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 **ISCAP,** Information Systems and Computing Academic Professionals. Publishing frequency is currently semi-annually. The first date of publication was 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. Special thanks to members of AITP-EDSIG who perform the editorial and review processes for JISAR.

JOURNAL OF
INFORMATION SYSTEMS APPLIED RESEARCH

Editors

Scott Hunsinger
Senior Editor
Appalachian State University

Thomas Janicki
Publisher
University of North Carolina Wilmington

2017 JISAR Editorial Board

Causes of cyberbullying in multi-player online gaming environments: Gamer perceptions

Jami Cotler
jcotler@siena.edu

Meg Fryling
mfryling@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

Abstract

Cyberbullying has received much attention in recent years due to a variety of resulting tragic events, including cyberbully victim suicides. However, research on cyberbullying in online gaming environments is relatively new and limited. Furthermore, existing research primarily focuses on young adolescents, leaving research gaps for analyzing cyberbullying among older gamers. This is particularly important since it is estimated that 68% of gamers are 18 years of age or older. Finally, causes of aggressive behavior, such as cyberbullying, are still unclear and cannot simply be linked to violent video games. Therefore, in order to develop strategies to mitigate cyberbullying in online gaming environments, we need to better understand what the biggest causes of cyberbullying are in this environment. This exploratory research investigates gamer perceptions regarding the causes of cyberbullying in online multi-player gaming environments. A survey was developed for this study and 936 respondents answered several open ended questions related to causes of cyberbullying in gaming environments. Content analysis of these questions revealed that gamers perceive the biggest causes are: anonymity, the cyberbully not seeing the real life effects of their behaviors, and no fear of punishment.

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

1. INTRODUCTION

Cyberbullying has received much attention in recent years due to a variety of resulting tragic events, including cyberbully victim suicides. Research in this problem space began in the 1990s as technology ownership and Internet access became more ubiquitous. Cyberbullying victimization and research has increased during the last decade as social networking use has skyrocketed (Beran & Li, 2005; Kowalski & Limber, 2007; Mesch, 2009; Ortega et al., 2009;

Patchin & Hinduja, 2006; Raskauskas & Stoltz, 2007; Ybarra, 2004). Early studies on cyberbullying focused primarily on prevalence in adolescent populations (Lenhart, 2010; Patchin & Hinduja, 2006; Yardi & Bruckman, 2011).

In recent years, cyberbullying research has extended to older populations (Aricak, 2009; Dilmac, 2009; Molluzzo et al., 2012; G. Rivituso, 2012; J. Rivituso, 2014; Smith & Yoon, 2012), including cyberbullying in college and in the workplace (Bond et al., 2010; Chapell et al., 2004; Cowie et al., 2002; De Cuyper et al., 2009; Keashly & Neuman, 2010; Lester, 2009; McKay et al., 2008; Privitera & Campbell, 2009). Additionally, cyberbullying in gaming environments has been investigated (Fryling et al., 2014, 2015). However, there is still limited research addressing the connections between bullying, cyberbullying, and gaming (Qing, 2015) so further investigation is warranted.

2. PSYCHOLOGICAL IMPACT OF CYBERBULLYING

It has been well established that cyberbullying triggers social problems and impacts victims negatively psychologically and emotionally (Blair, 2003; Fryling et al., 2014; Juvonen & Gross, 2008; Patchin & Hinduja, 2006; G. Rivituso, 2012; J. Rivituso, 2014). Cyberbullying may even more profoundly negatively impact its victims than traditional bullying because the details of such bullying behavior are publicly visible to a large audience for extended periods of time, allowing the victim to be re-victimized over and over (Campbell, 2005; G. Rivituso, 2012; J. Rivituso, 2014; Strom & Strom, 2005). Therefore, it is important for research to be conducted to offer a better understanding of why it happens and develop strategies to mitigate.

While theories vary as to the cause of cyberbullying (T. Anderson & Sturm, 2007) (Bandura, 1989, 1990; Diamanduros et al., 2008), it has been found that victimization not only has long-term negative psychological effects but it can cause victims to become cyberbullies themselves (Berthold & Hoover, 2000; Fryling & Rivituso, 2013; Katzer, 2009; Wong & Xio, 2012; Ybarra & Mitchell, 2004). Some research has found that young bullying victims are more than three times as likely to become a bully than individuals that have never been bullied themselves (Berthold & Hoover, 2000).

While cyberbullying research has focused primarily on adolescents, some researchers have begun to investigate cyberbullying in college and

in the workplace. Nonetheless, there is much still left to understand about adult cyberbullying behavior since there is much less research in this area (Lester, 2009). Organizations need to consider information systems that coworkers use in the workplace, including social media platforms (e.g. Facebook, Twitter), and the possibility of cyberbullying activity that may negatively impact employees.

3. CYBERBULLYING IN ONLINE GAMING ENVIRONMENTS

While there is little research on cyberbullying in gaming environments, there are many existing theories regarding video game violence and increased aggressive behavior. This section summarizes some of that research.

Some researchers believe that violence in video games has been shown to increase hostile behavior and decrease supportive behavior (C. A. Anderson & Bushman, 2001; C. A. Anderson et al., 2007; C. A. Anderson et al., 2010; Bushman & Anderson, 2002; Hasan et al., 2013; Power, 2009). Yang (2012) found an association between male adolescent online gamers and a preference for violent games, increased hostility, and aggressive behavior.

However, other researchers report that there is simply no concrete evidence of a causal relationship between violent video games and violent behavior (Ferguson, 2010; Ferguson & Kilburn, 2010; Przybylski et al., 2014; Sherry, 2007). Whether or not violence in video games is the cause of bad behavior continues to be hotly argued in the academic literature (Bushman et al., 2010; Ferguson & Dyck, 2012) and beyond (Valleskey, 2014).

The specific aggressive behavior of cyberbullying in online gaming environments has begun to be investigated by researchers (Fryling et al., 2015; Yang, 2012). Not only has cyberbullying been found to exist in online gaming environments but it can have negative psychological effects, similarly to other types of bullying (Fryling et al., 2015).

There is research to support that repeated online gaming cyberbullying victimization of male gamers increases the likelihood of observable aggressive behavior in their non-gaming lives (Yang, 2012). Lam et al. (2013) found that students that have been a cyberbully or a cyberbullying victim were twice as likely to have been exposed to violent video games.

Recent research has looked beyond a simple casual relationship between violent video games and aggression. For example, Przybylski et al. (2014) concluded that a lack of player competence in a game environment is a cause of aggressive behavior. While this research was not specifically on cyberbullying, there are opportunities to extend the concepts to this domain.

Other research on cyberbullying has identified sexual orientation and revenge as causes of victimization (Varjas et al., 2013). If victims are perceived as exhibiting bad behavior, bullies and bystanders may justify their cyberbullying victimization (Varjas et al., 2010). There is much room to explore additional causes of cyberbullying.

Research on cyberbullying in online gaming environments is relatively new and limited. Furthermore, existing research primarily focuses on young adolescents, leaving research gaps for analyzing cyberbullying among older gamers. This is particularly important since it is estimated that 68% of gamers are 18 years of age or older (MediaCT, 2013). Finally, causes of aggressive behavior, such as cyberbullying, are still unclear and cannot simply be linked to violent video games. Therefore, in order to develop cyberbullying mitigation strategies for online multi-player gaming environments, we need to better understand what the biggest causes of cyberbullying are in this somewhat unique and under-researched environment.

4. METHODOLOGY

This exploratory research investigates gamer perceptions regarding the causes of cyberbullying in online multi-player gaming environments.

A survey was developed by the researchers and it included questions from previous cyberbullying studies (Molluzzo et al., 2012; Smith & Yoon, 2012), in addition to questions specifically developed to address research objectives. Respondents were asked several open ended questions related to causes of cyberbullying in gaming environments.

The survey was pilot tested by a small group of online gamers prior to releasing and minor modifications were made based on feedback from the pilot testers. The final survey consisted of 42 questions, including demographics and several questions related to perceptions regarding potential cyberbullying mitigation strategies.

This paper presents the content analysis of an open-ended question which asked participants "What would you say are the biggest causes of cyberbullying within multi-player video games?" In addition, the results from one multi-answer question "Why do you think cyberbullying behavior within multi-player games occurs" containing nine categories are offered.

Population and Sample
The survey was distributed via an online gaming forum that hosted 564,166 total members and was completed by 1033 respondents. Of these respondents, 936 responded to the open-ended question "What would you say are the biggest causes of cyberbullying within multi-player video games?" and were used in the analysis presented in this paper. Participation was encouraged in two ways. First, forum members were presented with the survey link immediately upon login and respondents were offered a chance to win a $50 Amazon gift card. These two factors likely encouraged the high response rate.

The forum's gender distribution was approximately 62% female and 38% male. This distribution was similar to the sample used for this analysis (i.e. less than 1% difference). The average age of respondents was 22 with 73% of respondents reporting that they are 18 or older.

Results

Cause	Overall
Anonymity online	805 (86%)
The Cyberbully does not see real life effects on other players	713 (76%)
No fear of punishment	681 (73%)
Cyberbully crave attention	605 (65%)
No punishment for cyberbullying behavior available	524 (56%)
Cyberbully have stress anxiety and/or depression that is causing them to act out	432 (46%)
Cyberbully need to take out frustrations from being bullied themselves in real life	424 (45%)
Not enough methods to deal with it	285 (31%)
Too much freedom online	281 (30%)

Table 1: Results from multiple selection question "Why do you think cyberbullying behavior within multi-player games occurs?"

The results from the multi-answer question "Why do you think cyberbullying behavior within multi-player games occurs" presented in Table 1 show

that the highest percentage of participants reported anonymity (86%), the cyberbully not seeing the real life effects of their behaviors (76%) and no fear of punishment (73%) as causes of cyberbullying.

To supplement these findings a content analysis was performed on the 936 responses to the open ended question "What would you say are the biggest causes of cyberbullying within multi-player video games?" To start, two coders independently developed a code dictionary based on the literature and frequent word analysis of the survey responses and performed the initial coding of the a sample of 100 (11%) responses. This process yielded an 81% agreement between the coders. The coders then refined the coding dictionary (Appendix A) and completed another pass of 100 responses, this time yielding a 96% agreement, exceeding the suggested final intercoder agreement for qualitative data analysis of over 90% (Miles & Huberman, 1994). The rest of the responses were then systematically analyzed using NVivo 11 (QSR International) with the goal of gaining a deeper understanding of the participants' perception of the causes of cyberbullying specific to online gaming environments.

As shown in Table 2 (see Appendix A for full results), and somewhat similar to the multi-answer question, the most frequently cited reason for cyberbullying in online gaming environments is anonymity followed by personal character and then the lack of consequences. Twenty-four percent of the respondents mentioned anonymity in their response to this question. One participant reported that "Anonymity causes the illusion of no consequences creating (in the cyberbulliers opinion) a lawless environment where they can do whatever they want" while another participant expressed a deep concern surrounding anonymity citing that "I think it's because no one really knows who is who so the anonymous cyberbully can't really be identified. They don't really have anything to lose within a game." While others expressed concern by saying "The idea that you could be playing against anyone at any time is really quite frightening. The internet has grown so big that it is "unpoliceable" now." Other participants offered personal accounts of the impact of anonymity on cyberbullying and shared the following comments:

> I would say the biggest cause of cyberbullying within multi-player video games is the fact that most of the players who play online do not know each other in real life. I know this

for a fact, because my own relatives who play multi-player games are cyberbullies. They will ridicule and verbally abuse some person for not having as high of a score as them and place no boundaries as to what they say to them verbally, because they don't know the person so it doesn't matter to them. It's hard to convince them to stop, too, because they don't care whether they hurt someone they don't know (they don't feel it will affect their lives afterwards so they don't care).

> Lack of accountability in an anonymous environment. Someone actually said to me 'I don't have to be nice because it is online and doesn't count in real life'.

Eighteen percent of the respondents reported that it was the cyberbully's character that caused this behavior citing that "…lack of compassion, lack of understanding about others, selfishness" as well as "there are some people that just enjoy harassing others" as reasons for cyberbullying. Another participant said:

> Some people are just complete and utter inconsiderate innates who have not been taught properly, or rather have not learned properly on how to be respectful and polite to fellow people. These people troll and cyberbully often, perhaps, since they do not realize that it actually can hurt another person's feelings.

Sixteen percent of the respondants mentioned the lack of consequences as a reason for cyberbullying behavior in online gaming environments. Upon personal reflection one participant offered that:

> There is no accountability! I can call people names (or others can do the same to me), demeaning them for their sexual orientation, race, accent, GENDER, whatever. These are things I don't truly believe, but it doesn't matter because someone made me angry and I have no accountability to anyone on the internet. This is the truth I am ashamed of and want to commit to changing."

Others spoke of the lack of impact of not having any real consequences by stating "… the lack of 'real' consequences. Being banned from a game doesn't have the same impact that an arrest makes in real life for similar actions." Another participant shared that bad behavior is encouraged due to not having consequences by saying:

> There are no consequences. Cyberbullies are

free to do as they will, and they often are encouraged and receive praise from trolls. Victims are bullied more by the general trolls. Bad behavior is encouraged and applauded. It's been that way for years now.

Reason	Percentage Responding	Sample Quotes (Some excerpted)
Anonymity	248 (24%)	"The act of "trolling" is now the "in" thing." "The anonymous feeling people get makes them feel powerful." "I think people just get confident when their names and faces aren't attached to their words."
Character	169 (18%)	"People who ignore the rules of decency and courtesy online, and feel that since no one can "get them", they can get away with anything they want." "People just being jerks. Gaming and nerd culture is seriously jacked up right now in general."
No Consequences	147 (16%)	"People feel like they can do whatever they want because no one knows who they are." "People say

Reason	Percentage Responding	Sample Quotes (Some excerpted)
		things and do things they might not in real life- it's an outlet and there are no real consequences."

Table 2: Top three reasons cited for cyberbullying

5. DISCUSSION

Only 424 or 45% of participants in the multi-answer question and 31 (3%) of the participants responding to the short answer question support the findings of (Berthold & Hoover, 2000; Fryling & Rivituso, 2013; Katzer, 2009; Wong & Xio, 2012; Ybarra & Mitchell, 2004) in that being a cyberbully victim is a cause for bullying behavior in online gaming environment. However, 53 (6%) of the participants cited that the game design was a cause of cyberbullying which support the research of (C. A. Anderson & Bushman, 2001; C. A. Anderson et al., 2007; C. A. Anderson et al., 2010; Bushman & Anderson, 2002; Hasan et al., 2013; Power, 2009). Participants citing the game design and game company as a contributing factors of cyberbullying state that the games are designed in a way that encourages cyberbullying as a competitive advantage. Furthermore, one participant states that "The owners/programmers either allow for it, refuse to deal with complaints and in some cases engage in it themselves" and "Gaming companies are a big cause." Another study participant shared the perspective that having "sides" encourages cyberbullying by saying:

The implementation of "sides" in MMO's [Massively Multi-Player Online games] and open world combat was the stupidest thing they have ever done. In World of Warcraft, I am hunted down and killed repeatedly just because they can get a few points out of me for better items and gear. And then harassed if I call in my guild to back me up. If they got rid of the open world combat and being able to kill anyone who isn't on your "side" outside of pvp [player(s) versus player(s)] arenas. I quite WoW [World of Warcraft] because of being bullying by the "other side" and haven't gone back in fear of not being able to enjoy the game. I'm not a pvp person and you shouldn't have to be forced to pvp if you don't want to.

As previously discussed, Przbylski et al. (2014) found that the lack of play competence independent of the game design as a cause of aggressive behaviors. Expanding on this work, in this study 90 (10%) of the participants surveyed indicated that a player with a novice skill level attracted cyberbullying attention in gaming environments. Participants in this group cited that players classified as "NOOBS" along with newbie's are easy and frequent targets for cyberbullying. It should be noted here, that there is a distinction between the terms. NOOB's are characterized as "knowing little without any desire to learn more" (Urban Dictionary, 2016). However, a newbie is a beginner who is willing to learn and improve. One participant cited that "especially if they are new to the game they are generalized into the 'noob' category", while another participant shared that "If your new to the game, a PoC [Person of Color], or a sexuality other then straight, then you will most likely be bullied." Two repondents added "People who have more experience points pick on the newbies" and "The strong picking on the weak because they like that feeling of power." Finally, another respondent stated:

> The bully feels like they are hidden behind the internet, and they can take out their anger or feelings of superiority onto people they feel are "less worthy" or less experienced. Because advanced players can remain anonymous, they often bully weaker or newer players for the sake of fun and in-game rewards.

Limitations

All respondents were members of the Animal Crossing Community. While forum members report playing a wide variety of online games, it is hard to know if this population accurately represents the general online gaming population. The age and gender distributions of the Animal Crossings Community forum (62% female; Average age 22) is not the same as the online gaming populations (45% female; average age 30) (MediaCT, 2013). Additionally, as with all anonymous surveys, it is also unknown if respondents were completely truthful in their responses.

While there are a few generally accepted cyberbullying definition variations, the gaming environment is unique and may not align well with those definitions. Therefore, the respondents were asked to answer the survey questions based on what they perceive as cyberbullying in the gaming environment.

Future research will include additional analysis of the existing survey data to better understand the causes of cyberbullying in gaming environments, perceptions regarding what constitutes cyberbullying, and mitigation strategies. Further, the survey may be distributed via other channels. This may increase the diversity of respondents and reduce any unknown bias members of the Animal Crossing Community gaming forum may have.

6. CONCLUSIONS

This study sought to better understand the causes of cyberbullying in online gaming environments so that mitigation strategies can be developed and implemented. The analysis presented here revealed that gamers perceive the biggest causes are: anonymity, the cyberbully not seeing the real life effects of their behaviors, and no fear of punishment.

This study's findings extend existing research by exploring causes of cyberbullying in online gaming environments, including adult gamer populations. Future research can include developing strategies to help mitigate the three biggest causes of cyberbully, reported in this study. These mitigation strategies can include both technical implementations and policy enhancements. Once implemented in a gaming environment, cyberbullying activities can be re-evaluated to confirm the study's findings and access the value of these strategies.

7. ACKNOWLEDGEMENTS

8. REFERENCES

Anderson, C. A., & Bushman, B. J. (2001). Effects of Violent Video Games of Aggressive Behavior, Aggressive Cognition, Aggressive Affect, Physiological Arousal, and Prosocial Behavior: A Meta-Analytic Review of the Scientific Literature. Psychological Science, 12, 353-359.

Anderson, C. A., Gentile, D. A., & Buckley, K. E. (2007). Violent video game effects on children and adolescents: Theory, research, and public policy. New York, NY: Oxford University Press.

Anderson, C. A., Shibuya, A., Ihori, N., Swing, E. L., Bushman, B. J., Sakamoto, A., . . . Saleem, M. (2010). Violent video game effects on aggression, empathy, and prosocial behavior in Eastern and Western countries. Psychological Bulletin, 136, 151-173.

Anderson, T., & Sturm, B. (2007). Cyberbullying from playground to computer. Young Adult Library Services, 5(2), 24-27.

Aricak, T. (2009). Psychiatric symptomatology as a predictor of cyberbullying among university students. Eurasian Journal of Educational Research, 34, 167-184.

Bandura, A. (1989). Human Agency in Social Cognitive Theory. American Psychologist, 44(9), 1175-1184.

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

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(3), 256-277.

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

Blair, J. (2003). New breed of bullies torment their peers on the internet. Education Week, 22(21), 6.

Bond, S., Tuckey, M., & Dollard, M. (2010). Psychosocial Safety Climate, Workplace Bullying, and Symptoms of Posttraumatic Stress. Organization Development Journal, 28(1), 38-56.

Bushman, B. J., & Anderson, C. A. (2002). Violent Video Games and Hostile Expectations: A Test of the General Aggression Model. Personality & Social Psychology Bulletin, 28, 1679-1686.

Bushman, B. J., Rothstein, H. R., & Anderson, C. A. (2010). Much Ado About Something: Violent Video Game Effects and a School of Red Herring: Reply to Ferguson and Kilburn (2010). Psychological Bulletin.

Campbell, M. (2005). Cyber bullying: An Old Problem in a New Guise? Australian Journal Of Guidance & Counseling, 15(10), 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(153), 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(3), 206-224.

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(8), 693-704.

Dictionary, U. (2016). Retrieved from http://www.urbandictionary.com/define.php?term=Noob&defid=2568674

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

Ferguson, C. J. (2010). Blazing angels or resident evil? Can violent video games be a force for good? Review of General Psychology, 14(2), 68-81.

Ferguson, C. J., & Dyck, D. (2012). Paradigm change in aggression research: The time has come to retire the General Aggression Model. Aggression and Violent Behavior, 17, 220-228.

Ferguson, C. J., & Kilburn, J. (2010). Much ado about nothing: The misestimation and overinterpretation of violent video game effects in Eastern and Western nations: Comment on Anderson et al. (2010). Psychological Bulletin, 136(2), 174-178.

Fryling, M., Cotler, J., Rivituso, J., Mathews, L., & Pratico, S. (2014, November 6-9). Cyberbullying or normal game play? Impact of age, gender, and experience on perceptions regarding cyberbullying in multi-player online gaming environments. Paper presented at the Conference on Information Systems Applied Research (CONISAR), Baltimore, MD.

Fryling, M., Cotler, J., Rivituso, J., Mathews, L., & Pratico, S. (2015). Cyberbullying or normal game play? Impact of age, gender, and experience on perceptions regarding cyberbullying in multi-player online gaming environments. Journal of Information Systems Applied Research (JISAR), 8(1), 4-18.

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

Hasan, Y., Bègue, L., Scharkow, M., & Bushman, B. J. (2013). The more you play, the more aggressive you become: A long-term experimental study of cumulative violent video game effects on hostile expectations and aggressive behavior. Journal of Experimental Social Psychology, 49(2), 224-227. doi:http://dx.doi.org/10.1016/j.jesp.2012.10.016

Juvonen, J., & Gross, E. (2008). Extending the School Grounds?-Bullying Experiences in Cyberspace. Journal of School Health, 78(9), 496-505.

Katzer, C., Fetchenhauer, D., & Belschak, F. (2009). Cyberbullying: Who Are the Victims? Journal of Media Psychology, 2(1), 25-36.

Keashly, L., & Neuman, J. (2010). Faculty Experiences with Bullying in Higher Education. Administrative Theory & Praxis, 32(1), 48-70.

Kowalski, R. M., & Limber, S. P. (2007). Electronic Bullying Among Middle School Students. Journal of Adolescent Health, 41(6, Supplement 1), S22-S30. doi:10.1016/j.jadohealth.2007.08.017

Lam, L., Cheng, Z., & Liu, X. (2013). Violent Online Games Exposure and Cyberbullying/Victimization Among Adolescents. CyberPsychology, Behavior & Social Networking, 16(3), 159-165.

Lenhart, A. (2010). Cyberbullying 2010: What the research tells us Retrieved from http://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(5), 444-462. doi:10.1080/10668920902728394

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

MediaCT, I. (2013). Essential Facts About the Computer and Video Game Industry (Online). Retrieved from http://www.theesa.com/facts/pdfs/esa_ef_2013.pdf

Mesch, G. (2009). Parental Mediation, Online Activities, and Cyberbullying. CyberPsychology & Behavior, 12(4), 387-393.

Miles, M. B., & Huberman, A. M. (1994). Qualitative data analysis: An expanded sourcebook. Beverly Hills: Sage Publications.

