CONFIDENTIALITY, PRIVACY, ACCESSIBILITY AND SECURITY OF BIG DATA USAGES WITHIN MOBILE RECOMMENDER SYSTEMS, AS SOCIETY EMBRACES CLOUD BASED TECHNOLOGY

Jacob E, Mack, Keiser University, j.mack7@student.keiseruniversity.edu
Theophilus D. Owusu, Keiser University, towusu@keiseruniversity.edu
John J. Scarpino, Pittsburgh Technical College, scarpino.john@pti.edu

ABSTRACT

Recommender systems have a varied history concluding into modern day where more sophisticated automated integrated computer cloud based suggestion and analysis systems exist. As recommender systems become more contextual and deeper in learning analysis based on the data the systems, pulls there arises more confidentiality, privacy, accessibility, and security issues regarding end user’s data. Key identifying information from this data can be information that are confidential and of security in nature. Data information from movie preferences on Netflix for example via social media sites like Facebook can lead to both demographic data sharing to acquaintances and strangers as well as increase the risk for data hacking. Furthermore, social media sites are in part based upon recommender systems and not search engine technologies. A true recommender system filters even greater information than social media sites alone. There is a significant difference between a system that recommends choices for you and a system that does not. Google.com is a great example that based on your search and who you are the result of the data differs. This research paper will examine key confidentiality, privacy, accessibility, and security issues within the context of recommender systems with several key examples. The researcher’s use of Netflix as the key model for this aforementioned examination and analysis are combined with other relevant examples and are also investigated. A recent encryption method, homomorphic encryption is proposed and discussed.

Keywords: Confidentiality, Privacy, Cryptographic Security Issues, Mobile Recommender Systems, Mobile Information Systems, Social Media, Homomorphic Encryption, Machine Learning Algorithms, Artificial Neural Networks (ANNs), Mobile Searches, Mobile GPS, Mobile Location Searches.

INTRODUCTION

Recommender systems are becoming an ever more prevalent mainstay within a fast paced, information data base, and global community. Society now relies upon faster, more efficient and accurate information retrieval systems. This truth cannot be overstated for business at large, and social media, e-commerce sites that generate revenue on successful conversions of suggestions to final sales. Still, with more fast-paced technology that reveals more sensitive information from explored data, confidentiality, privacy, accessibility, and security outside of need to know stakeholders have become major research and industry issues. This research is meant to raise awareness of the subject matter, and current confidentiality, privacy, accessibility, and security issues.

RECOMMENDER SYSTEMS

Recommender systems are now considered as applications that can be software and cloud based technology acting in the middle tier between end users and servers. Often too, search engines employ in part recommender systems to create more relevant and specialized suggestions for searchers from a data pool from an inverted index listing. However, Recommender systems have a far broader and varied technological history within the use of: Personal Digital Assistants, (PDA’s), modern Smart Phones, tablets, web and e-commerce sites, Google New, Netflix, various other web and e-commerce sites, and various other mobile devices. In addition, recommender systems provide finer information data sharing than social media sites like Facebook with confidentiality, privacy, accessibility and security controls implemented, and recommender systems continue to collect more context conscious data.
Recommender systems might be found in the form of guided manual ratings, or more in the form of automated clustering of preferences as in the application of some Machine Learning Algorithms. For example when an end user manually rates a specific movie or shares it as generally recommended to friends on Facebook this is a guided item rating/selection. The Recommender system will then generate suggestions based upon the specific item the user chose to recommend or conversely indicates different movies based upon a low rating and/or lack of general recommendation.

Facebook and specifically within movie ratings as mentioned is a guide dominated in terms of recommendations. To its end users. However, the system may also generate additional suggestions outside the scope of the given genre, sub-genre, demographic data, and other item rating/recommendations made by the end user within the Facebook system. This introduction of randomness or less related recommendations maybe the result of using a Genetic algorithm, artificial neural network (ANNs), or some other Machine Learning Algorithm to uncover information data deeper within user selections, and not just manually driven, guided ratings. By selection. Though, such detailed data analysis goes into deeper information filtering and user choices beyond their item rating and selection. As such, there, arises, more confidentiality, privacy, accessibility, and security concerns in terms of vendor and Social Media storage practices of key information data without the end user’s knowledge or at least permission. In addition, as with any one or more complex systems more issues can potentially arise with data storage, retrieval and sharing of information data, which may lead into more uncontrollable privacy vulnerabilities and hacking exploits.

