Information Systems Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Competition Among Virtual Communities and User Valuation: The Case of Investing-Related Communities

Bin Gu, Prabhudev Konana, Balaji Rajagopalan, Hsuan-Wei Michelle Chen,

To cite this article:
Bin Gu, Prabhudev Konana, Balaji Rajagopalan, Hsuan-Wei Michelle Chen, (2007) Competition Among Virtual Communities and User Valuation: The Case of Investing-Related Communities. Information Systems Research 18(1):68-85. https://doi.org/10.1287/isre.1070.0114

Information Systems Research
Vol. 18, No. 1, March 2007, pp. 68–85
ISSN 1047-7047 | EISSN 1526-5536 | 07 | 1801 | 0068

informs

DOI 10.1287/isre.1070.0114

Competition Among Virtual Communities and User Valuation: The Case of Investing-Related Communities

Bin Gu, Prabhudev Konana
Department of Information, Risk, and Operations Management, Red McCombs School of Business, CBA 5.202; B6500, University of Texas at Austin, Austin, Texas 78712
{bin.gu@mccombs.utexas.edu, prabhudev.konana@mccombs.utexas.edu}

Balaji Rajagopalan
School of Business Administration, Department of Decision and Information Sciences, Oakland University, Rochester, Michigan 48309, rajagopa@oakland.edu

Hsuan-Wei Michelle Chen
Department of Information, Risk, and Operations Management, Red McCombs School of Business, CBA 5.202; B6500, University of Texas at Austin, Austin, Texas 78712, hsuan-wei.chen@phd.mccombs.utexas.edu

Virtual communities are a significant source of information for consumers and businesses. This research examines how users value virtual communities and how virtual communities differ in their value propositions. In particular, this research examines the nature of trade-offs between information quantity and quality, and explores the sources of positive and negative externalities in virtual communities. The analyses are based on more than 500,000 postings collected from three large virtual investing-related communities (VICs) for 14 different stocks over a period of four years. The findings suggest that the VICs engage in differentiated competition as they face trade-offs between information quantity and quality. This differentiation among VICs, in turn, attracts users with different characteristics. We find both positive and negative externalities at work in virtual communities. We propose and validate that the key factor that determines the direction of network externalities is posting quality. The contributions of the study include the extension of our understanding of the virtual community evaluation by users, the exposition of competition between virtual communities, the role of network externalities in virtual communities, and the development of an algorithmic methodology to evaluate the quality (noise or signal) of textual data. The insights from the study provide useful guidance for design and management of VICs.

Key words: network economics; computer-mediated communication and collaboration; virtual communities; IT diffusion and adoption
History: Sandra Slaughter, Senior Editor; Brian Butler, Associate Editor. This paper was received on September 30, 2005, and was with the authors 5 months for 3 revisions.

1. Introduction

Virtual communities provide unprecedented opportunities for individuals to share information and to interact with others even when no previous social ties exist (Bagozzi and Dholakia 2002, Butler 2001). *Business Week*'s cover story on June 20, 2005, noted that "companies are using Internet-powered services [virtual communities] to tap into the collective intelligence of employees, customers, and outsiders, transforming their internal operations" (Hof 2005). However, Skype CEO Niklas Zennstrom notes that, unlike a typical organization, a virtual community is "almost like an organism" where businesses have little control over their growth and direction (Hof 2005, p. 74). In this paper, we attempt to shed more light on the dynamics of virtual communities. To this end, we address two broad questions for businesses that operate virtual communities: (a) How do users value virtual community networks? (b) How do virtual communities differ in their value propositions,

and how do they attract users with different characteristics?

We address the above questions in the context of virtual investing-related communities (henceforth VICs) where users voluntarily participate to obtain or share investing-related information. A number of VICs are hosted by financial portals (e.g., Yahoo!Finance and Raging Bull), while others are pure play message boards (e.g., Silicon Investor).[1] The latest newcomer to the VIC world is Google Finance.[2] VICs host hundreds to thousands of discussion boards on individual stocks and compete to attract individual investors. They allow individual investors to discuss, debate, and evaluate stocks; exchange information; and interact socially (Das and Chen 2001, Antweiler and Frank 2004, Wysocki 1999, Konana and Balasubramanian 2005). Information posted on VICs affect users' attitudes and decisions (Das and Chen 2001). For example, false information posted about NEI Webworld Inc. on Internet chat rooms increased the stock price from 15¢ to more than $15 within a few hours (SEC Litigation 16620, July 6, 2000). In 2000, when an individual "knowingly and willfully" (as stated by the federal grand jury during trial) released a fraudulent negative news item about Emulex Corporation, the stock value plummeted 62% within a few hours. The inherent uncertainty and risks associated with investing activity make the information found on VICs an important resource for users. Furthermore, differing motives and incentives for economic gains (e.g., short-term versus long-term investing) induce varied levels of VIC participation. Thus, VICs offer an ideal context to examine the research questions of interest in this study.

First, we examine how users value VICs. In this context, VICs create value by providing information resources to their community members (Butler 2001).[3] There are two types of such information resources: bundled information from the VIC providers and postings from community members. Examples of bundled information offered by VIC providers include business news, company financials, stock charts, analyst reports, and other stock information. Each VIC provides different levels of bundled information; those levels may change over time. The second information resource is postings, which include messages from community members providing stock analysis, seeking investment advice, or discussing the implications of earnings or news releases.

We argue that postings on VICs exhibit both positive and negative externalities. Positive externalities arise when more postings increase the available information resource for the community. In response to the higher volume of information, the VIC attracts more users who are likely to respond by posting new information, providing analysis or opinions. This, in turn, further increases the extent of the information resource available to community members. The presence of positive externalities can be observed by comparing Yahoo!Finance and newly launched Google Finance. Although Google Finance provides extensive bundled information, its stock discussion boards have been rarely used so far. For the 14 stocks under consideration in this study, the number of total postings on Google between March 2006 and September 2006 is a meager 400. On the contrary, the number of postings on Yahoo!Finance between September 15, 2006, and September 19, 2006, (fewer than five days) for Microsoft stock alone is more than 750. Such a positive externality mechanism has been observed in peer-to-peer (P2P) electronic networks (Asvanund et al. 2004), although the type of resource considered in their study is different.

An increase in the posting volume, however, may not be costless. Additional posting increases processing cost on every user of the community, representing a classic case of negative externality (e.g. Butler 2001, Moreland et al. 1996). As a result, more postings may reduce the benefits of the virtual community, thereby increasing user attrition and reducing the number of postings in the future.

The type and nature of network externalities are influenced by information quality. High-quality postings provide better information and, thus, lead to better information resources and positive externalities. On the contrary, low-quality information or noise is often distracting and will result in an increase of a member's search and information-processing

[1] According to Global Comscore MediaMatrix, the top three finance portals attracted 64.4 million unique visitors in May 2006.

[2] Our study period ends before the launch of Google Finance.

[3] We focus only on the information resource in the paper because users are attracted to VICs mainly for their information value.

costs, thereby increasing the difficulty to convert the available information resource into benefits. Hence, low information quality is associated with negative externalities.

Second, given our understanding of user valuation of VICs, we explore how VICs differentiate and attract consumers with different characteristics. The driver for the differentiation strategies is the potential trade-off between the two underlying dimensions of the value of postings in VICs: posting volume and quality. Virtual communities are open to the public and welcome postings from any user. The open nature of virtual communities, however, allows noise, spam, and irrelevant or abusive postings into the communities. To improve quality, VICs implement a variety of strategies to monitor their discussion boards. Postings containing abusive, obscene, or commercial content can be removed, and the users who post such messages may be banned from the community. Raging Bull, for instance, provides moderators for discussion boards and technological capabilities that allow each user to establish an "ignore" list to screen out postings from unworthy posters. Such an operation is costly for Raging Bull because it requires real-time filtering every time a user visits the VIC; however, it is one of the most popular features for the community users. While maintaining quality is feasible in a community with a smaller number of postings, the challenge becomes substantial for virtual communities with a larger volume of postings. This is because the cost of maintaining quality is proportional to the number of postings, indicating an inherent trade-off between posting volume and quality.

