MIS REVIEW

An International Journal

Vol.20 No. 1 September 2014

MIS REVIEW Volume 20 Number 1 September 2014

airiti press.

GPN 2007800019
ISSN 1018-1393
DOI: 10.6131/MISR
9 771018 139006 0 9

MIS Review **Vol. 20, No. 1** Published in September 2014

Publisher: Samuel Chuen-Lung Chen

Published by: National Chengchi University, Department of Management Information Systems & Airiti Press Inc.

Editor-in-Chief: Eldon Y. Li

Executive Editor: Shari S.C. Shang

Assistant Editor: Laurence Fang-Kai Chang

Printed by: Sincere Digital Printing, Co.

Cover Designer: Thomes Chen

Typesetting: Hsiao-Hsuan Wang

Publication Office:

National Chengchi University, Department of Management Information Systems
No. 64, Sec. 2, ZhiNan Rd., Wenshan District, Taipei City 116, Taiwan

Airiti Press Inc.
18F., No. 80, Sec. 1, Chenggong Rd., Yonghe District, New Taipei City 23452, Taiwan

Order Information:

Airiti Press Inc.
18F., No. 80, Sec. 1, Chenggong Rd., Yonghe District, New Taipei City 23452, Taiwan
Tel: +886-2-29266006
Fax: +886-2-29235151
E-mail: press@airiti.com

Price: NT$ 400

e-Journal: http://www.airitilibary.com

DOI: 10.6131/MISR

ISSN: 1018-1393

GPN: 2007800019

Printed in Taiwan

Editorial Board

Editor's Introduction

This year marks the 30th anniversary of Department of Management Information Systems at National Chengchi University. Meanwhile, we also celebrate the 20th volume of MISR since its debut in 1988. In this festive issue, we are delighted to present four research papers. The summary of the four papers is as follows.

Mary K. Foster and Richard Michon in their paper "Insights into Motivation to Participate in Online Surveys" argue that more and more marketing research is being conducted using online surveys and the response rate is an issue because of the importance of these data for business decision-making. The study uses a sample of 1,501 from an existing opt-in online survey research panel to gain insight into the motivations of participating in online research, and the right incentives for participation. The findings suggest that respondents are motivated by their perceived level of expertise to offer relevant information, familiarity with and trust toward the sponsors of the survey, the propensity for sharing and participation in social media, sponsors' valuing their opinions through feedback, and sponsors' addressing privacy concerns appropriately. Further, the study segments responses by their type and frequency of social media use. Those with high participation and high information needs are motivated by all of the factors identified. In contrast, those who mostly socialize on social media are motivated by familiarity with sponsors, the opportunity to share online, and having privacy expectations met. Those who use social media mostly to seek information are motivated to participate by trust in sponsor, and having privacy expectations met. The types of incentives that work best to increase participation are consistent with the motivations identified: information about the nature and enforcement of privacy protection policies; ability to earn points toward rewards for quality of online contributions; and enforcing an online code of conduct. These results are of interest to marketing researchers who identify strategies for improving participation that are within managerial control and are not dependent on intrinsic characteristics of the participants.

Dinesh Kumar Saini and Sanad Al Maskari in their paper "Data Management Issues and Data Mining of Real Time System Application for Environment Monitoring" argue that environment pollution monitoring and control is a critical problem for the whole world. The aim of the paper is to present the challenges surrounding environmental data sets and to address these in order to develop solutions. Environmental data sets present a number of data management challenges including data collection, integration, quality and data mining. Environment data sets are also very dynamic and this presents additional challenges ranging from data gathering to data integration, particularly as these data sets are normally very large and expanding continuously. Statistical methods are an effective and economical way to analyze small, static data sets but they are not applicable for dynamic, real-time and large data sets. The use of data mining methods to discover hidden knowledge in large datasets therefore presents great potential to improve environmental management decisions. A representative environmental data set from quantitative air quality monitoring instruments has been assessed and will be used to demonstrate some of the issues in applying data mining approaches to monitoring data quality.

Jagvinder Singh, Adarsh Anand, Avneesh Kumar and Sunil Kumar Khatri in their paper “A Discrete Formulation of Successive Software Releases Based on Imperfect Debugging” state that software reliability is the major dynamic attribute of the software quality, so gaining reliability of software is a vital issue for software products. Due to intense competition, the software companies are coming with multiple add-ons to survive in the pure competitive environment by keeping an eye on existing systems in operational phase. Software reliability engineering is focused on engineering techniques for timely add-ons/upgrades and maintaining software systems whose reliability can be quantitatively evaluated. In order to estimate as well as to predict the reliability of software systems, failure data need to be properly measured by various means during software development and operational phases. Although software reliability has remained an active research subject over the past 35 years, challenges and open questions still exist. The paper presents a discrete software reliability growth modelling framework for multiple upgrades including the concept of two types of imperfect debugging during software fault removal process. The proposed model has been validated on real data set and provides fairly good results.

Prakash Kuppuswamy and Saeed Q. Y. Al-Khalidi in their paper “Securing E-Commerce Business Using Hybrid Combination Based on New Symmetric Key and RSA Algorithm” aim to explore how security in e-commerce is becoming more topical as the traditional shopping and transactions are moving away from physical stores to online stores. E-commerce has had a drastic effect on the global economy and has rapidly accelerated over the years into trillions of dollars each year. Protecting online payment users and application systems requires a combination of managerial, technical and physical controls. In the paper, they propose hybrid cryptographic system that combines both the symmetric key algorithm, and popular RSA algorithm. The symmetric key algorithm, which is based on integer numbers, and the RSA algorithm are widely used in all data security application. Efficiency of the combined security method is better than each individual method.

As the final note, we would like to thank all the authors and reviewers for their collaborative efforts to make this issue possible. It is our sincere wish that this journal become an attractive knowledge exchange platform among information systems researchers. Last but not least, to our loyal readers around the world, we hope you find the contents of the papers useful to your work or research.

Dr. Eldon Y. Li
Editor-in-Chief and University Chair Professor

Department of Management Information Systems
College of Commerce
National Chengchi University
Taipei, Taiwan
Fall 2014

MIS Review

September 2014 Vol.20 No.1

Contents

Research Articles

MIS Review Vol. 20, No. 1, September (2014), pp. 1-30
DOI: 10.6131/MISR.2014.2001.01

Insights into Motivation to Participate in Online Surveys

Mary K. Foster, Richard Michon
Ted Rogers School of Management, Ryerson University, Canada

ABSTRACT: *More marketing research is being conducted using online surveys. Response rate is an issue because of the importance of these data for business decision-making. This study uses a sample of 1,501 from an existing opt-in online survey research panel to gain insight into the motivations to participate in online research, and to identify the right participation incentives. The findings suggest that respondents are motivated by their perceived level of expertise to offer relevant information, familiarity with and trust toward the sponsors of the survey, the propensity for sharing and participation in social media, sponsors' valuing their opinions through feedback, and sponsors' addressing privacy concerns appropriately. Further, the study segments responses by their type and frequency of social media use. Those with high participation and high information needs are motivated by all of the factors identified. In contrast, those who mostly socialize on social media are motivated by familiarity with sponsors, the opportunity to share online, and having privacy expectations met. Those who use social media mostly to seek information are motivated to participate by trust in sponsor, and having privacy expectations met. The types of incentives that work best to increase participation are consistent with the motivations identified: information about the nature and enforcement of privacy protection policies; ability to earn points toward rewards for quality of online contributions; and enforcing an online code of conduct. These results are of interest to marketing researchers because they identify strategies for improving participation that are within managerial control and are not dependent on intrinsic characteristics of the participant.*

KEYWORDS: *Online Survey, Participation Motivation, Incentives to Participate, Increasing Online Response Rate*

1. Introduction

Internet penetration in North America is among the highest in the world. Marketing researchers are leveraging the advantages of these high levels of Web 2.0 access to transition research designs from traditional telephone surveys and personal interviews to online survey panels and communities. Given the increasing use of online technology to gather information, and the importance of consumer opinions and experiences in driving decisions about the type and range of products offered, marketers are interested in understanding how to engage consumers in sharing their opinions and increasing the quantity and quality of participation in online surveys.

There are concerns however about online surveys, namely non-probability samples and response rates affecting data quality. The purpose of this paper is to gain insight into the motivations to participate in online surveys and to identify the right participation incentives, which in turn may increase response rates. The paper is structured as follows:

- Changes in communication technology and administration of survey questionnaires are first analyzed.
- Survey participants from a national opt-in panel are initially segmented by the type and intensity of social media activities. Validated measurement scales describing a full range of social media activities are used in the clustering process (Foster, Francescucci & West, 2012; Li & Bernoff, 2008)
- The study captures the motivations of each segment to participate in online surveys. Measurement scales are submitted to psychometric analysis and aggregated into motivational constructs.
- The research explores various motivational incentives that might enhance segments' participation in online surveys.

2. Survey modes

Pollsters and marketing researchers have been forced to modify their survey practices and adjust to alternate data gathering platforms. For example, households' use of landlines is slowly but inexorably declining. According to National Center for Health Statistics, four-in-ten U.S. adults owned only a cellphone in 2013 (Blumberg & Luke, 2013). The decline of land line household penetration hurts random digit dialing (RDD) sampling frames, affecting coverage and introducing response biases. The Pew Research Center announced that 60 percent of national pools interviews are now administered by cellphones and 40% on landline phones (McGeeny & Keeter, 2014). However, cell phone surveys generate even lower response than traditional landline surveys, take twice longer to administer and cost 2.5 times more (American Association for Public Opinion Research [AAPOR], 2010).

Internet household penetration in North America is now higher than that of land lines, reaching nearly 80 percent of the population (Internet World Stats, 2012). There are many advantages associated with online surveys such as lower costs, flexible survey questionnaire designs and administration tools, personalized email pre-notification and reminders, and simplified data handling (Boyer, Adams & Lucero, 2010; Dillman, Smyth & Christian, 2009; Israel, 2011; Monroe & Adams, 2012). There are concerns, though, about the validity of non-probabilistic samples of opt-in panels and lower response rates to online surveys.

2.1 Non-probabilistic samples

Baker et al. (2013) have doubts about the validity of probability samples when coverage is low or non-response is high. These issues are not exclusive to online but to all forms of surveys. With the constant decline in coverage and non-response, some researchers (e.g., Groves, 2006; Savage & Burrows, 2007) wonder about the acceptability of non-probability sampling methods. In 2011, the American Association of Public Opinion Research (AAPOR) launched a task force "to examine the conditions under which various survey designs that do not use probability samples might still be useful for making inferences to a larger population" (www.aapor.org).

Probability sampling neutralizes exogenous covariates through randomization. The problem for non-probabilistic sampling is to identify and control exogenous variables associated with the object of the study in the sample selection. The validity of non-probabilistic approaches depends on the appropriateness of the theoretical frameworks and the quality of the variables used for respondent selection, and post hoc adjustment (Baker et al., 2010).

Ansolabehere and Schaffner (2014) explain that survey mode differences reported in the literature occur for a number of reasons. These studies are based on data collected five or more years ago when the techniques for constructing, matching and weighting opt-in Internet panels were not fully developed and that Internet usage among the public was not as it is today. Comparing simultaneous multi-mode national political surveys, Ansolabehere and Schaffner (2014) observe that a carefully executed opt-in Internet panel produces estimates that are as accurate as a telephone survey and that the two modes differ little in their estimates of other political indicators and their correlates. "Overall, our findings indicate that an opt-in Internet survey produced by a respected firm can produce results that are as accurate as those generated by a quality telephone poll and that these modes will produce few, if any, differences in the types of conclusions researchers and practitioners will draw in the realm of American public opinion" (Ansolabehere & Schaffner, 2014).

2.2 Response rates

Shih and Fan (2008) conducted a meta-analysis of thirty-nine studies to compare response rates from Web and mail surveys. Their findings reveal that mail surveys obtain a 10% higher response rate than Web surveys, although response rate differences vary considerably. In a another meta-analysis of 45 published and unpublished experimental comparisons between Web and other survey modes, Manfreda et al. (2008) note that Web surveys yield an 11% lower response rate compared to other modes. Similarly, Kim, Yu and Schwartz (2013) compare response rates of an online and face-to-face version of a daily visitor survey at five popular tourist attractions and find that the former have a 45% response rate and the latter a 62% response rate. This discrepancy goes beyond marketing research studies. In a study of course evaluations, Guder and Malliaris (2010) find that response rates drop by 26%, when the university switches from paper-based evaluations to online evaluations.

The completion rate gap between telephone and web survey has been narrowing in recent years, mainly because of the decline in telephone survey response rate, from 36 percent in 1997 to 9 percent in 2012 (Kohut et al., 2012). The difference in completion rate is significantly smaller with opt-in or panel members, as opposed to one-time participants (Manfreda et al., 2008). Many authors have looked at various methods to increase participation rate in online surveys (e.g., Bosnjak et al., 2008; Fan & Yan, 2010). Suggestions for enhancing response rates are applicable to all forms of surveys, including online questionnaire administration. They range from questionnaire design to personal invitation, reminders, and incentives.

Given the increasing use of online technology to gather information, and the importance of consumer opinions and experiences in driving decisions about the type and range of products offered, marketers are interested in understanding how to engage consumers in sharing their opinions and increasing the quantity and quality of participation in online surveys. This study looks at motivational incentives to complete online surveys.

3. Conceptual framework

The review of literature discusses research on motivation first, and then focuses on social media usage (e.g., Kahle & Valette-Florence, 2012; Lorenzo-Romero, Constantinides & Alarcón-del-Amo, 2012; Pagani, Hofacker & Goldsmith, 2011). Three streams of previous research provide the foundation for insights into motivation to participate in online surveys: (1) participation in online knowledge sharing communities of practice; (2) joining and participating in online social networks; and (3) joining web-based survey research panels. Although these three streams take different perspectives on online participation, they present similar results in that each identifies knowledge sharing, trust and reciprocity as strong drivers of joining and participating in a range of online activities. The review of literature is organized around these three common factors.

3.1 Knowledge sharing

The first strong driver identified by various researchers is knowledge sharing. The literature suggests that the knowledge sharing concept includes sub-topics related to the value of information, expertise, social connections and doing good. These sub-topics have application to motivation to participate in online social networks and online surveys.

3.1.1 Value of information

The first sub-topic under knowledge sharing focuses on the value of the information shared, and on the value of the information sharing process (Connolly & Bannister, 2008; Fitzgerald, 2004; Jun, Hu & Peterson, 2004). Bruggen and Dholakia (2010) investigate personality traits and their relationship with joining web-based survey panels, participating in online surveys, and expending effort on survey responses. They find that “need for cognition” or the enjoyment of thinking and learning (Cacioppo & Petty, 1982), “curiosity” or the need to investigate and seek information (Kashdan, Rose & Fincham, 2004), “openness” or the ability to adjust beliefs, and attitudes in light of new information attained (John, 1990) are positively associated with either or both of joining web panels and participating in online surveys.

3.1.2 Expertise

The second sub-topic under knowledge sharing is viewing it as a mechanism to demonstrate expertise. Wu and Sukoco (2010) conceptualize online participation in terms of McClelland’s (1987) work on achievement, affiliation and power. Online participants share knowledge as a way of expressing personal competency and expertise (Ardichvili, Page & Wentling, 2003; Brown & Duguid, 2000). Further, using the online space to demonstrate expertise increases an individual’s power, as he/she is more likely to gain recognition for his/her knowledge and to be able to influence others with the information shared (Bagozzi & Dholakia, 2006; Fuller, Jawecki & Muhlbacher, 2007; Sokolowski et al., 2000). A slightly different view on expertise and online participation is presented by Han et al. (2009) who position it in the context of self-perception theory (Tybout & Yalch, 1980), that is, whether individuals see themselves as the type who responds to surveys because of their ability to make a contribution.

Social cognitive theory also offers insights into expertise as part of the knowledge sharing motivation for participation in online surveys. The foundation of this theory is that social networks and a person’s expectations and beliefs influence behavior. Concepts at the core of the theory include self-efficacy and outcome expectation, and research shows that both are influential in knowledge sharing (Hsu et al., 2007; Kankangalli, Tan & Wei, 2005; Wang & Lai, 2006). Self-efficacy is the judgment of one’s ability to organize and execute given types of performances, while outcome expectation is the judgment of the likely consequences such performances produce (Chiu, Hsu & Wang, 2006). Self-efficacy has been applied to knowledge management to validate the effect of personal efficacy belief in knowledge sharing. According to knowledge sharing self-efficacy, a knowledge producer must have the perceived capability to contribute knowledge as the desire to share knowledge is not enough. Those with higher perceived self-efficacy are more willing to share knowledge (Pagani et al., 2011; Wang & Lai, 2006). Outcome expectation includes intrinsic factors such as recognition or pleasure derived from sharing knowledge, and extrinsic factors such as monetary reward. Researchers find that for those active in the online space, outcome expectation refers to expectations such as recognition, respect, reputation, and making friends. Results show that if participants believe that knowledge sharing increases their reputation or improves relationships they are more likely to share knowledge (Hsu et al., 2007).

3.1.3 Social connections

The third sub-topic within the knowledge sharing concept is making social connections. Wu and Sukoco (2010) investigate this as a motivator for online participation as affiliation (McClelland, 1987). They suggest that maintaining close and friendly relationships with others in the online space through knowledge sharing is an important motivator for participation, as is the perception of one's being responsible and co-operative (Han et al., 2009).

Chiu et al. (2006) use social cognitive and social capital theories to investigate the social perspective on the willingness of online participants to share knowledge. Results indicate that it is the features of social capital -- namely, ties between individuals, reciprocity and group identification through shared language and shared vision -- that increase quantity of knowledge sharing. The more social interactions undertaken by online participants, the greater intensity, frequency, and breadth of information exchange (Larson, 1992). Chiu et al. (2006) find that social capital factors such as social interaction and trust lead to a higher level of knowledge sharing in terms of both quality and quantity of knowledge shared. Improved social relationships also seem to motivate people to participate in the online space; the ability to interact with others online increases trust; and in turn, people are more comfortable in sharing knowledge.

3.1.4 Doing good

The final sub-topic in the knowledge sharing concept involves doing good or altruism. This construct is the degree to which a person is willing to increase other people's welfare without expecting returns. In terms of knowledge contribution, this means contributing knowledge without the outcome expectation of reciprocity. Research shows that those who feel good about contributing knowledge to help others tend to be more motivated to do so in an online environment. While Kankanhalli et al. (2005) and Wasko and Faraj (2005) find that enjoyment in helping others positively influences knowledge contribution, Wang and Lai (2006) are not able to replicate those findings.

3.2 Trust

Trust is the second of the common motivators identified across the three streams of literature. In terms of online participation, it means the reliability and confidence with which one views the online activity and those associated with it. Wasko and Faraj (2005) conceptualize online trust in terms of social capital. They define social capital as "resources embedded in a social structure that are accessed and/or mobilized in purposive action" (p. 38), and suggest that these social resources provide the conditions necessary for knowledge exchange to occur and can lead to greater knowledge sharing. Chiu et al. (2006) and Wasko and Faraj (2005) categorize relationships into three types: (1) *Structural* is the presence or absence of social interaction; (2) *relational* refers to trust, norms, reciprocity, and identification in that the person has a positive feeling toward the community; and, finally, (3) *cognitive* refers to a shared vision in terms of collective goals of members of a group and shared language through a common understanding of collective goals.

Kankanhalli et al. (2005) use relational capital to explain knowledge exchange and contend that trust, norms, and identification are social capital since they are organizational resources or assets rooted within social relationships. Broadly, social capital theory suggests that trust, shared norms and values among those engaged in online activities motivate knowledge sharing (Best & Krueger, 2006; Fassott, 2004, Song & Walden, 2007; Yoon, 2002), and this study proposes that these motivational constructs may also impact propensity to participate in online research.