The purpose of this research is to analyze and investigate deeper into modern day applications of recommender systems. We will review the general threats to confidentiality, privacy, accessibility and security, as well as, discuss specific threats while proposing a not yet realized but realistic solution, homomorphic encryption. This research relies upon both a robust literature review and the researcher’s own primary data taken from a survey of Information Technology, Information Systems, Computer Science Business Technology, and Computer Security professionals. As Recommender Systems become more accepted in practice and able to perform more widespread deep/big data analysis innovative issues arise in terms of privacy and security.

**LITERATURE REVIEW**

Lima J., Rocha C., Augustin I., & Dantas (2002) describe recommender systems in terms of management systems for information overload early on within its modern history beginning in the middle 1990’s. Such information overload can be the tens of thousands of movies and TV shows to choose from on Netflix or music catalogs on Apple’s iTunes. However the researchers in this paper also describe an underlying context that is more nuanced and complex in terms of the situation one might make a selection or series of selections. For example, choosing an arbitrary number of 10 movies from within 2 genres and 3 sub-genres on Netflix may generate 100 arbitrary recommendations from within the integrated system. However, there might be some underlying fundamental aspects and elements that can either reduce the number of recommendations to 65 movies with greater accuracy and precision. On the other hand, the context aware system may gain information from data analysis and suggest additional genres/sub-genres with high relevancy even if the movie total are increased, there might be more robust options for the viewer.

As to why these issues of recommender systems are important, even 9-10 years ago, Fesenmaier, Ober, and Werthner (2006) describe it well when they explore the myriad of recommendation systems that were utilized to analyze data within and on hundreds of different e-commerce sites and to support the many millions of consumers in their goals of finding both products and services. According to Bergemann and Ozmen (2006) recommender systems affect optimal pricing, and there is a two pronged approach to how they affect the market: generating information that reduces item/product uncertainties and offering recommender systems as a result of such reduced uncertainty which can accurately analyze external data. Now in 2014-2015 the figures for ubiquitous use of Recommender systems is only on the incline.

Significant research exists as to why context aware systems are being investigated and engineered to solve a myriad of issues in recommendation performance. For example, Adomavicus and Jannach (2013) state that even as recommender systems become increasingly fine-tuned and sophisticated from both academic and industry based
research, they still lack a robust analysis of specific contexts like: time, location, weather, user’s goals, mood, and mood influenced by others present. Also the type of device can pose specific information filtering performance reduction depending upon the specific hardware/software interface. In addition other variables like: one’s cultural milieu, subculture, one’s age, religious beliefs, or lack thereof, familial status, and personal ethics/morals can all be important features of analysis within a context aware system.

Demographics then would be an important feature in any recommender system analysis. Adomavicius & Tuzhilin (2005) refers to the list of features to as the feature space and in fact in the author’s professional and academic experience with working with recommender systems, this is a common naming convention and plays an important role in contextual information analysis. Another shortcoming within recommender systems is over specialization where recommendations for user items (as selections are often referred to, whether movies, music or search engine selections) are primarily based upon the former choices made by the user according to Adomavicius & Tuzhilin (2005).

However, such analysis of more contexts along with deeper variable (feature) extraction can open up new vulnerabilities for data leakage via: phishing, Trojans, Viruses, and out of date security protocol backdoors. Furthermore, with that information analysis, storage and retrieval it is not clear as to whether vendors will always use such information within the strictest confidence and not sell such valuable data Gentry, 2009; Stefania, Patrascu, & Simion, 2012). To make the context aware recommender systems more of a reality, robust and efficient in terms of time expenditure machine learning is often employed in some variant of a hybrid system according to Ghazanfar, & Prugel-Bennett (2010). Machine Learning is broadly defined as: programming a computer to handle computations that humans perform well as everyday tasks, but difficult to write out in the form of an algorithm (Ayodele, 2010). An additional broader and more commonly used definition might be summarized as: “a machine learns with respect to a particular task T, performance metric P, and type of experience E, if the system reliably improves its P at task T…” (Mitchell, p. 3, 2006). Essentially, a Machine Learning system must learn over time and with performing the relevant actions/transactions and learn from its own successes and failures.