Molluzzo, J. C., Lawler, J., & Manneh, J. (2012). A Comprehensive Survey on Cyberbullying Perceptions at a Major Metropolitan University – Faculty Perspectives. Paper presented at the Information Systems Educators Conference, New Orleans, LA.

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(4), 197-204.

Patchin, J., & Hinduja, S. (2006). Bullies Move Beyond the School Yard: A Preliminary Look at Cyberbullying: Sage Publications Inc.

Power, M. R. (2009). Video Games and a culture of conflict. Journal of Children and Media, 3, 90-94.

Privitera, C., & Campbell, M. (2009). Cyberbullying: The New Face of Workplace Bullying? CyberPsychology & Behavior, 12(4), 395-400.

Przybylski, A. K., Deci, E. L., Rigby, C. S., & Ryan, R. M. (2014). Competence-Impeding Electronic Games and Players' Aggressive Feelings, Thoughts, and Behavior. Journal of Personality and Social Psychology, 106(3), 441-457.

Qing, L. (2015). When Cyberbullying and Bullying Meet Gaming: A systemic Review of the Literature. Psychology & Psychotherapy, 5(4).

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

Rivituso, G. (2012). An Exploration of the Lived Experiences and the Psychological Impact of Cyberbullying Victimization Among College Students: An Interpretive Phenomenological Analysis. (Ed.D.), 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), 24(4), 71-75.

Sherry, J. L. (2007). Violent video games and aggression: Why can't we find effects? . In R. W. Preiss, B. M. Gayle, N. Burrell, M. Allen, & J. Bryant (Eds.), Mass media effects research: Advances through meta-analysis (Vol. xii, pp. 245-262). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Smith, J. A., & Yoon, J. (2012). Cyberbullying Presence, Extent, & Forms in a Midwestern Post-secondary Institution. Paper presented at the Information Systems Educators Conference, New Orleans, LA.

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

Valleskey, B. (2014). Aggression In Video Games Not Caused By Violent Content: Benzinga.com.

Varjas, K., Meyers, J., Kiperman, S., & Howard, A. (2013). Technology Hurts? Lesbian, Gay, and Bisexual Youth Perspectives of Technology and Cyberbullying. Journal of School Violence, 12(1), 27-44.

Varjas, K., Talley, J., Meyers, J., Parris, L., & Cutts, H. (2010). High school students' perceptions of motivations for cyberbullying: An exploratory study. Western Journal of Emergency Medicine, XI, 269-273.

Wong, Y. M., & Xio, B. S. (2012). An Empirical Investigation of Factors Instigating, Impelling, and Inhibiting Cyber-Bullying Behavior. Paper presented at the AMCIS 2012, Seattle, WA.

Yang, S. C. (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.

Yardi, S., & Bruckman, A. (2011). Social and technical challenges in parenting teens' social media use. Paper presented at the Proceedings of the 2011 annual conference on Human factors in computing systems, Vancouver, BC, Canada.

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.

Appendix A: Coding Dictionary

Percentage	Code	Description
What are the biggest causes of cyberbullying within multi-player video games?		
45 (5%)	accepted behavior	This is accepted behavior, bullying in gaming environments has become the norm.
44 (5%)	anger	The player exhibits cyberbullying behavior because of anger.
248 (26%)	anonymity	Cyberbullying is caused because the person's name is unknown.
4 (0%)	anonymity -> leading to feeling of untouchable	In situations where the cyberbullies name is unknown gives the feeling of being untouchable
22 (2%)	arrogance	The player exhibits cyberbullying behavior because of arrogance.
34 (4%)	attention seeking	The player exhibits cyberbullying behavior as a way to gain attention.
31 (3%)	being a victim of bullying	The player exhibits cyberbullying behavior because of being a victim.
10 (1%)	bias	The player exhibits cyberbullying behavior because of some sort of bias.
14 (1%)	bias/age	The player exhibits cyberbullying behavior has an age bias.
53 (6%)	bias/gender	The player exhibits cyberbullying behavior has a gender bias.
24 (3%)	bias/race	The player exhibits cyberbullying behavior has a race bias.
63 (7%)	boredom	The player exhibiting cyberbullying behavior is bored.
34 (4%)	can't see effects	The player exhibits cyberbullying behavior because of not seeing the effects on the person being bullied.
169 (18%)	character/ respect/ unethical	The persons' character, not caring or understanding the harm caused, or lack of respect of others is the cause of the bullying behavior.
69 (7%)	competition	The player exhibits cyberbullying behavior as a result of being competitive.
8 (1%)	confidence	The player exhibits cyberbullying behavior because of high levels of confidence.
6 (1%)	depression	The player exhibits cyberbullying behaviors because of depression.
27 (3%)	don't know	The participant responded with: I don't know
44 (5%)	frustration	The player exhibits cyberbullying behavior due to frustration.
53 (6%)	game design	The player exhibits cyberbullying behavior due to the design of the game.
3 (0%)	high self-esteem	The player exhibits cyberbullying behavior due to a high level of self-esteem.
8 (1%)	ignorance	The player exhibits cyberbullying behavior due to ignorance.
17 (2%)	lack of moderators	The player exhibits cyberbullying behavior due to lack of moderators.
4 (0%)	lack of self-control	The player exhibits cyberbullying behavior due to lack of self-control.
56 (6%)	low self esteem	The player exhibits cyberbullying behaviors due to low self-esteem.
69 (7%)	maturity	The player exhibits cyberbullying behavior due to a lack of maturity.
147 (16%)	no consequences	The player exhibits cyberbullying behavior due to lack of consequences.
6 (1%)	no one reporting	The player exhibits cyberbullying behavior due to no one reporting.

8 (1%)	not educated in what cyberbullying is	The player exhibits cyberbullying behavior due to lack of education of what cyberbullying is.
90 (10%)	novice player/NOOB	The player exhibits cyberbullying behavior due to the novice ability level of the player. The player exhibits cyberbullying behavior due to the novice ability of the player along with the lack of desire to learn.
5 (1%)	personality	The player exhibits cyberbullying behavior is cause by personality factors.
13 (1%)	personality -> aggressive	The player exhibits cyberbullying behavior due to aggression.
56 (6%)	power	The player exhibits cyberbullying behavior due to the desire for power.
7 (1%)	real life bully	The player exhibiting cyberbullying behavior is a real life bully.
121 (13%)	real life personal issues	The player exhibiting cyberbullying behavior has real life personal issues.
9 (1%)	too sensitive	The victims of cyberbullying are too sensitive.
11 (1%)	too serious	The reason players exhibiting cyberbullying behavior is due to taking the game too seriously.
47 (5%)	unsupervised	The reason there is cyberbullying is due to unsupervised minors.

An Interactive Toolbox
For Twitter Content Analytics

Musa Jafar
musa.jafar@manhattan.edu

Marc Waldman
marc.waldman@manhattan.edu

Manhattan College
Riverside, NY

Abstract

In this paper we present a simple and easy to use toolbox that can be used for social media content analytics in the world of Twitter. The toolbox was developed primarily for researchers with minimal computing background who wish to visually analyze the content of tweets (text and the associated metadata such as screen-names, hashtags, mentions, etc.) across the twitter-defined timeline or a user-specified timeline. The toolbox is open source and built on top of the R programming platform, R-Shiny and R-word cloud. The toolbox uses a word cloud approach to visualize both the metadata and the N-gram text sequences that make up the tweets collection (the tweets corpus). Filter mechanisms of the toolbox allow the researcher to control for the type and amount of data displayed in the associated word clouds – allowing for a finer resolution of analysis.

Keywords: Text-Analytics, Visual-Analytics, twitter, R-shiny, word-cloud, N-grams.

1. INTRODUCTION

Social media sites are a rich source of data for researchers and practitioners to analyze. This is especially true of Twitter as it provides a "real-time window into the opinions, hopes, beliefs, complaints and dreams of its users, and the insights that it aggregates can provide marketers, product developers, sales, digital journalists, sociologists, educators -- really, the entire enterprise -- with deep, rich and spontaneous feedback on virtually any topic" (Li et al. 2013).

Each tweet, although only 140 characters in length, has an associated collection of interesting metadata that includes the author's username (screen-name), timestamp, geo-location (if enabled), hashtags, retweet-count, favorite-count, etc. A researcher's dataset will typically consist of a large number of related tweets – typically related by hashtags, keywords or

authors. We will refer to this dataset of tweets as the tweets-corpus.

Typically acquiring the tweets to form the tweets-corpus is far from trivial. Twitter does not provide a non-programmatic mechanism to easily download the public tweets of any user (except your own tweets). A number of companies (Gnip, exporttweet.com, Twitonomy.com) do provide downloads of public tweets for a fee. However, a number of software tools are available that allow one to freely acquire tweets (within Twitter-specified limits). Tweets can also be captured by "screen-scraping" software or by utilizing the Twitter API – which provides access to tweets via a programming language such as Python or R. The Twitter Capture and Analysis Toolset (Borra et al. 2014) provides a freely available software distribution that can be used to capture tweets – however it requires users to have some familiarity with system administration concepts as the package is meant to be deployed in a Linux

environment. It also provides summaries in the form of pie charts and line graphs.

Once acquired, a tweets-corpus presents an analysis challenge to those individuals who are not comfortable with programming. Many freely-available software libraries are available to assist in the analysis of tweets (and other social media data and free-form text). However, many require the user to have extensive previous programming skills. We created our tweets-corpus analysis toolbox to help address this issue – allowing individuals without the programming background to visually analyze the corpus and associated metadata.

2. RELATED WORK

Analysis of twitter data is certainly not an unexplored topic. Many powerful tools have been developed to analyze social media data (including Twitter). Several of these tools require no programming background and provide rich insight into the data. However, this analysis is almost exclusively focused on the social media metadata – how often was your tweet retweeted, how many people are following you, how many people mentioned a specific hashtag, etc. To date, there has been little focus on easy-to-use tools for the analysis of tweets text - more specifically tools that allow a researcher to visualize the central themes of a set of tweets and how these themes evolved over time.

The paper by Zimmer and Proferes (Zimmer et al. 2014) provides an overview of how Twitter data is being used by researchers. It found that content analysis was the dominant form of analysis performed on tweets. Content analysis, as defined in their paper, is one "where text within a tweet was used in part of the analysis in some way". We envision our toolbox will be used mainly for content analysis but it also supports analysis of some of the metadata associated with the tweets corpus. Word clouds (sometimes also referred to as tag clouds) are the central visualization elements of our analysis tool.

Word Clouds
We use a word cloud approach (Viégas et al. 2007, 2008, 2009) to visualize the N-grams and the metadata of a tweets-corpus. An N-gram is defined as a continuous sequence of N-words in some block of text. For example, in the phrase "Mary had a little lamb" – there are five 1-grams (or sequences of one word) in the text. The word "Mary" is the first 1-gram and "lamb" is the fifth one-gram. During N-gram analysis "insignificant" words such as "the", "a" and "is" are frequently

dropped. "Mary had" and "little lamb" are among the more interesting 2-grams in the phrase.

There are numerous freely-available tools to generate word clouds. They vary from simple web-based tools such as wordle.net to sophisticated code libraries (e.g. the R-based word cloud package) that allow users to calibrate almost all aspects of cloud creation. We use the R-based word cloud in combination with RWeka and tm packages to facilitate the analysis beyond the 1-gram (Feinerer et al. 2008, 2015; Hornik, et al. 2009; Fellows 2014).

Word clouds, by their very nature, provide "big-picture" insight into a corpus of text. Words occurring with greater frequency are placed in a larger or more dominant font or color. Examples of word clouds used in our toolbox can be found in subsequent sections below. Almost all the word cloud tools we have examined operate solely at the 1-gram level – graphically illustrating the frequency of each word in a corpus. However, our toolbox supports word cloud visualization of larger N-grams. The toolbox provides user interface controls that allow the user to specify N-gram size (up-to 4-grams) with timeline and screen-name filter options. More information is provided in the implementation sections below.

Natural Language Processing
Although our toolbox is meant to assist the researcher to perform intelligent analysis on the tweets text we, as of yet, do not perform any form of sophisticated natural language processing on the text. This type of processing might take the form of sentiment analysis or other form of natural language understanding wherein the toolbox attempts to classify the text in some manner – for example, classify the kind of speech act intended by the creator of the tweet (Searle 1969) or labeling a tweet as either positive or negative toward a particular company, product, situation or an event.

Traditional theory, algorithms and tools that have been developed to analyze text from a corpus linguistics perspective (McEnry et al. 2012; Bird et al. 2009; Feinerer et al. 2008, 2015; Miller 1995) do not necessarily work well for social media based text. Social media text is typically grammatically incorrect, uses words that are not in the dictionary and embeds symbols, urls, hashtags, mentions and emoji(s). This makes it impractical to apply standard corpus linguistics algorithms and tools to analyze the content. (Maynard 2012) outlined many of the challenges that face social media text analytics.

Social Media Text Analytics

Social data is about the speech act itself, its background and its illocutionary effect. Until recently, except for the few that were documented and historically speaking, almost all speech acts went unnoticed. Those that have been archived are in forms and formats that are hard to access and subsequently hard to analyze. The internet and the underlying social media technologies however, provides us with platforms to express thoughts, say what is in our mind, make a comment, state a belief or express an opinion, an emotion or a desire. It also allowed for those thoughts to be captured and stored in digital formats that are retrievable, searchable, indexable, presentable and analyzable. At any given moment in time, those platforms allowed for the spawning of many social media networks and the forking of multiple communities.

Social media text bundles have a social active characteristic, are spur of the moment, do not easily conform to the traditional natural language processing rules, syntax and grammar. It uses language in a way not governed by traditional rules, it is free flowing, ambiguous and less rules bound. Social text is about the intention of the speaker, has a performative function and it is communication centric. In the case of Twitter, a tweet is created on the fly, it has a time component, a social aspect component, an intertextuality component and a para-textual component. When bundled together, a tweet collection becomes a corpus where each tweet is a rich document surrounded by a bundle of metadata (timestamp, hashtags, mentions, originality, status, geo-location, etc. within the bundle), the corpus is multi-dimensional and various aspects of it are analyzable (Ferragina et al. 2015; Metaxas et al. 2015).

3. THE CASE FOR VISUAL TEXT ANALYTICS

Well-designed visualizations are intuitive, insightful, hypothesis generating, help dispel myths, enable discoveries, emphasize a point of view or help discover patterns in almost every aspect of knowledge of our world. Terms like the "thinking eye" and the "seeing brain" date back to the Swiss-German artist Paul Klee and have been extensively used in the data visualization literature. The daunting challenge in social media text analytics is to make the content of a tweets-corpus visually available in a useful and presentable way for a researcher who is not a programmer, however expert in the domain content of the tweets-corpus. As John Tukey (Tukey 1977), the great statistician stated: "The greatest value of a picture is when it forces us to notice what we never expected to see".

Our toolbox allows an individual who is interested in performing ad-hoc analytics on a tweets-corpus to interactively and visually analyze it and its metadata, across multiple dimensions. We follow standard visualization principles. The user (browser-based) interface is built around the R-word cloud, R-Shiny package and its widgets (Chang et al. 2016), the computing engine (server back-end) is built around R, R text-analytics packages tm (Feinerer et al. 2008) and RWeka (Hornik et al. 2009).

As previously stated, we use a word cloud approach to visualize the N-grams and the metadata of a tweets-corpus. We apply timeline and metadata filters to allow the individual expert to get a better understanding of the context within which the content (tweets) was initially published. In our current implementation the word cloud itself is not yet interactive – you cannot directly manipulate and interact with the generated word cloud (drill down on an N-gram to view the corresponding tweets). This is a limitation of the R-word cloud package and the R-graphing system – which we may not use in future versions of the toolbox. However, we allow for the interactive control of the frequency ranges, N-gram count and font range. The toolbox allows the user to interactively and easily slice the content across a timeline and filter by the tweet author (screen-name).

Most forms of data analysis typically follow a three-phase process. First is the data collection phase wherein the necessary data is collected, possibly from multiple sources which can be a combination of API(s), screen-scrapping, purchase, etc. The second phase, which we will refer to as the data repurposing and cleaning phase, typically involves error checking and transforming the data into a format that can be cleanly loaded into the software being used for analysis. For example, the creation date and time of a tweet is usually in GMT(Greenwich Mean Time), converting it to the current locale may be necessary for meaningful analysis; hashtags and mentions are part of the body of the tweets, they may need to be extracted out for further analysis. The final phase is where the analysis actually takes place – data is typically filtered, summarized and visualized. Our toolbox is meant to assist in this third phase - to make the data analytics phase (of a tweets-corpus) generic, repeatable, interactive, intuitive and easy to use. The data collection and data repurposing phases are beyond the scope of this paper. The current implementation assumes that the tweets-corpus

has been created using the twitteR package (Gentry 2015). However, if the data has been acquired using other methods (Python-twitter API, purchased through a twitter subsidiary or screen-scraped from the twitter homepage) we do assume that the data has been converted into an R-dataframe format form that is compatible with twitteR reference 'status' class list (Gentry 2015). This format is shown in Figure 1 - the underlined attributes are the ones that are currently required by the toolbox, the hashtags and the mentions are extracted from the body of the tweet's text.

4. THE INTERACTIVE TOOLBOX

The retrieval, in-depth manipulation, text analytics, and presentation of the results require dexterity in programming which is lacking in most of us. Our main goal is to abstract the analytics phase and make it generic, repeatable, interactive, intuitive and easy to use. This allows a researcher from a non-computing discipline to achieve their goal of gaining an understanding, insight and knowledge of the content through iterative interactive analytics without needing to learn how to write or repurpose code. Allowing the none-algorithmic, non-programmer to visually analyze tweets-corpus across multiple dimensions.

Table 1 briefly summarizes some of the datasets (tweets-corpora) that were used during the development and testing (piloting) of the toolbox. In the following sections we present the various components of the toolbox and their functionality. All screenshots used in this paper were taken while utilizing the datasets in the previously mentioned table.

The User Interface
Upon startup, the toolbox presents the user with three tabs ("Load & View a Tweets Dataset", "N-Gram Analysis", and "Metadata Analysis") located near the top of the screen (Figures 2, 3, 4 and 5). The "Load & View a Tweets dataset" tab allows a user to import a tweets-corpus. Once loaded, the tweets-corpus is displayed in tabular form and can be filtered (to allow selective viewing of tweets) based on word or regular-expression pattern matching. Figure 2 shows the result of the regular expression filter "^I.* pray" being applied to Pope Francis tweets-corpus. This regular expression is filtering for all tweets (in the corpus) that include the capital letter "I" at the beginning of a tweet followed by the word "pray" somewhere in the text. Note that this regular expression allows an arbitrary number of words to appear between letter "I" and the word "pray".

Filtering can also be done via the tweets' author (screen-name) and the date of the tweets. This can be seen in Figure 3 which shows the "Dump Stoli, Dump Russian Vodka" tweets-corpus being filtered on the screen-name "fakedansavage".

The tool's "N-Gram Analysis" tab provides access to interface elements that allow the user to visualize the text of the tweets-corpus using word clouds. Figure 4 shows a 3-gram-based word cloud of the "Dump Stoli, Dump Russian Vodka" tweets-corpus. Note that the user interface elements below the word cloud allow the user to control the N-gram level (1, 2, 3 or 4-gram), as well as limits on (1) "Token Frequency Range" to allow for the filtering-out of very high and very low frequencies, (2) the Maximum Number of N-gram Tokens displayed to control for which tweets data will be included in the word cloud, and (3) "Font Size Scale" to penalize very high and reward very low frequency tokens in the display. Filtering can be done on screen-name and/or tweets corpus slice of the date range. Additional elements (font size and token frequency) govern how the word cloud is rendered – the font size of the cloud elements and the number of elements appearing in the word cloud, the most frequent tokens are displayed with the highest "Font Size" on the scale.

Varying the N-gram level can provide significant insight into the tweets-corpus. Two, three and four-gram analysis provides a more contextual usage of the words in the tweets and therefore more insight into a thread of interrelated tweets. We have found that 2, 3 and 4-gram analysis is particularly useful in the analysis of the tweets of competing groups – two or more groups with opposing messages. The toolbox facilitates this visual comparison of multiple word clouds via the "WordCloud in a new window" selection box that is found directly above the timeline filter in the user interface (see Figures 4 & 5). A use of this feature is discussed in greater detail in the application section below.

The tool's "Metadata Analysis" tab provides access to the interface elements that allow the user to visualize the metadata (screen-names, mentions, hashtags and applications used) associated with the tweets-corpus. Figure 5 shows a metadata analysis of the Narendra Modi tweets-corpus. In this figure we see a word cloud corresponding to the individual's (screen-names) that are referenced (mentioned) in the tweets. The prime minister mentioned himself (@narendramodi) most frequently (206 times). The screen-name @pmoinidia is mentioned 82 times while @un is mentioned only 37 times. The screen-name metadata analysis gives the user of

the toolbox insight into the inner circles and the interest of the prime minister. The toolbox also allows for the analysis and the comparison of the hashtags, and underlying application used to emit the tweet (iPhone, Galaxy, Blackberry, the kind of twitter-app, etc.).

5. THE APPLICATIONS OF THE TOOLBOX

While developing and testing the toolbox we collaborated with a number of our colleagues (across multiple disciplines) to utilize the toolbox in academic research and in the classroom. In this section we briefly describe some of this work in the hope that it will inspire others to use the toolbox for similar or new areas of research and pedagogy.

Colleagues in the Management Department utilized the toolbox in the development of a crisis management study on how executives at Stoli Group USA handled a social media crisis that began in the summer of 2013 when "Dan Savage", an LGBT activist and sex advice columnist (https://twitter.com/fakedansavage), called for a boycott of Stolichnaya vodka because of its perceived Russian origins. The Russian Government was being sharply criticized for the passing of discriminatory anti-gay laws. Savage's Twitter messages were intended to show solidarity with, and to draw international attention to the plight of, the gay, lesbian, bi-sexual and transgender (LGBT) community in Russia. The tweets that comprised the case study were purchased directly from Twitter's subsidiary, Gnip (https://www.gnip.com/). The dataset provided access to all relevant historical tweets (those containing hashtags such as #dumpstoli and #dumprussianvodka) over a 40 day period between July 23 - September 1, 2013. Utilizing the toolbox our colleagues found several trends in the tweets - who the influential tweeters were; what was the sentiment of the related tweets (those utilizing the relevant hashtags) and the main geographic locations that the tweets were emanating from. Figure 6 shows a 3-gram word cloud comparison of the Stoli company tweets and Dan Savage tweets during the same study-related time period. Note the sharp difference in the message and tone in the 3-grams of the two competing voices. An analyst can also compare the content in Figure 6 to the overall content in Figure 4 for the same time period. It is clear that the Dan Savage's message is winning over the Stoli-team message.

During the Pope's recent (2015) visit to the United States a colleague in the Religion department utilized the 3-gram word clouds (of the Pope's tweets) generated by our toolbox to facilitate classroom discussions. After showing the word clouds (containing 3-gram phrases) to the students, the instructor found that "An individual phrase or, more often than not, a number of closely connected phrases led them [the students] to begin a conversation about a topic that was forming in their minds but they had not yet articulated up to that point."

Another colleague in the Religion department is utilizing the toolbox to visualize the tweets of India's Prime Minister Narendra Modi. This colleague is planning on having his students utilize the toolbox in a class on topics in contemporary religion and science.