Researchers describe the trust construct as a significant motivator in online participation (Corritore, Kracher & Wiedenbeck, 2003; Lin, Hung & Chen 2009; Ridings, Gefen & Arinze, 2002; Usoro et al., 2007). Because of the lack face-to-face social cues in online activities, cultivating trust is both important and more difficult. When others confide personal information in online activities, trust is higher. In addition, in a trusting environment people are more inclined to help others and request help from others (Ridings et al., 2002; Usoro et al., 2007). Similarly, Lin et al.'s (2009) study shows that trust significantly affects knowledge sharing self-efficacy, which positively affects knowledge sharing. Ridings et al. (2002) explore the antecedents and effects of trust in online activities. The study measures two dimensions of trust -- ability and benevolence/integrity. Both dimensions increase through perceived responsive relationships in the online space, by a general disposition to trust, and by the belief that others confide personal information.

Most research on trust focuses on its role in promoting online engagement. Privacy concerns, in contrast, are about inhibiting participation in online communities and sharing knowledge because of a lack of trust. This includes issues related to security and the confidentiality and anonymity of information collected (Dommeyer & Gross, 2003; Han et al., 2009; Youn & Lee, 2009). Research shows that people want assurances that information and surveys are used for stated purposes only and that adequate measures exist to protect privacy and provide security. Not addressing these concerns may inhibit online sharing of knowledge and information and reduce trust (Ardichvili, 2008; Hsu et al., 2007; Phelps, D'Souza & Nowak, 2001).

3.3 Reciprocity

The final common factor identified in the literature is reciprocity, which can also be conceptualized as feedback. According to social exchange theory, "individuals engage in social interaction based on an expectation that it will lead in some way to social rewards" (Wasko & Faraj, 2005, p. 39). People share knowledge with the expectation that they will receive rewards, which may include approval, status, and respect (Kankanhalli et al., 2005; Wasko & Faraj, 2005). People share knowledge if they believe it increases their reputation (Wasko & Faraj, 2005). Likewise, increasing the benefits of and decreasing the cost of knowledge sharing encourages knowledge sharing (Dillman, 2000; Kankanhalli et al., 2005).

Participants share knowledge and provide feedback because there is an expectation that doing so will be useful to the knowledge sharer and at some point the favor will be returned (Han et al., 2009). The greater the anticipated reciprocity in a relationship, the more favorable is the attitude toward knowledge sharing. Further, receiving feedback from others through online participation provides mutual benefit thereby increasing the desire to share knowledge (Chiu et al., 2006). The link between the norm of reciprocity and trust is less clear. While Lin et al. (2009) find that the norm of reciprocity is a key determinant of trust in knowledge sharing, Wasko and Faraj (2005) and Chiu et al. (2006) find that reciprocity is not a significant predictor of knowledge contribution. Fahey, Vasconcelos and Ellis (2007) find that introducing rewards into online activities actually damages the exchange of knowledge because the economic self-interest rather than moral obligation becomes a more important motivator.

In summary, researchers identify a number of motivational constructs related to knowledge sharing that may have application to online survey research participation. These include the value of the information shared, trust and shared norms, self-efficacy and perceived expertise, feedback and reciprocity, altruism, social interaction and privacy concerns. One issue that has not been addressed by previous research is whether these motivational constructs are differentially important within online user groups.

3.4 Social media user groups

Social media is an emerging field and the tendency has been to dichotomize behavior into users and non-users, assuming that users represent one homogenous group. However, as participation has increased and as the options for networking and communicating online have proliferated in terms of type, scope and device, researchers are interested in investigating and understanding the nuances of online behavior. Almost three quarters of U.S. online adults use social networking sites, 71 percent of which are on Facebook (Duggan & Smith, 2013). Users of social networking sites and instant messages have been studied from a variety of perspectives, including both psychology and usage behavior.

3.4.1 Personality traits

Nadkarni and Hofmann (2012) systematically review the literature on the psychological factors contributing to Facebook use. They identify 42 studies focusing on the identity construction (demographics and personality characteristics) of registered Facebook users. They also look at the influence of the use of Facebook on narcissism and self-esteem, and the role of Facebook in acting as an avenue for self-presentation and self-disclosure. Facebook use is primarily determined by the need to belong and the need for self-presentation. These needs can act independently and are influenced by a host of other factors, including the cultural background, socio-demographic variables, and personality traits, such as introversion, extraversion, shyness, narcissism, neuroticism, self-esteem, and self-worth (Nadkarni & Hofmann, 2012).

Social media participation has also been studied under the Big-Five model for assessing broad level personality traits such as extraversion, neuroticism, openness to experiences, agreeableness, and conscientiousness (Ehrenberg et al., 2008; John & Srivastava, 1999). Research based on the Big-Five (e.g., Amichai-Hamburger, 2002; John & Srivastava, 1999; Ross et al., 2009), suggests that extraversion, emotional stability and openness to experience are related to uses of social applications on the Internet. Extraversion and openness to experiences are positively related to social media use. Controlling for socio-demographics (age and gender) and life satisfaction, emotional stability appears to be a negative predictor (Correa, Hinsley & De Zúñiga, 2010). Men with greater degrees of emotional instability are more likely to be regular users. The correlation between extraversion and social media is stronger among young adults. On the other hand, openness to new experiences appears as a significant predictor of social media use for the more mature segment (Correa et al., 2010).

3.4.2 Behavioral usage

Marketing and social media researchers conceptualize online behavior in terms of tasks performed (Li & Bernoff, 2008). Ip and Wagner (2008) base their framework on the frequency of using social media regardless of the type of task or activity. Others investigate usage in terms of motivation, such as exchange (Hersberger, Murray & Rioux, 2007) or benefits (Wasko & Faraj, 2005). Less than 10% of users, sometimes classified as Creators or Bloggers, are behind 75% of all user generated content. Creators are a new category of social influencers (Fennemore, Canhoto & Clark, 2011). Forrester Research (http://www.forrester.com) has developed a proprietary social technographics profile to segment users of technology and social media sites. The firm created a technographics hierarchy ladder where self-described Creators, Conversationalists, Critics, Collectors, Joiners, Spectators and Inactives are put on different rungs (Li & Bernoff, 2008)

Foster et al.'s study (2012) combines some of these approaches to identify four segments of social media user groups that differ in terms of the nature of their online activities and the frequency of participation. The first group, Social Media Technology Mavens participate highly in both information type and social type online activities when compared to the other segments. The second segment, the Minimally Involved group, is less likely to participate in all types online activities compared to others. The third segment, the Info Seekers are more likely to be involved in passive, information-search types of online activities such as reading the comments of others, but are less engaged when it comes to more active social activities. Socializers are higher than info seekers when it comes to more social interaction, such as posting comments to the social network pages of others, but are less involved in posting informational content. The previous research revealed that these four segments differ not only in terms of online behavior, but also in the reported motivations for engaging in online activities. This study examines the motivations of the four social media user group segments to participate in knowledge sharing through online survey research, and uses that understanding to test the appeal of various strategies for increasing online participation.

4. Research hypotheses and methodology

Social media activities include data browsing, socializing and user-generated contents. These activities can be represented on a continuum similar to Forrester's technographics ladder. Browsing social media sites is an antecedent to socializing and user-generated contents. Social media users form a diverse community and can be segmented along their social network behavior.

H_1: Social media activities enable well delimited actionable social media user segments.

H_2: Each social media user segment exhibits different motivations to participate in online surveys.

H_3: Each social media user segment is likely to respond to specific online survey participation incentives.

The research model is presented in Figure 1. Social media activities shape social network user segments. These segments have different motivations to join opt-in survey panels. They are also likely to respond to varied survey participation incentives.

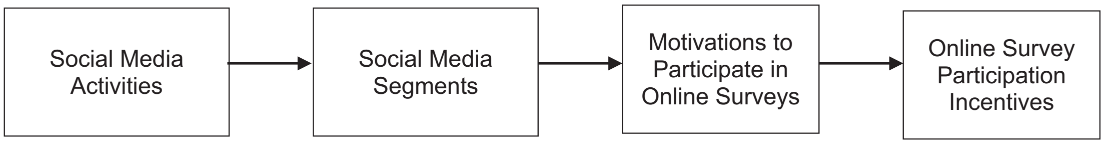
Social Media Activities
Social Media Segments
Motivations to Participate in Online Surveys
Online Survey Participation Incentives

Figure 1 Theoretical Framework

The total sample is 1,501. Respondents belong to an existing opt-in survey research panel comprised of reward plan members of a major airline and two large retailers. In return for participation in online surveys, respondents receive reward points in the program in which they are registered.

Those responding to the survey are representative of the demographics of the Canadian population in terms of regional distribution, age and gender, according to the latest information from Statistics Canada. Online respondents have been a member of this panel for an average of 47 months.

The survey instrument is designed to cover three areas: (1) social media activities, from which we derive our social media user group segments, and which are based on previous research; (2) motivation to participate, in which we test the applicability concepts described in the review of literature as drivers for an individual to participate online, using existing items where available; and (3) participation incentive items are taken from the industry best practice and are tested against underlying motivators.

5. Findings

5.1 Social media activities

In order to identify user segments, Foster et al.'s (2012) original measurement scale was updated for content sharing sites, microblogging services, location-based services, smartphones, and brand social media sites reflecting changes in technology. The initial set of core items for identifying user segments is from the list of online social technology activities developed by Li and Bernoff (2008). One of the limitations in using questions from previous studies on social media activities is that this is a dynamic field and in order to be relevant the items included have to be constantly updated in response to innovations in technological capabilities and accessibilities.

Table 1 describes the scale items with their psychometric properties, sources, factor loadings and cross-loadings. The three constructs, information seeking, socializing and active participation, have alpha coefficients of .77, .83, and .82, respectively. The measurement scales are subjected to a confirmatory factor analysis (Figure 2) using

Table 1 Social Media Activity Level

Measurement Scales Factor Loadings and Alpha Coefficients	Posting Contents	Socializing	Info Seeking
Active Participation/Posting Contents (Alpha = .817)			
1. Posting content to content sharing sites such as Tumblr, Digg, Reddit, Technorati or YouTube	**.814**	.178	.258
2. Posting to a micro-blogging service such as Twitter	**.801**	.223	.207
3. Publishing or updating your personal web page (excluding social networking sites)	**.783**	.231	.147
Socializing (Alpha = .829)			
1. Visiting social networking sites, such as Facebook, LinkedIn or MySpace	.069	**.894**	.139
2. Maintaining/updating your profile on a social networking site	.357	**.770**	.204
3. Posting comments to someone else's social networking page/account	.394	**.763**	.218
Information Seeking (Alpha = .765)			
1. Reading customer ratings and/or product/ service reviews	.232	.049	**.837**
2. Using a search engine to find information prior to a product or service purchase	.086	.241	**.822**
3. Reading online forums, blogs and discussion groups written by others	.335	.249	**.669**

Source: Foster et al. (2012); Li and Bernoff (2008).

AMOS, yielding an adequate fit ($\chi^2 = 35.59$, df = 16, CFI = .997, RMSEA = .026, RMR = .012).

Error correlations illustrate that: (1) socialization indicators correlate with information seeking; and (2) active participation indicators correlate to some extent with socialization. Error correlations emanating from the first socializing indicator ("*Visiting social networking sites*, *such as Facebook*, *LinkedIn or MySpace*") are negative. Proactive social media participants do more than just visit social networking sites, and those who browse social media are not necessarily proactive participants.

It appears that the three latent constructs are highly correlated and nested within each other: Information seeking → Socialization → Active Participation (Posting contents). Information seeking is likely to be an antecedent of socialization. Social media browsers can seek information without necessarily engaging is socialization. On the other

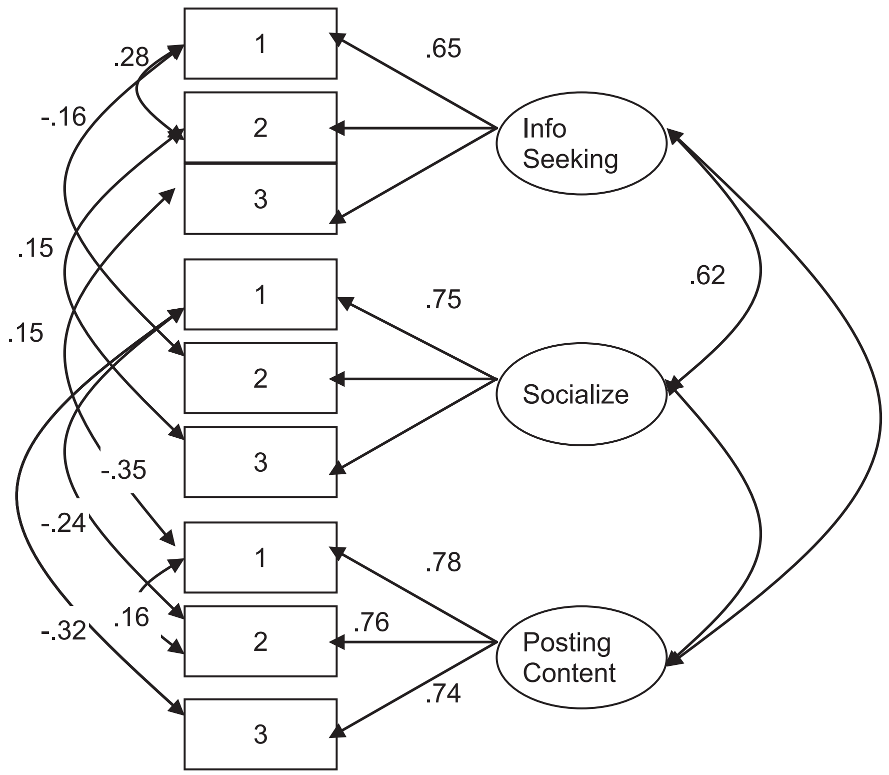

Figure 2 CFA on Social Media Activity Level

Note. Method: Maximum Likelihood, Standardized Coefficients, all *p*-values < .05. χ^2 = 32.592; DF = 16; χ^2/DF = 2.037; CFI = .997; RMSEA = .026; RMR = .012.

hand, active socializers are also likely to be information seekers. In turn, socialization is antecedent to active social media participation. Online community members who socialize do not have to be content creators. The inverse is not highly probable: community members who are regular content providers are more than likely to score high on all types of online social media activities.

5.1.1 Social media behavior segmentation

Having established three distinct types of online behaviors (information seeking, socializing, and creating contents), cluster analysis is performed on factor scores, using and comparing two clustering techniques: the two-step and K-means clustering algorithms. The Bayesian Information Criteria (BIC) from the two-step clustering confirms four optimal clusters. Table 2 shows the final cluster centers from K-means clustering.

The labels for the four clusters are: (1) Social Media Technology Mavens, (2) Info Seekers, (3) Socializers, and, (4) Minimally Involved. The maven group represents 7% of the sample and has the highest score on the active participation construct (posting and

Table 2 Social Media Activity Clusters

	Clusters			
	Mavens	**Socializers**	**Minimally Involved**	**Information Seekers**
N = 1,501	7.4%	27.4%	41.6%	23.6%
Social Media Activity Constructs				
Information Seeking	.589	-.141	-.645	**1.123**
Active on social media	.469	**1.188**	-.560	-.542
Active Participation (Posting Contents)	**2.842**	-.447	.008	-.386

creating contents). It is also high on all other factors. This group participates to a greater extent in all types of online activities and applications than the other three segments identified through clustering. Info Seekers represent 24% of the sample and score high on the need for information, but low on socializing and creating contents. Their focus is on seeking information from others, such as reading comments and reports posted by members of the online community. Socializers account for 27% of the cohort. The main focus of their online activities is to interact with others and maintain social connections on social networking sites. They score low on user-generated contents. Finally, the Minimally Involved group, representing 42% of the sample, is low on both information-seeking behavior and social interaction. The sample originates from an online opt-in panel; therefore, the label minimally involved is relative to this specific cohort.

5.1.2 Segment profiles

The average age of the Social Technology Mavens is 34; they are more likely to be male (71%), and engage in more types of online activities and applications (LinkedIn, Twitter, Slideshare, consumer information sites (e.g., Trip Advisor), and group buying (e.g., GroupOn), online professional and work groups, special interest groups) than other segments. Info Seekers, are more likely to be male (60%), have an average age of 49 and are not heavy users of any online activities or applications, but do report more engagement than other segments except for mavens. Socializers are more likely than other segments to be female (66%) and have an average age of 42. They are heavy users of Facebook, with 41% reporting spending 10 hours per week or more on Facebook, but not on other online activities or applications. The Minimally Involved group has an average age of 52 and is split almost equally between males and females. They are minimal or non-users of all types of online activities and applications.

5.2 Motivation to participate

Motivational constructs are extracted from the extant literature and adapted to the current technology environment. Table 2 describes the motivational constructs, and the supporting literature already reviewed (Chiu et al., 2006; Foster et al., 2012; Hsu & Lin, 2008; Hsu et al., 2007; Kankanhalli et al., 2005; Larson, 1992; Lin et al., 2009; Ridings et al., 2002). Emerging constructs are participants' self-perceived expertise (alpha = .93), familiarity and trust toward survey sponsors (alpha = .86), the propensity for sharing and participation in social media (alpha = .85), valuing sponsors or company feedback (alpha = .88), and concerns for privacy (alpha = .79).

First, the motivational constructs and indicators are validated on the entire cohort (χ^2 = 322.48, df = 122, CFI = .987, RMSEA = .033, RMR = .04). Subsequently, the same CFA is replicated on the various online member segments to investigate multi-group measurement invariance (Table 3 and Figure 3).

Table 3 Survey Participation Motivational Constructs

Measurement Scales Factor Loadings and Alpha Coefficients	**Factors**				
	1	**2**	**3**	**4**	**5**
Alpha coefficient	.928	.861	.845	.884	.793
A. Perceived Expertise (Hsu et al., 2007; Kankanhalli et al., 2005; Lin et al., 2009)					
1. When talking about new products and technologies with others, I am usually the one with the most detailed knowledge.	**.892**	.124	.182	.087	.045
2. I am known as the "go to" person for people who want to hear about the latest trends.	**.881**	.105	.210	.104	.013
3. I pride myself on usually being the first person among my friends to have heard of a new product or technology about to be offered on the market.	**.862**	.101	.231	.114	.016
4. People often ask me for product recommendations.	**.844**	.140	.125	.181	.062
B. Familiarity and Trust (Chiu et al., 2006; Hsu & Lin, 2008; Hsu et al., 2007; Lin et al., 2009; Ridings et al., 2002)					
1. I am more willing to give my honest opinion online when I trust the company I am being asked about.	.088	**.794**	.115	.240	.170
2. I am more likely to give my opinion online when I am familiar with the company whose product/service I am being asked about.	.143	**.790**	.170	.278	.125
3. I am more likely to give my opinion online when the company I am being asked about is revealed to me and not anonymous.	.093	**.770**	.197	.202	.171
4. I am more likely to share my opinion about a company/sponsor if I receive an incentive such as cash or loyalty points.	.169	**.676**	.200	.192	.140

Table 3 Survey Participation Motivational Constructs (continued)

Measurement Scales Factor Loadings and Alpha Coefficients	Factors				
	1	2	3	4	5
C. Sharing and Participation (Chiu et al., 2006; Hsu & Lin, 2008; Hsu et al. 2007; Larson, 1992)					
1. If someone posts personal information about themselves in an online community, I am more likely to trust them.	.153	.176	**.806**	.097	.061
2. Participating in an online community makes it easier for me to meet others who share my interests.	.209	.131	**.788**	.209	.117
3. I share personal information as part of being a member of an online community.	.136	.101	**.788**	.157	.011
4. If I post my opinions in an online community, I will make more friends in the online community.	.290	.276	**.704**	.064	-.021
D. Getting Company Feedback (Chiu et al., 2006; Hsu et al., 2007; Kankanhalli et al., 2005; Lin et al., 2009)					
1. I like it when a company tells me how my feedback has influenced their decisions and/or products after I share my opinion with them.	.169	.306	.188	**.816**	.141
2. When companies give me feedback on how my information influenced their decision-making, I am more likely to share my opinion with them in the future.	.180	.302	.207	**.803**	.145
3. I enjoy sharing my opinion with companies in order to contribute to the betterment of their product.	.153	.345	.160	**.752**	.147
E. Privacy (Foster et al., 2012; Hsu et al., 2007)					
1. I am concerned that companies who sponsor online communities are keeping track of my online activities and using that information for other purposes.	.026	.161	.083	.097	**.842**
2. Concerns about who has access to my posts on an online community make me less likely to share my opinions.	.021	.149	.050	.116	**.808**
3. I feel that it is an invasion of privacy for an online community to keep track of my online activities.	.047	.131	-.005	.100	**.808**

5.2.1 Motivation to participate by segment

The initial CFA on motivational scale items identifies five constructs: (1) Perceived Expertise, (2) Familiarity and Trust, (3) Sharing and Participation, (4) Feedback (Reciprocity), and (5) Privacy. The first order CFA invariance is tested against each social media segment. The multigroup model (Figure 3) outlines structural and loading invariance (Chi-square = 1,169.824, DF = 523, χ^2/DF = 1.767, CFI = .972, RMSEA = .023).

