ATTRIBUTE BASED PRIVACY CONTROL AND ANONYMIZATION FOR LOCATION PRIVACY

PRITI JAGWANI

AMARNATH & SHASHI KHOSLA SCHOOL OF INFORMATION TECHNOLOGY

INDIAN INSTITUTE OF TECHNOLOGY DELHI

JULY 2018

ATTRIBUTE BASED PRIVACY CONTROL AND ANONYMIZATION FOR LOCATION PRIVACY

by

PRITI JAGWANI

Amarnath and Shashi Khosla School of Information Technology

Submitted

in fulfillment of the requirements of the degree of Doctor of Philosophy

to the

INDIAN INSTITUTE OF TECHNOLOGY DELHI

JULY 2018

Certificate

The thesis entitled **"Attribute Based Privacy Control and Anonymization for Location Privacy"** being submitted by **Ms. Priti Jagwani** to the Indian Institute of Technology Delhi for award of the degree of **Doctor of Philosophy** is a record of original bonafide research work carried out by her. She has worked under my guidance and supervision, and has fulfilled the requirements for the submission of this thesis, which has attained the standard required for a Ph.D. degree of this institute.

The results presented in this thesis have not been submitted elsewhere for the award of any other degree or diploma.

Dr. Saroj Kaushik
Professor
Department of Computer Science and Engineering
Indian Institute of Technology Delhi

New Delhi

Acknowledgements

I would thank the Almighty for giving me the courage and determination to pursue the higher education in the field of Computer Science and Engineering. I would like to tender my heartfelt thankfulness to my supervisor **Prof. Saroj Kaushik** for her priceless supervision, continuous inspiration and enthusiastic support in every stage of the work. Her deep insight into problems, imaginative ideas, technical guidance and sufficient encouragement was a constant source of inspiration and motivation. Her critical comments on my technical writing, eye for detail and persistent encouragement strongly motivated me to strive hard to achieve the high targets set by her in completion of this thesis. Her religious, humanitarian and positive outlook towards the life has been a great source of inspiration even in the non-academic side of life.

I am extremely grateful to the members of my Student Research Committee (SRC) – Prof. Sanjiva Prasad, Prof. Brejesh Lal, Prof. Ponnurangam Kumaraguru, who have been very helpful by offering suggestions and advice. I thank my Ph.D. colleagues, especially, Dr. Sunita Tiwari, Dr. Shivendra Prasad Tiwari, and Kuntal Dey for their help on reviewing my work and offering useful suggestions. I am thankful to my friend Sushil Kumar Pandey for helping me throughout my research. I also extend the sincerest gratitude to the summer intern at IIT Delhi Shubham and most importantly my students who contributed in conducting experiments by registering in the proposed system and providing their valuable feedback. I express my sincere gratitude to my colleagues especially Harish Dhawan sir for providing the unconditional support. This work would not materialize without support from my student, Manoj Kumar who worked tirelessly with me to get this research done.

My parents Mr. Vasudev Goplani and Mrs. Rekha Goplani, my husband Uttam Jagwani and my children Naina and Kabeer as well as other family members showed immense patience and provided me great support during the course of my work. Their patience, sacrifice, inspiration and moral support are of special mention. I dedicated my thesis to them.

Priti Jagwani

Abstract

The increasing trend of embedding positioning capabilities (e.g., GPS) in mobile devices facilitates the widespread use of Location Based Services (LBS). It is certainly more convenient to use GPS of phone to locate the nearest places such as medical shop, restaurant, gas station etc. Extensive usage of these smart devices brought the ubiquitous computing on the fingertips of users. Popularity of these mobile phones equipped with position determination capabilities has revolutionized the living style of people. Also internet has occupied a large place in mass life style. With this tremendous growth of Internet and mobile phones the term “Location based services” has become a buzz word nowadays.

Although LBSs provide enhanced functionalities and convenience of ubiquitous computing, at the same time they also open up new vulnerabilities that can be exploited to cause security and privacy breaches. For LBS applications to succeed, privacy and confidentiality are key issues. For these applications, location of the individual is required. Consequently, they may pose a major privacy threat on its users because of revelation of location. “Privacy protection" for the users of location based services has originated as a relatively new domain for research. Several studies have been proposed for protecting location privacy of a user. Most of them try to prevent disclosure of unnecessary information by techniques that explicitly or implicitly control what information is given to whom and when.