Studies in economics (e.g., Shaked and Sutton 1983) have shown that the trade-off situation is ideal for differentiated competitive strategies. That is, some virtual communities could compete based on posting volumes, while others focus on posting quality. The differentiation also has implications for VICs' optimal strategies for providing bundled information. Virtual communities with high posting volume have economies of scale to provide extensive bundled information, which involves substantial fixed costs irrespective of community size. A case in point is Yahoo!Finance's recent decision to obtain exclusive use of refined and reviewed blog content for a wide range of stocks from Seeking Alpha. The decision involves hiring more than 200 analysts, a costly endeavor that few VICs can afford. On the contrary, virtual communities with low posting volume have lower economies of scale and hence, and can provide less bundled information.

Another implication of differentiation strategies is that VICs attract users with different characteristics. Users with higher information-processing costs are likely to prefer virtual communities with lower posting volume and higher information quality. On the other hand, users with lower information-processing costs are more likely to participate in communities with higher posting volume even though the average quality per posting is lower.

Based on the above discussion, we suggest that VICs do not compete simply to be the largest community with the most postings, a notion predominantly asserted in network externalities literature (e.g., see Katz and Shapiro 1995). Instead, we argue that multiple VICs could coexist due to their choice of differentiation between posting volume and quality.

While our approach to evaluating externalities is generally consistent with the extant literature, we use a different measure to quantify resources available to the community. Prior literature measures resource in a community by network size: that is, the number of users in a virtual community. The number of users in a network is an appropriate measure of resources when interactions among members are important. However, it may not be a good measure of resources in virtual communities where users mainly focus on information and the information contribution of each user is unequal. Generally, contributions in VICs follow a power law distribution (Kuk 2006): that is, a few top contributors provide most of the resources in a virtual community. Thus, we need to weigh each user in a virtual community by his contributions to the community. This is equivalent to measuring resources by the number of postings. We also note that users in traditional networks may decrease network value through free riding. For instance, more than 65% of Gnutella network users downloaded music for free without contributing to other members in the network, causing significant congestion (Adar and Huberman 2000, Krishnan et al. 2003). However, in VICs, free riders (e.g., passive readers) do not decrease user valuation of VICs. Instead, low-quality

postings decrease the benefit of VICs because the postings increase user information-processing costs. For these reasons, we consider the number of postings to be a better measure of network resource than the number of users.

An empirical challenge for researchers is how to identify the trade-off between posting volume and quality, since quality is not easily assessable on a large scale. We propose a novel approach to explicate these effects. By applying text mining algorithms, we discriminate high-quality from low-quality postings. Noting that the positive network externalities are primarily driven by high-quality postings and the negative externalities by volume of total postings, we use posting quality to identify the positive and negative externalities separately.

The remainder of this paper proceeds as follows: Section 2 provides the background and literature review, theoretical model, hypotheses, and empirical model. Section 3 discusses the data sources and research method. Section 4 discusses the results and implications. Section 5 presents contributions of the study to research and practice, limitations, and opportunities for future work.

2. Theoretical Model and Hypotheses

2.1. Background and Literature Review

Virtual communities are networks where members exchange information and resources (e.g., Bagozzi and Dholakia 2002, Asvanund et al. 2004, Butler 2001). Virtual communities have been studied in a number of disciplines including economics, social sciences, marketing, and MIS. For example, Sproull and Kiesler (1991) investigated virtual communities and argued that they can fundamentally change the ways people interact with one another. Such communities have also changed the dynamics of the relationship between a business and its customers. Chevalier and Mayzlin (2006) noted that consumers become empowered through online consumer communities that provide product reviews and advice, putting virtual communities at the center of online business strategy.

In virtual communities, such as P2P networks, the more resources provided by the members in the network, the more valuable the network is to all members. This in turn attracts more new members. That is, these networks exhibit positive externalities (Asvanund et al. 2004). Asvanund et al. (2004, p. 156) noted that "positive network externalities arise when users who choose to share their content bring new content, replicas of existing content, or other shared resources to the network." Such positive network externalities have been documented in other virtual communities for transactions (e.g., eBay), communication (e.g., Skype), and gaming (e.g., Amit and Zott 2001, Shankar and Bayus 2003). Network externalities in virtual communities are similar to those demonstrated in network products such as software or fax machines. The only difference is that the latter often has stand-alone value, while the former derives most of the value from the network itself.[4] Ceteris paribus researchers generally agree that positive network externalities imply that larger networks provide more value than smaller networks provide to members in the community (e.g., Economides 1996, Strahilevitz 2003, Katz and Shapiro 1995, Farrell and Saloner 1986, Kauffman et al. 2000).

Butler (2001), however, made an important point that resources alone do not provide benefits to community members. Members incur costs to convert resources into benefits because they need to expend effort to process information. This suggests that members of large communities experience disadvantages resulting from higher information-processing costs. After studying 206 e-mail-based listservs, Butler (2001) found that large communities are associated with a higher volume of e-mails and a greater variation in content and relevance. This results in substantial processing costs to users—a source of negative externalities. This notion of increased information-processing costs can also be explained with information overload (IO), a condition in which the amount of information an individual must process is larger than that individual's capacity for processing information. This condition (IO) could affect some or all of the cognitive processes (Fournier 1996) such as attention, storage, and retrieval (Lindsay and Norman 1977). For example, IO can result in a loss of information by forcing increased attention to new information, but

[4] In VICs, bundled information can be viewed as stand-alone value, while postings can be viewed as network value.

reduced attention to relevant information. This phenomenon is prevalent in today's environment of easy and inexpensive access or distribution of information (Fischer and Stevens 1991, Jones et al. 2004). For example, Yahoo!Finance's Google discussion board receives more than 1,000 postings a day, presenting a challenge for users to read and comprehend information on the board. In a similar vein, Asvanund et al. (2004) considered the negative externalities associated with P2P file downloading networks where users create substantial congestion on the network. These insights about the relationship between community size and user value indicate that a larger community may not necessarily be associated with a higher membership value.

There is another important stream of research that focuses on the value of virtual community as a social network, which is different from informational value considered in this research (e.g., Rheingold 1993, Kraut et al. 1996, Moreland and Levine 2001, Prentice et al. 1994, Sassenberg 2002). Researchers have argued that users gain value from social interactions in these communities that enable them to maintain existing ties and build new ones (Butler et al. 2005). Balasubramanian and Mahajan (2001) also emphasized the importance of social interaction in the economic system. Bagozzi and Dholakia (2002) described virtual communities as information neighborhoods (Burnett 2000) that exert significant influence on member perceptions about products and services. While we acknowledge the dimension of social interactions and bonding as being important in the context of virtual communities, it is beyond the scope of this paper.

2.2. Modeling Valuation of VICs

Central to much of the existing work is the evaluation of a community by current and potential users. As discussed earlier, VICs offer value to users from two sources: bundled information from VIC providers and postings from members. This value influences user decisions to visit and to contribute resources through postings. Let a_{kt} denote the value of bundled information provided by VIC k to a user at time period t. This value a_{kt} is assumed to be the same for all discussion boards hosted by a VIC at time period t, but varies across different VICs and over time.

