Volume 8, Issue 1
April 2015
ISSN: 1946-1836

JOURNAL OF INFORMATION SYSTEMS APPLIED RESEARCH

In this issue:

The **Journal of Information Systems Applied Research** (JISAR) is a double-blind peer-reviewed academic journal published by **EDSIG,** the Education Special Interest Group of AITP, the Association of Information Technology Professionals (Chicago, Illinois). Publishing frequency is currently semiannually. The first date of publication is December 1, 2008.

JISAR is published online (http://jisar.org) in connection with CONISAR, the Conference on Information Systems Applied Research, which is also double-blind peer reviewed. Our sister publication, the Proceedings of CONISAR, features all papers, panels, workshops, and presentations from the conference. (http://conisar.org)

The journal acceptance review process involves a minimum of three double-blind peer reviews, where both the reviewer is not aware of the identities of the authors and the authors are not aware of the identities of the reviewers. The initial reviews happen before the conference. At that point papers are divided into award papers (top 15%), other journal papers (top 30%), unsettled papers, and non-journal papers. The unsettled papers are subjected to a second round of blind peer review to establish whether they will be accepted to the journal or not. Those papers that are deemed of sufficient quality are accepted for publication in the JISAR journal. Currently the target acceptance rate for the journal is about 40%.

Questions should be addressed to the editor at editor@jisar.org or the publisher at publisher@jisar.org.

JOURNAL OF INFORMATION SYSTEMS APPLIED RESEARCH

Editors

JISAR Editorial Board

Cyberbullying or Normal Game Play? Impact of age, gender, and experience on cyberbullying in multi-player online gaming environments: Perceptions from one gaming forum

Meg Fryling
mfryling@siena.edu

Jami Cotler
jcotler@siena.edu

Computer Science
Siena College
Loudonville, NY 12211, USA

Jack Rivituso
rivitug@cobleskill.edu
Business and Information Technology
SUNY Cobleskill
Cobleskill, NY 12043, USA

Lauren Mathews
li08math@siena.edu

Shauna Pratico
sm22prat@alum.siena.edu

Computer Science
Siena College
Loudonville, NY 12211, USA

Abstract

This paper includes preliminary findings from a research study to investigate perceptions among adolescents and adults regarding prevalence, seriousness, and psychological impact of cyberbullying in multi-player online gaming environments. A survey was administered including questions regarding what gamers believe constitutes cyberbullying in online gaming environments, whether they have experienced cyberbullying in this space (i.e. witness, victim, or bully), and what, if any, psychological effects those experiences have had on them. The survey was posted to the Animal Crossing Community gaming forum and was completed by 1,033 respondents who report playing a variety of online games, in a variety of content levels (i.e. "early childhood" to "adult content"). Analyzing data from adolescent and adult respondents (ages 12-70) indicate that cyberbullying does occur in the online game space and can have negative psychological effects. In addition, an emergent theme from this research is that age, gender, and experience play an important role in perceptions regarding the frequency, seriousness, and impact of cyberbullying in online gaming environments.

Keywords: cyberbullying, online gaming, gender, MMORPG, cyber abuse, electronic bullying

1. INTRODUCTION

Research in the area of cyberbullying, especially in problem spaces such as social networking and texting has had a great deal of attention due to the increased number of tragic events resulting from cyberbullying using social media. Cyberbullying research began in the late 1990s and was largely in response to the growing use of technology among adolescents, as well as increased instances of cyber abuse among teenagers (Patchin & Hinduja, 2006; Yardi & Bruckman, 2011). These seminal studies primarily focused on the establishment of baseline information on prevalence of cyberbullying, as well as the various methods used by cyberbullies to harass their victims such as cell phone texting, YouTube videos, email, chat rooms, and online gaming.

The prevalence of the cyberbullying phenomenon has been researched among adolescents (Lenhart, 2010; Yardi & Bruckman, 2011). These studies have indicated that the burgeoning ownership of technology (e.g. smartphones, tablets, etc.) as well as the sharp increase in Internet and social networking site use has led to widespread cyberbullying victimization (Beran & Li, 2005; Kowalski & Limber, 2007; Mesch, 2009; Ortega, Elipe, Mora-Merchan, Calmaestra, & Vega, 2009; Patchin & Hinduja, 2006; Raskauskas & Stoltz, 2007; Ybarra, 2004). While much of the research has been focused on young adolescents, more recent work has investigated cyberbullying among older adolescents in college (Aricak, 2009; Dilmac, 2009; Molluzzo, Lawler, & Manneh, 2012; G. Rivituso, 2012; J. Rivituso, 2014; Smith & Yoon, 2012) as well as adults in the workplace (Keashly & Neuman, 2010; McKay, Arnold, Fratzl, & Thomas, 2008; Privitera & Campbell, 2009).

2. PSYCHOLOGICAL IMPACT OF CYBERBULLYING

Research has identified that cyberbullying causes severe psychological, emotional, and social problems among many of its victims (Blair, 2003; Juvonen & Gross, 2008; Patchin & Hinduja, 2006; G. Rivituso, 2012; J. Rivituso, 2014). Cyberbullying can have a long-lasting psychological impact on individuals; the result of which can include changes in self-efficacy, self-esteem and behavior. Researchers have offered varied theories as to the cause of these problems (Anderson & Sturm, 2007; Bandura, 1989, 1990; Diamanduros, Downs, & Jenkins, 2008). Additionally, research supports that such bullying has larger societal issues both inside and outside the cyber environments. In response to the negative stimuli of being cyberbullied, middle and high school student victims have been found to become cyberbullies themselves. (Berthold & Hoover, 2000; Fryling & Rivituso, 2013; Katzer, 2009; Wong & Xio, 2012; Ybarra & Mitchell, 2004). Berthold and Hoover (2000) reported that middle school student victims were more than three times as likely to bully others when compared to non-victims.

The psychological impact of cyberbullying may be more profound than that of traditional bullying because negative comments, threats, and accusations are often visible to a wide audience and are long-lasting. This content may be viewed repeatedly by the victim and their peers causing repeated victimization (Campbell, 2005; G. Rivituso, 2012; J. Rivituso, 2014; Strom & Strom, 2005). These factors generate a great deal of anxiety among victims and negatively impact their psychological state (Beale & Hall, 2007; DeHue, Bolman, & Vollink, 2008; Spear, Slee, Owens, & Johnson, 2009; Strom & Strom, 2005). The negative impact of cyberbullying leads to feelings of frustration, anger, and sadness that are detrimental to the victim's psychological well-being (Patchin & Hinduja, 2006). Victims of cyberbullying experience depressive symptoms, behavior problems, drug use, and negative attitudes toward school (Ybarra & Mitchell, 2004; Ybarra, 2004). Adolescent cyberbullying victims are likely to report behavioral issues, drinking alcohol, smoking, and depressive symptoms (Juvonen & Gross, 2008; Mason, 2008). Victims of cyberbullying quite often experience embarrassment, lowered self-esteem, and negative impacts on their academic, professional, personal and social life (Mesch, 2009) as well as an increased rate of suicidal thoughts (Kim, Koh, & Leventhal, 2005; Klomek, Sourander, & Gould, 2010; Patchin & Hinduja, 2007).

While elementary, middle, and high school students are the most researched groups regarding cyberbullying, researchers have found that older adolescents and adults can be victims of cyberbullying in college and the workplace (Bond, Tuckey, & Dollard, 2010; Chapell et al.,

2004; Cowie, Naylor, Smith, Rivers, & Pereira, 2002; De Cuyper, Baillien, & De Witte, 2009; Keashly & Neuman, 2010; Lester, 2009; Privitera & Campbell, 2009). However, scientific research on bullying and cyberbullying among older adolescents and adults within both college and the workplace is in its infancy with less than significant literature on the topic (Lester, 2009).

3. CYBERBULLYING IN ONLINE GAMING ENVIRONMENTS

More recently researchers have begun to investigate cyberbullying in online gaming environments. Yang (2012) examined 1,069 adolescent online game players in a quantitative study to explore the relationships between their gender, preference for video games, hostility, aggressive behavior, experiences of cyberbullying, and victimization. Participants were recruited from 16 elementary, middle, and high schools in three cities in Taiwan. Significant findings from this study indicate an association between male respondents and a preference for violent games, increased hostility, and aggressive behavior. Violent and bullying behavior in the online world does have significance outside of that environment as bullying behavior can cross-over between the online world (cyberbullying) and the physical world (traditional bullying). Yang (2012) found that male victims who had experienced repeated cyberbullying instances in online gaming, had a greater likelihood of observable aggressive behavior in his daily life.

A study conducted by Li (2006) involved 264 junior high school students from three Canadian schools. Findings from this study indicate that while boys and girls spend similar amounts of time online, there are distinct differences in behavior related to cyberbullying. The differences identified that boys were more likely to be involved in cyberbullying, but they were less likely to tell an adult if cyberbullying behavior was taking place. Leung and McBride-Chang (2013) conducted a study among 626 Hong Kong Chinese fifth and sixth grade students of both genders with the focus examining friendship and bullying experiences, both at school and in online computer gaming. Their findings indicate that while instances of cyberbullying are present in the online gaming environment, a positive development related to the social functioning of children was found, that being the development of friendships attributed to participation in online gaming.

Contemporary cyberbullying research has primarily focused on elementary, middle, and high school adolescents. As the phenomenon has grown, the research lens has changed from prevalence type studies to psychological type studies. Cyberbullying studies have begun to focus on the identification of the physical, emotional, and social problems associated with cyberbullying among adolescents. Findings have identified that victims of cyber abuse often suffer from a myriad of harmful stressors affecting the general well-being of the victim. There is still a need to further study the negative effects of cyberbullying victimization since research in this area still has a variety of gaps (Tokunaga, 2010). While some researchers have begun to examine this phenomenon among post-secondary students and adults, this work is somewhat limited. Most recently cyberbullying research has extended to the online gaming environment but these studies are also limited and have focused on elementary, middle, and high school students.

Additional research is warranted because of the limited amounts of empirical research on cyberbullying in gaming environments. In addition, an overlooked area in the existing research is the investigation of cyberbullying among older adolescents and adults (i.e. adult bullies and victims). A recent study conducted by Ipsos MediaCT for The Entertainment Software Association (2013) reported that 68% of gamers are adults (18 years or older). Therefore, it is important to not overlook this population in online gaming cyberbullying research.

As with traditional bullying, cyberbullying can have a long-lasting psychological impact on individuals; the result of which can include changes in self-efficacy, self-esteem and behavior. Therefore, strategies are needed to detect and mitigate cyberbullying wherever it may occur, including online gaming environments, and to whomever it may involve, including adult populations.

4. METHODOLOGY

This exploratory research investigates perceptions regarding cyberbullying prevalence and seriousness in gaming environments, with three primary research questions:

- Is there a perception that cyberbullying a problem in the online gaming environment among adolescents and/or adult populations?
- Have adolescent and/or adult gamers experienced cyberbullying in the game space, as a witness, victim, or bully?
- What, if any, psychological problems are resulting from cyberbullying in online gaming environments?

A survey instrument developed by the investigators of this study, including two undergraduate student researchers at a small liberal arts college, was used to address the research questions. This instrument incorporated questions from prior cyberbullying research studies (Molluzzo et al., 2012; Smith & Yoon, 2012) as well as a variety of new questions developed by the researchers to specifically address the study questions. The survey included questions spanning the participants belief of what constitutes cyberbullying in online gaming environments, whether they have experienced cyberbullying in this space (i.e. witness, victim, or bully), and what, if any, psychological effects those experiences have evoked. An expert online gamer was part of the team involved in developing the questions to ensure proper gaming "lingo" was followed (i.e. terms such as "aggroing" to refer to baiting monsters into attacking unprepared players and "griefing" to refer to deliberately irritating or harassing other players in the game).

Prior to releasing the final survey, it was pilot tested with a small group of online gamers. Minor modifications were made and the final survey, consisting of 42 questions, was posted to online games/forums and available to everyone in these environments. The questions broke down cyberbullying into specific behaviors that allowed participants to define cyberbullying and to identify specific behaviors they experienced as a witness, victim, and/or bully. Participants were asked specific questions addressing whether they believe cyberbullying is a problem in this environment and, for those who were victims or bullies, what psychological effect cyberbullying had on them. The survey also asked background information including age, gender, amount of time they spent playing, and their experience level.

Population and Sample

The surveys were initially distributed through a variety of online gaming discussion forums. In addition, a snowball sampling approach was attempted using social media. To encourage participation, respondents were offered the chance to win a $50 Amazon gift card upon completion of the survey. Email addresses of those interested in entering the drawing were collected and stored separately to ensure confidentially of responses. The most significant number of responses came from the Animal Crossing Community online gaming forum (http://www.animalcrossingcommunity.com/) and is the focus of this paper. The survey was posted in such a way that members were asked to participate upon login to the forum. We believe this, in addition to the gift card drawing, contributed greatly to the high response rate.

As of June 2014, the Animal Crossing Community gaming forum hosted 564,166 total members. The gender distribution of those completing the survey is similar (i.e. less than 1% difference) to that of forum members, for those that reported their gender (approximately 62% female and 38% male). Sixty-five percent of the individuals that report their age to the forum are over the age of 19. There was no response bias or known characteristic that predicted whether a participant responded to the survey.

One thousand four hundred and eighty-five participants started and 1033 (70%) completed the survey. One thousand twenty-five surveys were used in the analysis presented in this paper. Six responses under the age of 12 and two responses with an age of 116 were eliminated from the sample due to validity concerns. The population age range is 12 to 70 with an average age of 22.04, median age of 19 and mode age of 18. When asked to select a category that best describes the use of online gaming, 46% of the sample indicated that they most often participate as an Explorer (see Figure A1 in Appendix A).

Respondents may play games from multiple content types and were asked to select all categories that applied. Participants of the survey most often played games classified as **Everyone** (76%), **Everyone 10+** (58%), **Teen** (74%), and **Mature** (58%). Respondents play games from a variety of content levels, from content suitable for all ages to mature content. (see Table A1 in Appendix A). A majority (66%)

of the participants self-reported an experience level of advanced player and 14% classified themselves as an expert player.

Results

Before describing the data analysis, a brief overview of distributions of perceptions will be presented. 38% of respondents (43% of females and 28% of males) have avoided a multi-player video game because they were concerned about cyberbullying behavior. 54% (57% of females and 49% of males) have left a multi-player video game because someone was exhibiting cyberbullying behavior. 63.51% of the participants either agreed or strongly agreed that cyberbullying is a serious problem in the online gaming environment (see Figure 1), with females reporting higher at 68% then males at 55%.

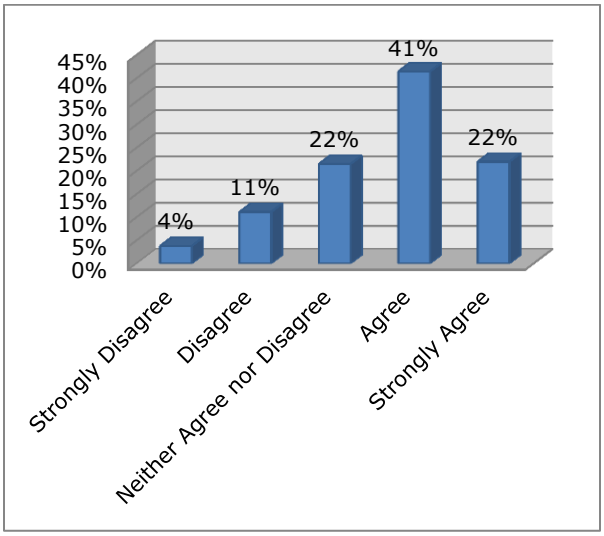

Figure 1: Cyberbullying is a serious issue within multi-player video games.

62.12% of the participants reported that cyberbullying occurs often to all of the time in online gaming environments. Again, female respondents reporting higher at 67% than male respondents at 53%.

We further broke down cyberbullying into categories of being a victim and witness of the bully. 78% of respondents have been a victim of cyberbullying (79% of females and 73% of males) in multi-player online gaming environments, 91% (same for both males and females) have witnessed cyberbullying, and 35% (29% of females and 42% of males) admit to exhibiting cyberbullying behavior (see Figure A2 in Appendix A). Female respondents were slightly more inclined to report cyberbullying in the gaming environment (49%) versus male respondents (45%).

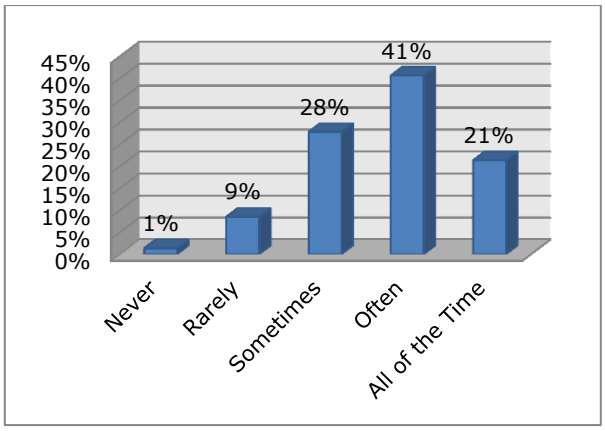

Figure 2: What degree does cyberbullying occur within multi-player video games?

There is also a perception that females are more likely to be the victim of cyberbullying and less likely to be the perpetrator of cyberbullying than males. 26% of respondents reported that females are more likely to be cyberbullied, 10% think males are more likely to be cyberbullied, and 64% think both are equally likely to be cyberbullied. 58% believe males are more likely to be a cyberbully, 1% believe females are more likely to be a cyberbully, and 41% think both are equally likely to be cyberbully.

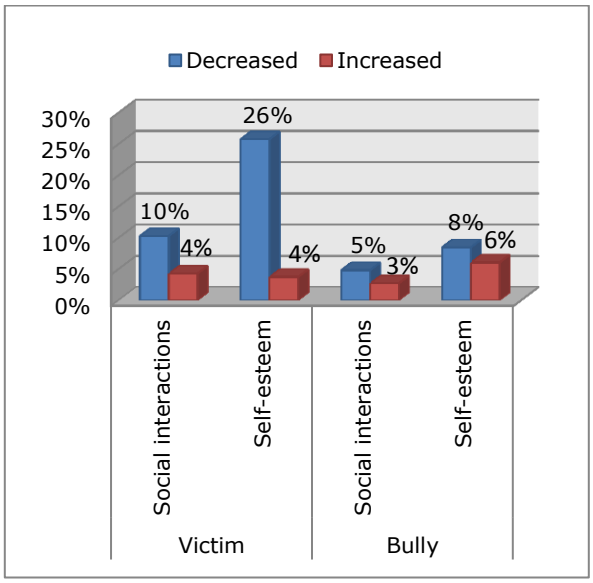

Figure 3: Comparison of psychological impact between victim and bully

Psychological impact (victim and bully)
When bullying behavior occurs both the victim and bully are negatively impacted psychologically. Both groups report a net decrease in both social interactions and self-esteem (see Figure 3). Female respondents experienced a greater negative psychological impact. For example, the net decrease in the self-esteem of female cyberbullying victims is 27% versus male victims at 12% (see Figure 4).

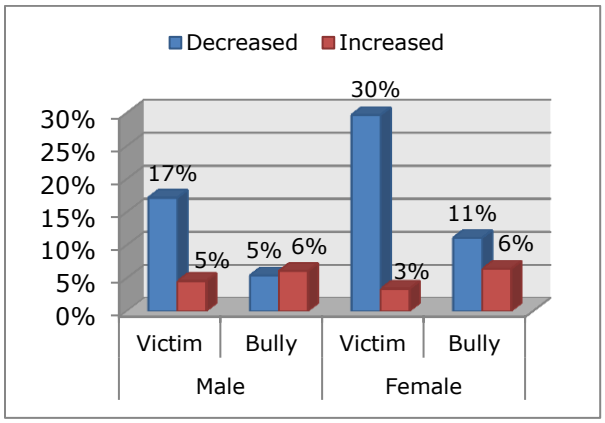

Figure 4: Comparison of impact on self-esteem between victim versus bully and male versus female

Additional negative factors such as aggressiveness, stress, anxiety, anger, and depression have a net increase for both the victim and bully (see Figure A3 in Appendix A). Male and female respondents had similar negative impacts on these factors as the victim of cyberbullying. However, female respondents have notably greater negative consequences when acting as the bully (see Figures A4 and A5 in Appendix A).

Data Mining
Further analysis was conducted using the Waikato Environment for Knowledge Analysis (WEKA). WEKA is a free data mining tool consisting of a collection of machine learning algorithms that can applied directly to a dataset (see http://www.cs.waikato.ac.nz/ml/weka/). The WEKA workbench allows for automatic analysis of large datasets to identify which data are most relevant. WEKA contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. We used WEKA to perform both supervised (linear regression) and unsupervised (clustering) data-mining techniques.

Linear Regression
Linear regression was used as a form of supervised data mining to test hypotheses that emerged from the initial data analysis, described in the previous section. For example, based on a given age and experience level, how likely will the person feel cyberbullying is serious?

WEKA develops a regression model by only using the independent variables that statistically (measured in R-squared) contribute to the accuracy of the model. The following independent variables were provided to WEKA for consideration in the regression model: age, gender, and gaming experience level (see Table 1)

Independent Variable	Values
Age	Coded as actual age
Gender	1=Male 2=Female
Experience	1=None 2=Beginner 3=Intermediate 4=Advanced 5=Expert

Table 1: Independent Variable Coding

Using WEKA's linear regressions allowed us to analyze the data by predicting a dependent numerical value for the given set of independent variables. The dependent variables that were analyzed corresponded to the three questions below.

- Is cyberbullying a serious problem?
- How often is cyberbullying occurring in online gaming?
- How often have you experienced, witnessed or participated in cyberbullying?

Several regressions were run and the most significant ones are reported here. The results of the following questions served as a dependent variable.

Question 1: Based on your definition of cyberbullying, select the degree to which you agree with the following statement:

Cyberbullying is a serious issue within multi-player video games. (1=Strongly Disagree to 5=Strongly Agree)

WEKA produced the following regression formula:

Serious Issue =
*0.0109 * Age +*
*0.3578 * Gender +*
*-0.0905 * Experience +*
3.19

To summarize the results, older females (ages 63 – 70) with little experience are more likely to strongly agree that cyberbullying a serious problem. Young males (12 – 37), who are more experienced gamers, are more likely to neither agree or disagree that cyberbullying is a serious problem.

Question 2: Based on your definition of cyberbullying, to what degree would you say cyberbullying occurs within multi-player video games? (1=Never to 5=All of the Time)

WEKA produced the following regression formula:

Degree Occurs =
*0.2426 * Gender +*
3.3209

This formula reveals that both genders perceive cyberbullying occurring *sometimes* to *often*, with more females stating cyberbullying happens *often*. Age and experience were not determinate factors.

Question 3: Based on your definition of cyberbullying, please estimate how often you have experienced cyberbullying (as a victim) within multi-player video games. (1=Never to 5=All the time)

WEKA produced the following regression formula:
Victim Frequency =
*-0.0063 * Age +*
*0.2254 * Gender +*
*0.1587 * Experience +*
1.4405

This formula predicts that young females with high gaming experience are the most likely to be a cyberbullying victim in the online gaming environment. As age increases the likelihood of being a victim decreases.

Question 4: Based on your definition of cyberbullying, please estimate how often you have experienced cyberbullying (as a witness) within multi-player video games. (1=Never to 5=All the time)

WEKA produced the following regression formula:

Witness Frequency =
*-0.0099 * Age +*
*0.1988 * Gender +*
*0.1622 * Experience +*
2.3523

This model predicts that all respondents are likely to witness some cyberbullying in the online gaming environment. Similarly to the victim frequency regression model, young females with high gaming experience are the most likely to witness cyberbullying.

Question 5: Based on your definition of cyberbullying, please estimate how often you have experienced cyberbullying (as individual exhibiting bullying behavior) within multi-player video games. (1=Never to 5=All the time)

WEKA produced the following regression formula:

Exhibit Cyberbullying behavior Frequency =
*-0.0081 * Age +*
*-0.1289 * Gender +*
1.8599

This formula reveals that experience is not a factor in determining whether or not an individual will exhibit cyberbullying behavior. The model predicts that older females are least likely to exhibit cyberbullying behavior and young males are most likely.

