutral).

The remainder of this paper proceeds as follows. Section 2 provides background and literature review on virtual communities and financial message boards. In Section 3, we present the empirical model. Section 4 discusses the data sources and research method. In Section 5, we discuss the implications, limitations, and potential generalizations. Section 6 presents concluding remarks and future research directions.

2. Background and Literature Review

2.1 Virtual Investing-related Community

Virtual communities based on message boards and chat rooms have emerged as important social networks with implications for economic activities. This is particularly true in the context of VICs. They provide platforms to seek, disseminate, and discuss stock-related information. VICs also offer consumers a broad range of comprehensive online financial services essential to managing one's financial interests through an integrated service.

There are numerous VICs including the major ones offered by Yahoo!Finance, Raging Bull, Silicon Investor, Motley Fool, and MorningStar. While some communities were specifically created within existing financial networks (e.g., Raging Bull or Morning Star), others like Yahoo!Finance are part of a larger set of services. Information from these VICs spreads rapidly to thousands of investors within and across virtual communities, and has the potential to influence the attitudes and

decisions of these investors (Das and Chen, 2001). VICs may also help spread rumors and enable explicit manipulation of stock price. For example, in November 1999, false information about NEI Webworld Inc. posted in Internet chat rooms pumped the stock up from 15 cents to more than $15 within few hours (SEC Litigation number 16620, July 6, 2000). During 2000, Emulex Corporation's stock value plummeted 62% within few hours when an individual "knowingly and willfully" (as stated by the Federal Grand Jury during trial) released a fraudulent, negative news release regarding the company through discussion forums. Thus, VICs offer an interesting context to study how users value such forums and how their valuation influences the growth of these communities.

Researchers have focused efforts to understand why investors participate in various communities, how such information is interpreted, and how interactions impact financial markets. For example, Antweiler and Frank (2004) showed that activities on message boards predict market volatility and, while economically small, the effect on stock returns is statistically significant. Using self-reported sentiments (e.g., buy, sell, hold) Tumarkin and Whitelaw (2001) also found similar results. Bagnoli et al. (1999) found that the unofficial whispers from VICs and other Internet websites are often more accurate than analysts in predicting company earnings. Wysocki (1999) analyzed message postings and relationship to stock characteristics (e.g., speculative stock) and found that stocks with significant growth uncertainty attracted larger message postings. The study also found that the number of postings was related to stock price volatility and trading volume. On further analysis, Wysocki noted that a majority of the messages are posted by a small fraction of active participants who respond to other participants enthusiastically and post messages even after trading hours.

A few studies have investigated the size of VICs and information quality. Raging Bull, one of the smaller VICs, provides excellent online posting quality by offering extensive monitoring and online screening tool that allows any investor to establish a ban list to filter messages. However,

Yahoo!Finance, the largest VIC, is known to have higher noise (lower quality), but provides extensive complementary services. However, we are not aware of any studies that investigated the trade-off between integrated services on size and quality and the role they play in consumer valuation of VICs.

In summary, prior work has largely focused on the relationship between VIC postings and stock price with few additional parameters, such as impact of news release and activity levels. But, there is little research addressing the question of how investors (members) value VICs (online communities) or how various VICs differentiate to attract and/or retain investors. By addressing fundamental questions relating to user valuation of VICs and the competition among them, this study attempts to examine these issues.

2.2 Related Literature Review

At its core, virtual communities are communication systems that allow for many-to-many communications within a collection of users (Butler 2001). Sproull and Kiesler (1990) show that a virtual community can fundamentally change the ways people interact with one another, and lead to more flexible and efficient organizations. It has also changed the dynamics of the relationship between a business and its customers. Chevalier and Mayzlin (2003) note that empowering consumers through online consumer communities for product reviews and advise has causal impact on their purchasing behavior and put virtual communities at the center of online business strategy. These communities lower information asymmetry that has been historically exploited by firms.

In general, the larger a virtual community is, the more valuable it becomes to the provider – an issue studied extensively in the network externality literature (e.g., Katz and Shapiro 1985; Farrell and Saloner 1986). The network externality effect has been studied in a number of contexts including adoption and diffusion of products and switching costs (Katz and Shapiro 1986, Riggins,

Kriebel, and Mukhopadhyay 1994)

Butler (2001), however, suggested that large communities also have their disadvantages due to information processing costs. After studying 1,066 email-based listserv, he found that large communities are associated with high volume of emails and more variations in contents. This means users need to read more emails and the contents of these emails are less likely to be relevant to their interests, resulting in substantial processing cost to users – a source of negative externality. In a similar vein, Asvanund et al. (2004) consider the negative externality in the context of peer-to-peer (P2P) file download network. Their findings suggest that increase in the size of the virtual network creates substantial congestions on the network and may decrease the value of the network to individual users. These findings on the relationship between community size and intermediate cost variables can be interpreted to indicate that a larger community may not necessarily be associated with higher value to its membership.

One of the reasons that higher value does not follow from a larger community is because information processing costs increase with information overload (caused by increased membership). This impacts the types of consumers who will be attracted to a community and the desirable size of a community. *Information overload (IO)*, a condition in which the amount of data an individual must process is larger than that individuals' capacity for processing information, is widely recognized as a scenario stemming from the easy access to a wide network like the Internet (Fischer and Stevens, 1991). This condition of overload could impact any or all of the cognitive processes (Fournier, 1996) such as *attention*, *storage*, and *retrieval* (Lindsay and Norman, 1977). For example, IO can result in loss of information by increased *attention* to new information. In addition to IO, today's easy and inexpensive distribution of information also causes *data smog (*Shenk, 1997) – the rise in low quality information which could also influence the cognitive mechanisms in information processing by making it difficult for the user to pay attention to relevant information. Virtual community boards

present both the overload and data smog problem as evident from the volume and quality of postings. Lindsay and Norman's (1977) human information processing model elaborates on how humans process information to build knowledge.

Drawing from and building on prior research on VICs, network externalities and information processing costs, this study seeks to explicate the relationships among network size, posting quality and integrative services.

2.3 Automated Text Mining

Analyzing text messages and classifying the emotive content of messages posted on VICs poses several challenges. The main challenge is arriving at a common understanding or interpretation of the content. Messages can be *noise*, *signal*, or *neutral*. The subjective nature of some messages may lead to disagreement among readers as to whether a given message is truthful, important, or reliable. For instance, it is common to find messages with postings "XYZ sucks" (XYZ refers to some stock symbol) without any elaboration. While it appears such messages would belong to the noise category, the posting could be argued as providing some useful information. Such problems have also been encountered in previous text classification research. Foltz et al (1999) used classifiers for automated grading of student projects where there was disagreement among three human graders with a correlation of only 0.73.

