Trusted Fog Based Mashup Service for Multimedia IoT based Smart Environmental Monitoring

Ahmed M. Elmisery[1,*], and Mirela Sertovic[2]

Abstract

Data mashup is a web technology that combines information from multiple sources into a single web application. Mashup applications create a new horizon for new services, like environmental monitoring. Environmental monitoring is a serious tool for the state and private organizations, which are located in regions with environmental hazards and seek to gain insights to detect hazards and locate them clearly. These organizations utilize a data mashup to merge datasets from different Internet of multimedia things (IoMT) context-based services in order to leverage its data analytics performance and the accuracy of the predictions. However, mashup different datasets from multiple sources is a privacy hazard as it might reveal citizens specific behaviors in different regions. The ability to preserve privacy in mashuped datasets and at the same time provide accurate insights becomes a key success for the spread of mashup services. In this paper, we present our efforts to build a fog-based middleware for private data mashup (FMPM) to serve a centralized environmental monitoring service. The proposed middleware is equipped with concealment mechanisms to preserve the privacy of the merged datasets from multiple IoMT networks involved in the mashup application. Also, these mechanisms preserve the aggregates in the dataset to maximize the usability of information to attain accurate analytical results. We also provide a scenario for IoMT-enabled data mashup service and experimentation results.

Key Words: IoMT Networks, Environmental Monitoring, Data mashup, Multimedia data.

I. INTRODUCTION

Environmental hazards of natural origin involve large extensions of land such as earthquakes, tsunamis, volcano eruptions, landslides and forest fires are common in countries like Chile and produce emergency scenarios where roads are often saturated or damaged and power supplies are down, disrupting connectivity. These hazards can easily affect a large number of people and isolate them from their surrounding environment. While information and storage capabilities are becoming virtually limitless, in such situations, accessing the right information at the right time by the right organization is a crucial requirement to take proper decisions and to publish highly relevant information to the affected communities and helpers in charge of handling the emergency situations [1]. Decision makers usually require access to highly accurate information servers and data application to estimate the number of affected citizens in a certain region and the best available ways to support them.

Environmental monitoring is one of the areas, which attracts public concern. The advance of cloud computing and Internet of things reshaped the manner in which the sensed information is being managed and accessed. The advances in sensor technologies have accelerated the emergence of environmental sensing service. These new services grasp the significance of new techniques in order to understand the complexities and relations in the collected sensed information. Particularly, it utilizes portable sensing devices to extend the sensing range, and cloud-computing environments to analyse the big amount of data collected by various Internet of multimedia things (IoMT) networks in a productive form. Various kinds of sensors are being deployed in the environment as the physical foundation for most of the environmental sensing services. It is highly desirable to link the sensed data with external data collected from different services in order to increase the accuracy of the predictions [2] In regions with environmental hazards, a large number of citizens makes intensive observations about these regions using their mobile phone during their daily activities. This massive data is expected to be generated from different sources and published on various Internet of multimedia things (IoMT) context-based

services such as Facebook®, Waze® and Foursquare®. In such situation, it is beneficial to include such data in the decision-making process of environmental monitoring services. In this context, Data Mash-up services appear as a promising tool to accumulates this data and manage in an appropriate way.

Data mashup [3] is a web technology that combines information from multiple sources into a single web application for specific task or request. Mashup technology was first introduced in [4] and since then it creates a new horizon for service providers to integrate their data to deliver highly customizable services to their customers [3]. Data mashup can be used to merge datasets from external IoMT context-based services to leverage the monitoring service from different perspectives like providing more precise predictions and performance, and alleviating cold start problems [5] for new environmental monitoring services. Due to that, Providers of the next generation environmental monitoring services keen to gain accurate data mash-up services for their systems. However, privacy is an essential concern for the application of mashup in IoMT-enabled environmental monitoring, as the generated insights obviously require the integration of different behavioural and neighbouring environment data of citizens and from multiple IoMT context-based services. This might reveal private citizens' behaviours that were not available before the data mashup. A serious privacy breach can occur if the same citizen is registered on multiple sites, so adversaries can try to deanonymize the citizen's identity by correlating the information contained in the mashuped data with other information obtained from external public databases. These breaches prevent IoMT context based services to reveal raw behavioural data of the citizen to each other or to the mashup service. Moreover, divulgence citizens' data represent infringement against personal privacy laws that might be applied in some countries where these sites operate. As a result, if the citizens know their raw data are revealed to other parties, they will absolutely distrust this site. According to surveys results in [6, 7] the users might leave a service provider because of privacy concerns.