Overall, using a supervised method of data mining for the regression analysis provided an opportunity to further answer the research questions and offered a deeper understanding of some of emergent themes central to age, gender and experience.

Clustering

In addition to answering the initial research questions, running unsupervised data mining techniques helps develop future research questions and hypotheses. Clustering is an unsupervised form of data-mining that does not test a hypothesis but rather it lets patterns emerge from the data. In clustering every

attribute is used to analyze the data. For example, what age groups or genders are most likely to perceive cyberbullying as a serious problem?

Considering the psychological effects of cyberbullying on both the victim and bully, learning more about the bully may offer a scaffold for future inquiry. A cluster analysis provides the framework to create behavioral models. Table 2 below highlights the groups as they emerge as a bully or non-bully. The characteristics are identical with the exception of age. For both male and females the younger counterpart, with all other characteristics equal, emerges as the bully. While the age difference is only a few years, it is worth further investigation in future research.

Cluster 0 (187 – 18%) Age 21 Female Intermediate player Plays 10–39 hrs/week **Bully? Yes**	Cluster 1 (467 – 46%) Age 23 Female Intermediate player Plays 10–39 hrs/wk Bully? No
Cluster 2 (219 – 21%) Age 22 Male Intermediate player Plays 10–39 hrs/week Bully? No	Cluster 3 (152 – 15%) Age 20 Male Intermediate player Plays 10–39 hrs/wk **Bully? Yes**

Table 2: Cluster Analysis (Bully)

Building on the cluster analysis above, frequency of being a victim or witnessing cyberbullying behaviors was added. By adding the additional dependent variables of victim and witness, behavior models displaying the characteristics of the cyberbully may become apparent. The objective of using these variables is to have a better understanding of the behaviors that may cause cyberbullying behaviors. For example, if a person is a victim do they become a cyberbully? Or if a gamer is a witness does this also contribute to cyberbullying behaviors?

Cluster 0 (341 – 33%) Age 23 Female Advanced Player Plays 40-69 hrs/wk Rarely a victim Rarely a witness **Never a bully**	Cluster 1 (313 – 31%) Age 23 Female Advanced Player Plays 10–39 hrs/wk Sometimes a victim Often a witness **Rarely a bully**
Cluster 2 (197 – 19%) Age 21 Male Advanced Player Plays 0 – 9 hrs/wk Rarely a victim Rarely a witness **Never a bully**	Cluster 3 (174 – 17%) Age 21 Male Advanced Player Plays 10–39 hrs/wk Sometimes a victim Often a witness **Rarely a bully**

Table 3: Cluster Analysis (Victim, Witness, & Bully)

The cluster analysis above (Table 3) shows that for both males and females if they have minimal exposure to being a victim or witness they are likely to not engage in bullying behaviors. In contrast, if a gamer is more exposed to being a victim and witness they are more likely to exhibit cyberbullying behaviors themselves.

5. DISCUSSION

Limitations

While the survey had over 1000 respondents it is important to note that the entire sample used for this paper were all members of the Animal Crossing Community gaming forum. While individuals in this community report playing a variety of online games, in a variety of content levels (i.e. "early childhood" to "adult content"), further research is needed to determine whether or not our findings can be generalized to the online gaming community at large.

While the age and gender distributions of respondents were reflective of the Animal Crossing Community forum members, they were not with the total population of the online gaming community. 62% of the respondents indicated that they are female, while ESA reports that only 45% of gamers are female. The average age of our respondents was 22, while the ESA reports that the average age of gamers is 30 (Ipsos MediaCT, 2013).

Future research will include distributing the survey through other channels to increase the diversity of respondents, decrease any unknown bias towards members of the Animal Crossing Community gaming forum, and augment the findings outlined in this paper. Additional investigation will contain further data mining to look for more patterns, other predictors of bullying behavior, factors contributing to negative psychological impacts, and dynamics that may contribute to the mitigation of cyberbullying.

Conclusions

The objective of this study was to investigate perceptions among adolescents and adults regarding prevalence, seriousness, and psychological impact of cyberbullying in multi-player online gaming environment. Preliminary analysis of the data supports prior research that suggests that there are instances of cyberbullying in online gaming environments (e.g. Leung and McBride-Chang, 2013; Li, 2006; Yang, 2012) and extends that work by including adult populations.

This study also supports the hypothesis that there are negative psychological consequences of cyberbullying in online gaming. Similar to the findings by Li (2006) our male respondents were slightly more likely to exhibit bullying behavior and slightly less likely to report cyberbullying incidents than female respondents.

Yang (2012) found that male victims who had experienced repeated cyberbullying instances in online gaming, had a greater likelihood of observable aggressive behavior in his daily life. Our study did not find a notable difference between male and female victims in regard to increased aggressive behavior but both showed a net increase.

Cluster analysis revealed that cyberbully victims and witnesses may be more likely to exhibit cyberbullying behavior. This finding supports prior research (e.g. Fryling & Rivituso, 2013; Shu Ching Yang, 2012) that suggests cyberbullying victimization increases the likelihood of exhibiting cyberbullying behavior.

Overall our male respondents were slightly less negatively impacted psychologically by being bullied than female respondents. However, female respondents reported a notable greater net increase over male respondents in aggressiveness, stress, anxiety, anger, and depression after exhibiting cyberbullying behavior.

While individuals of all genders, age groups, and experience levels may be impacted by cyberbullying in online gaming environments, perceptions regarding the seriousness of such activities varied among these groups. Older females with less gaming experience reported the highest perception of cyberbullying occurrence, seriousness, and victimization. Conversely, younger male respondents with more gaming experience report the lowest perception of cyberbullying occurrence, seriousness, and victimization. Females are more likely to be negatively impacted psychologically, particularly when exhibiting cyberbullying behavior, and are more likely to avoid or leave a game due to cyberbullying behavior.

This research serves to enhance the understanding of the general public by identifying that cyberbullying activities transcend the social networks, cell phones, email, and chat rooms. The study aims to identify that computer gaming, often sought by users of all ages as a means of entertainment and even relaxation, has as an inherent risk and participates are vulnerable to cyberbullying activities. The work sought to begin to understand the social norms of bullying behavior in gaming environments by investigating perceptions regarding cyberbullying prevalence, seriousness, and psychological impact.

The findings from this research add to the academic and scientific understanding of cyberbullying in the problem space of gaming. Findings add to the growing database of empirical knowledge on this construct for both adolescents and adults. Future research will explore triggers of cyberbullying behavior in the online gaming environment and mitigation strategies, including technological enhancements to monitor and to mitigate cyberbullying. Our ultimate objective in future research is to better understand under what conditions cyberbullying occurs and to provide some best practices in prevention with possible human-computer interaction interventions.

ACKNOWLEDGEMENTS

Thank you to Jerad Rose from Animal Crossing Community for your support in distributing the survey.

This research was partially funded by the Siena College Center for Undergraduate Research and Creative Activity.

6. REFERENCES

Anderson, T., & Sturm, B. (2007). Cyberbullying from playground to computer. *Young Adult Library Services*, *5*, 24–27.

Aricak, T. (2009). Psychiatric symptomatology as a predictor of cyberbullying among

university students. Eurasian Journal of Educational Research, 34, 167-184., *34*, 167–184.

Bandura, A. (1989). Human Agency in Social Cognitive Theory. *American Psychologist*, *44*, 1175–1184.

Bandura, A. (1990). Some Reflections on Reflections. *Psychological Inquiry*, *1*, 101–105.

Beale, A., & Hall, K. (2007). Cyberbullying: What School Administrators (and Parents) Can Do. *The Clearing House - Heldref Publications*, *81*, 8–12.

Beran, T., & Li, Q. (2005). Cyber-harassment: A Study of a New Method for an Old Behavior. *Cyber-Harassment: A Study of a New Method for an Old Behavior*, *32*, 256–277.

Berthold, K., & Hoover, J. (2000). Correlates of Bullying and Victimization among Intermediate Students in the Midwestern USA. *School Psychology International*, *21*, 65–78.

Blair, J. (2003). New breed of bullies torment their peers on the internet. *Education Week*, *22*, 6.

Bond, S., Tuckey, M., & Dollard, M. (2010). Psychosocial Safety Climate, Workplace Bullying, and Symptoms of Posttraumatic Stress. *Organization Development Journal*, *28*, 38–56.

Campbell, M. (2005). Cyber bullying: An Old Problem in a New Guise? *Australian Journal Of Guidance & Counseling*, *15*, 68–76.

Chapell, M., Casey, D., De la Cruz, C., Ferrell, J., Forman, J., Lipkin, R., … Whittaker, S. (2004). Bullying in College by Students and Teachers. *Adolescence*, *39*, 54–64.

Cowie, H., Naylor, P., Smith, P., Rivers, I., & Pereira, B. (2002). Measuring Workplace Bullying. *Aggression and Violent Behavior*, *7*, 35–51.

De Cuyper, N., Baillien, E., & De Witte, H. (2009). Job Insecurity, perceived employability and targets' and perpetrators' experiences of workplace bullying. *Work & Stress*, *23*, 206–224.

DeHue, F., Bolman, C., & Vollink, T. (2008). Cyberbullying: Youngsters' Experiences And Parental Perception. *CyberPsychology & Behavior*, *11*, 217–222.

Diamanduros, T., Downs, E., & Jenkins, S. (2008). The role of school psychologists in the assessment, prevention, and intervention of cyberbullying. *Psychology in the Schools*, *45*, 693–704.

Dilmac, B. (2009). Psychological Needs as a Predictor of Cyber bullying: a Preliminary Report on College Students. *Educational Sciences: Theory & Practice*, *9*, 1307–1325.

Fryling, M., & Rivituso, G. (2013). Investigation of the Cyberbullying Phenomenon as an Epidemic. Presented at the 31st International Conference of the System Dynamics Society, Cambridge, MA.

Ipsos MediaCT. (2013). *Essential Facts About the Computer and Video Game Industry* (Online). Entertainment Software Association (ESA). Retrieved from http://www.theesa.com/facts/pdfs/esa_ef_2013.pdf

Juvonen, J., & Gross, E. (2008). Extending the School Grounds?-Bullying Experiences in Cyberspace. *Journal of School Health*, *78*, 496–505.

Katzer, C. (2009). Cyberbullying: Who Are the Victims? *Journal of Media Psychology*, *2*, 25–36.

Keashly, L., & Neuman, J. (2010). Faculty Experiences with Bullying in Higher Education. *Administrative Theory & Praxis*, *32*, 48–70.

Kim, Y., Koh, Y., & Leventhal, B. (2005). School Bullying and Suicidal Risk among Korean Middle School Students. *Pediatrics: The Official Journal of the American Academy of Pediatrics*, *115*, 357–363.

Klomek, A., Sourander, A., & Gould, M. (2010). The Association of Suicide and Bullying in Childhood To Young Adulthood: A Review of Cross-Sectional and Longitudinal Research Findings. *The Canadian Journal of Psychiatry*, *55*, 282–288.

Kowalski, R. M., & Limber, S. P. (2007). Electronic Bullying Among Middle School Students. *Journal of Adolescent Health*, *41*,

S22–S30. doi:10.1016/j.jadohealth.2007.08.017

Lenhart, A. (2010). *Cyberbullying 2010: What the research tells us*. Retrieved from www.pewinternet.org/Presentations/2010/May/Cyberbullying-2010.aspx

Lester, J. (2009). Not Your Child's Playground: Workplace Bullying Among Community College Faculty. *Community College Journal of Research and Practice*, *33*, 446–464.

Leung, A. N., & McBride-Chang, C. (2013). Game on? Online Friendship,Cyberbullying, and Psychosocial Adjustment in Hong Kong Chinese Children. *Journal of Social and Clinical Psychology*, *32*(2), 159–185.

Li, Q. (2006). Cyberbullying in schools: A research of gender differences. *School Psychology International*, *27*, 157–170.

Mason, K. (2008). Cyberbullying: A Preliminary Assessment for School Personnel. *Psychology in the Schools*, *45*, 323–348.

McKay, R., Arnold, D., Fratzl, J., & Thomas, R. (2008). Workplace Bullying In Academia: A Canadian Study. *Employee Responsibilities Rights*, *20*, 77–100.

Mesch, G. (2009). Parental Mediation, Online Activities, and Cyberbullying. *CyberPsychology & Behavior*, *12*, 387–393.

Molluzzo, J. C., Lawler, J., & Manneh, J. (2012). A Comprehensive Survey on Cyberbullying Perceptions at a Major Metropolitan University – Faculty Perspectives. Presented at the Information Systems Educators Conference. Retrieved from http://proc.isecon.org/2012/pdf/1918.pdf

Ortega, R., Elipe, P., Mora-Merchan, J., Calmaestra, J., & Vega, E. (2009). The Emotional Impact on Victims of Traditional Bullying and Cyberbullying A Study of Spanish Adolescents. *Journal of Psychology*, *217*, 197–204.

Patchin, J., & Hinduja, S. (2006). Bullies Move Beyond the School Yard: A Preliminary Look at Cyberbullying. *Journal of School Violence*, *4*(2), 123–147.

Patchin, J., & Hinduja, S. (2007). Offline Consequences of Online Victimization: School Violence and Delinquency. *Journal of School Violence*, *6*, 89–112.

Privitera, C., & Campbell, M. (2009). Cyberbullying: The New Face of Workplace Bullying? *CyberPsychology & Behavior*, *12*, 395–400.

Raskauskas, J., & Stoltz, A. (2007). Involvement in traditional and electronic bullying among Adolescents. *Developmental Psychology*, *43*, 564–575.

Rivituso, G. (2012, July 17). *Cyberbullying: An Exploration of the Lived Experiences and the Psychological Impact of Victimization among College Students An Interpretive Phenomenological Analysis* (Dissertation). Northeastern University, Boston, MA.

Rivituso, J. (2014). The Lived Experiences and Psychological Impact of Cyberbullying Victimization Among College Students. *Journal of Information Systems Education (JISE)*, *Forthcoming*.

Shu Ching Yang. (2012). Paths to Bullying in Online Gaming: The Effects of Gender, Preference for Playing Violent Games, Hostility, and Aggressive Behavior on Bullying. *Journal of Educational Computing Research*, *47*(3), 235–249.

Smith, J. A., & Yoon, J. (2012). Cyberbullying Presence, Extent, & Forms in a Midwestern Post-secondary Institution. Presented at the Information Systems Educators Conference. Retrieved from http://proc.isecon.org/2012/pdf/1945.pdf

Spear, B., Slee, P., Owens, L., & Johnson, B. (2009). Behind the Scenes and Screens Insight Into the Human Dimension of Covert and Cyberbullying. *Journal of Psychology*, *217*, 189–196.

Strom, P., & Strom, R. (2005). Cyberbullying by Adolescents: A Preliminary Assessment. *The Educational Forum*, *70*, 21–32.

Tokunaga, R. . (2010). Following you home from school: A critical review and sythesis of research on cyberbullying victimization. *Computers in Human Behavior*, *26*(3), 277–287.

Wong, Y. M., & Xio, B. S. (2012). An Empirical Investigation of Factors Instigating, Impelling, and Inhibiting Cyber-Bullying

Behavior. Presented at the AMCIS 2012, Seattle, WA. Retrieved from http://aisel.aisnet.org/amcis2012/proceedings/HCIStudies/29

Yardi, S., & Bruckman, A. (2011). Social and technical challenges in parenting teens' social media use (pp. 3237–3246). Presented at the Proceedings of the 2011 annual conference on Human factors in computing systems, 1979422: ACM. doi:10.1145/1978942.1979422

Ybarra, M. L. (2004). Linkages between depressive symptomatology and Internet Harassment among young regular Internet users. *CyberPsychology & Behavior*, *7*, 247–257.

Ybarra, M. L., & Mitchell, K. J. (2004). Online aggressor/targets, aggressors, and targets: A comparison of associated youth characteristics. *Journal of Child Psychology and Psychiatry*, *45*, 1308–1316.

Editor's Note:

This paper was selected for inclusion in the journal as the CONISAR 2014 Best Paper The acceptance rate is typically 2% for this category of paper based on blind reviews from six or more peers including three or more former best papers authors who did not submit a paper in 2014.

Appendix A – Figures and Tables

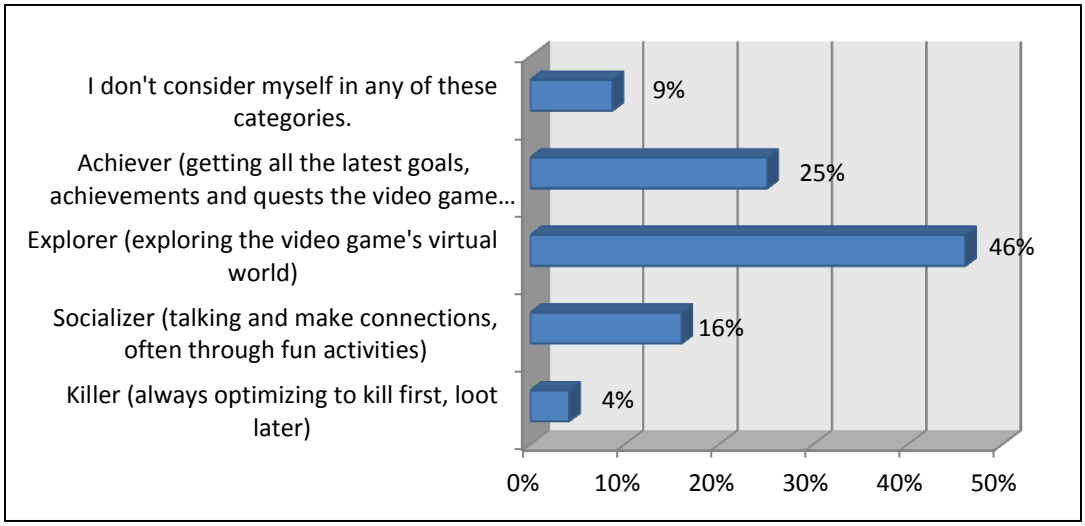

Figure A1: Which of the following categories best describes you?

Entertainment Software Rating Board (ESRB) rating	Percent Play
Early Childhood: Content is intended for young children.	3%
Everyone: Content is generally suitable for all ages. May contain minimal cartoon, fantasy or mild violence and/or infrequent use of mild language.	76%
Everyone 10+: Content is generally suitable for ages 10 and up. May contain more cartoon, fantasy or mild violence, mild language and/or minimal suggestive themes.	58%
Teen: Content is generally suitable for ages 13 and up. May contain violence, suggestive themes, crude humor, minimal blood, simulated gambling and/or infrequent use of strong language.	74%
Mature: Content is generally suitable for ages 17 and up. May contain intense violence, blood and gore, sexual content and/or strong language.	58%
Adults Only: Content suitable only for adults ages 18 and up. May include prolonged scenes of intense violence, graphic sexual content and/or gambling with real currency.	8%

Table A1: Game Content Types Played

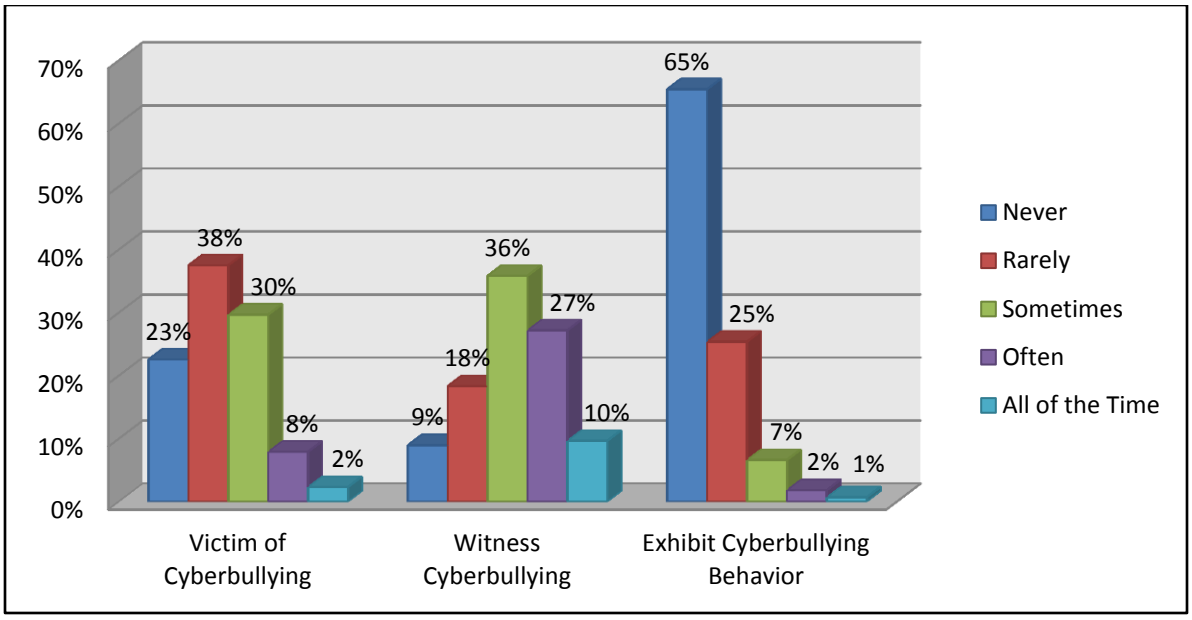

Figure A2: Comparisons of perceptions of frequency of being a victim, witness and/or a bully

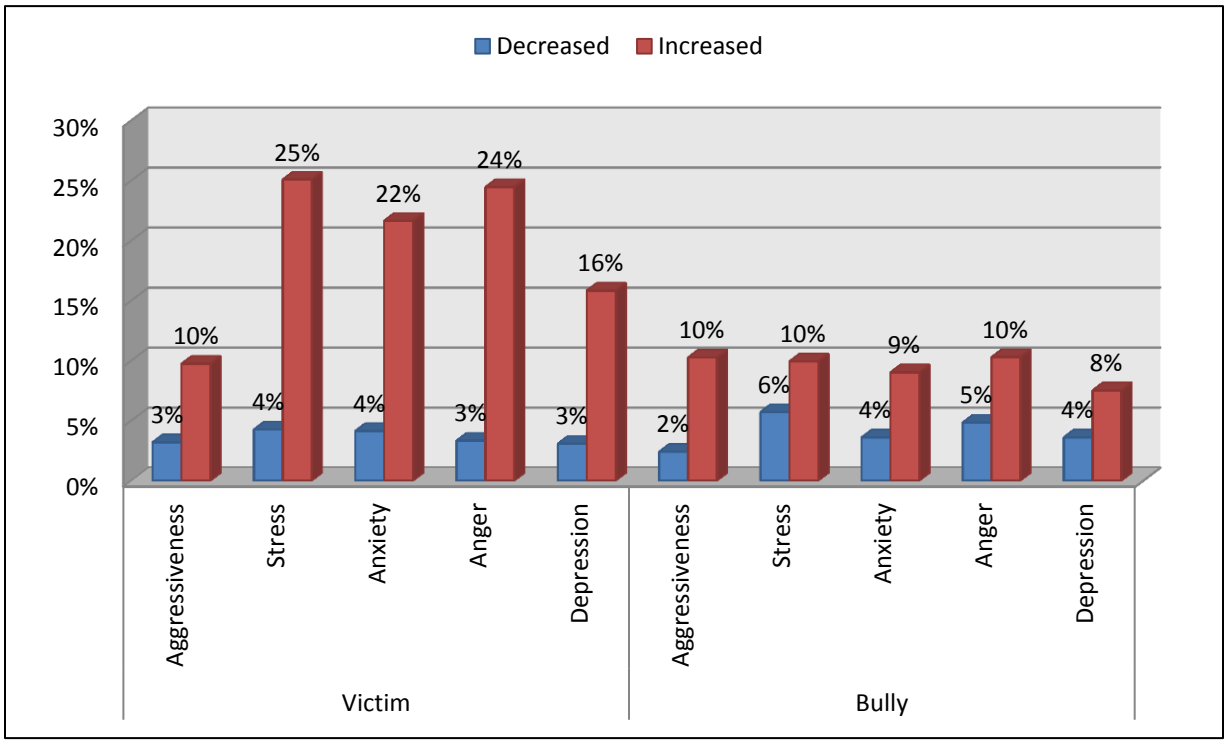

Figure A3: Comparison of psychological impact between victim and bully (All Respondents)