Another challenge with community messages is that they generally do not observe proper grammatical rules and spelling, and therefore, readability analysis approaches used in isolation are less likely to be effective. Users frequently use abbreviations for many words (e.g., "u" for "you", "L8er" for "Later") and generally ignore spelling errors. The automated content analysis is compounded by semantic differences based on the context. For example, in drug companies much of the conversation centers on potential products that are "in the pipeline" of research and development. However, in the energy sectors the term *pipeline* takes on a very different connotation.

Each industry and company has a rather different combination of words that are often used within relevant conversations. Thus, finding a common set of words for classification is challenging.

In developing our automated classifier, we build on the approach used by Das and Chen (2001). Their method uses multi-algorithmic technique to classify messages based on sentiments: spam messages or messages that are neither bullish nor bearish are considered neutral, while bullish or bearish on a particular stock were classified as positive or negative, respectively. Their study developed stock-specific classifiers for each stock discussion board to take into account the unique characteristics of the postings. While we borrow the underlying approach, there are important differences. First, different from Das and Chen (2001) study, we develop generic classifiers that can be applied to a broad range of virtual investing-related postings. Second, we implement a decision tree classifier based on readability analysis (Foltz, Laham, & Landauer 1999) to categorize messages. Third, we apply evolutionary computing methods (Holland, 2000) to induce classification rule sets.

3. Empirical Model and Hypotheses

3.1. Modeling Valuation of VICs

VICs provide two sources of value to investors (i.e., consumers): 1) integrated services such as email, news, stock quotes, financial analysis, company profile, analyst reports, etc; and 2) information posted from other investors in the discussion forum.

We denote a_i the value of integrated service provided by VIC *i* to investors. Value from integrated services within each VIC is the same for all investors. However, this value is different across VICs since each community provides different levels of integrated services.

The value of discussion boards depends on member participation and the quality of information (i.e., less noise and more signal) and it varies overtime. VICs have some influence on quality by filtering noise (e.g., by deleting profanity) or hiring moderators. The value of a network

depends on the number of useful postings; however, the cost to users depends on the total number of postings in a network.

For the empirical model, we denote the total number of postings for stock board *j* on VIC *i* as n_{ij}. We denote q_{ij} to be the probability that a given posting contains useful information. The value of the discussion board, therefore, is a function of $n_{ij}q_{ij}$, the number of useful postings in the stock board. The more useful postings a board contains, the more value it provides to its users[3]. However, like all economic resources, we posit that postings on VICs exhibit diminishing marginal value. This indicates value is a concave function of the number of useful postings. We therefore adopt a quadratic form for the value function, i.e., the value from a stock board with $n_{ij}q_{ij}$ useful postings can be expressed as:

$$b_i\left(n_{ij}q_{ij}\right)+c_i\left(n_{ij}q_{ij}\right)^2 \qquad (1)$$

Here b_i stands for the marginal value of the useful posting. c_i represents the change in marginal value as the number of useful postings increase. The diminishing marginal value suggests that c_i is negative. This is consistent with the finding of Asvanund et al. (2004), who show that resource availability increases with the number of users in a P2P network, but the marginal contribution of each additional user decreases with the network size.

Thus, the total value V_{ij} received by investors visiting *i*'s discussion board on stock *j* is:

$$V_{ij} = a_i + b_i\left(n_{ij}q_{ij}\right)+c_i\left(n_{ij}q_{ij}\right)^2 \qquad (2)$$

We build upon Asvanund et al. (2004) results by explicitly estimating the value of online postings in utility terms. Our approach explores the difference between useful postings which affects the value and the total postings which affects the cost. By doing so, we will be able to quantify both the benefits and costs in terms of consumer valuation and to derive the optimal

[3] Cost of reading these postings will be considered separately.

community size. We first state the hypotheses on consumer valuation as follows:

Hypothesis 1a: The value of VIC increases with the number of useful postings available on the discussion board.
Hypothesis 1b: The value of a VIC is concave with regard to the size of useful postings. That is, the marginal contribution of each additional useful posting decreases with the size of the discussion board.

While bringing benefits to investors, the use of VICs is not costless. Das and Chen (2001) show that a large fraction of postings consist of noises and rumors. Investors often need to process hundreds of postings and ferret out useful information. Such large volume of postings contributes to information overload, which affects investors' cognitive ability to analyze postings and reduces their attention to useful information (Lindsay and Norman 1977; Shenk 1997). The cost of information overload depends on the size of the community. When the community size is large, the information processing costs increase substantially. As a result, investors are more likely to end participation in communities when their size increases above a particular threshold (Butler 2001; Jones et. al. 2004).

Thus, the cost of using VICs consists of two components. The first part is the opportunity cost, d_i, of reading postings, which increases linearly with the number of postings. The second part is the cost of information overload, e_i, which is increasing and convex in the number of postings. The combination of the two costs suggests that we can use a quadratic form of cost functions:

$$C_{ij} = d_i n_{ij} + e_i n_{ij}^2 \qquad (3)$$

The cost function is similar to the one used by Asvanund et al (2004), which shows that network congestion and download time increases with the number of users in a P2P network, and the marginal cost increases with the network size. We build upon their results by simultaneously estimating the value and cost of using online VICs in terms of consumer utility. By linking cost and value to consumer utility, we can quantify the information processing costs in terms of consumer

utility and show how consumers make trade-off between costs and benefits. We state the hypotheses on information processing as follows:

Hypothesis 1c: The cost of using a VIC increases with the number of postings in the discussion board.
Hypothesis 1d: The cost of using a VIC is convex with regard to the size of the postings. That is, the marginal cost of an additional posting increases with community size.

3.2. Competition between VICs

The second issue explored in this study is how VICs compete against each other for investors. Each VIC offers a discussion board for each publicly traded stock. The investors, therefore, face a choice among different discussion boards on the same stock. The competition between these discussion boards depends on the net utilities received by consumers. Given the value and cost discussed in the previous section, the utility a consumer receives from the discussion board on stock *j* in VIC *i* is the difference between value received and cost incurred:

$$U_{ij}\left(n_{ij}, q_{ij}, a_i, b_i, c_i, d_i, e_i\right) = V_{ij} - C_{ij} = a_i + b_i\left(n_{ij}q_{ij}\right) + c_i\left(n_{ij}q_{ij}\right)^2 - d_i n_{ij} - e_i n_{ij}^{\ 2} \qquad (4)$$

To increase the net utility received by consumers, VICs have two options: they can either provide better integrated service, or provide discussion boards with higher quality. The two options, however, have different cost implications. Better integrated services involve collecting extensive company news, purchasing analyst reports and company SEC filings, and setting up online systems to distribute the information. These services all involve substantial fixed costs, but the marginal costs of providing this service to one more user is negligible. Thus, VICs with a large membership base benefit from economies of scale. On the other hand, improving posting quality requires active monitoring services which involve little fixed costs, but substantial variable costs for each additional user, which makes quality improvement on large VICs more expensive. The difference between the cost structures of the two options suggests that large VICs have cost advantage in providing better integrated services, while small VICs have cost advantage in providing

higher quality discussion boards. Thus, large VICs are more likely to provide better integrated services, indicating a_i, a positive function of the community size of VIC i, N_i. We rewrite a_i as $a(N_i)$. Likewise, the above discussion suggests that the quality q_{ij} provided by VICs is a negative function of the community size N_i.