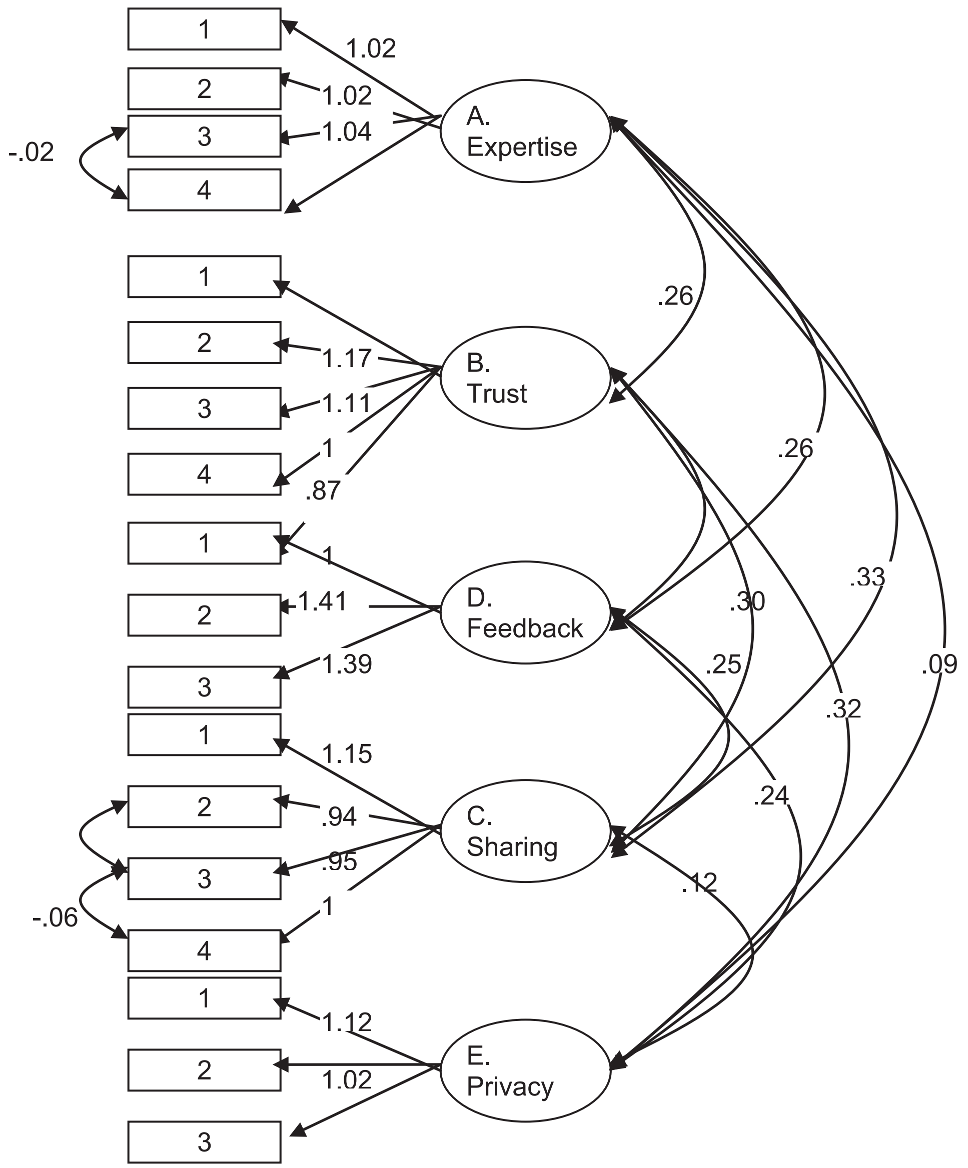

Figure 3 Multigroup CFA on Motivational Constructs

Note. Method: Maximum Likelihood, Unstandardized Coefficients, all *p*-values < .05. χ^2 = 1169.824; DF = 523; χ^2/DF = 1.767; CFI = .972; RMSEA = .023.

All parameters constrained, except for the cross-loading path from Trust to Feedback first indicator ("*I like it when a company tells me how my feedback has influenced their decisions and/or products after I share my opinion with them*") that was set free.

The model is constrained on all parameters with one exception. In the case of the minimally involved, a cross-loading path between Trust and one of the Feedback indicators ("*I like it when a company tells me how my feedback has influenced their* *decisions and/or products after I share my opinion with them*") is set up. The cross-loading path decreases the chi-square by as much as 70, and intuitively links Trust and Feedback together.

Table 4 shows multigroup model comparisons. The unconstrained multigroup has a superior fit (χ^2/DF = 1.82, RMSEA = .023) than the single group CFA (χ^2/DF = 2.64, RMSEA = .033), which supports the multigroup approach. The multigroup CFA with constrained measurement weight (χ^2/DF = 1.77, RMSEA = .023) is marginally better than the unconstrained model. The Akaike Information Criterion (AIC) also indicates a more parsimonious model. The model slightly deteriorates when considering structural covariance invariance (χ^2/DF = 2.06, RMSEA = .027). However, it is more than adequate and still offers a better fit than the single group CFA. The conclusion is that motivational constructs are structurally invariant and quasi measurement invariant across all segments. Therefore, there is a valid measurement tool with which to compare segments.

An ANOVA comparing summated latent motivations to participate in online surveys (Table 5) shows significant differences among social media clusters. For the total sample,

Table 4 Motivation Multigroup CFA

	χ^2	DF	χ^2/DF	CFI	RMSEA	AIC
Single group	322.48	122	2.64	.987	.033	
Multiple groups (4):						
Unconstrained	886.76	487	1.82	.972	.023	1,280.764
Measurement Weights	924.29	523	1.77	.972	.023	1,246.292
Structural Covariance	1,169.82	568	2.06	.958	.027	1,401.824
Saturated model	.000	0		1.000		1,368.000
Independence model	14,929.01	612	24.39	.000	.125	15,073.012

Table 5 Motivational Constructs by Segments

Cluster s	Perceived Expertise	Familiar& Trust	Sharing and Participation	Feedback	Privacy
1. Mavens (7%)	**3.65**	**3.81**	**3.51**	**3.98**	3.52
2. Socializers (27%)	2.57	**3.45**	**2.59**	3.56	3.58
3. Minimally Involved (42%)	**2.20**	**2.78**	**1.93**	**2.99**	**3.26**
4. Info Seekers (24%)	**2.88**	**3.26**	**2.28**	3.51	3.59
Total	2.57	3.15	2.31	3.34	3.44

Note. Summated motivation scales from 1 to 5, where 1 is "Does not describe me at all," and 5 is "Very much describes me." Mean values in bold face, significantly different at p-value < .05.

the motivational factors that score highest include familiarity and trust (3.15 out of 5 -- knowing the name of the sponsor, being familiar with the sponsor, and trusting the sponsor), feedback (3.34 out of 5 -- knowing how the information shared influenced decisions), and privacy (3.44 out of 5 -- ensuring that information is protected and identity is secure when sharing information online). Individuals categorized as mavens perceive themselves as experts on new products, technologies and trends (3.65 out of 5 vs. 2.57 for the entire cohort). As self-perceived experts, they are more than willing to provide their opinions. Mavens, socializers, and info seekers are more likely to participate if they feel familiar (3.81, 3.45 and 3.26 vs. 3.15) with the sponsoring organization. Mavens and socializers also have a greater propensity to share their personal information with other community members (3.51 and 2.59 vs. 2.31). Further, mavens appreciate feedback from companies or sponsoring organizations (3.98 vs. 3.34). On the other hand, the minimally involved group is less likely to value feedback (2.99 vs. 3.34). Finally, the minimally involved group appears less concerned than any other segment with privacy issues (3.26 vs. 3.44).

5.3 Participation incentives

Firms maintaining online panels for surveys have toolboxes of incentives to promote participation. These are empirical techniques developed in the trade that have not been validated in the scientific literature. After conducting key informant interviews with ten senior executives in online marketing research companies to understand current practices and strategies to enhance participation, thirteen possible non-monetary incentives (Table 6) to promote online participation are identified. The next section investigates the overall appeal of each incentive and then the appeal by social media user group.

Online panel and survey managers have batteries of incentives to promote survey participation. The incentives with the highest likelihood of increasing participation in online surveys for the total sample include: (1) Having the policy about the protection of personal information prominently displayed (3.35 out of 5); (2) Allowing members to earn points toward rewards for the quality of their contributions to the online community (3.36 out of 5); and (3) the online community enforcing a code of online conduct (3.39 out of 5).

Social media segments are liable to react differently to non-monetary participation incentives offered by online survey sponsors. Table 6 displays average response of each segment to various incentives. Mavens are self-motivated, and do not report that any of non-monetary incentives will influence their level of participation in online surveys. Socializers and the minimally involved report welcoming: (1) policies about the protection of personal information (means > 3.5 out of 5); (2) enforced codes of online conduct (means > 3.5); (3) the choice to reveal true identity or use an avatar (means > 3.2); and (4) earning rewards for the quality of contributions (means > 3.4). Finally, none of the

Table 6 Non-Monetary Participation Incentives and Social Media Segments

	1 Mavens	2 Socializers	3 Minimally Involved	4 Info Seekers	Total
1. Online community members provide detailed information about interests and buying habits.	2.74	2.67	2.74	2.68	2.69
2. Policy about the protection of personal information is prominently displayed	3.08*	**3.49***	**3.53***	3.18*	3.35
3. Members can earn points toward rewards for the quality of their contributions to the online community.	3.10*	**3.42***	**3.58***	3.24*	3.36
4. Members can spend reward points to access information not available to the general membership.	2.80	3.03	3.00	2.90	2.95
5. Online community provides different ways for members to communicate with each other	2.93*	3.00*	3.11*	2.82*	2.95
6. Online community provides tips about how to make your contribution meaningful	2.82*	3.03*	3.13*	2.98*	3.01
7. Online community allows members with like-minded interests to contact each other.	2.88*	2.95*	2.96*	2.72*	2.87
8. Members can view other members' personal information.	2.61	2.35	2.35	2.40	2.39
9. Members have the choice as to whether they reveal their true identity or use an avatar.	2.92*	3.19*	**3.33***	2.98*	3.12
10. Members are not limited to written posts to share their opinions, but can post pictures, videos etc.	2.97*	2.99*	3.12*	2.75*	2.93
11. The online community enforces a code of online conduct.	3.06*	**3.48***	**3.59***	3.26*	3.39
12. Members have the ability to rate the contributions of others on a specific topic.	2.84*	2.95*	3.06*	2.80*	2.91
13. Sponsor recognizes outstanding contributions online for the whole community to see.	2.95*	3.04*	3.09*	2.90*	2.99

Note. * Sig < .05. Scales 1 to 5, where 1 is much less likely to share opinion and 5, much more likely to share opinion. Values in bold underscore mean responses above 3.3 out of 5.

incentives appeals to info seekers who appear to be relatively impermeable to the various industry offers.

6. Discussion and managerial implications

The research supports H_1 by delineating well defined social media user segments. It replicates the findings of an earlier study by Foster et al. (2012) on a limited sample of university students, which identified four distinct segments among social media users. Because the sample for this study is broader in terms of age and regional distribution than those previous studies, it provides greater ecological validity, as the findings can be generalized to a wider social media community. This study also demonstrates that these constructs are highly correlated and nested within each other. Involvement in social media progresses from information seeking to online socializing to content creation. There is an implied hierarchy in that those who post comments also socialize and seek information; those who socialize also seek information.

The social media segmentation suggests that the younger the individual, the higher the engagement in social media activities. This is not surprising, given that the adoption of new technology and the ease with which technology is integrated into daily life is associated with younger age groups (Ispos Reid, 2012).

Research findings partially support H_2, which posits that each social media user segment exhibits different motivations to participate in online surveys. Some motivations factors apply to all, while others are segment specific.

The data suggest that the top motivators for the entire sample are: privacy -- ensuring privacy concerns are adequately addressed; feedback -- knowing how their information is used by the company and that it was useful; and finally familiarity and trust -- knowing and trusting the sponsor of the survey. The emergence of “Feedback” as a strong motivator is consistent with the literature on Social Exchange Theory and reciprocity in that knowledge sharing increases when the sharer receives some benefit as a result of sharing (Chiu et al., 2006; Han et al., 2009; Kankanhalli et al., 2005; Wasko & Faraj, 2005). “Familiarity and trust” confirms previous research indicating that trust is a significant predictor of virtual community member’s desire to exchange information (Corritore et al., 2003; Lin et al., 2009; Ridings et al., 2002; Usoro et al., 2007). Likewise, it is not surprising that privacy concerns emerge as a deterrent to online survey participation, as other researchers have documented a variety of security, confidentiality, safety and anonymity issues related to using and sharing information through the Internet and particularly social media sites (Han et al., 2009; Hsu et al., 2007; Levin et al., 2008).

What is interesting about the top motivators is that they are within the sponsors' control. They are not dependent on peer response or intrinsic motivators within individuals. If sponsors of online surveys take steps to protect the information provided to them, make their privacy policies prominent and transparent, provide feedback about how information is being used for decision-making, and recognize contributions through rewards and incentives, the findings suggest that these actions positively reinforce motivation to participate.

This study reveals that different motivations are important for different user group categories. Mavens are a highly motivated group and identify all of the motivations -- perceived expertise, familiarity and trust, sharing and participation, feedback and privacy as triggers for their participation in online surveys. This broad range of intrinsic and extrinsic factors is consistent with Foster et al.'s (2012) finding that mavens are motivated to participate in social media as a form of self-expression. They feel competent and confident and believe they have important information to share. This perceived expertise is what differentiates mavens from the other user groups. Likewise, they have high standards relating to the protection of their privacy, the need for meaningful feedback if they provide information, their desire to know to whom they are providing information, and their commitment to enhancing the broader community through online sharing. None of the other segments is motivated as highly or by as many of the factors identified as are the mavens. Two of the five factors are greater than 3.5 out of 5 for Socializers and Info Seekers: feedback and privacy, but neither are statistically significant. None of the factors is greater than 3.5 for the minimally involved. In terms of statistical significance, the data indicate that socializers are significantly more motivated than the total sample by familiarity and trust, and sharing and participation, even though the average score is less than 3.5 out of 5. Info seekers are significantly more motivated than the total sample on perceived expertise, familiarity and trust, and sharing and participation even though the average score is less than 3.5 out of 5. The minimally involved group is significantly lower on all factors than the rest of the sample. The findings suggest that it is not one factor that motivates online participation, but rather a combination of factors that work together. This makes the development and design of promotional and engagement material more complex because all of the factors must be included in the positioning of the importance of participation.

Finally, this study focuses on identifying effective strategies for promoting the quality and quantity of participation on online survey research. Motivations are important to consider in that they can inform strategy necessary to encourage participation in the online space (Lorenzo-Romero et al., 2012). The findings suggest partial confirmation of H_3 (Each social media user segment is likely to respond to specific online survey participation incentives). Some incentives are likely to apply to all, while others would be specific to some segments.

The most effective incentives are linked to the most important motivations. Participants are motivated by “familiarity and trust” and this relates to their report of an increased likelihood to share opinions online if sponsors prominently display how they protect respondents’ personal information. “Feedback” is linked to the reported positive impact of the opportunity for respondents to earn points for the quality of their contributions, which can later be used for rewards. Finally, adequately addressing privacy issues as a motivator are consistent with the positive response to enforcing an online code of conduct (Rao & Quester, 2006). Previous research shows an interesting contradiction. Although social media users report high concerns related to privacy, they do not indicate any intention of changing their online behavior to protect their privacy (Levin et al., 2008), so it is not clear that addressing privacy concerns will increase online participation. For those who have already agreed to be part of an online panel and thus have accepted that some information will be shared, it may mean that failing to present a privacy policy, regardless of its contents, may deter participation, as opposed to needing specific elements of a privacy policy in order to promote participation.

Specific motivations are also linked to specific incentives. Those who are motivated by the need to participate and share are more likely to participate more if the format provides additional opportunities for online sharing and connecting with those of like interests. Those, for whom trust and familiarity are important, are most influenced by options that address privacy concerns and provide rewards for the quality of contributions. Adequately responding to privacy concerns are important for those who report valuing privacy. This suggests that when designing new incentives to increase participation that marketing research firms should start with motivations because there is consistency between motivations and the appeal of particular incentives.

Mavens appear intrinsically motivated to participate online, and thus are not particularly incented to increase their level of participation by any of the options tested. Socializers and the minimally involved are both significantly motivated by displaying privacy protection information, rewards for quality contributions and enforcing an online code of conduct. Info seekers are also motivated by the same items more than other options, but at a lower level than for the other two groups. This suggests that marketing research companies should focus their attention on incentives in these three areas (trust, privacy, feedback) because: (1) they are consistent with the most important motivations for survey respondents; (2) they are consistently appealing, albeit at different levels, across three of the four social medial users groups; and (3) they are within the control of the marketing research company for development and implementation and not dependent on the intrinsic motivations of individuals to be effective. In terms of limitations, while this study was conducted with a representative sample in terms of age, gender and regional distribution, it only includes respondents who are current members of an online survey panel. A broader sample may have yielded different results.

Acknowledgements

The authors acknowledge funding support for this project from the Applied Research and Commercialization Initiative of the Federal Economic Development Agency for Southern Ontario (Fed Dev Ontario).

References

American Association for Public Opinion Research. (2010), 'New considerations for survey researchers when planning and conducting RDD telephone surveys in the U.S. with respondents reached via cell phone numbers', available at www.aapor.org/Cell_Phone_Task_Force_Report.htm (accessed 25 March 2014).

Amichai-Hamburger, Y. (2002), 'Internet and personality', *Computers in Human Behavior*, Vol. 18, No. 1, pp. 1-10.

Ansolabehere, S. and Schaffner, B.F. (2014), 'Does survey mode still matter? Findings from a 2010 multi-mode comparison', *Political Analysis*, Vol. 22, No. 3, pp. 285-303.

Ardichvili, A. (2008), 'Learning and knowledge sharing in virtual communities of practice: motivations, barriers and enablers', *Advances in Developing Human Resources*, Vol. 10, No. 4, pp. 541-554.

Ardichvili, A., Page, V. and Wentling, T. (2003), 'Motivation and barriers to participation in virtual knowledge-sharing communities of practice', *Journal of Knowledge Management*, Vol. 7, No. 1, pp. 64-77.

Bagozzi, R.P. and Dholakia, U.M. (2006), 'Open source software user communities: a study of participation in Linux user groups', *Management Science*, Vol. 52, No. 7, pp. 1099-1115.

Baker, R., Brick, J.M., Bates, N.A., Battaglia, M., Couper, M.P., Dever, J.A., et al. (2013), 'Summary report of the AAPOR task force on non-probability sampling', *Journal of Survey Statistics and Methodology*, Vol. 1, pp. 90-143.