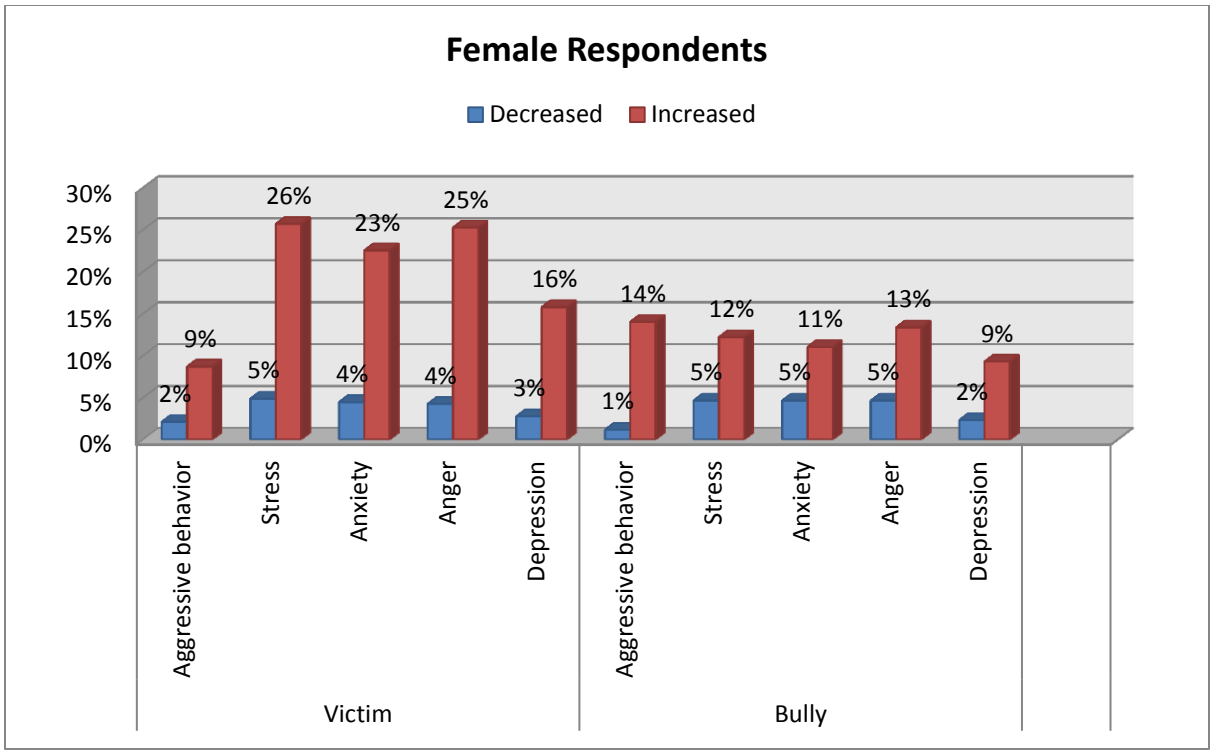

Figure A4: Comparison of psychological impact between victim and bully (Females only)

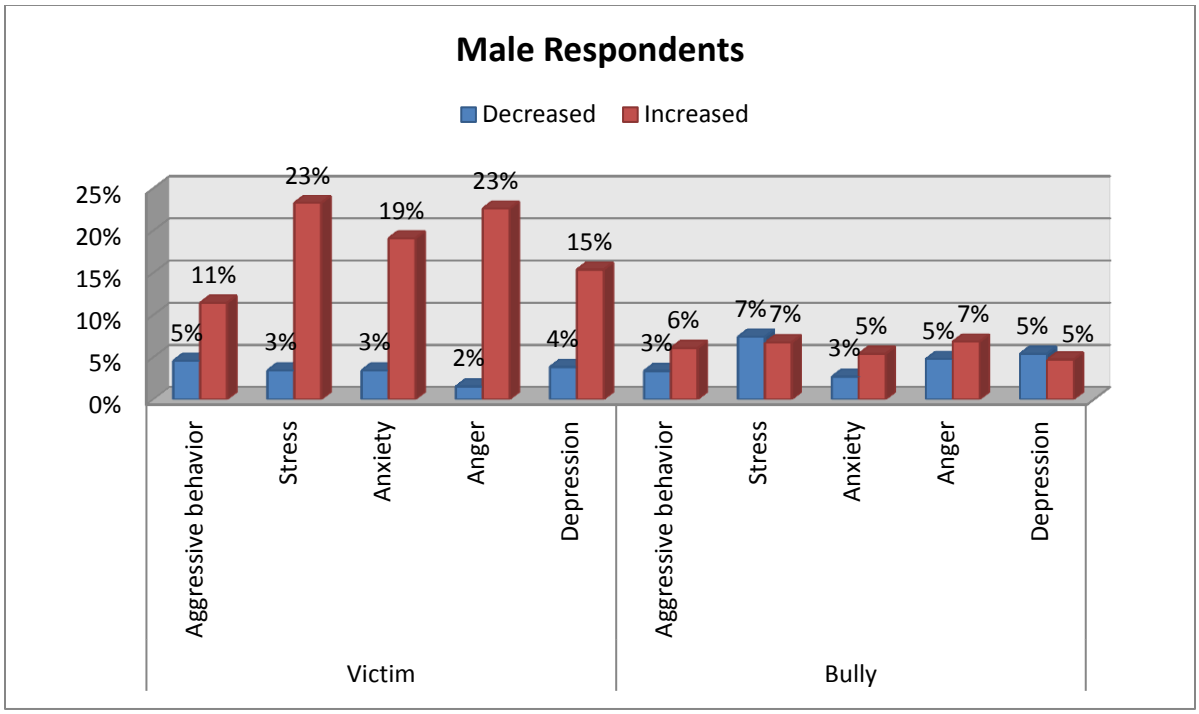

Figure A5: Comparison of psychological impact between victim and bully (Males only)

The Silent Treatment in IT Projects: Gender Differences in Inclinations to Communicate Project Status Information

Melinda Korzaan
melinda.korzaan@mtsu.edu

Nita Brooks
nita.brooks@mtsu.edu

Computer Information Systems
Middle Tennessee State University
Murfreesboro, TN 37128, USA

Abstract

Incomplete and inaccurate information in Information Technology project status reporting results in a project becoming vulnerable to unexpected problems and potentially blindsiding stakeholders to impending project failure. The research presented in this study extends current knowledge of project status reporting by focusing on the inclination of project team members to communicate key project status information to members of upper management. A sample of 222 individuals currently working on IT projects were surveyed and both individual and work climate variables were tested in a simple direct effects model to predict inclination to report project status information to upper management (IRPI). To investigate potential individual differences based on gender the model was also run for the sample of male worker and female workers. Results show that there are differences in the relationships in the model based on gender. For males the factors that significantly predict IRPI include a sense of responsibility for the project, over optimism of project success, and potential negative consequences for reporting status information (NC). For females the factors that significantly predicted IRPI were the project development phase and NC. Although NC predicted IRPI for both genders, the effect was stronger for men than women. Implications for practice research are discussed.

Keywords: Project Management, Whistleblowing, Project Status Reporting, Software Development

1. INTRODUCTION

Accurate and timely reports to upper management about Information Technology (IT) project status is vital for avoiding costly calamities, yet when the reports involve bad news there is a reluctance to relay that information to those who have power and authority to take corrective actions (Keil, Smith, Pawlowski, & Jin, 2004). Research on resistance to communicate bad news is explained primarily in IT research through whistleblowing theory (Keil et al, 2004, Smith, Keil, & Depledge, 2001). Such resistance contributes to the problem of misreporting IT project status and may be caused by factors in the following categories: individual traits, work climate, and cultural differences (Keil, Smith, Iacovou, & Thompson, 2014). The goal of this study is to expand previous research by examining gender differences in the relationships between the inclination to report project information (IRPI) and both individual and work climate factors.

An exploratory study is conducted to investigate predictors of individual inclinations to discuss project status information with member(s) of upper management. Separate predictive models are generated for both male and female participants to reveal potential gender differences in these reporting inclinations.

2. LITERATURE REVIEW AND HYPOTHESES

Although this study is primarily exploratory in nature, the hypotheses and overall predictive model is based on prior research in IT project status reporting and whistleblowing. Studies have shown that IT projects usually give advanced warning signals of imminent failure. However, warning signals are often ignored or reported with a biased positive spin (Keil et al, 2014, Cueller, Keil, & Johnson, 2006). Research in project management and project status reporting have found that individual assessment of whether the project status ought to be reported along with an assessment of personal responsibility to report project status information influences individual reluctance to report status information. Additional indirect influential factors include perceived information asymmetry and organizational climate (Keil et al, 2004). Furthermore, Keil et al. (2014) published a summary of research findings from over 14 studies over the past 15 years in five key truths about why status reports go wrong. A succinct description of these five truths are: (1) Executives can't rely on staff to accurately report problems, (2) Causes for misreporting project status include personal traits, work climate, and culture, (3) An audit team cannot offset the effects of misreporting and withholding project status information (4) A senior executive placed in charge of a project may increase misreporting in project status information, and (5) Executives frequently ignore negative information about projects (Keil et al, 2014).

Whistleblowing and Project Status Reporting

Project status reporting literature has relied significantly on the theoretical backdrop of organizational whistle-blowing. Whistle-blowers are described as "organization members who disclose information about dysfunctional organizational activities to either people or organizations who may be able to address the problems." (Keil et al, 2004, p.66). The dysfunction in the context of IT projects is when there is information indicating a significant problem or impending project failure yet nothing is being done to address the problem or redirect the project from its current failing path. Cueller et al. (2006) also identify reporting bad news as theoretically similar to whistle-blowing. Individuals may resist reporting bad news in order to avoid any negative repercussions and some may avoid speaking up due to personal perceptions such as feeling they lack confidence in their understanding of the trouble the project is experiencing or feeling like it is not their place or responsibility to report the information (Smith et al, 2001, Keil et al, 2004, Cuellar et al, 2006).

Individual and Work Climate Factors

The aforementioned literature (Keil et al, 2014) identifies individual characteristics, work climate, and culture influences as factors influencing project status reporting. For this study, the scope is delimited to individual and work climate factors. Abbreviation for all variables included in this study are identified in Table 1: Construct Abbreviations located in the Appendices. Individual characteristics investigated include: age, education, number of years working for the organization, number of years in IT, a feeling of responsibility and accountability for the project, and optimism of the project ultimately being successful. It is proposed that there will be a significant positive relationship between these individual characteristics and IRPI. The proposed positive relationships with age, number of years in IT, number of years in the organization, education, and IRPI may be explained in part by the logic in the following sentences. As individuals gain more experience in IT, their organization, and life in general they main gain confidence in their assessment and interpretation of project information. As confidence is gained in the assessment of whether the information would be important to communicate, individuals would be likely to go ahead and report such information to members of upper management (Keil et al, 2004).

*H1: The number of years an individual is at the organization (**YO**) will have a positive effect on **IRPI**.*

*H2: The number of years an individual has worked in the field of Information Technology (**YIT**) will have a positive effect on **IRPI**.*

*H3: Age will have a positive effect on **IRPI**.*
*H4: Education (**EDU**) will have a positive effect on **IRPI**.*

Keil et al (2004) found that perceptions of responsibility toward reporting status information decreased reluctance to report information. This study explores the perceptions of responsibility toward the project itself rather than toward status reporting. If an individual feels personally responsible for the outcome of a project they may be more inclined to talk candidly to members of upper management about the project status. It may be reasoned that an individual would be more likely to discuss project status especially to those who would have the authority to allocate the resources needed to address problems and ultimately improve the outlook for project success.

*H5: Perceptions of responsibility (**RES**) for project success will have a positive effect on **IRPI**.*

Optimistic beliefs about the probability of project success may increase the likelihood that a worker would be willing to discuss the status of the project with upper management. If the information to be relayed isn't negative and one is optimistic of project success, then it is logical that the individual would not be hesitant in discussing the project status with upper management. However, if the information is negative then being optimistic about project success may soften the blow of relaying bad news. Believing that the project will eventually be successful (even in spite of a troubled project status) may help communicators offer a positive note to offset a negative message. Furthermore, it may deflect negative consequences of delivering bad news if the overall impact on the project can be minimized and presented as not killing the project's overall likelihood of success. Such a communication tactic is analogous to using a politeness strategy to minimize the threat of the bad news and may be used as a communication approach to lessen the impact of a negative message (Lee, 1993).

*H6: Optimistic belief in project success (**OPS**) will have a positive effect on **IRPI**.*

Factors related to the project and work climate include the project development phase, negative information about the project, and negative consequences for communicating project status. The project development phase is placed in the work climate category, primarily because the development process along with social and the work environment connected to the development process are most closely connected to work climate as opposed to either individual or cultural factors. It is hypothesized that the later in the development cycle of a project the more likely an individual to go to upper management to discuss project status. Part of this may be due to the fact that the closer the project is to completion and the deadline, the less likely workers are to hold to a biased belief that there is still plenty of time for a problem to work itself out. According to Keil et al. (2014) and Cueller et al. (2006), many IT projects exhibit warning signs in advance of problems. However, addressing the warning signs are typically ignored when the problems are still in the preventative state. This phenomenon seems to indicate that it is not until a later project phase that information gets communicated and addressed. Reasons for such delay may be due to a work environment that is not conducive to bringing forth negative information when a project is still in its early phases.

*H7: The later the phase in the project development cycle (**PDP**) the more likely workers will be inclined to report project information (**IRPI**).*

The dependent variable (IRPI) is measured generically as an individual's inclination to report and discuss project status related information with upper management. It does not differentiate between positive and negative status information. Therefore, negative information (NI) is included partially as a control variable. According to the "mum" effect, when individuals are faced with bad news they will likely choose to remain silent and not communicate the negative message (Lee, 1993). In addition, whistle-blowing theory, which was discussed earlier, also supports reluctance to report negative information. Therefore, it is believed that if the project status information is negative there will be less of an inclination to report the information to upper management.

*H8: Negative project status information (**NI**) will have a negative effect on **IRPI**.*

If there is a threat of negative consequences for relaying information to upper management then a worker may be less inclined to communicate. Even if the message to be relayed is not negative, there may be an organizational climate where higher levels of management have closed doors, there may be a fear of wasting a manager's time, or there may be perceptions that communication would not be welcomed at

higher levels of management. If the message is negative and the organizational climate is one that may "shoot the messenger" of bad news then workers may be inclined to remain silent for fear of retaliation from management or even from colleagues (Keil et al, 2004, Mesmer-Magnus & Viswesvaran, 2005).

*H9: Perceptions of negative consequences (**NC**) for sharing project information will have a negative effect on **IRPI**.*

Gender Differences

Based on Cuellar et al. (2006) gender differences were found when participants in a research experiment were faced with making a decision to de-escalate a project. Results revealed that women were more likely to delay projects in the face of negative information than men. An explanation offered for why women would be more inclined to delay projects is that women may be less likely to be sensitive to personal negative consequences if it means preventing negative impacts on the organization. Because gender differences have been identified previously in the context of project management, it is believed that there will also be gender differences in the relationships between the individual and work climate factors and IRPI. It is hypothesized that the relationship between negative consequences and IRPI will become non-significant when the model is tested for women. If women are less likely to be concerned about personal negative consequences than men then they will be less likely influenced by the existence of negative consequences when faced with the decision to speak to upper management about a project's status.

*H10: The relationships in the proposed model will be difference for men than for women. The relationship between **NC** and **IRPI** will be non-significant for the sample of women.*

Refer to Figure 1: Proposed Model located in the Appendices for the direct effects model showing the proposed relationships.

3. METHOD AND RESULTS

A sample survey was administered to individuals currently working on IT projects. Project team members are the individuals most closely involved with the project, influence information on status reports, and are more likely to be some of the first to know when projects are heading for trouble (Keil & Robey, 1999, Snow & Keil, 2002). A sample of 232 survey responses were gathered, ten responses were removed due to incomplete demographic information, leaving a usable sample of 222 responses. Survey questions are listed in Table 2: Survey Items in the Appendices. Multiple item constructs such as IRPI, responsibility, and optimism of project success were adapted from existing measures (Korzaan, 2009, Smith et al., 2001, Simon & Houghton, 2003, Schoorman & Holahan, 1996). Three models were tested as simple direct effects regression models. The first model was tested for the full sample of 222 participants, the second model was run for the sample of 87 female participants, and the third model was run for the sample of 135 male participants. The final results models for all three of these scenarios are shown in the Appendices in Figure 2: Final Model-All Data, Figure 3: Final Model–Females, and Figure 4: Final Model–Males.

The results from running the model with all data reveal that the only hypotheses supported were H5 (RES→IRPI), H6 (OPS->IRPIR), H7 (PDP-IRPI), and H9 (NC->IRPI). The more one believes that they are responsible for the project, the more optimistic one is about the overall success of the project, and the later the project development phase then the more likely one is to go to upper management and discuss project status information. The stronger the perception of experiencing negative consequences for discussing project status information then the less likely an individual will be to discuss that information with upper management. The amount of variance explained in the dependent variable was 28%. None of the demographic information (age, number of years IT experience, number of years' experience at the organization, and education) was significant in predicting IRPI.

When the direct effects model was run for females and then for males H10 was found to be partially supported. There are some differences in the models between the sample of males and females. However, although it was hypothesized that negative consequences would be significant for men and not significant for women, this hypothesis was not supported. Negative consequences were found to be significant for both men and women; however, the effect for men was β=-.35 and the effect for women was β=-.2. So although negative consequences were significant for both genders, it is not as strong of an effect for women as it is for men. Other

relationships, which were not hypothesized to be significantly different between the genders, were found to be significantly different. For men, a sense of responsibility for the project and optimism of project success were significant positive predictors of IRPI. However, for women, neither factor was significant. For women, the later the development phase the more likely they were to discuss project information; however, for men the development phase was not significant. The model explained 40% of the variance in the dependent variable for men and 22% of the variance in the dependent variable for women.

4. DISCUSSION

This study contributes to IT project status reporting literature by identifying individual and work climate variables that predict when individuals are more willing to discuss project related information with members of upper management. This extends current knowledge in the research stream of IT project management and project status reporting. Another significant contribution to research is the demonstration of the differences in the influential factors in predicting willingness to discuss project information with upper management for men and women. Because project development phase was significant in predicting IRPI in both the model with all data and the model for women, it is recommended that mechanisms be implemented in the development life cycle to promote project status communication early in the development life cycle when there is still enough time to prevent potential problems and address trouble areas in the project before they spiral out of control.

The study highlights the potential importance for upper management to pay close attention to project team composition. According to Keil et al. (2014), team composition is a key factor in accommodating cultural differences. The findings of this study support the concept as well as recognize the implications related to gender within the team environment

A consistent and strong negative predictor of IRPI is negative consequences for reporting status information. This work climate variable is something that upper management has control over and it is recommended that management implement policies and promote a culture that encourages open communication about project status. It is also recommended that they guard against any potential of backlash to an individual reporting negative information. Instead, an open door policy that fosters open communication is encouraged.

5. LIMITATIONS AND FUTURE RESEARCH

This study was an initial and exploratory endeavor in identifying key individual and work climate factors that influence individuals' willingness to report project status information to upper management. Future research is needed to confirm this study's findings and to enhance the rigor of the research method. For example, some measures are one item constructs and further development and validation of measurement items and constructs is needed in future studies. Furthermore, future research is needed to help understand additional gender differences and how these differences may be balanced in project team composition. It is important to also consider the gender of individuals in the role of upper management and how that might impact the likelihood of communicating project status. Finally, there is a call for future research to investigate cultural factors that influence individuals' inclinations to report project status information. Implications of this could also impact the content of project management training and education.

6. CONCLUSIONS

Although many members of senior management believe that employees will communicate when problems arise with IT projects, the reality is that most will not speak up or if they do will bias the information in a positive direction (Keil et al., 2014). This research has helped address the silent treatment from IT project workers by identifying key factors that help predict when individuals will be more likely to discuss project information with upper management. For management this means to promote a corporate culture that does not "shoot the messenger" for bearing bad news but instead shields employees from potential negative consequences for communicating project information. It is also important for project teams to be comprised of a balanced representation of males and females. However, statistics still show a shortage of women in the technology workforce with only 20% computer jobs and 7% CIO positions held by women (Fisher, 2013). Perhaps it is time for the field of IT project management to open more discussion on promoting and supporting women

in the IT project management profession as well as build more awareness of gender issues.

7. REFERENCES

Cuellar, M., Keil, M, & Johnson, R. (2006). The deaf effect response to bad news reporting in Information Systems projects. E-Service Journal, 5(1), 75-97.

Fisher, A. (March 11, 2013). Why are there still so few women in science and tech. *Fortune,* Retrieved July 5, 2014 from http://fortune.com/2013/03/11/why-are-there-still-so-few-women-in-science-and-tech/

Keil, M. & Robey, D. (1999). Turning around troubled software projects: An exploratory study of the de-escalation of commitment to failing courses of action. *Journal of Management Information Systems*, 15(4), 63-88.

Keil, M., Smith, J., Iacovou, C., & Thompson, R. (2014). The pitfalls of project status reporting. *MIT Sloan Management Review*, 55(3), 56-65.

Keil, M., Smith, J., Pawlowski, S., & Jin, L. (2004). 'Why didn't somebody tell me?' Climate, information asymmetry, and bad news about troubled projects. *The DATA BASE for Advances in Information Systems*, 35(2), 65-84.

Lee, F. (1993). Being polite and keeping MUM: How bad news is communicated in organizational hierarchies. *Journal of Applied Social Psychology*, 23(14), 1124-1149.

Mesmer-Magnus, J. & Viswesvaran, C. (2005). Whistleblowing in organizations: An examination of correlates of whistleblowing intentions, actions, and retaliation. *Journal of Business Ethics*, 62, 277-297.

Korzaan, M. (2009). The influence of commitment to project objectives in Information Technology (IT) projects. *Review of Business Information Systems*, 13(4), 89-98.

Schoorman, F. & Holahan, P. (1996). Psychological antecedents of escalation behavior: Effects of choice, responsibility, and decision consequences. *Journal of Applied Psychology*, 81(6), 786-795.

Simon, M. & Houghton, S. (2003). The relationship between overconfidence and the introduction of risky products: Evidence from a field study. *Academy of Management Journal*, 46(2), 139-149.

Smith, J., Keil, M, & Depledge, G. (2001). Keeping mum as the project goes under: Toward an explanatory model. *Journal of Management Information systems*, 18(2), 189-227.

Snow, A. & Keil, M. (2002). The challenge of accurate software project status reporting: A two-stage model incorporating status error and reporting bias. IEEE Transactions on Engineering Management, 49(4), 491-505.