The effect of VIC size on integrated services and posting quality not only holds across VICs, but is also evident over time. When Raging Bull first started, it offered no integrated services but only a collection of stock discussion boards. With increase in size over time, it added integrated services such as stock quotes, charts, news, and even email accounts. At the same time, posting quality on Raging Bull declined. For example, the signal to noise ratio of Raging Bull's Dell discussion board was 47% in 1998 with a size of about 30 postings per week, but reduced to 42% in 1999 when its size increased to 60 postings per week. It then rebounded to 56% in 2000 as Dell board's size decreased to 20 postings per week. Thus, we formally state the hypotheses as follows:

Hypothesis 2a: The value of integrated services offered by VICs increases with their size.

Hypothesis 2b: The posting quality of VICs decreases with their size.

3.3. Consumer Self-Selection

Given that VICs are differentiated along their integrated services and their posting quality, it is natural that different types of consumers are attracted to different VICs. In particular, consumers with low information processing cost thresholds and high valuation for information gravitate towards VICs with high quality postings, and value the size less. On the other contrary, consumers with high information processing cost thresholds and low valuation for information would likely self-select VICs with large number of postings without regard to quality. That means, value factor b is a negative function of the community size, while cost factor d_i is a positive function of community size. We rewrite then as $b(N_i)$ and $d(N_i)$, respectively. We therefore have the following

hypotheses:

Hypothesis 3a: Investors with higher valuation of postings prefer higher quality VICs.

Hypothesis 3b: Investors with lower information processing cost thresholds prefer higher quality VICs.

Given Hypotheses 2 and 3, the consumer utility function of the stock discussion board in VIC *i* on stock *j* can be written as:

$$U_{ij} = a(N_i) + b(N_i)(n_{ij} q_{ij}) + c(n_{ij} q_{ij})^2 - d(N_i) n_{ij} - d n_{ij}^{\ 2} \qquad (5)$$

3.4 Empirical Model

Given the utility function above, the empirical estimation model can be written as follows:

$$U_{ij}(N_i, n_{ij}, q_{ij}; \beta) = \beta_o + \beta_1 \lg N_i + \beta_2 q_{ij} n_{ij} + \beta_3 \lg N_i n_{ij} q_{ij} + \beta_4 q^2{}_{ij} n^2{}_{ij} + \beta_5 n_{ij} + \beta_6 \lg N_i n_{ij} + \beta_7 n_{ij}^{\ 2} + \varepsilon_{ij} \quad (6)$$

Hypothesis 1a suggests that β_2 should be positive since the utility increases with the size of useful postings. Hypothesis 1b indicates that β_4 shall be negative as the utility increase is concave and the marginal value is decreasing. Hypothesis 1c predicts that β_5 will be negative as the cost is increasing with community size. Hypothesis 1d expects β_6 to be negative due to accelerating information processing costs as a result of information overload. Hypothesis 2a indicates that β_1 shall be positive as larger VICs prefer to improve their integrated services. Hypothesis 3a indicates that β_3 should be negative as consumers which high valuation of postings will prefer high quality but smaller communities. Likewise, Hypothesis 3b indicates that β_4 should be positive as consumers with lower information processing cost would prefer larger communities.

If we know consumers' utility for each of the three VICs, the above empirical model can be estimated by ordinary least-squares (OLS) regression. However, in practice, we only observe consumers' choices of the VICs. To analyze their choice decisions and to infer utilities of the VICs, we use multinomial logit (MNL) regression model. The underlying assumption of MNL model is

that consumers always choose the offer with the largest latent utility. The latent utility has two components: deterministic and stochastic components. The deterministic component is determined by attributes of VIC *i* with regard to stock discussion board *j* (N_i, n_{ij}, q_{ij}) as well as consumer valuation of these attributes (βs). The stochastic component is due to measurement errors and unobserved variations in consumer preferences (ε_{ij}). Given the VIC attributes and consumer preference, the probability P_i that a consumer chooses VIC *i* over other VICs is:

$$P_i = \frac{U_{ij}\left(N_i, n_{ij}, q_{ij}; \beta\right)}{\sum_k U_{kj}\left(N_k, n_{kj}, q_{kj}; \beta\right)}. \qquad (7)$$

Not all hypotheses are tested using the MNL model. Hypothesis 2b deals with the relationship between posting quality and community size. Given that we can measure posting quality directly, we run OLS regression of posting quality on community size.

$$q_{ij} = \delta_o + \delta_1 \lg N_i + \delta_2 n_{ij} + \zeta_{ij} \qquad (8)$$

Hypothesis 2b suggests that large VICs will be associated with low quality. Therefore, δ_1 in (8) above is posited to be negative. Likewise, within the same VIC, it is easier to maintain quality on smaller discussion boards than larger discussion boards. Therefore, δ_2 will be negative as well.

4. Data & Methods

4.1 VIC Selection

To empirically test our model, we collected a dataset of online postings from three large VICs: Yahoo! Finance, Silicon Investor, and Raging Bull from January 1, 1998 to January 10, 2002. These three VICs have been widely acknowledged as the leading investment communities during that time and used in other studies (Antweiler and Frank 2004). Each of the three VICs hosts thousands of discussion boards for various stocks.

As discussed earlier, Yahoo!Finance offers the most comprehensive integrated services for

individual stocks. Yahoo!Finance provides moderate monitoring of its investment community. Postings containing abusive, obscene, or commercial content can be removed and the investors who post such messages may be banned from the community.

Unlike Yahoo!Finance which started as part of Yahoo! portal, Silicon Investor started primarily as a VIC. At its inception in late 1995, Silicon Investor mainly targeted technology stocks and was well-known for its users who were largely employees of technology firms with inner working knowledge. Over the years, it has substantially broadened its coverage, but maintained its root as a virtual investment community for investors interested in high-tech companies. Silicon Investor charges investors an annual fee to post messages on its discussion boards, but allows anyone to read postings for free. The posting fee serves two purposes. It foremost serves as a revenue source in addition to the traditional revenue source of banner ads as in Yahoo!. Second, there is a control on the types of participants in the community.