Aim of present research is to focus on attribute based privacy control and improving anonymization techniques for location privacy. In the proposed research, focus is on trajectories and single location anonymization. The overall work is divided into four sub problems as follows.

First problem is focused on trajectory anonymization. In the proposed work, trajectories are anonymized using Earth Mover’s Distance (EMD) as a metric rather than Euclidean distance. A trajectory graph has been constructed which simulates spatial and temporal proximity of trajectories. Further, K- anonymity subsets of the graph have been constructed through graph split using EMD. K-anonymity means that the user is indistinguishable amongst K users. Trajectories in those subsets are best candidate trajectories for anonymization. The proposed approach shows an average improvement of 10% in information loss. Also a significant improvement in total computation time has been observed as compared to Euclidean approach.

Other problems in the proposed research work are handling privacy issues for single location queries and instances.

Second problem focuses on finding value of K for K- anonymity using context parameters. These parameters are taking care of spatial as well as temporal constraints/conditions and are fuzzy in nature. A Fuzzy inference system (FIS) takes these parameters as inputs and calculates value of location disclosure. Value of K will be determined by using this location disclosure. To evaluate the feasibility of the proposed approach, Fuzzy C means clustering (FCM) and genetic algorithm (GA) have been applied. These approaches have been used in order to reduce the size of rule base of FIS for scalability and optimization of the proposed approach. Results of both the approaches have shown that size of rule base can be decreased significantly with a Root Mean Square Error (RMSE) value < 10%.

The third problem is about enhancing the quality of location based services. It includes adoption of real time context obtained from users available in the cloaking area through crowd sourcing model. Based on the blend of nearest distance and real time context parameters, a new rank for the service is generated. It has been statistically proven that the proposed approach outperforms in terms of user feedback as compared to other systems.

The last problem proposes the concept that cloaking area with K-1 similar profile users is more secure as compared with K-1 random users and proved statistically. The proposed approach also quantifies the gain obtained in privacy using KL Divergence as a metric. It has been proved that cloaking area with similar profile users provides a privacy gain of 33% as compared to the cloaking area of random profile users.

For the purpose of design, evaluation and experiments related to above stated problems, Adult census dataset (32560 profiles) by UCI ML repository and GPS trajectory dataset (178 users) collected by Microsoft Research Asia's GeoLife project have been used. To prove the effectiveness of the proposed ideas different prototype systems have been implemented. Evaluations of these prototypes are done using various well known metrics. The results of the proposed approaches are compared with state of the art research and it is found that proposed approached perform better than the existing ones.

सार

बड़े पैमाने पर अवस्थिति आधारित सेवाओं के उपयोग के कारण भूमंडलीय अवस्थिति पद्धति (जी.पी.एस.) क्षमता की सुविधाओं का मोबाइल उपकरणों में विस्तार हुआ है। निश्चित रूप से आसपास के स्थानों,जैसे-औषधालय,रेस्त्रां,गैस स्टेशन आदि को खोजने के लिए फोन की भूमंडलीय अवस्थिति पद्धति का उपयोग अधिक सुविधाजनक है।इन स्मार्ट उपकरणों का उपयोग बहुत अधिक बढ़ने के कारण कम्प्यूटिंग देशव्यापी स्तर पर लोगों की अंगुलियों के नीचे आ गया है। अवस्थिति निर्धारण की सुविधा से पूर्ण मोबाइल फोन की लोकप्रियता ने लोगों की जीवन शैली को क्रांतिकारी तरीके से बदल दिया है।इसके अलावा इंटरनेट बड़े पैमाने पर लोगों के जीवन का हिस्सा बन चुका है।