Think through Netflix as the central recommender system model. Netflix has a wide range of movies, documentaries and TV shows to choose from. Within these general headings lie genres, sub-genres and how other users who rated similar to you rated them as well. Just from this general explicitly programmed analysis there already exists a myriad of data as personal identifiers and showing demographic/browsing habits. Now add an advanced ML system that can break apart smaller components of a viewer’s ratings/choices and actual search engine browsing habits and it becomes clear that at the very least privacy becomes a major issue. With Facebook storing and analyzing Netflix data even without the end user’s consent there exists a leak potential even if Facebook/Netflix respects a do not share data request from the user. What has become more prevalent in recent years are hacker exploits of personal data like: home addresses, private emails, chat messages, and bank account numbers. As two recent events in the news consider, the hacking of PayPal by the group Anonymous, and Sony emails with all the data taken and shared publicly from their servers.

Context aware systems then in general can analyze, store, retrieve, disseminate, and apply more vast amounts of data, filter more information from said data and provide more specifically robust recommendations based upon said methodology (Arora, Kumar, Devre, & Ghumare, 2014; Baltrunas, & Ricci, 2010; Ekstrand, Riedl, & Konstan, 2011). Whether machine learning based or some newer hybrid system context aware systems also present privacy and security issues that must be explored further and new solutions must be presented with robust testing. While free/low cost Ad blocker, Disconnect and light search engine scanners all provide some measure of protection, more complex engineered systems are still required.

Modern day cryptographic and other computer security methodologies are now being applied to better ensure both privacy and security. There are a number of encryption methods, protocol key exchanges, privacy filtering approaches, and Business Intelligence data firewalls that are and can be implemented. However, there exists variability in knowledge/skill sets and budgets applied within organizations to such necessary applications. For example some IT/IS/CS professionals may not be aware of more effective systems for recommendation and their related security/privacy insurance. Moreover, organizations may have some awareness but prefer higher short term profit margins to more expensive and time consuming integration of more secure systems in the present. In fact,
According to Owusu and Hoffman (2014) there exists considerable business bias in terms of how a recommender system is employed in data collection, analysis and cryptographic protection. In the short run it may not be profitable for a given company to employ or research the best possible data integrity protection techniques as such developments and implementations would be costly in terms of financial, time and marketing resources. Imagine if Netflix had to go offline for a month because it decided to fix known vulnerabilities to the cloud server system, security monkey, designed to make Amazon Web services public cloud service. While there is no doubt code updates will continue to be necessary, as real world recommender system security is always less than theoretical or ideal, there now exists exciting technologies and well developed research to close the gap further on recommender system security and privacy issues Ekstrand, 2014; Subrata, & Gorman, 2013).

 Nonetheless, with the rise of mobile device hacking, theft and sharing of personally identifiable data, modern day security and specifically cryptography has become a more relevant and hot topic for both mobile devices in general and within the context of recommender systems that can be exploited in a number of ways Gentry, 2009; Lima, Rocha, Augustin, & Dantas, 2002). The researchers will provide a brief summary of some of the potentially more effective solutions.

 A recent development in encryption not only solves many encryption specific issues but also reduces the load placed upon protective protocols like SSL/TSL and some forced HTTPS everywhere scheme. The encryption application is known as homomorphic encryption and it was developed by Gentry (2009) but it has since become a widespread experimental and less costly area of engineering research since its inception. Homomorphic encryption though, in part theorized long before Gentry’s work within his PhD dissertation and later in conjunction with other researchers at IBM was not possible to create and test prior to Gentry’s (2009) work. Now Microsoft/IBM and other technology companies offer free open source code for development, a myriad of research articles on the subject and provide a framework for more practical applications. Conversely, there are still engineering, IS, IT, and CS applications not yet exploited or fully investigated. One such practical application would involve a streaming site like Netflix. To understand how this, work it is first necessary to understand the fundamentals of homomorphic encryption.

 Homomorphic can be broken down to its constituent roots, which refers to same shape and this is exactly how homomorphic encryption works. Data sent from a client side user is encrypted in transit as it usually is over an insecure channel, however; in this case when it reaches a vendor server it remains encrypted. However, the “black box” hiding the personal data retains the same shape as it had in transit and it reveals nothing to the server side of its specific data types, in this case streaming preferences (Chung, Kalai, & Vadhan,2009; Gentry, 2009). Yet the server can still perform mathematical operations upon the encrypted data and send it back to the client/end user where they can decrypt the message and see novel recommendations (Gentry, 2009). Thus, there can be a robust context recommender system without the need for data leakage/interception and decryption of the message data at least in transit and server side.