As discussed in the introduction, the value of postings depends on the number of postings (i.e., available information) and the quality of these postings in a particular discussion board; posting value varies over time. Let n_{jkt} denote the total number of postings for stock discussion board j on VIC k at time period t and q_{jkt} denote the probability that a given posting contains useful or relevant information (i.e., an indicator of quality). The value of the stock discussion board, therefore, is a function of $n_{jkt}q_{jkt}$ number of quality postings in the discussion board. However, like most economic resources, we posit that the number of quality postings on VICs exhibit diminishing marginal value for a given user (Katz and Shapiro 1986). The rationale is that for a given event (e.g., earnings announcements), the first few postings often have enough information to inform a user. Subsequent postings are less valuable than the first few postings.[5] Thus, the value of a VIC to a user is an increasing but concave function of the number of quality postings.[6] We adopt a quadratic form[7] for the value function, which allows us to identify separately the baseline value of the first quality posting and the marginal value of additional quality postings. This indicates that the value of a stock discussion board with $n_{jkt}q_{jkt}$ quality postings can be expressed as

$$b(n_{jkt}q_{jkt}) + c(n_{jkt}q_{jkt})^2. \tag{1}$$

The marginal value of a quality posting is therefore $b + 2cn_{jkt}q_{jkt}$. Here, b represents the baseline value of a quality posting. If there is a diminishing return to the number of quality postings, c would be negative. Our model is consistent with Asvanund et al. (2004), who hypothesized that resource availability increases with the size of a P2P network, but also demonstrated that the marginal contribution of a network to a given member decreases with its size.

[5] Our cursory analyses confirms that in active discussion forums there are a number of postings that often repeat or agree with information reported earlier.

[6] The value of the VIC to a user is distinctive from the total network value of the VIC, which equals the sum of the VIC value over all its users.

[7] Quadratic forms are often used in empirical research to capture diminishing marginal returns. A key benefit of using quadratic forms is easy interpretation of the regression results.

Thus, the total value V_{jkt} received by a user visiting VIC k's discussion board on stock j at time period t is

$$V_{jkt} = a_{kt} + b(n_{jkt}q_{jkt}) + c(n_{jkt}q_{jkt})^2. \quad (2)$$

Our approach expands on extant empirical models by incorporating not only the influence of the information quantity (i.e., posting volume), but also the influence of its quality. By doing so, we will be able to quantify the value of VICs accurately and to differentiate between positive and negative externalities. Based on the above discussion, we state the hypotheses on user valuation as follows:

Hypothesis 1A. *The value of a discussion board to a user increases with the number of quality postings available on the discussion board in a given time period.*

Hypothesis 1B. *The value of a discussion board to a user is concave with regard to the number of quality postings on the discussion board in a given time period. That is, the marginal contribution of each additional quality posting decreases with the total number of quality postings in the discussion board in a given time period.*

While offering a range of benefits to users, VICs also entail costs. Das and Chen (2001) showed that a large fraction of postings in VICs consists of noise and rumors. The users of these communities often need to process hundreds of postings to extricate useful information. Large volumes of postings contribute to IO, which in turn affects users' cognitive ability to analyze postings and reduces the attention they pay to useful information (Lindsay and Norman 1977, Shenk 1997). The cost of using VICs increases disproportionately with posting volume as users reach or exceed their information-processing capabilities. As a result, users are more likely to end participation in these communities or choose a smaller community when the posting volume increases above a particular threshold (Butler 2001, Jones et al. 2004). Thus, the cost of using VICs is expected to be increasing and convex in the total number of postings. This suggests that we can use a quadratic form for the cost function:

$$C_{jkt} = dn_{jkt} + en_{jkt}^2. \quad (3)$$

The cost function is similar to the one used by Asvanund et al. (2004). Their study showed that network congestion and download time increase with the number of users in a P2P network, and the marginal cost increases with the network size. We extend their model by recognizing that all postings affect costs, but only quality postings affect value. This approach allows us to identify simultaneously the value and the cost of using VICs. We state the hypotheses as follows:

Hypothesis 1C. *The cost incurred by a user increases with the total number of postings on a discussion board in a given time period.*

Hypothesis 1D. *The cost incurred by a user is convex with regard to the total number of postings on a discussion board in a given time period. That is, the marginal cost of an additional posting to a user increases with the total number of postings on a discussion board in a given time period.*

The above hypotheses implicitly assume that users have to process every posting to find information. In reality, users may use search technologies to screen out useless postings. The use of such search techniques, in general, may reduce some information-processing costs. If we assume that search helps users remove a fraction of low-quality postings, their information-processing costs will be reduced accordingly. In this case, the coefficients in the equation will decrease and will be interpreted as the joint effect of processing costs and search technologies.

2.3. Differentiation Among VICs

The second issue we explore is how VICs differentiate in their value propositions. The differentiation among discussion boards depends on the benefits received by users. Given the specifications of value and cost functions discussed in the previous section, the net value a user receives from stock discussion board j in VIC k at time period t is the difference between the value received and the cost incurred:

$$\begin{aligned} U_{jkt} &= V_{jkt} - C_{jkt} \\ &= a_{kt} + b(n_{jkt}q_{jkt}) + c(n_{jkt}q_{jkt})^2 - dn_{jkt} - en_{jkt}^2. \end{aligned} \quad (4)$$

To increase the benefits received by the users, VICs have two options: they can either provide better bundled information (a_{kt}) or improve posting quality in the community ($n_{jkt}q_{jkt}$). The two options of increasing value, however, have different cost implications

for the VIC providers. Providing better bundled information involves collecting extensive company news, purchasing analyst reports and company SEC filings, and setting up online systems to distribute the information, all of which have substantial fixed costs. However, the marginal cost of providing this information to an additional user is minimal. Thus, large VICs would benefit from economies of scale, which suggests that as a VIC grows it has more incentive to provide better bundled information.[8]

While VIC providers can control the level of bundled information, it is a challenge to control posting quality, since postings are created by community users in an open domain. This is especially true for large VICs. While small VICs can often rely on social norms to enforce posting quality (Honeycutt 2005), large discussion boards have weaker social ties and hence are likely to find it more difficult to maintain a quality standard through emergent social mechanisms (Cummings et al. 2002). One approach VIC providers use to improve posting quality is to actively monitor and filter low-quality postings, thereby reducing the information-processing costs for the users. Other strategies to improve quality include allowing users to report abusive postings, which are then investigated manually by VIC providers; or offering customized automatic filtering mechanisms to screen postings based on keywords or user IDs. All these activities involve substantial variable costs, which grow proportionally to posting volume.

The difference between the cost structures of the two options to VIC providers suggests that as a VIC grows larger it is better positioned to provide more bundled information due to economies of scale. At the same time, when a VIC grows larger it has a greater cost disadvantage in moderating the community and maintaining posting quality. Given these dynamics, a VIC is likely to capitalize on its advantages. This implies that larger VICs will focus more on leveraging their economies of scales, while smaller VICs will focus more on improving posting quality. That is, the value of a stock message board is influenced not only by the volume and quality of postings within that discussion board, but also by the scale of VIC it belongs to. When everything else is equal, larger VICs are likely to provide more value to users than do smaller VICs, but smaller VICs are likely to maintain higher posting quality.

Hypothesis 2A. *Ceteris paribus, larger VICs are likely to provide more value to users than do smaller VICs.*

Hypothesis 2B. *Smaller VICs are likely to maintain higher posting quality than do larger VICs.*

2.4. User Differentiation

There is substantial literature in economics and strategy showing that consumer heterogeneity can reduce direct competition and increase the profitability of differentiation strategies (e.g., Shaked and Sutton 1983, Vandenbosch and Weinberg 1995). In the context of VICs, we assume that users have different levels of information-processing costs and different levels of tolerance for quality information. Users having higher information-processing cost per posting are likely to gravitate toward VICs that offer fewer but higher-quality postings. On the contrary, users having lower information-processing cost per posting are likely to process more postings. Thus, although the average quality per posting in larger VICs is lower than in smaller VICs, these users can leverage greater levels of postings in larger VICs to derive a high aggregate value. Such users have greater motivation to join a larger VIC. Hence,

Hypothesis 3A. *Larger VICs attract users with lower valuation of high-quality postings.*

Hypothesis 3B. *Larger VICs attract users with lower information-processing cost.*

2.5. Empirical Model

Besides the factors identified in the above hypotheses, other factors may affect the value of a stock discussion board. In particular, structural differences across VICs, such as fee structure, may simultaneously affect posting volume and information quality in a stock discussion board, leading to potentially biased estimates.[9] Similarly, differences across stocks, such as

[8] We do not conduct a test for the economies of scale directly. However, if we observe large VICs are more likely to be associated with higher a_{kt}, it would be consistent with predictions based on economies of scale.