इंटरनेट और मोबाइल फोन में बड़े पैमाने पर हुई वृद्धि के कारण अवस्थिति आधारित सेवा (एल. बी.एस.) शब्द आजकल चर्चा में है। वैसे तो अवस्थिति आधारित सेवाओं ने कम्प्यूटिंग की सेवाओं में विस्तार किया है, परंतु उन्होंने नए प्रकार के जोखिमों को भी उत्पन्न किया है जो समय के साथ सुरक्षा और निजता के लिए खतरा बन सकते हैं। अवस्थिति आधारित सेवाओं (एल.बी.एस.) से संबंधित एप्स की सफलता के लिए निजता एवं गोपनीयता मुख्य मुद्दे हैं। चूंकि इन एप्स के लिए किसी व्यक्ति के स्थान की जानकारी की आवश्यकता होती है,इस कारण से उपयोगकर्ता के स्थान के बारे में खुलासा हो जाने से उसकी निजता के लिए बड़ा खतरा हो सकता है। अवस्थिति आधारित सेवाओं के उपयोगकर्ताओं के लिए निजता संरक्षण का विषय शोध के लिए अपेक्षाकृत नया क्षेत्र है। किसी उपयोगकर्ता की स्थान गोपनीयता की सुरक्षा के लिए कई अध्ययन प्रस्तावित किए गए हैं। इनमें से अधिकतर उन तकनीकों द्वारा बिना वजह की सूचनाएं देने से रोकने के बारे में है, जिनके पास इसका नियंत्रण है कि सूचना किसे और क्यों दी जाए?

इस शोध का लक्ष्य स्थानों की निजता की संरक्षा के लिए ऐसी तकनीकों को बेहतर बनाए जाने से संबंधित है जो सूचनाओं को गुप्त रख सके।इस बात के उपर ध्यान दिया गया है कि क्षेत्र विशेष तथा स्थान को गुमनाम रखा जा सके। इस शोध-कार्य को अग्रलिखित चार भागों में विभक्त किया गया है।

पहले भाग का संबंध निर्धारित किए गए मार्ग को नामरहित रखे जाने पर केन्द्रित है।इस प्रस्तावित शोध में तय मार्ग को अर्थ मूवर्स डिस्टेंस (इ.एम.डी.) को मानक के रूप में अपनाकर नामरहित किया गया है न कि इसमें यूक्लीडियन डिस्टेंस का उपयोग किया गया है। मार्ग का जो आरेख तैयार किया गया है उसमें मार्ग की स्थानिक तथा समय संबंधी पहचानों को लिया गया है। इसके बाद के नामरहित उपखंड बनाए गए हैं जिसमें इ.एम.डी.के टुकड़ों का उपयोग किया गया है।आगे,ग्राफ का के-गुमनाम सबसेट इ.एम.डी.का उपयोग कर ग्राफ स्प्लिट के माध्यम से बनाया गया है। के - नामरहित का मतलब है कि उपयोगकर्ताओं के मध्य मुख्य उपयोगकर्ता नामरहित बना रहे।उन उपखंडों में जो मार्ग दिए गए हैं वे नामरहित किए जाने के लिए सबसे उपयुक्त हैं।इस प्रस्तावित उपाय के कारण सूचना के खोने की दर में औसत 10%की कमी दिखाई दे रही है।इसके अलावा यूक्लीडियन दृष्टिकोण की तुलना में कुल गणना समय में भी एक महत्वपूर्ण सुधार देखा गया है।

इस शोध का दूसरा मुद्दा किसी स्थान विशेष की निजता से जुड़े प्रश्न और उदाहरण हैं।दूसरा मुद्दा "के- नामरहित" के उपयोग के मानकों में "के" के मूल्य निर्धारण पर प्रकाश डालता है।ये मानक देश तथा काल से जुड़े मानकों का ध्यान रखते हैं और अपनी प्रकृति में धुंधले हैं।एक अस्पष्ट अनुमान प्रणाली (फजी इन्फेरेंस सिस्टम) इन मानकों को अंतर्निवेश के रूप में अपनाता है और स्थान प्रकटीकरण के मूल्य की गणना करता है।'के' का मान इन स्थान प्रकटीकरण के प्रयोग द्वारा निर्धारित किया जाएगा।इस प्रस्तावित उपाय की संभावना के लिए फजी सी का मतलब है कि उसके लिए (एफ. सी.एम.) और जेनेटिक एल्गोरिथ्म(जी.ए.) को आजमाया गया है।एफ.आई.एस. के नियम आधार को कम करने के लिए प्रस्तावित पद्धति में उसकी धारणा को बदल दिया गया है। इससे इसके आधार में रूट मीन स्क्वायर त्रुटि(आर.एम.एस.ई.) मूल्य में 10%से कम का अंतर आया।