An ad hoc Session with the ToolBox
While analyzing the Stoli-tweets, a colleague asked the question: What was Mr. Savage tweeting about the day before he published the "dumpstoli" "dumpRussianVodka" hashtags and Avatar? The 2013-07-23 tweets were not part of the Gnip dataset that we acquired, also when we acquired the dataset, our queries filtered on hashtags and not screen-name(s). Using the advanced search feature of twitter.com we captured these tweets. We then parsed the html and scrapped the tweet-id(s). Knowing the ids of the tweets and using the twitteR API we acquired the data and loaded it into the toolbox. The process provided us with an insight into Mr. Savage's interests at large and what triggered his interest in creating the "dumpStoli" and "dumpRussinaVodka" hashtags and avatar (Figure 3). Visualizing the 2013-07-23 tweets in the toolbox (Figure 7), allowed us to find answers to questions we did not previously propose to investigate. We were also able to compare the content with his tweets as they related to Stoli-case (Figure 4). We were able to track the announcement of the "dumpRussianVodka" and "dumpstoli" avatar, we were also able to find out where it all started, as it was part of Mr. Savage's timeline (Figure 3). On "2013-07-24 01-39-50" when he retweeted "it's time to call a BOYCOTT OF THE 2014 OLYMPICS IN RUSSIA…" from @BrianKentMusic, 12 hours later (2013-07-24 13:51:54) he announced his avatar. These are the types of capabilities we want to give to the investigator of the content to facilitate their analysis. To allow them to perform "Intelligence Analysis" through the "Search and Filter", "Schematize", "Build Case", "Tell Story", "Make a Presentation" (Pirolli et al. 2005).

Another ad hoc Session with the Tool
Another question that was raised: Has Pope Francis' message changed between Christmases? With our toolbox and using filters (Dec. 20-31st)

of 2013, 2014, and 2015 (Figure 8); the data provided an abundance of insight into the overarching themes. In 2015 messages about prosecution and suffering predominated as compared to the 2013 and 2014 messages.

The Implementation of the Toolbox
In designing the toolbox we paid a close attention to the "Sense making Model for Intelligence Analysis" (Pirolli et al. 1999, 2005; Card et al. 1999). A Model-View-Controller (MVC) pattern approach was used to build the toolbox. The user interface (the View), is built on top of R-Shiny, R-data table and R-word cloud packages. The R-word cloud package is used to render the N-grams of the text and the metadata. The back-end server (the Controller), is built on top of basic R, RWeka (Hornik et al. 2009), R-tm text mining package (Feinerer et al. 2013) and R-helper packages. The R-Shiny server (Chang et al. 2016) manages the reactivity of the user interface and encapsulates the set of utility functions needed to interface with the data (Model) to perform the text-analytics functions on the tweets-corpus. The Model encapsulates the data and the set of functions used to retrieve, query, clean and manipulate a tweets corpus.

6. FUTURE WORK

Work continues on improving the usability and the feature set of our text analytics toolbox. Our future plans include accepting a tweets-corpus in both CSV and JSON formats, enabling live querying of twitter so that a twitterer and their circle of friends information can be displayed for a given screen-name, incorporating a higher resolution view of the timeline to the hour and possibly minute or even the second level is a high priority, including user interface elements to allow the exclusion of selected screen-names from a tweets-corpus analysis and geo-location analysis where geo-location is provided. The R generated word cloud is currently not interactive. As part of the long term enhancement of the toolbox, we want to implement an interactive Web-GL, SVG-based word cloud wherein a researcher can drill down using the tokens displayed in the word cloud itself to further investigate the underlying content and corresponding tweets.

7. CONCLUSIONS

Twitter data is playing an increasing role in research conducted across a variety of fields. In 2010, the Library of Congress and Twitter signed an agreement for the Library of Congress to retrieve and store all of the public tweets and on an ongoing basis. Eventually the tweets will become available "in a comprehensive useful way" (Library of Congress 2013). However, it is not clear as to when and how this data trove will become available. This also emphasizes the need for new approaches and different analytics tools to be developed.

We have implemented an extensible toolbox for the visual intelligent analysis of a tweets-corpus. This toolbox makes the analysis process repeatable, and the results are replicable. It was primarily designed for individuals who typically possess no programming background such as social scientists and digital journalists. The tool's contributions include visual N-gram (1, 2, 3 and 4 -gram) analysis of a tweets-corpus in combination with filter mechanisms that can be used to refine the granular level of analysis. The toolbox is open source and freely available.

Finally, one common misunderstanding of the framework is that it just produces pretty word clouds. While the framework does indeed display aesthetically appealing word clouds, its power however is in the analytics back-end that enables the visualization of content. Social events fade away as fast as they come to life. This framework allows for the instant analytics of social media content within the context of Twitter. An article, an event, a comment that spurred a storm of Twitter-based reaction can be instantly analyzed and its background investigated by a digital journalist, a commentator, or an opinion article writer.   Our   toolbox   facilitates   the contextualization of the parts.

8. REFERENCES

Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python. *O'REILLY Media*.

Borra, E., & Rieder, B. (2014). Programmed method developing a toolset for capturing and analyzing tweets. *International Journal of Information Management* 66(3), 262-278.

Card, S., Mackinlay, J., & Shneiderman, B. (1999). Readings in Information Visualization: Using Vision to Think. (1st Ed.). *Morgan Kaufmann*.

Chang, W., Cheng, J., et al. (2016, January 12). Shiny: Web Application Framework for R. *R package version 0.12.2. CRAN R-Project:* Retrieved January 14, 2016 from https://cran.r-project.org/web/packages/shiny/index.html

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text Mining Infrastructure in R. *Journal of Statistical Software* 25(5), 1-54.

Feinerer, I., & Hornik, K. (2015, July 03). tm: Text Mining Package. *R package version 0.6*. *CRAN R-Project:* Retrieved January 2016 from https://cran.r-project.org/web/packages/tm/index.html

Fellows, I. (2014, June 13). wordcloud: Word Clouds. *R package version 2.5*. *CRAN R-Project:* Retrieved January 14, 2016 from https://cran.r-project.org/web/packages/wordcloud/index.html

Ferragina, P., & Santoro, F. (2015). On Analyzing Hashtags in Twitter. *9th International AAAI Conference on Weblogs and Social Media (ICWSM)*. AAAI, Oxford, UK, 110-119.

Gentry, J. (2015, July 29). twitteR: R Based Twitter Client. *R package version 1.1.9*. *CRAN R-Project:* Retrieved January 14, 2016 from https://cran.r-project.org/web/packages/twitteR/index.html

Hornik, K., Buchta, C., and Zeileis, A. (2009). Open-Source Machine Learning: R Meets Weka. *Computational Statistics* 24(2), 225-232.

Li, C., Etlinger, S., Live, R., Jones, A., et al. (2013). Twitter's IPO: An Analysis of Opportunities and Threats. Retrieved January 23, 2016 from http://www.altimetergroup.com/2013/10/twitters-ipo-an-analysis-of-opportunity-and-threats/

Library of Congress. (2013), January 31. "Update on the Twitter Archive at the Library of Congress," Retrieved January14, 2016 from https://www.loc.gov/today/pr/2013/files/twitter_report_2013jan.pdf

Maynard, D., Bontcheva, K. & Rout, D. (2012). Challenges in developing opinion mining tools for socialmedia. Proceedings of @ NLP can u tag# usergenerat-edcontent ?! Workshop at LREC 2012, Turkey.

McEnery, T., & Hardie, A. (2012). Corpus Linguistics. *Cambridge, UK: Cambridge University Press*.

Metaxas, P. T., Mustafaraj, E., Wong, K., et al. (2015). What do Retweets indicate? Results from User Survey and Meta-Review of Research. *International AAAI Conference on Weblogs and Social Media (ICWSM)*. AAAI, Oxford, UK, 658-661.

Miller, G. A. (1995). WordNet: A Lexical Database for English. *Communications of the ACM* 38(11), 39-41.

Pirolli, P., & Card, S. (1999). Information Foraging. *Psychological Review* (106), 643-657.

Pirolli, P., & Card, S. (2005). The Sensemaking Process and Leverage Points for Analyst Technology. *International Conference on Intelligence Analysis* MITR, McLean, VA, 1-6.

Searle, J. R. (1969). Speech Acts. An Essay in the Philosophy of Language. Cambridge: Cambridge University Press.

Tukey, J. (1977). Exploratory Data Analysis. Readings: *Adison Wesley*.

Viégas, F. B., Wattenberg, M. H., et al. (2007). Many Eyes: A Site for Visualization at Internet Scale. *IEEE Transactions on Visualization and Computer Graphics* 13(6), 1121-1128.

Viégas, F. B., & Martin Wattenberg, M. (2008). Tag Clouds and the Case for Vernacular Visualization. *ACM Interactions* 15(4), 49-52.

Viégas, F. B., Wattenberg, M., & Feinberg, J. (2009). Participatory Visualization with Wordle. *IEEE Transactions on Visualization and Computer Graphics* 15(6), 1137-1144.

Zimmer, M., & Proferes, J. N. (2014). A topology of Twitter research: disciplines, methods, and ethics. Journal of Information Management 66(3), 250-261.

Editor's Note:

This paper was selected for inclusion in the journal as a CONISAR 2016 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 2016.

Appendices and Annexures

Anatomy of a Tweet

```
$ :Reference class 'status' [package "twitteR"] with 17 fields
 ..$ text         : chr "I had 15,000 people in Phoenix but @politico said \"the rooms capacity is
                     just over 2000.\" But said Bernie  Sanders had 11,000"| __truncated__
 ..$ favorited    : logi FALSE
 ..$ favoriteCount: num 3239
 ..$ replyToSN    : chr(0)
 ..$ created      : POSIXct[1:1], format: "2015-08-22 20:13:21"
 ..$ truncated    : logi FALSE
 ..$ replyToSID   : chr(0)
 ..$ id           : chr "635182993334792193"
 ..$ replyToUID   : chr(0)
 ..$ statusSource : chr "<a href=\"http://twitter.com/download/android\" rel=\"nofollow\">
                     Twitter for Android</a>"
 ..$ screenName   : chr "realDonaldTrump"
 ..$ retweetCount : num 1970
 ..$ isRetweet    : logi FALSE
 ..$ retweeted    : logi FALSE
 ..$ longitude    : chr "-74.6971706"
 ..$ latitude     : chr "40.6546378"
 ..$ urls         :'data.frame': 0 obs. of  4 variables:
 .. ..$ url        : chr(0)
 .. ..$ expanded_url: chr(0)
 .. ..$ dispaly_url : chr(0)
 .. ..$ indices    : num(0)
```

Figure 1: The structure of a twitteR tweet

Tweet Subject	Tweet Count	Timeline	Acquisition Method	Comments
Pope Francis	795	2013-03-17→ 2015-12-31	twitteR API	Pope Francis tweets
Narendra Modi	4,034	2014-12-10→ 2015-12-31	twitteR API	Prime Minister Modi tweets
John Stewart	3,010	2015-08-08→ 2015-08-09	twitteR API	Last Night of the Daily Show
Elton John	4,099	2015-03-16	twitteR API	Dolce & Gabbana Feud
Keith Olberman	1,238	2015-02-17→ 2015-02-24	twitteR API	Feud with Penn State & suspension by ESPN
Saudi Leaks	5,671	2015-6-20	twitteR API	Comments on Saudi leaks
Dump Stoli, Dump Russian Vodka	53,954	2013-07-23→ 2013-09-01	Purchased+ Python +R+ ScreenScrapping	Tweets were Purchased from GNIP for a case study
Sarah Palin	16,347	2009-11-19→ 2015-09-11	twitteR API	Mrs. Palin timeline & Palin related tweets.
Donald Trump	30,264	2015-06-28→ 2015-09-10	twitteR API	Mr. Trump's timeline & Trump related tweets
1600 Pennsylvania Ave Washington DC	16,061	2016-01-01→ 2016-01-13	Google Maps +Twitter API(s)	We used Google Maps API & Twitter geo-location tags

Table 1: Tweets-corpora used in testing the toolbox

Figure 2: Load & View a Tweets dataset tab: Pope Francis tweets, with a regular expression filter
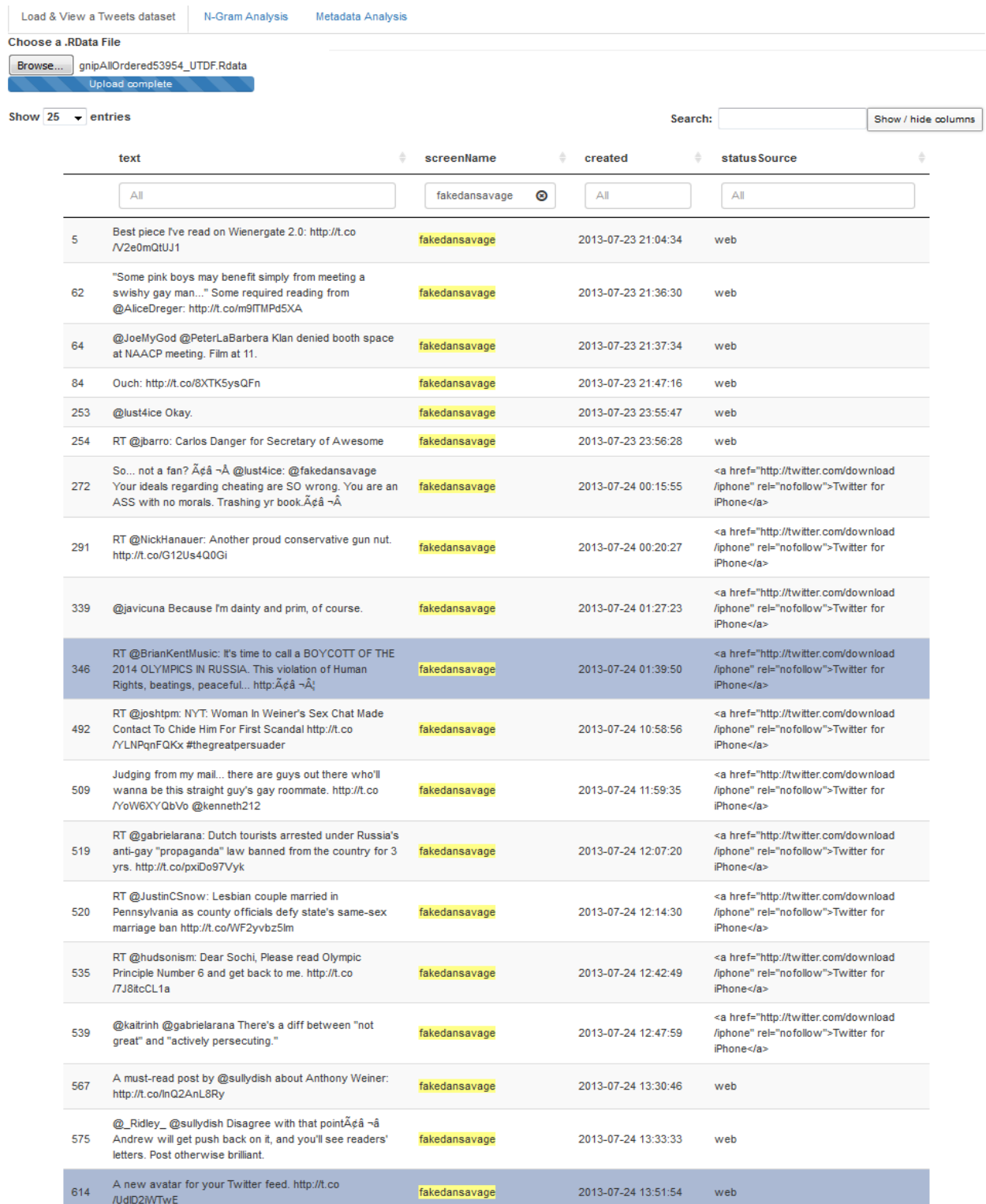
Figure 3: Mr. Savage's timeline when dumpstoli, dumpRussianVodka avatar was tweeted
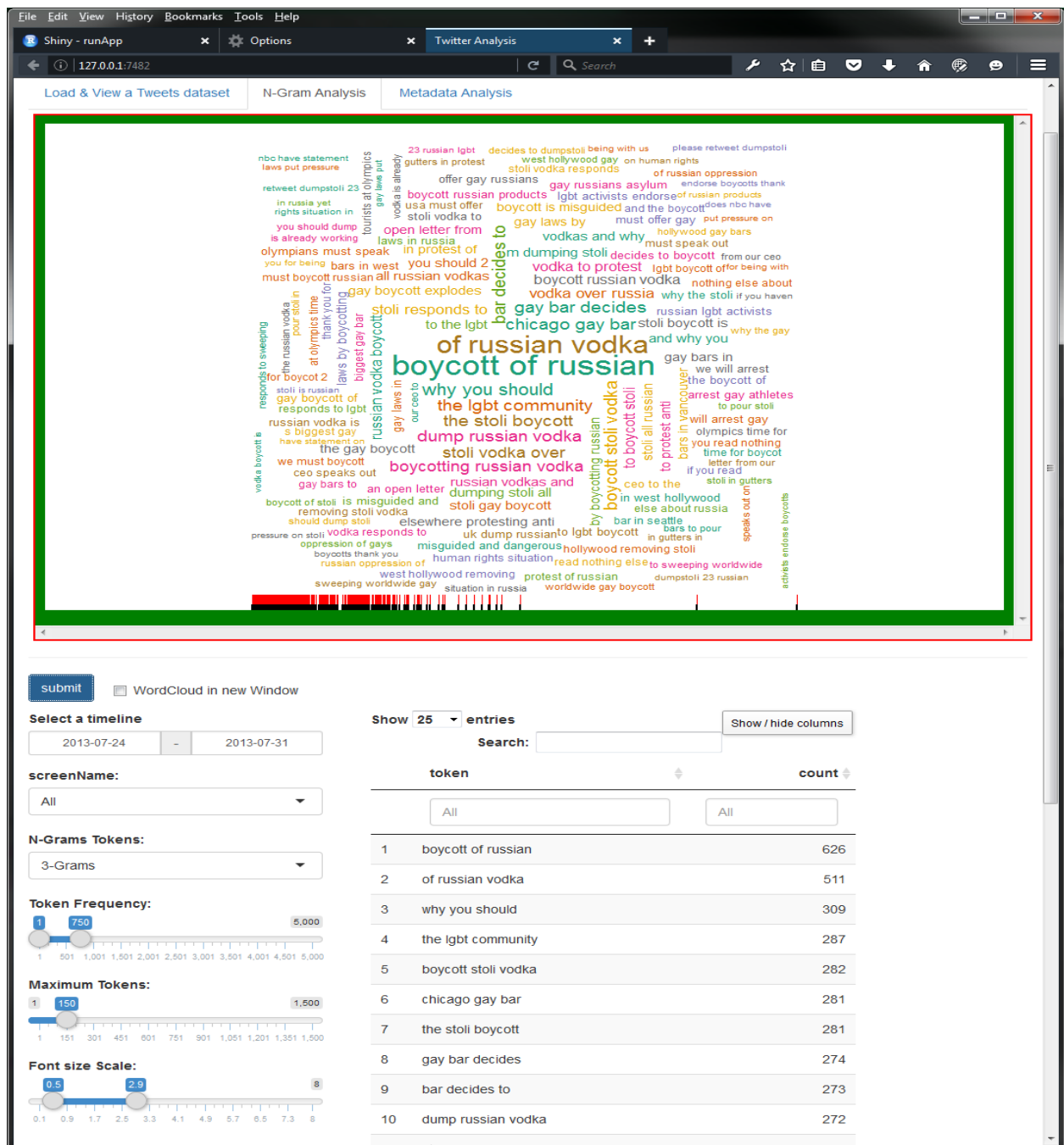
Figure 4: N-grams Analysis Tab: View all 3-grams of text in the dumpstoli tweets dataset

Figure 5: Metadata Analysis tab: P.M. Narendra Modi mentions

Stoli Team Tweets

Dan Savages Tweets
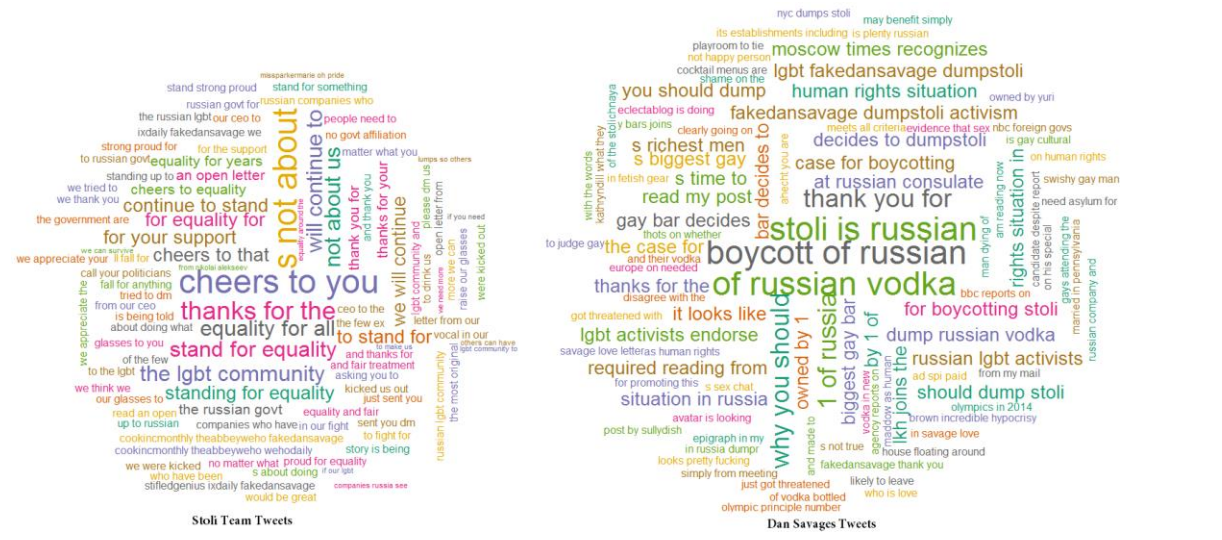
Figure 6: Comparing 3-grams of relevant tweets between Jul-24 & Jul-31 of 2013