In this work, we proposed Fog-based middleware for private data mashup (FMPM) that bear in mind privacy issues related to mashup multiple datasets from IoMT context-based services for environmental monitoring purposes. We focus on stages related to datasets collection and processing and omit all aspects related to environmental monitoring, mainly because these stages are critical with regard to privacy as they involve different entities. We present two concealment algorithms to protect citizens' privacy and preserve the aggregates in the mashuped datasets in order to maximize usability and attain accurate insights. Using these algorithms, each party involved in the mashup is given a complete control on the privacy of its dataset. In the rest of this paper, we will generically refer to behavioural and neighbouring environment data as Items. Section II describes some related work. In section III we introduce IoMT-enabled data mashup network scenario landing our FMPM. In section IV introduces the proposed concealment algorithms used in our FMPM. In section V describes some experiments and results based on concealment algorithms for IoT context-based services. Finally, Section VI includes conclusions and future work.

II. RELATED WORK

The majority of the literature addresses the problem of privacy on third-party services [8-13]; Due to it is a potential source of leakage of personally identifiable Information. However, a few works have studied the privacy for mashup services. The work in [3] discussed a private data mashup system, where the authors formalize the problem as achieving a k-anonymity on the integrated data without revealing detailed information about this process or disclosing data from one party to another. In [14] it is proposed a theoretical framework to preserve the privacy of customers and the commercial interests of merchants. Their system is a hybrid recommender that uses secure two-party protocols with public key infrastructure to achieve the desired goals. In [15, 16] it is suggested another method for privacy preserving on centralized services by adding uncertainty to the data, using a randomized perturbation technique while attempting to make sure that necessary statistical aggregates don't get disturbed much. Hence, the server has no knowledge about true values of individual data for each user. They demonstrate that this method does not decrease essentially the obtained accuracy of the results. But recent research work [17, 18] pointed out that these techniques don't provide levels of privacy as it was previously thought. In [18] it is Pointed out that arbitrary randomization is not safe because it is easy to breach the privacy protection it offers. They proposed a random matrix based spectral filtering techniques to recover the original data from perturbed data. Their experiments revealed that in many cases random perturbation techniques preserve very little privacy.

III. DATA MASHUP IN IOT-ENABLED ENVIRONMENTAL MONITORING SCENARIO

We consider the scenario where the IoMT-enabled data mashup service (IoMT-enabled DMS) integrates datasets from multiple IoMT context-based services for the IoT-enabled environmental monitoring; figures (1) and (2) illustrates the scenario used in this work. We assume all the involved parties follow the semi-honest model, which is a realistic assumption because each party needs to accomplish some business goals and increases its revenues. Also, we assume all parties involved in the data mashup have similar items set (activities' catalogue) but the users' sets are not identical. Each IoT context-based service has its own ETL (Extract, Transform, Load) service that has the ability to learn behavioural and neighbouring environment data of citizens.

The data mashup process based on FMPM can be summarized as follows; the environmental cognition service sends a query to the IoMT-enabled DMS to gather information related to behavioural and neighbouring environment data of citizens in a specific region to leverage its predictions and performance. The coordinator Agent in IoMT-enabled DMS lookup in its providers' cache to determine the providers could satisfy that query then it transforms query of the environmental cognition service into appropriate sub-queries languages suitable for each provider's database. The manager agent unit sends each sub-query to the candidate providers to incite them about the data mashup process. The provider who decides to participate in that process, forwards the sub-query to its manager agent to refine it considering its privacy preferences. This step allows the manager agent to audit all issued sub-queries and prevent ones that can extract sensitive information. The resulting dataset sent to the local concealment agent (LOA) to hide real participants' data using the appropriate concealment algorithm. Then, every synchronization agents at each provider along with the coordinator agent engage in distributed joint process to identify frequent and partially frequent items in each dataset, then send the joined results to the coordinator. The coordinator agent builds a virtualized schema for the datasets and submits it to each provider involved in the mashup process. Based on this virtualized schema, the providers incite their global concealment agent (GOA) to start the appropriate concealment algorithm on the locally concealed datasets. Finally, the providers submit all the resulting datasets to IoMT-enabled DMS that in turn unites these results and delivers them to the environmental cognition service. The environmental cognition service uses these datasets to accomplish the required data analytics goals. We use anonymous pseudonyms identities to alleviate providers' ide

ntity problems, as the database providers does not want to reveal their ownership of the data to competing providers moreover the IoMT-enabled DMS will keen to hide the identities of providers as a business asset.