[9] We thank the anonymous reviewer for raising this issue and prompting us to incorporate this in our model.

volatility and market capitalization, could affect the quality and posting volume on a stock discussion board. As a control for unobserved structural differences among VICs and stocks, we include fixed effects γ_{jk} for every stock discussion board in each VIC in our model. By doing so, our coefficients are estimated from time-series variation in discussion boards as they increase or decrease in posting volume over time. We also note that a user's decision to post on a stock discussion board depends on the individual's prior usage of the hosting VIC. Once a user makes a decision to use the discussion boards in one VIC, he is less likely to switch to another VIC. To control for this dependence, we add a user's past usage of the VICs into our empirical model: x_{ikt} takes a value of 1 if user i had used VIC k before time period t. Otherwise, it takes a value of 0. The coefficient on the variable captures the dependence between a user's decision in the current time period and his prior usage of the VIC.[10]

Hypotheses 2 and 3 indicate that the coefficients a, b, and d in Equation (4) vary with the total posting volume of the hosting VIC. Larger VICs are more likely to provide more bundled information (higher a), but are less likely to offer higher-quality postings. As a result, larger VICs are more likely to attract users with lower information-processing costs (lower d) and lower valuation for high-quality information (lower b). The relationship between VIC posting volume and the coefficients, however, is not linear. The lack of economies of scale makes it economically challenging for smaller VICs to provide bundled information. However, when VICs reach a critical mass, they are likely to offer bundled information; subsequent growth in VICs may have little impact on future offerings. This indicates that bundled information offers have an increasing but concave relationship with VIC posting volume. The same argument of diminishing impact of VIC posting volume can be applied to other factors, as well. To model the nonlinear relationship, we use the natural log of the total number of VIC postings ($\log N_{kt}$) in our regression model as the factor that influences coefficients a, b, and d. We express the coefficients as follows:

$$a = \beta_0 + \beta_1 \log N_{kt}, \tag{5}$$

$$b = \beta_2 + \beta_3 \log N_{kt}, \tag{6}$$

$$d = \beta_5 + \beta_6 \log N_{kt}. \tag{7}$$

Equation (4) also assumes that users make decisions based on the discussion board and VIC characteristics in the current period. This raises an empirical concern. The estimation could be influenced by reverse causality. That is, while the equation suggests that high posting volume leads to higher user valuation of VICs, it is also true that higher valuation leads to more postings. The results of empirical estimation could be biased if we do not take this into consideration. The concern can be addressed by using lagged variables (i.e., variables from the previous period) in the empirical model. By doing so, we are able to make a stronger inference about the direction of the causality. We revise Equation (4) to address these issues. We also follow the convention for empirical models by reversing the sign for d and e in Equation (4), renaming coefficients c and e to make them consistent with the other coefficients, and substituting Equations (5)–(7) into Equation (4). Therefore, our final empirical model is

$$\begin{aligned} U_{ijkt} = {} & \gamma_{jk} + \beta_1 \log N_{k,t-1} + \beta_2 q_{jk,t-1} n_{jk,t-1} \\ & + \beta_3 \log N_{k,t-1} n_{jk,t-1} q_{jk,t-1} + \beta_4 q^2_{jk,t-1} n^2_{jk,t-1} \\ & + \beta_5 n_{jk,t-1} + \beta_6 \log N_{it-1} n_{j,t-1} + \beta_7 n^2_{jk,t-1} \\ & + \beta_8 x_{ikt} + \varepsilon_{ijkt}. \end{aligned} \tag{8}$$

Table 1 summarizes the hypotheses and predicted direction of the corresponding coefficients in Equation (8). Hypothesis 1A suggests that β_2 should be

Table 1 Summary of Hypotheses and Predicted Direction of Coefficients

Hypotheses	Coefficients	Direction
1A	β_2	+
1B	β_4	−
1C	β_5	−
1D	β_7	−
2A	β_1	+
3A	β_3	−
3B	β_6	+

[10] Our approach to control for user experience is similar to the approach used in the switching cost literatures (e.g. Chen and Hitt 2002).

positive, since the user valuation increases with the number of quality postings. Hypothesis 1B indicates that β_4 would be negative, since the marginal value of quality postings is decreasing. Hypothesis 1C predicts that β_5 will be negative, because the cost is increasing with posting volume. Hypothesis 1D expects β_7 to be negative, due to accelerating information-processing costs. Hypothesis 2A suggests that β_1 shall be positive, since larger VICs have an advantage in providing more bundled information and, therefore, are more likely to do so. Hypothesis 3A indicates that β_3 should be negative, given that larger VICs appeal to users with lower valuation of quality postings. Likewise, Hypothesis 3B indicates that β_6 should be positive, suggesting that larger VICs appeal to users with lower information-processing costs.

If we know user valuations for every discussion board in each of the VICs, the above empirical model can be estimated by linear regressions. In practice, however, we only observe users' VIC choices when they post messages. These choices do not necessarily imply that a user makes a conscious decision about where to post the message; rather, most users post messages on their preferred VICs. As a result, the postings reveal user preferences among VICs. We analyze these revealed choices among VICs and infer user utilities of the VICs using a multinomial logit (MNL) regression model (Gensch and Recker 1979, McFadden 1987). The underlying assumption of the MNL model is that a user always chooses the VIC that provides the highest latent benefit. We also allow users' decisions to be dependent on their past usage of the VIC because a user is more likely to stay with a VIC once she has already participated in it. The latent benefit has two components: deterministic and stochastic. The deterministic component is determined by the attributes of the stock discussion board j within VIC k ($N_{jk,t-1}$, $n_{jk,t-1}$, $q_{jk,t-1}$), user prior usage of VICs (x_{ikt}), and population characteristics (βs). The stochastic component is due to measurement error and unobserved variation in user preferences (ε_{ijkt}). Given the VIC attributes and user preference, the probability P_{ijkt} that user i chooses VIC k over other VICs for stock j during time period t is

$$P_{ijkt} = \frac{U_{ijkt}(N_{k,t-1}, n_{jk,t-1}, q_{jk,t-1}, x_{ikt}; \beta)}{\sum U_{ijkt}(N_{k,t-1}, n_{jk,t-1}, q_{jk,t-1}, x_{ikt}; \beta)}. \quad (9)$$

We do not test all hypotheses using the MNL model and the user posting data. Hypothesis 2B deals with the relationship between posting quality and VIC posting volume. Given that we can measure posting quality directly, we run an OLS regression of posting quality on VIC posting volume and discussion board posting volume.

$$q_{jkt} = \delta_o + \delta_1 \log N_{kt} + \delta_2 n_{jkt} + \zeta_{jkt} \quad (10)$$

Hypothesis 2B suggests that larger VICs have a disadvantage when trying to maintain quality and, therefore, are more likely to be associated with lower quality. Therefore, δ_1 in (10) above is posited to be negative. Likewise, within the same VIC, it is easier to maintain quality on smaller discussion boards than on larger discussion boards. Hence, δ_2 is also expected to be negative.