Figure 7: 3-gram tweets of Mr. Savage on 2013-07-23

To preserve space, we only displayed the word clouds
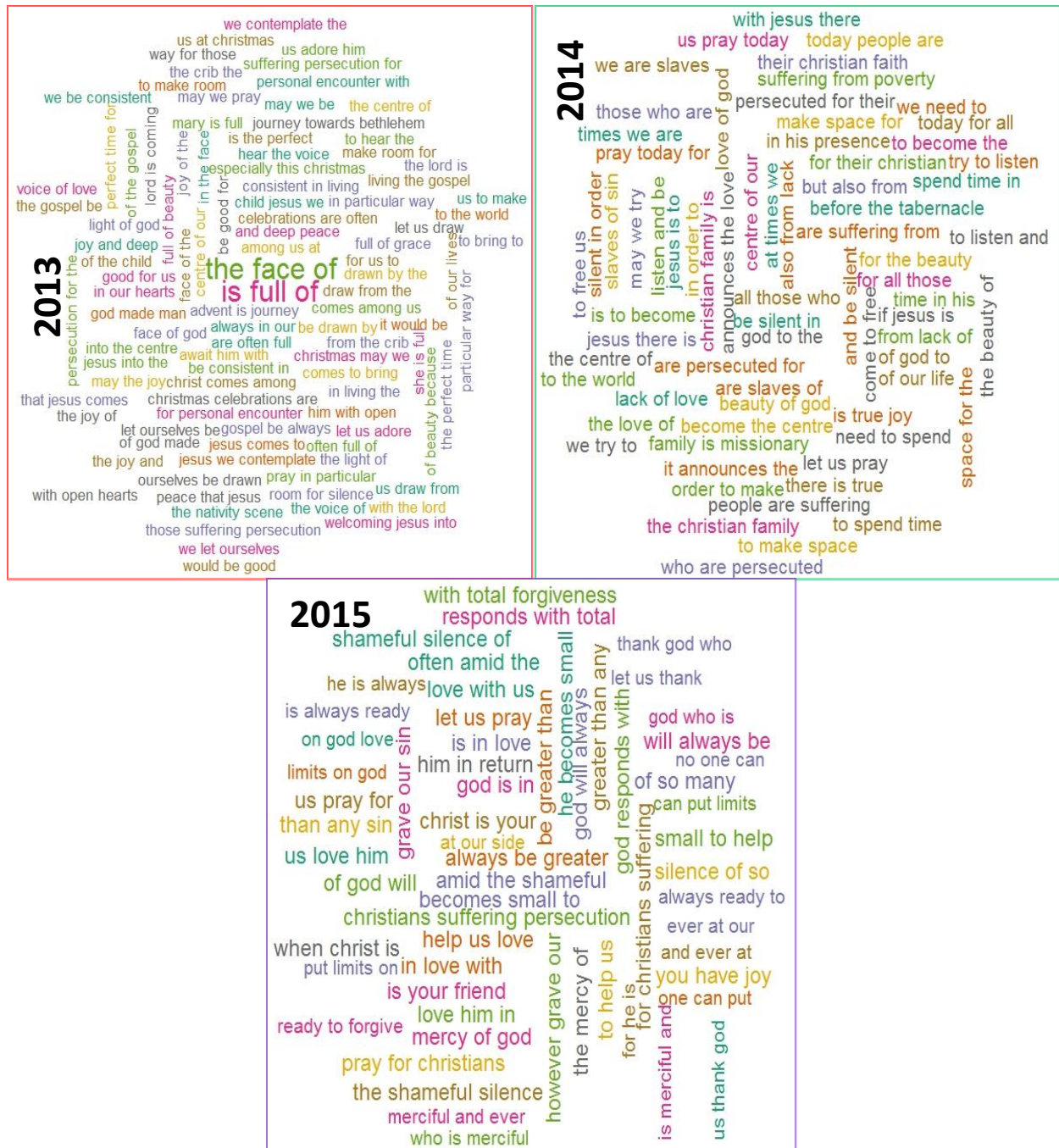
Figure 8: A 3-gram tweets of Pope Francis around Christmas 2013, 2014 & 2015

Gateway to Clinical Intelligence and Operational Excellence through a Patient Healthcare Smart Card System

Sadath Hussain
hussains1@xavier.edu

Thilini Ariyachandra
ariyachandrat@xavier.edu

Mark Frolick
frolick@xavier.edu

Management Information Systems
Xavier University
Cincinnati, Ohio 45207

Abstract

The proliferation of technology in our daily lives has widely changed the way in which information is being captured, processed, stored and analyzed. Information systems have become an integral part of the healthcare system in the developed world. However, the patient journey through the numerous routes in the health care system can make the patient data integrity, profiling, reporting and analysis extremely challenging. It is imperative to design a model to capture the patient's medical records from various health care contexts (i.e. Family doctors, Free Clinics, Emergency room visits, Social Services routes, Senior care facilities and Hospital systems). This article proposes a distributed healthcare information system database which captures and synchronizes information for all patients routing throughout the various services in the US health care system that uses a Unified Medical Record Access and Analysis (UMRAA) card. The proposed system acts as an integrated data solution of a patient's medical history for use in daily operations and decision support analytics.

Keywords: medical record, healthcare analytics, data privacy, data security, unified access, multi-payer system

1. INTRODUCTION

The use of information technology in our daily lives has been on the rise, patient care information systems have now become an integral part of the patient care support in the developed world (Ash, Berg, & Coiera, 2014). In modern healthcare systems, "automation systems in hospitals and medical centers serve the purpose of providing an efficient working environment for healthcare professionals" (Kardas & Tunali, 2006). Some argue that the use of information technology is essential for keeping patients' records (Dick & Steen, 1991;

Armony, Israelit, Mandelbaum, Marmor, Tseytlin, & Yom-Tov, 2015). The use of information technology is believed to have increased the quality of health care services, and decision support system for health care management, health education and research (Jones, Rudin, Perry, & Shekelle, 2014).

Information technology in healthcare has been greatly instrumental in enhancing the ability to apply the vast resources of information technology in complex and sustained health management situations. Healthcare providers that have seen great success in the area of

research and treatment have tremendously benefitted from the use of big data and analytics (Kayyali, Knott, & Van Kuiken, 2013). IT capabilities have empowered providers to be able to maneuver the medium to their advantage. The widespread adoption of IT has also led to a better grasp on handling unintended and unexpected issues when they arise (Bates, Saria, Ohno-Machado, Shah, & Escobar, 2014).). Additionally, IT capabilities encourage more abstract thinking about health care data; especially for research purposes. Yet this data is not fully integrated for access during regular patient visits or for decision support and research analytics (Jensen, Jensen, & Brunak 2012). The United States continues to have large integrated transactional and decision support databases but they are NOT integrated across the nation (Koebnick, Langer-Gould, Gould, Chao, Iyer, Smith, & Jacobsen, 2012).

This paper proposes the creation of a distributed healthcare information system that is grounded in past literature. It also lays out the process through which the adoption of an universal medical record access and analytics (UMRAA) system would be created. The implementation of such a system will involve a pilot study in a single US state. The potential success of the proposed plan in this particular state could be leveraged to help sell the idea to the rest of the nation. The paper first discusses the state of healthcare and IT as presented in past literature done in the field. The next part of the paper describes smart card use in healthcare, the proposed system, advantages and challenges as well as the plan for the development and implementation of the Unified Medical Record Access (UMRAA) card in detail. Finally the potential challenges that the proposal could face are discussed.

2. STATE OF HEALTHCARE AND IT

Keeping in mind the current state of the US healthcare system, it is very timely and paramount for a information systems and analytics to continue to be a major contributor to every decision-making process that goes on in the healthcare industry (Himss 2014). The growth of data collection and the inclusion of information technology in healthcare continues to grow at a rapid pace. It is expected to reach $31.3 billion by 2017 (Bernie Monegain, 2013). The current state of healthcare predominantly revolves around the following issues: (1) actual costs associated with getting quality care, (2) the accessibility and availability of healthcare across the continental US, and (3) the education provided to individuals about maintaining sound health and obtaining precautionary health check-ups in order to prevent major medical costs (Shi, & Singh,

2014). Another major issue with healthcare would be the potential inability to cater to the needs of the next wave of senior citizens (Ou, Shih, Chin, Kuan, Wang, & Shih 2013).

Additionally, the rise of medical errors and the potential for maltreatment due to lack of availability of complete patient health information is also one of the major issues in the healthcare industry (Agha 2014). The dissatisfaction with the healthcare system could increase over the next few years as a result of increased out-of-pocket expenses associated with the weakening economy and increasing prescription drug prices (Haren, McConnell, & Shinn, 2009). The gradual increase in uninsured individuals is only going to add to the increasing costs of health care in the future. Keeping these factors in mind, bolstering the US healthcare structure with the support of an integrated information technology solution for access and analysis would be the potential solution moving forward (IOM, 2009).

One such change in the past that has shown enormous success is switching from paper medical records to electronic medical records that provided a centralized location for storing patient information that in turn streamlined healthcare management (Frolick, 2005; Jacobus, Braun, & Cobb, 2014). The implementation of information system in health care practices is fraught with numerous risks. The stored data on multiple locations in health practices can be a challenge to report essential and strategic information for various stakeholders (AHRQ, 2006). The healthcare information management system in many developed countries like the US, Canada, and many European countries is not well integrated (Brown 2003). Due to various patient information flows and routes, there has been an inherent difficulty to integrate and report the data essential for the management, clinicians, policy makers and researchers (Poon, Jha, Christino, Honour, Fernandopulle, Middleton, & Kaushal, 2006).
.

3. SMART CARDS AND MEDICAL HEALTH RECORDS

The idea of having a complete medical record on a smart card based system has been considered for several years now (Smart Card Alliance 2012). Computer systems that could store medical histories on a smart card were invented more than a decade ago in countries such as Hungary, France, and Spain (Naszlady & Naszlady 1998). However, the US has yet to implement such a system at a national level. In 1998, a study involving an electronic chip card was carried out where 5000 chronically ill patients throughout Hungary received a smart

card that had entire patients' medical history stored in it (Naszlady & Naszlady 1998). The goal was to achieve complete patient information and also to support the growing need for an integrated healthcare delivery model.

Currently, there is no national health care smartcard system in place in the US. However, in some European countries such as Britain and France, pilot programs had been established over a decade ago. These pilot programs have proved to be highly useful and easily implementable (Neame, 1997; Marschollek & Demirbilek 2006; Liu, Yang, Yeh, & Wang. 2006). Today's, health smartcards in France have served the purpose of carrying information related to health insurance, and some ongoing health records and basic emergency health information. Furthermore, strengthening the evidence of their usefulness, the Exeter Care Card (ECC) pilot program that was funded by Britain's Department of Health and carried out by Exeter University (Hopkins, 1990), showed tremendous advantages of having such a smart card system in the health care industry. The advantages of the ECC pilot were the reduction in the cost of prescribing; reduced cost of carrying out investigations; reduction in risk of iatrogenic cases of illness; reduced times taken for data communication; ready access to necessary medical records (Neame, 1997). A valuable addition to the result of the study was that it also showed high patient satisfaction levels.

In 2006, another study illustrated to the US healthcare industry the possibility of solving one of the major hurdles to smartcard technology - which is interoperability (Marschollek & Demirbilek, 2006). Since healthcare organizations do not use the same health information system software, there are multiple sets of ways to code for the same information based on the type of software that is being used. These challenges can easily be overcome using standardized software and technologies in order to facilitate interoperability with multiple healthcare information systems, such as used in the German Health Card pilot program (Marschollek & Demirbilek, 2006).

Another study conducted by Wei Chen et al (2012), proposed to establish a portable electronic medical record system that applied streaming media technology to access medical images and transmit them via the Internet. This is an example of a distributed information management systems in healthcare. Figure one shows a graphical representation of the structure of the portable electronic medical record (EMR) system. The study proposed a

system that is composed of the EMR query system, data exchanging, and the EMR streaming media system. The proposed architecture provided local hospital users the ability to acquire EMR text files from a previous hospital. It also helped access medical images as reference for clinical management. The proposed architecture shown in figure one provides a diagrammatic illustration of what a distributed information system could look like (Wei Chen, 2010). One major limitation to the system shown in the study is the system's dependency on the internet for its data transfer functionality. However, the concept proposed in this paper does not require the internet for its operability and functionality.

The factors that have been referenced from all the various studies described provides a compelling argument to implement a comprehensive, consolidated and secure model for healthcare information system that can be easily and quickly made available and accessible to healthcare providers. Adding portability to the electronic medical record system in the form of the UMRAA card maximizes efficiency and streamlines the whole patient-doctor experience at a national level.

4. UNIFIED MEDICAL RECORD ACCESS AND ANALYTICS (UMRAA) DISTRIBUTED INFORMATION SYSTEM

The work presented here proposes a theoretical approach to building a distributed information management system in the form of a Unified Medical Record Access and Analytics (UMRAA) card. The UMRAA system will not only store patient's entire medical history, but it will also update, synchronize and store all information at every point of a patient's encounter with healthcare (i.e. Hospital, Family Doctor, etc.). This system enables any doctor in the US to view a patient's entire medical history thereby increasing overall efficiency and reducing the possibility of potential medical errors or mall treatments.

Historically, smart cards have supported an impressive variety of applications, and this variety will expand as the cards have become smaller, cheaper, and more powerful (Shelfer & Procaccino, 2002). With this innovative outlook and compelling need for a comprehensive, consolidated and secure model for healthcare information system, this research describes a patient information system, where the data is captured on all the respective information systems redundantly. A key player in implementing the UMRAA Card system is the federal government.

The United States has one of the highest per capita health care spending among all the developed countries (Anderson, Frogner, Johns & Reinhardt, 2006), yet healthcare remains complicated and expensive. Highest per capita healthcare spending has not translated into more resources; the problem here is that the US health care system does not have the proper allocation of funds. The German Health Card (Marschollek & Demirbilek, 2006), the Taiwanese national health insurance card – which also had medical records (Liu, C et al., 2006) - and the Exeter Care Card (Hopkins, 1990) all have one thing in common. They have government funding. Out of the three aforementioned programs, the ECC and German Health Card were both well accepted by the patients and healthcare providers. These studies suggest that such a program can be implemented in the US and that the UMRAA card system will be well received by the general population.

However, it is worth mentioning that there are other proposals for integrating health care such as cloud based, block chain and personal health records. The characteristics of each of them are presented in table one.

5. INFORMATION TECHNOLOGY INNOVATION AND HEALTHCARE

Efforts to develop and optimize the integration of healthcare data have been continually evolving. To create a state of the art continuum of care there has been great demand for complete integration of the delivery of care and operational processes. The increased adoption of Electronic Health Records across the continuum has led to the accumulation of diverse healthcare data in silos. Given the growing opportunities and challenges with regard to achieving clinical and operational excellence, it is paramount to develop the ability to integrate all the diverse healthcare data in a meaningful way. This would enable healthcare organizations to utilize data in a more inclusive and innovative fashion. The acquisition of healthcare data that span various domains within the continuum of care must be captured and consolidated in a manner that is readily reusable for higher order functioning without the need for manual transformation of data. One major barrier to integrating data generated from diverse applications is its complexity which is inherent due to the nature of individual applications not being designed for synchrony (Lenz, R., Beyer, M., & Kuhn, K. A., 2007). The development of a comprehensive distributed information systems infrastructure would be the ideal solution for smooth information flow and knowledge generation. In spite of potential barriers, the demand for

clinical and operational efficiency, and the adaptation of state of the art technology, has enabled a few organizations to succeed in garnering benefit from data integration. Healthcare decision makers would be forward thinking if they could predict and understand the extraordinary potential and value that UMRAA program could provide.

Data integration has led the way to the implementation of real-time business intelligence. Analyzing healthcare data to predict patient volumes, the cost of care and clinical outcomes in order to improve the quality, safety, and productivity of point of services has always been an integral part of running a competitive healthcare system (Mettler, T., & Vimarlund, V., 2009). In the past, organizations addressed this in the form of monthly and yearly data analysis and also often depending on market analysis and independent external resources. However, in recent years, this trend has changed due to the growing pace of delivery of care transition from volume based to value based care models. With ever-increasing healthcare market share competition and growing healthcare customer demands and changing needs, progressive decision makers are demanding dynamic data analytics reports and real-time business intelligence. Real-time business intelligence has become a great necessity in today's evolving healthcare environment in which organizations are operating in a constant state of evolution due to the continual change in the healthcare landscape (Raghupathi & Raghupathi 2014). Changing policies and customer demands are at their historic peak. Patients and stakeholders are more educated and technologically advanced, and therefore more demanding. Advances in technology along with growth in innovative mindfulness have also provided a favorable landscape for the growth of real-time business intelligence. Real-time business intelligence in the form of dynamic dashboards is becoming a common must-have for today's decision makers. Adaptation of revolutionary concepts such as UMRAA in healthcare are going to be the drivers of advanced future real-time business intelligence platforms (Azvine, B., Cui, Z., & Nauck, D. D., 2005). UMRAA has the capability to harness extraordinary diverse healthcare data and would, in turn, provide organizations with most current and state of the art information and knowledge that would allow them to make both business and clinical decisions in a more timely, efficiently manner and with a high degree of accuracy and success.

In addition to real-time business intelligence, big data in healthcare is another domain that has the potential to provide extraordinary advantages to the way we provide care to

healthcare customers (Raghupathi & Raghupathi 2014). With the advent of the widespread adaptation of electronic medical records, healthcare organizations are in possession of large volumes of complex and variable healthcare payer, provider and patient information. Today's advanced electronic health records have the ability to capture, store, manage and distribute healthcare big data in a timely and efficient manner. Existing analytical techniques have the capacity to derive deeper information and knowledge from this vast amount of health and business-related big data in terms of quality, safety, access and outcomes of providing care (Banerjee, A., Bandyopadhyay, T., & Acharya, P., 2013). With the introduction and implementation of UMRAA, benefits that arise from large volumes of organizational data can be factored exponentially to its potential applicability. One study states that big data has the potential to save more than $300 billion per year in US healthcare, which is predominantly coming from reductions of approximately eight percent in national healthcare spending. The two largest areas for savings with $165 billion and $108 billion in waste respectively are coming from Clinical Operations and Research Big data analytics (Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H., 2011).

Another promising avenue belonging to the data analytics realm that can influence healthcare is the system of interconnected applications that yield information from the environment through sensing and also interact with the surrounding physical world by using the internet to provide advanced services for applications, communication, and business intelligence (Atzori, L., Iera, A., & Morabito, G., 2010). Internet of Things has gradually transitioned into the mainstream by transforming the traditional Internet into a fully integrated Internet of tomorrow wherein lies the potential for not leaving any single digital device to work in solitude. The digital world of the future is poised to embark on a journey where each and every possible device on earth would communicate to the other in order to create an unmanned equilibrium of processes around us that would automatically sense the needs of the environment that they reside (Aggarwal, C. C., Ashish, N., & Sheth, A., 2013). Internet of Things has been vastly growing and successful in recent years due to the growth and development of devices that are generically embedded by open wireless technology such as Wi-Fi, radio frequency identification, and Bluetooth as well as actuator nodes and embedded sensor (Gubbi, J., Buyya, R., Marusic, S., & Palaniswami, M., 2013). UMRAA has the potential to take delivery of

healthcare to a new dimension with the potential to transform into "Health Internet" as the Internet of Things for healthcare applications. As UMRAA is designed to work as an integrated data platform, it has the potential to interact and function to command necessary actions that may be warranted in given situations that presently require human intervention.

Finally, we discuss smart cities as the potential universal reality of tomorrow. Smart cities are geographic boundaries that are encapsulated in an integrated information and communication infrastructure that has multiple applications that are programmed to cooperate and coordinate with each other in order to understand and potentially predict our needs even before we realize them (Solanas, A., Patsakis, C., Conti, M., Vlachos, I. S., Ramos, V., Falcone, F., ... & Martínez-Ballesté, A., 2014). The widespread adaptation and implementation of UMRAA can provide a fundamental framework for future smart cities to replicate the structure and design of UMRAA to other required domains of human necessities in today's world such as banking, education, transportation, and communication.

6. PROPOSED SYSTEM

Although the use of information systems increased the quality of health care services, it still faces numerous challenges (Shekelle, Morton, & Keeler, 2006). The existence of multiple information systems in various healthcare facilities can make it a real challenge to compile and report the data for different purposes (Heathfield, Pitty, &Hanka, 1998). Historically, a number of data modeling techniques have evolved from recent research to improve reporting, analytics and quality of health care. In a recent study (McGregor, 2012), Patient Journey Modeling architecture (PaJMa) was used as a modelling technique for data representation. PaJMa was an approach used for process flow modeling that was designed specifically to understand the nuances of patient journeys during a patient-provider encounter.

UMRAA is a portable unified medical record system which is a part of a distributed healthcare information system. It maintains a continuously-growing list of data records that each patient encounters. Each encounter refers to a previous encounter on the smart card and is thus the patient data on the smart card is comprehensive and hardened against tampering and revision by unauthorized users.

The diagrammatic representation of the proposed system is represented in Figure 2

which shows the data flow in the proposed distributed information system. To phase in the system to the current US population, the first point of patient's data entry will be determined by the circumstances of each individual patient. Current patients can be given their UMRAA card on their next doctor or ER visit, while new-borns can be registered in conjunction with the issuance of their social security card and/or birth certificate. The proposed system could work very similarly to a CarFax (Barnett 1991) report that consumers can obtain for knowing the complete history of a car based on its vehicle identification number.

CarFax is a database search and reporting program where used car consumers enter a vehicle identification number (VIN). A CarFax application searches a database, and provides a report for a small fee. The business model for Carfax is as follows: used car consumers want to know if an automobile has a tarnished history, and CarFax provides an instant answer. CarFax is able to provide an answer because they have compiled a huge database of more than 1 billion vehicle registration records. By searching the database it can quickly provide detailed information on just about any vehicle sold in the US. The system proposed in the paper can be thought of as a "CarFax for patients."

Similar to a CarFax report, a healthcare provider such as a physician or nurse practitioner in any part of the United States may have access to a patient's entire medical history, regardless of which part of the country and what type of service the patient receives. For instance, a family doctor's reports from Anchorage, Alaska and free clinic reports from Atlanta, Georgia can all be accessible to a nurse practitioner in Amelia, Ohio). The UMRAA Card will synchronize (i.e. Download new data and upload previous data or even just allow access to the information stored) all information every time it is accessed at a healthcare facility.

Even after the Affordable Care Act (ACA), private healthcare organizations have the tendency to work in silos. Therefore the idea of having a Unified medical record would not be well accepted in the US. For the US healthcare system to adopt the proposed plan nationwide, it is important that the federal government push this agenda as an addition to the ACA which would help further the cause of moving from fee based performance to value based system. Individual health organizations would not be willing to adopt such a system as they fear that it would lead to patient poaching by their competitors. The data available in the smart card could be available to competitors if patients come in contact with competitors.

Therefore, in order to implement and reap the benefits of the proposed plan, it needs to be adopted by the federal government which can make it a mandate as they did it in case of the ACA.

In addition, the adoption of the proposed plan would require the development of a common vocabulary. A champion from the highest echelon of government will be required to encourage adoption of such a system by healthcare providers, patients and insurance providers. High system adoption and system success through a strong champion from upper management is well recognized in large IT implementations in organizations (Cresswell, Bates, & Sheikh, 2013). Such a champion would encourage all stakeholders and consumers to adopt and share a common vocabulary/taxonomy. These stakeholders and consumers will be able to reap the benefits only if they are willing to talk the same language and if they are willing to do adopt a common taxonomy. In other words, a champion from the highest levels of government and the emergence and widespread adoption of a shared taxonomy is essential to implementing the proposed plan.

The implementation of the UMRAA Card system can be straight forward. Figure three shows the step by step illustration of the UMRAA process. Patient registration is triggered on the very first visit of a patient to any health care facilities or at birth. The proposed system requires patients to register at a point of interest in any possible routes such as primary care practice, hospital facility, social services or emergency room. In emergency cases, healthcare takes precedence over recording patient information and in some cases the formalities are relaxed in an emergency situation. The prioritization of registration and healthcare support is largely dependent upon the individual patient circumstance. Therefore, the patient registration can take place before or after the treatment for first-time visitors.

The UMRAA record for each patient is created in a database and stored on a portable electronic card. The next step of the process would be to encrypt the UMRAA card with a secure password. Now the UMRAA card is ready to be carried by the individual patient and presented at all health care facilities for data synchronization on each visit. After the essential clinical and health care activities, the patient data is recorded in the information system and transferred over to the health care card. If the patient paid a visit to another health care facility and the card is holding new data, it is transferred over to the primary care information system. This will ensure

completeness of patient data in all health care facilities. The process is visually presented in Figure 4.