Raging Bull is our third VIC. It opened later than Yahoo!Finance and Silicon Investor, but experienced significant growth after CMG's initial investment in September 1998. Raging Bull initially focused on small stocks, especially OTC stocks. Over time, it grew into the third-largest "stock talk" site with 2 million daily page views, surpassed only by Yahoo!Finance and Silicon Investor. A key differentiation of Raging Bull is its technology that can allow investors to screen out obscene or disruptive postings. Raging Bull lets investors to rate postings – a technique now adopted by other VICs – and allows users to ignore certain members of the community. It also has dedicated moderators to monitor discussion boards.

We collected message postings from a random sample of 14 stocks that were common to all the three communities on a weekly basis from January 1, 1998 to January 10, 2002. We relied on a stratified sample to cover stocks of different risk (i.e., beta) levels. Stocks such as General Motors (GM), General Electric (GE), IBM, Disney (DIS), McDonalds (MCD), and Microsoft (MSFT) are

widely held and less speculative. Stocks such as Brocade communications (BRCD), CMGI, JDSU, Inktomi (INKT), and eBay were more speculative with wide fluctuation in stock movement during the time period for which data was collected. Cisco (CSCO) and Dell were relatively less speculative, but exhibited significant growth. For each message the following attributes were acquired: message number (MsgNo), author, subject, posting date, posting time, and message content. The summary statistics are provided in Table 1.

For each stock, we then calculate the market share of the three VICs on a weekly basis. The market share is determined by number of postings regarding that stock on a particular VIC divided by the total number of posting on that stock during the week. The change of market share over time shows the flow and dynamics of competition among the VICs on that particular stock. We also calculate an alternative measure of market share by using unique number of posters instead of number of postings. The results using the alternative measure yielded qualitatively similar findings. In the interest of brevity, the reported results are based on the number of postings.

Our analyses also required membership size of VICs. It was not feasible to count all postings across all discussion boards within a particular VIC. We, therefore, looked at MediaMatrix data, a large collection of online click-stream activities for a random sample of approximately 10,000 Internet users. The data records every URL visited by a user and the time and duration of the visit. We can infer membership size by the number of times the user visited each of the three VICs, the cumulative duration of such visits, or the number of users who have visited the VICs. We experimented with these different measures and they all yield qualitatively the same result. In the analysis that follows, we use the number of visits as the measure.

4.2 Text Processing and Classification for Quality Assessment

A key derived measure in this study is the posting quality of individual stock discussion boards. Messages are classified into three categories: *signal*, *noise* and *neutral* (Rajagopalan et al. 2004) We

consider a message as a *signal* carrier if and only if it is relevant to the stock, and a discernable sentiment – positive or negative – is expressed toward the stock. A message is categorized as *noise* if the content is spam, flame, or completely unrelated to the message board topic. We consider a message to be *neutral* if the content relates to the stock in particular and/or the market in general with implications for the stock, but with no specific signal such as buy, hold, or sell presented.

Figure 2 provides an algorithmic methodology derived from Das and Chen (2001). We used Network Query Language (NQL) (for details, see http://www.nqltech.com/nql.asp) to download messages from three message boards, which were then fed into the five classification algorithms to classify them into Noise, Neutral, or Signal according to Figure 2. The methodology to extract relevance was carried out in three steps: classifier development, testing & validation, and application.

In the *first stage*, five relevant extraction algorithms widely used in text categorization (Das, S. S. Martinez-Jerez, A. and Tufano, P., 2004; Antweiler, W. and M. Frank, 2004) were used: Lexicon-based Classifier (LBC), Readability-based Classifier (RBC), Weighted Lexicon Classifier (WLC), Vector Distance Classifier (VDC) and Differential Weights Lexicon Classifier (DWLC) (See Appendix for details on the classifiers). A sixth classifier combining the outputs of each of the five classifiers was designed and implemented to provide the final categorization. Two supplementary databases, *lexicon* and *grammar rule set*, were developed to support the classification algorithms.

The domain-specific lexicon is a set of frequently occurring keywords constructed based on an extensive examination of a random subset of messages in each of the categories – Noise (C_1), Neutral (C_2), and Signal (C_3). The lexicon aids in the classification of message postings into one of the three categories of interest and is used by four classifiers (LBC, WLC, VDC and DWLC). A sample of the keywords is displayed in Table 8. The grammar rule set used by VDC, different from individual keywords, is a vector representation of a set of words occurring together (a sentence) representing a category of interest. For example, consider the grammar rule set representing the

category C_3 - "This is definitely a buy and quite possibly more so than Compaq consider 1-5 year growth numbers for each company". Several such sequences representing the categories constitute the grammar rule set for each category. When a new message posting is evaluated for categorization, VDC will classify it as C_3 if the computed distance between the new message and the grammar rule set for C_3 is minimal. The grammar rule set was also constructed based on a manual analysis of one hundred messages from all categories.

Once the classifiers were designed, the algorithms were trained using 300 messages. Human coding was carried out by two business graduate students who were informed of the definitions of the categories. A high degree of consensus emerged (inter-rater reliability > 90%) after a few pilot coding sessions. A small number of messages that were classified differently by the students were revisited and a consensus reached regarding their categorization. The training data set was deliberately kept small so as to prevent over-fitting the data (leading to poor out-of-sample performance), a common problem in classification problems. A simple majority voting of the five algorithms served as a way to combine the inputs from all the classifiers. The messages are then classified into Noise, Neutral, or Signal.

The *second phase* involved testing the classifier on a subset of the sample (800 messages as holdout) quarantined and not used for inducing the rule sets. Performance of the classifiers, based on classification ratio, was analyzed for their ability to classify a message into the three categories. Classification ratios of the algorithms were compared with prior attempts like Das and Chen (2001). For each input message tested, the statistical accuracy, the corrected classification ratio (*CR*), is computed as

$$CR = \frac{\#\text{ of messages correctly classified into their actual category } i}{\#\text{ of attempted classifications}}$$

The *final step* involved applying the classification method to the entire data set and computing the posting quality as the percentage of relevant postings. The key statistics of the three VICs are presented in Table 2. Our first experimental results are shown in the following. The results for all classifiers are presented in Table 9.

5. Results and Discussion

Overall, results indicate that users differ in their choices of virtual communities, and the virtual communities have different value propositions and strategies for users. We first report on the virtual community posting data characteristics to lay the foundation for hypotheses testing.

5.1 Descriptive Statistics

Table 1 shows the number of postings and the market share of the three VICs for each stock during the period January 1, 1998 to January 10, 2002. A close examination of the table reveals two interesting patterns. First, market shares of VICs vary substantially across stocks. For example, Yahoo!Finance received 92% of all postings with regard to BRCD, but only 16% of all postings on CSCO. Likewise, Silicon Investor's market share on BRCD is merely 1%, but it accounts for 59% of all postings on Dell. The substantial variation is more visible when we examine the distribution of market share across stocks. Figure 2 shows a histogram of the market share distribution, which indicates that a VIC either commands a dominant market share ($\geq$75%), or becomes negligible ($\leq$25%). Relatively few VICs are in the middle ground. This pattern is consistent with the presence of positive network externality.