IV. PROPOSED CONCEALMENT ALGORITHMS

In the next sub-sections, we introduce our proposed algorithms used to preserve the privacy of the resulting datasets with minimum loss of accuracy.
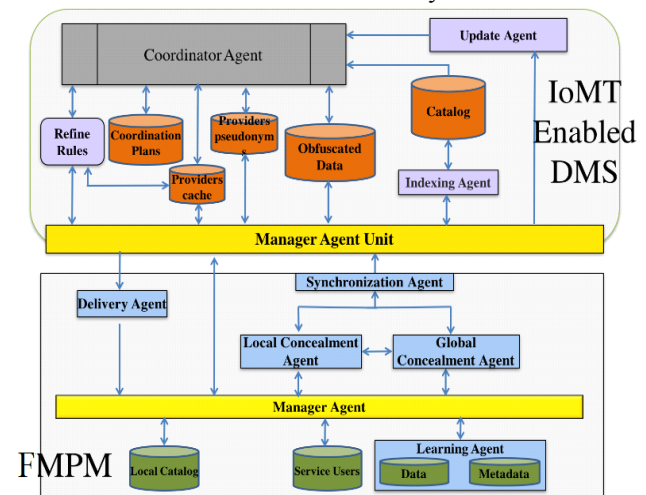

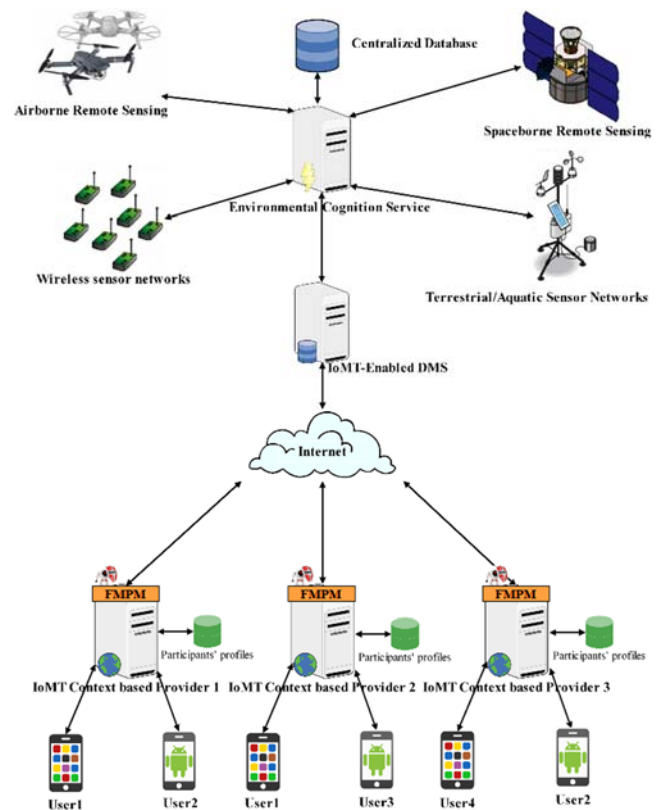Fig.1. The Building Blocks of IoMT-enabled DMS and FMPM.

Fig.2. IoMT-enabled data mashup service with Third Party IoMT context-based services.

A closer look at the attack model proposed in [19] reveals that, if a set of behavioural and neighbouring environment data of certain citizen is fully distinguishable from the data of other citizens in the dataset with respect to some features. This citizen can be identified if an attacker correlates the revealed data with data from other publicly accessible databases. Therefore, it is highly desirable that the dataset has at least a minimum number of items should have a similar feature vector to every real item released by each participant. A real item in the released dataset can be described by a certain number of features in a feature vector, such as place of activity, type of activity, duration, time, date and so on. Both implicit and explicit ways can be used to extract this information and to construct these feature vectors and to maintain them. Additionally, the data sparsity problem associated with ETL services can be used to formulate some attacks as also shown in [19]. Before starting, we introduce a couple of relevant definitions.