3. Data and Methods

3.1. VIC Selection

To empirically test our model, we collected a data set of online postings from three large VICs—Yahoo!Finance, Silicon Investor, and Raging Bull—from January 1, 1998, to March 31, 2002. These three VICs were widely acknowledged as the leading communities for investors during that time and have been used in other studies (e.g., Antweiler and Frank 2004). Yahoo!Finance is part of Yahoo! portal and is the largest VIC. Unlike Yahoo!Finance, which started as part of the Yahoo! portal, Silicon Investor started primarily as a stand-alone VIC. At its inception in late 1995, Silicon Investor mainly targeted technology stocks. Over the years, it has substantially broadened its coverage, but maintained its reputation as a VIC for users interested in high-tech companies. Silicon Investor charges users an annual fee to post messages on its discussion boards, but allows anyone to read postings for free. Besides serving as a revenue source for Silicon Investor, the fee limits postings to serious users who are more likely to provide higher-quality postings. Raging Bull began later than Yahoo!Finance and Silicon Investor, but experienced significant growth after CMGI's initial investment in September 1998. Raging Bull initially focused on small stocks, especially over-the-counter (OTC) stocks. Over time, it grew into the third-largest "stock talk" site,

Table 2 Number of Stock Messages Collected

	Message boards					
	Yahoo!Finance		Raging Bull		Silicon Investor	
Stock ticker	# 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	65,536	44	19,218	13
CNET	34,782	80	7,837	18	962	2
CSCO	16,082	16	65,525	66	17,019	17
DELL	64,361	31	33,827	17	106,267	52
DIS	65,122	54	5,068	4	49,478	41
EBAY	65,444	72	18,429	20	6,926	8
GE	9,237	16	44,981	80	1,912	3
GM	23,155	77	1,067	4	5,940	20
IBM	57,399	77	10,602	14	6,317	8
INKT	60,796	85	8,519	11	1,932	3
JDSU	22,136	20	65,536	60	21,586	20
MCD	13,410	17	65,472	82	561	1
MSFT	49,399	28	65,337	37	59,517	34
Average	**43,725**	**45**	**33,020**	**34**	**21,330**	**22**

with 2 million daily page views. A key distinction of Raging Bull is its focus on posting quality through moderators and technological features.

We collected message postings on a weekly basis from a random sample of 14 stocks common to all three communities. Given prior findings that stocks with significant growth uncertainty and beta value attract larger number of postings (Wysocki 1999), 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), McDonald's (MCD), and Microsoft (MSFT) are widely held and less volatile. Stocks such as Brocade Communications (BRCD), CMGI, JDS Uniphase (JDSU), Inktomi (INKT), and eBay carry more risk and are considered volatile with wide fluctuation in stock movement during the time period in which we collected the data for this study. Cisco (CSCO) and Dell were relatively less speculative, but exhibited significant growth. We acquired the following attributes for each posting: message number, author, subject, posting date, posting time, and message content. We provide the summary statistics in Table 2.

Our analysis requires estimating discussion board posting volume, overall VIC posting volume, and posting quality. Discussion board posting volume is directly observable from our data, but it was not feasible to measure overall VIC posting volume because it requires counting the total number of 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 approximate VIC posting volume with the number of times a user visited each of the three VICs, the cumulative duration of the visits, or the number of users who have visited the VICs. These measures are highly correlated. In the analysis that follows, we use the number of visits as the measure. To explore the time-series variation in VIC volume, we collected data on weekly visits to each of the three VICs recorded by MediaMatrix.

3.2. 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 discussion board topic. We consider a message to be *neutral* if the

content relates to the stock in particular or the market in general with implications for the stock, but presents no specific signal such as buy, hold, or sell. In our analysis, we count both signal and neutral messages as quality postings. We also tried an alternative measure of counting only signal messages as quality postings. Both measures yield qualitatively similar results. Our measure of quality focuses on the informational aspect of the posting quality and does not measure posting quality from the social interaction perspective.

We use a multialgorithmic technique (Das and Chen 2001) to classify messages. We developed investing-related classifiers to take into account the unique characteristics of the postings (see appendix for detailed descriptions of classifiers and algorithms). We used Network Query Language (NQL)[11] to download postings from discussion boards in three VICs, which we then fed into the five classification algorithms to classify them as noise, neutral, or signal. We performed the methodology used to determine posting quality in three steps: classifier development, testing and validation, and application.

In the first stage, we used five relevant extraction algorithms widely employed in text categorization (Das et al. 2005, Antweiler and Frank 2004). A sixth classifier extracted the final categorization based on majority category among the five relevant extraction algorithms. Once the classifiers were designed, we trained the algorithms using 300 messages. Two business graduate students who were informed of the definitions of the categories carried out the human coding. A high degree of consensus emerged (inter-rater reliability > 90%) after a few pilot coding sessions. A small number of messages that the students classified differently from each other were revisited and a consensus reached regarding their categorization. We deliberately kept the training data set small so as to prevent overfitting the data (leading to poor out-of-sample performance), which is 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 were then classified into noise, neutral, or signal.

[11] For details, see Network Query Language: http://www.nqltech.com/nql.asp.

The second phase involved testing the classifier on a subset of the sample not used for inducing the rule sets. We analyzed the classifiers' performance for their ability to classify a message into the three categories. Three authors of this study independently evaluated a random sample of 100 messages for accuracy of the classification. The majority evaluation among the three was considered as the true classification. The accuracy was computed by comparing the true category and the predicted category for each input message tested, that is,

$$\text{Accuracy} = (\text{Number of messages that predicted category is the same as its true category } i) \cdot (\text{Number of total messages})^{-1}.$$

The result of our validation reveals an average accuracy rate of 78% for the three-classification scheme and 86% for the two-classification scheme (i.e., noise or signal, where signal includes neutral messages). Among the 14% incorrect predicted category, 2% of the messages were wrongly classified as signal (i.e., false signal) and 12% were wrongly classified as noise (i.e., false noise). The overall classification accuracy is superior or comparable to that obtained in the literature on extracting sentiments from unstructured textual data with significant noise levels. The accuracy levels in the literature for two-classification sentiment extraction ranged from 50% to 90%, which drop to 50% to 70% for three-classification sentiment extraction (Pang et al. 2002, Hu and Liu 2004, Gamon and Aue 2005). Substantial research in sentiment extraction in the computer science literature recognizes the difficulty of sentiment classification, suggesting that our results are at least on par if not superior to much of the extant literature. For example, our classification accuracy is significantly better than that reported by Das and Chen (2001).

The final step involved applying the classification method to the entire data set and computing the posting quality as the percentage of quality postings.

4. Results

4.1. Descriptive Statistics

Table 2, which shows the number of postings and the market share of the three VICs for each stock during the period January 1, 1998, to March 31, 2002, reveals

two interesting patterns. First, the market share of VICs for each stock varies substantially, and the dominance in one stock does not carry over to other stocks. For example, Yahoo!Finance's BRCD discussion board received 92% of all postings for that stock, but its CSCO discussion board received only 16% of all postings for that stock. Second, the table reflects that 7 out of the 14 stocks have one dominant discussion board with a market share of over 75%. This is consistent with the presence of positive network externalities where one network dominates all others. The lack of dominance in other discussion boards indicates the possible presence of negative externalities. A close examination of the VICs reveals that most fragmented discussion boards have large posting volume. This suggests that, with the increase of posting volume, the negative externalities increase. As a result, users choose other VICs, leading to fragmented stock discussion boards. The presence of externalities will be tested in §4.2.

Table 3a illustrates the differences among the three VICs in their average weekly *discussion board posting volume*, information quality, and *VIC posting volume*. Yahoo!Finance is the largest VIC in terms of the average weekly *discussion board posting volume* and *VIC posting volume*, followed by Raging Bull and Silicon Investor. However, Yahoo!Finance's posting quality is the lowest. The signal and neutral postings account for less than 30% of the total in Yahoo!Finance. In contrast, the posting quality on the smaller VICs is much higher: more than 40% of the postings on these two VICs contain useful information. The correlation matrix of the three variables is reported in Table 3(b). We note that the posting quality is negatively correlated with the *discussion board posting volume* and the *VIC posting volume*. All the correlations are below 0.40, which indicates that multicollinearity is not a major concern.