Appendices

Construct	Construct Description
IRPI	Inclination to report project information with upper management
YO	Number of years at organization
YIT	Number of years in IT
Age	Age
EDU	Education
RES	Sense of responsibility for the project
OPS	Optimism of project success
PDP	Project development phase
NI	Negative information about project status
NC	Negative consequences for reporting project status information

Table 1: Construct Abbreviations

Variable	Survey Questions
IRPI Inclination to report project information	How likely would you be to go directly to upper management by yourself to discuss the status of this project? How likely would you be to try and persuade members of the development team to go to upper management and discuss as a group the status of this project?
NI Negative information	There are many challenges that must be overcome before this project can succeed. This project will need to overcome several obstacles.
RES Responsibility for project	The project's performance is a reflection of me personally I am responsible for the project's outcome I am accountable for the project's success
OPS Optimism of project success	I am completely sure the project will finish successfully. I am absolutely positive that this project will be a success.
NC Negative consequences	If you went directly to upper management by yourself and discussed the status of this project, how likely is it that you would suffer negative consequences?
YO Number of years at organization	How many years have you been employed at your current organization?
YIT Number of years in IT	How many years of experience do you have with IT development projects?
Age	_____ 20-29 years _____ 30-39 years _____ 40-49 years _____ 50-59 years _____ 60 or more
PDP Project Development Phase	_____ Analysis _____ Design _____ Programming _____ Testing
EDU Education	_____ High school _____ Some college _____ Associates Degree _____ High school _____ High school _____ High school

Table 2: Survey Items

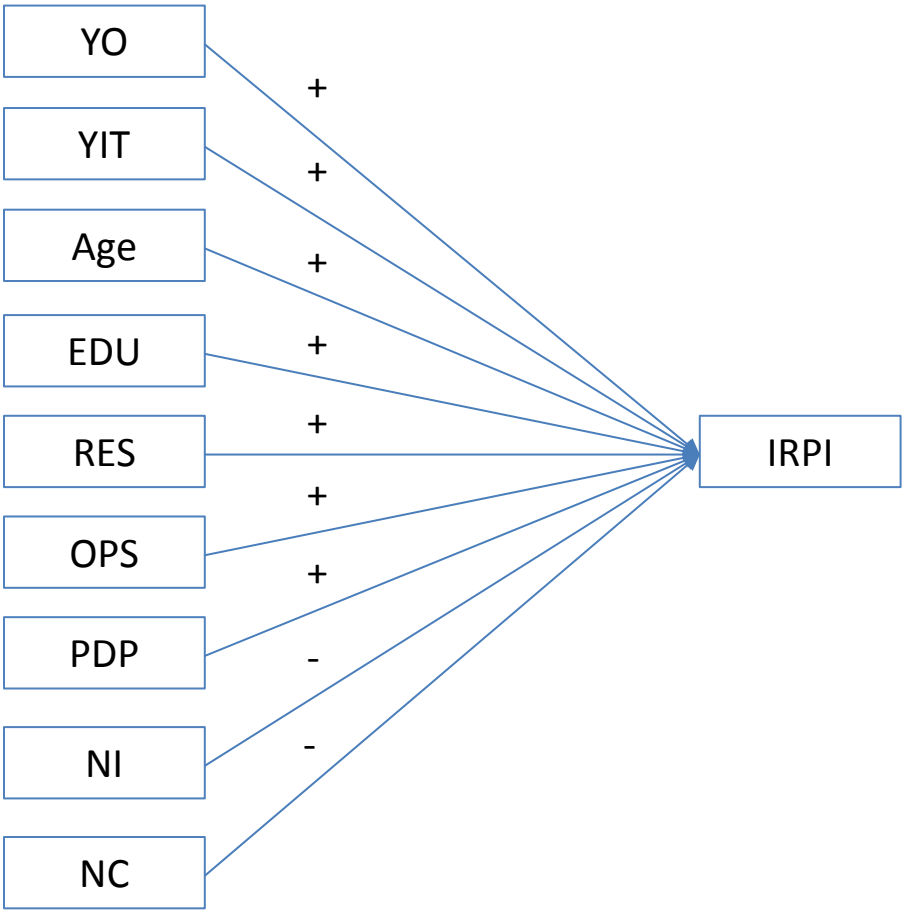

Figure 1: Proposed Model

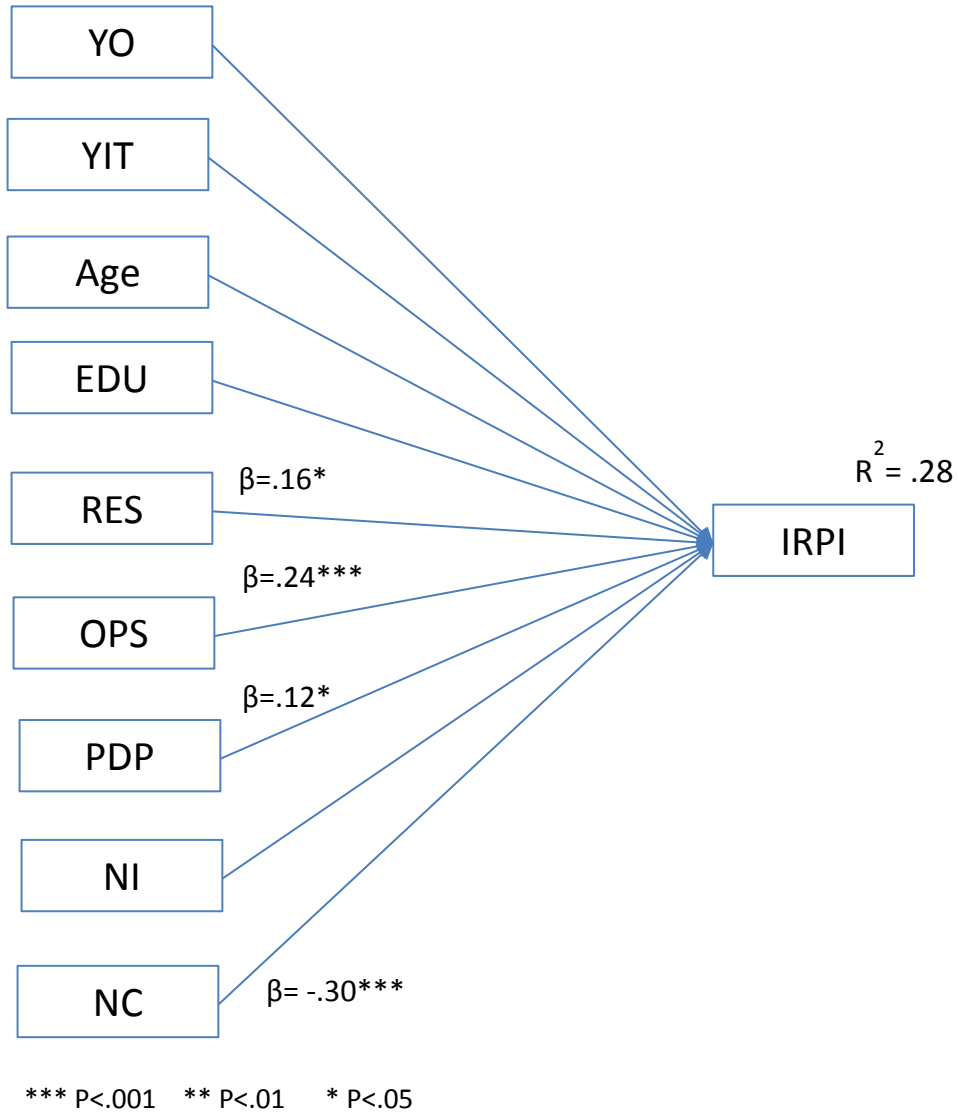

Figure 2: Final Model – All Data

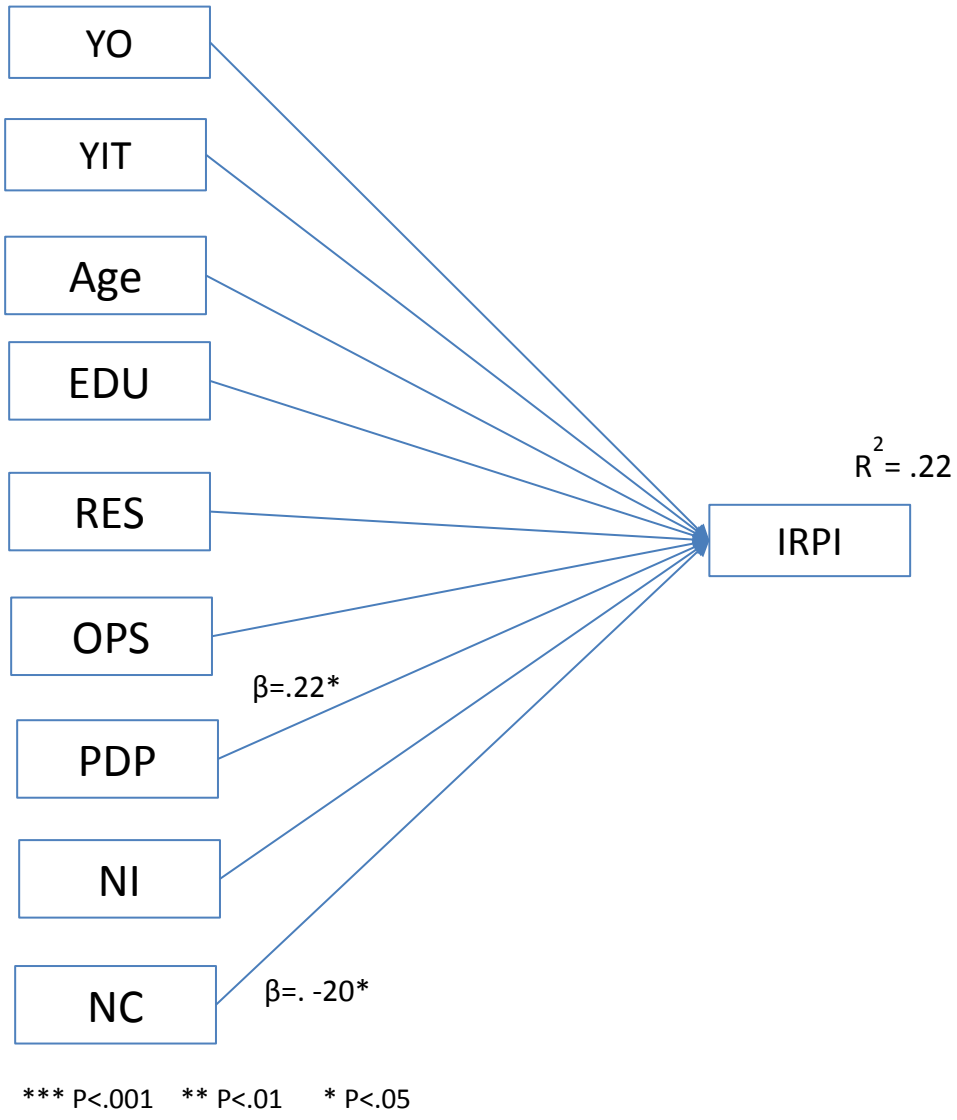

Figure 3: Final Model – Females

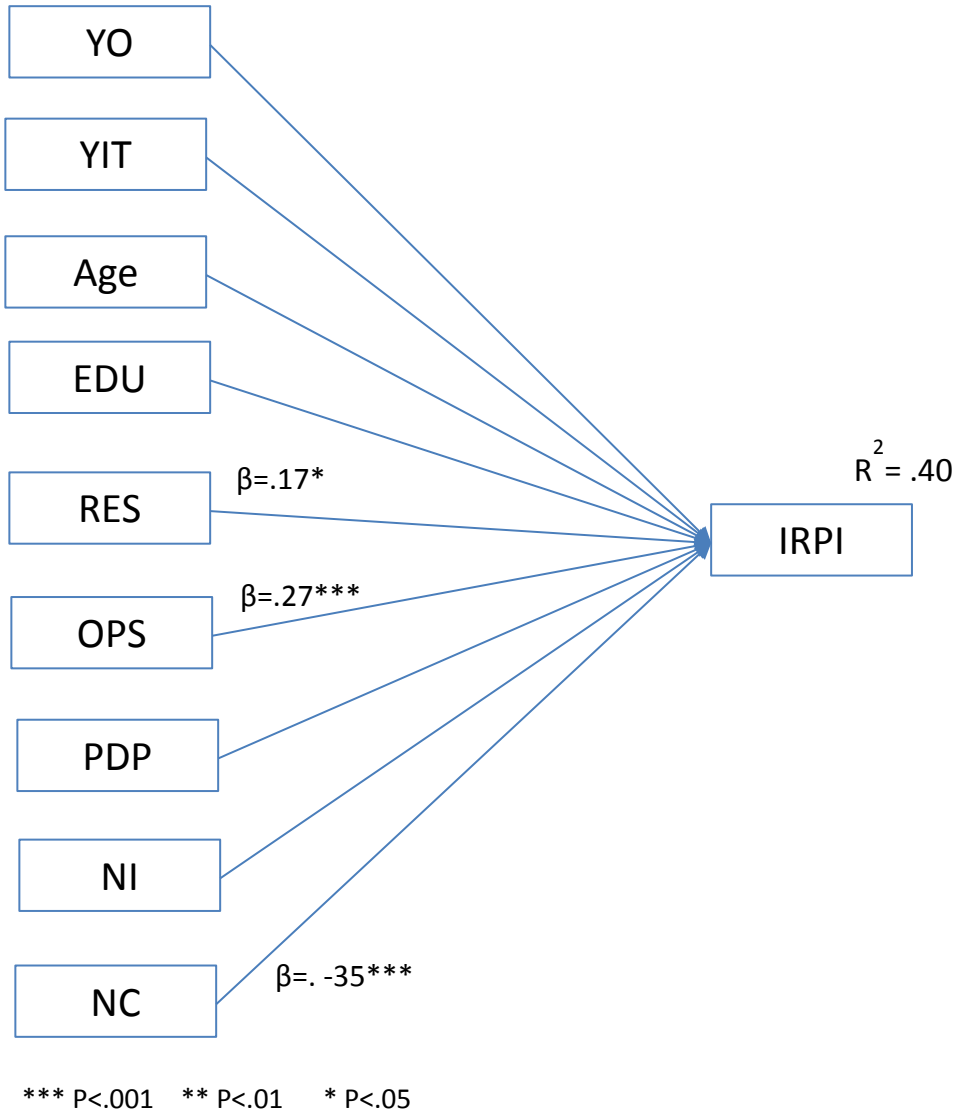

Figure 4: Final Model – Males

Building a Better Stockbroker: Managing Big (Financial) Data by Constructing an Ontology-Based Framework

Logan Westrick
westricl@mail.gvsu.edu
Epic
Verona, Wisconsin 53593 USA

Jie Du
dujie@gvsu.edu

Greg Wolffe
wolffe@gvsu.edu
School of Computing & Information Systems
Grand Valley State University
Allendale, MI 49401 USA

Abstract

Financial investment decision making is a complex process, in which decision makers utilize specific techniques to analyze a large volume of noisy time-series data in order to arrive at a final decision. Collecting and managing the enormous amount of available financial data is an important task in this process, for both researchers and end-user investors. This paper proposes an ontology-based framework for effectively managing big financial data. It further describes the steps required to implement such a framework, and reports the results of a feasibility study into implementing the proposed framework. A Financial Statement Ontology (FSO) is created using the Web Ontology Language (OWL) in the Protégé knowledge framework together with a data acquisition driver written in Perl. The use of an ontology adds a layer of abstraction to Big Data, alleviating the need for end-users to concern themselves with added complexity. The framework thus allows researchers and investors to spend more time on problem-solving and less time managing Big Data. In addition to the described application to finance, the proposed framework has the potential to be applied to any other domain in which relevant data is distributed across multiple systems or is accessed using different formats or names, such as is common in medical research.

Keywords: big data, ontology, financial decision support systems, knowledge base

1. INTRODUCTION

The "3Vs" model defines Big Data as "high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization" (Beyer, 2014). Researchers have proposed various methods for addressing the difficulties caused by the unique properties of Big Data. Examples include using visual analysis tools (Jafar, Babb, & Dana, 2014) and using ontologies (Buitelaara, Cimianob, Frankc, Hartungc, & Racioppa, 2008).

The explosive growth of Big Data has led to problems in managing data in such a way that it remains easily accessible to users (McAfee & Brynjolfsson, 2012). One of the disciplines in which this is most evident is the area of finance.

There is a wealth of financial information available on today's Internet. For example, Google, Yahoo, and MSN provide extensive financial data including financial statements, information on domestic and international financial markets, and business news relating to companies. Investors increasingly rely on these data to inform their financial investing decisions, such as predicting stock behavior (Deller, Stubenrath, & Weber, 1999). Given the rapid growth of accessible real-time financial data, the problem of effectively collecting and managing this data has become a challenging task. Given that context, this paper strives to answer the following research question:

How should one effectively manage big financial data so as to better make an informed financial investing decision?

Recently, Du and Zhou (2012) proposed the use of an ontology-based framework to address the problem of normalizing financial data obtained from multiple sources. They defined several data oriented concerns and then presented an ontology mapping mechanism designed to mitigate these problems. Building on their work, this paper describes the process and conducts a feasibility study into the creation of an ontology-based knowledge base for financial data. The goal is to allow investors to easily compare and use financial information obtained from heterogeneous online sources, and thus make intelligent and well-informed analytical decisions.

To enable this, a Financial Statement Ontology (FSO) is created using the latest version of the Web Ontology Language (OWL 2) in the Protégé knowledge framework. All key attributes found in balance sheets, income statements, and cash flows are captured in the FSO. Then, a Perl-based data acquisition driver is used to seamlessly access online sources. It combines the online data with the base ontology to produce an ontology containing individual entries.

As a case study, the practical implications of our findings show the promise of applying the proposed framework and knowledge base, especially to other domains.

The remainder of this paper is organized as follows: Section 2 provides relevant background information on ontologies, the semantic web, and financial decision making. Section 3 presents specifics of the proposed ontology-based framework, while Section 4 documents the implementation details. Section 5 describes the use of the proposed ontology-based framework and provides a sample workflow. Section 6 discusses our findings as to the maturity of the technology, lessons learned, and potential pitfalls. Section 7 concludes the paper.

2. BACKGROUND

An ontology is a formal specification of a set of concepts, agents, and relationships. Ontologies find their basis in the Semantic Web. The term *Semantic Web* was coined by Tim Berners-Lee and is defined as "a web of data that can be processed directly and indirectly by machines" (Berners-Lee, Hendler, & Lassila, 2001). The foremost purpose of a semantic web, or net, is to encapsulate knowledge and its representation so that machines can "understand" and respond to complex human requests. The Semantic Web movement is led by the World Wide Web Consortium (W3C) with the goal of embedding semantic data (data about what things mean; as opposed to syntax, which is simply their structure) into the current unstructured web to create a network of data that can be navigated by machines. This would enable intelligent agents to conduct the tedious work of finding and processing data, freeing humans to do more important (or at least less menial) tasks.

One of the central components of the Semantic Web is the ontology. Ontology was introduced by Gruber as meaning an explicit specification of conceptualization (Gruber, 1993). An ontology as the term is used in computer science is essentially an extension of the time-tested relational database to include semantic data in addition to syntactic data.

The standard language for ontologies in the Semantic Web is the Web Ontology Language (OWL); version 2 is the current standard (OWl, 2014). The OWL specification includes several variants that differ in their expressiveness. Some of these sublanguages, or profiles, allow for faster computer reasoning by restricting the set of allowed statements. OWL 2 DL (Direct Logic) is the most permissive subset that remains computable (use of the abbreviation OWL will refer to OWL 2 DL unless otherwise stated). The central idea of OWL is that any relational database can be simplified to three

fields; subject, predicate, and object. A single entry is thus known as a relation. Other names for an entry include axioms (as they are the data assumed to be true by a computer reasoner) and triples. By specifying a set of default verbs and providing the ability to define more verbs within the ontology, OWL allows for embedding a knowledge base's semantics within itself. This gives a computer program the ability to read the ontology and reason with it without prior knowledge of its structure.

This constitutes a substantial improvement over current development processes, in which both the structure and semantics of data must be explicitly programmed into a program before meaningful work can be accomplished. As an additional benefit, an ontology that follows the proper OWL restrictions can be reasoned on by a computer reasoner, which can make deductive inferences based on the already-stated relations in the knowledge base. In theory, this can allow for faster and more precise development, since programmers no longer need to specify every single relation.

Similar to a database, an OWL ontology can be accessed by using a query language known as the SPARQL Protocol and RDF Query Language (SPARQL) (W3C, 2014b), a variant of Structured Query Language (SQL) that is customized for use with triple stores and ontologies.

Ontology plays a key role in the field of Information Systems, such as improving information consistency and reusability, systems interoperability and knowledge sharing (Kishore, Ramesh, & Sharman, 2007). The crucial research issues surrounding ontology focus on two aspects: ontology generation and ontology mapping (Ding & Foo, 2002a, 2002b). An ontology is generated to provide a shared framework of the common understanding of a specific domain. The creation process can be manual, semi-automated, or fully automated. Ontology mapping expands and combines existing ontologies to support communication and interaction between existing and new domains.

Ontologies have previously been applied to the field of finance (Chenga, Lub, & Sheu, 2009; Fensel & Brodie, 2003). For example, Wand & Wang (1996) focused on improving data quality, and Du and Zhou (2012) endeavored to mitigate data quality problems. When the data quality has been compromised, their framework is used to fix the data quality problem by retrieving the correct data from another data source. We extend their work and propose our own ontology-based framework which provides a knowledge base for end users, transparently allowing them to access multiple data sources as deemed necessary.

3. FRAMEWORK

This paper proposes an ontology-based framework for effectively managing Big Data. The basic structure of the framework is illustrated in Figure 1 (see Appendix). There are three key components to the framework:

- a base ontology,
- online data sources,
- a data acquisition driver.

The first component is the base ontology, called the FSO, in which the key financial concepts from balance sheets, income statements, and cash flows are defined. Their relationships are also captured in the FSO.

The second component is the abundant financial data available on the Internet. At the time this project was implemented, Google Finance, Yahoo!Finance, and MSN Money Central each provided free financial data for investors. Typically each of these data sources represents the data using their own unique knowledge representation structure.

The final component in the framework is the data acquisition driver. The role of the driver is to access online sources and to combine their data with the base ontology to create an ontology containing individual entries.

It is important to note the separation between information supplied by the base ontology and information added by the driver. In this step, different names for the same type of object are stored as labels to help with the ontology mapping of the framework.

For illustration purposes, and to graphically convey the complexity of the interconnections, the expanded diagram of an FSO populated with a single set of statements from a single company (Google) is given in Figure 2. The ontology used for framework testing and debugging has roughly 10x this amount of data; typical production ontologies could easily have 1000x as much data. This graphically illustrates the potential benefit of using an ontology to help manage Big Data. Abstraction and automation can relieve end-users of the burden of dealing with this level of detail and complexity.

4. IMPLEMENTATION

In order to implement an ontology-based framework, a number of important decisions need to be made:

- Decide whether to use a development environment and choose an ontology syntax.
- Decide which OWL reasoner to incorporate into the framework.
- Determine how to handle inconsistent or incomplete data.
- Decide how to store and access the ontology from within the driver.
- Determine driver language and database access protocol.
- Decide how best to access the driver-generated ontology from the user's perspective.

Development Environment / Syntax

Two choices for the development environment are the NeOn toolkit and the Protégé ontology editor. The NeOn Toolkit is the ontology engineering environment developed as part of the NeOn Project (NeOn, 2006). The toolkit is based on the well-known Eclipse platform and provides an extensive set of plug-ins. While the NeOn toolkit has a sleek and intuitive UI, it lacks support for some of the most recent features of OWL, including the ability to specify keys. The Protégé ontology editor is a free, open source ontology editor. It is referred to as "the leading ontological engineering tool" (Gašević, Djurić, & Devedžić, 2009). It completely supports all OWL features and allows for saving an ontology in all of the various OWL syntaxes, and includes several convenient visualization tools. For most applications, the ease of adding classes and the built-in visualization tools also make Protégé a far better solution than writing OWL by hand. Therefore, the decision was made to use Protégé for ontology development.

OWL supports several different syntaxes for creating ontologies. For this project, Functional syntax was chosen because of its conciseness and relative ease of specification (OWL's own formal specification uses the same syntax). Below is a sample giving the flavor of the Functional syntax:

```
ClassAssertion (:Company :GOOG)
DataPropertyAssertion (:hasName
    :GOOG "Google"^^xsd:string)
ObjectPropertyAssertion (:hasEntry
    :GOOG
    :GOOG_NonRecurringOpEx_2013-12-
    31_2014-04-02_03-46-51Z)
```

Other common syntaxes include the RDF/XML syntax (W3C, 2014a), which is very verbose but is the most widely supported OWL syntax, and the Manchester syntax (W3C, 2008), which is specifically designed to be easily readable by non-logicians. Less commonly used are the Turtle (W3C, 2012) and OWL/XML syntaxes (W3C, 2013).

Reasoner Selection

One of the benefits of using an ontology is to take advantage of computer reasoning. For example, a reasoner can exploit the knowledge embedded in a transitive relationship without direct user coding or intervention.

At this time, the choice of reasoner is straightforward: the HermiT reasoner is the only one that fully supports the newest specification of OWL (OWL 2), and it is also the fastest of the free reasoners (KRR, 2014). HermiT is written in Java; it can work with other languages (e.g. Perl), but naturally works best with Java. It can be imported directly in Java, it can be accessed via the OWL API, or it can be run from the command line. The Protégé development environment supports the use of different reasoners via a plugin system, so incorporating the HermiT reasoner was straightforward.

Inconsistency Handling

The most difficult part of this project was determining how to handle the occurrence of inconsistent data. An OWL reasoner can easily determine whether or not an ontology is consistent. It can make inferences, provided the ontology is consistent. But at this time, reasoning with an inconsistent ontology seems to be impossible without writing a custom reasoner (Rosati, Ruzzi, Graziosi, & Masotti, 2012) or requiring direct human intervention.