However, the patient may also choose only to have new reports downloaded and have their previous history be only viewable to the healthcare provider they are seeing. This method will prevent possible privacy leaks and also prevent a littering of data in healthcare databases throughout the country where the patient was only visiting and had needed unplanned emergency care.

7. POTENTIAL BENEFITS

The UMRAA card would have a number of potential benefits with regards to many different stakeholders in the continuum of care. From a patient's perspective, the patient has complete and comprehensive information about his/her health status. This would immensely help in terms of patient compliance and satisfaction. Furthermore, most of the benefits that have been outlined for other stakeholders seem also to apply to individual patients as well.

From the providers' perspective, the healthcare provider will have complete knowledge and information that would be required to deliver a complete management plan. This vital information would also help in avoiding duplication of health services and at the same time avoid redundancies. The major benefit from the UMRAA card would be in the event of an emergency at a healthcare facility that never had any prior encounter with the patient seeking emergency medical care. The information on the card would be critical in making life and death decisions that would impact the health of the patient. Another significant benefit would be the reduction in readmission rates with the help of the data available from UMRAA to the first clinical point of contact which would potentially help the healthcare provider to take necessary steps to avoid readmission while planning the management.

Looking from a payer's perspective, UMRAA could lead to care coordination and reduction in over utilization leading to cost savings. Lesser administrative burden to process claims. Customization of health plans to determine the best combination of benefits for covered lives.

To put things in perspective, US healthcare system is one of the biggest spenders in comparison to its other developed counterparts worldwide. In spite of having extraordinary expenditure geared towards healthcare, outcomes do not match the expenditure or are less than some of the countries that spend far less than the US on healthcare. In view of these unfavorable trends in US healthcare, politicians and law maker have systematically steered healthcare from volume to value based service with the aid of various policies and reforms over the years (Mayes, R. 2011). Moving forward, having such new unexplored territories in the area of population health and value based care, would be detrimental for organizations with limited access to needed data. Having access to information that would help them control cost as well as provide quality care in order to be in alignment with the policies and reforms, is crucial for success. The kind of timely analytical information that can be derived from the data on UMRAA can be the paramount factor providing competitive edge to healthcare providers seeking growth in business and increased market share.

UMRAA can provide instant predictive information to healthcare personnel with regards to impending acute health episode based on the trend in the chronic condition of a patient. It can potentially reduce the emergency department visits as well as the thirty day readmission post discharge from a healthcare facility (Amarasingham, R., et al, 2010). As the US healthcare system is pacing towards a shift from volume to value based care as well as reimbursement; much of the financial gains are going to come from keeping patients out of emergency departments as well as preventing 30 day readmissions. Integrated systems like UMRAA are going to greatly impact such areas with reliable and accurate predictive information. This will enable payers and providers to act in time and steer towards desirable health outcomes that are in line with the requirements of the centers for Medicare and Medicaid services (CMS).

Healthcare organizations having access, disease, outcomes and analytics data, derived from UMRAA, can be tremendously instrumental to organizations. They can better align with the requirements of the CMS, as a result. CMS is the largest healthcare payer in the US., An organization's ability to align with CMS's requirements with the help of predictive analytics coming from UMRAA can greatly impact the financial viability of organizations. With enhanced reimbursements, organizations can gain growth and sustainability. Furthermore, with regards to BI benefits from UMRAA, avoiding duplication and reduction in redundancies can potentially give rise to economies of scale in relation to the use of supplies and cost of care. Additionally, having mass quantities of information coming from patient encounters across the continuum of care can facilitate the development of benchmark and best practices.

In addition to the above, CMS and healthcare in general are slowly steering towards population health and bundled payments. Historically, healthcare organizations, and the different entities within have worked in silos. This kind of operational silos in terms of providing and managing care will make predicting risk for bundled payments very difficult. UMRAA can be helpful in assessing time and costs in terms of episodes of care. This will help organizations be more cost effective in terms of predicting costs based on historical expenditures. Having the advantage of data and information in terms of cost of the care during the continuum can help them predict ball park risk. This will help us better negotiate payer contracts as well as deliver efficient and timely care leading to increased financial returns.

The vast amount of information and data that would be gathered from a complete and comprehensive healthcare history of patients' would help in running clinical and non-clinical healthcare analytics that would greatly help in research and development of clinical as well as healthcare management protocols. The vast amount of data and knowledge that would be generated from the advanced analytics of patient data will improve the efficiency of healthcare management and reduce healthcare costs, both, to the individual patient as well as the US healthcare system in general (Raghupathi, Raghupathi 2014). This kind of business and clinical intelligence framework can lead to a major breakthrough in healthcare and can give rise to a completely new approach to the way the US delivers healthcare in the future (Foshay, & Kuziemsky, 2014).

8. CHALLENGES

While there are a number of advantages of UMRAA, there are some challenges to the proposed approach. The UMRAA card system will incur costs of hardware, portable electronic health care cards, and software. In addition to the numerous costs associated with healthcare, the added costs of this system will be a burden that can be felt across the board, ranging from increasing insurance premiums to increasing tax rates.

The costs associated with the initiation, implementation, and maintenance of this system will have to be shared between the governments (i.e. City, State & Federal), insurance companies and local healthcare systems. As mentioned earlier, the costs associated with the initiation of this program will have to be funded via a federal stimulus package. However, to accelerate the implementation, the program can be started state wise, with the least economically weak states implementing the program first. Furthermore, the maintenance costs of this program will have to be shared in some proportion by all parties responsible for the healthcare needs of a US resident, namely, the governments at all levels, insurance companies, and even the individual hospital system or primary practice at a local level. This ensures that there are checks and balances across the board, and everyone is held accountable.

The system requires a compatible system at each patient healthcare touch point. Words on any online material can be recognized by software as text that can be editable. Similarly, the UMRAA database will have to be able to recognize all the data associated with medical records across various health care check points and be able to store that data in an effective way for daily access as well as decision support analytics.

The damaged, lost or stolen cards can occur and bring about additional costs. Such cards may also raise privacy and security concerns. These costs will, to a certain extent, be the responsibility of the patients. Just as social security cards and expensive jewellery are protected, the UMRAA card will also become an indispensable part of an individual. Providing a secure way of accessing the information stored on the card can be a challenge. A combination of a fingerprint scan along with a swipe of the UMRAA card at health care facilities should suffice for the security of the stored information. In order to access the information and/or synchronize with the UMRAA card, the patient will have to provide a fingerprint scan along with swipes of the healthcare providers ID card and the UMRAA card itself. Although, fingerprint technology for security has been around for ages, it is still considered one of the many key methods of providing reliable security features to a smart card (Butler, A. 2016). Moreover, this transaction can only be completed at healthcare facilities, thus preventing unauthorized access.

Restricting access of the UMRAA information to only essential healthcare providers (i.e. Physicians, Nurse Practitioners, etc.) will ensure that the information stored on the UMRAA card remain confidential between the patient and doctor. Therefore, non-essential healthcare providers who would not need access to a patient's entire medical history such as a pharmacist or a licensed practical nurse should not be allowed into the UMRAA database. Only a physician and possibly a nurse practitioner should have unlimited access to the UMRAA information.

The overall integration of patient health records is the main success factor of UMRAA. However, there is a need for the healthcare system stakeholders to realize that it is important to either have a single payer system or a federal mandated multi payer system for the proposed plan to be adopted and implemented nationwide. Besides the factors that were discusses earlier, the main factor that stops different healthcare organizational silos from adopting something like this is that they fear losing patients to their competitors. This paper has highlighted this fundamental deterrent that needs to be addressed in order to have buy-in from the payers and providers of healthcare. If not there may be a need to make it a federal mandate, just as it was done in the case of the "Affordable Care ACT."

Finally, there are a number of countries that have implemented a healthcare smart card system. Many US healthcare organizations have also implemented smart card pilot programs with their organizations. However, most of these programs are implemented with the idea of stream lining the reimbursement system as well as obtaining medical information and also for the purpose of biometric authentication. Although the implementation of smart card technology has been discussed in the US for many years, its adoption in the US has been slow due to key issues, such as privacy, interoperability, and security of patients' data. However, some healthcare institutions such as Mt Sinai Hospital in New York have incorporated the use of storing medical information, health insurance contact details, and even electrocardiogram results on a secure and private smart card (Smith, & Barefield, 2007). In spite of all these efforts, the US feels the need to conduct more research on the use of smart card technology due to previously mentioned reasons.

Smart Card Alliance which is a non-profit organization whose purpose is to develop an understanding and explain the use of smart card technology is trying to bring awareness and therefore stays connected to industry leaders through educational programs, marketing research, and open forums (Alliance 2015). They have also been trying to ease the industries fear that revolves around the security and privacy issues that come with it. However, Smart Card Alliance fails to understand that in US, it is going to be very difficult to implement such a system due to a multiple payers (Hussey & Anderson 2003), in spite of all the benefits that have been discussed already in this paper.

There are healthcare smart cards in the US that are currently being proposed for nationwide implementation. However, the players proposing such smart card implementation are failing to take into account the biggest barrier to nationwide system implementation: the multi payer system. The US healthcare system is comprised of multiple payers. Medicare is the single largest payer in the US and the rest are multiple private insurers. Therefore, the US healthcare has been mostly working in silos up until the affordable care act (ACA) came into play.

9. CONCLUSION

The use of heterogeneous patient information system in various health care facilities can make it a challenge to report and analyse patient data. The data integrity and completeness can be challenging for the employers, clinicians, and researchers. The UMRAA card distributed information system approach can combine and integrate the pertinent patient data in all the healthcare facilities to support quality of health care, reporting, treatment and management. Besides the innumerable benefits of this system, there are challenges to overcome. These challenges come in the shape of privacy, security, and costs associated with the initiation, implementation, and maintenance of this system.

Smart cards are not new to the healthcare industry. Current health smart card is in use in many parts of Europe as well as here in the US such as the one that was implemented by the Mount Sinai health system in New York. Most of these health smart cards are used of reimbursement feasibility. In addition to making reimbursement easy, there are a few that are more geared towards improving the care of patients as well as improving quality and reducing cost.

There are endless possibilities and benefits that could come along with this program, such as globalization of healthcare and uniformity in the quality of services offered to the patients. An UMRAA system has the potential to reduce malpractices, delayed decision making, etc. which usually occur as a result of healthcare providers not having enough information. Other future possibilities in health care with UMRAA would be with the use of analytics. Analytics for research purposes using the data derived from the widespread use of UMRAA card would potentially lead to better health care process reengineering. The initiation and implementation of this program is indeed a herculean task, but the advantages far outweigh the disadvantages. All in all the UMRAA system is at the center of providing the most efficient and standardized patient care

within the United States (Raghupathi & Raghupathi 2014).

In addition to traditional analytics, emerging and innovative applications and platforms such as Big Data Analytics, Real-time Business Intelligence, Internet of Things and Smart Cities are frameworks that would most benefit from the UMRAA program. UMRAA has the potential to generate data from numerous patient care points of contact and integrate it with global healthcare communities. Various innovative platforms mentioned earlier may stumble into potential pitfalls without a well-engineered data integration support system. However, UMRAA program with its current robust design can act as a promising instrument for laying the foundation and widespread realization future mass data-centric applications.

This paper proposes an ambitious plan to assist in the modern day US healthcare landscape that is currently undergoing a swift transformation from the traditional volume based to a value based model of care. Some of the benefits and challenges discussed in this paper are extremely pertinent to the evolving US healthcare cosmos.   We need to have a targeted vision in terms of what tools and techniques that can be helpful to us to be able to provide the best in class value based care to our patients. The paper lays ground work to continue the discussion around the technical as well as logistical functions involve in the actual implementation of the UMRAA. There is a need for ongoing research in this space; specifically, on security as well as system integration. The information provided hereprovides an outlook on the continuum of care afforded to patients and how UMRAA intersects at various points throughout the continuum. It can help in the development of a stream of research that shows how healthcare can go the distances in bringing about excellence in providing cost effective quality healthcare.

10. REFERENCES

Agha, L. (2014). The effects of health information technology on the costs and quality of medical care. *Journal of health economics*, *34*, 19-30.

Aggarwal, C. C., Ashish, N., & Sheth, A. (2013). The internet of things: A survey from the data-centric perspective. In *Managing and mining sensor data* (pp. 383-428). Springer US.

Alliance, S. C. (2015). Smart Card Alliance.". *INSIDE Contactless Offers Free, Downloadable, Open NFC API and Source Code on SourceForge*.

Amarasingham, R., Moore, B. J., Tabak, Y. P., Drazner, M. H., Clark, C. A., Zhang, S., ... & Halm, E. A. (2010). An automated model to identify heart failure patients at risk for 30-day readmission or death using electronic medical record data. *Medical care*, *48*(11), 981-988.

Anderson, G. F., Frogner, B. K., Johns, R. A. & Reinhardt, U. E. (2006). Health Care Spending And Use Of Information Technology In OECD Countries. Health Affairs, 25, no.3 (2006):819-831

Armony, M., Israelit, S., Mandelbaum, A., Marmor, Y. N., Tseytlin, Y., & Yom-Tov, G. B. (2015). On patient flow in hospitals: A data-based queueing-science perspective. *Stochastic Systems*, *5*(1), 146-194.

Ash, J. S., Berg, M., &Coiera, E. (2014). Some unintended consequences of information technology in health care: the nature of patient care information system-related errors. Journal of the American Medical Informatics Association, 11(2), 104-112.

Atzori, L., Iera, A., & Morabito, G. (2010). The internet of things: A survey. *Computer networks*, *54*(15), 2787-2805.

Azvine, B., Cui, Z., & Nauck, D. D. (2005). Towards real-time business intelligence. *BT Technology Journal*, *23*(3), 214-225.

Barnett III, E. H. (1991). *U.S. Patent No. 4,989,144*. Washington, DC: U.S. Patent and Trademark Office.

Bates, D. W., Saria, S., Ohno-Machado, L., Shah, A., & Escobar, G. (2014). Big data in health care: using analytics to identify and manage high-risk and high-cost patients. *Health Affairs*, *33*(7), 1123-1131.

Banerjee, A., Bandyopadhyay, T., & Acharya, P. (2013). Data analytics: Hyped up aspirations or true potential. *Vikalpa*, *38*(4), 1-11.

Brown, L. D. (2003). Comparing health systems in four countries: lessons for the United States. *American Journal of Public Health*, *93*(1), 52-56.

Butler, A. (2016). *U.S. Patent No. 20,160,104,161*. Washington, DC: U.S. Patent and Trademark Office.

Chen, W., & Shih, C. C. (2012). Architecture of portable electronic medical records system integrated with streaming media. *Journal of medical systems*, *36*(1), 25-31.

Cresswell, K. M., Bates, D. W., & Sheikh, A. (2013). Ten key considerations for the successful implementation and adoption of large-scale health information technology. *Journal of the American Medical Informatics Association : JAMIA*, *20*(e1), e9–e13. http://doi.org/10.1136/amiajnl-2013-001684

Dick, R. S., & Steen, E. B. (1991). The computer-based patient record: an essential technology for health care: Natl Academy Pr. Eberhardt, S. P. (1998). U.S. Patent No. 5,832,488. Washington, DC: U.S. Patent and Trademark Office.

Foshay, N., & Kuziemsky, C. (2014). Towards an implementation framework for business intelligence in healthcare. *International Journal of Information Management*, *34*(1), 20-27.

Gubbi, J., Buyya, R., Marusic, S., & Palaniswami, M. (2013). Internet of Things (IoT): A vision, architectural elements, and future directions. *Future Generation Computer Systems*, *29*(7), 1645-1660.

Haren, M. C., McConnell, K., & Shinn, A. F. (2009). Increased patient cost-sharing, weak US economy, and poor health habits: implications for employers and insurers. *American health & drug benefits*, *2*(3), 134.

Heathfield, H., Pitty, D., & Hanka, R. (1998).Evaluating information technology in health care: barriers and challenges. BMJ: British Medical Journal, 316(7149), 1959.

Himss A (2014) State of the U.S. Healthcare Information Technology Industry [White paper], retrieved from http://apps.himss.org/foundation/docs/FINAL%20COPY%20STATE%20OF%20THE%20ART%20REPORT.pdf

Hopkins, R. (1990). Exeter Care Card evaluation report. London: HMSO. Frolick, M. (2005).Using electronic medical records to improve patient care: The St. Jude Children's Research Hospital Case. Information Science Today.

Hussey, P., & Anderson, G. F. (2003). A comparison of single-and multi-payer health insurance systems and options for reform. *Health policy*, *66*(3), 215-228

IOM (2009). America's Uninsured Crisis: Consequences for Health and Health Care [White paper]. Retrieved January 23, 2016, from http://iom.nationalacademies.org/Activities/HealthServices/InsuranceStatus.aspx

Jacobus, C. J., Braun, J. C., & Cobb, P. (2014). *U.S. Patent Application No. 14/282,949*.

Jensen, P. B., Jensen, L. J., & Brunak, S. (2012). Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, *13*(6), 395-405.

Jones, S. S., Rudin, R. S., Perry, T., & Shekelle, P. G. (2014). Health information technology: an updated systematic review with a focus on meaningful use. *Annals of internal medicine*, *160*(1), 48-54.

Kardas, G. & Tunali, E. T. (2006). Design and implementation of a smart card based healthcare information system. Computer Methods and Programs in Biomedicine. 81(3), 66-78.

Kayyali, B., Knott, D., & Van Kuiken, S. (2013). The big-data revolution in US health care: Accelerating value and innovation. *Mc Kinsey & Company*, 1-13.

Koebnick, C., Langer-Gould, A. M., Gould, M. K., Chao, C. R., Iyer, R. L., Smith, N., & Jacobsen, S. J. (2012). Sociodemographic characteristics of members of a large, integrated health care system: comparison with US Census Bureau data. *The Permanente Journal*, *16*(3), 37

Liu, C., Yang, P., Yeh, Y. & Wang, B. (2006). The impacts of smart cards on hospital information systems - An investigation of the first phase of the national health insurance smart card project in Taiwan. International Journal of Medical Informatics. 75(2), 173-181.

Lenz, R., Beyer, M., & Kuhn, K. A. (2007). Semantic integration in healthcare networks. *International journal of medical informatics*, *76*(2), 201-207.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity.

Marschollek, M. &Demirbilek, E. (2006).Providing longitudinal health care information with the new German health Card - a pilot system to track patient pathways. Computer Methods and Programs in Biomedicine. 81(3), 266-271.

Mayes, R. (2011). Moving (realistically) from volume-based to value-based health care payment in the USA: starting with Medicare payment policy.*Journal of Health Services Research & Policy*, *16*(4), 249-251.

McGregor, C., Steadman, A., Percival, J., & James, A. (2012).A Method for Modeling Health Informatics Capacity in Patient Journeys Supported by Interprofessional Teams. Paper presented at the System Science (HICSS), 2012 45th Hawaii International Conference on.

Mettler, T., & Vimarlund, V. (2009). Understanding business intelligence in the context of healthcare. *Health informatics journal*, *15*(3), 254-264.

Monegain, B. (2013, December 30). Big growth forecast for health IT market. Retrieved January 22, 2016, from http://www.healthcareitnews.com/news/big-growth-forecast-health-it-market

Naszlady, A. &Naszlady, J. (1998), Patient Health Record on A Smart Card. International Journal of Medical Informatics. 48(1-3), 191-194.

Neame, R. (1997). Smart cards - the key to trustworthy health information systems. Information in practice, BMJ: British Medical Journal. 314, 573-577.

Ou, Y. Y., Shih, P. Y., Chin, Y. H., Kuan, T. W., Wang, J. F., & Shih, S. H. (2013, October). Framework of ubiquitous healthcare system based on cloud computing for elderly living.

.

In *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2013 Asia-Pacific* (pp. 1-4). IEEE.

Poon, E. G., Jha, A. K., Christino, M., Honour, M. M., Fernandopulle, R., Middleton, B., ... & Kaushal, R. (2006). Assessing the level of healthcare information technology adoption in the United States: a snapshot. *BMC Medical Informatics and Decision Making*, *6*(1), 1.

Race, Ethnicity, and Language Data: Standardization for Health Care Quality Improvement. (2006). Retrieved January 23, 2016, from http://www.ahrq.gov/research/findings/final-reports/iomracereport/reldata5.html

Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. *Health Information Science and Systems*, *2*(1), 1.

Shekelle, P., Morton, S. C., & Keeler, E. B. (2006).Costs and benefits of health information technology.Shelfer, K. M., &Procaccino, J. D. (2002).Smart card evolution. Communications of the ACM, 45(7), 83-88.

Shi, L., & Singh, D. A. (2014). *Delivering health care in America*. Jones & Bartlett Learning.

Smith, S. P., & Barefield, A. C. (2007). Patients meet technology: The newest in patient-centered care initiatives. *The health care manager*, *26*(4), 354-362.