 **PROBLEM ANALYSIS**

 The past 4-5 years has seen a multitude of improved and more sophisticated context aware recommender systems (Baltrunas & Ricci, 2013). More recent developments over the past two years has led to new engineering process of actual intelligent knowledge management systems that better filter information in novel ways via hybrid systems (Adomavicius, & Jannach, 2013; Ayodele, 2010). Such systems often introduce complex machine learning algorithms to improve the behavior of learning from experience to improve performance (Ekstrand, 2014). However, with the advent of big data analytics deep learning techniques and larger cloud based server storage of data, information leakage becomes a larger concern and ongoing issue (Sangwan, & Rathi, 2014).

 In fact there has been an explosion of the engineering of machine learning and genetic algorithm based context aware recommender systems. Such aforementioned systems are capable of both learning more from user item ratings and browsing. The integrated systems can as well as, add additional recommendations based upon a similarity/dissimilarity index (Ekstrand, 2014; Kim & Kim, 2010).. Adomavichius and Jannach (2013) highlight the importance of intelligent systems in how they behave at filtering retrieved data as useful information. With Gentry’s (2009) work we now have novel ways in which to explore security solutions within context aware Recommender systems.
RESEARCH METHODOLOGY

The researchers designed a survey with 8 questions in Likert scale in form with two open ended fill in questions, making this research both quantitative and qualitative in nature. This aforementioned approach is known as mixed methodology. The sampling frame was not randomized and only specifically professionals within: Information Technology, Information Systems, Computer Science, Business Technology, and Computer Security, were asked to fill out the survey. The survey was created using Google Docs. A combination of snowball sampling, and promoting the survey on Facebook and LinkedIn Technology groups was employed. There were a total of 44 participants. Being that this survey is largely Likert Scale in basis it is then ordinal in design and not categorical so in general, non-parametric statistics or percentage frequencies are the preferred method for quantitative analysis. In this instance due to the sample size and data already provided in the Google document instruments, frequencies are shown in the results section from said instrument. As such Initial data was collected and analyzed from the Google Docs Spreadsheet, Google Analytics and related histograms. The two fill in questions asked about the participant’s professional background and the other regarding how much experience if any within Computer Security. Our findings show significant concerns about: emerging technologies and the importance of recommender system research, privacy/security issues with recommender systems, mobile device security, and to a lesser extent, current encryption methods.

RESULTS

The researchers found significant data from the survey to support the claims that recommender system privacy and security are important areas to perform further research into. The literature review of both historical and recent peer reviewed research papers also strongly supports this claim. The researchers, therefore rejects the null hypothesis that there is no need for more research into new and improved security/cryptographic methods. From both the survey and the literature review it becomes apparent there is a strong need to continue to research, engineer, test, and implement various more secure integrated recommender systems and security measures. Below is the complete summary results of the data taken by Google Docs and analyzed by Google Analytics:

**Question 1:** The research around mobile device security is still an ongoing professional pursuit in the face of emerging technologies.
Question 2: Recommender systems present new challenges to privacy and security concerns regarding personal data.

| 1 | 0 | 0% |
| 2 | 1 | 4% |
| 3 | 2 | 8% |
| 4 | 12 | 50% |
| 5 | 9 | 38% |

Question 3: As computing power increases, devices like smartphones and tablets are becoming increasingly susceptible to hackers and various malicious software.

| 1 | 0 | 0% |
| 2 | 0 | 0% |
| 3 | 3 | 13% |
| 4 | 8 | 33% |
| 5 | 12 | 50% |

Question 4: Current encryption methods are insufficient to protect sensitive data.
Question 5: Describe your current professional work role(s) and technology background.