A further question arises when dealing with financial applications: whether or not it is even desirable to repair inconsistent data? After all, inconsistent data might be a sign of fraud.

Given the nature of this study, it made sense to simply leave inconsistent data as is. As long as the ontology contains data about a subject from at least one source, a properly formed query will return that result transparently. If data is inconsistent, the same query will return all of the multiple results. It is left to the user to determine the significance of the inconsistency and how best to handle it.

Another common problem found in the online financial data sources is terminological

ambiguity (Du & Zhou, 2012). Specifically, different financial sources use different names for the same data. One possible solution to this general problem is to create and use a thesaurus to handle the terminological ambiguity. For example, Mannette-Wright (2009) developed a shared knowledge environment for automating the document transformation problem in the medical field, in which the HL7 industry standard thesaurus is used to resolve terminological ambiguity. In the finance field, a recent study combines Semantic Web technologies and linked data principles to increase interoperability and comparability of business reports represented with XBRL markup (O'Riain, Curry, & Harth, 2012). XBRL US GAAP Taxonomies v1.0 defines concepts and their relationships in financial statements and can be used as a thesaurus to resolve discrepancies among reports that are prepared by accountants who are using different accounting principles. In our study, the terminology ambiguity problem is solved by adding label annotations to the various entry classes in the ontology based on the XBRL US GAAP Taxonomies. For example, by creating an appropriate label in the FSO, the reasoner can infer that the concept of "Cost of Revenue, Total" found in Google Finance is equivalent to the concept of "Cost of Revenue" as given in Yahoo!Finance.

A diagram of the final implementation of the ontology can be seen in Figure 3. The "base" ontology, specified by the user, is given in blue. The remainder of the ontology, filled in by the accompanying driver and inferred by the included reasoner, is in red. The ontology contains several types of entities:

- Class: a set or category, represented by a taller rectangle (e.g. "`GrossProfit`")
- Individual: a member of a class, represented by a shorter rectangle (e.g. "`:GOOG`")
- Relation: a connection between concepts or objects, represented by a rounded box (e.g. "`hasVal`")
- Literal: a fixed value, represented by an ellipse (e.g. "'`2013-12-31`"^^`xsd:date`"')

Ontology Storage/Access

Since it captures concepts and relationships, an ontology is meant to be persistent, implying the need to store the ontology for future access. In a full production environment, as might be found in a brokerage, the expectation would be substantial redundancy and security. It would employ triple-store servers and authentication tools. As part of a research project, this prototype simply leaves the base ontology as a file in OWL functional syntax. The file can either be explicitly supplied to the driver or the driver can be configured to access the default base ontology online.

Driver Design

Most OWL-related software is written in Java, including the OWL application programming interface (API). However, for this project, Perl was chosen as the language for driver implementation. This was due partly to local expertise with the language but primarily because of Perl's advanced capabilities for text-processing. APIs send all data as text, requiring it to be parsed; Perl was well suited to that task.

User Interface

Many commercial systems include an API for a SPARQL engine. The API contains hooks which allow a developer to easily build a GUI on top of it, abstracting the actual SPARQL code away from the user. In other words, the user constructs a SPARQL query by selecting options from drop-down lists; essentially interacting with the system by filling out forms.

Due to time and budget constraints, the prototype described here uses the free SPARQL engine built into Protégé. It is basically a command-line interface. SPARQL is the SQL-based language used to construct queries to the Resource Description Framework (RDF), the standard model for Web-based data interchange. Figure 4 illustrates the command-line interface included with Protégé. A sample SPARQL query shows the SQL-like nature of the language. The figure also shows the outcome of executing the query – multiple results are matched and returned.

5. USAGE

A diagram of a typical workflow can be seen in Figure 5. Rectangles are user-directed actions, ellipses are automated framework activities. The diagram is also color-coded to match the ontology structure diagram in Figure 3, showing interactions with the base ontology and the fully-populated ontology.

The first step is initiated by the user; simply run the driver. The driver expects:

- a list of names and ticker symbols of companies whose data should be retrieved,
- the location of the base ontology ("`def`" tells the driver to access and use the default base ontology stored online),

- the name of the file to write/store the completed ontology to,
- whether to access quarterly or annual reports, and
- at least one group number.

The syntax is therefore "driver.pl tickerList {localFile | def} outfile {annual|quarterly} $group_1$ ($group_2$... $group_n$)". This command will automatically generate an ontology populated with all of the available online data for the companies specified in the ticker list. The data will be normalized as per the specifications in the ontology, and the final ontology will be written to backing store.

In the next major step, the user opens the populated ontology in the Protégé knowledge editor.

Protégé includes a console-based SPARQL engine. The user can use this command-line interface to query the ontology, similar to querying a database. The reasoner in the framework helps make inferences regarding the relationships in the ontology; e.g. it can reconcile different names for the same entity.

Budget and time permitting, future enhancements might involve developing GUI-based interaction with the SPARQL engine.

6. DISCUSSION / LESSONS LEARNED

The proposed ontology-based framework goes well beyond traditional databases. It:

- Encodes *semantics* in addition to *syntax.*
- Facilitates knowledge pooling and inter-departmental communication.
- "Knows" that different terms mean the same thing, allowing different users to query the ontology using the terminology most familiar to them.
- Infers unstated relations that logically follow from stated ones, shortening development time and reducing error.
- Allows both researchers and business people to spend more time on the core problems in their fields and less time managing Big Data.

Experience with this project demonstrates that the ontology model is definitely ready for practical use. OWL supports either adding labels to classes in an ontology or adding multiple names to an individual as data properties. This means that a user can query a properly constructed ontology using the terminology that they are familiar with and receive data from all disciplines, even if other disciplines added their data to the ontology under a different name. The most practical immediate application for such a feature would be in constructing research databases.

For example, a genetics research database might have numerous papers and publications about a particular gene. The gene symbol used by the Mouse Genome database is *Pax6*, while the homolog in the Human Genome database would be called *PAX6*. The entry in the ontology could be given both names, and a scientist studying the gene in humans can then add a paper to the ontology, tagging it with the keyword "*PAX6*". If a mouse researcher later queried the ontology for all papers written about "*Pax6*", the ontology would also return papers concerned with that gene and written by human gene researchers.

This study did expose several limitations with the current implementations of the ontology model. Perhaps because of its origins in philosophy, logic, and linguistics, the area where the ontology framework still has the most maturing to do is its handling of numerical data. For this reason, a financial knowledge base is not really the best subject for a case study on ontologies. While this project was successful in creating a working ontology framework, there were several places where it became readily apparent that OWL was being used in ways it was not quite designed for. Most notably, there is no way to specify that the value of one entry should be the "sum of", or "difference of", or is otherwise numerically related to other entries. For instance, gross profit is defined as the difference between total revenue and cost of revenue in GAAP. But the current ontology cannot specify this kind of relationship between numerical data. To get around this limitation, the framework was configured with a property such that a user could still see these relationships, but a machine reasoner would have to be explicitly told what the properties meant before it could use them. This is perhaps the most important task at which a standard relational database is still better than an ontology.

Another potential pitfall to those considering a similar project in the area of finance is the lack of free financial APIs. In the recent past there were several free APIs including Yahoo! Finance, MSN Money, and Google Finance. At the time of this writing, only Yahoo! Finance remains free. MSN Money and Google Finance still exist as websites accessible to human users. However,

they no longer have APIs for developer use, and their current terms of service prohibit using an automated program to access their data. Therefore, to assess the ontology's performance at normalizing data from multiple sources, it was necessary to manually enter and use mock data for eventual population into the complete ontology. Although this method worked and demonstrated a proof-of-concept, it is obviously not the same as using multiple online sources.

7. CONCLUSION

This research project proposes an ontology-based framework to facilitate the management of Big (financial) Data. While still nascent, the ontology model shows great promise for managing enormous amounts of data in an efficient, transparent, and user-friendly way.

The main contribution of this research is to implement the proposed ontology framework and to assess its performance in a financial application. The implementation is described in detail, ranging from a discussion of available tools to consideration of important issues. Ironically, OWL's focus on object classification rather than numerical data means that at present a financial knowledge base is not really the best choice of subject for an ontology; our success in spite of that bodes well for the future of OWL and the Semantic Web in general.

8. REFERENCES

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web: Scientific American.

Beyer, M. (2014). Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data Retrieved June 6, 2014

Buitelaara, P., Cimianob, P., Frankc, A., Hartungc, M., & Racioppa, S. (2008). Ontology-based information extraction and integration from heterogeneous data sources. International Journal of Human-Computer Studies, 66(11), 759–788.

Chenga, H., Lub, Y.-C., & Sheu, C. (2009). An ontology-based business intelligence application in a financial knowledge management system. Expert Systems with Applications, 36(2), 3614–3622.

Deller, D., Stubenrath, M., & Weber, C. (1999). A survey on the use of the Internet for investor relations in the USA, the UK and Germany. *European Accounting Review,* 8(2), 351-364.

Ding, Y., & Foo, S. (2002a). Ontology research and development. Part 1 - a review of ontology generation. Journal of Information Science, 28(2), 123-136.

Ding, Y., & Foo, S. (2002b). Ontology research and development. Part 2 - a review of ontology mapping and evolving. Journal of Information Science, 28(5), 375-388.

Du, J., & Zhou, L. (2012). Improving the quality of financial data using ontologies, Decision Support Systems. Decision Support Systems, 54(1), 76-86.

Fensel, D., & Brodie, M. (2003). Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce (2 ed.). New York, NY: Springer.

Gašević, D., Djurić, D., & Devedžić, V. (2009). Model Driven Engineering and Ontology Development (2nd Edition): Springer.

Gruber, T. (1993). A transitional approach to portable ontology specifications. Knowledge Acquisition, 5(2), 39-41.

Jafar, M., Babb, J. S., & Dana, K. (2014). Decision-Making via Visual Analysis using the Natural Language Toolkit and R. Journal of Information Systems Applied Research, 7(1), 33-46.

Kishore, R., Ramesh, R., & Sharman, R. (Eds.). (2007). Ontologies: A Handbook of Principles, Concepts and Applications in Information Systems: Springer–Verlag.

KRR. (2014). HermiT OWL Reasoner Retrieved June 6, 2014, from http://www.hermit-reasoner.com/

Mannette-Wright, A. (2009). A Feasibility Study of Ontology-based Automatic Document Transformation: Pace University.

McAfee, A., & Brynjolfsson, E. (2012). Big Data: The Management Revolution. Harvard Business Review.

NeOn. (2006). NeOn Toolkit User's Guide Retrieved June 6, 2014, from http://www1.cs.unicam.it/insegnamenti/reti_2008/Readings/neon_users_guide.pdf

O'Riain, S., Curry, E., & Harth, A. (2012). XBRL and open data for global financial ecosystems: A linked data approach," International Journal of Accounting Information Systems. International Journal of Accounting Information Systems, 13(2), 141-162.

OWl. (2014). OWL 2 Web Ontology Language Document Overview (Second Edition) Retrieved April 4, 2014, from http://www.w3.org/TR/owl2-overview

Rosati, R., Ruzzi, M., Graziosi, M., & Masotti, G. (2012). Evaluation of Techniques for Inconsistency Handling in OWL 2 QL Ontologies. In P. Cudré-Mauroux, J. Heflin, E. Sirin, T. Tudorache, J. Euzenat, M. Hauswirth, J. Parreira, J. Hendler, G. Schreiber, A. Bernstein & E. Blomqvist (Eds.), The Semantic Web – ISWC 2012 (Vol. 7650, pp. 337-349): Springer Berlin Heidelberg.

W3C. (2008). OWL 2 Web Ontology Language: Manchester Syntax Retrieved September 7, 2014, from http://www.w3.org/TR/2008/WD-owl2-manchester-syntax-20081202/

W3C. (2012). Turtle - Terse RDF Triple Language Retrieved September 7, 2014, from http://www.w3.org/TR/2012/WD-turtle-20120710/

W3C. (2013). OWL Web Ontology Language: XML Presentation Syntax Retrieved September 7, 2014, from http://www.w3.org/TR/owl-xmlsyntax/

W3C. (2014a). RDF/XML Syntax Specification Retrieved September 7, 2014, from http://www.w3.org/TR/rdf-syntax-grammar/

W3C. (2014b). SPARQL Query Language for RDF Retrieved June 6, 2014, from http://www.w3.org/TR/rdf-sparql-query/

Appendix (Figures)

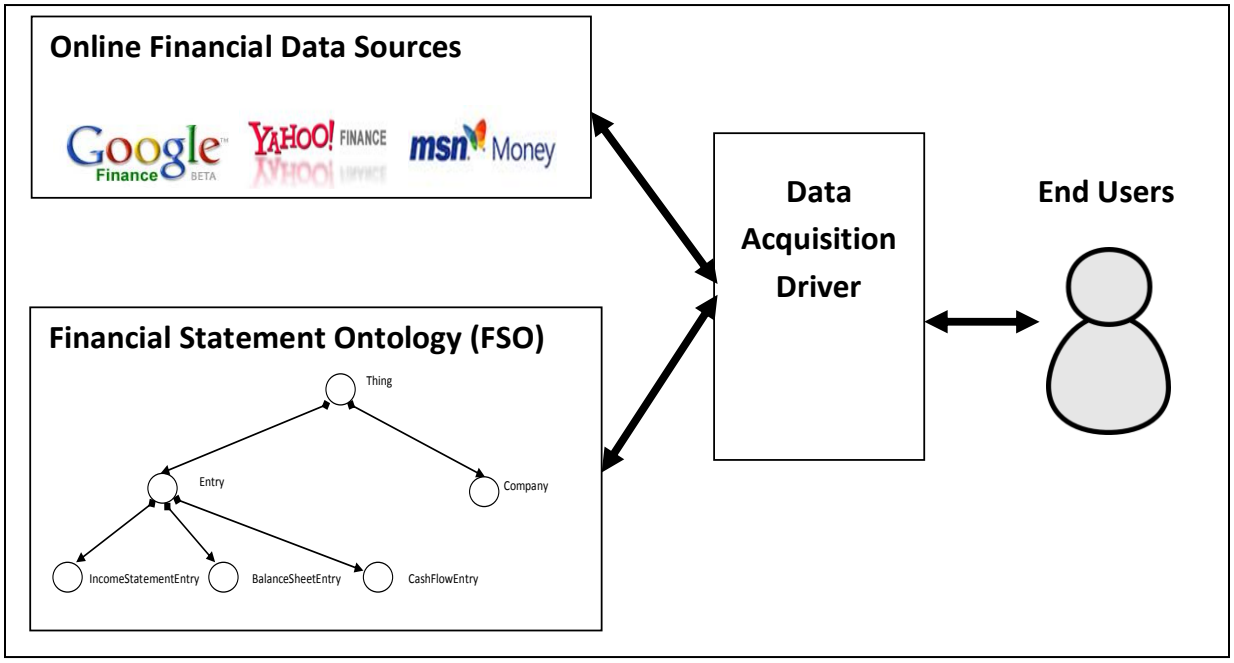

Figure 1: The proposed ontology-based framework for managing Big (financial) Data

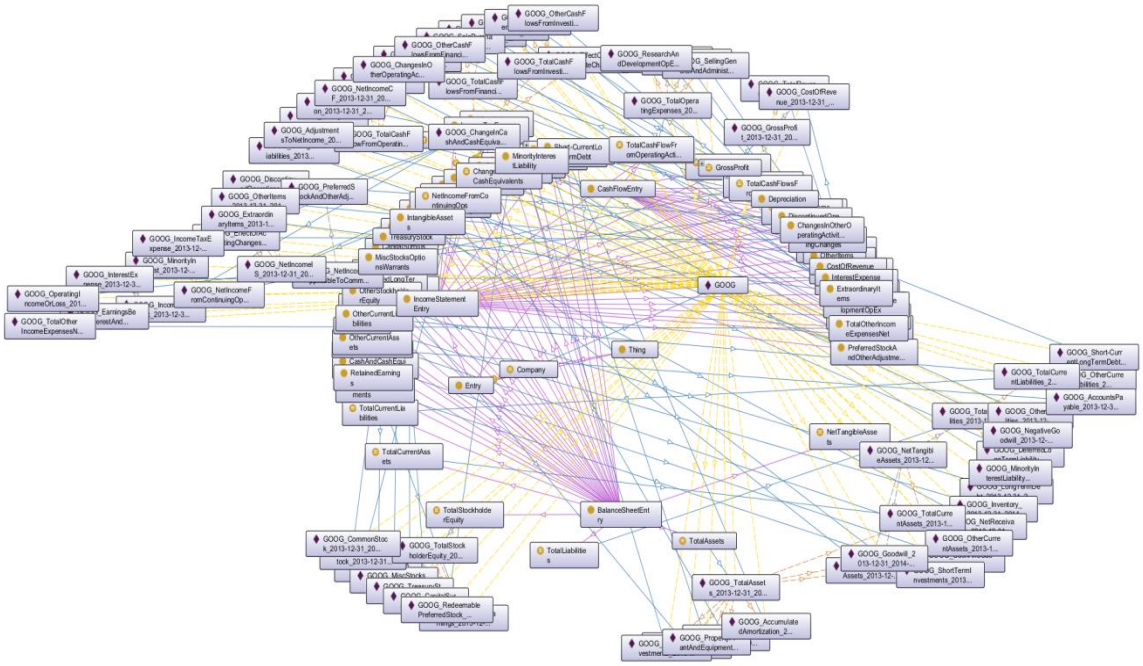

Figure 2: Fully-expanded diagram of the financial ontology populated with a single set of statements from a single company. Typical systems would be many orders of magnitude more complex.

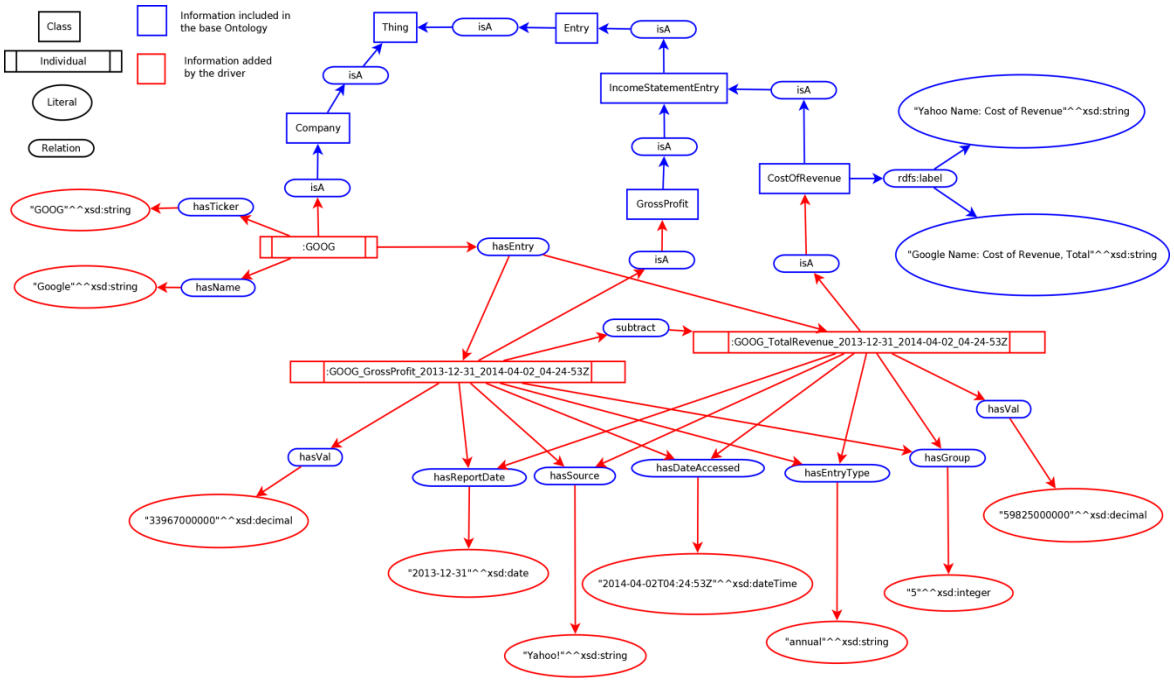

Figure 3: Diagram of the final design of the ontology. Blue represents the "base" ontology. Red entities are obtained via the driver and/or inferred by the reasoner.

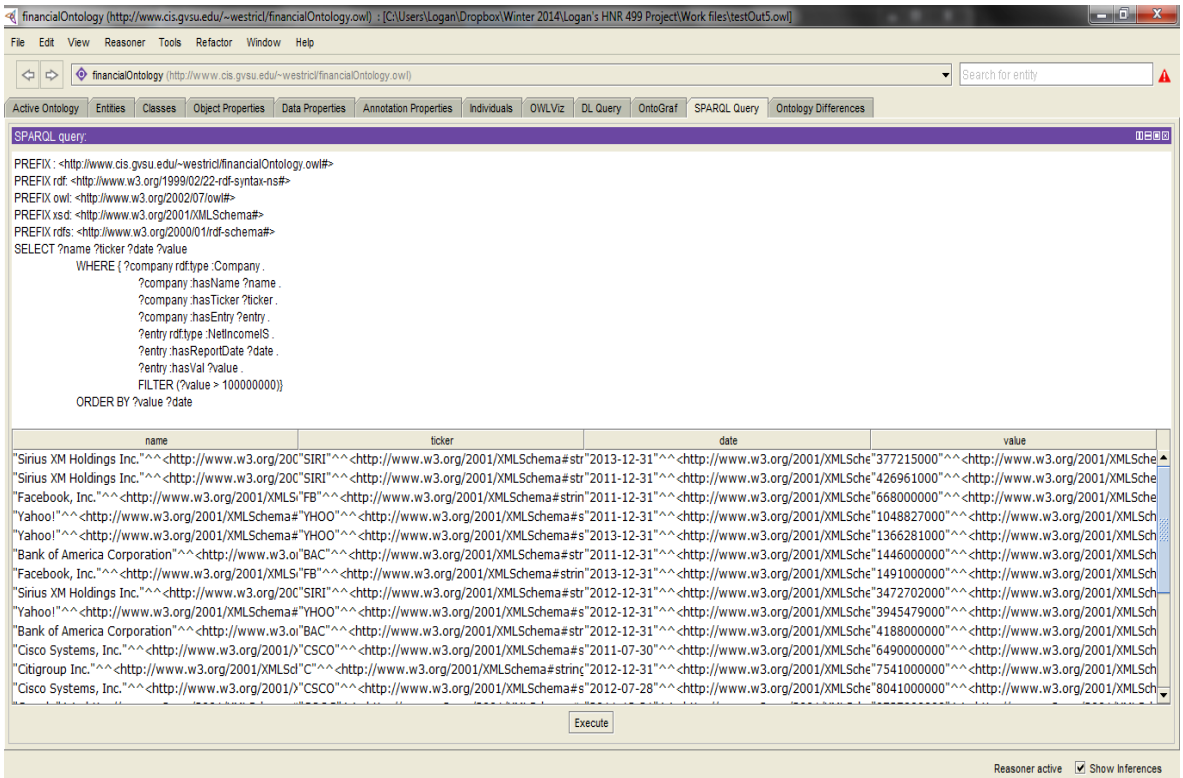

Figure 4: Sample SPARQL query. Pictured is the SPARQL console included in the *Protege* ontology editor; it uses an SQL-based language to query RDF graphs.