The second pattern in the table is evident when we average the market shares of each VIC across stocks. The table shows that the average market share for Yahoo!Finance is above 50% and that for Silicon Investor is only 16%. But, if we look at the total number of postings for each VIC, we observe a different pattern. Silicon Investor has a total number of 23,804 postings, accounting

for 24% of all postings collected. This contrast is puzzling at first glance. How can a VIC with 24% of all postings have only an average market share of 16%? The answer is that smaller VICs like Silicon Investor attract fewer postings for volatile, low capitalization, stocks (e.g., BRCD, Cnet), but more postings for large stable stocks (e.g., DELL, DIS, MSFT). Consequently, they command a close-to-zero market share for small capitalization stocks, but only a mediocre market share for large stocks. A closer examination of our data suggests that this is indeed the case. For speculative stocks like BRCD or INKT, Silicon Investor accounts for only 1% market share; however, for large, widely held, stocks such as DIS, MSFT or DELL, Silicon Investor has a much larger market share.

These patterns indicate that users value VICs differently, and different VICs have different value proposition for various stocks. A smaller VIC can have disproportionately higher number of users for certain type of stocks. This phenomenon is consistent with the hypothesis that a VIC's marginal value decreases with its size and marginal cost increases with its size. As such, users of popular stock boards are more likely to split and join smaller VICs.

Table 2 provides summary statistics of the variables used in this study. Given our interest in the differences across VICs, summary statistics are presented for each of the three VICs. The table shows that Yahoo!Finance is the largest in terms of both the number of postings per week and the total number of monthly visits by users. Raging Bull follows closely as the second largest VIC despite its late start. Silicon Investors has the lowest membership size among the three, although it is one of the earliest VICs. Table 2 also shows that Yahoo!Finance's discussion boards contain a large percentage of noisy postings. According to our data analysis (see Section 4.2 for details) only 29% of Yahoo!Finance's postings are meaningful with the remaining 71% being noise. In contrast, the posting quality on Raging Bull and Silicon Investor is much better; more than half of the postings on these two VICs have signal (i.e., useful information). This is consistent with our arguments that it is more difficult and costly to maintain quality in large VICs.

5.2 Hypotheses Testing

We analyzed online investors' valuation of VICs using the multinomial logit model discussed in Section 3.4. Tables 3 and 4 present our results. Column 1 shows the estimates of the coefficients. Column 2 identifies the corresponding hypothesis for each of the estimates, and Column 3 shows whether the hypothesis is supported by the empirical result. Except for H1b, all our hypotheses were supported.

To provide intuitive illustrations, we present in Table 5 the utilities of discussion boards for the three VICs. As we mentioned earlier, the typical discussion board varies significantly across VICs: Yahoo's stock discussion boards are much noisier than other two VICs; however, it is larger than the others. Table 5 presents a hypothetical typical stock discussion board for each VIC. We then use the estimated coefficients to calculate the utilities of these stock discussion boards. We further break down the utility into three components: value from integrated services, value from discussion boards, and the cost from using discussion boards.

Our results suggest that discussion boards present substantial value to users. Table 1 shows that the coefficient on useful posting (i.e., β_2) is positive and significant. This suggests that consumers value useful postings and their utility increases with number of useful postings. This validates our Hypothesis 1a. But, surprisingly, the coefficient on the quadratic form of useful posting (i.e., β_4) is positive as well, which indicates that consumer utility from useful postings increases exponentially with the number of useful postings. This is not consistent with Hypothesis 1b of decreasing marginal value of useful postings. We suspect that the marginal value may tend to increase in the beginning when the number of useful postings are low, but after certain critical mass the marginal value tend to decrease. However, this phenomenon may not be observable in the data. Table 5 compares the value of discussion boards across the three VICs. It reveals that the values differ substantially. The value of Raging Bull's stock discussion board has the highest value (1.45)

due to its high quality and relatively large number of postings. On the other hand, Yahoo!Finance has the lowest value (0.73) despite being the largest stock discussion board.

We also find that online investors incur substantial information processing costs in using VICs which supports Hypothesis 1c. The coefficient of number of postings (i.e. β_5) is negative, which indicates that consumer utility of discussion boards decreases with the number of postings after controlling for the effect of useful postings. Moreover, the information processing cost is convex since the coefficient of the quadratic term of number of postings (i.e., β_7) in Table 3 is negative (note: the low value of the estimate is not a concern since the estimate is for the square of the number of postings). This is consistent with Hypothesis 1d and the information overload literature (e.g., Fournier 1996; Shenk 1997) where it is argued that the cost of processing information increases substantially with the volume of information due to *data smog*. Table 5 presents the cost of using discussion board across different VICs. We find that the cost of using a typical Yahoo!Finance discussion boards is the most significant – it accounts for 91% of the total value created by the discussion boards. This is largely due to the low quality postings on Yahoo!Finance's discussion boards. But, even for high quality VICs, such as Raging Bull, the cost of using the discussion board accounts for more than half (57%) of the value created.

The differences in attributes of VICs make them to pursue different strategies to attract online users. First, we find that integrated services play a major role in the competition among VICs and are especially important for large VICs (Hypothesis 2a). The coefficient on VIC size (i.e. β_1) is positive in Table 3 indicating that VICs are associated with more integrated services when they grow larger. This is because, with increase in size, VICs have the advantages of economies of scale and, therefore, can afford to spend resources on integrated services with the cost spread over thousands of users. On the other hand, small VICs do not enjoy such economies of scale and the cost (per user) of providing high quality integrated services becomes prohibitive.

But, small VICs have other advantages. A key aspect of small VICs is their low costs to maintain quality as suggested by Hypothesis 2b. The result presented in Table 4 supports this hypothesis. In essence, analysis shows that quality is negatively related to the overall VIC size as well as the size of the particular discussion boards. The coefficient of -0.12 on the log of membership size suggests that every time the VIC doubles its size, the signal ratio of the VIC reduces by 12%. This is consistent with our hypothesis that VICs have more ability to improve posting quality when they are small, but such an effect becomes increasingly insignificant when they grow larger.

Finally, the result suggests that users adopt a clear self-selection strategy in choosing VICs. The coefficient on the moderating effect of VIC size on number of useful postings (β_3) is negative. This suggests that users of large VICs value useful postings less than those of smaller VICs. Likewise, the coefficient on the moderating effect of VIC size on number of postings (β_6) is positive, indicating that users of large VICs incur less costs reading messages. Putting the two effects together, it suggests that users of large VICs have higher information processing thresholds in reading postings and less valuation for useful postings, supporting Hypotheses 3a and 3b.