A.) Grad student
B.) Computer Science
C.) Computer Science Research, AI
D.) I am currently leading a group at Mob.Net. I am the CEO of this company.
E.) Involved with IT/IS Computer Science
F.) Developer for Web and Native Platforms
G.) yes it is
H.) I am a PhD student in Computer Science
I.) Business.
J.) I work as trainee in the area of support in a College.
K.) Software Engineer.
L.) Director of Production in IT/IS industry
M.) IT advisor for IT system acquisition in Defense area.
N.) web development
O.) Software Engineer
P.) My bachelor degree is Computer Science and I am PhD student in Electrical Engineering. I have worked in Information Systems Management and Computer Programming around 7 years.
Q.) I have 4 years of software programmer experience, 1.2 years of java programmer, 2.8 years of android programmer, I am working as a technical design and research analyst in mobile app development.
R.) Hassan Mahmoud is a Lecturer assistant since 2006 and PhD researcher in computing & bioengineering. He taught more than 20 courses such as, data analysis, computer vision, image processing, neural networks, system analysis & design, and object oriented programming. Moreover, he supervised various graduation projects for finger print identification, mobile applications, Tele-radiology, wireless communication, image retrieval, and data mining. He is a reviewer of several international journals and conferences, such as, Science Direct-ELSEVIER, IEEE. He assisted in organizing more than 4 conferences and workshops in Italy. He is a member of Egyptian university staff syndicate, and Egyptian software engineering association. In 2013 he became member of INdAM: Italian National group of advanced mathematics, and GNCS: national group of scientific computing, He won the following grants by competition 1) PhD fellowship from Italy in 2012, 2) Won International Association for Pattern Recognition (IAPR) grant in 2014, 3) He won 2 grants from GNCS competitions in 2013, 2014. He is a Team Leader (9 years practical development experience), Oracle certified professional (OCP), R&D specialist, and CMMI configuration controller. He worked in 4 software development houses based in UK,
Saudi Arabia, Italy, and Egypt since 2005 specialized in developing ERP solutions and web development, especially in medical pattern analysis & radiological applications and other fields such as tourism, Engineering ERP, Oil ERP, and E-learning medical solutions. He led the development team in an international medical imaging software development house, “Sama advanced technologies” for 5 years and developed complete ERP e-learning system for faculty of medicine, radiology department, Cairo university, and developed Radiology Information System “RIS” and Package Archiving and communication system “PACS” for Ain Shams University specialized hospital. He published more than 15 papers & posters in international journals and conferences. He presented his work in various scientific events and attended more than 60 scientific seminars and workshops, and more than 10 PhD specialized courses. He published more than 15 papers & posters in international journals and conferences. He presented his work in various scientific events and attended more than 60 scientific seminars and workshops, and more than 10 PhD specialized courses.

Question 6: How much experience do you have computer security?

A.) limited
B.) minimal
C.) A little, I search about apps for security and etc.
D.) 0
E.) I have a little experience about cryptography and computer security.
F.) I don’t know exactly
G.) I worked around data encryption, secured databases in mobile devices, and content copyright protection. Encrypt and decrypt data on the fly was a good option to protect user data.
H.) More than sufficient to answer these questions.
I.) Little
J.) Zero.
K.) Basic understanding but no professional experience.
L.) None
M.) Hassan Mahmoud has more than 9 years of research and technical experience as: a) Academic Researcher & Lecturer assistant b) Software development Team leader, mainly, his experience was about (but not limited to) the following aspects: Machine Learning, data analysis, Web development, Windows development, ERP solutions, and mobile development with security aspects.
N.) 4 Year
O.) 6 months
P.) Not much
Q.) Been active in computer security research since 2005.
R.) Not much but I understand Hashing, Cookies, and Encryption. Every time I learn something about security, I learn that there are another five things that I don't know.
S.) I used to teach encryption but am now more than a decade out of date. I now rely on commercial software to cover the non-human part of my, and our, security
T.) Too less. No technical deep knowledge in fact.

Question 7: How likely are you to make professional advice to upgrade or change security protocols for a mobile device?
Question 8: How likely are you to make professional suggestions on how to improve an online media site's security or improve data privacy?

CONCLUSION

The full raw data from the survey can be viewed above as well as the Google analytics results. While the sample size was not a random design or calculated for effect size, the results from professionals in related technology and business fields indicate a significant proportion favor stronger research into and implementation of: confidentiality, privacy, accessibility, and security measures. Some of these measures include encryption and closing holes in
protocols. To date no mechanism has been developed with the same level of potential of homomorphic encryption to protect end users from breaches of: confidentiality, privacy, accessibility, and security. While more research is needed, this is precisely the goal of this research paper; to increase awareness and test long-term solutions to ongoing problems. The researcher has uncovered a number of gaps in the privacy and security literature as well areas for software engineering/cryptographic security research. The need to bring down the computational and financial costs for certain algorithmic encryptions still exist, but with present research and new methods being introduced in both cryptography and recommender systems the means to accomplish such goals now exists.

REFERENCES