Definition 1 (Dissimilarity measure): this metric measures the amount of divergence between two items with respect to their feature vector. We use the notation $\mathcal{D}_m(I_u, I_n)$ to denote the dissimilarity measure between items I_u and I_n based on the feature vector of each item. $\mathcal{D}_m(I_u, I_n) < \delta \Rightarrow I_u \sim I_n$ [I_u is similar to I_n], δ is a user defined threshold value.

Definition 2 (Affinity group): the set of items that are similar to item I_u with respect to pth attribute A_p of the feature vector and it is called affinity group of I_u and denoted by $C_{A_p}(I_u)$.

$$C_{A_p}(I_u) = \{I_n \in D_n | (I_u \sim I_n) \wedge (A = A_p)\}$$
$$= \{I_n \in D_n | \mathcal{D}_m(I_u, I_n) < \delta\}$$

Definition 3 (K-Similar item group): Let D_ϖ be the real items dataset and $\widetilde{D_\varpi}$ its locally concealed version. We say $\widetilde{D_\varpi}$ satisfies the property of k-Similar item group (where K is defined value) provided for every item $I_u \in D_\varpi$. There is at least k-1 other distinct fake items $I_{n_1}, \dots I_{n_{(k-1)}} \in D_n$ forming affinity group such that:

$$FV\left(I_{n_i}\right) \sim FV\left(I_u\right), \qquad \forall \ 1 \leq i \leq k - 1$$

A. Local Concealment using Clustering Based Obfuscation (CBO) Algorithm

Our motivation to propose CBO is the limitation of the current anonymity models. The current anonymity models proposed in the literature failed to provide an overall anonymity as they don't consider matching items based on their feature' vectors. CBO uses the feature vectors of the current real items to select fake items highly similar to real items to create homogeneous concealed dataset. Using fake transactions to maintain privacy was presented in [3], [20, 21], the authors considered adding fake transactions to anonymise the original data transactions. This approach has several advantages over other schemes including that any off-the-shelf data analytics algorithms can be used for analysing the concealed data and the ability to provide a high theoretical privacy guarantee. The locally concealed dataset obtained using CBO should be indistinguishable from the original dataset in order to preserve privacy. The core idea for CBO is to split the dataset into two subsets, the first subset is modified to satisfy K-Similar item group definition, and the other subset is concealed by substituting real items with fake items based probabilistic approach. CBO creates a concealed dataset D_P as following:

1. The sensitive items are suppressed from the dataset based on provider preferences thereafter we will have the suppressed dataset D as the real dataset.
2. Selecting a ϖ percent of highest frequent items in dataset D to form a new subset D_ϖ. This step aims to reduce the substituted fake items inside the concealed dataset D_P. Moreover, it maintains data quality by preserving the aggregates of highly frequent preferences.
3. CBO builds affinity groups for each real item $\forall \ I_u \in D_\varpi$ through adding fake items to form *K-Similar items group*. We implemented this task as a text categorization problem based on the feature vectors of real items. We also implemented a bag-of-words naive Bayesian text classifier [22] that extended to handle a vector of bags of words. The task continues until all items in D_ϖ are belonging to different affinity groups, then we get a new dataset $\widetilde{D_\varpi}$.
4. For each $I_u \in D_u = D - D_\varpi$, CBO selects a real item $\{I_u\}$ from real item set D_u with probability α or selects a fake item $\{I_n\}$ from the candidate fake item set D_n with probability $1 - \alpha$. The selected item I_P is added as a record to the concealed dataset D_P. This method achieves the desired privacy guarantee because the type of selected item and α are unknowns to external parties. The process continues until all real items in D_u are selected.
5. Finally, the concealed dataset D_P is merged with the subset $\widetilde{D_\varpi}$, which obtained from step 3.