For illustration purpose, we consider how users of different VICs would value a hypothetical discussion board with 400 weekly postings and a signal ratio of 40%. If the users of different VICs have the same characteristics, then their valuations for the discussion board would be the same. But, if consumers are differentiated, then their valuations will differ. Table 6 shows the result. The valuation suggests that consumers are heterogeneous across VICs. The value and cost are substantially different for exactly the same discussion board. We observe that Yahoo!Finance's users receive the lowest value from the discussion boards. They also incur the lowest costs for the discussion boards. This is again consistent with Hypotheses 3a and 3b that Yahoo!Finance attracts consumers with the higher information processing cost thresholds, but also with the lowest value for information. This result is due to consumer self-selection. Investors who have low information

processing cost threshold – that is, high cost reading postings – and high valuation for quality information will have more incentive to choose small VICs, which offer less postings, but high quality. On the contrary, online consumers who choose large VICs have high information processing cost threshold – that is, low cost reading postings – and low valuation for high quality postings.

5. Contributions

Although all VICs use similar technologies and attempt to attract members to their community, they are important differences between them, and they adopt distinct strategies to attract and retain users. The underlying cause of the differentiation is economics: a VIC cannot be all things to all users. Thus, VICs need to make a trade-off between size and quality. This paper addressed these tradeoffs, and discussed how virtual communities differentiate and compete. We also investigated the role of user valuation in networked communities. An understanding of these dynamics has important implications for research and practice.

5.1 Contributions to Research

This research makes three important empirical and theoretical contributions. First, the research empirically shows the presence of both positive and negative externalities in virtual communities. The negative externalities due to higher noise level (i.e., lower quality postings) put a cap on the network size. These results are consistent with negative externalities observed in P2P networks and the Internet due to congestion (e.g., Mackie-Mason and Varian 1994; Asvanund et. al. 2004). However, the source of negative externalities in VICs is from noise levels and higher information processing costs rather than congestion observed in other networks. This inherent trade-offs between size and quality forces a VIC to make a choice to compete effectively in the market.

Second, our study shows that the network size for VICs is dependent on the characteristics

of the stock (i.e., the context) and type of integrated (bundled) services provided. That is, different stocks exhibit different behavior in each VIC. The users take into account the type of stock, extent of integrated services, and the network size in their valuation. Thus, a network that is strong in one particular stock (i.e., context) does not necessarily carry over to other contexts within the same network. The availability of integrated service appears to complement participation in stocks that are volatile or speculative, where noise level is much higher. Thus, users appear to make trade-offs between integrated services and the quality/information processing costs due to higher noise. We are not aware of existing network literature providing these finer details in network formation.

Third, the study sheds light on how users value communities and shape the competition. We find that network size by itself is a cost factor and more postings require more information processing from consumers. What really makes a network valuable is the number of quality postings. By directly measuring both cost and value of messages in utility terms, we are able to show how consumers make trade-offs between the value and cost of using VICs. We show that the value of a useful posting is about three times the cost of reading a posting. Moreover, we show that the cost of using VICs exhibits increasing marginal costs, indicating there is a limit on network size.

5.2 Contributions to Practice

Our research results provide guidance for the design and control of VICs, since monitoring quality and information processing costs has implications to the network size. Our research shows that both size and quality are both important factors in consumer valuation of communities. In fact, the result reflects that size by itself is more a cost factor than a value factor for consumer valuation. It may be noted that size has implications for community owner due to higher revenue sources from advertisers, but not necessarily of value to users. Thus, virtual communities (especially smaller ones) need to focus on improving quality. While most VICs have simple mechanism to improve quality

and to lower information processing costs by providing capabilities to filter offensive users or language, or by posting sentiments, advanced methods can be incorporated to evaluate each postings to be noise, signal or relevant. A user can choose to filter any potential noise. The insights here can help develop software or software functionalities embedded within virtual communities that can improve quality and lower information processing costs.

The algorithmic methodology provided here is generic enough that it can be applied to a broad variety of applications. For example, for user reviews on Yahoo!Movie, this framework could be used to build predictive models based on the review critiques and ratings in order to see whether the reviews actually have corresponding reflections on the movie box office. This approach could also apply to Amazon.com book reviews and ratings. Forum comments could be further analyzed and classified as noise, signal, or neutral to help make qualitative feedback scores more precise and reliable. As the use of unstructured text becomes increasingly critical for research, the algorithmic methodology presented here could be an important component of data gathering strategies by researchers.

6. Conclusion

While a significant body of research examines the content of virtual community postings and motivations of these postings, much less is known about how these message boards grow and compete with each other and how positive and negative network externalities interplay in the competition. Our analysis fills this gap in the competitive analysis of virtual communities. We show that network externality is not the only factor in the competition. Quality is also an important factor that determines the value of VICs to consumers. Moreover, the level of network externality and quality depends on the strategic decisions made by virtual communities, which are remarkably differentiated in their value propositions and their attractiveness to consumers.

This paper has several limitations. First, we consider a relatively small set of stratified stocks. Although, the current list may not compromise the results, a more extensive study would increase the confidence in the results. However, due to data collection restrictions imposed by VICs (e.g., blocking of software agents) and finding balanced common stock, there is a limitation on adding more stocks. Second, our text processing and classification approach can be further refined to yield stronger results. However, since all VICs are equally affected, any improvement in the methodology further improves the significance of the results. Third, our logistic model implicitly assumes that VIC users are aware of the existence of the three VICs, while in reality not all online investors are aware of the presence of three VICs. This issue is similar to the presence of consumers' consideration set that has been widely considered in marketing research.

Our study also opens up several opportunities for future studies. First, future research can focus on the role of VICs on market efficiency. That is, research can explore the extent to which information signal correlates with market performance. Second, virtual communities in different domains have different characteristics. For instance, eBay appears to be dominating the virtual community for online auction, while peer-to-peer networks show substantial number of competing networks. It is worthwhile to find out why competition is feasible in some virtual communities, but not in others. Third, it will be interesting to understand user switching behavior from one community to another. Finally, replication of our study in other domains (e.g., user support communities) would be valuable.

References

Antweiler, W. and Frank, M. 2004. Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards," *Journal of Finance*, v59(3) 1259-1295.

Asvanund, A, Clay K., Krishnan, R., and Smith, M. 2004. An Empirical Analysis of Network Externalities in Peer-To-Peer Music Sharing Networks. *Information Systems Research* 15(2) 155-174.

Bagnoli, M., Beneish, M. and Watts, S. 1999. Whisper Forecasts of Quarterly Earnings per Share. *Journal of Accounting and Economics.* 28(1) 27-50.

Butler, B. S. 2001. Membership Size: Communication Activity, and Sustainability: A Resource-Based Model of Online Social Structures. *Information Systems Research* 12(4) 346-362.