तीसरा मुद्दा है स्थान आधारित सेवाओं की गुणवत्ता को बेहतर बनाना। इसमें क्राउड सोर्सिंग माडल पर उपयोगकर्ताओं से वास्तविक समय संदर्भ लिया जाता है जो कि उस इलाके में उपलब्ध रहता है। कम-से-कम दूरी को आधार बनाकर वास्तविक समय में कमी लाई गई है। सांख्यिकीय रुप से यह सिद्ध किया गया है कि प्रस्तावित पद्धति अन्य प्रणालियों की तुलना में उपयोगकर्ता की प्रतिक्रिया के संदर्भ में बेहतर प्रदर्शन करता है।

आखिरी मुद्दा इस अवधारणा का प्रस्ताव करता है कि "के-1" यादृच्छिक उपयोगकर्ताओं के साथ क्लौकिंग क्षेत्र "के-1" यादृच्छिक उपयोगकर्ताओं की तुलना में अधिक सुरक्षित है और यह सांख्यिकीय रुप से सिद्ध हुआ है।के.एल. विचलन के उपयोग के माध्यम से निजता का मामला भी प्रस्तावित पद्धति में खरा उतरता है।औचक प्रोफाइल वाले उपयोगकर्ताओं के मुकाबले एक समान उपयोगकर्ताओं के प्रोफाइल का मानक 33% सुरक्षित है।

उपर्युक्त समस्याओं से संबंधित डिजाइन, मूल्यांकन और प्रयोगों के उद्देश्य के लिए माइक्रोसॉफ्ट रिसर्च एशिया के जियो लाइफ प्रोजेक्ट द्वारा एकत्रित यू.सी.आई.एम.एल. रिपोजिटरी और जी.पी.एस. प्रक्षेपण डेटासेट(178 उपयोगकर्ता) द्वारा वयस्क जनगणना डेटासेट(32560 प्रोफाइल) का उपयोग किया गया है।इन विचारों के प्रभाव को सिद्ध करने के लिए अलग-अलग तरह के प्रोटोटाइप अपनाए गए हैं। जाने-माने मानकों के आधार पर इन प्रोटोटाइप का मूल्यांकन किया गया है। उपलब्ध पद्धतियों की तुलना में इस पद्धति को अधिक बेहतर प्रदर्शन करने वाला पाया गया है।

Table of Contents

List of Figures

List of Tables

Acronyms

AGPS	Assisted Global Positioning System
ANNC	Adaptive Nearest Neighborhood Cloaking
ASR	Anonymizing Spatial Region
CA	Cloaking Area
CGI	Cell Global Identity
CR	Cloaking Region
EMD	Earth Mover's Distance
FCM	Fuzzy C Means
FIS	Fuzzy Interface System
GA	Genetic Algorithm
GIS	Geographic Information System
GPS	Global Positioning System
GSM	Global System for Mobile Communication
GSM	Global System for Mobile Communication
IETF	Internet Engineering Task Force
kNN	K Nearest Neighbour
LBRS	Location Based Recommender System
LBS	Location Based Services
LSP	Location Service Provider
MBB	Minimum Bounding Box
MBR	Minimum Bounding Rectangle
MNTG	Minnesota Web Based Traffic Generator
NNC	Nearest Neighbour Cloaking

P3P	Privacy Preference Project
PDRM	Personal Digital Rights Management
PDS	Plausible Deniable Search
PIR	Private Information Retrieval
POI	Point of Interest
QoS	Quality of Services
RFID	Radio Frequency Identification
RMSE	Root Mean Square Error
TG	Trajectory Graph
TTP	Trusted Third Party
ZKP	Zero Knowledge Proof