Table 3(a) Descriptive Statistics

	Message boards		
Variables	Yahoo!Finance	Raging Bull	Silicon Investor
Discussion board posting volume (*weekly*)	387	222	109
Posting quality (%)	28	49	42
VIC postings volume (*log*)	4.49	3.64	3.63

Table 3(b) Correlation Matrix

	Variables		
Variables	Posting volume (weekly)	Information quality	VIC size (log)
Discussion board posting volume (*weekly*)	1.00	−0.19	0.21
Posting quality	−0.19	1.00	−0.33
VIC postings volume (log)	0.21	−0.33	1.00

4.2. Hypotheses Testing

We analyzed online users' valuation of VICs using the MNL model discussed in §2.5. Our model includes control variables such as fixed effects for individual stock discussion boards, and a user's prior usage of VICs. By including these variables, we control for the observable and unobservable heterogeneity among VICs, stocks, and users. We explore the time-series nature of our data and consider how changes in the discussion board posting volume affect posting quality, and how they jointly affect user valuation of the boards. By doing so, we separately identify the magnitude of the positive and negative externalities and the different value propositions offered by VICs. Tables 4 and 5 summarize our hypotheses testing results. Except for Hypothesis 1B, all our hypotheses are supported.

Table 4 shows the results for the MNL model. As hypothesized, user valuations depend on the posting volume and quality. The coefficient of the number of quality postings (i.e., β_2) is positive and significant, suggesting that user valuation increases with the number of quality postings. This indicates that more quality postings attract more users and hence more postings, suggesting a positive feedback mechanism. This supports Hypothesis 1A and is consistent with prior findings of positive externalities in virtual networks such as P2P networks (e.g., Asvanund et al. 2004).

Surprisingly, though, the coefficient on the quadratic form of quality posting (i.e., β_4) is not significant, which indicates that the marginal value of an additional quality posting is a constant. This is inconsistent with Hypothesis 1B. It also contradicts the findings of marginal diminishing return on network size in earlier empirical studies (e.g., Asvanund et al. 2004) and a common assumption in theoretical

Table 4 Network Externality and Competition Among VICs

	Estimates	Hypotheses	Supported
# of quality postings (β_2)	0.030** (0.001)	H1A	Yes
(# of quality postings)^2 (β_4)	0.796E−7 (0.909E−7)	H1B	No
# of all postings (β_5)	−0.010** (0.000)	H1C	Yes
(# of all postings)^2 (β_7)	−2.587E−7** (0.134E−7)	H1D	Yes
Log VIC postings (log) (β_1)	0.203** (0.042)	H2A	Yes
# of quality postings × log VIC postings (β_3)	−0.007** (0.000)	H3A	Yes
# of all postings × log VIC postings (β_6)	0.003** (0.000)	H3B	Yes
Prior VIC usage	31.135 (253.570)		
Other control variables not reported+			
Observations	571,456		
Log likelihood	−26,205		

+To control for inherent cross-sectional difference across VICs and stocks, we include fixed effects in the above model. This captures all time-invariant factors such as stock type that may affect users' evaluation of a particular discussion board. We also include user prior usage of VICs in the above model to control for dependence among user decisions.

**$p < 0.05$.

network externality literature (e.g. Katz and Shapiro 1995). We posit that the constant marginal value could be due to the dynamic nature of the stock market and the discussion boards.[12] New events occur frequently, and a large discussion board may be able to capture more information without much duplication. Moreover, it is likely that additional postings may result in better intelligence, clarification, debate, or interpretation for an inherently uncertain stock market. These reasons may contribute to the nondecreasing marginal value of an additional posting.

We also find that users incur substantial information-processing costs in using VICs, which supports Hypothesis 1C. The coefficient of the number of postings (i.e., β_5) is negative, which indicates that the user valuation of discussion boards decreases with the number of postings after controlling for posting quality. Comparing the coefficient of the number of postings ($\beta_5 = -0.010$) with the coefficient on quality posting ($\beta_2 = 0.030$), we find that the baseline cost of processing a posting is about one-third the baseline value embedded in a quality posting. Moreover, the marginal cost for information processing increases with the posting volume, because the coefficient of the quadratic term of posting volume (i.e., β_7) in Table 4 is negative.[13] This supports Hypothesis 1D and is aligned with the IO literature (e.g., Fournier 1996, Shenk 1997), where it is argued that the cost of processing information increases substantially with the volume of information. This is also consistent with earlier findings that large social networks result in increased costs for accessing resources in these communities (Butler 2001).

Table 5 Quality and Discussion Board Size

Variables	Estimates	Hypotheses	Supported
VIC posting volume (log) (δ_1)	−0.12** (0.01)	H2B	Yes
Discussion board posting volume (δ_2)	−0.00008** (0.00)	H2B	Yes
Observations	2,603		
R-square	11.35%		

The above results indicate that both posting volume and quality play an important role in determining user valuation of VICs. Ideally, a VIC provider would like to increase posting volume as well as to improve posting quality to attract users. However, our results indicate that there is a trade-off between posting volume and quality (Hypothesis 2B). Table 5 presents the results for this trade-off. The analysis shows that quality is negatively related to the overall *VIC posting volume* and the individual *discussion board posting volume*. The coefficient of −0.12 on the log of *VIC posting volume* suggests that, every time the VIC increases its total posting volume by 10%, the quality of postings is reduced by 1.2%. This is consistent with our hypothesis that VICs have a greater ability to improve posting quality when they are small, but that their ability to do so decreases significantly as they grow larger.

The trade-off between posting volume and quality suggests that VICs may pursue differentiated strategies in attracting users. The differentiation among

[12] We thank an anonymous reviewer for suggesting this explanation.

[13] The low value of the estimate is not a concern since the estimate is for the square of the number of postings.

VICs could be gauged by examining the coefficient of *VIC posting volume* (i.e., β_1) in Table 4. We find that the coefficient is positive, indicating that user valuation of a discussion board increases with the number of postings in the hosting VIC after controlling for individual discussion board characteristics, such as posting volume and quality. This suggests larger VICs provide additional value beyond what is available through their discussion boards.[14] As we explained earlier, this is because larger VICs have the advantages of economies of scale and can, therefore, afford to provide more value and services.[15]

The differentiation among VICs suggests that various VICs may attract users with different characteristics. We note from Table 4 that the coefficient of the moderating effect of *VIC posting volume* on the number of quality postings (β_3) is negative. This suggests that users of larger VICs value quality postings less than do those using smaller VICs. Likewise, the coefficient of the moderating effect of *VIC posting volume* on number of postings in a stock discussion board (β_6) is positive. This suggests that users of larger VICs have lower information-processing costs than those who use smaller VICs. Putting the two effects together, we can infer that as a result of VIC differentiation, they attract different types of users, supporting Hypotheses 3A and 3B.

The results of our analysis suggest that competition among VICs is more nuanced than a simple head-on battle for posting volume. An increase in the posting volume brings both positive and negative externalities. Moreover, there is an inherent trade-off between posting volume and quality. Consequently, VIC providers need to balance the benefits of increased posting volume with higher information-processing costs and lower quality. The trade-off leads VICs to adopt differentiation strategies. Some choose to focus on posting volume, while others choose to emphasize higher quality. We find that the differentiation strategies work well for VIC users by attracting different types of users.

[14] It is also possible that a larger VIC may provide one-stop experience or high brand reputation, which may also lead to higher valuation. Unfortunately, it is impossible to separate these effects from the informational effect.

[15] There may be an alternative explanation. The finding could also be explained by users' desire to join a larger VIC with more people to interact with. However, we note that if users are interested in interacting with other users, the effect should be more relevant at the stock message board level than at the VIC level. Given that the finding is established at the VIC level after controlling for effects at the stock message board level, we believe the alternative explanation is not plausible.