Chevalier J. A. and Mayzlin, D. 2003. The Effect of Word of Mouth on Sales: Online Book Reviews. *NBER Working Papers* 10148, National Bureau of Economic Research, Inc

Das, S. R. and Chen, M. Y. 2001. Yahoo! for Amazon: Sentiment Parsing from Small Talk on the Web. *Proceedings of the 8th Asia Pacific Finance Association Annual Conference,* 2001.

Das, S. S., Martinez-Jerez, A. and Tufano, P. 2004. e-Information: A Clinical Study of Investor Discussion and Sentiment. November 2004. Accessed July 27, 2005 at http://www.people.hbs.edu/ptufano/einfo_Nov2004.pdf

Farrell J. and Saloner, G. 1986. Installed Base and Compatibility: Innovations, Product Pre-announcements, and Predation. *American Economic Review.* 76: 940-955

Fischer, G. and Stevens, C. 1991. Information access in complex, poorly structured information spaces. In *Reaching Through Technology: CHI 91 Conference Proceedings,* edited by Robertson, S. P., Olson, G. M., and Olsen, J. S.. ACM Press. New York, NY. pp 63-70

Foltz, P. W., Laham, D. and Landauer, T. K. 1999. Automated Essay Scoring: Applications to Educational Technology. In *Proceedings of EdMedia '99.*

Fournier, J. 1996. Information Overload and Technology Education. *Technology and Teacher Computing Annual* 1996

Hof, R. 2005. The Power of Us – Mass Collaboration of the Internet is Shaking Up Business. *Businessweek.* June 20, 2005.

Holland, J. 2000. Building Blocks, Cohort Genetic Algorithms, and Hyperplane-Defined Functions. *Evolutionary Computation* 8(4): 373-391

Jones, Q, Ravid, G. and Rafaeli, S. 2004. Information Overload and the Message Dynamics of Online Interaction Spaces: A Theoretical Model and Empirical Exploration. *Information Systems Research.* 15(2) 194-210

Katz M. L..and Shapiro, C. 1985. Network externalities, competition, and compatibility. *American Economic Review* 75: 424-440.

Katz, M. L. and Shapiro, C. 1986. Technology Adoption in the Presence of Network Externalities. *Journal of Political Economy.* 94(4) 822-41.

Konana, P. and Balasubramanian, S. 2005. The Social–Economic–Psychological Model of Technology Adoption and Usage: an Application to Online Investing. *Decision Support Systems.* 39(3) 505-524

Lindsay P. and Norman, D. A. 1977. *Human Information Processing*. Academic Press. New York, NY. 1977.

MacKie-Mason, J. K. and Varian, H. 1994. The Economic FAQs About the Internet," *Journal of Economic Perspectives*, Summer 8: 75-96.

Rajagopalan, B., Konana, P., Lee, C. and Wimble, M. 2004. Extracting Relevance from Virtual Investing-Related Community Postings. *Proceedings of Americas Conference on Information Systems* 2004.

Riggins, F. J., Kriebel, C. H., and Mukhopadhyay, T. 1994. The Growth of Interorganizational Systems in the Presence of Network e=Externalities. *Management Science.* 40(8): 984-998.

Shenk, D. 1997. *Data Smog: Surviving the Information Glut.* HarperEdge. New York, NY.

Sproull, L. and Kiesler, S.. 1991. *Connections, New Ways of Working in the Networked Organization* Cambridge, Massachusetts: The MIT Press.

Tumarkin, Robert, and Robert F. Whitelaw. 2001. News or Noise? Internet Postings and Stock Prices. *Financial Analysts Journal* 57: 41-51.

Wysocki, P.D. 1999. Cheap Talk on the Web: The Determinants of Postings on Stock Message Boards. Working Paper No.98025, University of Michigan

Appendix

Lexicon-based Classifier (LBC, CL1)

LBCs have been effectively used in earlier studies (Das and Chen, 2001). To build a LBC, first we develop a set of frequently occurring keywords for our three categories – Noise (C_1), Neutral (C_2), and Signal (C_3). These keywords together form a domain-specific lexicon. LBCs then categorize a message m_l, where $l = \{1, 2, \ldots M\}$, by matching each message content against this lexicon for each category and classifying the message as belonging to a category with the highest degree of matches. Formally, Category(m_l) = C_i, where i is the index for which:

$$\sum Count(m_l, Key_{ij}) = Max_{k=1,2,3}\{\sum Count(m_l, Key_{kj})\}.$$

Key_{ij} is the keyword j in the lexicon for Class I, and $Count(m_l, Key_{ij})$ returns the number of occurrences of Key_{ij} in the message m_l.

Readability-based Classifier (RBC, CL2)

This classifier is based upon research on readability analysis, and can be valuable when used in conjunction with other classifying methods. We first randomly select a subset of messages from the sample and use genetic algorithm (GA) to induce a rule set based on three variables: word count, mean word length, and number of unique words. GA will then attempt to find the range of values for the three variables to define a class. The resulting "if then…" decision tree is applied to a test set for validation.

Weighted Lexicon Classifier (WLC, CL3)

This classifier is a variation of LBC. WLC overcomes the drawback of LBC which induces a bias for classes with higher number of keywords. To eliminate the bias, WLC bases its classification on $Max[\frac{n(k_i)}{N_i}]$, where N_i represents the total number of keywords in class *i*. More formally, Category(m_l) = C_i, where i is the index for which $\sum Count(m_l, \frac{Key_{ij}}{N_i}) = Max_{k=1,2,3}\{\sum Count(m_l, \frac{Key_{kj}}{N_k})\}$.

Vector Distance classifier (VDC, CL4)

According to Chen and Das (2001), VDC treats each message as a word vector in D- dimensional space, where D represents the size of the lexicon. The proximity between a message m_l and grammar rule G_j is computed by the cosine angle of the $Vector(m_l)$ and $Vector(G_j)$, where $Vector(V)$ is the D-dimensional word vector for V. Message m_l belongs to class of G_k when the computed angle is the minimum, which means that the proximity is the maximum.

Differential Weights Lexicon Classifier (DWLC, CL5)

This classifier represents another variation of the LBC, which assigns differential weights to each word in the lexicon. DWLC recognizes the varying importance of each keyword in classification and

overcomes the equal weight bias in LBC. More formally, Class(m_l) = C_i, where i is the index for which $\sum Weight_{ij} \times Count(m_l, Key_{ij}) = Max_{k=1,2,3}\{\sum Weight_{kj} \times Count(m_l, Key_{kj})\}$.

The sixth classifier is designed by combining the outputs of the five classifiers using a simple majority voting mechanism. If we assume that each classifier categorizes (votes) message m_l as belonging to category C_i, this combination classifier simply relies on the number of votes each message gets to decide which category message m_l belongs to.