Solanas, A., Patsakis, C., Conti, M., Vlachos, I. S., Ramos, V., Falcone, F., ... & Martínez-Ballesté, A. (2014). Smart health: a context-aware health paradigm within smart cities. *IEEE Communications Magazine*, *52*(8), 74-81
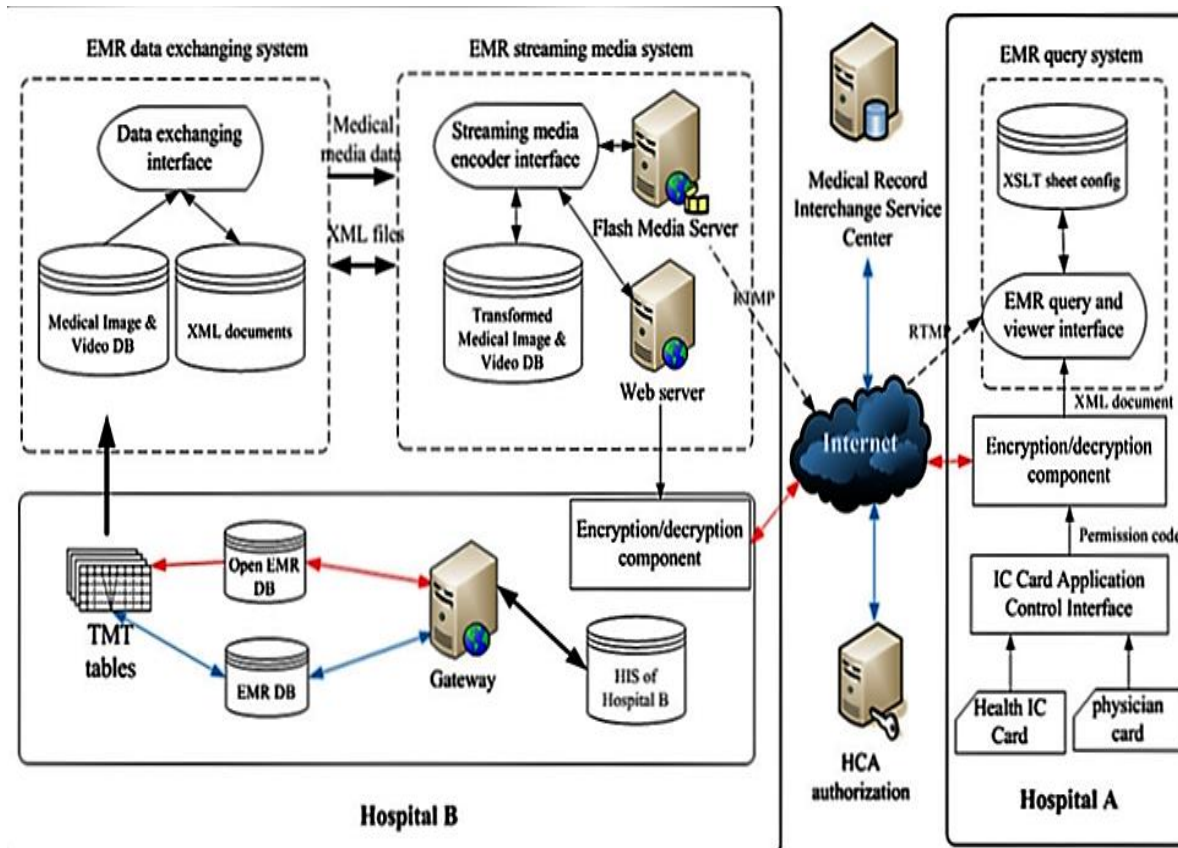
Appendices and Annexures

Figure 1: System architecture that integrates streaming media into portable EMR system (Wei Chen, 2010)

Cloud-based health records	Personal health records	Blockchain
o Internet required for access. o Interoperability would be a question. o Ownership of complete health record not in the hands of patients as in the case of smart card.	o Integration of complete medical history o Portability barrier Internet needed.	o Portability barrier o Irreversible transactions, even on incorrect ones

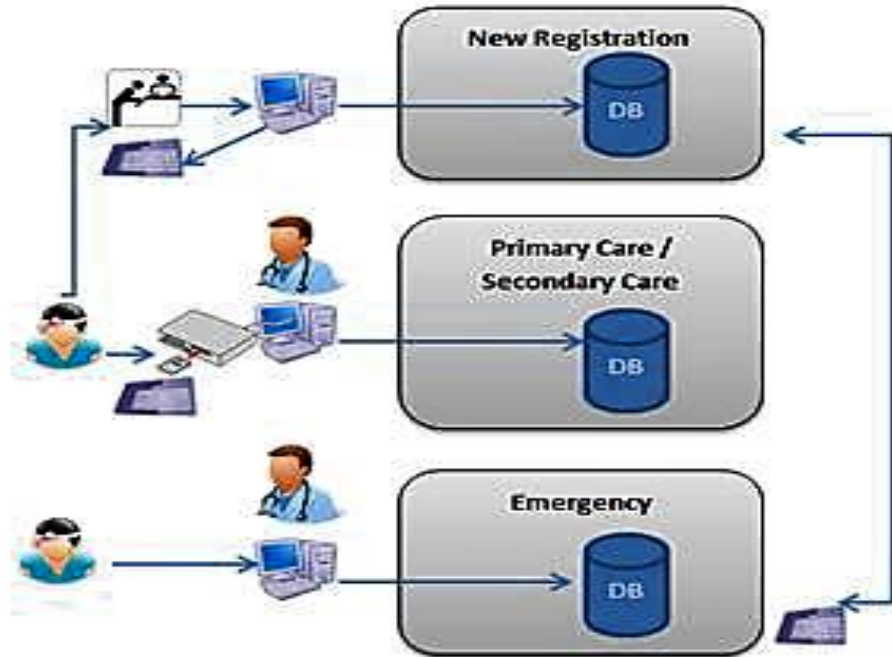Table 1: Alternative means of storing healthcare records other than smart cards

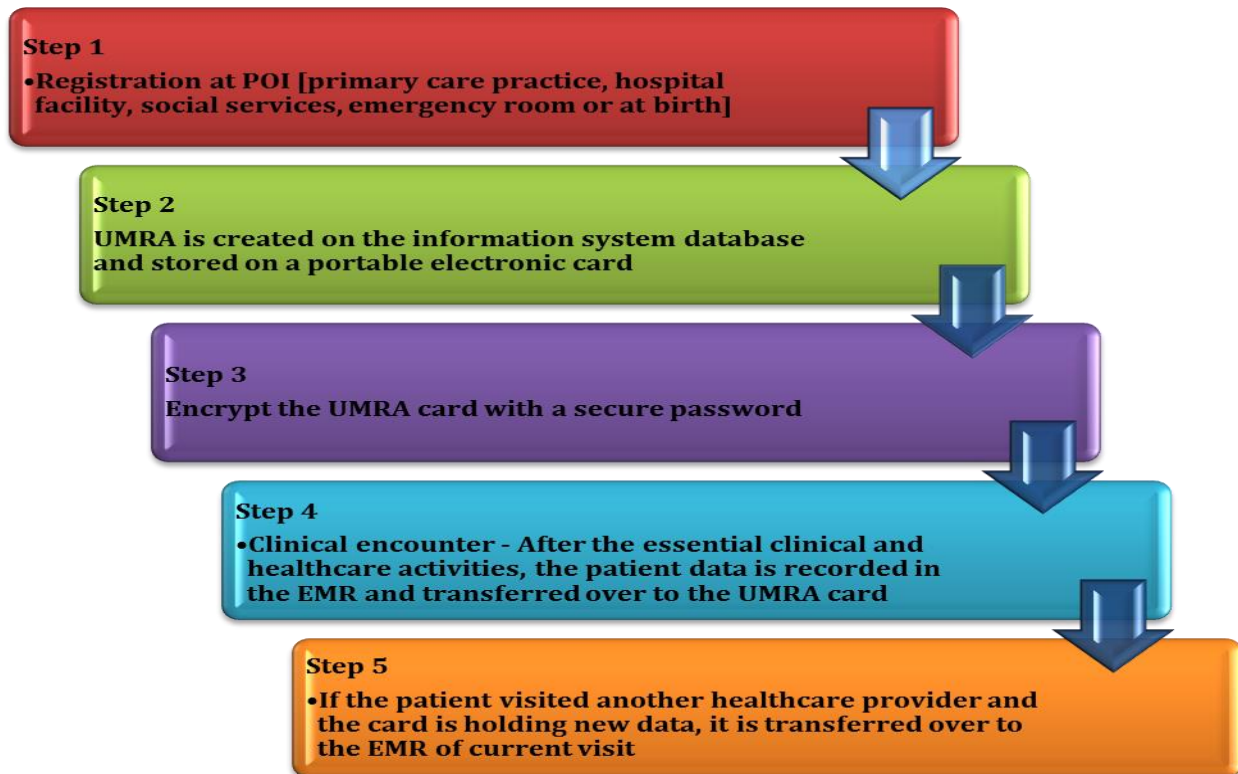Figure 2.Data flow in the proposed distributed information systemn[UMRAA]

Step 1
•Registration at POI [primary care practice, hospital facility, social services, emergency room or at birth]

Step 2
UMRA is created on the information system database and stored on a portable electronic card

Step 3
Encrypt the UMRA card with a secure password

Step 4
•Clinical encounter - After the essential clinical and healthcare activities, the patient data is recorded in the EMR and transferred over to the UMRA card

Step 5
•If the patient visited another healthcare provider and the card is holding new data, it is transferred over to the EMR of current visit

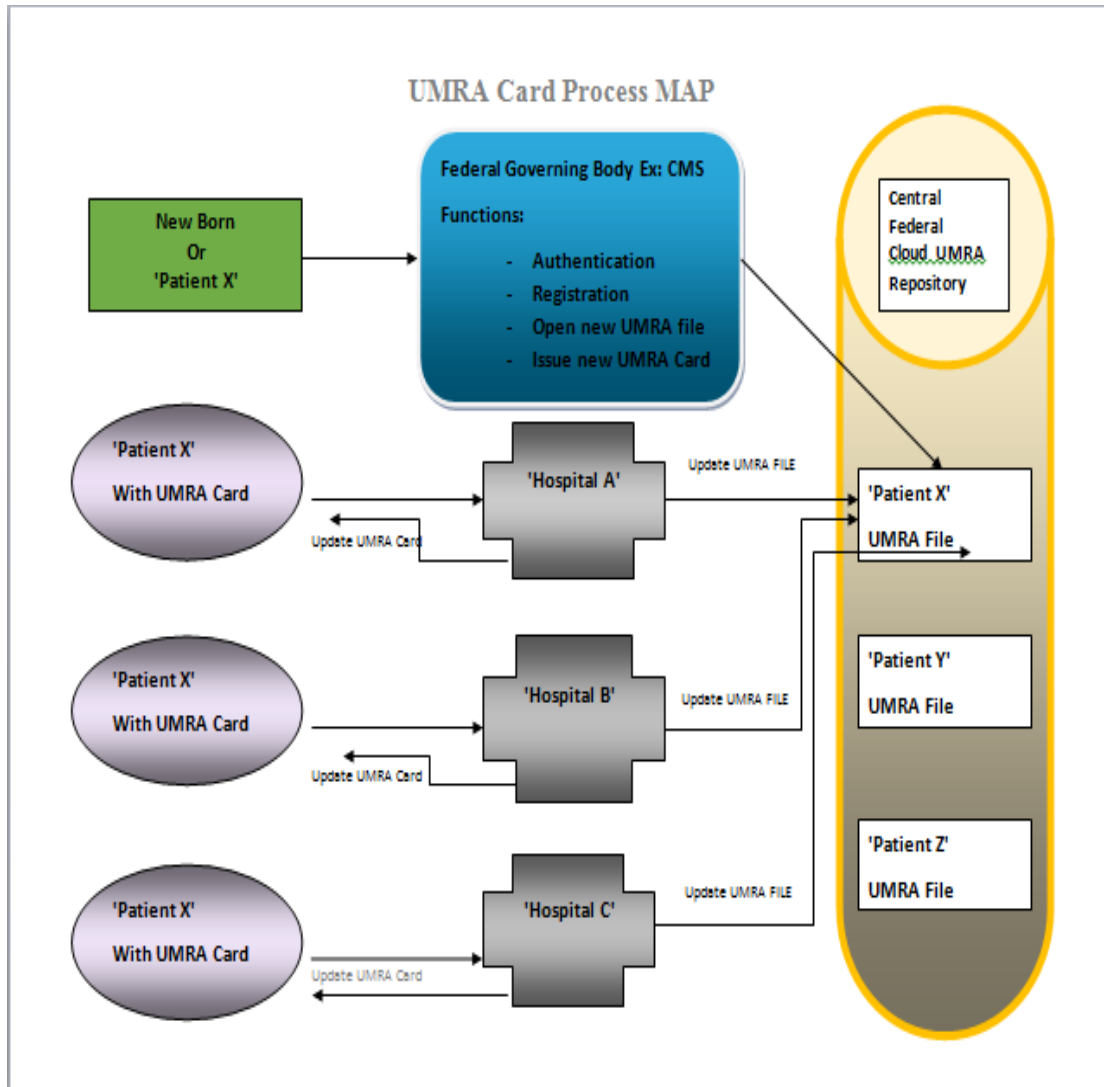Figure 3: Steps by step illustration of UMRAA process

Figure 4: The Overall UMRAA Process Map

Crowdsourcing Surveys:
Alternative Approaches to Survey Collection

Jeff Cummings
cummingsj@uncw.edu

Christopher Sibona
sibonac@uncw.edu

Information Systems and Operations Management Dept.
University of North Carolina Wilmington
Wilmington, NC 28403

Abstract

A challenge often facing survey researchers is finding an appropriate sample that is representative of the target population. Researchers are often constrained by resources (e.g., cost) and time which often limits them to using convenience sampling (e.g., student population). With the increase availability and use of crowdsourcing services, many researchers are finding a resource that provides a larger, random population to target. This paper discusses two approaches to crowdsourcing, social networking site and Amazon Mechanical Turk, to help researchers understand the options available for crowdsourcing surveys. Techniques are presented as well as benefits and potential issues to these approaches for survey research.

Keywords: Surveys, Samples, Crowdsourcing, Amazon Mechanical Turk, Social Networking Sites

1. INTRODUCTION

Surveys provide an important tool for testing various research models or retrieving the opinions of participants on certain topics. While there has been numerous studies on designing surveys, the challenge many researchers face is finding the appropriate population. Much of survey research has a common goal: provide a sampling of the population to draw conclusions to the broader population at large (Cooper & Schindler, 2013). The challenge with achieving this goal is finding the appropriate sampling of participants that would be representative to the broader population. This is exasperated by numerous constraints such as the cost of sampling and the time it takes to achieve an appropriate sample size.

The purpose of this paper is to discuss some of the challenges facing researchers conducting surveys using traditional sampling approaches followed by the popularity of alternatives approaches such as crowdsourcing. New approaches such as using social networking sites or Amazon Mechanical Turk (i.e., MTurk) provide researchers with an opportunity to collect a broader, random sample, cost effectively. These two approaches are discussed including how to setup, distribute and effectively collect data.

2. SAMPLING

The primary goal of survey research is to sample a population to make inferences based on their responses. Sampling involves the selection of subset from a particular population with the intended purpose of generalizing the results from the subset to the greater population being studied (Blumberg et al., 2014). Aims of sampling include to make an inference about an unknown parameter or to test a statistical hypothesis

relating to a certain population (Krishnaswami & Satyaprasad, 2010). Since the target population may be too broad (e.g., professionals worldwide), sampling may involve contacting local companies or professionals to participate in a survey (Bhattacherjee, 2012).

Types of sampling can include both probability (random sampling) and non-probability sampling. Random sampling typically yields more generalization and a greater accuracy of estimation of population parameters (Cooper & Schindler 2013). Conversely, non-probability sampling is often chosen based on convenience and can be used to make general inferences under specific conditions. Non-probability samples may be very helpful to understand a phenomenon that is happening at the moment in an exploratory manner (Cooper, 2008). For example, a student population may be considered a common, non-probability sample used within research when trying to generalize to the greater population. This can also be useful when the goal is instrument testing or measurement validation such as pilot testing (Bhattacherjee, 2012).

Characteristics of a good sample include (Krishnaswami & Satyaprasad, 2010):

- **Representativeness**: is the sample representative of the population you are interested in or in other words, is this a valid sample?
- **Accuracy**: is bias absent from the sample (i.e., is it free from any influence that may cause a difference b/w sample and population)?
- **Precision**: standard error and standard deviation of the sample estimate (the smaller error/estimate, higher the precision)
- **Size**: is the size of the sample large enough that an inference can be drawn with a certain level of confidence (i.e. reliability)?

Other factors/constraints to consider when choosing the appropriate survey approach include sampling precision, budget/cost, facility availability, and time (Sreejesh et al., 2014). Given the constraints and characteristics needed for a good sample, researchers must determine the most appropriate approach to collecting the sample using various survey approaches.

3. TRADITIONAL SURVEY APPROACHES

Researchers have a variety of approaches to choose from for sampling. This can be accomplished through both voluntary and paid approaches. A common approach for researchers has been mass emails to an intended subset of a population of interest. Challenges with this approach include the intended recipient may not receive the email, response rate is best within only the first few days and researchers often have to monitor/remind participants through repeated emails.

Alternatively, phoned interviews have been successful for companies such a Pew Research or McKinsey. While this can be effective, this is often expensive and does not always render higher response rates. However, both of these approached do render a more random (or probabilistic) sample. Unfortunately, due to many of the limitations listed previously, many researchers opt for a convenience (non-probabilistic) sample such as a student population which may limit the generalizability of the results.

While the previous approaches are often based on voluntary participation, an alternative approach may be to pay an external firm to collect data. Companies such as Survey Monkey price their respondents based on survey length, targeting (or worker qualifications), incidence & complexity. This again is often price prohibitive especially for researchers with limited budgets. Other companies do offer similar, more limited services often restricted to only a few questions (e.g., Pollfish and Survata). Google has even gotten into the survey collection service, starting at $0.10 for a single question, $1 for two questions and larger projects starting at $2000 per project

4. CROWDSOURCING SURVEYS

Due to the numerous constraints with survey research (e.g., budget, time, etc.), researchers have begun to use alternative approaches to the more traditional sampling methods such as mass emails or hiring external companies. The focus of this paper is on exploring the use of crowdsourcing as an alternative to data collection. Specifically, the use of social networking sites and Amazon Mechanical Turk (MTurk) as a means of crowdsourcing survey data collection.

The use of crowdsourcing is not a new topic as researchers use this technique as an approach to survey research. However, many researchers have limited knowledge of how to start using this technique as well as potential issues. The goal of the current paper is to discuss details concerning the initial setup and administration of these approaches. The benefits and drawbacks are also discussed.

Survey Tool

With either approach, a researcher must first decide on what platform to create their survey. This may be accomplished through specific university resources available (e.g., university survey system), an external company providing either a free or paid survey hosting service (e.g., Qualtrics or Survey Monkey) or through the specific service being used (e.g., Amazon Turk offers a basic survey creation tool). Much like traditional survey research, the tool chosen will depend upon the functionality needed in the survey. For example, Qualtrics has a built in function that will display certain questions dependent upon a participant's previous answers (i.e., Display Logic).

There are some considerations researchers must be aware of that may vary from traditional survey research. One is device compatibility. Since SNSs are commonly accessed via mobile devices, researchers may opt for a survey tool that is mobile ready (e.g., Qualtrics). Another consideration is the ability to interface with other sites such as Mechanical Turk. To pay participants in mTurk, a random number needs to be generated and tracked. These approaches will be discussed further in the subsequent sections concerning the approach to survey collection.

5. SOCIAL NETWORKING SITES

Because of their increased popularity and widespread use, social networking sites (SNS) have become new outlets for the recruitment of survey participants. A SNS has been broadly defined as "*a web-based service that allows individuals to (1) construct a public or semi-public profile within a bounded system, (2) articulate a list of other users with whom they share a connection, and (3) view and traverse their list of connections and those made by others within the system*" (Boyd and Ellison 2008, p.211). There are a number of different approaches depending on the SNS chosen. This paper will discuss the use of the currently popular sites Facebook and Twitter.

Facebook

Facebook currently has 1.09 billion daily active users worldwide in over 70 different languages. This provides a very diverse subject pool for sampling. A commonly used survey approach is the use of Facebook Ads for survey recruitment. Ads allow researchers target a specific audience by location, demographics, interests, behaviors and connections (details can be found at: https://www.facebook.com/business/products/ads/ad-targeting/). Ads then appear in a FB user's news feed and when clicked, will redirect the user to the external survey site. Ads are charged per click no matter if the survey is taken or not and can be set to limited number of clicks per day.

Other recruitment strategies include the creation of fan pages and groups within FB. These tend to be less effective than the placement of ads. While LinkedIn is not specifically discussed in the current paper, it does provide another avenue for targeting more professional users through similar methods such as LinkedIn groups.

Twitter

Twitter currently has over 310 million monthly active users with 79% of accounts outside the US and supports over 70 different languages (see https://about.twitter.com/company). Twitter is a micro-blogging site limiting the user to 140 character messages (Java, et al., 2007). Much like Facebook, Twitter can provide a random sample of a large population.

A popular approach is to set up an account and begin to follow organizations affiliated with the topic of interest being explored (e.g., for researchers interested in gaining insights from project managers, following PMI institute's twitter feed) or individuals prominent in field (e.g., following Bruce Schneier, a prominent security analyst, to find more users in the security community). Researchers then send tweets directly to the organization or individual asking them to retweet the survey. Setting up a hashtag for the survey also allows users to follow the number of retweets.

Sibona and Walczak (2012) used a similar approach in which they used directed messaging from one user to another (@reply mechanism) with a request to take the survey. Recruitment consisted of searching for tweets containing specific words and then sending a direct message to those users. Their research found that of 7,327 tweets/retweets concerning the survey, 2,865 users started the survey with 1,544 users completing the survey.

There are a number of suggestions for sending out a successful tweet concerning a survey. As mentioned earlier, messages are limited to 140 characters so there is a need to be concise. Miller (2011) suggests the following tweet format:

- Start with a short message about survey
- Ask users to participate
- Use a Bit.ly or TinyURL link to conserve space
- Include relevant Hashtags
- Finish by always asks for a retweet

Benefits and Drawbacks to SNS

There are a number of benefits to using SNS over traditional approaches. Researchers are able to go beyond a convenience sample to survey a more random sample of the population. Additionally, depending on the research question, Facebook allows one to target a specific population with ads. While some research has successfully used SNS for recruitment (see Sibona & Walczak, 2012), it is not without drawbacks.

Prior research has found mixed results using Ads with some finding recruitment for medical surveys to be effective (Yuan, et al., 2014) while others found this to be a more expensive approach compared to other recruitment strategies such as traditional mailings with survey information (Gu, et al., 2016). Others have concern around the representativeness of the population (especially the offline population) as well as the unknown of actual/fictitious accounts (Ahmed, 2015).

6. AMAZON MECHANICAL TURK

Amazon Mechanical Turk (MTurk) is a crowdsourcing application which allows "Requesters" to submit tasks for online "Workers" to complete. This can be a variety of tasks from software development projects (e.g., building applications) to solving business problems (e.g., data processing). It relies on a network of Workers who sign up to complete tasks called "Human Intelligence Tasks" or HITs.

The demographics of MTurk workers are estimated to be ~65% female, average age of 36, and a majority of workers in the US (47%) with a significant number, 34%, from India (Paolacci, Chandler & Ipeirotis, 2010). While not publically published by Amazon, it is estimated that there are approximately 7500 full time Workers with some estimates suggesting close to 500,000 part-time workers (Guarino, 2015).

MTurk has been used for a variety of applications in academic research. This technique has been used in a number of fields including political sentiment analysis (Diakopoulos & Shamma, 2010) and behavioral marketing studies (Collier & Barnes, 2015). Within IS research, MTurk has been used in perception studies concerning security (Kelley, 2010) or privacy (Liu et al., 2011). The goal of the following subsections is to provide researchers with an overview of conducting research on MTurk and benefits/issues with using MTurk.

Creating a New Project

Before creating a new project within MTurk, a researcher needs to setup a requester account through the website which will allow the researcher to post new projects or HITs to the system. This will link to your existing Amazon account and will also link to a credit card for purchasing prepaid HITs. The system is setup to only allow a researcher to create a project with payment that can be covered with the available balance from the prepaid HITs (can be found under "My Account").

Once an account has been created and approved, the researcher can move on to the creation of a project (or HIT) through MTurk. While the focus of this section is on the details concerning how to create an effective project, a detailed outline of creating projects is included in Appendix A. The first task is to decide on the type of project to be used. Built in functionalities of a new Turk project include general data collection, writing, transcription categorization, sentiment analysis and other various approaches. Survey options include either a link to an external survey site ("Survey Link" option) or the use of the built in survey function on Turk ("Survey" option). Since the MTurk built-in survey functionality is limited, most researchers chose external sites such as Qualtrics or Survey Monkey to use for data collection. These sites offer more functionality and data can be downloaded into various formats.

Entering Properties

The next step is defining the properties for the HITs used in the project. This includes the description, setting up the HIT and worker requirements. As part of setting up the HIT, the researcher must decide on the appropriate reward per assignment (i.e., how much will paid for completing the survey). There is much debate on the appropriate amount to pay workers for the assignment. Recent research suggests that incentivizing workers does result in higher quality data with studies suggesting the target per hour rate be close to minimum wage (Litman et al. 2015; Mitra et al., 2015). Thus, the researcher should pilot the survey to get an estimate of the time of completion to calculate the amount to reward per assignment. For example, a survey taking approximately 5 - 6 minutes to complete might be rewarded $0.50. MTurk does have a surcharge of 10%, thus the final cost per subject is effectively $0.55. This is still a much debated aspect of using MTurk especially between academicians and workers who feel they are underpaid. However, research has found the average pay rates of $0.50 to $0.75 per respondent in MTurk (Bernisky, Huber & Lenz,

2012). This is much less compared to the average non-student sample cost which can be as high as $15 to $20 per respondent (Bernisky, Huber & Lenz, 2012).