Best, S.J. and Krueger, B.S. (2006), 'Online interactions and social capital: distinguishing between new and existing ties', *Social Science Computer Review*, Vol. 24, No. 4, pp. 395-410.

Blumberg, S.J. and Luke, J.V. (2013) 'Wireless substitution: early release of estimates from the national health interview survey, January-June 2013', available at www.cdc.gov/nchs/data/nhis/earlyrelease/wireless201312.pdf (accessed 23 March 2014).

Bosnjak, M., Neubarth, W., Couper, M.P., Bandilla, W. and Kaczmirek, L. (2008), ‘Prenotification in web-based access panel surveys: the influence of mobile text messaging versus e-mail on response rates and sample composition’, *Social Science Computer Review*, Vol. 26, No. 2, pp. 213-223.

Boyer, C.N., Adams, D.C., and Lucero, J. (2010), ‘Rural coverage bias in online surveys?: evidence from Oklahoma water managers’, *Journal of Extension*, Vol. 48, No. 3, Article 3TOT5, available at http://www.joe.org/joe/2010june/tt5.php (accessed 24 March 2014).

Brown, J.S. and Duguid, P. (2000), *The Social Life of Information*, Harvard Business School Press, Boston, MA.

Bruggen, E. and Dholakia, U.M. (2010), ‘Determinants of participation and response effort in web panel surveys’, *Journal of Interactive Marketing*, Vol. 24, pp. 239-250.

Cacioppo, J.T. and Petty, R.E. (1982), ‘The need for cognition’, *Journal of Personality and Social Psychology*, Vol. 42, No. 1, pp. 116-131.

Chiu, C.M., Hsu, M.H. and Wang, E.T.G. (2006), ‘Understanding knowledge sharing in virtual communities: an integration of social capital and social cognitive theories’, *Decision Support Systems*, Vol. 42, pp. 1872-1888.

Connolly, R. and Bannister, F. (2008), ‘Factors influencing Irish consumers’ trust in internet shopping’, *Management Research News*, Vol. 31, No. 5, pp. 339-358.

Correa, T., Hinsley, A.W. and De Zúñiga, H.G. (2010), ‘Who interacts on the Web?: the intersection of users’ personality and social media use’, *Computers in Human Behavior*, Vol. 26, No. 2, pp. 247-253.

Corritore, C.L., Kracher, B. and Wiedenbeck, S. (2003), ‘On-line trust: concepts, evolving themes, a model’, *International Journal of Human-Computer Studies*, Vol. 58, No. 6, pp. 737-758.

Dillman, D.A. (2000), *Mail and Questionnaire Internet Surveys: The Tailored Design Method*, 2nd ed., Wiley, New York, NY.

Dillman, D.A., Smyth, J.D. and Christian, L.M. (2009), *Internet*, *Mail*, *and Mixed-Mode Surveys: The Tailored Design Method*, 3rd ed., Wiley, New York, NY.

Dommeyer, C.J. and Gross, B.L. (2003), ‘What consumers know and what they do: an investigation of consumer knowledge, awareness, and use of privacy protection strategies’, *Journal of Interactive Marketing*, Vol. 17, No. 2, pp. 34-51.

Duggan, M. and Smith, A. (2013), ‘Social media update 2013’, *Pew Research*, available at http://pewinternet.org/Reports/2013/Social-Media-Update.aspx (accessed 20 March 2014).

Ehrenberg, A., Juckes, S., White, K.M. and Walsh, S.P. (2008), 'Personality and self-esteem as predictors of young people's technology use', *CyberPsychology & Behavior*, Vol. 11, No. 6, pp. 739-741.

Fahey, R., Vasconcelos, A.C. and Ellis, D. (2007), 'The impact of rewards within communities of practice: a study of the SAP online global community', *Knowledge Management Research & Practice*, Vol. 5, pp. 186-198.

Fan, W. and Yan, Z. (2010), 'Factors affecting response rates of the web survey: a systematic review', *Computers in Human Behavior*, Vol. 26, No. 2, pp. 132-139.

Fassott, G. (2004), 'CRM tools and their impact on relationship quality and loyalty in e-tailing', *International Journal of Internet Marketing and Advertising*, Vol. 1, No. 4, pp. 331-349.

Fennemore, P., Canhoto, A.I. and Clark, M. (2011), *An Investigation of Existing and Emerging Segmentation Practices in Online Social Media Networks*, The Henley Centre for Customer Management, Reading, UK.

Fitzgerald, L. (2004), 'The influence of social communication networks on intentions to purchase on the web', *International Journal of Internet Marketing and Advertising*, Vol. 1, No. 2, pp. 137-154.

Foster, M.K., Francescucci, A. and West, B.C. (2012), 'Different strokes for different folks: why different user groups participate in online social media', *International Journal of Internet Marketing and Advertising*, Vol. 7, No. 2, pp. 103-119.

Fuller, J., Jawecki, G. and Muhlbacher, H. (2007), 'Innovation creation by online basketball communities', *Journal of Business Research*, Vol. 60, No. 1, pp. 60-71.

Groves, R.M. (2006), 'Nonresponse rates and nonresponse bias in household surveys', *Public Opinion Quarterly*, Vol. 70, No. 5, pp. 646-675.

Guder, F. and Malliaris, M. (2010), 'Online and paper course evaluations', *American Journal of Business Education*, Vol. 3, No. 2, pp. 131-137.

Han, V., Albaum, G., Wiley, J.B. and Thirkell, P. (2009), 'Applying theory to structure respondents' stated motivations for participating in web surveys', *Qualitative Market Research: An International Journal*, Vol. 12, No. 4, pp. 428-442.

Hersberger, J.A., Murray, A.L. and Rioux, K.S. (2007), 'Examining information exchange and virtual communities: an emergent framework', *Online Information Review*, Vol. 31, No. 2, pp. 135-147.

Hsu, C.L. and Lin, J.C.C. (2008), 'Acceptance of blog usage: the roles of technology acceptance, social influence and knowledge sharing motivation', *Information & Management*, Vol. 45, pp. 65-74.

Hsu, M.H., Ju, T.L., Yen, C.H. and Chang, C.M. (2007), 'Knowledge sharing behavior in virtual communities: the relationship between trust, self-efficacy, and outcome expectations', *International Journal of Human-Computer Studies*, Vol. 65, No. 2, pp. 153-169.

Internet World Stats. (2012), 'Top 20 countries with the highest number of internet users', available at http://internetworldstats.com/top20.htm (accessed 8 March 2013).

Ip, R.K.F. and Wagner, C. (2008), 'Weblogging: a study of social computing and its impact on organizations', *Decision Support Systems*, Vol. 45, pp. 242-250.

Ipsos Reid. (2012), 'The Ipsos Canadian inter@ctive Reid report 2012', available at http://www.ipsos.ca/common/dl/pdf/Ipsos_InteractiveReidReport_FactGuide_2012.pdf (accessed 21 December 2012).

Israel, G.D. (2011), 'Strategies for obtaining survey responses from extension clients: exploring the role of e-mail requests', *Journal of Extension*, Vol. 49, No. 3, Article 3FEA7, available at http://www.joe.org/joe/2011june/a7.php (accessed 15 March 2014).

John, O.P. (1990), 'The big five factor taxonomy: dimensions of personality in national language and in questionnaires', in Pervin, L.A. (Ed.), *Handbook of Personality: Theory and Research*, Guilford, New York, NY, pp. 66-100.

John, O.P. and Srivastava, S. (1999), 'The big five trait taxonomy: history, measurement, and theoretical perspectives', in Pervin, L.A. and John, O.P. (Eds.), *Handbook of Personality: Theory and Research*, 2nd ed., Guilford, New York, NY, pp. 102-138.

Jun, M., Hu, J. and Peterson, R.T. (2004), 'A comparison of information searchers and e-shoppers on the perceptions of e-shopping factors: an assessment', *International Journal of Internet Marketing and Advertising*, Vol. 1, No. 2, pp. 204-228.

Kahle, L.R. and Valette-Florence, P. (2012), *Marketplace Lifestyles in an Age of Social Media: Theory and Methods*, M.E. Sharpe, Armonk, NY.

Kankanhalli, A., Tan, B.C.Y. and Wei, K.K. (2005), 'Contributing knowledge to electronic knowledge repositories: an empirical investigation', *MIS Quarterly*, Vol. 29, No. 1, pp. 113-143.

Kashdan, T.B., Rose, P. and Fincham, F.D. (2004), 'Curiosity and exploration: facilitating positive subjective experiences and personal growth opportunities', *Journal of Personality Assessment*, Vol. 82, No. 3, pp. 291-305.

Kim, N., Yu, X. and Schwartz, Z. (2013), 'Can online surveys substitute traditional modes? An error-based comparison of online and onsite tourism destination surveys', *Tourism Review International*, Vol. 17, pp. 31-45.

Kohut, A., Keeter, S., Doherty, C., Dimock M. and Christian L. (2012), *Assessing the Representativeness of Public Opinion Surveys*, Pew Research Center, Washington, DC.

Larson, A. (1992), 'Network dyads in entrepreneurial settings: a study of governance of exchange relationships', *Administrative Science Quarterly*, Vol. 37, No. 1, pp. 76-104.

Levin, A., Foster, M.K., Nicholson, M.J., West, B., Hernandez, T. and Cukier, W. (2008), *The Next Digital Divide: Online Social Network Privacy*, Ryerson University, Toronto, Canada.

Li, C. and Bernoff, J. (2008), *Groundswell: Winning in a World Transformed by Social Technologies*, Harvard Business Press, Boston, MA.

Lin, M.J.J., Hung, S.W. and Chen, C.J. (2009), 'Fostering the determinants of knowledge sharing in professional virtual communities', *Computers in Human Behavior*, Vol. 25, pp. 929-939.

Lorenzo-Romero, C., Constantinides, E. and Alarcón-del-Amo, M.d.C. (2012), 'Segmenting the social networking sites users: an empirical study', *International Journal of Internet Marketing and Advertising*, Vol. 7, No. 2, pp. 136-156.

Manfreda, K.L., Bosnjak, M., Berzelak, J., Haas, I. and Vehovar, V. (2008), 'Web surveys versus other survey modes: a meta-analysis comparing response rates', *International Journal of Market Research*, Vol. 50, No. 1, pp. 79-104.

McClelland, D.C. (1987), *Human Motivation*, Cambridge University Press, New York, NY.

McGeeney, K. and Keeter, S. (2014), 'Pew Research increases share of interviews conducted by cellphone', *Pew Research*, available at www.pewresearch.org/fact-tank/2014/01/15/pew-research-increases-share-of-interviews-conducted-by-cellphone (accessed 27 February 2014).

Monroe, M.C. and Adams, D.C. (2012), 'Increasing response rates to web-based surveys', *Journal of Extension*, Vol. 50, No. 6, pp. 6-7.

Nadkarni, A. and Hofmann, S.G. (2012), 'Why do people use Facebook?', *Personality and Individual Differences*, Vol. 52, No. 3, pp. 243-249.

Pagani, M., Hofacker, C.F. and Goldsmith, R.E. (2011), 'The influence of personality on active and passive use of social networking sites', *Psychology and Marketing*, Vol. 28, No. 5, pp. 441-456.

Phelps, J.E., D'Souza, G. and Nowak, G.J. (2001), 'Antecedents and consequences of consumer privacy concerns: an empirical investigation', *Journal of Interactive Marketing*, Vol. 15, No. 4, pp. 2-17.

Rao, S. and Quester, P. (2006), 'Ethical marketing in the internet era: a research agenda', *International Journal of Internet Marketing and Advertising*, Vol. 3, No. 1, pp. 19-34.

Ridings, C.M., Gefen, D. and Arinze, B. (2002), 'Some antecedents and effects of trust in virtual communities', *Journal of Strategic Information Systems*, Vol. 11, No. 3-4, pp. 271-295.

Ross, C., Orr, E.S., Sisic, M., Arseneault, J.M., Simmering, M.G. and Orr, R.R. (2009), 'Personality and motivations associated with Facebook use', *Computers in Human Behavior*, Vol. 25, No. 2, pp. 578-586.

Savage, M. and Burrows, R. (2007), 'The coming crisis of empirical sociology', *Sociology*, Vol. 41. No. 5, pp. 885-889.

Shih, T.H. and Fan, X. (2008), 'Comparing response rates from web and mail surveys: a meta-analysis', *Field Methods*, Vol. 20, No. 3, pp. 249-271.

Sokolowski, K., Schmalt, H.D., Langens, T.A. and Puca, R.M. (2000), 'Assessing achievement, affiliation, and power motives all at once: the multi-motive grid (MMG)', *Journal of Personality Assessment*, Vol. 74, No. 1, pp. 126-145.

Song, J. and Walden, E.A. (2007), 'How consumer perceptions of network size and social interactions influence the intention to adopt peer-to-peer technologies', *International Journal of E-Business Research*, Vol. 3, No. 4, pp. 49-66.

Tybout, A. and Yalch, R.F. (1980), 'The effect of experience: a matter of salience?', *Journal of Consumer Research*, Vol. 6, pp. 406-413.

Usoro, A., Sharratt, M.W., Tsui, E. and Shekhar, S. (2007), 'Trust as an antecedent to knowledge sharing in virtual communities of practice', *Knowledge Management Research & Practice*, Vol. 5, pp. 199-212.

Wang, C.C. and Lai, C.Y. (2006), 'Knowledge contribution in the online virtual community: capability and motivation', *Lecture Notes in Computer Science*, Vol. 4092, pp. 442-453.

Wasko, M.M. and Faraj, S. (2005), 'Why should I share? Examining social capital and knowledge contribution in electronic networks of practice', *MIS Quarterly*, Vol. 29, No. 1, pp. 35-57.

Wu, W.Y. and Sukoco, B.M. (2010), 'Why should I share? Examining consumers' motives and trust on knowledge sharing', *The Journal of Computer Information Systems*, Vol. 50, No. 4, pp. 11-19.

Yoon, S.J. (2002), ‘The antecedents and consequences of trust in online-purchase decisions’, *Journal of Interactive Marketing*, Vol. 16, No. 2, pp. 47-63.

Youn, S. and Lee, M. (2009), ‘The determinants of online security concerns and their influence on e-transactions’, *International Journal of Internet Marketing and Advertising*, Vol. 5, No. 3, pp. 194-222.

About the authors

Mary K. Foster is a Professor of Marketing at Ted Rogers School of Management at Ryerson University in Toronto Canada. Her research interests include social media, online social networks, cyberbully and privacy.

Corresponding author. Ted Rogers School of Management, Ryerson University, 350 Victoria Street, Toronto ON, M5B 2K3. Tel: 416-970-5000 ext. 6734. E-mail address: mfoster@ryerson.ca

Richard Michon is an Associate Professor of Marketing at Ted Rogers School of Management at Ryerson University in Toronto Canada. His research interests include big data, research design, retail atmospherics and retail marketing. E-mail address: rmichon@ryerson.ca

MIS Review Vol. 20, No. 1, September (2014), pp. 31-43
DOI: 10.6131/MISR.2014.2001.02

Data Management Issues and Data Mining of Real Time System Application for Environment Monitoring

Dinesh Kumar Saini[1,2], Sanad Al Maskari[3]
[1]*Faculty of Computing and Information Technology, Sohar University, Oman*
[2]*Research Fellow and Adjunct Faculty, School of ITEE, University of Queensland, Australia*
[3]*School of ITEE, University of Queensland, Australia*

ABSTRACT: *Environment pollution monitoring and control is very big problem for the whole world. Taking decision in the environment is becoming more challenging. The aim of this paper is to present the challenges surrounding environmental data sets and to address these in order to develop solutions. Environmental data sets present a number of data management challenges including data collection, integration, quality and data mining. Environment data sets are also very dynamic and this presents additional challenges ranging from data gathering to data integration, particularly as these data sets are normally very large and expanding continuously. Statistical methods are very effective and economical way to analyze small, static data sets but they are not applicable for dynamic, real-time and large data sets. The use of data mining methods to discover hidden knowledge in large datasets therefore presents great potential to improve environmental management decisions. A representative environmental data set from quantitative air quality monitoring instruments has been assessed and will be used to demonstrate some of the issues in applying data mining approaches to poor data quality.*

KEYWORDS: *Data Management, Real Time Systems, Data Mining, Environment Monitoring Systems*

1. Introduction

Huge amount of data are generated by environmental sensors every day across the globe. There is tremendous need for data analysis systems which are able to mine massive and continuous stream of real world data applications such as temperature monitoring, air pollution, stock market, network security, etc. Data generated by environmental sensors are recorded at time intervals of seconds through to minutes and over time these sensors will create datasets that need to be mined in real time in a way that takes into considerations the dynamic nature s of the real world changes that are being measured. Without appropriate analysis methods that allow inferences to be derived based on patterns observed within these data sets it will not be possible to lead to new knowledge discoveries (Fayyad, Piatetsky-Shapiro & Smyth, 2006) defines knowledge Discovery in

Databases as “the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.” In this research we are attempting to identify best possible data mining solutions for environmental data using various algorithms, focusing on streaming data mining methods as these types of data are increasingly common now as a result of rapidly emerging technology developments. In order to do this we have developed an environmental data mining model (EDMM) which covers all critical stages needed to produce high quality data mining results. A key tenant of the model is that data analysis methods must take into consideration the dynamic behaviors of the incoming data streams. This means that statistical methods are not always appropriate for the large, dynamic data sets characterizing many environmental variables. The dynamic nature of such data means that their behaviors change over time and can lead to concept drift. The occurrence of concept drift (Helmbold & Long, 1994) will hinder the accuracy of the original data model thus make it increasingly inaccurate over time. In our model we will take into consideration the effects of concept drift to reflect provide greater data model accuracy.

Air pollution, marine pollution, and sand storms are three major problems that affect agriculture and fishing, which need to be addressed and treated proactively (Al-Maskari, Saini & Omar, 2010). Air pollution is a very critical problem that can lead to catastrophic outcomes if not managed and monitored proactively.

2. Research problem

We took as case of Oman at Sohar Port for the current study of our research. The new born industrial zone introduced a mixture of chemical and petrochemical plants that leads to odorous emissions. Such emissions in a populated area cause nuisance discomfort to citizens, and health issues. The government has installed a network of sensors mobile and static to monitor odorous emissions in the industrial zone. The electronic sensors are provided by Common Invent, Delft, The Netherlands All sensors communicate by wireless or telephone links with a central database management system maintained by Common Invent which handles over 2 million new data entries per day. The data base management system reads and stores the incoming raw data, as well as providing some interpretation of the data and coordinates automated event handling for clients (see http://www.comon-invent.com).

Electronic nose (or e-nose) is the name given to a wide class of instruments capable of measuring odor information in different environments (Saini & Yousif, 2013). The data from the e-noses provided by Comon Invent has been used to distinguish odorous petrochemical vapors, like fuel oil, naphtha, gasoline, jet fuel, from manure, from sewage gas or from VOCs like toluene, benzene, styrene etc. (Bootsma et al., 2012). In Sohar Port the objective is to develop the e-nose data assessment system to be trained to associate digital fingerprints with various odors. Once trained, the e-nose database system can then recognize smells as they arise and inform the user as to their origin once this has been established. The current e-noses deployed at Sohar Port create a record every 3 minutes and stores the raw data in a remote server for the eight different sensors located in each unit. The system also records wind direction, speed, and air temperature, as well as GPS location and time. Currently a total of 7 static sensors and one mobile sensor are installed in the industrial zone with another two sensors installed in an adjacent urban community. Consequently data collected from these distributed sensors forms a large data source, which needs to be cleaned, mined and analyzed in different dimensions to create a usable prediction models and for consumption by different applications.