5. Discussion and Conclusions

While a significant body of research examines the content of virtual community postings and the motivations of these postings, much less is known about how these discussion 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 posting volume is not the only factor in the competition. Quality is an equally important factor in determining the user valuation of VICs and the competition among VICs.

This research makes two important contributions: First, the study sheds light on how users value communities and how their valuations affect both positive and negative externalities in virtual communities. We find that the posting volume by itself is a cost factor since more postings require more information processing from users. A network only becomes more valuable with the number of quality postings. By simultaneously including the number of quality postings and the total number of postings, we are able to directly compare the value and cost of using VICs. Our computations estimate the value of a quality posting to be about three times the cost of processing a posting. We also show that discussion boards exhibit increasing marginal costs, indicating that there is a limit to the growth of discussion boards by balancing the positive and negative externalities. Our results extend prior literature on networks and virtual communities (e.g. Butler 2001, Mackie-Mason and Varian 1994, Asvanund et al. 2004) by showing the important role that information quality plays in determining the direction of network externalities.

Second, we extend the understanding of the competition in virtual communities by examining the differentiation among them. The presence of trade-off between posting volume and quality indicates that a virtual community will have difficulty dominating on all fronts. This requires virtual community providers to pursue differentiation strategies. Our

results provide guidance for both established and new VICs. While posting volume has implications for community providers because of the increased revenue from advertisers, larger volume does not necessarily imply increased value to users. We note that larger and established virtual communities need to pay close attention to improving quality. Our findings indicate that the overall quality on large virtual communities is surprisingly low, even though many VICs have adopted processes to manually or automatically screen out abusive or offensive postings. This implies that an endeavor limited to removal of abusive or offensive postings has a minor impact on posting quality, because many low-quality postings are neither offensive nor abusive. More advanced methods need to be incorporated to evaluate and identify quality postings and give users more options for filtering the irrelevant ones. Our analysis also suggests that there are opportunities for new entrants and small VICs to compete effectively with established virtual communities, but this requires them to pursue a differentiated strategy by focusing on information quality.

Although not discussed extensively in the paper, the algorithmic methodology provided here is generic enough to be applied to a broad variety of other applications. For example, for user reviews on Yahoo!Movie, this framework could be used to identify the quality of movie reviews. Similarly, this approach could be applied to analysis of Amazon reviews and ratings or to forums with only qualitative feedback. As the use of unstructured text becomes increasingly critical for research, the algorithmic methodology presented here could be an important component of researchers' data-gathering strategies.

Results of our study are to be interpreted carefully in the presence of 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 confidence in the findings. Second, our MNL model implicitly assumes that VIC users are aware of the existence of the three VICs, while in reality not all online users are aware of the choices. Controlling for a user's prior VIC usages mitigates this issue; however, more detailed data on users' awareness of different virtual communities will improve the significance of the results. Third, we hypothesized strategic behavior of VIC providers without direct knowledge of the actual strategies employed by these providers. Instead, we infer strategic behaviors indirectly from user valuation of virtual communities. While this approach is widely used in economic studies, knowledge of the actual strategies employed by VIC providers is likely to provide additional insights.

Finally, the generalizability of results of this study across different types of communities will depend on the characteristics of that virtual community. VICs are information-exchange communities with weak ties among community members. The lack of social pressure is one of the main obstacles in maintaining quality in these communities. We expect our findings to offer insights into other information exchange–based virtual communities with weak ties among community members, such as online product review communities, which are known for attracting false reviews from manufacturers, retailers, and competitors. On the other hand, our findings are less applicable to online communities, such as open source communities, with strong ties among community members and where reputation and peer review process act as powerful mechanisms to prevent irrelevant postings. These social ties and powerful screening mechanisms deter members from self-interest behavior that could be damaging to the community (Roberts et al. 2006).

Our study also opens up several opportunities for future studies. First, future research can validate and extend the findings of this study by examining competition in virtual communities that are largely focused on providing social benefits (e.g. MySpace, Xanga) to their members. Second, in the context of VICs, 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. Third, the analysis can be expanded to virtual communities in different domains with different competitive characteristics. For instance, in P2P auctions there is only one dominant player, while other P2P networks compete with a substantial number of other networks. It is worthwhile to study why competition is prominent in some virtual communities, but not in others. Finally, information quality in virtual communities appears to

have a significant impact on the value of virtual communities. However, prior studies have focused more on the quantity than on the quality of interaction and information sharing in virtual communities. Future work that looks into the information quality dimension within virtual communities could be fruitful.

Acknowledgments
Prabhudev Konana and Balaji Rajagopalan acknowledge support from the National Science Foundation's Information Technology Research Grants IIS-0218988 and IIS-0219107, respectively. The authors acknowledge initial help from Chan-Gun Lee (researcher at Intel), Dharmaraj Karthikeyan, and Matt Wimble, and feedback from participants in the Americas Conference on Information Systems, 2004. The authors thank the senior editor, associate editor, and anonymous reviewers for helpful comments and suggestions throughout the review process.

Appendix
Analyzing text messages and classifying the emotive content of messages posted on VICs pose several challenges: (a) It is difficult to arrive at a common understanding or interpretation of the content due to the subjective nature some messages have even for humans (Foltz et al. 1999); (b) messages generally do not observe proper grammatical rules and spelling. In developing our automated classifier, we built on the approach used by Das and Chen (2001). We implemented a decision tree classifier based on readability analysis (Foltz et al. 1999) to categorize messages. Furthermore, we applied evolutionary computing methods (Holland 2000) to induce classification rule sets.

We use five classifiers to classify posting quality: lexicon-based classifier (LBC), readability-based classifier (RBC), weighted lexicon classifier (WLC), vector distance classifier (VDC), and differential weights lexicon classifier (DWLC). The details of these classifiers are provided below.

Lexicon-Based Classifier (CL1)
LBCs have been effectively used in earlier studies (Das and Chen 2001). To build an 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

$$\sum \text{Count}(m_l, \text{Key}_{ij}) = \max_{k=1,2,3} \left\{ \sum \text{Count}(m_l, \text{Key}_{kj}) \right\}.$$

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 (CL2)
RBCs are based on 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. Notice that the thresholds were empirically derived in our study.

Weighted Lexicon Classifier (CL3)
WLCs are a variation of LBCs. WLC overcomes the drawback of LBC, which induces a bias for classes with a higher number of keywords. To eliminate the bias, WLC bases its classification on Max$[n(k_i)/N_i]$, 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

$$\sum \text{Count}\left(m_l, \frac{\text{Key}_{ij}}{N_i}\right) = \max_{k=1,2,3} \left\{ \sum \text{Count}\left(m_l, \frac{\text{Key}_{kj}}{N_k}\right) \right\}.$$

Vector Distance Classifier (CL4)
According to Das and Chen (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 (CL5)
DWLCs represent 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

$$\sum \text{Weight}_{ij} \times \text{Count}(m_l, \text{Key}_{ij}) = \max_{k=1,2,3} \left\{ \sum \text{Weight}_{kj} \times \text{Count}(m_l, \text{Key}_{kj}) \right\}.$$

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.

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

Table A1 A Sample 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

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 A1. 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. The grammar rule set was also constructed based on a manual analysis of 100 messages from all categories.

References

Adar, L., B. Huberman. 2000. Free riding on Gnutella. Technical report, Xerox ARC (August).

Amit, R., C. Zott. 2001. Value creation in e-business. *Strategic Management J.* **22** 493–520.

Antweiler, W., M. Frank. 2004. Is all that talk just noise? The information content of Internet stock discussion boards. *J. Finance* **59**(3) 1259–1295.

Asvanund, A., K. Clay, R. Krishnan, M. Smith. 2004. An empirical analysis of network externalities in peer-to-peer music sharing networks. *Inform. Systems Res.* **15**(2) 155–174.

Bagozzi, R., U. M. Dholakia. 2002. Intentional social action in virtual communities. *J. Interactive Marketing* **16**(2) 2–21.