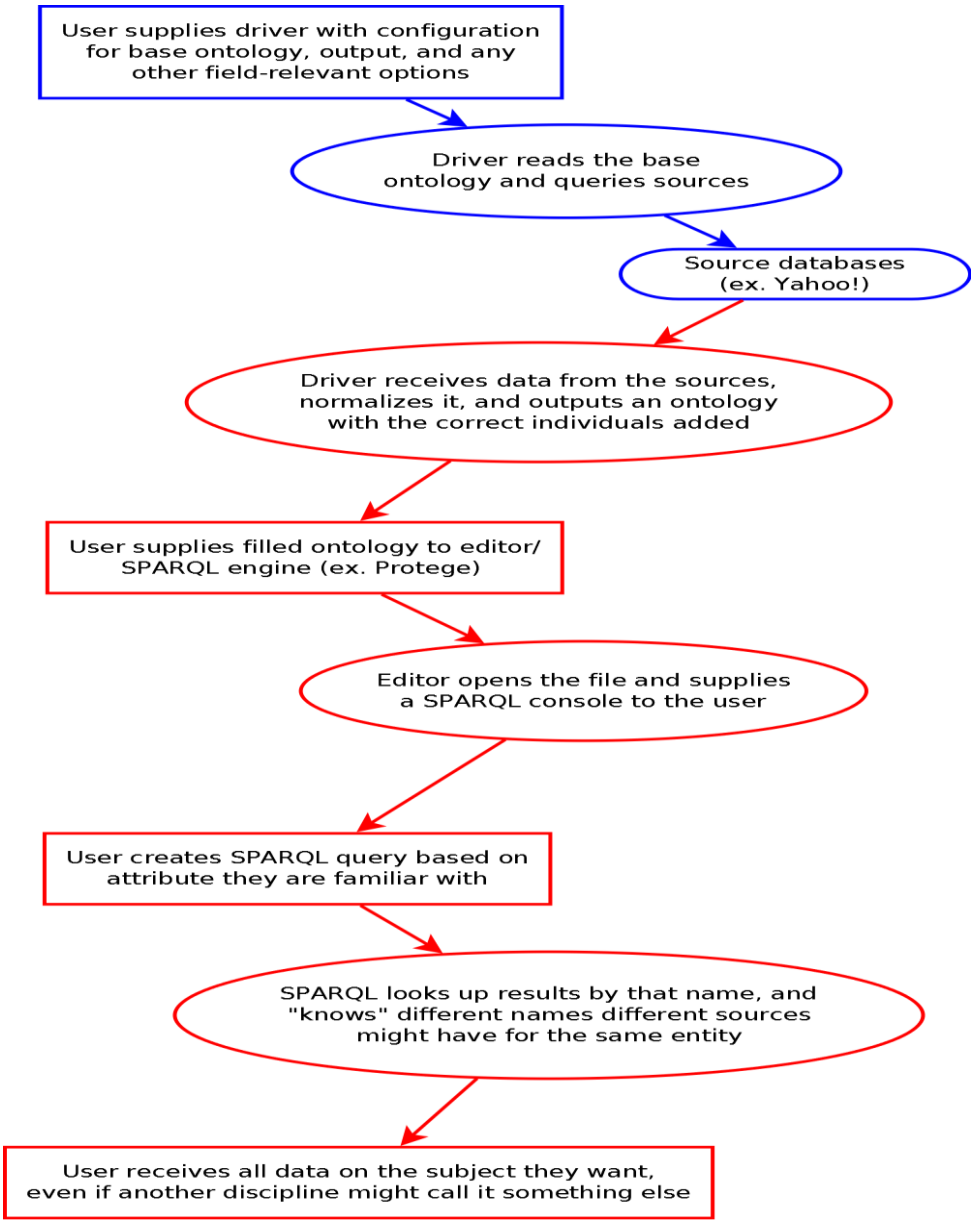

Figure 5: Workflow diagram. Rectangles represent user tasks, ellipses represent automated tasks performed by the ontology framework. Blue represents processes interacting with the base ontology; red represents interactions with the completely populated ontology.

On Adapting a Military Combat Discrete Event Simulation with Big Data and Geospatial Modeling Toward a Predictive Model Ecosystem for Interpersonal Violence

Fortune S. Mhlanga
fsmhlanga@lipscomb.edu
College of Computing and Technology
Lipscomb University
Nashville, Tennessee 37204, U.S.A.

E. L. Perry
eperry@faulkner.edu
Department of Computer Science
Faulkner University
Montgomery, Alabama 36109, U.S.A.

Robert Kirchner
kirchnerr@sbcglobal.net
USAF Retired
Cape Girardeau, Missouri, U.S.A

Abstract

The United States leads industrialized countries in rates of interpersonal violence with homicide being the second leading cause of death for people aged 15 to 24 years. In 2010, more than 4,800 youths (ages 10 to 24) received emergency treatment at hospitals due to injuries caused by physical assaults. This problem has taken epidemic proportions with 33% of high school students reporting physical altercations within the last year, 20% reporting being bullied on school grounds, 16% reporting electronic bullying, and 5% declaring that they had taken a weapon to school within the last 30 days prior to completing a survey conducted by the Centers for Disease Control in 2012. This paper presents an approach to adapting a military combat discrete event and gaming simulation with big data and geo-spatial modeling towards construction of a predictive model ecosystem for interpersonal violence. The ecosystem will be designed and tested using United States data on interpersonal violence collected over the past 20 years. Spatio-temporal data on interpersonal violence will be collected across the entire United States and stored in a Big Data management and analytics facility that will provide the basis for mapping the patterns of historical and current interpersonal violence. The facility will contain both analytical and simulation tools that collectively allow the researcher to input a strategy and observe predicted future states. The adapted discrete event simulation is envisioned to use a predictor-corrector method which will make the ecosystem a self-improving model for interpersonal violence prediction.

Keywords: Discrete Event Simulation, Interpersonal Violence, Predictive, Decision Support Systems, Regression.

1. INTRODUCTION

Interpersonal violence (IPV) among youth can result in significant physical, psychological, social, educational and economic consequences. Although rates of violence among youths have been in decline, IPV death remains the number one cause of death among youths aged 10 – 24. Furthermore, according to the Centers of Disease Control (CDC) Fact Sheet of 2012 on understanding youth violence, treatment of nonfatal injuries sustained from assaults caused more than 700,000 emergency room treatment visits in 2011 (CDC Fact Sheet 2012). Although no state is immune to this issue, Tennessee is ranked among the states with highest homicide rates among youth aged 10 - 24.

The U.S. Department of Justice has indicated that the predictors of youth violence can be grouped into five domains: (1) individual, (2) family, (3) school, (4) peer-related, and (5) community and neighborhood factors (Hawkins et. al 2000). Data from the long-term studies that have identified predictors of youth violence can ultimately help determine violence prevention policy and practice.

Despite the fact that Big Data is still a loosely defined term used to describe data sets so large and complex that they become awkward to work with using standard statistical software (Snijders 2012), it is a natural choice platform for our envisaged predictive model ecosystem. Although it has been defined in many different ways (Teradata 2013), a generally accepted definition of Big Data comes from Gartner (Sicular 2013). Gartner defines Big Data in terms of high-volume, high-velocity, and high-variety. IPV data meets all of these criteria:

- From a velocity perspective, there was a violent crime roughly every 30 seconds in the United States in 2011, and one third of high school students reported physical altercations in 2011 (CDC Fact Sheet 2012). Based on the 2008/2009 public high school enrollment (Agus 2010) that translates to an altercation roughly every 6 seconds.

- From a volume perspective, crime data is a clear example of Big Data. The FBI Unified Crime Report program dates back to 1930. In 2012 it included law enforcement agencies representing 308 million United States inhabitants (98.1 percent of the total population) (UCR 2013). This represents just one of many large sources of data on crime and aggression.

- From a variety perspective, multiple facets of crime data must be interrelated to understand IPV. Farrington cites multiple causes of IPV including antisocial behavior, aggressiveness, hyperactivity, parental criminality, poor family management, poverty, delinquent peers and more (Farrington 1998). Bringing together data which can represent this wide variety of potential causes of violence is a quintessential big data challenge.

IPV is a by-product of social, economic and political structure (Saenger 2000). Regardless of whether they are, for example, adult or youth, or male or female, people who unfortunately become exposed to this type of violence often find it difficult to put their traumatic experiences behind them. Thus, studies of IPV now go well beyond describing the physical injuries of victims and survivors to include analyses of psychological and emotional impacts (Kaukinen 2004 and Walker 2014). Because violence takes place at particular locations and times, many studies are also looking into the spatio-temporal patterns of this problem (Walker 2014 and Sparks 2011).

In mathematics, particularly numerical methods, a predictor–corrector method is an algorithm that proceeds in two steps. First, the prediction step calculates a rough approximation of the desired quantity. Second, the corrector step refines the initial approximation using another means. It is common to use an explicit method for the prediction and an implicit method for the correction. For example, in the solutions of ordinary differential equations, a simple predictor–corrector method (known as Heun's method) can be constructed from the Euler Method (an explicit method) and the trapezoidal method (an implicit method).

When a system is driven by the laws of Physics, a predictor-corrector methods that is often used is the Kalman Filter (Kalman 1960), named for Rudolf E. Kalman, one of the primary developers of its theory. Kalman filters are often used in guidance, navigation and control of vehicles, particularly aircraft and spacecraft. The filter forms a prediction of the defining state variables for the system using a time series of noisy input

data from radar, telemetry and on-board sources. The correction is done using a weighted average, with more weight being given to the estimates with higher certainty.

This paper is organized as follows. In Section 2 we describe the military combat discrete event simulation which we are adapting towards construction of our envisioned predictive model ecosystem. We also argue that a discrete event simulation is the correct high-level model for our predictive model ecosystem. Section 3 presents and discusses the high-level components of our Big Data and geospatially-enabled predictive model ecosystem for IPV. Section 4 presents our progress to date. Section 5 presents various ideas that we are currently exploring to support the architecture of our predictive model ecosystem. Finally, Section 6 concludes the paper with an outlook for future work.

2. THE MILITARY COMBAT DISCRETE EVENT AND GAMING SIMULATION

To avoid "reinventing the wheel", we have chosen to begin with an existing military combat discrete event and gaming simulation (originally called SIMWAR XXI), which was developed at Maxwell Air Force Base in Montgomery, Alabama for war games. The discrete event simulation is flexible enough to model other types of interactions such as the spread of a disease or the spread of an ideology or doctrine. The system contains a unique variable-resolution model of the earth's surface in which the surface is "tiled" with hexagons and (a few) pentagons. The hexagons can be scaled to different sizes depending on the requirements of the simulation. In addition, there are two built-in expert systems that can be used to model interactions and movements of players and units in the simulation.

The discrete event simulation comprises approximately 200,000 lines of C++ code that is flexible and portable to many different computer architectures. It uses a priority queue as the fundamental data structure for its event queue. It models the operation of any system as a discrete sequence of events in time. Each event occurs at a particular instant in time and marks a change of state in the system. Between consecutive events, no change in the system is assumed to occur; thus the simulation can directly jump in time from one event to the next. All events are stored on the event-queue and ordered by time. The basic cycle of operation is: (1) extract the next event from the queue; (2) update the simulation clock to the time of this event; and (3) execute the event, putting future events on the queue as necessary.

This contrasts with continuous simulation in which the simulation continuously tracks the system dynamics over time. Instead of being event-based, this is called an activity-based simulation; time is broken up into small time slices and the system state is updated according to the set of activities happening in the time slice. Because discrete-event simulations do not have to simulate every time slice, they can typically run much faster than the corresponding continuous simulation.

3. THE ENVISIONED SYSTEM

Figure 1 presents our first cut at depicting the architectural framework for our predictive model ecosystem for IPV. Our envisaged predictive model ecosystem comprises fundamental components: (i) a generic discrete event simulation facility (DES Facility) which will be adapted to spatio-temporal data on IPV, and (ii) a Big Data management and analytics facility (BDM&A Facility) which will be integrated to the DES Facility. The BDM&A Facility will be designed and built to integrate diverse and aggregated spatio-temporal data on IPV. This input data will represent an aggregation of populations along social, economic or demographic lines. The data will be characterized by a set of attributes that collectively will describe lifestyle, interactions and general quality of life of populations. The BDM&A Facility will be used to facilitate and discover new and unanticipated types of analysis and new information and knowledge pertaining to IPV. The BDM&A Facility is thus an integral component of the knowledge discovery process fostering prediction of future behavior of attributes, identification of the existence of subtle activities or events, and enacting strategies to blunt surprises, which may emerge as unanticipated consequences relative to IPV.

The BDM&A Facility will be used both as input to the initial models of the spatio-temporal data on IPV and as a real world picture of current conditions. When this real world picture is compared with the predicted picture from our model, statistical methods are used to obtain corrections to parameters within the model. This iterative predictor/corrector technique is used to

make AIM an ever improving model for the spread of infectious disease.

The DES Facility will be designed and built to have capability of modeling the entire world as one play box, with detailed terrain models of special areas of interest. We will utilize scenario generation tools along with visualization tools to model and visualize the dynamics of IPV in any part of the world. We envisage for the DES Facility to use a predictor-corrector method, which we hope to eventually automate. From our study of spatio-temporal data on IPV, we will be able to identify:

- initial values for a set of parameters used in a discrete event simulation to predict the state of IPV over a given region of study for a future date (perhaps a few weeks or months in the future);
- plausible ranges for each of the parameters; and
- the sensitivity of the predicted state to each parameter. This last factor, the sensitivity of the predicted state to each parameter, is computed using the internal equations of the discrete event simulation.

When the correct amount of time has elapsed, we will do another study to get the actual state of IPV over the region. The differences in the parameters for the actual state and predicted state can be used (along with the sensitivity data) to update the parameters of the discrete event simulation to give a self-improving feature to our predictive model.

4. CURRENT WORK AND INITIAL RESULTS

To avoid "reinventing the wheel," we have chosen to begin with adapting, and running some tests on, an existing generic discrete event simulator (SIMWAR XXI 2004) which will subsequently become the DES Facility for the ecosystem. In the current study, data from The Texas Almanac 2014-2015 (Texas A&M University Press, 2014) on IPV and population statistics of Texas counties has been collected and analyzed using linear regression. We wanted to see if IPV could be predicted from standard population statistics. While the data management, analytical methods and algorithms for the BDM&A Facility have not yet been crystallized and defined, the DES Facility together with historical data from the BDM&A Facility will form an initial model of the IPV.

To date, we have made progress, toward the ability to model and predict levels of IPV, in two areas. First, we have identified an initial set of objects and events for our discrete event simulation. Second, we analyzed some data from a set of Texas counties using linear regression in order to identify a basic predictor equation for IPV within each county.

Initial Set of Objects and Events for the DES

The full set of objects and events in our discrete event simulation is not shown due to page constraints. These objects are designed so that the simulator can be used for (more general) studies related to pandemics (Mhlanga 2013) as well as this initial study on IPV. At present, we have identified more than 35 objects, and their associated attributes, for the discrete event simulation. The overarching (root or superclass) object, called Ecosystem, encapsulates the entire set of objects specific to the study being conducted. It is described by general attributes (such as unique identifier, or ID, for the object together with its long name, or LongName) of all (subclass) objects with a description. Such subclass objects include, for example:

(i) APU (Autonomous Population Unit) – a section of the population that is treated as a single entity. It encapsulates the general information that describes the general attributes of all APUs. (The ID and /or LongName could possibly be formatted such that it maintains a pedigree of its ancestor APUs, e.g., US_TN_NASHVILLE_LIPSCOMB.);

(ii) Demographic – general information that describes the general attributes of all data concerning a specific aspect of the population of an APU, such as gender, marital status, ethnicity, or age range. It can be broken down into subtypes such as male and female for gender;

(iii) Enabler – something that allows a population to affect the ecosystem in the studied way, such as a rifle, knife, personal capability (use arms and legs as weapons), a belief, or a belief system that advocates violence;

(iv) Contributor – Ecosystem specific thing that causes an individual to be more likely to resort to a studied activity. If someone were abused as a child, they are unemployed, they are impulsive, their father uses drugs,

etc., that person may be more likely to engage in IPV;

(v) Influence – what affects one object, A, has on another object, B, from the point of view of object A. For example, a long hot spell (A) could cause an increase in the number of physical assaults (B);

(vi) IncidentType - a possible result including incidents such as murder, rape, assault, death, etc., of the use or activation of an Enabler in the study. This will most likely be a class hierarchy. This is because we need to be able to model deaths, because they change the demographics of the APUs. Also, other incidents may affect the data of other objects – so, modeling this as a class hierarchy will allow those types of incidents to be processed differently;

(vii) Zone – a defined geographic area. It can be used to model area-wide things such as weather, economic conditions, political conditions, etc., that would not necessarily be attributed to a single APU;

(viii) Condition – a physical, environmental, social, etc., set of circumstances in effect in a Zone or for a specific APU at a specific time. This could include weather conditions, social conditions, political conditions, etc.;

(ix) Impact – the effect (i.e., impact) an incident has on other objects. For example, a death incident should at least decrease the population of an APU but it may also have an effect on certain demographics within the APU.

Possible events at this time include the following:

(i) Interact – one APU or Actor has some kind of interaction with another APU or Actor;

(ii) EnablerEvaluationEvent – an event to evaluate a specific Enabler of a specific APU to determine if it is to be used or activated;

(iii) IncidentEvent – a result of an enabler being used or activated. For example, a murder, rape, assault, etc.;

(iv) ImpactEvent – makes an impact effective;

(v) Move – when one APU moves from one place to another;

(vi) Spawn – when part of an APU breaks off into a separate, independent APU;

(vii) Merge – when an APU joins another APU to become a single APU;

(viii) Condition Change – one or more attributes of a Condition changes.

Initial Results

For this initial study, our index of IPV in a county is the sum of the number of murders, the number of assaults and the number of rapes during a given time period. Table 1 shows a portion of the initial data set collected from the Texas Almanac (Texas Almanac 2014).

The full set contains data from a randomly selected set of 36 out the 254 counties in the state of Texas. The last column, labeled Tot IPV, is the sum of the number of murders, rapes and assaults in each of the 36 counties during 2012. This column represents the dependent variable for our study. We want to predict it from the independent (or explanatory) variables shown in columns 2 – 7. These variables represent the population of the county, the percentage of the population that is Anglo, the percentage that is Black, the percentage that is Hispanic, the per-capita income of the county, and the percentage of the population that is unemployed. Linear regression runs using these initial data did not produce acceptable results. The page limitations on this paper do not allow enough space to describe the general process of Stepwise Multiple Regression in detail. However, it is described in most of the textbooks (see (Garson 2013), for example) on the subject. (In general, one begins with the independent variable best correlated with the dependent variable. In Stage 2, the remaining independent variable with the highest partial correlation to the dependent is entered and a new regression is completed. This process continues until either (1) the addition of the new variable does not significantly increase r-squared; or (2) all variables are used. If the process terminates and the value of r-squared is not sufficiently high then the researcher looks for new independent variables. The ultimate goal is to get a set of independent variables that are not highly correlated among themselves but are highly correlated to the dependent variable and have the R-squared value above 0.95.)

While we hoped to find a predictive equation with r-squared above 0.95, we found that it could not be done using combinations of the variables from Table 1. This led to a series of experiments in which we added new independent variables and dropped old ones in our regression runs. During these experiments, we used data from the 2014 edition of the Texas Almanac (Texas Almanac 2014), the Texas Department of Public Safety Databases, and the Texas Education Agency Public Records.

A portion of the most successful of these experiments is shown in Table 2 and Figure 2. We used county population (x_1), per-capita income (x_2), public school drop-out rate per 100 students (x_3), the number of incarcerated persons in the county for 2012 (x_4), and the number of concealed carry weapon permits issued in 2012 (x_5) in the county to predict our index for IPV for 2012 (y).

Figure 2 shows a portion of the summary of this regression. Note that the r-squared value is 0.97, indicating that 97% of the variation in the index of IPV is explained by Equation 1.

$$y = (0.001352*x_1) + (0.000599*x_2) + (22.83063*x_3) + (1.396537*x_4) - (0.38155*x_5) + error$$

Equation 1. Predictor equation

Further research is needed to see if this is unique to Texas. (We suspect that it is but much more work is to be done.) In Equation 1, *error* is a random variable which is normally distributed about 0.

We refer to Equation 1 as the *predictor equation*. It leads to the following observations:

(i) IPV can be expected to increase by 1 for each 1000 person increase in population.

(ii) Each $10000 increase in per-capita income will lead to an average of about 6 new cases of IPV.

(iii) An increase (or decrease) of 0.1 in the dropout rate in the public schools will lead to a corresponding increase (or decrease) of two cases of IPV per year.

(iv) An increase (decrease) of 1 in the county prison population will lead to and increase (decrease) of 1 in the county IPV cases.

(v) The negative sign in the coefficient for x_5 indicates that for each increase of 10 in the number of concealed weapon permits in a given year, one can expect a decrease of about 4 in the cases of IPV in the county.

With reference to the predictor equation observations (i) – (v) above, it is important to note the big difference between "prediction" and "causation" and that we have simply observed that, in the Texas data, there is, from (ii) for example, a positive relationship between per-capita income and IPV. This does not mean that increasing income causes violence. We suspect that this is unique to Texas. In the past few years, there has been a large increase in per-capita income in the oil regions of West and South Texas. Crime has also increased dramatically as oil field workers from around the world have scrambled for jobs in these areas. The same comments are appropriate for observation (iv). Crime is on the rise in many areas of Texas due to the big money brought into the state by the oil companies. We are not suggesting that putting people in jail causes increased violence. However, there is a positive relationship that could be used for predictive purposes. Our overall goal in the study was to produce a predictive equation for the state of Texas. It was not to determine the causes of IPV.

5. IMPLEMENTATION IDEAS

At this early stage of project conception, we are still exploring appropriate storage and management, implementation, modeling and simulation approaches befitting to support our architectural framework. We are currently exploring a Hadoop cluster for the BDM&A Facility along with other statistical tools for the analytics. We are also studying available tools that normalize data, exclude outliers and determine correlations, especially for the so-called "*data munging*" and for extracting the behavior model for the environment.

One approach that has come to mind is to treat IPV as a dynamic system. In such a model, the reference would be the current socio-economic environment and bullying (or being bullied) comprising the output. Measurement of the inputs and output of the previous state would lead to a model able to predict future state. The model could be trained on individual measurements from students in temporal order.

After training, an input function could be established able to predict methods which would end the literal cycle of violence. Implementation could be fairly simple as, once data was organized by subject first, then year, the data could be sequentially parsed by the model with very little data existing in memory at any one time.

We are also considering a graph-based human behavior prediction model in which a Bayesian behavior graph could be created based on observed action paths taken by subjects attempting to obtain goals. These paths would be combined into a Bayesian network or even a partially observable Markov decision process (POMDP) if probability becomes a big part of the calculations. The values of probabilities in the conditional probability tables associated with each node could be learned through analysis of the paths taken by the subjects and their associated attributes. As the graph would be relatively small, but would need traversing repeatedly, it's storage as a simple program object would facilitate calculation. Each subject's temporal activities could be projected onto the graph. This approach would require knowledge of subjects' sequential actions.

Another means of looking at bullying is to look at it as an economics transaction. In this case, the bully purchases power from the bullied. The 'cost' of bullying is effectively an externality on the bullied. The unit of the purchase is 'power'. Based on the survey data, bullying or being bullied could be modeled as a transaction and a wealth of 'power' possessed by each subject could be recorded. In such a model, indicators could be considered as factors effecting the volume of 'power' transferred during the bullying transaction or as source of outside 'power' affecting the wealth of an individual. As each subject would need to have their personal wealth of 'power' tracked, persistent storage will be required in the BDM&A Facility. An initial model could be postulated with testing of the sensitivities to specific indicators used to refine the model.

Bullying could also be modeled as a disease which spreads from subject to subject. A dynamic network representing the subjects would be created and updated for each time slice of the data. The probability of spread of the bullying 'disease' would then be calculated based on the indicators also resident on the subject. To accomplish this model, inter-personal relationships would have to be somehow established. This could be gleaned from social network data, or geographic location data. The storage of a large dynamic network is problematic as most graph tools are not well suited for large dynamic networks. Instead, the graph and node attributes may best be represented in a relational database. If the model were to simply promulgate the disease over the graph without changing the underlying architecture, the graph could be stored in a graph database with node attributes identified both by their attribute and time slice. Either way, the relational database or graph database will be a sub-component of the BDM&A Facility.
Although geographic information systems (GIS) are de facto tools for analyzing, interpreting and presenting spatio-temporal information (Longley et. al., 2011), few studies have exploited the capabilities of GIS to help understand, address problems, predict the distribution of and make decisions concerning IPV. One reason is that there are still some unknowns regarding the application of GIS concepts and methods in studies of the spatio-temporal patterns and causes of violence (Pridemore 2010). GIS have, however, been employed in many studies involving crime in general (Wang 2005). Pain et al. (Pain 2006), for example, used GIS to address simple but important "[w]hen, where, if, and but" questions about the effects of street lighting on crime and the fear of crime occurrence. Through GIS, Walker et al. (2014) conducted an exploratory spatio-temporal analysis of the distribution of violent trauma hotspots many of which were correlated with night club areas and Saturday night times.