Other properties for the HITs include the number of assignment (i.e., the number of survey participants wanted), time allotted to complete the HIT once started (provide more time to complete the task than what you calculated in the pilot), time the project will be available to complete and the auto-approve of payment to workers. MTurk enables the requester to approve or reject a worker's assignment within a given period of time.   This ensures that workers respond to the survey appropriately and don't give the same value for all answers (more hints will be discussed further to limit this from occurring).   If you don't approve/reject the worker assignment, it will automatically be approved after a time set in this section.

The last section, Worker requirements, should be evaluated more closely during the creation of the project. MTurk has a classification of workers known as "Mechanical Turk Masters." These are considered to be high performing workers across multiple types of tasks in MTurk. A challenge with using this category of workers is the reward typically needs to be higher to attract these workers and surveys using these workers take longer to complete due to the limited number of MTurk Masters.   An alternative approach to attracted more qualified workers is to add additional criteria or qualifications that workers must meet to respond to the HIT. A generally accepted practice is to set "HIT Approval Rate (%)" to greater than 95 and "Number of HITs Approved" to greater than or equal to 50. This ensures the requester gets experienced workers with a have high approval rate.   A final qualification to consider is location. The requester can limit the workers to a specific (e.g., US worker only) if the target is a specific country. Aside from the system qualifications provided by MTurk, requesters also have the option of creating custom qualifications they can assign to specific workers. For example, if a specific worker consistently does quality work on surveys, the requester could create a qualification called "Quality Worker" and assign the worker this qualification. This allows the requester to include the qualification "Quality Worker" into a specific project to limit only those workers that have been assigned this qualification by the requester.   A warning against custom qualifications is it may limit the number of workers for a specific project. MTurk currently allows up to 5 qualifications.

Design Layout
In design layout, the requester provides a brief summary/instructions for the worker who wants to accept the HIT assignment. Provide enough instruction that the worker knows what needs to be accomplished but not so much information that they are overwhelmed. There is no specific limit to the amount of instructions provided but being succinct will encourage users to finish the task. Assuming an external site is being used, this is where the worker will be able to click on the link to the external site as well as provide a survey code. The survey code is used to ensure the worker completed the task and can be used by the requester to ensure the responses were appropriate.

There are two options that can be used to validate the worker completed the task. One option is to include an area in the actual survey (e.g., a input textfield in Qualtrics) which requires the worker to enter the MTurk worker ID for verification purposes. The researchers should also state that they must completely finish the HIT to receive payment. This is a simple way of verifying the worker's survey quality in which the need for a verification code is not needed. However, the preferred approach by most researchers is to include a survey code displayed at the end of the survey within their external survey tool. This requires additional work within the survey site (e.g., Qualtrics or Survey Monkey). There are links in Appendix B which walk through the steps of including a randomly generated survey code in both Qualtrics and Survey Monkey. In either option, the output from the survey results will include a unique code (either worker ID or survey code) that can be used to pay workers in MTurk.

The final step is to preview the design layout and then "Publish Batch" under the "Create" heading. Once the batch is published, the requester will receive daily reports of the workers accepting the HIT. Any emails from workers having difficulty with the assignment will also be sent so checking email frequently at the beginning of a batch is suggested in case there are any issues.

Benefits and Drawbacks to MTurk
MTurk provides solutions for many of the issues faced with traditional methods by providing a larger, more diverse subject pool with a faster experiment cycle at a lower cost (Mason & Suri 2012). However, researchers have to be aware of some of the potential quality drawbacks such as less experienced participants, awareness of manipulations and "data" quality (Chandler et al. 2014; Horton et al. 2011). Based on their research concerning naiveté of workers, Chandler

et al. (2014) suggests that researchers be aware that many workers may have seen similar experiments by other researchers on MTurk. They suggest trying to avoid common paradigms that may have already been used by other researchers and attempt to measure whether workers have engaged in a similar experiment in the past. Much like traditional survey approaches, the use of attention checks should also be used to ensure workers are engaged in the survey.

As for questions of validity, replication research comparing a traditional lab setting and an MTurk subject pool found similar results with internal and external validity present in the MTurk results (Horton, et al. 2011). Another benefit may be in the ability to do longitudinal studies. Since researchers have the ability to track and store quality workers, they have the ability to selectively choose who to respond to a subsequent survey using the custom qualification discussed in the previous sections (Paolacci, Chandler & Ipeirotis, 2010). Additionally, studies have found that MTurk workers were more attentive to instructions and responded correctly to attention checks more often than a student subject pool (Hauser & Schwarz, 2015). Thus, while there are some concerns of quality, studies have found that MTurk workers can consistently provide valid and reliable data.

7. CONCLUSION

Traditional approaches to recruiting survey participants has often been limiting based on cost and time, especially for researchers at smaller institutions with limited funding. This has led many researchers to employ convenience samples (e.g., student population) causing issues with generalizability of results.

Crowdsourcing provides an alternative approach of gaining a broader sample that can achieve the characteristics suggested for a good sample. This technique allows researchers to reach a broader sample of participants. While this does not provide an exhaustive review of all literature related to these techniques, this paper provides guidelines for researchers considering the use of crowdsourcing techniques including social networking sites and Amazon Mechanical Turk.

Survey research is often about compromises. It can be challenging and expensive to get a representative sample with any technique used. The use of MTurk may provide researchers a better representative sample for a given research questions compared to other techniques (e.g., student samples).

8. REFERENCES

Ahmed, W. (2015, September 28). Challenges of using Twitter as a data source: An overview of current resources. Retrieved from http://blogs.lse.ac.uk/impactofsocialsciences/2015/09/28/challenges-of-using-twitter-as-a-data-source-resources/

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon. com's Mechanical Turk. *Political Analysis, 20*(3), 351-368.

Bhattacherjee, A. (2012). *Social Science Research: Principles, Methods, and Practices*.

Blumberg, B., Cooper, D. R., & Schindler, P. S. (2014). *Business Research Methods*: McGraw-Hill Education.

Boyd, D., and Ellison, N. 2008. "Social Network Sites: Definition, History, and Scholarship," *Journal of Computer-Mediated Communication* (13:1), pp 210-230.

Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior research methods, 46*(1), 112-130.

Collier, J.E., Barnes, D.C. (2015) Self-Service Delight: Exploring the Hedonic Aspects of Self-Service. *Journal of Business Research,* 68 (5), 986-993.

Cooper, D., & Schindler, P. (2013). *Business Research Methods: 12th Edition*: McGraw-Hill Higher Education.

Diakopoulos, N.A., Shamma, D.A. (2010). Characterizing Debate Performance Via Aggregated Twitter Sentiment, *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Atlanta, Georgia, USA, pp. 1195-1198.

Gu, L., Skierkowski, D., Florin, P., Friend, K., Ye, Y. (2016) Facebook, Twitter, &Amp; Qr Codes: An Exploratory Trial Examining the Feasibility of Social Media Mechanisms for Sample Recruitment. *Computers in Human Behavior*, *60,* 86-96.

Guarino, B. (2015, October 15). How Many Full-Time Mechanical Turks Work for Amazon? Retrieved from

https://www.inverse.com/article/7066-how-many-full-time-mechanical-turks-are-there

Hauser, D. J., & Schwarz, N. (2015). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. Behavior research methods, 1-8.

Horton, J. J., Rand, D. G., & Zeckhauser, R. J. (2011). The online laboratory: conducting experiments in a real labor market. *Experimental Economics, 14*(3), 399-425.

Java, A., Song, X., Finin, T., & Tseng, B. (2007, August). Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*.

Kelley, P.G. (2010). Conducting Usable Privacy & Security Studies with Amazon's Mechanical Turk, *Symposium on Usable Privacy and Security (SOUPS)*, Redmond, WA.

Liu, Y., Gummadi, K.P., Krishnamurthy, B. & Mislove, A. (2011). Analyzing Facebook Privacy Settings: User Expectations Vs. Reality, *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*. ACM, Berlin, Germany, pp. 61-70.

Krishnaswami, O. & Satyaprasad, B. (2010). *Business Research Methods*. Mumbai, IN: Himalaya Publishing House.

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior research methods, 44*(1), 1-23.

Miller, J. (2011, April 27). How to Use Twitter for Crowdsourcing and Simple Market Research. Retrieved from http://www.socialmediatoday.com/content/how-use-twitter-crowdsourcing-and-simple-market-research

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on amazon mechanical turk. Judgment and Decision making, 5(5), 411-419.

Sibona, C. & Walczak, S. (2011). Unfriending on Facebook: Friend Request and Online/Offline Behavior Analysis. *44th Hawaii International Conference on System Sciences*, Kauai, HI, pp. 1-10.

Sreejesh, S., Mohapatra, S., & Anusree, M. R. (2014). *Business Research Methods: An Applied Orientation*: Springer International Publishing.

Yuan, P., Bare, G. M., Johnson, O. M., & Saberi, P. (2014). Using Online Social Media for Recruitment of Human Immunodeficiency Virus-Positive Participants: A Cross-Sectional Survey. *J Med Internet Res, 16*(5).

Editor's Note:

This paper was selected for inclusion in the journal as a CONISAR 2016 Meritorious Paper. The acceptance rate is typically 15% 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 2016.

Appendix A:
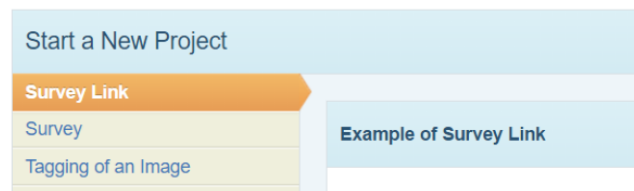**Setting up a Project in Amazon Mechanical Turk**

Requester home page (when logged-in as a Requester)

1.) Click on Create and select New Project
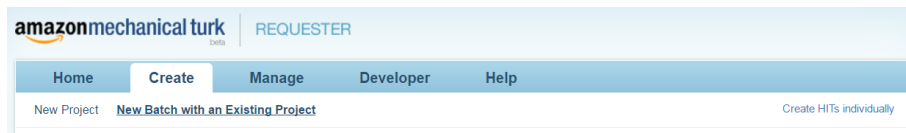
2.) The next page will be the selection of the type of project you plan on conducting. This paper has focused on the use of a "Survey Link."

3.) Enter the project title, description and keywords of the project. This is what the worker will see when searching for HITs.

4.) The next section under properties is how to set up your HIT.

5.) The last section of properties allows you to set up the worker requirements. The image below captures the suggested settings mentioned in the paper.

6.) The last step is to create the list of instructions, include the link to the survey and provide an area for entering the survey code (discussed in paper).
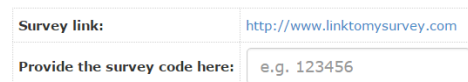

7.) Once you have saved the project, you will be taken back to the page listing your projects and the last step is to "Publish Batch." This will make the HIT available for Workers to complete.

Appendix B:
Additional Resources

Random Survey Code Generator – these are step by step guidelines to including a survey code at the end of the survey.

Qualtircs

You will need to do the following to have a survey code generated. These steps are all done within Qualtrics. This will allow the researchers to verify the survey code generated in Qualtrics with the code entered in MTurk. This can be used to ensure you only pay workers who completed the survey appropriately.

1.) Create an end of survey message (under Library tab). Select End of Survey Messages as the Category. Include the following in the message box after you have thanked the participant for completing the survey:

   MTurk Survey Code:
   ${e://Field/MTurkCode}

2.) Next, select the Edit Survey tab and then click on the Survey Options button. This will allow you to add a "Custom end of survey message…".

3.) The last step is to modify the Survey Flow which can be found under the Edit Survey Tab. Add a new element (Web Service) to the end of the survey. The URL to enter is http://reporting.qualtrics.com/projects/randomNumGen.php. Then, test the link which will allow you to "Add Embedded Data" and set "Embedded Data" to MTurkCode (the same as the piped text in step 1). You will also need to "Add parameters to send to web service" to generate the random number.

A more detailed outline of this process is outlined here:
https://tylerburleigh.com/MTurk/survey-completion-codes-in-qualtrics/

Survey Monkey:

Survey Monkey offers a similar way of generating a random survey code for workers to enter into MTurk upon completion of the survey. A detailed description of generating this code can be found at: http://nicholasnicoletti.com/blog/2015/06/survey-monkey-and-mechanical-turk-the-verification-code/

The Effects of Discount Pricing Strategy on Sales of Software-as-a-Service (SaaS): Online Video Game Market Context

Hoon S. Choi
choihs@appstate.edu

B. Dawn Medlin
medlinbd@appstate.edu

D. Scott Hunsinger
hunsingerds@appstate.edu

Computer Information Systems and Supply Chain Management
Appalachian State University
Boone, NC 28608, United States

Abstract

A discount pricing strategy is one of the most effective marketing tools to enhance sales of products in various market domains. Although it is also extensively used in the information technology (IT) industry, little prior research has examined its effects in the context of IT products. This research investigated the effects of the discount pricing in the context of the online video game market based on Software-as-a-Service (SaaS). Analyzing a large empirical panel data, this research found that the strategy has a positive effect on the sales of online video games. It also illustrated that discount rate and the amount of a discounted price have positive effects on the sales while the number of the competitors utilizing a discount pricing strategy has a negative effect.

Keywords: SaaS, Online Video Game, Discount Pricing, Panel Data, Price Fairness, Perceived Value, Utility Theory, Competition Theory

1. INTRODUCTION

Cloud computing refers to the IT-related services offered through the Internet. It includes the services for networks, servers, storage, and applications (Mell et al. 2011), and infrastructures that provide these services (Armbrust et al. 2010). Due to its effectiveness and convenience, the cloud computing services are extensively adopted in various business domains. A recent survey for IT professionals reported that 95% of the organizations are currently relying on cloud computing services for their business (RightScale 2016). According to Gartner Inc., a research group specialized in IT, and Forbes (2016), while the global market revenue of cloud computing was $58.6 billion in 2009, it reached $175 billion in 2015, recording approximately 300% growth for the last six years.

Cloud computing services can be categorized into three service models by the capability allowed to

users: Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS) (Mell et al. 2011). IaaS provides the highest level of control on the infrastructure to the users. It offers the capability to control the infrastructure including storage, networks, and other underlying computing resources. PaaS offers users the capability to build, manage, and refine the applications in the datacenters of the providers while not offering access for changing the infrastructure. The users of SaaS have the least level of control on the cloud computing infrastructures. They simply have a capability to access to applications of the service providers via a thin client interface.

Among the three cloud computing services, SaaS is the most widely used, and as such is driving the overall growth of the industry. In 2015, approximately 61% of the cloud computing industry revenue was generated from SaaS products such as human capital management, emailing, web conferencing, and web documenting services (Technology Business Research 2015).

The video game industry is one of the IT domains that actively adopt SaaS due to its capability to provide interactions among users and to update game contents. For example, a large portion of mobile game apps based on SaaS allow the users to access to the servers of game service providers via thin clients on mobile devices (Lowthorpe et al. 2013). Conventional console and PC video game markets are also moving from tangible software packages (e.g., CD) to SaaS. As an example, most game software distributed at Steam.com, which generated approximately 15% of the PC video game industry revenue in 2014, are operated through SaaS (Vellanki 2016).

Discount pricing is one of the most commonly employed marketing strategies for increasing product sales (Chen et al. 2012; Dawson et al. 2009; Sheng et al. 2007; Yin et al. 2014). In the literature, it has been an extensively investigated topic in various product domains such as apparel (Alford et al. 2002), food (Mishra et al. 2011), electronics (Della Bitta et al. 1981; Sheng et al. 2007), and automobiles (Goldberg 1996). These studies reported that discount pricing has a positive impact on the sales by affecting the consumer's value perception of the products (Alford et al. 2002; Della Bitta et al. 1981), intention to purchase and purchase incidence (Mishra et al. 2011), and net profit of the product (Lee et al. 1986; Monahan 1984).

Although the effect of discount pricing on sales has been examined in various domains, few studies investigated in the context of the SaaS industry, particularly the online video game domain. Additionally, the prior studies mainly adopted survey methods in a controlled experimental setting (Alford et al. 2002; Mishra et al. 2011; Sheng et al. 2007) or analytical modeling (Lee et al. 1986; Monahan 1984), which could not test its actual impact on sales.

In order to fill the above gap in the extant literature, this study investigates the effects of discount pricing on sales of SaaS products in the context of the online video game industry. Adopting a large panel data including 188,546 observations of 5,867 online video games, particularly, it attempts to address the following questions; (1) *"Is discount pricing effective in increasing sales of SaaS products?"*, (2) *"Does a higher level of discount rate and the amount of a discounted price induce more sales?"*, and (3) *"As more competitors offer their SaaS products at discounted prices, does the effect of discount pricing decrease?"*.

The rest of this paper is organized as follows; the literature review section discusses the literature concerning discount pricing and cloud computing. The hypothesis development introduces theories and proposes a set of hypotheses to address the major question of this study. The hypothesis test section describes the data source, empirical models, and analysis results. The discussion and conclusions section discusses the major findings of this study. Lastly, the limitations and contribution section discusses the limitations of this research and contributions to both academia and field practitioners.

2. LITERATURE REVIEW

The purpose of this research is to investigate the effects of discount pricing on the sales of a cloud computing service, SaaS. Therefore, the extant studies on discount pricing and cloud computing are discussed in this section.

Research on Impact of Discount Pricing
The primary focus of the prior studies on discount pricing is the impact of the discount on the perceived value of the product and the buying intention, employing an experimental method or an analytical modeling approach. They indicated that a higher level of price discount rate induces a higher value perception on a certain product and a higher buying intention (Alford et al. 2002; Della Bitta et al. 1981; Nusair et al. 2010). They

also reported that the impact can differ by product type (Mishra et al. 2011), whether bundled with multiple products (Sheng et al. 2007), demographics such as gender, age, and ethnicity (Goldberg 1996), initial price before discount (Coulter et al. 2007), and the level of discount rate (Coulter et al. 2007; Lee et al. 1986; Monahan 1984). However, the contexts of these studies were conventional, tangible product domains, which are different from IT services such as SaaS. Although Ghose and Han (2014) studied the impact of discount pricing on sales of mobile apps, little research investigated it in the context of the SaaS market.

Research on Cloud Computing
The major stream on cloud computing study is the conceptual discussions on a new technology, cloud computing (Armbrust et al. 2010; Buyya et al. 2009; Mell et al. 2011; Qian et al. 2009; Vouk 2008; Weinhardt et al. 2009). These studies introduced novel concepts, systems structures, stakeholders, and potential topics for cloud computing research. In the computer engineering discipline, many studies focused on the technical aspects of cloud computing. They introduced technologies for cloud computing (Ekanayake et al. 2009; Foster et al. 2008; Marinos et al. 2009; Yan et al. 2009; Zhang et al. 2010) and examined their technical performance (Calheiros et al. 2011; Jackson et al. 2010; Ostermann et al. 2009; Yu et al. 2010). In the management information systems discipline, researchers studied the perception of business practitioners on cloud computing (Leavitt 2009; Marston et al. 2011; Pearson et al. 2009), its adoption in business (Behrend et al. 2011; Ercan 2010; Kim 2011; Low et al. 2011; Sultan 2010), and its privacy and security issues (Kaufman 2009; Li et al. 2009; Subashini et al. 2011; Takabi et al. 2010). However, few studies tested the impact of discount pricing on the sales of cloud computing services.

In summary, although numerous studies in multiple disciplines investigated discount pricing and cloud computing, few covered the effects of discount pricing on the sales of cloud computing products or services. Therefore, the results of this study would provide the clarification of the relationship between discount pricing and the sales of cloud computing service, particularly SaaS based video games.

3. HYPOTHESIS DEVELOPMENT

In order to address the aforementioned research purposes, this section introduces specific hypotheses developed on the basis of the theoretical foundations adopted in the prior studies concerning discount pricing such as price fairness, perceived value, utility theory, and competition theory.

Discount Pricing and Sales of SaaS
Prior studies illustrated the effectiveness of discount pricing to increase sales (Chen et al. 2012; Dawson et al. 2009). In the digital marketplaces where SaaS products are distributed, discount pricing is known to enhance the purchase intention of online shoppers (Chevalier et al. 2003; Earl et al. 2000; To et al. 2007) as well as increasing the actual sales of the products (Ghose et al. 2014). They explained the effectiveness with two theoretical viewpoints: price fairness evaluation and utilitarian motivation of consumers.

In the evaluation of fairness of product price, consumers may use two types of price: perceived price and internal reference price (Sheng et al. 2007). Perceived price refers to the price recognized by a consumer, which is generally a listed price of a product, while internal reference price means a price which plays as a scale to evaluate the appropriateness of the perceived price. If the perceived price is lower than the internal reference price, consumers believe it is inexpensive (Kalyanaram et al. 1995; Maxwell 2002). Discount pricing can directly decrease the perceived price so that reduce its distance from the internal reference price. Therefore, consumers are more willing to buy a product when it is offered at a discounted price. In addition, it can indirectly affect the internal reference price. Consumers would perceive a product at a regular price more expensive than a product at a discounted price due to their decreased internal reference price. In both scenarios, discount pricing can provide products a higher chance to be chosen by consumers.

Another standpoint for the effectiveness of discount pricing is a utilitarian motivation of consumers, which is a critical determinant of intention to purchase. Utilitarian motivation refers to a tendency to seek for a rational, efficient, and goal driven decision to complete a task (Batra et al. 1991; Hirschman et al. 1982). Therefore, consumers with the motivation are more likely to purchase a desired product when it is offered at a discounted price because they can satisfy their need at a lower cost.

The discussion above predicts that consumers of SaaS products are more likely to purchase the

service offered at discounted prices and consequently, such products would have higher sales. Therefore, the following hypothesis is suggested;

H1: Discount pricing has a positive effect on the sales of a SaaS product.

Discount Rate and Sales of SaaS

Discount rate is a critical factor to drive the purchase decision of consumers (Chen et al. 1998; Coulter et al. 2007; Heath et al. 1995) by closely addressing the gap between the perceived price and the internal reference price of a desired product. For instance, when the internal reference price of a consumer is $100 for a SaaS product and its initial price (i.e., the perceived price) is $150, 30% discount reduces the difference from the internal reference price more than 10% discount does. As a result, the product would be more likely to be purchased when offered at 30% discount than 10% discount. In terms of utilitarian motivation, likewise, a higher discount rate would encourage the consumer to purchase it more than a lower rate since it allows the consumer to attain benefits from the product at a lower cost.

For the aforementioned reasons, the consumers of SaaS products would perceive a desired product more attractive when it has a higher discount rate. Particularly, given that the SaaS market is a highly competitive domain where multiple vendors provide similar products (Murphy 2015), a high discount rate should be a critical factor to encourage consumers to purchase and to increase the sales of a SaaS products. Thus, the following hypothesis is proposed;

H2: Price discount rate has a positive effect on the sales increase of a SaaS product.

Discounted Price and Sales of SaaS

Another important dimension of discount pricing is the amount of a discounted price, which is an actual saving in the perspective of consumers. The amount is known to generate more interest from potential consumers by increasing the perceived value of the product (Della Bitta et al. 1981). As well as increasing the purchase intention of potential customers, the discounted price stimulates market demands on the product. Prior research concerning the impact of price discounts on supply and demand illustrated that the amount of discounted price has a proportional relationship with the quantity of the product ordered (Hui-Ming et al. 1997; Lee et al. 1986) and overall sales volume (Raju 1992).