Balasubramanian, S., V. Mahajan. 2001. The economic leverage of the virtual community. *Internat. J. Electronic Commerce* **5**(3) 103–138.

Burnett, G. 2000. Information exchange in virtual communities: A typology. *Inform. Res.* **5**(4). http://informationr.net/ir/5-4/paper82.html.

Butler, B. S. 2001. Membership size: Communication activity, and sustainability: A resource-based model of online social structures. *Inform. Systems Res.* **12**(4) 346–362.

Butler, B., L. Sproull, S. Kiesler, R. Kraut. 2005. Community effort in online groups: Who does the work and why? S. Weisband, L. Atwater, eds. *Leadership at a Distance: Interdisciplinary Perspectives*. Lawrence Erlbaum Associates, Mahwah, NJ.

Chen, P., L. Hitt. 2002. Measuring switching costs and the determinants of customer retention in Internet-enabled businesses: A study of the online brokerage industry. *Inform. Systems Res.* **13**(3) 255–274.

Chevalier, J. A., D. Mayzlin. 2006. The effect of word of mouth on sales: Online book reviews. *J. Marketing Res.* **43**(3) 345–354.

Cummings, J. N., B. Butler, R. Kraut. 2002. The quality of online social relationships. *Comm. ACM* **45**(7) 103–108.

Das, S. R., M. Y. Chen. 2001. Yahoo! for Amazon: Sentiment parsing from small talk on the Web. *Proc. 8th Asia Pacific Finance Assoc. Annual Conf.* (July 22–25), Bangkok, Thailand.

Das, S. S., A. Martinez-Jerez, P. Tufano. 2005. e-information: A clinical study of investor discussion and sentiment. *Financial Management* **34**(3) 103–137.

Economides, N. 1996. The economics of networks. *Internat. J. Indust. Organ.* **16**(4) 673–699.

Farrell, J., G. Saloner. 1986. Installed base and compatibility: Innovations, product pre-announcements, and predation. *Amer. Econom. Rev.* **76** 940–955.

Fischer, G., C. Stevens. 1991. Information access in complex, poorly structured information spaces. S. P. Robertson, G. M. Olson, J. S. Olsen, eds. *Reaching Through Tech.: CHI 91 Conf. Proc.*, ACM Press, New York, 63–70.

Foltz, P. W., D. Laham, T. K. Landauer. 1999. Automated essay scoring: Applications to educational technology. *Proc. EdMedia '99*, Seattle, WA (June 19–24).

Fournier, J. 1996. Information overload and technology education. *Proc. 7th Internat. Conf. Soc. Inform. Tech. Teacher Educ.*, Phoenix, AZ, 380–385.

Gamon, M., A. Aue. 2005. Automatic identification of sentiment vocabulary: Exploiting low association with known sentiment terms. *Proc. ACL Workshop Feature Engrg. Machine Learn. NLP*, Ann Arbor, MI (June 29–30) 57–64.

Gensch, D. H., W. W. Recker. 1979. The multinomial, multiattribute logit choice model. *J. Marketing Res.* **16**(1) 124–132.

Hof, R. 2005. The power of us: Mass collaboration of the Internet is shaking up business. *Business Week* (June 20) 74–82.

Holland, J. 2000. Building blocks, cohort genetic algorithms, and hyperplane-defined functions. *Evolutionary Comput.* **8**(4) 373–391.

Honeycutt, C. 2005. Hazing as a process of boundary maintenance in an online community. *J. Comput.-Mediated Comm.* **10**(2).

Hu, M., B. Liu. 2004. Mining and summarizing customer reviews. *Proc. 10 ACM SIGKDD Internat. Conf. Knowledge Discovery and Data Mining*, Seattle, WA (Aug. 22–25) 168–177.

Jones, Q., G. Ravid, S. Rafaeli. 2004. Information overload and the message dynamics of online interaction spaces: A theoretical model and empirical exploration. *Inform. Systems Res.* **15**(2) 194–210.

Katz, M. L., C. Shapiro. 1986. Technology adoption in the presence of network externalities. *J. Political Econom.* **94**(4) 822–841.

Katz, M. L., C. Shapiro. 1995. Network externalities, competition, and compatibility. *Amer. Econom. Rev.* **75** 424–440.

Kauffman, R., J. McAndrews, Y.-M. Wang. 2000. Opening the "black box" of network externalities in network adoption. *Inform. Systems Res.* **11**(1) 61–82.

Konana, P., S. Balasubramanian. 2005. The social-economic-psychological model of technology adoption and usage: An application to online investing. *Decision Support Systems* **39**(3) 505–524.

Kraut, R. E., W. Scherlis, T. Mukhopadhyay, J. Manning, S. Kiesler. 1996. Homenet: A field trial of residential Internet services. *Comm. ACM* **39**(12) 55–63.

Krishnan, R., M. Smith, Z. Tang, R. Telang. 2003. The virtual commons: Why free-riding can be tolerated in peer-to-peer networks. *Workshop Inform. Systems Econom.*, College Park, MD (May 29–30).

Kuk, G. 2006. Strategic interaction and knowledge sharing in the KDE mailing list. *Management Sci.* **52** 1031–1042.

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

MacKie-Mason, J. K., H. Varian. 1994. The economic FAQs about the Internet. *J. Econom. Perspectives* **8**(Summer) 75–96.

McFadden, D. 1987. Regression-based specification tests for the multinomial logit model. *J. Econometrics* **34** 63–82.

Moreland, R. L., J. M. Levine. 2001. Socialization in organizations and work groups. M. E. Turner, ed. *Groups at Work: Theory and Research*. Lawrence Erlbaum Associates, Mahwah, NJ, 69–112.

Moreland, R. L., L. Argote, R. Krishnan. 1996. Socially shared cognition at work: Transactive memory and group performance. J. Nye, A. Brower, eds. *What's Social About Social Cognition? Research on Socially Shared Cognition in Small Groups*. Sage, Thousand Oaks, CA.

Pang, B., L. Lee, S. Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. *Proc. Conf. Empirical Methods Natural Assoc. Computational Linguistics Language Processing* (*EMNLP*), Philadelphia, PA (July 6–7) 79–86.

Prentice, D. A., D. T. Miller, J. R. Lightdale. 1994. Asymmetries in attachments to groups and to their members: Distinguishing between common-identity and common-bond groups. *Personality Soc. Psych. Bull.* **20**(5) 484–493.

Rajagopalan, B., P. Konana, C. Lee, M. Wimble. 2004. Extracting relevance from virtual investing-related community postings. *Proc. 10th Americas Conf. Inform. Systems*, New York (Aug. 5–8).

Rheingold, H. 1993. *The Virtual Community: Homesteading on the Electronic Frontier*. Harper Collins, New York.

Roberts, J., I. H. Hann, S. Slaughter. 2006. Understanding the motivations, participation, and performance of open source software developers: A longitudinal study of the Apache projects. *Management Sci.* **52**(7) 984–999.

Sassenberg, K. 2002. Common bond and common identity groups on the Internet: Attachment and normative behavior in on-topic and off-topic chats. *Group Dynam.* **6**(1) 27–37.

Shaked, A., J. Sutton. 1983. Natural oligopolies. *Econometrica* **51** 1469–1483.

Shankar, V., B. L. Bayus. 2003. Network effects and competition: An empirical analysis of the home videogame industry. *Strategic Management J.* **24** 375–384.

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

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

Strahilevitz, L. J. 2003. Charismatic code, social norms, and the emergence of cooperation on the file-swapping networks. *Virginia Law Rev.* **89**(3) 505–595.

Vandenbosch, M., C. B. Weinberg. 1995. Product and price competition in a two-dimensional vertical differentiation model. *Marketing Sci.* **14**(2) 224–249.

Wysocki, P. D. 1999. Cheap talk on the Web: The determinants of postings on stock discussion boards. Working Paper 98025, University of Michigan, Ann Arbor, MI.