The beginning point for the discrete event simulator in our study of IPV is an existing simulation which was previously used for combat simulation and war games within the United States Air Force. This simulation is completely data driven, which makes it extensible to other domains. Although we will discard much of the combat portion of this model, we plan to retain the ground and terrain model which can be used to model the entire world (or any portion of it) as a tiled region of variable-sized hexagons and pentagons. The new objects and events will be general enough to facilitate the use and extension of this tool for other studies including world pandemics (Avian-flu, AIDS, etc.), political issues (greenhouse gases, fresh water, etc.), drugs and drug trafficking, and others. The existing simulation also includes two expert systems that may be useful for defining rules for

population interaction in all of the studies and as a basis for the corrector method. We will add code to the discrete event simulator to allow a feedback or predictor / corrector loop as shown in Figure 1. A finite set of parameters $\{x_1, x_2, \ldots, x_n\}$ will determine the "state" of the system. For example, we might choose x_i as the percent of IPV relative to population Pi. This state can then be easily compared to the actual state at a given time. The differences between the actual and predicted values will determine corrections which can be applied to model parameters and processes to get better predictive capability in the next time cycle (Gershenfeld 1999).

If the existing discrete event simulator proves difficult to adapt, we can consider implementing a neural network or classification system such as the support vector machine which could be used to perform the predictive aspect of the ecosystem. Both of these systems support the predictor/corrector technique or method.

The predictive system is also envisioned to follow a process that employs reinforcement or machine learning techniques. Such systems manipulate the data to produce a model and predict the behavior of the environment and then display the results in a simulation. When real data comes in and comparisons are made, the systems receive feedback that may require them to re-analyze the data and predict a more accurate model.

6. CONCLUSION

We have presented ideas towards development of a Big Data and geospatially-enabled predictive model ecosystem for IPV. This approach in which we combine a robust big data management and analytics capability with predictor / corrector methods to forecast IPV is unique and novel. It extends the use of the Kalman Filter predictor / corrector methods to new domains and the use of data mining and knowledge discovery technologies (commonly used in areas of retail and marketing, banking and finance, manufacturing, and healthcare).

The capabilities of GIS to handle large volumes of data can be augmented through geovisualization tools and techniques. Unlike GIS which are powerful largely in the area of geocomputational analysis, geovisualization places the user squarely at the center of geospatial data analysis, interpretation and sense-making (Hodza 2009). The goal is to exploit the over 50% of our brain neurons that are primarily for supporting our visual sense. The goal is also to augment human cognition through the use of highly interactive, dynamic and multidimensional visual displays like maps, charts, tables and graphs. This in itself is important because there are many cases where the human eye-mind combination is more effective and efficient at uncovering spatio-temporal patterns, relationships and trends embedded in large and complex data (Byrne 1999). Heer (2013) cites John Tukey, the famous mathematician as having said "Nothing – not the careful logic of mathematics, not statistical models and theories, not the awesome arithmetic power of modern computers – nothing can substitute here for the flexibility of the human mind."

Our goal is to facilitate geospatial and geovisual thinking by exploiting the combination of the geocomputational capabilities of GIS and the geovisual analytics power of geovisualization. Both GIS and geovisualization are useful tools and techniques in geospatial data mining (Valencio 2013) which is also of primary importance in this study.

Although we plan to initially study the problem of IPV within some region, our methods and tools are general enough to apply to other medical-socio problems involving the spread of disease (Mhlanga 2013) and (possibly) other domains.

While we have some interesting results from our initial study, we need to extend our data set to include randomly selected counties for randomly selected other states and see if we get similar results. We also realize that the predictor equation in Section 4 does not allow for population dynamics. This will come from the use of the discrete event simulation. Population movement, weather, other dynamic local situations can increase or decrease IPV in the given locality. We view the discrete event simulation as the ideal tool for dynamic analysis. While our initial study used the county as the basic population unit (APU), this may not be the best one for our simulation.

Our collection of objects and events will continue to evolve as we collect more and more data on IPV and begin to build the BDM&A Facility to test these objects and events. We are also garnering a better understanding of the objects and events themselves. For instance, we are looking at

ways to specify what a change to an attribute of an object does to other objects when there is an Influence relationship between the objects, or what affect an Incident occurring would have on other objects. For example, a death Incident should affect the population count of one or more APUs, and may also have an effect on one or more Demographics of the population of the APU, affect some aspect of a Contributor, Influence, Condition, Enabler, etc.

Once determined, the BDM&A Facility will also define the schema to accommodate the real-world and intermittent results of the simulation. We will also gradually get a better grasp at defining how the simulation would actually work. We are currently entertaining ideas to determine how close the result of a simulation are to real-world conditions, and what data changes would need to be made to get the results of a simulation run closer to the real-world results. Our goal is get a simulation where we can test strategies for reduction in IPV.

As we continue this work, we also plan to add more geo-temporal analysis techniques to better understand IPV.

7. REFERENCES

Agus, Jessica (2010). "National High School Center at AIR", High Schools in the United States, Quick Stats Fact Sheet, December 2010, (http://www.betterhighschools.org/pubs/documents/HSInTheUS_1210.pdf).

Byrne, Christina A. and Heidi S. Resnick and Dean G. Kilpatrick and Connie L. Best and Benjamin E. Saunders (1999). "The Socioeconomic Impact of Interpersonal Violence on Women", Journal of Consulting and Clinical Psychology, Vol. 67 (3), pp. 362-366, 1999. doi: 10.1037/0022-006X.67.3.362.

CDC Fact Sheet (2012). "Understanding Youth Violence", National Center for Injury Prevention and Control, Division of Youth Violence, Centers for Disease Control, (http://www.cdc.gov/violenceprevention/pdf/yv-factsheet-a.pdf), 2012.

Farrington, David P (1998). "Predictors, Causes, and Correlates of Male Youth Violence", Journal of Crime and Justice, Vol. 24, pp. 421-475, The University of Chicago Press (http://www.jstor.org/stable/1147589), 1998.

Garson, G. David (2013). "Multiple Regression". Statistical Associates Publishing. Blues Book Series, 2013, (http://www.statisticalassociates.com)

Gershenfeld, Neil (1999). "The Nature of Mathematical Modeling", Cambridge University Press, 1999.

Hawkins, J. David and Todd Herrenkohl and David Farrington and David Brewer and Richard Catalano and Tracy Harachi and Lynn Cothern (2000). "Predictors of Youth Violence", Juvenile Justice Bulletin, Office of Juvenile Justice and Deliquency Prevention, U.S. Department of Justice, April 2000.

Heer, Jeffrey (2013). "Interactive Visualization of Big Data", O'Reilly Strata, (http://strata.oreilly.com/2013/12/interactive-visualization-of-big-data.html), December 20, 2013.

Hodza, Paddington (2009). "Evaluating user experience of Experiential GIS", Transaction in GIS, Vol. 13 (5-6), pp. 503-525, 2009.

Kalman, Rudolph E. (1960). "A New Approach to Linear Filtering Prediction Problems", Transactions of the ASME - Journal of Basic Engineering, Vol. 82 (Series D), pp. 35-45, 1960.

Kaukinen, Catherine (2004). "Status Compatibility, Physical Violence, and Emotional Abuse in Intimate Relationships", Journal of Marriage and Family, Vol. 66 (2), pp. 452-471, 2004.

Longley, Paul A. and Michael F. Goodchild and David J. Maguire and David W. Rhind (2011). "Geographic Information Systems and Science", 3rd edition, Wiley, London, 2011.

Mhlanga, Fortune S. and E.L. Perry and C-S Wei, and Peter A. Ng (2013). "Towards a Predictive Model Architecture for Current or Emergent Pandemic Situations", 2013 Summer Simulation Multi-Conference (Summer Sim'13), Society for Modeling & Simulation International, Toronto, Canada, ACM DL 2013 ISBN 978-1-62748-276-9 (57), July, 2013.

Pain, Rachel and Robert MacFarlane and Keith Turner and Sally Gill (2006). "'When, Where, If, and But': Qualifying GIS and the Effect of Street Lighting on Crime and Fear", Environment and Planning, 38, 2055-2074, 2006.

Pridemore, William A. (2010). "Using GIS and Spatial Analysis to Better Understand Patterns and Causes of Violence", 2010 Annual Meeting of the American Association for the Advancement of Science (AAAS), 18-22 February, 2010, San Diego, USA.

Saenger, Sieglinde A (2000). "Family Violence : A Review of the Dysfunctional Behavior Patterns", Minnesota Center Against Violence and Abuse, MINCAVA electronic clearinghouse, 2000, (http://www.mincava.umn.edu/documents/familyviolence/familyviolence.html#idp30185040).

Schiller, Daniel and Ingo Liefner (2007) "Higher Education Funding Reform and University-Industry Links in Developing Countries: The Case of Thailand." International Journal of Higher Education and Educational Planning, vol. 54, no. 4, pp. 543-556, October 2007.

Sicular, Svetlana (2013). "Gartner's Big Data Definition Consists of Three Parts, Not to be Confused with Three "V"s", Gartner, Inc., (http://www.forbes.com/sites/gartnergroup/2013/03/27/gartners-big-data-definition-consists-of-three-parts-not-to-be-confused-with-three-vs/), March 27, 2013.

SIMWAR XXI (2004). "Simulation Engine Analyst Manual", Air Force Wargaming Institute, Maxwell AFB, AL. April 2004.

Snijders, Chris and Uwe Matzat and Ulf-Dietrich Reips (2012). "'Big Data': Big Gaps of Knowledge in the Field of Internet Science'", International Journal of Internet Science, Vol. 7 (1), pp. 1-5, 2012.

Sparks, Corey (2011). "Violent Crime in San Antonio, Texas: An Application of Spatial Epidemiological Methods", Spatial and Spatio-Temporal Epidemiology, 2 (4), pp. 301-309, DOI:10.1016/j.sste.2011.10.001, 2011.

Teradata (2013). The Big Data Conundrum: How to Define It? (http://www.technologyreview.com/view/519851/the-big-data-conundrum-how-to-define-it/), October 3, 2013.

Texas Almanac (2014). Published by Texas Historical Association, 1155 Union Circle, #311580, Dention, Texas, 76203.

UCR (2013). "Law Enforcement Officers Killed and Assaulted, 2012", Summary of the Uniform Crime Reporting (UCR) Program, U.S. Department of Justice – Federal Bureau of Investigation, (http://www.fbi.gov/about-us/cjis/ucr/leoka/2012/standard-ucr-info/about_ucr_2012.pdf), Released Fall 2013.

Valencio, Carlos R. and Thatiane Kawabata and Camila A. de Medeiros and Rogeria C. G. de Souza and José M. Machado (2013). "3D Geovisualisation Techniques Applied in Spatial Data", MLDM'13 Proceedings of the 9th international conference on Machine Learning and Data Mining in Pattern Recognition, 57-68, Springer-Verlag Berlin, Heidelberg, 2013.

Walker, Blake B. and Nadine Schuurman and S. Morad Hameed (2014). "A GIS-based Spatiotemporal Analysis of Violent Trauma Hotspots in Vancouver, Canada: Identification, Contextualisation and Intervention", BMJ Open, 4 (2), DOI: 10.1136/bmjopen-2013-003642, 2014.

Wang, Fahui (2005). "Geographic Information Systems and Crime Analysis", Hershey, P.A.: Idea Group Publishing, 2005.

Editor's Note:

This paper was selected for inclusion in the journal as a CONISAR 2014 Distinguished Paper. The acceptance rate is typically 7% for this category of paper based on blind reviews from six or more peers including three or more former best papers authors who did not submit a paper in 2014.

APPENDICES

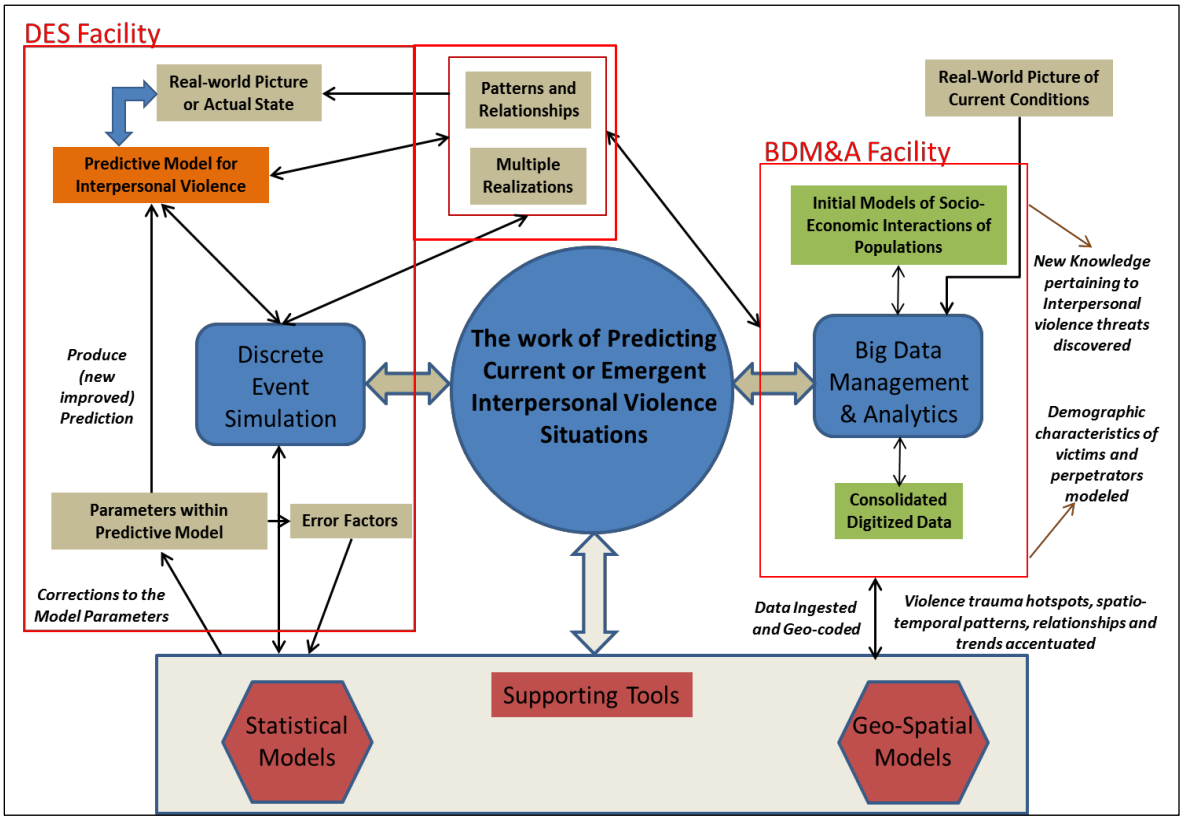

Figure 1. Predictive Model Ecosystem for IPV

Factors							Interpersonal Violence Study			
County	**Pop**	**%Anglo**	**%Black**	**%Hisp**	**PC Income**	**%Unemp**	**Murder**	**Rape**	**Assault**	**Tot IPV**
Bowie	93148	65.72	24.03	7.05	35360	6.7	7	13	454	474
Brazos	200665	58.75	10.54	23.86	29045	5.7	5	52	670	727
Briscoe	1561	69.11	2.48	26.35	27769	8.1	1	0	0	1
Castro	8164	36.11	2	60.79	48285	5.4	0	0	7	7
Colorado	20696	59.05	12.57	26.97	39030	5.8	3	12	39	54
Crane	4562	39.43	3.31	55.51	36362	5.1	9	1	8	18
Deaf Smith	19360	29.59	0.96	68.36	35880	5.2	0	1	26	27
Denton	707304	63.71	8.46	18.73	42371	5.9	5	133	484	622
Falls	17610	52.21	24.71	21.47	28073	8.9	1	4	23	28
Freestone	19515	67.98	15.88	14.49	31573	6	1	1	39	41
Grimes	26783	59.87	16.08	22.2	31418	6.6	1	7	85	93
Hall	3293	57.86	6.84	34.08	23662	8.4	0	0	5	5
Hamilton	8307	86.69	0.86	10.93	20238	5.8	0	2	11	13
Hill	35115	72.55	6.57	18.93	32266	7.1	1	10	71	82
Leon	16803	76.6	7.54	13.9	35114	7.3	1	4	9	14
Matagorda	36547	46.72	10.95	39.22	33287	10.1	1	11	90	102
Maverick	55365	3.2	0.22	95.24	22188	14.8	3	4	156	163
Menard	2240	62.94	0.75	35.42	30157	7.2	0	0	4	4
Montague	19565	86.92	0.57	10.3	40161	5.1	0	4	20	24
Montgomery	485047	70.26	4.36	21.43	48508	5.8	12	48	547	607
Moore	22313	37.32	1.76	53.02	34060	4	1	8	43	52
Morris	12787	66.24	22.44	8.65	34904	9.5	0	1	53	54
Orange	82977	82.22	8.68	6.36	38163	11	1	16	231	248
Panola	24020	72.94	16.26	8.82	39654	6	0	7	57	64
Potter	122335	48.36	9.82	35.99	33714	5.7	10	111	897	1018
Rains	10943	86.3	2.62	8.36	30131	7.4	0	7	14	21
Randall	125082	76.49	2.63	17.58	40001	4.3	1	2	53	56
Real	3369	70.34	0.82	26.5	30296	7.7	4	6	1	11
Reeves	13798	19.71	4.99	74	23505	9.9	2	0	14	16
San Saba	6002	66.66	3.45	28.36	31384	8.4	1	0	12	13
Scurry	17126	56.91	4.64	39.96	37970	4.3	1	9	49	59
Taylor	133473	50.42	0.73	46.77	37132	5.3	3	44	335	382
Titus	32663	48.16	9.29	40.58	28542	7.3	0	0	62	62
Travis	1095584	50.29	8.09	33.87	43198	5.7	33	246	2703	2982
Upton	3283	46.2	1.64	50.36	45030	3.7	0	0	1	1
Williamson	456556	63.09	6.13	23.61	40067	5.9	2	89	349	440
Wilson	44370	58.2	1.72	38.52	34810	6.2	3	6	34	43
Yoakum	8075	38.08	0.94	59.39	41060	3.5	1	5	0	6
Zapata	14290	6.4	0.36	92.83	25162	6.9	0	1	28	29

Table 1. Initial data from Texas counties (Texas Almanac 2014)

County	Pop	PC Income	DORate/100	Incar2012	CCPermits2012	Tot IPV
Bowie	93148	35360	0.7	338	687	474
Brazos	200665	29045	2.4	585	1177	727
Briscoe	1561	27769	0	1	7	1
Castro	8164	48285	1.1	16	29	7
Colorado	20696	39030	1	52	178	54
Crane	4562	36362	1.2	11	40	18
Deaf Smith	19360	35880	0.8	72	101	27
Denton	707304	42371	0.7	1093	4716	622
Falls	17610	28073	2.7	34	105	28
Freestone	19515	31573	0.8	53	113	41
Grimes	26783	31418	1.7	73	186	93
Hall	3293	23662	0	9	14	5
Hamilton	8307	20238	0.7	15	75	13
Hill	35115	32266	0.4	145	214	82
Leon	16803	35114	0.4	22	190	14
Matagorda	36547	33287	0.6	125	234	102
Maverick	55365	22188	1.3	70	46	163
Menard	2240	30157	0.6	6	9	4
Montague	19565	40161	0.3	63	157	24
Montgomery	485047	48508	0.1	1145	4223	607
Moore	22313	34060	1	42	114	52
Morris	12787	34904	0.2	31	79	54
Orange	82977	38163	1.5	181	767	248
Panola	24020	39654	1.2	52	174	64
Potter	122335	33714	2.3	491	551	1018
Rains	10943	30131	0.4	26	80	21
Randall	125082	40001	0.5	277	1191	56
Real	3369	30296	4.1	6	43	11
Reeves	13798	23505	0.9	33	10	16
San Saba	6002	31384	0.2	10	52	13
Scurry	17126	37970	0.7	45	98	59
Taylor	133473	37132	2	523	944	382
Titus	32663	28542	0.2	105	150	62
Travis	1095584	43198	2	2314	4546	2982
Upton	3283	45030	0	9	7	1
Williamson	456556	40067	0.6	574	3022	440
Wilson	44370	34810	0.6	66	361	43
Yoakum	8075	41060	0.2	13	44	6
Zapata	14290	25162	2.2	36	48	29

Table 2. Final data from Texas counties

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.98633
R Square	0.972847
Adjusted R Square	0.968732
Standard Error	90.70097
Observations	39

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*	*Upper 95.0%*	*Upper 95.0%*
Intercept	-31.276	89.61255992	-0.349013202	0.729296596	-213.5940905	151.0422	-213.594	151.0422	207.2783	312.9804
X Variable 1	0.001352	0.000382242	3.538227445	0.001220996	0.000574782	0.00213	0.000575	0.00213	0.005241	0.005448
X Variable 2	0.000599	0.002562329	0.23396538	0.816457299	-0.004613601	0.005813	-0.00461	0.005813	0.013961	0.012561
X Variable 3	22.83063	18.07397649	1.263176848	0.2153778	-13.941153	59.60241	-13.9412	59.60241	133.862	-0.31876
X Variable 4	1.396537	0.144064212	9.693849717	3.50639E-11	1.103435974	1.689638	1.103436	1.689638	-0.29273	
X Variable 5	**-0.38155**	**0.04080004**	**-9.351587732**	**8.44079E-11**	**-0.464553459**	**-0.29854**	**-0.46455**	**-0.29854**		

Figure 2. Regression summary output

Measuring Algorithm Performance With Java: Patterns of Variation

Kirby McMaster
kmcmaster@weber.edu
Computer Science
Moravian College
Bethlehem, PA 18018, USA

Samuel Sambasivam
ssambasivam@apu.edu
Computer Science
Azusa Pacific University
Azusa, CA 91702, USA

Stuart Wolthuis
stuart.wolthuis@byuh.edu
Computer & Information Sciences
BYU-Hawaii
Laie, HI 96762, USA

Abstract

Textbook coverage of algorithm performance emphasizes patterns of growth in expected and worst case execution times, relative to the size of the problem. Variability in execution times for a given problem size is usually ignored. In this research study, our primary focus is on the empirical distribution of execution times for a given algorithm and problem size. We examine CPU times for Java implementations of four sorting algorithms: selection sort, insertion sort, bubble sort, and quicksort. We measure variation in running times for these sorting algorithms. We show how the sort time distributions change as the problem size increases. With our methodology, we compare the relative stability of performance for the different sorting algorithms.

Keywords: algorithm, sorting, performance, variation, order-of-growth, Java.

1. INTRODUCTION

The performance of algorithms is addressed at different levels throughout the computing curriculum. In introductory programming courses, informal comparisons of alternative algorithms are presented without a rigorous theoretical framework (Lewis and Loftus, 2011; Liang, 2012).

In Data Structures textbooks (Koffman & Wolfgang, 2010; Lafore, 2003), the emphasis is on how to implement algorithms to support data structures of varying complexity, such as stacks, priority queues, binary search trees, and

weighted graphs. A casual introduction to "Big-Oh" notation is included to relate problem size to execution time for various types of algorithms.

In Analysis of Algorithms textbooks (Cormen, Leiserson, Rivest, & Stein, 2009), the discussion of algorithm performance places greater emphasis on mathematical reasoning. A formal examination of algorithm efficiency based on resources required (primarily CPU time) looks at best case, worst case, and average case situations.

Most of the discussion centers on worst case analysis because the mathematical arguments are simpler. *Order-of-growth* is defined to ignore constants and lower order terms, so average case results are often proportional to the worst case. Worst case examples provide an upper bound on the execution time for an algorithm.

Sedgewick & Wayne (2011) present a mathematical analysis of algorithms, and then relate their mathematical models to empirical results obtained from algorithm run times on a computer. They give several algorithms for finding three numbers (from a large input file) that sum to zero. They ran each algorithm once for each input file, assuming that the only source of variation was the actual data. However, in our research we experienced situations where repeated execution of the same algorithm on the same data resulted in different execution times.

Some textbooks briefly mention that running times can vary for different inputs. However, they include no discussion of the nature of the *distribution* of execution times for random inputs. Variation includes not only *dispersion* (how spread out the scores are from a central value), but also *skewness* (how unbalanced the scores are at each end of the distribution).

Variation can be of greater importance than averages when consistency/dependability of execution time is a major requirement. This is true in systems having strict time constraints on operations, such as manufacturing systems, real-time control systems, and embedded systems (Jones, 2009).