Analysis of Local Concealment using CBO

In terms of performance, CBO requires supplementary storage costs and computations costs. The supplementary storage costs can be reduced by clustering items in the resulting dataset into C clusters and use the feature' vectors of top N items with high rates in each cluster for CBO algorithm. Thus supplementary storage costs will be in

order of $O(CN)$. The computation costs for CBO are divided between computational complexities required to create affinity groups and adding fake items. Obviously, the computation overhead in creating affinity groups dominates, and it can be reduced by selecting lower values for ϖ.

B. Global Concealment using Random Ratings Generation (RRG) Algorithm

After executing CBO, the synchronization agents build a virtualized schema with the aid of the coordinator agent at IoMT-enabled DMS then the global concealment agent starts executing the RRG algorithm. The coordinator agent will not be able to know the real items in the merged datasets as they already concealed locally using CBO algorithm. The main aim for the RRG is to alleviate data sparsity problem by filling the empty cells in such a way to improve the accuracy of the predictions at the environmental monitoring side and increase the attained privacy for providers. The RRG algorithm consists of following steps:

1. The global concealment agent finds the number of majority frequent items I_r and partially frequent items by all users $I - I_r$, where I denotes the total number of items in merged datasets.
2. The global concealment agent randomly selects an integer ρ between 0 and 100, and then chooses a uniform random number ξ over the range $[0, \rho]$.
3. The global concealment agent decides ξ percent of the partially frequent items in merged datasets and uses the KNN to predicate the values of the empty cells for that percentage.
4. The remaining empty cells are filled by random values chosen using a distribution reflecting the frequent items in the merged datasets.

Analysis of Global perturbation using RRG

The privacy of the merged datasets is maintained because all the processing is done on the datasets that previously processed using CBO. The global concealment agent improves the overall privacy and accuracy by increasing the density of the merged datasets due to the filled cells. With increasing ρ values, the RRG reduces the randomness in the frequencies. That might increase the accuracy of the predictions while decreases the privacy level. So, RRG should select ρ in a way to achieve the required balance between privacy and accuracy.

V. EXPERIMENTAL RESULTS

The proposed algorithms are implemented in C++, we used message-passing interface (MPI) for a distributed memory implementation of RRG algorithm to mimic a distributed network of nodes. In order to evaluate the effect of our proposed algorithms on mashuped datasets used in problem solving. A dataset pulled from the SportyPal® network that was linked to another dataset containing behavioural and neighbouring environment data of 8000 students in the University of Zagreb in Croatia in the period of 2006 to 2008. For the purpose of this work, we intended to measure two aspects in this dataset, which are privacy breach levels and accuracy of results. We divide the dataset into a training set and testing set. The training set is concealed then used as a database for the monitoring service. To evaluate the accuracy of the generated predictions, we used the mean average error (MAE) metric proposed in [23]. To measure the privacy breach levels, we used mutual information as a measure for the notion of privacy breach of D_u through D_P.
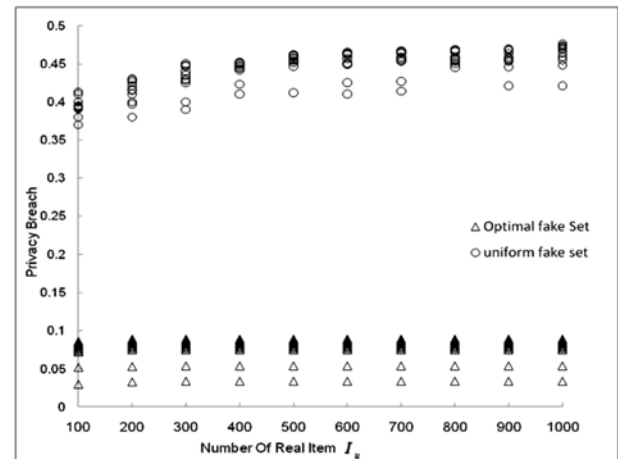

Fig. 3. Privacy breach for optimal and uniform fake sets.

In the first experiment, we want to measure the relation between the quantity of real items in the concealed dataset and privacy breach, we select α in a range from 1.0 to 5.5, and we increased the number of real items from 100 to 1000. We select fake items set using uniform distribution as a baseline. As shown in figure (3), our generated fake set reduces the privacy breach and performs much better than uniform fake set. As the number of real items increase the uniform fake set get worse as more information is leaked while our optimal fake set does not affect with that attitude.