Data gathered from different environmental sensors needs to be analyzed and assessed appropriately. Any system that handles critical data that can reduce air quality impacts on the adjacent communities has to be credible and effective. Credibility demands that the system is accurate when dispatching a poor air quality alert.. The system will need to monitor and predict odors emissions in a distributed area taking into accounts external elements such as wind, direction and speed, rainfall and dust storms. The system has to be smart enough to issue an alarm based on area of effect and the likely degree of impact. In developing such a system the following are the main challenges faced in this research:

(1) Data integration and data quality issues need to be addressed before applying data mining techniques.

(2) Multiple data sources from different stake holders that need to be integrated.

(3) Odor sensors data sets have limited finger prints which make it hard to create an accurate prediction model.

(4) Current data mining methods used in environmental analysis do not take into considerations the surrounding environment changes and their dynamic behaviors.

(5) Algorithm evaluation in the scope of environmental monitoring is complicated by the complex nature of environment systems.

(6) Using statistical methods to analyze dynamic, real-time and large data sets are not appropriate (Arasu et al., 2003).

3. Research approach

The combination of emerging new semantic web technologies and web services to access, process and integrate data and models held within both centralized and distributed hydrological databases allows:

(1) Identification and prioritization of the key stakeholder user requirements, queries and datasets.

(2) Data quality and data cleansing processes.

(3) Development of the common data models and ontologies to integrate both static and real-time data streams, visual, spatial and temporal data, legacy databases and newly generated datasets.

(4) Prediction of the Air pollution: Using the captured data we intend to create models to identify the air pollution sources, delineate affected areas and estimate next areas to be polluted. A learning system will be employed to enable a better sensing of odors. Beside the data collected from sensors, feedback from experts, researchers, and community will be assessed to see if these improve the predictions. A community enabled sensor network will be considered to improve the odor prediction model.

(5) The data collected by the e-noses will be analyzed using various data mining and machine learning techniques. A prediction model for the air pollution effects will be the core focus of research. In the prediction model we will take into consideration the metrological variables (temperature, wind speed, wind direction, and rain). Clustering, Euclidean, Cosine similarities, ANN, CVFDT, Fuzzy logic and other techniques will be used and compared to achieve best prediction model.

(6) The output of the prediction model will be evaluated against criteria developed for an industry and community alert system.

4. Motivation

Information technology has to be proven to be cost effective and provide a reliable solution if it is to be deployed effectively in environmental management prevention and control (Abdelzaher et al., 2010). The deployment of distributed network sensor has grown rapidly in the past ten years (Hill et al., 2000). Many applications have been implemented using network sensors such as environment monitoring, intrusion detection (Wood & Stankovic, 2000), habitat monitoring (Crumiere, 1999), defense, transportation, and scientific exploration. These sensors produce very large volumes of data that need to be cleaned, stored, and retrieved for analysis and decision making (Li et al., 2000). The huge amount of data produced by these sensors need to be monitored and controlled to provide a meaningful and useful information and it can be used as a feedback loop to management and control systems.

Environment monitoring applications can be very complex because they involve many variables with different dimensions and different scales. Managing, accessing and analyzing data generated from distributed environmental sensors remains a serious challenge for researchers and scientists. The complex data set created by these sensors present many challenges including visualization, data storage and retrieval, data quality and data integrity as well as data analysis and mining.

The ability of sensor network to deliver large amount of data in real-time create a data mining challenge. In our research we are attempting to cover these challenges and provide solutions to such issues. Most data mining algorithm doesn't take into considerations the dynamic behaviors of multi dimensions data generated by sensor networks especially for environmental data. Concept drift must be taken into considerations when analyzing environmental data in real time.

5. Environmental network sensors

The new developments in the area of electronics and wireless network have led to the creation of Environmental Network Sensors (ENS). ENS are large distributed systems communicate through wireless, telephone or satellite networks to stream the data to a central location for processing and analysis (Sohn et al., 2003). A fundamental aspect of environmental network sensors deployment is therefore data management and analysis of these data streams. To enable environmental organizations and authorities to be able to make constructive decisions regarding likely environmental impacts environmental informatics need to be in well shape and in place.

6. Why data mining

Traditional methods for analyzing data using statistical methods are limited to very small data sets and to single users dealing with them directly. Although statistical methods are very economical, simple and effective but they are complicated when trying to apply them to new applications (Friedman, n.d.). Unlike data mining methods, they are not meant to deal with huge, dynamic and real time data sets. Most statisticians will consider 1,000 point as a large data set but in data mining world millions of transactions can be analyzed using data mining algorithms (Roppel et al., 1998).

The evolution of data mining techniques started with the advent of the first computer devices within the university environment and more broadly with the subsequent development of personal computing devices (Pan & Yang, 2007). The improvements of data communication, networks, databases and the ability for users to ubiquitously access data and services in real time has revolutionized the data mining field. Each new technology and development of numerical techniques was based on the previous one. The following shows how we have developed through the data mining era:

(1) 1970s era: Data collection
In this era data was collected using computers, tapes and discs.

(2) 1980s & 1990s: Data Access
The introduction of Relational DBMS, Relational data was implemented, Data RDBMS, advanced data models (extended-relational, object oriented [OO], deductive, etc.) Application-oriented DBMS (spatial, scientific, engineering, etc.)

(3) 2000s: Data mining era
The introduction of the WWW and the maturity of the internet have created massive databases. Data mining algorithms started to emerge and advanced data mining algorithms has been introduced. New applications in multimedia, web mining, streaming data management and mining were deployed and still apply (Crumiere, 1999).

Data mining is mature enough to be applied in environmental applications and other large data applications due to:

(1) Massive data collected for instance large distributed sensors are used to collect large and complex data about the nature and pollution.

(2) We can now access powerful multiprocessors and data storage technologies at reasonable prices.

(3) Many data mining algorithms have been developed.

(4) Data mining is concerned with creating knowledge and information from dynamic and huge data sets. It is a blended field of statistics, machine learning and data bases. Data mining also referred to as Knowledge Discovery in Database (KDD). The definition of data mining varies between different authors based on their own background, experience and views. For example:

- Data mining is the process of exploration and analysis, by automatic or semiautomatic means, of large quantities of data in order to discover meaningful patterns and rules.
- Data mining is the process of extracting previously unknown, comprehensible and actionable information from large database and using it to make crucial business decisions.
- Data mining is finding interesting structure (patterns, statistical models, relationships) in databases.
- Data mining is a decision support process where we look in large databases for unknown and unexpected patterns of information.

The use of data mining techniques in e-nose odor monitoring is still in an early stage. Few modern data mining techniques have been used in e-nose odor monitoring and analysis. Artificial Neural Networks have been applied to predict Sulphur dioxide concentration in Delhi (Green, Chan & Goubran, 2009), to process the signal from odor sensor arrays for near-real-time odor identification (Kahn, Katz & Pister, 1999), using electronic nose to identify the age of spoiled food based to predict piggery odor concentrations (Foster, 2002). In his paper (Aberer et al., 2010) describes the vision of community based sensing using a mobile geo-sensor network (Gehrke & Madden, 2004). The OpenSense project is aiming to investigate air pollution monitoring using community-driven sensing (European Commission, 2001).

7. Environmental data mining model

In this section we will describe our Environmental Data Mining Model (EDM) that we will use to create a prediction model using various data mining algorithms. The EDM has eight essential stages starting from data collection and ending up with decision making and result monitoring (refer to Figure 1). The use of this model will optimize the efficiency and accuracy of our data mining models.

7.1 Data collection

Data collection can be done manually or automatically using sensors. The Sohar Environment Unit (SEU) uses Mobile Air Quality Monitoring Station (MAQMS) which

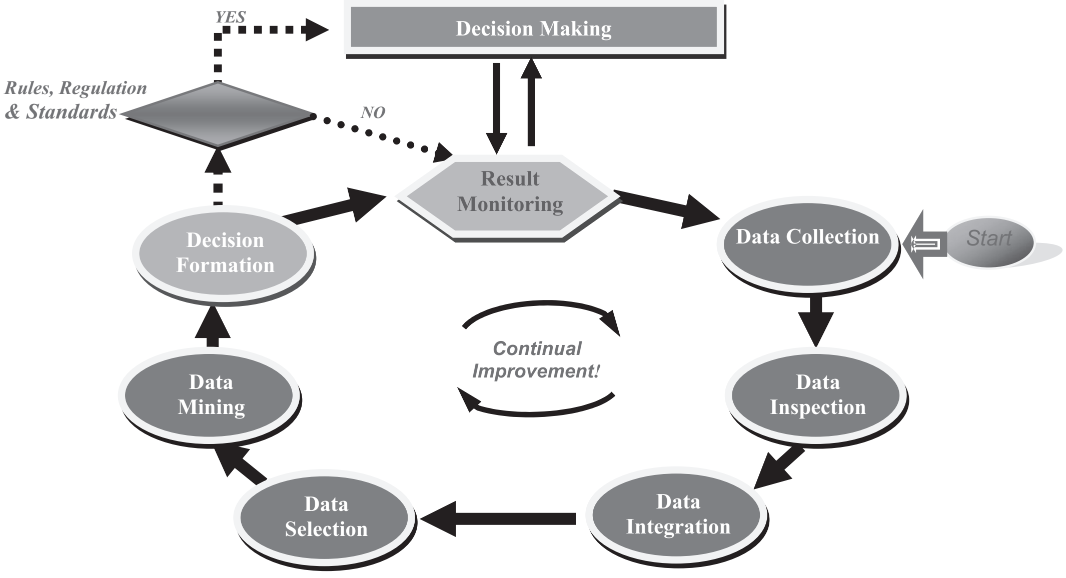

Figure 1 EDDM Model

record data hourly. MAQMS record PM10, O3, CO, SO2, H2S, NO, NOx concentrations with the corresponding meteorological data. The new distributed network sensors (e-Nose) records data every 3 minutes. The E-Nose data are designed to monitor odors around Sohar Industrial Port (SIP). The reading from the E-Nose sensors doesn't refer to any gas instead it writes some values based on the sensor reaction (refer to Table 1).

7.2 Data inspection

Current data collected by SEU are not hosted in centralized warehouses rather it's saved in Excel spreadsheets. This present an issue with data cleansing and integration between distributed sensors. Raw environmental data i is usually not clean, incomplete (empty values), noisy (contains error) and may have inconsistencies (Al-Maskari et al., 2010). The following table illustrates the multi source problems identified in SEU environmental datasets:

The employees of the system have all the rights to expect that the data they are dealing with are indeed correct. Otherwise wrong decisions will be made and the consequences of such decisions can be significant to them and to the environment

Table 1 Data Quality and Cleaning Issues

Problem		Dirty Data	Remarks
Attribute	Missing Values	O3 = "No Data"	Value unavailable during data gathering (null value)
	Different values with same meaning	O3 = Zero, O3 = 0	Both refer to the same value
Record	Violated data type	O3 = Zero, O3 = Span	The value should be numeric
Record Type	Duplicate records	Sensor 1 (Date = 1/7/08, time = 11:00, dust = 0.025) Sensor 1 (Date = 1/7/08, time = 11:00, dust = 0.025)	Duplicate records from different data sources
	uniqueness violation	Sensor 1 (Date = 1/7/08, time = 12:00, dust = 0.016) Sensor 1 (Date = 1/7/08, time = 12:00, dust = 0.025)	The primary key used in both sensors are date and time
	Different time format	Sensor 1 (Date = 29/02/08, time = 0:00) Sensor 2 (Date = 02/29/08, time = 24:00)	Different date and time format is used by different sensors
	Different unit formats	Sensor 1 Dust unit ug/m^3 Sensor 2 Dust unit mg/m^3	Different units are used by different sensors
	Noisy	e-nose S1 = 0, -0.12	Errors and outliers

Poor data quality management of environmental data can lead to the following:

(1) Increase cost as more time will be spent correcting errors rather than performing critical operations.

(2) Poor data quality may lead to poor decision making which lead to incorrect estimate and predictions. E-environment is so sensitive that we should not tolerate any compromise when it comes to decision making or we will endanger or people and planet earth.

(3) More difficult to set strategy and execute it. Environmental strategic decision requires data gathering from various data sources with some uncertain quality. This makes it harder to develop a sound strategic decision. Executing the strategy becomes difficult as inaccurate results become evident.

To overcome the above concerns the following operations will be conducted to the data sets:

(1) Fill in missing values.

(2) Identify outliers and smooth out noisy data.

(3) Correct inconsistent data.

(4) Duplicate identification.

7.3 Data integration

Environmental data is heterogonous by nature therefore combining data from different sources can be a very challenging task. Legacy data, heterogonous data, time synchronization can be a source of problems when integrating data from different sources. Before integrating multiple data sources they must pass data quality checks, otherwise data quality problems will be inherited from the source databases. By looking at Sohar industrial region environmental data we can observe the multi-source cleansing problem (refer to Tables 1 and 2).

Table 2 Data Collection for Various Gases on the Same Day at Different Time

Date	Time	Dust	O3	CO	nCH4	SO2
		mg/m^3	ppb	ppm	ppm	PPB
01/07/2008	10:00	0.37	S <	S <	RS232	3.608
01/07/2008	11:00	0.025	36.75	0.86	RS232	1.249
01/08/2008	0:00	No Data	15.82	0	RS232	0.125

7.4 Data selection

Data inspection and integration is considered to be the most difficult and longest stages in EDMM. After creating integrated clean data sets it's necessary to define application domain for the data mining algorithm. Meta data information, prior knowledge and application goals must be defined. Once they are defined a target data set will be generated. One of our objectives is to predict odors in the SIP area. We will use E-Nose data combined with social network feedback dataset.

7.5 Data mining

In this stage an appropriate data mining approach will be selected. Various data mining approaches exist including association rule mining, clustering, induction and streaming data mining. Once the appropriate approach is selected then a suitable implementation will be applied. Data mining can focus on a variety of areas.

Typical areas to examine are:

(1) habits and behavior,

(2) demographics,

(3) time,

(4) product characteristics.

8. Conclusion

In this paper we introduced our Environmental Data Mining Model which addressed all aspects surrounding environmental datasets. The EDM introduced eight essential stages necessary to create an accurate data mining results. EDM is a continuous improvement model aiming to improve the process of environmental data mining. EDM will help in the improvement of decision making and environment pollution control.

Acknowledgements

The authors of the paper are thankful for the University Research Council and Sohar University, Oman. Authors are thankful for the University of Queensland, school of ITEE. Thanks to Prof Lance Bode and Prof Tony for the Valuable Input.

References

Abdelzaher, T., Blum, B., Cao, Q., Evans, D., George, J., George, S., et al. (2010), 'EnviroTrack: towards an environmental computing paradigm for distributed sensor networks', *Proceedings of the 24th International Conference on Distributed Computing Systems*, Tokyo, Japan, pp. 582-589.

Aberer, K., Sathe, S., Chakraborty, D., Martinoli, A., Barrenetxea, G., Faltings, B., et al. (2010), 'OpenSense: open community driven sensing of environment', *Proceedings of the ACM SIGSPATIAL International Workshop on GeoStreaming*, San Jose, CA, pp. 39-42.

Al-Maskari, S.S., Saini, D.K. and Omar, W.M. (2010), 'Cyber infrastructure and data quality for environmental pollution control in Oman', *Proceeding of International Conference on Data Analysis*, *Data Quality & Metadata Management*, Mandarin Orchard, Singapore, doi: 10.5176/978-981-08-6308-1_D-038.

Arasu, A., Babcock, B., Babu, S., Datar, M., Ito, K., Nishizawa, I., et al. (2003), 'STREAM: the Stanford stream data manager', *IEEE Data Engineering Bulletin*, Vol. 26, No. 1, pp. 19-26.

Bootsma, R.J., Marteniuk, R.G., Mackenzie, C.L. and Zaal, F.T.J.M. (1994), 'The speed-accuracy trade-off in manual prehension: effects of movement amplitude, object size and object width on kinematic characteristics', *Experimental Brain Research*, Vol. 98, No. 3, pp. 535-541.

Crumiere, M. (1999), 'Artificial neural network prediction of ground-level ozone concentration in Palm Beach City', Unpublished master thesis, Folrida Atlantic University, Boca Raton, FL.

European Commission. (2001), *IST 2001: Technologies Serving People*, European Commission, Rue de la Loi, Belgium.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (2006), 'From data mining to knowledge discovery: an overview', in Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. and Uthurusamy, R. (Eds.), *Advances in Knowledge Discovery and Data Mining*, AAAI/MIT Press, Menlo Park, CA, pp. 1-34.

Foster, I. (2002), 'The grid: a new infrastructure for 21st century science', *Physics Today*, Vol. 55, No. 2, pp. 42-47.

Friedman, J.H. (n.d.), 'Data mining and statistics: what's the connection?', available at http://statweb.stanford.edu/~jhf/ftp/dm-stat.pdf (accessed 21 November 2014).

Gehrke, J. and Madden, S. (2004), 'Query processing in sensor networks', *IEEE Pervasive Computing*, Vol. 3, No. 11, pp. 46-55.

Green, G.C., Chan, A.D.C. and Goubran, R.A. (2009), ‘Identification of food spoilage in the smart home based on neural and fuzzy processing of odour sensor responses’, *Proceedings of Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Minneapolis, MN, pp. 2625-2628.

Helmbold, D.P. and Long, P.M. (1994), ‘Tracking drifting concepts by minimizing disagreements’, *Machine Learning*, Vol. 14, pp. 27-45.

Hill, J., Szewczyk, R., Woo, A., Hollar, S., Culler, D. and Pister, K. (2000), ‘System architecture directions for network sensors’, *ASPLOS*, Vol. 35, No. 11, pp. 93-104.

Kahn, J.M., Katz, R.H. and Pister, K.S.J. (1999), ‘Next century challenges: mobile networking for “smart dust”’, *Proceedings of ACM/IEEE International Conference on Mobile Computing and Networking*, Seattle, WA, pp. 271-278.

Li, Y., Callahan, T., Darnell, E., Harr, R., Kurkure, U. and Stockwood, J. (2000), ‘Hardware-software co-design of embedded reconfigurable architectures’, *Proceedings of the Design Automation Conference*, Los Angeles, CA, pp. 507-512.

Pan, L. and Yang, S.Y. (2007), ‘A new intelligent electronic nose system for measuring and analyzing livestock and poultry farm odours’, *Environment Monitoring and Assessment*, Vol. 135, pp. 399-408.

Roppel, T.A, Padgetta, M.L, Waldemark, J. and Wilson, D. (1998), ‘Feature-level signal processing for near-real-time odor identification’, in Dubey, A.C. and Harvey, J.F. (Eds.), *The SPIE Conference on Detection and Remediation Technologies for Mines and Minelike Targets III*, Orlando, FL, pp. 13-17.

Saini, D.K. and Yousif, J.H. (2013), ‘Environmental scrutinizing system based on soft computing technique’, *International Journal of Computer Applications*, Vol. 62, No. 13, pp. 45-50.

Sohn, J.H., Smith, R., Yoong, E., Leis, J.W. and Galvin, G. (2003), ‘Quantification of odours from piggery effluent ponds using an electronic nose and an artificial neural network’, *Biosystems Engineering*, Vol. 86, No. 4, pp. 399-410.

Wood, A. and Stankovic, J.A. (2000), ‘Denial of service in sensor networks’, *IEEE Computer*, Vol. 35, No. 10, pp. 54-62.

About the authors

Dinesh Kumar Saini is working as Associate Professor in faculty of computing and Information Technology Sohar University which is affiliated with university of Queensland. He received his Ph.D. in software Systems, Prior to that he has done his Masters of Engineering in software systems. He is member of major professional bodies. He has been actively engaged in teaching and research since last 16 years. His main research interests are in Environmental Informatics, Learning Content Management Systems, Searching & Recommending Techniques, Mathematical Modeling, Simulation, Cyber Defense, Network Security, Computational Intelligent Techniques, Software Testing and Quality.