The price range of SaaS products varies, from free to higher than $7,000. In the dataset employed in this study, it ranges from $0.5 to $199. Therefore, the impact of discount rate highly differs by its initial price. For example, the saving from 20% discount on a SaaS product at $0.5 is simply $0.1 while it is $1,400 for a SaaS product at $7,000. Therefore, SaaS consumers should consider the amount of discounted price when making their purchase decision. They will be more likely to purchase a SaaS product when its discounted price is larger and therefore, the sales of the product would increase. This discussion introduces the following hypothesis;

H3: The amount of discounted price has a positive effect on the sales of a SaaS product.

Number of Competitors Offering Price Discounts and Sales of SaaS

Competitive intensity refers to the degree of competition in a product category. It should be considered in estimating the impact of price discounts on sales, particularly for the highly competitive SaaS market. It generally has a negative relationship with potential sales increase (Raju 1992), suggesting that if there are more competitors, it is more difficult to achieve sales increase. The effects of discount pricing would be subject to the competitive intensity. As more competitors offer their products at discounted prices, consumers would perceive the discount promotion less attractive and consequently, each product will have less chances to increase its sales (Kopalle et al. 1999).

In the SaaS market, consumers can easily find multiple products offered at discounted prices at their point of purchase such as Amazon Web Services and Microsoft Azure. Similar to consumers in conventional markets, they would perceive discount pricing less attractive as more discounted products are available. Therefore, the following hypothesis shown in Figure 1 is suggested:
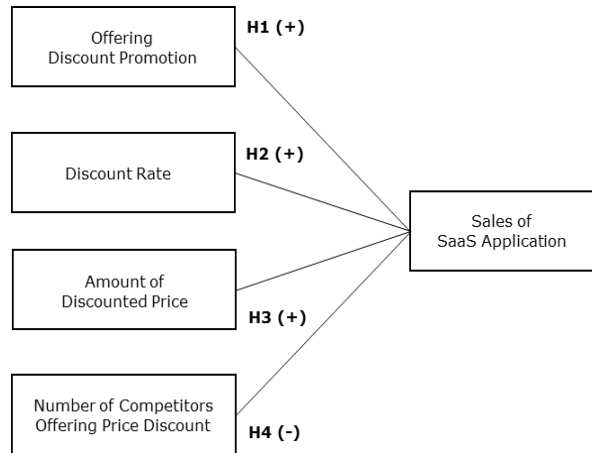
Figure 1: Research Model

Figure 1 illustrates the research model to summarize the hypotheses proposed.

H4: The number of current SaaS products offered at discounted prices has a negative effect on the sales of a SaaS product.

4. HYPOTHESIS TEST

Dataset
The data for this research were collected from two sources: steamspy.com and steamdb.info. Steamspy.com provides a sales tracking service for online video games served by Steam, which is the world largest online video game service provider based on SaaS. The number of active user accounts is almost 40 million, which accounts for more than 50% of downloadable PC games (Chiang 2011; Mudgal 2012; Reinhardt 2012). Steamspy.com has various data including daily sales, total number of owners, price, active players, and average playtime. Although the data are collected by a sampling approach using approximately 100,000 to 150,000 user accounts per day, they are known to be highly accurate within a 0.33% error margin (Gilbert 2015; Orland 2015). The data concerning discount rate and discounted price were collected from steamdb.info offering various information about video games. The data adopted in this study were collected daily for four months, from November 13 2015 to March 11 2016. The dataset includes 188,546 observations of 5,867 online video games based on SaaS platforms.

Empirical Models
Two econometric models are developed to examine the proposed hypotheses. The variables used in the models are $DailySales_{i,t}$, $DiscountDummy_{i,t}$, $Price_{i,t}$, $UserScore_{i,t}$, $Owners_{i,t}$,

$DiscountRate_{i,t}$, $DiscountedPrice_{i,t}$, and $TotalPromotions_t$. Table 1 illustrates the definitions of these variables.

Model 1 tests Hypothesis 1, testing the difference in sales between SaaS products offered at discounted prices and those at original prices. The dependent variable of Model 1 is $DailySales_{i,t}$, while its independent is $DiscountDummy_{i,t}$. It also includes three control variables, $Price_{i,t}$, $UserScore_{i,t}$, and $Owners_{i,t}$

Model 1
$DailySales_{i,t} = a_0 + a_1DiscountDummy_{i,t} + a_2Price_{i,t} + a_3UserScore_{i,t} + a_4Owners_{i,t} + \varepsilon_{i,t}$

Variables	Definitions
$DailySales_{i,t}$	Daily sales volume of a SaaS based video game i at time point t
$Discount$ $Dummy_{i,t}$	Whether a SaaS based video game i at time t at a discounted price (c.f., discounted price =1, original price=0)
$Price_{i,t}$	Actual price listed of a SaaS based video game i at time t
$UserScore_{i,t}$	The average score of consumer evaluation on a SaaS based video game i at time t
$Owners_{i,t}$	The total number of owners of a SaaS based video game i at time t
$Discount$ $Rate_{i,t}$	The percentage of price discount applied to a SaaS based video game i at time t
$Discounted$ $Price_{i,t}$	The amount of price discounted for a SaaS based video game i at time t
$Total$ $Promotions_t$	The total number of SaaS based video games offered at discounted prices at time t

Table 1: Variable Definitions

Model 2 tests Hypotheses 2, 3, and 4 to examine the impact of discount pricing. While the dependent is $DailySales_{i,t}$, the independents are $DiscountRate_{i,t}$ (H2), $DiscountedPrice_{i,t}$ (H3), and $TotalPromotions_{i,t}$ (H4). It includes $UserScore_{i,t}$ and $Owners_{i,t}$ as control variables.

Model 2
$DailySales_i = \beta_0 + \beta_1DiscountRate_{i,t} + \beta_2DiscountedPrice_{i,t} + \beta_3TotalPromotions_t + \beta_4UserScore_{i,t} + \beta_5Owners_{i,t} + \varepsilon_{i,t}$

Analysis Results

OLS (Ordinary Least Squares) estimation is initially adopted to test the proposed hypotheses. However, Breusch-Pagan test and Durbin-Watson test indicated heteroscedasticity and serial correlation in the models. This is understandable since the dataset adopted in the analysis is a panel dataset, highly subject to the violation of OLS assumptions. Therefore, GLS (General Least Squares) estimation and OLS with robust standard errors (a.k.a., robust OLS) are conducted to address these issues (Freedman 2012).

The analysis results of Model 1 are illustrated in Table 2, including R-squared, coefficients, and P-values. R-Squared of OLS and robust OLS is 0.3203, indicating that 32.03% of total variance of $DailySales_{i,t}$ is explained by the independent variables of Model 1. The results are consistent in the OLS, GLS, and OLS with robust standard errors. The coefficient for $DiscountDummy_{i,t}$ (a_1) testing Hypothesis 1 is positive and significant at the 1% level. This indicates that a SaaS product tends to have higher sales when offered at a discounted price. Therefore, Hypothesis 1 is supported.

Dependent : DailySales	OLS	GLS	Robust OLS
R-Squared	0.3203	-	0.3203
Constant	1056.88**	1056.88**	1056.88**
Discount Dummy	144.66**	144.66**	144.66**
Price (control)	37.81**	37.81**	37.81**
UserScore (control)	2191.62**	2191.62**	2191.62**
Owners (control)	0.0025**	0.0025**	0.0025**

*p < 5%, ** p < 1%

Table 2: Analysis Results of Model 1

Table 3 shows the analysis results of Model 2, testing Hypotheses 2, 3, and 4. R-squared of the OLS and robust OLS of Model 2 is approximately 0.43, indicating that the independents explain 43% of total variance of $DailySales_{i,t}$. The overall hypothesis test results remain constant in OLS, GLS, and Robust OLS. For Hypothesis 2, the coefficient for $DiscountRate_{i,t}$ (β_1) is positive and significant at the 5% level. It suggests that a higher discount rate for a SaaS application has a positive impact on its sales increase, supporting Hypothesis 2. Testing Hypothesis 3, the coefficient for $DiscountedPrice_{i,t}$ (β_2) is positive and significant at the 1% level, indicating a positive relationship between the amount of discounted price and the sales of a SaaS product. It suggests that as the difference between original price and discounted price is larger, its sales tends to increase. This supports Hypothesis 3.

With regard to Hypothesis 4, testing the relationship between the number of competitors offered at discounted prices and sales of a SaaS product, the coefficient for $TotalPromotions$ (β_3) is negative and significant at the 1% level. It suggests that as more SaaS products at discounted prices are available for consumers, each SaaS product tends to have lower sales. Therefore, Hypothesis 4 is supported.

Dependent : DailySales	OLS	GLS	Robust OLS
R-Squared	0.4296	-	0.4298
Constant	1018.5**	1018.05**	1018.05**
Discount Rate	558.38*	558.38*	558.38*
Discounted Price	12.55**	12.55**	12.55**
Total Promotions	-0.074**	-0.074**	-0.074**
UserScore (control)	25.28**	25.28**	25.28**
Owners (control)	0.0032**	0.0032**	0.0032**

*p < 5%, ** p < 1%

Table 3: Analysis Results of Model 2

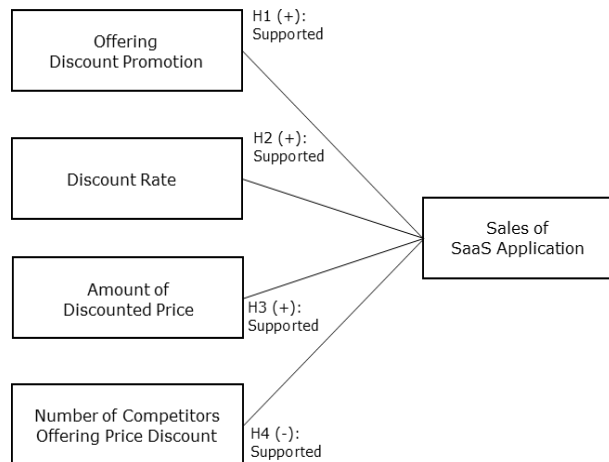Figure 2 summarizes the analysis results of the above hypothesis test.

Figure 2: Hypothesis Test Results

5. DISCUSSION AND CONCLUSION

This research investigated the impact of price discounts on the sales of SaaS products in the context of the online video game market. Specifically, it examined the effects of discount rate and discounted price, as well as the number of the competitors offered at discounted prices. The hypothesis test results concerning these factors suggest the following findings.

First, support for Hypothesis 1 shows that SaaS products offered at discount prices tend to have higher sales than those at non-discounted prices. This result is consistent with the findings of the extant literature based on the theories of price fairness evaluation and utilitarian motivation. This also implies that consumers in the SaaS market have a similar consumer behavior with those in conventional market sectors. This also suggests that discount pricing would be an effective strategy to increase sales in the SaaS market. Therefore, practitioners in the domain may consider the strategy to increase the sales of their products.

Concerning the level of price discounts, both discount rate and the amount of discounted price are found to have a positive impact on the sales of SaaS products. Support for Hypothesis 2 indicates that discount rate has a positive relationship with SaaS application sales. Support for Hypothesis 3 implies that the amount of discounted price has a positive relationship with the sales. As well as discount rate, therefore, the amount of discounted price plays an important role in increasing sales of SaaS products. Practitioners in the market need to consider both promotional factors, therefore, when planning their discount pricing strategies.

Support for Hypothesis 4 suggest that the number of competitors offering price discounts has a negative relationship with the sales of SaaS products. Therefore, as the number of the products offered at discounted prices increases, the sales of each product would decrease. This finding provides a meaningful implication to practitioners who consider at what time they should offer a price discount. Discount pricing would be more effective when less competitors are using the same strategy than when more are.

6. LIMITATIONS AND FUTURE RESEARCH

Although this is one of first empirical studies to investigate the impact of discount pricing in the SaaS domain, there are several limitations, particularly with regard to the data used. This study adopted a dataset for online video games based on SaaS. Although they are a type of SaaS, they are categorized into hedonic products, consumed for entertainment and enjoyment. However, the typical SaaS products are used for practical purposes, such as conferencing, and web documenting services, and emailing. Therefore, the analysis results of this study concerning the impact of discount pricing may not be applicable to the other SaaS product types. Testing Hypothesis 4, this study did not distinguish direct and indirect competitors in its analysis. For instance, a video game in RPG (Role Playing Game) genre would not directly compete with those in different genres, nor significantly related to the sales of the other genres of video games. However, they were not separated in the analysis and therefore, the result may be different if estimated with more thoroughly categorized data. Finally, the dataset includes the sales for only four months. Future research may adopt a more comprehensive, large dataset that can provide more generalizable findings and implications from the analysis.

7. REFERENCES

Alford, B. L., and Biswas, A. 2002. "The effects of discount level, price consciousness and sale proneness on consumers' price perception and behavioral intention," *Journal of Business research* (55:9), pp 775-783.

Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Konwinski, A., Lee, G., Patterson, D., Rabkin, A., Stoica, I. O. N., and Zaharia, M. 2010. "A View of Cloud Computing,"

Communications of the ACM (53:4), pp 50-58.

Batra, R., and Ahtola, O. T. 1991. "Measuring the hedonic and utilitarian sources of consumer attitudes," *Marketing letters* (2:2), pp 159-170.

Behrend, T. S., Wiebe, E. N., London, J. E., and Johnson, E. C. 2011. "Cloud computing adoption and usage in community colleges," *Behaviour & Information Technology* (30:2), pp 231-240.

Buyya, R., Yeo, C. S., Venugopal, S., Broberg, J., and Brandic, I. 2009. "Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility," *Future Generation computer systems* (25:6), pp 599-616.

Calheiros, R. N., Ranjan, R., Beloglazov, A., De Rose, C. A., and Buyya, R. 2011. "CloudSim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms," *Software: Practice and Experience* (41:1), pp 23-50.

Chen, H., Marmorstein, H., Tsiros, M., and Rao, A. R. 2012. "When more is less: The impact of base value neglect on consumer preferences for bonus packs over price discounts," *Journal of Marketing* (76:4), pp 64-77.

Chen, S.-F. S., Monroe, K. B., and Lou, Y.-C. 1998. "The effects of framing price promotion messages on consumers' perceptions and purchase intentions," *Journal of retailing* (74:3), pp 353-372.

Chevalier, J., and Goolsbee, A. 2003. "Measuring Prices and Price Competition Online: Amazon.com and BarnesandNoble.com," *Quantitative Marketing and Economics* (1:2), pp 203-222.

Chiang, O. 2011. "The Master of Online Mayhem," Forbes: Forbes.

Coulter, K. S., and Coulter, R. A. 2007. "Distortion of price discount perceptions: The right digit effect," *Journal of Consumer Research* (34:2), pp 162-173.

Dawson, S., and Kim, M. 2009. "External and internal trigger cues of impulse buying online," *Direct Marketing: An International Journal* (3:1), pp 20-34.

Della Bitta, A. J., Monroe, K. B., and McGinnis, J. M. 1981. "Consumer Perceptions of Comparative Price Advertisements," *Journal of Marketing Research* (18:4), pp 416-427.

Earl, P. E., and Potts, J. 2000. "Latent demand and the browsing shopper," *Managerial and Decision Economics* (21:3-4), pp 111-122.

Ekanayake, J., and Fox, G. 2009. "High performance parallel computing with clouds and cloud technologies," in *Cloud Computing*, Springer, pp. 20-38.

Ercan, T. 2010. "Effective use of cloud computing in educational institutions," *Procedia-Social and Behavioral Sciences* (2:2), pp 938-942.

Forbes, G. a. 2016. "Global revenue generated from public cloud computing services from 2009 to 2016 (in billion U.S. dollars)," Statista.

Foster, I., Zhao, Y., Raicu, I., and Lu, S. Year. "Cloud computing and grid computing 360-degree compared," Grid Computing Environments Workshop, 2008. GCE'08, Ieee2008, pp. 1-10.

Freedman, D. A. 2012. "On the so-called "Huber sandwich estimator" and "robust standard errors"," *The American Statistician*).

Ghose, A., and Han, S. P. 2014. "Estimating demand for mobile applications in the new economy," *Management Science* (60:6), pp 1470-1488.

Gilbert, D. 2015. "Let's talk steam spy," Wadjeteye Games: Wadjeteye Games.

Goldberg, P. K. 1996. "Dealer Price Discrimination in New Car Purchases: Evidence from the Consumer Expenditure Survey," *Journal of Political Economy* (104:3), pp 622-654.

Heath, T. B., Chatterjee, S., and France, K. R. 1995. "Mental accounting and changes in price: The frame dependence of reference dependence," *Journal of Consumer Research*), pp 90-97.

Hirschman, E. C., and Holbrook, M. B. 1982. "Hedonic consumption: emerging concepts, methods and propositions," *The Journal of Marketing*), pp 92-101.

Hui-Ming, W., and Yu, J. 1997. "A deteriorating inventory model with a temporary price discount," *International Journal of Production Economics* (53:1), pp 81-90.

Jackson, K. R., Ramakrishnan, L., Muriki, K., Canon, S., Cholia, S., Shalf, J., Wasserman, H. J., and Wright, N. J. Year. "Performance analysis of high performance computing applications on the amazon web services cloud," Cloud Computing Technology and Science (CloudCom), 2010 IEEE Second International Conference on, IEEE2010, pp. 159-168.

Kalyanaram, G., and Winer, R. S. 1995. "Empirical generalizations from reference price research," *Marketing science* (14:3_supplement), pp G161-G169.

Kaufman, L. M. 2009. "Data security in the world of cloud computing," *Security & Privacy, IEEE* (7:4), pp 61-64.

Kim, W. 2011. "Cloud computing adoption," *International Journal of Web and Grid Services* (7:3), pp 225-245.

Kopalle, P. K., Mela, C. F., and Marsh, L. 1999. "The dynamic effect of discounting on sales: Empirical analysis and normative pricing implications," *Marketing Science* (18:3), pp 317-332.

Leavitt, N. 2009. "Is cloud computing really ready for prime time?," *Computer*:1), pp 15-20.

Lee, H. L., and Rosenblatt, M. J. 1986. "A Generalized Quantity Discount Pricing Model to Increase Supplier's Profits," *Management Science* (32:9), pp 1177-1185.

Li, H., Dai, Y., Tian, L., and Yang, H. 2009. "Identity-based authentication for cloud computing," in *Cloud computing*, Springer, pp. 157-166.

Low, C., Chen, Y., and Wu, M. 2011. "Understanding the determinants of cloud computing adoption," *Industrial management & data systems* (111:7), pp 1006-1023.

Lowthorpe, C., Taylor, S., and White, G. 2013. "Stop just making stuff! Listening, co-creation and sustainability in independent game development," *Journal of Audience & Reception Studies* (10:2), pp 275-297.

Marinos, A., and Briscoe, G. 2009. "Community cloud computing," in *Cloud Computing*, Springer, pp. 472-484.

Marston, S., Li, Z., Bandyopadhyay, S., Zhang, J., and Ghalsasi, A. 2011. "Cloud computing—The business perspective," *Decision support systems* (51:1), pp 176-189.

Maxwell, S. 2002. "Rule-based price fairness and its effect on willingness to purchase," *Journal of economic psychology* (23:2), pp 191-212.

Mell, P., and Grance, T. 2011. *The NIST definition of cloud computing. [electronic resource]*, (Gaithersburg, MD : Computer Security Division, Information Technology Laboratory, National Institute of Standards and Technology, [2011].

Mishra, A., and Mishra, H. 2011. "The influence of price discount versus bonus pack on the preference for virtue and vice foods," *Journal of Marketing Research* (48:1), pp 196-206.

Monahan, J. P. 1984. "A Quantity Discount Pricing Model to Increase Vendor Profits," *Management Science* (30:6), pp 720-726.

Mudgal, K. 2012. "Valve releases PR; Steam userbase doubles in 2011, Big picture mode coming soon," gamingbolt.

Murphy, L. 2015. "SaaS Market Positioning: How to Compete in Crowded Markets."

Nusair, K., Jin Yoon, H., Naipaul, S., and Parsa, H. 2010. "Effect of price discount frames and levels on consumers' perceptions in low-end service industries," *International Journal of Contemporary Hospitality Management* (22:6), pp 814-835.

Orland, K. 2015. "Introducing Steam Gauge: Ars reveals Steam's most popular games," ars technica.

Ostermann, S., Iosup, A., Yigitbasi, N., Prodan, R., Fahringer, T., and Epema, D. 2009. "A performance analysis of EC2 cloud computing services for scientific computing," in *Cloud computing*, Springer, pp. 115-131.

Pearson, S., Shen, Y., and Mowbray, M. 2009. "A privacy manager for cloud computing," in *Cloud Computing*, Springer, pp. 90-106.

Qian, L., Luo, Z., Du, Y., and Guo, L. 2009. "Cloud computing: an overview," in *Cloud computing*, Springer, pp. 626-631.

Raju, J. S. 1992. "The Effect of Price Promotions on Variability in Product Category Sales," *Marketing Science* (11:3), pp 207-220.

Reinhardt, J. 2012. "Indie Distribution Platforms that are not Steam," Gamasutra.

RightScale 2016. "Cloud Computing Trends: 2016 State of the Cloud Survey," Cloud Management.

Sheng, S., Parker, A. M., and Nakamoto, K. 2007. "The Effects of Price Discount and Product Complementarity on Consumer Evaluations of Bundle Components," *Journal of Marketing Theory and Practice* (15:1), pp 53-64.

Subashini, S., and Kavitha, V. 2011. "A survey on security issues in service delivery models of cloud computing," *Journal of network and computer applications* (34:1), pp 1-11.

Sultan, N. 2010. "Cloud computing for education: A new dawn?," *International Journal of Information Management* (30:2), pp 109-116.

Takabi, H., Joshi, J. B., and Ahn, G.-J. 2010. "Security and privacy challenges in cloud computing environments," *IEEE Security & Privacy*:6), pp 24-31.

Technology Business Research 2015. "The State of SaaS in 2015: Global Market Trends, Growth Forecasts, Industry Reports," Technology Business Research Inc.

To, P.-L., Liao, C., and Lin, T.-H. 2007. "Shopping motivations on Internet: A study based on utilitarian and hedonic value," *Technovation* (27:12), pp 774-787.

Vellanki, M. 2016. "This Is The Current State Of The Video Game Industry."

Vouk, M. 2008. "Cloud computing–issues, research and implementations," *CIT. Journal of Computing and Information Technology* (16:4), pp 235-246.

Weinhardt, C., Anandasivam, D.-I.-W. A., Blau, B., Borissov, D.-I. N., Meinl, D.-M. T., Michalk, D.-I.-W. W., and Stößer, J. 2009. "Cloud computing–a classification, business models, and research directions," *Business & Information Systems Engineering* (1:5), pp 391-399.

Yan, L., Rong, C., and Zhao, G. 2009. "Strengthen cloud computing security with federal identity management using hierarchical identity-based cryptography," in *Cloud Computing*, Springer, pp. 167-177.

Yin, X. U., and Jin-Song, H. 2014. "Effects of Price Discounts and Bonus packs on Online Impulse Buying", *Social Behavior & Personality: an international journal* (42:8), pp 1293-1302.

Yu, S., Wang, C., Ren, K., and Lou, W. Year. "Achieving secure, scalable, and fine-grained data access control in cloud computing," Infocom, 2010 proceedings IEEE, Ieee2010, pp. 1-9.

Zhang, Q., Cheng, L., and Boutaba, R. 2010. "Cloud computing: state-of-the-art and research challenges," *Journal of internet services and applications* (1:1), pp 7-18.