In the second experiment, we measured the relation between the quantity of fake items in the subset D_{ϖ} and the accuracy of the classification results. We select a set of real items from our dataset, then we split it into two subsets D_{ϖ} and D_u. We concealed subset D_u with fixed value for α to obtain the subset D_p. We append the subset D_{ϖ} with either items from optimal fake set or

uniform fake set. Thereafter, we gradually increased the percentage of real items in D_ϖ that are selected from our dataset from 0.1 to 0.9. Figure (4) shows MAE values as a function of the concealment rate for the whole concealed dataset D_p. The IoMT context-based service can select a concealment rate based on its privacy preferences. Hence, with a higher value for the concealment rate, higher accurate predictions can be attained by the monitoring service. Adding items from the optimal fake set have a minor impact on MAE of the results without having to select a higher value for the concealment rate.
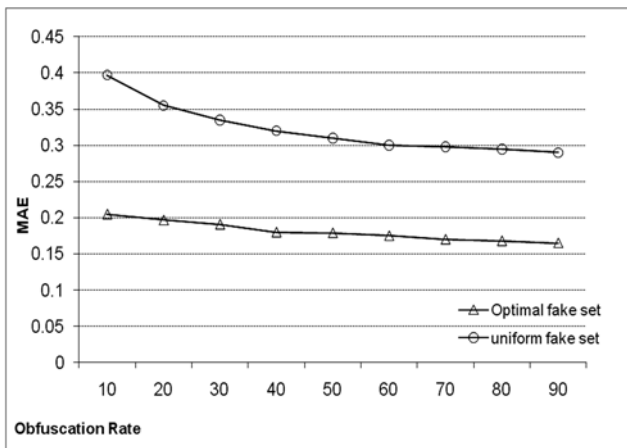

Fig. 4. MAE of the generated predictions vs. concealment rate.

IV. CONCLUSION

In this work, we presented our ongoing work on building a fog-based middleware for private data mashup (FMPM) to serve centralized IoT-enabled environmental monitoring service. We gave a brief overview over the mashup process and two concealment mechanisms. The experiments show our approach reduces privacy breaches and attains accurate results. We realized many challenges in building an IoMT-enabled data mashup service. As a result, we focused on environmental monitoring service scenario. This allows us to move forward in building an integrated system while studying issues such as a dynamic data release at a later stage and deferring certain issues such as virtualized schema and auditing to future research agenda.

REFERENCES

[1] T. Catarci, M. de Leoni, A. Marrella, M. Mecella, B. Salvatore, G. Vetere, et al., "Pervasive software environments for supporting disaster responses," *IEEE Internet Computing*, vol. 12, pp. 26-37, 2008.

[2] J. San-Miguel-Ayanz, E. Schulte, G. Schmuck, A. Camia, P. Strobl, G. Liberta, et al., "Comprehensive monitoring of wildfires in Europe: the European forest fire information system (EFFIS)," *Approaches to Managing Disaster – Assessing Hazards, Emergencies and Disaster Impacts*, 2012.

[3] T. Trojer, B. C. M. Fung, and P. C. K. Hung, "Service-Oriented Architecture for Privacy-Preserving Data Mashup," *presented at the Proceedings of the 2009 IEEE International Conference on Web Services*, 2009.

[4] R. D. Hof., Mix, Match, And Mutate. BusinessWeek. Available: http://www.businessweek.com/print/magazine/content/05_30/b3944108_mz063.htm?chan=gl, 2005

[5] M. d. Gemmis, L. Iaquinta, P. Lops, C. Musto, F. Narducci, and G. Semeraro, "Preference Learning in Recommender Systems," *presented at the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD)*, Slovenia, 2009.

[6] L. F. Cranor, "'I didn't buy it for myself' privacy and ecommerce personalization," presented at the Proceedings of the 2003 ACM workshop on Privacy in the electronic society, Washington, DC, 2003.

[7] C. Dialogue, "Cyber Dialogue Survey Data Reveals Lost Revenue for Retailers Due to Widespread Consumer Privacy Concerns," *in Cyber Dialogue*, 2001.

[8] A. M. Elmisery and D. Botvich, "Enhanced middleware for collaborative privacy in IPTV recommender services," *Journal of Convergence*, vol. 2, p. 10, 2011.