Corresponding author. Faculty of Computing and Information Technology, Sohar University, Oman. Research Fellow and Adjunct Faculty, School of ITEE, University of Queensland, Australia. P.O.Box.-44, PC-311, Sohar, Sultanate of Oma. Tel: (+968) 26720101. E-mail address: dinesh@soharuni.edu.om, dkssohar@gmail.com

Sanad Al Maskari is working as lecturer in the faculty of computing and information technology Sohar University. Currently he is on study leave for pursuing his Ph.D. in the university of Queensland Brisbane Australia. E-mail address: sanad.almaskari@uqconnect.edu.au

MIS Review Vol. 20, No. 1, September (2014), pp. 45-58
DOI: 10.6131/MISR.2014.2001.03

A Discrete Formulation of Successive Software Releases Based on Imperfect Debugging

Jagvinder Singh[1], Adarsh Anand[2], Avneesh Kumar[3], Sunil Kumar Khatri[4]

[1]*Maharaja Agrasen College, University of Delhi, India*

[2]*Department of Operational Research, University of Delhi, India*

[3]*Integrated Academy of Managment Technology, MTU, India*

[4]*Amity Institute of Information Technology, Amity University Uttar Pradesh, India*

ABSTRACT: *Software reliability is the major dynamic attribute of the software quality, so gaining reliability of software product is a vital issue for software products. Due to intense competition the software companies are coming with multiple add-ons to survive in the pure competitive environment by keeping an eye on existing system i.e. system in operational phase. Software reliability engineering is focused on engineering techniques for timely add-ons/Up-gradations and maintaining software systems whose reliability can be quantitatively evaluated. In order to estimate as well as to predict the reliability of software systems, failure data need to be properly measured by various means during software development and operational phases. Although software reliability has remained an active research subject over the past 35 years, challenges and open questions still exist. This paper presents a discrete software reliability growth modeling framework for multi-up gradations including the concept of two types of imperfect debugging during software fault removal phenomenon. The Proposed model has been validated on real data set and provides fairly good results.*

KEYWORDS: *Software Reliability, Non-Homogeneous Poisson Process (NHPP), Software Testing, Successive Software Releases, Imperfect Debugging*

1. Introduction

The Computer systems now pervade every aspect of our daily lives. While this has benefited society and increased our productivity, it has also made our lives more critically dependent on their correct functioning. Software reliability assessment is important to evaluate and predict the reliability and performance of software system. Several SRGMs have been developed in the literature to estimate the fault content and fault removal rate per fault in software. Goel and Okumoto (1979) have proposed NHPP based SRGM assuming that the failure intensity is proportional to the number of faults remaining in the software. The model is very simple and can describe exponential failure curves. Ohba (1984) refined the Goel-Okumoto model by assuming that the fault detection / removal rate increases with time and that there are two types of faults in the software. SRGM proposed by Bittanti et al. (1988) and Kapur and Garg (1992) have similar forms as that of Ohba (1984) but are developed under different set of assumptions. Bittanti et al. (1988) proposed an SRGM exploiting the fault removal (exposure) rate during the initial and final time epochs of testing.

Kapur and Garg (1992) described a fault removal phenomenon, where they have assumed that during a removal process of a fault some of the remaining faults may also be removed. These models can describe both exponential and S-shaped growth curves and hence are termed as flexible models.

NHPP based SRGMs are generally classified into two groups. The first group contains models, which use the execution time (i.e., CPU time) or calendar time. Such models are called continuous time models. The second group contains models, which use the test cases as a unit of fault removal period. Such models are called discrete time models, since the unit of software fault removal period is countable (Kapur & Garg, 1999; Kapur et al., 2011; Musa, Iannino & Okumoto, 1987; Pham, 2006; Yamada, Ohba & Osaki, 1984). A test case can be a single computer test run executed in an hour, day, week or even month. Therefore, it includes the computer test run and length of time spent to visually inspect the software source code. A large number of models have been developed in the first group while fewer are there in the second group due to the difficulties in terms of mathematical complexity involved.

The utility of discrete reliability growth models cannot be under estimated. As the software failure data sets are discrete, these models many times provide better fit than their continuous time counterparts. Therefore, in spite of difficulties in terms of mathematical complexity involved, discrete models are proposed regularly. Most of discrete models discussed in the literature seldom differentiate between the failure observation and fault removal processes. In real software development scene, the number of failure observed can be less than or more than the number of error removed. Kapur and Garg (1992) has discussed the first case in their Error removal phenomenon flexible model which shows as the testing grows and testing team gain experience, additional number of faults are removed without them causing any failure. But if the number of failure observed is more than the number of error removed then we are having the case of imperfect debugging. Due to the complexity of the software system and the incomplete understanding of the software requirements, specifications and structure, the testing team may not be able to remove the fault perfectly on the detection of the failure and the original fault may remain or replaced by another fault. While the first phenomenon is known as imperfect debugging, the second is called fault generation (Kapur et al., 2011; Kapur, Singh, et al., 2010; Pham, 2006). In case of imperfect debugging the fault content of the software is not changed, but because of incomplete understanding of the software, the detected fault is not removed completely. But in case of error generation the fault content increases as the testing progresses and removal results in introduction of new faults while removing old ones.

The concept of imperfect debugging was first introduced by Goel (1985). He introduced the probability of imperfect debugging in Jelinski and Moranda (1972). Kapur, Garg and Kumar (1999) and Kapur et al. (2011) introduced the imperfect debugging in Goel and Okumoto (1979). They assumed that the FRR per remaining faults is reduced due to imperfect debugging. Thus the number of failures observed by time infinity is more than the initial fault content. Although these two models describe the imperfect debugging phenomenon yet the software reliability growth curve of these models is always exponential. Moreover, they assume that the probability of imperfect debugging is independent of the testing time. Thus, they ignore the role of the learning process during the testing phase by not accounting for the experience gained with the progress of software testing. All these models are continuous time models. Kapur, Singh, et al. (2010) and Kapur, Tandon and Kaur (2010) have proposed three discrete models taking into account imperfect fault debugging and fault generation phenomena separately. But even that framework was restricted to single release of the software. Overcoming this, Kapur, Singh, et al. (2010) and Kapur, Tandon, et al. (2010) developed many multi release models but they were formulated in continuous time framework. In this paper, a general discrete SRGM for multi releases incorporating fault generation and imperfect debugging with learning has been proposed.

2. Multi up-gradation of software

The present software development environment is very competitive and advanced. Many independent and well established developing organizations are competing in the market with similar products and capabilities to attain the maximum market share and brand value. As such software delivered with full functionalities and very high reliability built over a period of time may turn out to be unsuccessful due to technological obsolescence. Therefore now a day's the software are rather developed in multiple releases where the latest releases might be developed by improving the existing functionality and revisions, increasing the functionality, a combination of both or improving the quality of the software in terms of reliability etc. (Kapur, Singh, et al., 2010; Kapur, Tandon, et al., 2010). For example we can see the various software in the market named as Windows 98, Windows 2000, Windows ME, Windows XP, Windows Vista, Windows 7 etc. For another illustration consider a development firm developing antivirus software. Such a firm can begin with releasing the product that detects and remove viruses and spywares from the computer system. In their second release they can provide the feature of protecting the system from virus infected emails. Next, they can add the trait of blocking spyware automatically for the next release. Finally, the fourth release can provide the root kit protection along with removing hidden threats in the operating system.

This step by step release (base software with features enhancement) is advantageous for the developing firms in various contexts. Firstly, if a firm implements the complete characteristic capabilities in first release, than that would delay the product release in the market in the desired time window. Secondly, launching of new software product may bring the developing firm in limelight, but the stream of subsequent product releases is the source of their bread and butter. Moreover releasing different versions of the product lengthen the market life of product, protect competitive advantages and sustain crucial maintenance revenue streams.

Software products aren't static and each release has a limited existence. As soon as a software product reaches the market, a variety of factors begin creating demand for changes (Figure 1). Defects require repairs. Competitors offer software with added features and functions. Evolving technology requires upgrades to support new hardware and updated versions of operating software. Sales demands new features to win over prospects. Customers want new functionality to justify maintenance expenditures. These demands accumulate over time, eventually reaching the point when the software product must be upgraded with a new version to remain viable in the market. As soon as the new version is released, the cycle begins again.

For software developing organizations it is not an easy task to design software in isolation. Developing reliable software is one of the most difficult problems faced by them. Timely development, resource limitations, and unrealistic requirements can all negatively impact software reliability. Moreover, there is some interdependence between their developments. The interdependence between their developments exists in many

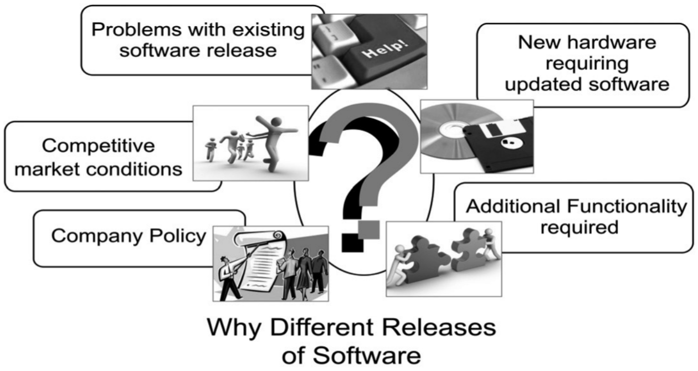
Problems with existing software release
Help!
New hardware requiring updated software
Competitive market conditions
Additional Functionality required
Company Policy
Why Different Releases of Software

Figure 1 Need for Different Releases

ways, which also affects their reliability. A new release (an upgraded version of the base software) may come into existence even during its development, at the time of release or during its operation. The code and other documents related to a release may be some modification of the existing code and documents and/or addition of new modules and related modification in the documents. The dependence of the development process of successive releases necessitates considering this dependence in the reliability growth analysis.

Kapur, Tandon, et al. (2010) developed multi up-gradation reliability model. In this paper, the fault removal process for multiple releases is modeled with respect to testing time (CPU time). This is a continuous time model. But in real life situations most of the data is collected at discrete time instants so there arises a need for the modeling framework which is also discrete in nature. The discrete models relate fault removal process to either number of test cases executed or number of testing weeks etc. These models many a times provide a better fit as compared to continuous time models. Recently, Kapur, Aggarwal and Nijhwan (2014) have developed a modeling framework for multi up-gradations in discrete environment. But they have considered the fault removal process to be depending on all the previous releases. In the proposed model the fault removal process is related to the number of testing periods. The assumption of fault removal process to be depending on all the previous releases has been relaxed and we have considered the dependency on just previous release. Furthermore, Due to complexity and incomplete understanding of the software, the testing team may not be able to remove/correct the fault perfectly on observation/detection of a failure and the original fault may remain resulting in the phenomenon known as imperfect debugging, or get replaced by another fault causing error generation. This paper develops a mathematical relationship between features enhancement and software faults removal process incorporating the aforesaid concepts of imperfect debugging. The model is developed for four software releases in the software. It assumes that when the software is upgraded for the first time, some new functionality is added to the software. The new code written for the software enhancement leads to some new faults in the software which are detected during the testing of the software. During the testing of the newly developed code, there is a possibility that the certain faults were lying dormant in the software which were not removed or detected in the previously released software version. The testing team also removes these faults before releasing the up-graded version of software in the market.

The paper has been organized as follows: Section 3 provides the basic Structure for single release of software which is also the framework for modeling multiple releases; Section 4 contains the parameter estimation values along with the data description.

3. Software reliability modelling for single release: framework for multi-releases

3.1 Model development

During debugging process faults are identified and removed upon failures. In reality this may not be always true. The corrections may themselves introduce new faults or they may inadvertently create conditions, not previously experienced, that enable other faults to cause failures. This results in situations where the actual fault removals are less than the removal attempts. Therefore, the FRR is reduced by the probability of imperfect debugging. Besides, there is a good chance that some new faults might be introduced during the course of debugging (Kapur et al., 2011; Pham, 2006; Yamada et al., 1984).

3.2 Assumptions

The developed below is based upon the following basic assumptions:

(1) Failure observation / fault removal phenomenon is modeled by NHPP.

(2) Software is subject to failures during execution caused by faults remaining in the software

(3) Each time a failure is observed, an immediate effort takes place to decide the cause of the failure in order to remove it.

(4) Failure rate is equally affected by faults remaining in the software.

(5) The debugging process is imperfect.

3.3 Notations

a : Initial Fault content of the software $\sum_{i=1}^{4} a_i = a$.

$a(n)$: Total fault content of the software dependent on the number of testing periods.

b_i : Proportionality constant.

$m_i(n)$: Mean number of faults removed by n number of testing periods.

F_i : Probability distribution function for the number of testing periods.

p_i : The probability of fault removal on a failure (i.e., the probability of perfect debugging).

α_i : The rate at which the faults may be introduced during the debugging process per detected fault.

In all the above notations, i = release 1 to 4.

3.4 Formulation

The software testing phase a software system is executed with a sample of test cases to detect and correct software faults, which cause failures. A discrete counting process $[N(n), n \geq 0]$, $(n = 0, 1, 2, ...)$ is said to be an NHPP with mean value function $m(n)$, if it satisfies the following conditions (Kapur et al., 2011):

There are no failures experienced at $n = 0$, that is, $N(0) = 0$.

The counting process has independent increments, that is, the number of failures experienced during $(n, n+1)^{th}$ testing period is independent of the history and implies that $m(n+1)$ of the process depends only on the present state $m(n)$ and is independent of its past state $m(n)$, for $x < n$.

In other words, for any collection of the numbers of testing periods $n_1, n_2, ..., n_k$ $(0 < n_1 < n_2 < ... < n_k)$ the k random variables $N(n_1)$, $N(n_2) - N(n_1)$, ..., $N(n_k) - N(n_{k-1})$ are statistically independent.

For any of two numbers of test cases n_i and n_j where $(0 \leq n_i \leq n_j)$, we have:

$$\Pr\{N(n_j) - N(n_i) = x\} = \frac{\{m(n_j) - m(n_i)\}^x}{x\,!}\exp\left[-\{m(n_j) - m(n_i)\}\right] \quad (1)$$

The mean value function $m(n)$ which is a non-decreasing in n represents the expected cumulative number of faults detected by n testing periods. Then the NHPP model with $m(n)$ is formulated by:

$$\Pr\{N(n) = x\} = \frac{\{m(n)\}^x}{x\,!}\exp\left[-\{m(n)\}\right] \quad (2)$$

Therefore, under the above assumptions, the expected cumulative number of faults removed between the n^{th} and $(n+1)^{th}$ testing period is proportional to the number of faults remaining after the execution n^{th} test run, satisfies the following difference equation

$$m(n+1) - m(n) = bp(a(n) - m(n)) \quad (3)$$

where an increasing a(n) implies an increasing total number of faults expressed as

$$a(n) = a + \alpha m(n) \quad (4)$$

Substituting Equation (4) in Equation (3) we have

$$m(n+1) - m(n) = bp(a + \alpha m(n) - m(n)) \quad (5)$$

Solving Equation (5) under the initial condition $m(n = 0) = 0$ we get

$$m(n) = \frac{a}{1-\alpha}\left[1 - \left(1 - bp(1-\alpha)\right)^n\right] \quad (6)$$

This Equation (6) can be rewritten as:

$$m(n) = a * F(n) \tag{7}$$

where,

$$\begin{cases} a* = \dfrac{a}{1-\alpha} \\ F(n) = 1 - \left(1 - bp(1-\alpha)\right)^n \end{cases} \tag{8}$$

Release 1:

A primary purpose of testing is to detect software failures so that defects may be discovered and corrected. Testing starts once the code of software is written. Before the release of the software in the market the software testing team tests the software rigorously to make sure that they remove maximum number of bugs in the software. The first release is the foundation of the software so testing team are bound to give their best effort. Although it is not possible to remove all the bugs in the software practically. Therefore, when the software is tested by the testing team, there are chances that they may detect a finite number (less than the total content of the faults) of bugs in the code developed.

So finite numbers of bugs are then removed and mathematical equation for it is given as under:

$$m_1(n) = a_1 * F_1(n) \qquad 0 < n \le n_1 \tag{9}$$

where,

$$\begin{cases} a_1 * = \dfrac{a_1}{1-\alpha_1} \\ F_1(n) = 1 - \left(1 - b_1 p_1 (1-\alpha_1)\right)^n \end{cases} \tag{10}$$

Release 2:

After first release, the company has information about the reported bugs from the users; hence in order to attract more customers, a company adds some new functionality to the existing software system. Adding some new functionality to the software leads to change in the code. These new specifications in the code lead to increase in the fault content. Now the testing team starts testing the upgraded system, besides this the testing team considers dependency and effect of adding new functionalities with existing system. In this period when there are two versions of the software, $a_1 * (1 - F_1(n_1))$ is the leftover fault content of the first version which interacts with new portion of detected faults i.e.,

$F_2(n-n_1)$. In addition a fraction of faults generated due to enhancement of the features are removed with new rate. Here it may be noted that there is a change in the fault detection rate. This change in the fault detection may be due to change in time, change in the fault complexity due to new features or change in testing strategies etc. The mathematical equation of these finite numbers of faults removed can be given by:

$$m_2(n) = a_2 * F_2(n-n_1) + a_1 * (1 - F_1(n_1))F_2(n-n_1) \quad n_1 < n \leq n_2 \tag{11}$$

where,

$$\begin{cases} a_2 * = \dfrac{a_2}{1-\alpha_2} \\ F_2(n-n_1) = 1 - \left(1 - b_2 p_2 (1-\alpha_2)\right)^{n-n_1} \end{cases} \tag{12}$$

Release 3:

The modeling for release 3 is done on the basis of the arguments similar to given in second release along with taking into consideration the fact that the next release will not contain the remaining faults of all previous releases, rather it will be dependent on the just previous release. A proportion of faults get removed when the testing team tests the new code and these faults are removed with the detection proportion $F_3(n-n_2)$. During the testing of newly integrated code, apart from the faults lying in the new code, a number of bugs which have remained undetected i.e., $a_2 * (1 - F_2(n_2 - n_1))$ are also removed with the detection proportion $F_3(n-n_2)$. The resulting equation is as following:

$$m_3(n) = a_3 * F_3(n-n_2) + a_2 * (1 - F_2(n_2-n_1))F_3(n-n_2) \quad n_2 < n \leq n_3 \tag{13}$$

where,

$$\begin{cases} a_3 * = \dfrac{a_3}{1-\alpha_3} \\ F_3(n-n_2) = 1 - \left(1 - b_3 p_3 (1-\alpha_3)\right)^{n-n_2} \end{cases} \tag{14}$$

Similarly for release 4, the corresponding mathematical expression can be given by:

$$m_4(n) = a_4 * F_4(n-n_3) + a_3 * (1 - F_3(n_3-n_2))F_4(n-n_3) \quad n_3 < n \leq n_4 \tag{15}$$

where, a_4 * and $F_4(n-n_3)$ can be defined as done in previous steps.

4. Parameter analysis

Parameters estimation is of primary concern in software reliability prediction. For this, the failure data is collected and is recorded in either of the following two formats-the first approach is to record the time between successive failures while second way is to collect the number of failures experienced at regular testing intervals. The mean number of faults detected/removed by testing periods $m(n)$ is mostly described by the non-linear functions and once the analytical solution is known for a given model, the parameters in the solution are required to be determined. Parameter estimation is done by Non-linear Least Square (NLLS). For this nonlinear regression (NLR) module of SPSS has been used.

4.1 Model validation

To check the validity of the proposed model to describe the software reliability growth, it has been tested on Tandem Computers (Wood, 1996). The data set includes the failure data from 4 major releases of the software at Tandem Computers. In the First Release, 100 faults were detected after testing for 20 weeks. The Second Release was done after detecting 120 faults for 19 weeks. The Third and Forth Release were done after testing for 12 and 19 weeks, removing 61 and 42 faults respectively. Table 1 gives the value of the parameters and Table 2 provides the comparison criteria's.

The performance of SRGM is judged by their ability to fit the past software failure occurrence / fault removal data and to predict satisfactorily the future behavior of the software failure occurrence / fault removal process. Figures 2 ~ 5 give the Goodness of Fit curves for the four releases.