Research Plan

The primary objective of this research is to examine how algorithm execution time distributions depend on problem size, randomness of data, and other factors. We limit our study to sorting algorithms for arrays of integers. In the next section, we list potential sources of variation for execution times. We then describe our experimental design to control sources of variation beyond algorithm structure and problem size. Our results and conclusions are summarized later in the paper.

2. SOURCES OF VARIATION

There are many system features which can affect algorithm performance. In this research, we use CPU time as our primary measure of performance. A layered list of sources of variation in sort times is outlined below.

1. Computer hardware components: (a) CPU clock speed, pipelines, number of cores, internal caches, (b) memory architecture, amount of RAM, interleaved RAM, external caches.

2. Operating system features: (a) process scheduling algorithms, multi-tasking, parallel processing, (b) memory allocation algorithms, and virtual memory.

3. For Java programs: (a).Java JIT compiler, (b) Java run-time options, (c) Java run-time behavior, especially automatic garbage collection.

4. Application program: (a) choice of algorithm, and how it is implemented, (b) size of problem, (c) amount of memory required by the algorithm, (d) data type and data source.

Our main focus in this paper is on patterns of variation in execution times due to features in the *application program*. We limit our research to *sorting algorithms*, including selection sort, insertion sort, bubble sort, and quicksort. We examine a range of array sizes, and repeatedly fill the arrays with random integers.

To minimize algorithm performance effects from the lower hardware and software layers, we ran all final results on a single computer. This computer had an Intel Core2 Duo CPU, Windows 7 operating system, and Version 7 of the Java compiler and run-time.

Unexpected Variation

In our research environment, we assumed that algorithm execution times would depend almost entirely on:

1. the sorting algorithm
2. the size and data type of the array
3. the randomness of the generated data

Surprisingly, this assumption was *not* supported by our test data. Unexpected patterns of variation in performance were encountered throughout our research study.

For example, early in the exploratory phase of our study, we performed the *selection* sort algorithm 7 times on an array size of 100. For each sort operation, independent random values of type *int* were generated to fill the array. The execution times in nanoseconds (ns) for the sort module were:

```
113827
320489
 16328
 15394
 14928
 14928
 14462
```

A statistical summary of CPU times to sort these arrays is:

```
Minimum =  14462
Median  =  15394
Maximum = 320489

Mean    =  72908
Std dev = 115195
```

Several patterns in this data can be noted:

1. The maximum sort time is more than 20 times larger than the median. This is due to the presence of *outliers* (large sort times) in the sample.

2. The median sort time is only slightly larger than the minimum.

3. The average sort time is much larger than the median, suggesting a *positively-skewed* distribution.

4. The standard deviation of the sort times is larger than the mean. This measure of variation is greatly inflated by outliers.

3. METHODOLOGY

The above example containing outliers was not atypical in our study. Because of these unexpected patterns in execution time data, we developed a *methodology* for generating and analyzing performance data that is relatively immune to outlier effects.

CPU time measurement does not provide an "exact" performance value for an algorithm. Karl Pearson theorized that measurements represent samplings from a *probability distribution* of values (Salsburg, 2001). For example, to answer the question of "how fast is a sprinter?", his/her running times in 100-meter dash events over a season provide a partial answer in the form of a distribution of sample values.

For a given hardware/software environment, sorting algorithm, and array size, our methodology assumes that the distribution of execution times is a *mixture* of two components: (a) *normal* variation due to randomness of the data, and (b) other sources of variation that result in outliers.

Our methodology attempts to extract the normal variation component from the combined distribution. This requires being able to *detect* possible outliers and *remove* them from the sample.

Our sort time data often contained a relatively large number of outliers. Therefore, we did not perform statistical tests to detect individual outliers. Instead, we used two general approaches for removing outliers:

1. Set limits on the perceived "normal" data, and *trim* off values outside these limits. In particular, we examine trimmed means and trimmed standard deviations.

2. Use statistics such as the median that are less susceptible to outliers.

Our performance analysis approach was developed first for the *selection sort* algorithm. Samples of execution times for selection sort were obtained for a range of array sizes starting with 100.

Our Java data generation program, initially written for selection sort, performs the following steps:

1. Input the array size (N) and number of algorithm repetitions (R).

2. For each repetition:
   a. fill the data array with random integers.
   b. sort the array, and place the execution time (collected using the Java System nanoTime function) in a SortTime array.

3. After all repetitions are completed, sort the execution times in the SortTime array.

4. Calculate various statistical summaries of the execution times. This part of the Java program was modified frequently throughout the study.

As data were collected for the sorting algorithm, we evaluated how well different statistics summarized essential features of the sort time distributions. When the methodology began to provide consistent results for selection sort, we applied the methodology to the remaining sorting algorithms.

Sample Case

The following sample case demonstrates much of the process in developing our methodology. In this case, the array size is 100, and the number of repetitions is 1000. A frequency distribution of the 1000 sort times obtained from running our Java program once is shown below.

Table 1: Selection Sort Distribution.
Sort Time in nanoseconds (ns)
Size N = 100, Repetitions R = 1000

SortTime	Freq	CumFreq	Diff
14461	36	36	---
14462	68	104	1
14928	472	576	466
14929	78	654	1
15394	124	778	465
15395	194	972	1
15861	17	989	466
15862	4	993	1
16328	1	994	466
17261	1	995	933
19127	1	996	1866
37320	1	997	18193
108695	1	998	71375
111028	1	999	2333
113827	1	1000	2799

Several unusual features appear in the above distribution:

1. The sample of sort times contains many repeat values. Only 15 distinct values appear in the 1000 repetitions of the sorting algorithm.

2. Among the smaller sort times, most appear in "pairs", differing only by 1 nanosecond. This is probably due to rounding, since the nanoTime function returns an integer.

3. If we consider pairs differing by 1 as a single value, over 99% of the distribution is concentrated in 4 sort time pairs.

4. Again considering pairs differing by 1 as a single value, the difference between consecutive pairs is between 466 and 467. We can interpret this difference as the resolution of the "clock tick" for our nanoTime clock. Oracle's Java documentation (Oracle, 2014) states that the System.nanoTime method "returns the current value of the most precise available system timer, in nanoseconds." Apparently, our recorded sort times are not accurate to 1 nanosecond. In tests on other computers, we observed that the clock increment is hardware specific.

5. The three largest values--113827, 111028, and 108695--are clearly outliers. But are there other outliers? The distribution is slightly skewed, even without top three values.

6. The median of the distribution is 14928, which is close to the minimum value.

We now ask the most important question for our methodology. "What characteristics of the sort time distribution are relevant for describing patterns of variation?" We will be generating sort time distributions for different sorting algorithms and various array sizes. The patterns of variation we are trying to explain should be observable within each of these separate distributions.

A related research question is: "What statistical measures best summarize the variation in sort time distributions, without being distorted by outliers?" Three characteristics of distributions are of particular interest:

1. *central tendency*: Where is the "center" of the distribution? Outliers can distort the mean of the distribution, but not the median.

2. *dispersion*: How widely spread are the values from the central value? For "normal" variation, dispersion should not be inflated by outliers.

3. *skewness*: How "unbalanced" is the distribution on both sides of the central value? Skewness can be exaggerated by outliers.

Central Tendency and Skewness

Given a sorting algorithm and an array size, we want to estimate the center of the distribution of "normal" sort times. This distribution does not

include outliers. Our main statistic is the *trimmed mean*.

We must decide which scores to "trim" from the sample of sort times. We want to trim enough values so that the trimmed mean approaches the median and is not influenced by extreme values.

In Table 2, we present several trimmed mean candidates and compare them to the median. The data is from the sample of sort times described in Table 1. The untrimmed mean is based on the entire sample, including outliers. The 99/01 trimmed mean removes the largest and smallest 1% (approximately) of the sample before calculating the mean. Other trimmed means remove the top and bottom 5%, 10%, and 20% of the sample. The median can be interpreted as the mean obtained by removing the largest and smallest 50%, but leaving the middle score(s).

Table 2: Selection Sort Trimmed Means.
Size N = 100, Repetitions R = 1000

Trim Percent	Mean	vs. Median
Untrimmed	15367	439
99/01	15048	120
95/05	15053	125
90/10	15053	125
80/20	15099	171
50/50 (Median)	14928	-0-

Note that the median remains unchanged for all trimmed samples because we removed the same number of values from both ends of the sorted list of values. For this sample of data, removing the top and bottom 1% seems to be sufficient to remove the effect of outliers on the mean.

Dispersion
The main topic of interest in this research is patterns of variation in algorithm performance. The dispersion in the distribution of sample sort times provides a measure for performance variation. We want to determine the variation for the "normal" sort times, apart from outlier effects.

The most common measure of variation for quantitative variables is the standard deviation. However, the standard deviation is very sensitive to outliers.

As with trimmed means, we calculate standard deviations from trimmed samples, hopefully with outliers removed. Since we are not testing for individual outliers, we trim different percentages of larger and smaller values from the sample.

Standard deviations, both untrimmed and trimmed, are presented in Table 3. The sample data is again from Table 1.

Table 3: Trimmed Standard Deviations.
Size N = 100, Repetitions R = 1000

Trim Percent	Std Devn
Untrimmed	5318
99/01	311
95/05	273
90/10	273
80/20	225
Quartile Deviation	233

It is apparent that trimming the top 1% (containing the outliers) and bottom 1% leads to a substantial reduction in the standard deviation. Additional trimming has relatively little effect on the standard deviation in this case.

The *quartile deviation* is included in Table 3 for comparison purposes. The *interquartile range* (IQR) is a well-known measure of the spread of scores in a distribution. It is defined to be difference between the third quartile Q3 (75th centile) and the first quartile Q1 (25th centile). The quartile deviation is *half* the interquartile range (IRQ/2).

Higher Repetitions
The data from Table 1 represents a sample of 1000 sort times. In the early development of our methodology, we generated samples of this size for array sizes between 100 and 1000. We performed statistical analyses on data for these sample sizes.

As we became more comfortable with our methodology, we increased the number of repetitions to 10000. Each time we ran our Java data generation program, we obtained a sorted array containing 10000 execution times. With larger samples, we got a clearer picture of the stability of our results.

In Table 4, we present a frequency distribution for one sample of 10000 sort times, based on selection sort of arrays of size 100. This

distribution of 10000 values is similar to the previous distribution of 1000 values.

Table 4: Selection Sort Distribution.
Size N = 100, Repetitions R = 10000

SortTime	Freq	CumFreq	Diff
* 13995	58	58	---
* 14461	2295	2353	466
* 14928	5669	8222	467
* 15394	1655	9877	466
* 15861	84	9961	467
other	38	9999	---
1866018	1	10000	---

* Consecutive values combined
(e.g. 13995 -- 55, 13996 -- 3)

1. The sample of sort times contains thousands of repeat values.

2. The smaller sort times appear in "pairs" that differ by 1 nanosecond (shown with asterisks). The lowest five pairs comprise over 99% of the distribution. Perhaps we need a better "clock" than the one provided by Java's nanoTime method.

3. The minimum value of 13995 is one clock tick below 14461, which was the minimum value in the smaller sample. The maximum value of 1866018 is an order of magnitude larger than the earlier maximum of 113827. In our methodology, generating random data that include large sort times is not unusual.

4. The median of this second distribution remains at 14928, which is again close to the minimum value.

4. ANALYSIS OF DATA

In this section, we analyze performance variation for four sorting algorithms: selection sort, insertion sort, bubble sort, and quicksort. For each algorithm, we examine six array sizes: 200, 400, ... , 1200. Patterns of mean variation across array sizes for a given algorithm is comparable to order-of-growth models covered in algorithm textbooks.

We extend our research to describe sort time distributions *within* each algorithm/array size combination. We measured central tendency, dispersion, and skewness for these distributions. Each test case involved 10000 repetitions of one sorting algorithm for a single array size.

Sort Time Central Tendency
We measured central tendency with trimmed means and the median. Our early work with arrays of size 100 suggested that trimming the top and bottom 1% is sufficient to remove outliers. However, for larger array sizes, the amount of variation increases. We made a conservative decision to trim the top and bottom 5% of the scores from each distribution.

Trimmed means for all six array sizes for each sorting algorithm are listed in Table 5. All times are in nanoseconds.

Table 5: Sort Time Distribution - Trimmed 95/05 Mean

Size	Select	Insert	Bubble	Quick
200	49979	21306	86374	14991
400	177308	79163	313611	32643
600	378867	173847	677236	51110
800	654887	304813	1118399	70304
1000	1004657	471508	1698413	89862
1200	1427205	674708	2415186	109896

Looking at each row separately, we see that the largest mean execution times are for bubble sort, followed by selection sort. Insertion sort are less than half the values for selection sort. Quicksort times are much smaller, especially for large array sizes.

This computer generated data is consistent with the nature of each of these sorting algorithms. For random data, bubble sort performs a large number of comparisons and swaps, while insertion sort performs many comparisons and shifts. In selection sort, the number of comparison operations is almost constant, regardless of the values in the array. The insertion sort and bubble sort algorithms can terminate early, depending on how fully sorted the data are initially. Quicksort is fastest because of its recursive design.

If we look down each column at the pattern of increasing mean execution times, the results follow traditional order-of-growth models. For selection sort, when the array size doubles (e.g. 400 -> 800), the mean sort time is approximately four times larger (177308 -> 654887). This supports an $O(N^2)$ order-of-growth model. A similar pattern occurs for insertion sort and bubble sort. Quicksort displays a noticeably smaller growth rate.

We prepared a summary table containing untrimmed means, but do not include it in this paper. With sample sizes of 10000, removing the top and bottom 5% (presumably containing outliers) had relatively little effect on the means. The trimmed means are about 1% to 2% smaller than the untrimmed means. Correlations between trimmed and untrimmed means is above 0.999 for each algorithm. As we shall see, trimming has a much greater effect on measures of dispersion.

We provide in Table 6 the medians for the sort time distributions for each algorithm/array size combination.

Table 6: Sort Time Distribution - Median

Size	Select	Insert	Bubble	Quick
200	49916	21459	86302	14928
400	177271	79305	313487	32655
600	378801	173539	677355	51314
800	654500	304627	1117730	70441
1000	1004374	471169	1698052	90034
1200	1426570	674566	2414594	110093

When the medians are compared to the trimmed means, there are minor differences, but the pattern is almost identical. This suggests that the trimming has successfully removed outliers, and the trimmed distributions are less skewed.

Sort Time Dispersion

We remind the reader that the values in the tables are not absolute. They are the results of random sampling of an algorithm. With means, the results are relatively stable, even in the presence of a small number of outliers.

The same claim cannot be made for measures of dispersion. Statistics such as the standard deviation and the range can be greatly distorted when even a few outliers are in the sample. Our main objective in this study is to characterize variation in sort time distributions. With judicious trimming, we can avoid the problem of having an unreasonable number of outliers. Even so, occasional bizarre values appeared in our data sets.

The most common measure of dispersion for a distribution is the standard deviation. To illustrate how volatile standard deviations can be with outliers, in Table 7 we present *untrimmed* standard deviations using complete samples of 10000 sort times.

In this table, untrimmed standard deviations for selection sort range in value from 6182 to 201238. Observe that increasing the array size does not always result in a larger standard deviation. The size of each standard deviation is heavily influenced by outliers. Similar irregular patterns occur for each sorting algorithm.

Table 7: Sort Time Distribution - Untrimmed Standard Deviation

Size	Select	Insert	Bubble	Quick
200	6182	10747	30301	2088
400	74580	39323	55776	3226
600	62073	34151	37991	20779
800	201238	104351	68866	36104
1000	57680	73650	43354	52899
1200	188333	32091	64344	24352

In the next table, we show how volatile variation statistics can be "tamed" with the careful use of trimming. Table 8 lists trimmed standard deviations obtained by removing the 5% largest and 5% smallest values from the sample. We chose 5% limits to be consistent with the previous trimming of means. In practice, 5% trimming might not always be enough.

Table 8: Sort Time Distribution - Trimmed 95/05 Standard Deviation

Size	Select	Insert	Bubble	Quick
200	558	761	1118	326
400	1243	2127	2838	494
600	1496	3782	3976	632
800	1712	6091	5559	794
1000	2141	8421	7520	955
1200	2852	10123	10096	1155

For the trimmed standard deviations in Table 8, the pattern in each column shows an increase in dispersion as the array size increases. These results are representative of what we *usually* obtained with 5% trimming.

The variation patterns for the four sorting algorithms is instructive. The greatest rates of increase in dispersion are for insertion sort and bubble sort. The smallest rate of increase is for quicksort.

Selection sort, as we showed in Table 5, has the second largest mean sort times. But the rate of increase in dispersion is less than for insertion and bubble sort. Why? We let the reader answer that question. Quicksort has a lower rate of increase in dispersion than selection sort.

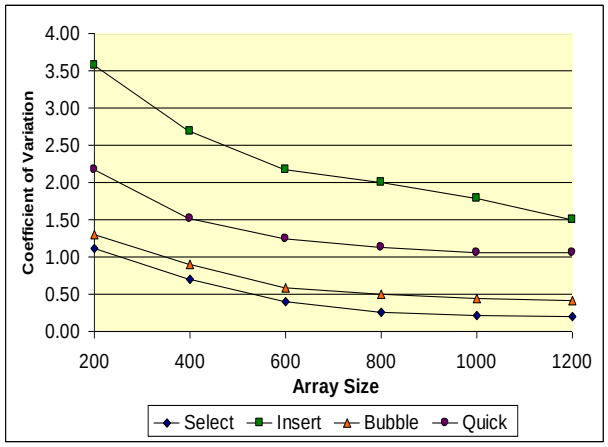

Figure 1: Sort Time Relative Variation - 95/05 Coefficient of Variation (%)

Coefficient of Variation

Another way of comparing dispersion among similar distributions is by measuring relative variation. In this case, we divide the trimmed standard deviation by the corresponding trimmed mean. The statistic is called the *coefficient of variation*. To make the value of the statistic easier to interpret, we multiplied it by 100, so that we express the standard deviation as a *percentage* of the mean.

Measures of relative variation for our sorting algorithms and array sizes are displayed in Figure 1. Both means and standard deviations are trimmed at the top and bottom 5% levels.

Selection sort has the smallest values for the coefficient of variation, followed closely by bubble sort. The selection sort means are more than twice as large as the times for insertion sort, but selection sort standard deviations are smaller. The result is less relative variation for selection sort.

Quicksort has smaller standard deviations and smaller means. The ratios fall in between the high and low values of the other algorithms. One interesting feature revealed by Figure 1 is that, for all four algorithms, the coefficient of variation *decreases* as the array size increases. Although the standard deviation increases for larger arrays, the mean increases at a faster rate.

It is tempting to conjecture that the ratios approach a lower limit for very large arrays. That is a question for future research.

In any case, the fact that the relative variation is small for large arrays might justify the emphasis on mean execution times in textbooks. Sort time variation could be viewed as less important for large arrays.

Sort Time Skewness

Throughout our research, we used the difference between the mean and median as a crude measure of skewness. One criteria for choosing a trim level for the sort time distributions was based on this difference being small. A comparison of the 95/05 trimmed means in Table 5 with the medians in Table 6 shows the closeness of each mean to the corresponding median. This indicates that the skewness in the trimmed distributions is relatively minor.

Our decision for the recommended amount of trimming was guided more by its effect on the standard deviation. Trimming the top and bottom 1% would be satisfactory to remove the skewness effects due to outliers. However, standard deviations are more affected by outliers, so we chose to trim 5% from the top and bottom of samples. This often led to a ten-fold reduction in the sample standard deviation.

Unexplained Variation

In our research design, we generated separate execution time distributions for specific sorting algorithm and array size combinations. The variation within these distributions was assumed to consist of a "normal" component and outliers.

We assumed that the normal component of variation would be due primarily to the randomness of the data. Measurement of this source of variation was not very accurate because of the granularity of the Java nanoTime clock. A clock increment of 466.5 nanoseconds is almost half of a microsecond. With the speed of the processor (GHZ), much of the effect of random data on algorithm performance is hidden within these 0.466 microsecond intervals.

The most puzzling aspect of our performance measurement was the frequent appearance of outliers. Outliers can have multiple causes. In our study, the "chief suspect" is the Java runtime environment. This software performs various actions to improve the performance of a running program. The feature most relevant seems to be Java's *automatic garbage collection* (Boyer, 2008; Wicht, 2011).

At various points during the execution of a program, the Java runtime chooses to free memory that is currently unreferenced. Generally, this is considered a good thing. However, automatic garbage collection makes it difficult to benchmark program performance.

The simple solution for running benchmark programs with Java would be to turn off Java's garbage collection feature. That is not an option. Our solution is to remove outliers from our sample. Garbage collection takes varying amounts of time. In our samples, the largest times were often 10 to 100 times larger than normal sample values.

5. SUMMARY AND CONCLUSIONS

The primary purpose of this study was to analyze variation in the performance of sorting algorithms written in Java. Most of the emphasis in algorithm textbooks is on average and worst case performance. We are more interested in the *distribution* of execution times when an algorithm is run multiple times.

We designed a methodology to control hardware, operating system, and Java runtime effects. We wanted processing time variation to result primarily from the sorting algorithm selected, the size of the array, and the randomness of the data. We wrote a Java test program to repeatedly fill an array, sort it, and record and save the execution times. The execution time data was then used to calculate statistics that summarize the distribution in terms of central tendency, dispersion, and skewness.

Our experiment was performed for four sorting algorithms: selection sort, insertion sort, bubble sort, and quicksort. For each algorithm, a range of array sizes were examined. A number of results were reported, including the following:

1. Execution time distributions were discrete, with relatively few distinct values. This was primarily due to the limited resolution of the Java nanoTime function.

2. Distributions were positively skewed and included a few very large outliers. As a result, samples had to be trimmed to remove outliers before calculating statistics.

3. For all sorting algorithms, the mean sort time increased as the array size increased. This was expected. The differing observed rates of increase were consistent with well-known order-of-growth models for the algorithms.

4. For each sorting algorithm, the standard deviation of execution times increased with array size. The algorithms differed in the amount variation and the pattern of growth. These patterns can be explained in terms of the structure of each algorithm.

5. For each algorithm, the standard deviation grew at a slower rate than the mean. This was demonstrated by a decreasing coefficient of variation as the array size grew larger.

Three conclusions can be drawn from our results. *First*, sort time variation exists and may be an important factor in systems with real-time constraints. *Second*, sort time variation is less important for very large arrays because the amount of variation is small compared to the mean. *Third*, beware of outliers in the data, especially when using the Java runtime environment for benchmarks.

Future Research

A good research study generates more questions than it answers. That was true in this study. Our planned future research activities include:

1. Extend our analysis of variation to other sorting algorithms, such as merge sort and shell sort.

2. Use our methodology on algorithms written in other programming languages. An obvious next language is C++. One problem is that C++ provides different timer functions in different operating environments.

3. Study the behavior of Java's nanoTime function in different hardware and software environments. The statement by Oracle that nanoTime provides the "most precise available system timer" is intriguing and suggests a number of practical questions for further research.

6. REFERENCES

Boyer, Brent (2008). Robust Java benchmarking, Part 1: Issues. IBM DeveloperWorks.

Cormen, Thomas H., Leiserson, Charles E., Rivest, Ronald L., & Stein, Clifford (2009). *Introduction to Algorithms* (3rd ed). MIT Press.

Jones, Nigel (2009). Sorting (in) embedded systems. Stack Overflow.

Koffman, Elliot, & Wolfgang, Paul (2010). *Data Structures: Abstraction and Design Using Java* (2nd ed). Wiley.

Lafore, Robert (2003). *Data Structures and Algorithms in Java* (2nd ed). Sams Publishing.

Lewis, John, & Loftus, William (2011). *Java Software Solutions, Foundations of Program Design* (7th ed). Addison-Wesley.

Liang, Y. Daniel (2012). *Introduction to Java Programming* (9th ed). Prentice Hall.

Oracle (2014). java.lang Class System. www.docs.oracle.com

Salsburg, David (2001). *The Lady Tasting Tea*. W. H. Freeman.

Sedgewick, Robert, & Wayne, Kevin (2011). *Algorithms* (4th ed). Addison-Wesley.

Wicht, Baptiste (2011). Java Micro-Benchmarking: How to write correct benchmarks. www.javacodegeeks.com