Table 1: Number of Stock Messages Collected

Message Boards / Stock Ticker	Yahoo!Finance		Raging Bull		Silicon Investor	
	# of postings	Market share	# of postings	Market share	# of postings	Market share
BRCD	65,430	92%	4,540	6%	978	1%
CMGI	65,392	44%	65,576	44%	19,219	13%
CNET	34,782	80%	7,837	18%	963	2%
CSCO	16,082	16%	65,527	66%	17,109	17%
DELL	64,486	27%	33,827	14%	142,703	59%
DIS	65,122	54%	5,068	4%	49,478	41%
EBAY	65,444	72%	18,429	20%	6,947	8%
GE	9,237	16%	44,993	80%	1,912	3%
GM	23,155	77%	1,067	4%	5,940	20%
IBM	57,399	78%	10,606	14%	5,365	7%
INKT	60,796	87%	8,519	12%	952	1%
JDSU	22,136	20%	65,572	60%	21,586	20%
MCD	13,410	17%	65,476	82%	561	1%
MSFT	49,399	28%	65,340	37%	59,547	34%
Average	**43,734**	**51%**	**33,027**	**33%**	**23,804**	**16%**

Figure 1: Market Share Distribution of Discussion Boards

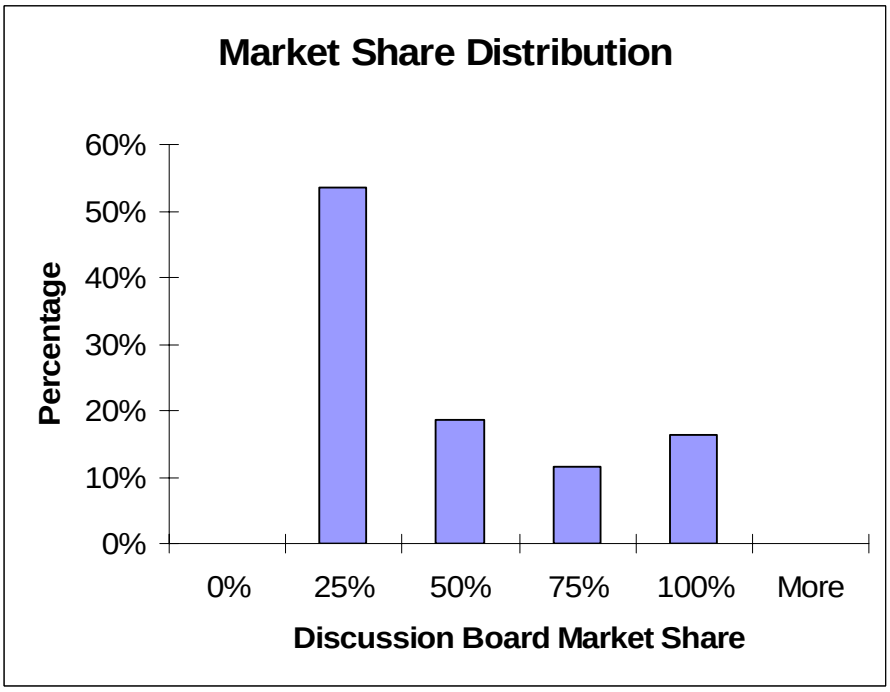

Table 2: Descriptive Statistics

Message Boards / Variables	Yahoo!Finance	Raging Bull	Silicon Investor
# of postings/week	492	433	226
Signal ratio	29%	52%	51%
# of useful postings	127	206	104
VIC size (log)	3.60	3.11	2.78

Table 3: Network Externality and Competition among VICs

	Estimates	Hypotheses	Supported
# of useful postings (β_2)	0.02** (0.00)	H1a	Yes
(# of useful postings)^2 (β_4)	1.60 E-7** (0.24E-7)	H1b	No
# of all postings (β_5)	-0.01** (0.00)	H1c	Yes
(# of all postings)^2 (β_7)	-2.86 E-7** (0.00E-7)	H1d	Yes
VIC Size (log) (β_1)	0.45 ** (0.01)	H2a	Yes
# of useful postings × log VIC Size (β_3)	-0.005** (0.000)	H3a	Yes
# of all postings × log VIC Size (β_6)	0.002** (0.000)	H3b	Yes
Control variables not reported			
Observations	520,712		
Log Likelihood Ratio	-335,397		

Table 4: Quality and Discussion Board Size

	Estimates	Hypotheses	Supported
VIC Size (log) (δ_1)	-0.12** (0.01)	H2b	Yes
Discussion Board Size (δ_2)	-0.00008** (0.00)	H2b	Yes
Observations	2603		
R-square	11.35%		

Table 5: Value of VICs (for typical boards of individual VICs)

Message Boards / Value Components	Yahoo!Finance	Raging Bull	Silicon Investor
Integrated services	0.28	0.15	0.00*
Value of discussion boards	0.73	1.45	0.88
Cost of discussion boards	-0.66	-0.82	-0.67
Cost as a percentage of Value	91%	57%	76%
Net value of discussion boards	0.07	0.63	0.21
Net value of the VIC	0.35	0.78	0.21

* The value of VICs is calculated in utility terms. A well-known feature of utility function is that it represents consumer preference and is only useful in making comparison between choices. It is therefore necessarily to choose a reference point. In this table, we denote Silicon Investors' integrated services as the reference point.. It does not mean that Silicon Investors' integrated services have no value. Rather, the number in the table represents the value of a service as compared to the integrated services offered by Silicon Investors.

Table 6: Value of VICs (for the same board)

Message Boards / Value Components	Yahoo!Finance	Raging Bull	Silicon Investor
Integrated services	0.28	0.15	0.00*
Value of discussion boards	0.92	1.12	1.36
Cost of discussion boards	-0.53	-0.76	-1.02
Cost as a percentage of Value	57%	67%	75%
Net value of discussion boards	0.39	0.37	0.34
Net value of the VIC	0.67	0.52	0.34

Table 7: A message example for GOOGLE from Yahoo! Finance.

MsgNo	Author	Subject	Content	Post Date
27	western_unionmans	Strong Sell Also	I say its worth something like $20 per share.	08/20/01

Table 8: A sample of Lexicon of Stock Messages for Three Classes

Noise	Neutral	Signal
idiot	P/E	Buy
moron	margin	Sell
stupid	ratio	Hold
retard	demand	short
fool	inventory	long

Table 9: Classifier Performance Statistics

Classifier	Total Correct	Correct Noise	Correct No Signal	Correct Signal
LBC	0.4254	0.4488	0.2564	0.6667
RBC	0.5912	0.7087	0.4359	0.0
WLC	0.4365	0.4724	0.2308	0.6667
VDC	0.3481	0.3150	0.3333	0.6667
DWLC	0.5249	0.6457	0.0769	0.6667
Combined (Simple Majority Voting)	0.4254	0.4488	0.2051	0.8

Figure 2: Methodology Design and Flow Chart for Stock Sentiment Extraction.