[9] A. M. Elmisery and D. Botvich, "Agent based middleware for private data mashup in IPTV recommender services," *in 2011 IEEE 16th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*, pp. 107-111, 2011.

[10] A. M. Elmisery and D. Botvich, "Multi-agent based middleware for protecting privacy in IPTV content recommender services," *Multimedia Tools and Applications*, vol. 64, pp. 249-275, 2012.

[11] A. M. Elmisery, "Private personalized social

recommendations in an IPTV system," *New Review of Hypermedia and Multimedia*, vol. 20, pp. 145-167, 2014.

[12] A. M. Elmisery, S. Rho, and D. Botvich, "A Fog Based Middleware for Automated Compliance With OECD Privacy Principles in Internet of Healthcare Things," *IEEE Access*, vol. 4, pp. 8418-8441, 2016.

[13] A. M. Elmisery, S. Rho, and D. Botvich, "A distributed collaborative platform for personal health profiles in patient-driven health social network," *Int. J. Distrib. Sen. Netw.*, vol. 2015, pp. 11-11, 2015.

[14] A. Esma, "Experimental Demonstration of a Hybrid Privacy-Preserving Recommender System," pp. 161-170, 2008.

[15] H. Polat and W. Du, "Privacy-Preserving Collaborative Filtering Using Randomized Perturbation Techniques," *presented at the Proceedings of the Third IEEE International Conference on Data Mining*, 2003.

[16] H. Polat and W. Du, "SVD-based collaborative filtering with privacy," *presented at the Proceedings of the 2005 ACM symposium on Applied computing*, Santa Fe, New Mexico, 2005.

[17] Z. Huang, W. Du, and B. Chen, "Deriving private information from randomized data," *presented at the Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, Baltimore, Maryland, 2005.

[18] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar, "On the Privacy Preserving Properties of Random Data Perturbation Techniques," *presented at the Proceedings of the Third IEEE International Conference on Data Mining*, 2003.

[19] A. Narayanan and V. Shmatikov, "Robust De-anonymization of Large Sparse Datasets," *presented at the Proceedings of the 2008 IEEE Symposium on Security and Privacy*, 2008.

[20] J.-L. Lin and J. Y.-C. Liu, "Privacy preserving itemset mining through fake transactions," *presented at the Proceedings of the 2007 ACM symposium on Applied computing*, Seoul, Korea, 2007.

[21] J.-L. Lin and Y.-W. Cheng, "Privacy preserving itemset mining through noisy items," *Expert Systems with Applications*, vol. 36, pp. 5711-5717, 2009.

[22] D. D. Lewis, "Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval," *presented at the Proceedings of the 10th European Conference on Machine Learning*, 1998.

[23] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J.

T. Riedl, "Evaluating collaborative filtering recommender systems," *ACM Trans. Inf. Syst.*, vol. 22, pp. 5-53, 2004.

Authors

Ahmed M. Elmisery: is currently working as an assistant professor at the electronic engineering department of Federico Santa María Technical University (Chile). He is also a research fellow at Internet of Things and People Research Center, Malmö University (Sweden), and Adjunct Assistant Professor at Computer Science Department, Technical College (Egypt). He visited the Center for Advanced Technology in Telecommunications and Secure Systems at Monash University, Australia from May 2014 to June 2015. Before that, he was working as a Researcher in computer security at Telecommunications Software and Systems Group, Department of Computing, Mathematics and Physics, Waterford Institute of Technology, (Ireland). He received his B.S. degree in computer science from the Faculty of Computer Science, Mansoura University, Egypt (2001), M.S. degree in computer science from the Arab Academy for Science & Technology, Egypt (2007), and Ph.D. degree in computer science from Waterford Institute of Technology, Ireland (2014). He has published over 40 research papers in national and international conferences. His research interests include security, cryptography, penetration testing, malware analysis and machine learning. He is conducting research on privacy and security for future telecommunication services. His work has been grounded to develop privacy-enhanced algorithms for Internet of things.

Mirela Sertovic: received the B.S. degree in Education from University of Tuzla (Bosnia and Herzegovina) in 2004. She is currently a Ph.D. candidate at University in Zagreb (Croatia). Her research interests include Social humanistic informatics, Social Networks, Blended Learning, Technology integration in teacher education, Application of technology in language learning and teaching, and Technology-assisted language learning.