Table 1 Parameter Estimates

Release 1		Release 2		Release 3		Release 4	
a_1	109.24	a_2	118.58	a_3	62.187	a_4	43.316
b_1	0.3286	b_2	0.2777	b_3	0.3332	b_4	0.0374
p_1	0.7119	p_2	0.7028	p_3	0.6691	p_4	0.6728
α_1	0.1899	α_2	0.1838	α_3	0.1739	α_4	0.0097

Table 2 Comparison Criteria

Criteria	R2	Bias	Variation	MSE
Release 1	.990	0.403	2.81	7.71
Release 2	.995	0.214	2.159	7.065
Release 3	.995	0.050	1.490	1.909
Release 4	.992	0.075	1.106	1.163

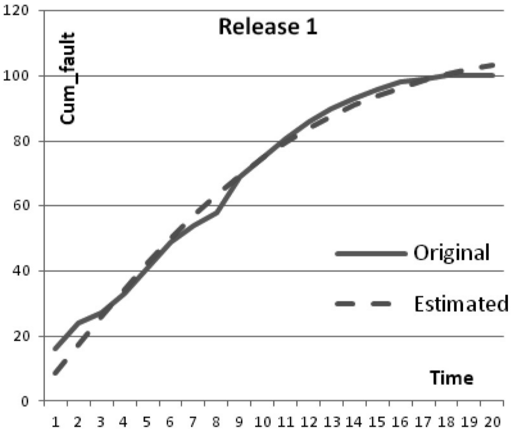

Figure 2 Goodness of Fit for Release 1

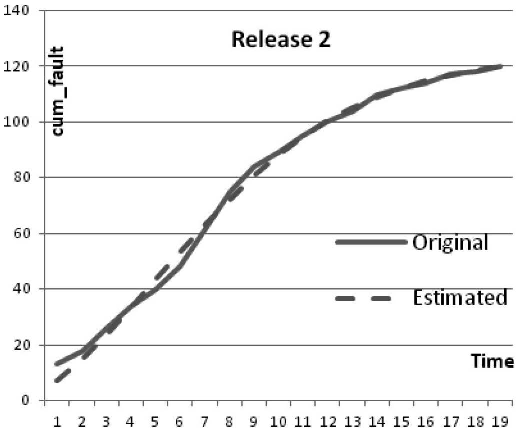

Figure 3 Goodness of Fit for Release 2

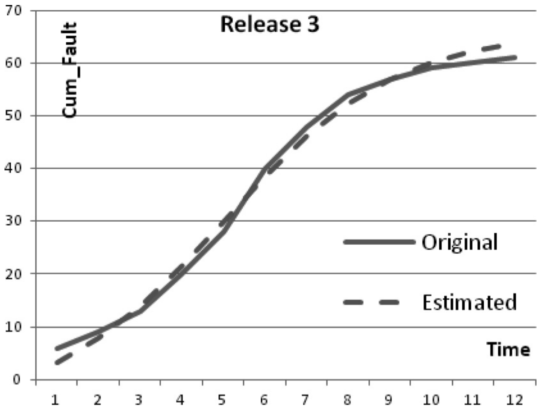

Figure 4 Goodness of Fit for Release 3

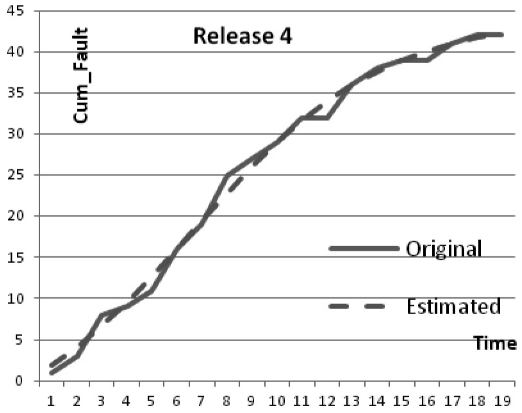

Figure 5 Goodness of Fit for Release 4

5. Conclusion

Making reliable software is the need of an hour. Every customer needs a more efficient and bug free software. Software products in general face a fierce competition in the market and therefore have to come up with upgraded versions of the software. But the matter of the fact is that up-gradation is a complex and difficult process. The add-ons can result in distorting the fault removal rate and can cause change in the number of fault contents. The software reliability multi up-gradation model developed in this paper incorporates this concept of Imperfect Debugging and is based on the assumption that the overall fault removal of the new release depends on the faults generated in that release and on the leftover faults of just previous release (for each release). This helps us in removing more and more faults in the software and produces highly reliable software. The proposed multi up gradation model is estimated on a four release data set.

Acknowledgements

The work done in this paper is supported by grants to the second author via grant No DRCH/R&D/2013-14/4155 and RC/2014/6820 from University of Delhi, India. Further, we would like to acknowledge that this paper is an extended version of a paper titled “A Discrete Formulation of Successive Software Releases Based on Imperfect Debugging” presented at an International Conference on Reliability, Infocom Technologies and Optimization, January 29 ~ 31, 2013

References

Bittanti, S., Bolzern, P., Pedrotti, E. and Scattolini, R. (1988), ‘A flexible modelling approach for software reliability growth’, in Goos, G. and Hartmanis, J. (Eds.), *Software Reliability Modelling and Identification*, Springer Verlag, Berlin, Germany, pp. 101-140.

Goel, A.L. (1985), ‘Software reliability models: assumptions, limitations and applicability’, *IEEE Transactions on Software Engineering*, Vol. SE-11, pp. 1411-1423.

Goel, A.L. and Okumoto, K. (1979), ‘Time-dependent error-detection rate model for software reliability and other performance measures’, *IEEE Transactions on Reliability*, Vol. 28, No. 3, pp. 206-211.

Jelinski, Z. and Moranda, P.B. (1972), ‘Software reliability research’, in Freiberger, W. (Ed.), *Statistical Computer Performance Evaluation*, Academic Press, New York, NY, pp. 465-497.

Kapur, P.K., Aggarwal, A.G. and Nijhawan, N. (2014), ‘A discrete SRGM for multi release software system’, *International Journal Industrial and System Engineering*, Vol. 16, No. 2, pp. 143-155.

Kapur, P.K. and Garg, R.B. (1992), ‘A software reliability growth model for an fault removal phenomenon’, *Software Engineering Journal*, Vol. 7, pp. 291-294.

Kapur, P.K., Garg, R.B. and Kumar, S. (1999), *Contributions to Hardware and Software Reliability*, World Scientific, Singapore.

Kapur, P.K., Pham, H., Gupta, A. and Jha, P.C. (2011), *Software Reliability Assessment with OR Applications*, Springer, London, UK.

Kapur, P.K., Singh, O., Garmabaki, A.S. and Singh J. (2010), ‘Multi up-gradation software reliability growth model with imperfect debugging’, *International Journal of Systems Assurance Engineering and Management*, Vol. 1, pp. 299-306.

Kapur, P.K., Tandon, A. and Kaur, G. (2010) ‘Multi up-gradation software reliability model’, *Proceedings of the 2nd IEEE International Conference On Reliability, Safety & Hazard*, Mumbai, Indian, pp. 468-474.

Musa, J.D., Iannino, A. and Okumoto, K. (1987), *Software Reliability: Measurement, Prediction, Application*, McGraw-Hill, New York, NY.

Ohba, M. (1984), ‘Software reliability analysis models’, *IBM Journal of Research and Development*, Vol. 28, No. 4, pp. 428-443.

Pham, H. (2006), *System Software Reliability*, Springer, New York, NY.

Wood, A. (1996), ‘Predicting software reliability’, *Computer*, Vol. 29, pp. 69-77.

Yamada, S., Ohba, M. and Osaki, S. (1984), ‘S-shaped software reliability growth models and their applications’, *IEEE Transactions on Reliability*, Vol. 33, No. 4, pp. 289-292.

About the authors

Jagvinder Singh is working as Assistant Professor in Maharaja Agrasen College, University of Delhi. He received his Doctoral Thesis in 2012 from Department of Operational research, University of Delhi. His area of research is Software Reliability Modeling. He has published several papers in International/National Journals and Proceedings. He is a lifetime member of the Society for Reliability Engineering, Quality and Operations Management (SREQOM). E-mail address: jagvinder.singh@gmail.com

Adarsh Anand is doing research in the area of Innovation Diffusion Modeling and Software Reliability Assessment. Presently he is working as an Assistant Professor in the Department of Operational Research, University of Delhi (INDIA). He did his PhD. And M Phil in Operational Research in 2013 and 2010 respectively. He has published several papers in International/National Journals and Proceedings. He is a lifetime member of the Society for Reliability Engineering, Quality and Operations Management (SREQOM).
Corresponding author. Room No. 208, 2nd Floor, Department of Operational Research, University of Delhi, Delhi 110007, India. Tel: 011-27666960. E-mail address: adarsh.anand86@gmail.com

Avneesh Kumar is pursuing his PhD from Jiwaji University, MP. Currently he holds the position of a Lecturer in INMANTEC, UP. He has been active member of the Society for Reliability Engineering, Quality and Operations Management (SREQOM). E-mail address: avn119@rediffmail.com

Sunil Kumar Khatri is working as Director in Amity Institute of Information Technology, Amity University, Noida, India. He is a Fellow of IETE, Sr. Member of IACSIT and of Computer Society of India and Member of IEEE. He has been conferred “IT Innovation & Excellence Award for Contribution in the field of IT and Computer Science Education” by *Knowledge Resource Development & Welfare Group* Dec, 2012 and the award for “Exceptional Leadership and Dedication in Research” during the *4th International Conference on Quality, Reliability and Infocom Technology* in the year 2009. He has edited three books, two special issues and published several papers in international and national journals and proceedings. His areas of research are Software Reliability, Modeling and Optimization, Data Mining and Warehousing, Network Security, Soft Computing and Pattern Recognition. E-mail address: sunilkkhatri@gmail.com

MIS Review Vol. 20, No. 1, September (2014), pp. 59-71
DOI: 10.6131/MISR.2014.2001.04

Securing E-Commerce Business Using Hybrid Combination Based on New Symmetric Key and RSA Algorithm

Prakash Kuppuswamy[1], Saeed Q. Y. Al-Khalidi[2]
[1]*Department of Computer Engineering and Networks, Jazan University, KSA*
[2]*Deanship of Libraries Affairs, King Khalid University, KSA*

ABSTRACT: *Security in e-commerce is becoming more topical as the shift from traditional shopping and transactions move away from physical stores to online. E-commerce has had a drastic effect on the global economy and has rapidly accelerated over the years into the trillions of dollars a year. Protecting payment web application users and application systems requires a combination of managerial, technical and physical controls. In this paper, we propose hybrid cryptographic system that combines both the symmetric key algorithm, and popular RSA algorithm. The symmetric key algorithm based on integer numbers and RSA algorithm widely using in all data security application. Efficiency of the security methods are dignified and such competence increases as we combined security methods with each other.*

KEYWORDS: *E-Commerce, Hybrid Security, RSA Algorithm, Simple Symmetric Key*

1. Introduction

Electronic commerce is buying and selling of goods and services across the internet. Commercial activities over the internet have been growing in an exponential manner over the last few years. When it comes to payment, one needs to establish a sense of security. Customers must be able to select a mode of payment and the software must verify their ability to pay. This can involve credit cards, electronic cash, encryption, and/or purchase orders. The more of these techniques are supported by an E-commerce package, the more secure the system can be, and therefore the more customers are benefits from E-commerce abilities (Al-Slamy, 2008; Greenberg, 2001; Olkowski, 2001).

Security issues are an important topic in e-commerce. How to protect the security of an e-commerce system and data is its core research area (Davis, 2003). There are many sensitive financial data and asset data in e-commerce databases, such as transaction records, commercial transactions, user account and market scheme and so on. The data are very important to the parties involved in e-commerce, so we must assure their security completely (Hou, 2009).

At present the security technologies used in e-commerce databases are Web access control, user authentication, authorization control, safety audit, backup and recovery, data encryption and so on. These technologies can assure general database security, but it is difficult to assure their security for important databases. Encryption technology is one of the most effective technologies of database security. However a simple encryption technology, such as symmetrical encryption or asymmetrical encryption, is very difficult to guarantee the security of network databases. We must combine the both and through hybrid encryption we can create a safe, efficient e-commerce database system (Hou, 2009)

Secure communication is an intrinsic requirement for many popular online transactions such as e-commerce, stock trading and e-banking. E-commerce and m-commerce transactions are growing at an explosive rate. The success of these depends on how transactions are carried out in the most secured manner. The prime requirements for any e-commerce and m-commerce transactions are Privacy, Authentication, Integrity maintenance and Non-Repudiation. Cryptography helps us in achieving these prime requirements. Today, various cryptographic algorithms have been developed. These are broadly classified as symmetric key (Rasmi & Paul, 2011).

A hybrid cryptosystem is a protocol using multiple ciphers of different types together, each to its best advantage. One common approach is to generate a random secret key for a symmetric cipher, and then encrypt this key via an asymmetric cipher using the recipient’s public key (Rasmi & Paul, 2011).

The message itself is then encrypted using the symmetric cipher and the secret key. Both the encrypted secret key and the encrypted message are then sent to the recipient. The recipient decrypts the secret key first, using his/her own private key, and then uses that key to decrypt the message (Janakiraman, Ganesan & Gobi, 2007).

It is clear that electronic commerce will revolutionize businesses, and customers will be offered new and exciting services. As E-commerce businesses are growing, more secure technologies are being developed and improved every day. The current internet security polices and technologies fail to meet the needs of end users. The success or failure of an E-commerce operations hinges on myriad factors, including but not limited to the business model, the team, the customers, the investors, the product, and the security of data transmissions and storage. Any business that wants to have a competitive edge in today’s global marketplace should adopt a comprehensive security policy in consultation with partners, suppliers, and distributors that will provide safe environment for the coming proliferation of E-commerce (Chaffey, 2004; Greenberg, 2001). In Figure 1 shows the features of E-commerce security and Figure 2 show the simple architecture of E-commerce system.

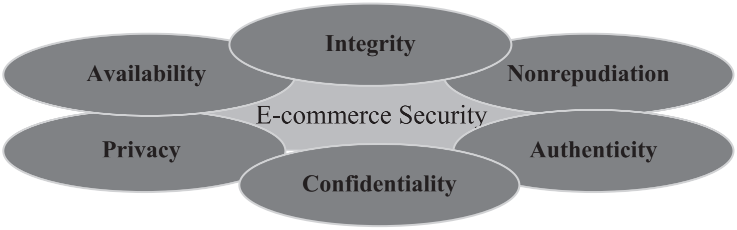

Figure 1 Dimension of E-Commerce Security

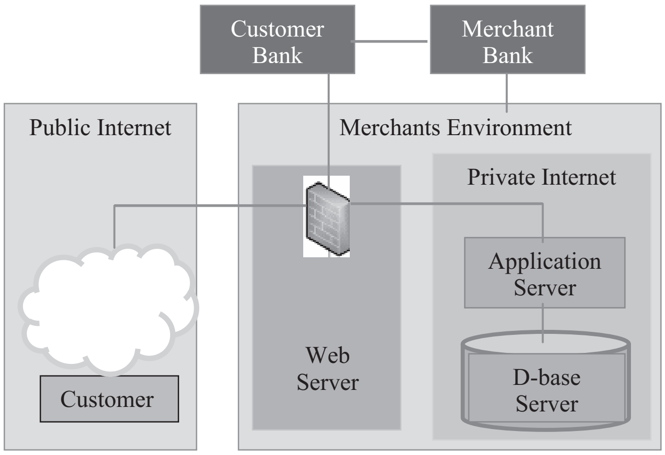

Figure 2 Simple E-Commerce Structure

2. Literature review

Kherad et al. (2010), in this research they proposed a new self-developed symmetric algorithm called FJ RC-4, which is derived from RC4. They investigated and compared the robustness of the RC4 and FJ-RC4 and shown that FJ-RC4 is stronger than RC4 against the attacks. In addition, it takes more time to find key in FJ-RC4 and requires more resources (Kherad et al., 2010).

Rasmi and Paul (2011), the circle symmetric key algorithm is based on 2-d geometry using property of circle, and circle-centered angle. It is a block cipher technique but has the advantage of producing fixed size encrypted messages all cases. The asymmetric algorithm is RSA with CRT which improves the performance of the basic RSA algorithm by four (Rasmi & Paul, 2011).

Palanisamy and Jeneba Mary (2011), the Rijndael algorithm mainly consists of a symmetric block cipher that can process data blocks of 128, 192 or 256 bits by using key lengths of 128, 196 and 256 bits. This work also generating two pairs of keys; public and private key. Using Public key it encrypts the data key and other one is public and private key pair ,which will send to other person, so that opposite person can decrypt the encrypted key using his public and private key (Palanisamy & Jeneba Mary, 2011)

Kuppuswamy and Al-Khalidi (2012), proposed new symmetric key algorithm based on integer and modular 37 and select any number and calculate inverse of the selected integer using modular 37. The symmetric key distribution should be done in the secured manner. This study's main goal is to reflect the importance of security in network and provide the better encryption technique for currently implemented encryption techniques in simple and powerful method (Kuppuswamy & Al-Khalidi, 2012).

Yasin, Haseeb and Qureshi (2012), they suggested E-commerce has presented a new way of doing transactions all over the world using internet. Organizations have changed their way of doing business from a traditional approach to embrace E-commerce processes. The purpose of this paper is to explain the importance of E-commerce security digital signature and certificate based cryptography techniques in E-commerce security (Yasin et al., 2012).

Nanehkaran (2013), electronic commerce is supporting of customers, supplying of services and commodities, portion of business information, manages business transactions and maintaining of bond between suppliers, customers and vendors by devices of telecommunication networks. In this research article paper is to review of principles, definitions, history, frameworks, steps, models, advantages, barriers and limitations of electronic commerce (Nanehkaran, 2013).

3. Problem statement

Over the years, the methods used by ecommerce or web commerce sites to process and store credit card/master card information has become much more sophisticated than the early days of online shopping business. This progress has helped online business overcome one of its greatest obstacles, customer faith. As showed by the amount of money transaction online every year, people feel much more secure in online shopping than they ever have. Regrettably for businesses, the methods used by cyber criminals trying to steal their customer's information have made it easier than ever for them to compromise a web application.

Security threats to web sites and web applications come in many ways. Data centres and other resources used for hosting web sites and their associated systems need to be protected from all type of vulnerable activity. Threats should be identified using application threat modeling and then evaluated with a vulnerability assessment. Susceptibilities can be removed or reduced and counter measures put in place to mitigate the effects of an incident should the threat be realized.

Figure 3 shows money spending for information/data security between 2009 and 2013. In the year of 2011 and 2012, 2.6 million US$ only spend for information security, but, in the year of 2013 spend more than 4 million dollar, It is the huge margin comparing to the previous years.

4. Proposed hybrid algorithm

4.1 Simple symmetric key algorithm

Symmetric key is implemented in two ways either as a block cipher or stream cipher. Block cipher transforms a fixed length block of plaintext say a fixed size of 64 data into a block of ciphertext (encrypted text) data of the same length. We know that, whatever user ID consist of Alphabets between A to Z and numbers which is between 0 ~ 9. Here, In New symmetric key algorithm, we introduce synthetic data, which is based on the user ID. Normally the synthetic data value consists of equivalent value of alphabets and numbers. Alphabet value A is assigned as integer number 1 and B = 2 ... so on. Next we consider integer value 0 assigned as 27 and 1 = 28 ... 9 = 36 also the space value considers as an integer number 37.

4.1.1 Key generation method

(1) Select any natural number say as n

(2) Find the inverse of the number using modulo 37 (key 1) say k.

(3) Again select any negative number (for making secured key) n1.

(4) Find the inverse of negative number using modulo 37 (key 2) k1.

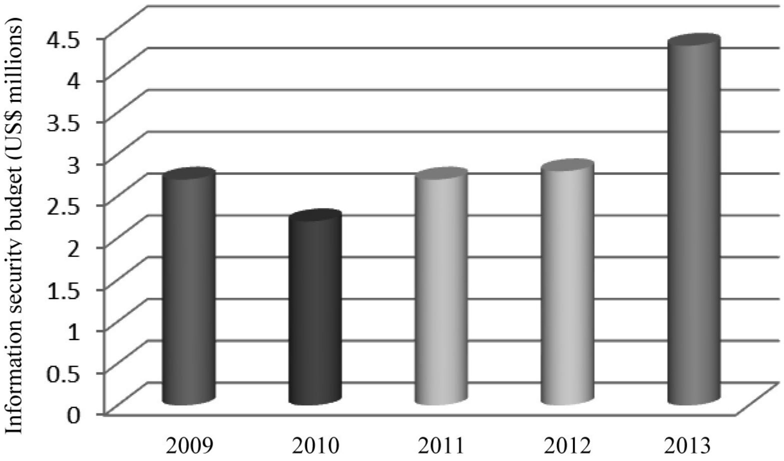

Figure 3 Budget Used for Information Security

Source: The Global Information Security Survey 2014.

4.1.2 Encryption method

(1) Assign synthetic value for user ID.

(2) Multiply synthetic value with random selected natural number.

(3) Calculate with modulo 37.

(4) Again select random negative number and multiply with it.

(5) Again calculate with modulo 37 $CT = (PT \times n \times n1) \bmod 37$.

4.1.3 Decryption method

(1) Multiply received text with key 1 & key 2.

(2) Calculate with modulo 37.

(3) Remainder is Revealed Text or Plain Text $PT = (CT \times n^{-1} \times n1^{-1}) \bmod 1$.

4.2 RSA asymmetric key algorithm

The RSA algorithm is based on the assumption that integer factorization is a difficult problem. This means that given a large value n, it is difficult to find the prime factors that make up n. It is most popular asymmetric key algorithm.

4.2.1 Key generation

(1) Choose two very large random prime integers p and q.

(2) Compute n and $\varphi(n)$: $n = pq$ and $\varphi(n) = (p-1)(q-1)$.

(3) Choose an integer e, $1 < e < \varphi(n)$ such that: $\gcd(e, \varphi(n)) = 1$(where gcd means greatest common denominator).

(4) Compute d, $1 < d < \varphi(n)$ such that: $ed \equiv 1 \pmod{\varphi(n)}$, the public key is (n, e) and the private key is (n, d).

The values of p, q and $\varphi(n)$ are private; e is the public or encryption exponent; d is the private or decryption exponent.

4.2.2 Encryption

ciphertext $CT = M^e \pmod{n}$.

4.2.3 Decryption

Message M $= CT^d \pmod{n}$.

5. Proposed hybrid architecture

The following hybrid architecture design using, symmetric cipher and familiar RSA public key algorithm. It is basis of the protocol that enables to provide security while accomplishing an important system or network security. A protocol is an agreed-on hierarchical sequence of actions that leads to desirable results. Both the encrypted secret key and the encrypted message are then sent to the Merchant. The recipient decrypts the private key first, using his own private secret key, and then uses that secret key to decrypt the message. Figure 4 shows the block diagram of a hybrid crypto system which takes the advantages of both shared secret and public key algorithms. That means it combines both the symmetric key algorithm and asymmetric-key algorithm to take the advantage of the higher speed of symmetric ciphers and the ability of asymmetric ciphers to securely exchange keys.

6. Implementation with sample message

Our E-commerce implementation test measures efficient of implementing E-commerce on real time application. Ecommerce security is responsible for identifying

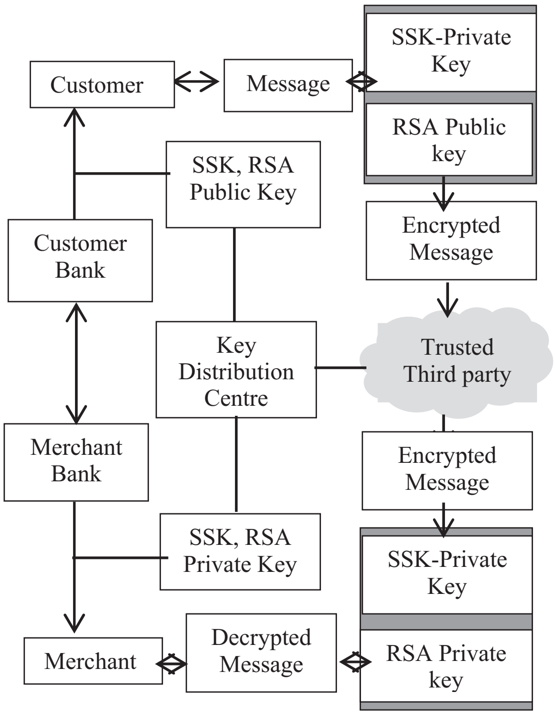

Figure 4 Proposed E-Commerce Architecture

network security threats, coordinating threat response, secure payment transaction. It will be responsible for business transaction between customer and merchant using external networks. In this implementation process the sample message "PRODUCT" mentioned in the Table 1 has taken for the experiment.

Table 1 Sample Message

P	R	O	D	U	C	T
16	18	15	4	21	3	20

6.1 SSK key generation

(1) We are selecting random integer number n = 3.

(2) Then inverse of 3 = 25 (verification 3 × 25 mod 37 = 1). So, Key 1 = 25.

(3) Again we are selecting random negative number n1 = -8.

(4) Then inverse of -8 = 23 (verify -8 × 23 = -184 mod 37 = 1). So, Key 2 = 23.

Encryption using SSK shows in the Table 2.

Table 2 Symmetric Key Encryption

Plain Text	Integer Value	CT = (M × n) mod 37 (n = 3)	CT = (CT × n1) mod 37 (n = -8)	Cipher Text
P	16	11	23	W
R	18	17	12	L
O	15	8	10	J
D	4	12	15	O
U	21	26	14	N
C	3	9	2	B
T	20	23	1	A

6.2 RSA key generation

Encryption using RSA.

We choosing here.

P = 7; q = 13; Therefore n = 91 Øn = 72.

Selecting e = 5 then inverse of e or d = 29 (verification 5 × 29 mod 72 = 1).

Public key is e, n = 5, 91.

Private key "d" = 27.

6.3 RSA encryption

Now we receive the cipher text message from above table "WLJONBA" i.e., equivalent integer value 23, 12,10, 15, 14, 2, 1. The encryption process of RSA algorithm mentioned in Table 3.

$(m)^e$ mod n i.e., $(2)^7$ mod 33 = 29.

Table 3 RSA Encryption Using Public Key

W	23	$(23)^5$ mod 91 =	4
L	12	$(12)^5$ mod 91 =	38
J	10	$(10)^5$ mod 91 =	82
O	15	$(15)^5$ mod 91 =	71
N	14	$(14)^5$ mod 91 =	14
B	2	$(2)^5$ mod 91 =	31
A	1	$(1)^5$ mod 91 =	1

6.4 Decryption using RSA & SSK

The decryption process of RSA algorithm and symmetric key algorithm mentioned in Table 4 and Table 5 respectively.

$(m)^d$ mod n

Table 4 RSA Decryption Using Private Key

4	$(4)^{29}$ mod 91 =	23	W
38	$(38)^{29}$ mod 91 =	12	L
82	$(82)^{29}$ mod 91 =	10	J
71	$(71)^{29}$ mod 91 =	15	O
14	$(14)^{29}$ mod 91 =	14	N
31	$(31)^{29}$ mod 91 =	2	B
1	$(1)^{29}$ mod 91 =	1	A

Table 5 Symmetric Key Decryption Using Private Key

Cipher Text	Integer Value	PT = (M × k1 × k2) mod 37	Plain Text
W	23	16	P
L	12	18	R
J	10	15	O
O	15	4	D
N	14	21	U
B	2	3	C
A	1	20	T

7. Result analysis

Here we have encrypted customer message “PRODUCT” into numbers using private and RSA public key and hence decrypted the keys to obtain the final character and the final message. Here we analyzed with existing DES, 3DES and AES algorithm to find out our new hybrid combination performance. The algorithm executes on PC computer of CPU Intel Pentium 4, 2.2 MHz Dual Core. The programs implemented using MATLAB and messages are stored in 3 different arrays for Key generation, Encryption and Decryption scheme. It is tested with the length of 100 bits.

Here we are examining two types of facts for consideration of performance. First one is computational performance and second one is communication performance. Computational performance refers to the speed of computation required to perform cryptographic operations. Communication performance indicates the total security required for transmission of data between two parties.

DES is the old “data encryption standard” from the seventies. Its key size is too short for proper security. 3DES is believed to be secure up to at least “2^{112}” security. But it is slow, especially in software. AES is the successor of DES, and it accepts keys of 128, 192 or 256 bits. Our proposed Hybrid combination of algorithm based on Simple symmetric key and RSA algorithm, which has been using in many application. The key size of RSA algorithm is standard and compatible for all application also encryption/decryption time of the Hybrid is less than comparing to the other algorithms. It is more secure than others using by the combination of two different algorithm. Table 6 and Figure 5 show the performance of DES, 3-DES, AES and our proposed hybrid algorithm.

Table 6 Performance Comparison

Algorithm	Key Size	Encryption/Decryption Timing (100 bits)
DES	64 bits	10 Sec
3-DES	2^{112}	20 Sec
AES	256 bits	15 Sec
SSK+RSA	2,048	10 Sec

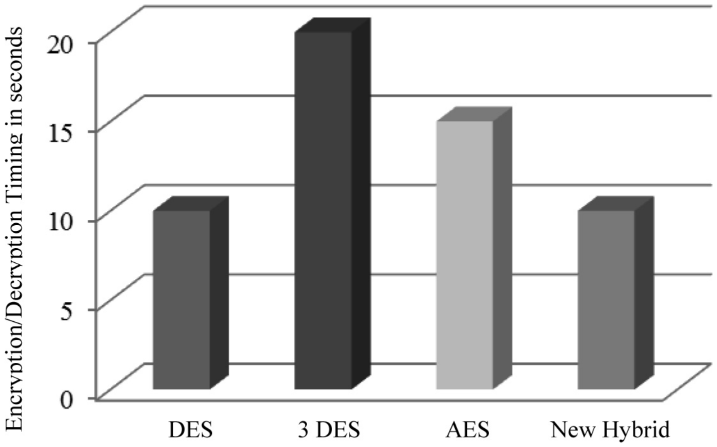

Figure 5 Performance Analysis of Various Algorithms

8. Conclusion

The proposed hybrid encryption algorithm used in this paper can also be used to enhance the security of other network. This work using simple symmetric key algorithm based and natural numbers and modular 37 cryptography used to data encryption/decryption and RSA cryptography asymmetric algorithm. On implementation of this combination of hybrid algorithm, we concluded several points. The encryption and decryption of any data has a secret or private key, which is used for data encryption. For this purpose asymmetric key or public key system is used. We introduced here the version of RSA which was resistant against security attack. Finally we illustrated the new directions for the future research.

References

Al-Slamy, N.M.A. (2008), ‘E-commerce security’, *International Journal of Computer Science and Network Security*, Vol. 8, No. 5, pp. 340-344.

Chaffey, D. (2004), *E-Business and E-Commerce Management*, 2nd ed., Prentice Hall, Harlow, UK.

Davis, Z. (2003), 'E-commerce', *Software World*, Vol. 30, pp. 207-212.

Greenberg, P.A. (2001), 'In E-commerce we trust ... not', available at http://www.ecommercetimes.com (accessed 12 February 2014).

Hou, J. (2009), 'Research on database security of E-commerce based on hybrid encryption', *Proceedings of the 2009 International Symposium on Web Information Systems and Applications*, Nanchang, China, pp. 363-366.

Janakiraman, V.S., Ganesan, R. and Gobi, M. (2007), 'Hybrid cryptographic algorithm for robust network security', *ICGST-CNIR*, Vol. 7, No. I, pp. 1141-1146.

Kherad, F.J., Naji, H.R., Malakooti, M.V. and Haghighat, P. (2010), 'A new symmetric cryptography algorithm to secure e-commerce transactions', *Proceedings of the International Conference of Financial Theory and Engineering*, Dubai, United Arab Emirates, pp. 234-237.

Kuppuswamy, P. and Al-Khalidi, S.Q.Y. (2012), 'Implementation of security through simple symmetric key algorithm based on Modulo 37', *Council for Innovative Research International Journal of Computers & Technology*, Vol. 3, No. 2, pp. 335-338.

Nanehkaran, Y.A. (2013), 'An introduction to electronic commerce', *International Journal of Scientific & Technology Research*, Vol. 2, No. 4, pp. 190-193.

Olkowski, D.J. (2001), 'Information security issues in E-commerce', available at http://www.sans.org/reading-room/whitepapers/ecommerce/information-security-issues-e-commerce-37 (accessed 10 March 2015).

Palanisamy, V. and Jeneba Mary, A. (2011), 'Hybrid cryptography by the implementation of RSA and AES', *International Journal of Current Research*, Vol. 33, No. 4, pp. 241-244.

Rasmi, P.S. and Paul, V. (2011), 'A hybrid crypto system based on a new circle-symmetric key algorithm and RSA with CRT asymmetric key algorithm for E-commerce applications', *Proceedings of International Conference on VLSI, Communication & Instrumentation*, Kerala, India, pp. 14-18.

Yasin, S., Haseeb, K. and Qureshi, R.J. (2012), 'Cryptography based E-commerce security: a review', *International Journal of Computer Science Issues*, Vol. 9, No. 2, pp. 132-137.

About the authors

Prakash Kuppuswamy, Lecturer, Computer Engineering & Networks Department in Jazan University, KSA. He is research Scholar-Doctorate Degree yet to be awarded by Dravidian University. He has published 25 International Research journals/Technical papers and participated in many international Conferences in Maldives, Libya and Ethiopia. His research area includes Cryptography, Bio-informatics and E-commerce security etc.

Corresponding author. College of Computer Science & Information system, Jazan University, KSA. Sathuvachary, Vellore, Tamil Nadu, India. Tel: +966 532883941. E-mail address: prakashcnet@gmail.com

Saeed Q. Y. Al-Khalidi, Dean, Deanship of Libraries Affairs at King Khalid University, Abha. KSA. He published many National & International papers, Journals. Also, he participated as a Reviewer in many international conferences worldwide. He completed Master Degree and Doctor of Philosophy in University of East Anglia. His research interests include: Information System development, approaches to systems analysis and the early stages of systems development process, IT/IS evaluation practices, E-readiness assessment. E-mail address: prakashcnet@gmail.com, salkhalidi@yahoo.com

CALL FOR PAPER

MIS Review: An International Journal

Published 2 Issues Annually by Airiti Press Inc.

MIS Review is a double-blind refereed academic journal published jointly by Airiti Press Inc. and Department of Management Information Systems, College of Commerce, National Chengchi University in Taiwan. The journal is published both in print and online. We welcome submissions of research papers/case studies in the areas including (but not limited to):

1. **MIS Roles, Trends, and Research Methods**
   Roles, positioning and research methods of management information systems, and the impacts & development trends of information technology on organizations.

2. **Information Management**
   Information infrastructure planning and implementation, information technology and organizational design, strategic applications of information systems, information system project management, knowledge management, electronic commerce, end-user computing, and service technology management.

3. **Information Technologies**
   Database design and management, decision support systems, artificial intelligence applications (including expert systems and neural networks), software engineering, distribution systems, communication networks, multimedia systems, man-machine interface, knowledge acquisition & management, data mining, data warehouse, cooperative technology, and service science & engineering.

4. **Information Applications and Innovations**
   The applications and innovations of business functional information systems (e.g., production, marketing, financial, human resources, and accounting information systems), enterprise resource planning, customer relationship management, supply chain management, intellectual capital, geographic information systems, and integrated information systems.

5. **Information Technology Education and Society**
   Information education, e-learning, and information impacts on society.

6. **Others**
   Other MIS-related topics.

INSTRUCTIONS FOR SUBMISSION

1. Papers can be prepared in either Chinese or English. If your paper is written in Chinese, it will be translated into English once it is accepted for publication.

2. There is no submission deadline for MIS Review. All papers will be double-blind reviewed by at least two reviewers, who will be recommended by the Editorial Board. The processing time for the first-round formal reviews is about six weeks. Subsequently rounds of reviews tend to be faster.

3. To simplify file conversion effort, PDF or Microsoft Word 2000/2003 (for Windows) format is advised. Then, please submit your paper via the MIS Review website (URL: http://www.icebnet.org/misr/).

4. MIS Review is an academic journal. According to international practice, once an article is accepted and published, MIS Review will not give or take any payment for the publishing. An electronic copy of the paper will be sent to the article author(s) for non-profit usage.

5. The submitted and accepted paper should follow the author guidelines for paper submission format provided on the MIS Review website.

The submitted paper should include the title page, abstract, key words, the paper body, references, and/or appendices. You must submit three files. The information of author(s) should not appear anywhere in the paper body file, including page header and footer.

1. On a separate (cover letter) file, please follow the author guidelines provided on the MIS Review website to prepare the letter.

2. On a separate (title page) file, please note the title of the paper, names of authors, affiliations, addresses, phone numbers, fax numbers, and E-mail addresses.

3. On a separate (paper body) file, please include the paper title, an abstract, a list of keywords, the paper body, the references, and/or appendices. The abstract must contain the research questions, purposes, research methods, and research findings. The abstract should not exceed 500 words and the number of keywords must be 5-10 words.

4. The submitted and accepted paper should follow the author guidelines for paper submission format provided on the MIS Review website.

CONTACT

Editorial Assistant
Department of Management Information Systems
College of Commerce, National Chengchi University
No. 64, Sec. 2, ZhiNan Road, Wenshan District,
Taipei 11605, Taiwan R.O.C.
Phone: +886-2-29393091 ext.89055
E-mail: misr@mis.nccu.edu.tw

airiti press Subscription Form

MIS Review

You may subscribe to the journals by completing this form and sending it by fax or e-mail to

Address: 18F., No. 80, Sec. 1, Chenggong Rd., Yonghe District, New Taipei City 23452, Taiwan (R.O.C.)
Tel: +886-2-29266006 ext. 8301 Fax: +886-2-29235151 E-mail: press@airiti.com Website: http://www.airitipress.com

PERSONAL				**LIBRARIES / INSTITUTIONS**			
	Europe	**US/CA**	**Asian/Pacific**		**Europe**	**US/CA**	**Asian/Pacific**
1 Issue	€ 34	US$ 41	US$ 38	**1 Issue**	€ 53	US$ 65	US$ 62

Vol.		No.		~	Vol.		No.		Copies		US$		Total US$	

*All Price include postage

PLEASE NOTE

- Issues will be sent in two business days after receiving your payment.
- Please note that all orders must be confirmed by fax or email.
- Prices and proposed publication dates are subject to change without notice.
- Institutions include libraries, government offices, businesses, and for individuals where the company pays for the subscription.
- Personal rates are available only to single-user personal subscribers for personal and non-commercial purposes.
- Airiti Press reserves its right to take appropriate action to recover any losses arising fromany intended or unintended misrepresentation of the term "Personal Subscriber".

BILLING INFORMATION

Name	
Company	
Tel	Fax
E-mail	
Shipping Address	

INTERNATIONAL PAYMENTS

Pay by Credit Card	
Card Type ☐JCB ☐MasterCard ☐Visa	
Card Name	
Card Number	
Expiry Date ________ / ________	CVV number
Signature	

Direct Bank Transfer	
Beneficiary	AIRITI INC.
Address	18F., No. 80, Sec. 1, Chenggong Rd., Yonghe District, New Taipei City 23452, Taiwan (R.O.C.)
Bank Name	E.Sun Commercial Bank, Ltd.Yong He Branch
Account No	0107441863017
Swift Code	ESUNTWTP
Bank Address	No.145, Zhongzheng Rd., Yonghe District, New Taipei City 23454, Taiwan (R.O.C.)