An Analysis of Privacy Preserving Data Aggregation Protocols for WSNs

Irfana Memon

An Analysis of Privacy preserving Data Aggregation protocols for WSNs

Irfana MEMON

ERISCS Research Group, Aix-Marseille Université, France

Irfana.MEMON@univ-amu.fr

Abstract. Wireless sensor network (WSN) technology has the potential to change the way we live, work, protect and do business, with applications in entertainment, travel, industry, telemedicine, disaster and emergency management. Data aggregation is key technique for power-efficient information acquisition in WSNs. However, data privacy during data aggregation is an important issue when the WSN is deployed in sensitive data applications, such as telemedicine. If the issues associated with data privacy are not seriously considered, the technology would not be trustingly used for many valuable applications. The existing privacy preserving data aggregation protocols provide a method to sustain privacy of collected sensor's data from external and internal adversaries during data aggregation in WSNs. The basic aim of the paper is to investigate the critical aspects of the existing privacy preserving data aggregation protocols for WSNs and highlight their major limitations. We claim that in future such limitations can be corrected. Our ongoing work is to propose an alternative solution to overcome such limitations, but this will be presented in a future paper.

1 Introduction

A WSN can be generally described as a network of nodes that cooperatively monitor environmental or physical conditions such as temperature, vibration, pressure,location or motion from unattended locations. Recently, WSNs have found their way into a wide variety of applications like disaster relief, emergency rescue operation, military surveillance, habitat monitoring, remote health care applications, environmental monitoring [1]. WSNs are usually deployed in hostile environments, where the deployment of sensors, their maintenance, recharging or replacing their batteries are not always feasible. Sensor nodes are resource constrained (i.e., have limited resources in terms of power, memory and transmission range). Previous studies such as [2] have shown that data communication between sensor nodes requires a large portion of the total energy consumption of the WSNs, thus finding an optimal approach to minimize the messages transmitted in the network is particularly important.

Data aggregation is a key feature for power-efficient information acquisition in resource-constrained WSNs. It is the process of combining data from different sensor nodes by using some functions such as suppression or filtering (eliminating duplicates), min, max and average; to minimize data transmission and removing redundan-

cy. Several studies address data aggregation schemes in WSNs to minimize data transmission [[3]-[6]].

If privacy of the collected sensor's information is not preserved, it is not safe to deploy the WSNs for sensitive data applications, such as telemedicine and military applications. Data privacy in WSNs has to be guaranteed end-to-end (that is, each node should know its own data, but has no knowledge about neighbors data in the network and only base station could read final aggregation result). Very little work has been proposed to address Privacy preserving Data Aggregation (PDA) in WSNs. The purpose of this paper is to review the ewisting PDA protocols and to highlight their limitations. Up to now, to the best of our knowledge, there have been two surveys of this literature so far, one by Na Li et al., [7] and another by Rabindra Bista et al., [8]. However, they do not provide analysis of the protocols in the terms of communication and computation cost, security, and fault tolerance. The rest of the paper is organized as follows. In section 2, we give a critical analysis of existing PDA protocols. Section 3 presents basic requirements for designing PDA protocol. Section 4 presents the conclusion and future work.

2 An analysis of existing Privacy preserving Data Aggregation (PDA) protocols

In this section, we review the existing PDA protocols for WSNs and classify them in three classes according to schemes adopted to satisfy data privacy:
privacy homomorphism, perturbation, and shuffling.

2.1 Privacy homomorphism

Privacy homomorphism scheme allows arithmetic operations to be performed on encrypted data without need of decryption. The protocols proposed in [9], [11], and [12] are based on this scheme.

2.1.1 Concealed data aggregation (CDA) [9]: In this scheme, sensor nodes share a common symmetric key with the base station that is kept hidden from intermediate aggregators. Aggregators carry out aggregation functions that are applied to encrypted data without decrypting. This provides the advantage of aggregators not to carry out costly decryption and re-encryption operations. CDA employs the privacy homomorphic encryption function proposed by Domingo-Ferrer [10]. In CDA, each sensor node splits its sensed data into 'd' parts ($d \geq 2$), encrypts each part and sends encrypted data to the aggregator node. The aggregator node aggregates received encrypted data without decryption and sends it to the base station. Base station decrypts the encrypted aggregated data to derive the original data.

Complexity analysis:

Communication and Computation Cost: In CDA, each node divides its data into d parts, encrypt each part and sends it to aggregator node. As a result, CDA suffers from excessive computational complexity and communication overhead.

Security: CDA ensures data privacy against aggregators but does not guarantee the privacy of individually sensed data against other nodes because all sensor nodes share the same encryption key with the base station.

2.1.2 Efficient Aggregation of encrypted data [11]: This scheme is essentially a stream cipher and its homomorphic property relies on the synchronization among the key-stream generators, i.e., all sensors in the field must share the same key-stream generator. In this scheme, aggregator node can aggregate received data from its children without decryption. The main idea of this approach is to use modular addition (+) instead of xor (Exclusive-OR) operation that is found in the stream ciphers. To minimize trust assumptions, this scheme assumes that each sensor n_i share a distinct long-term key k_i with the BS. This key is originally derived, using a pseudo-random function (PRF), from the master secret K, which is only known to the BS. Sensor nodes encrypt their data using key k and then forwards it to aggregator, who aggregates all received encrypted data of its children without decrypting. The base station can decrypt and derive original data.

Complexity analysis:

Communication Cost: In this scheme, BS needs to know the IDs of all nodes. Therefore, each aggregator needs to append the IDs of its children that did not reply to the query to the aggregate, which generates high communication overhead.

Security: Although the proposed scheme provides end-to-end privacy, but it is vulnerable to false data injection attacks. An external attacker can add an arbitrary value to an aggregate cipher text. Moreover, it is difficult to ensure the confidentiality of the commonly shared key-stream generator in this scheme.

2.1.3 Efficient and Provably Secure Aggregation of encrypted data [12]: This is an improved scheme of [11] for privacy-preserving additive aggregation based on homomorphic encryption. Compared to earlier work [11], this paper provides the details of a concrete construction using a pseudo-random function (PRF).

In this scheme, the key is generated by a certain deterministic algorithm (with an unknown seed) such as a pseudo-random function [13]. Two components are used in construction: a pseudo-random function 'f' and a length-matching hash function 'h'. The output of the pseudo-random function can be hashed down by some length-matching hash function. The purpose of 'h' is to shorten a long bit-string, rather than to produce a fingerprint of a message. For instance, 'h' can be implemented by truncating the output of a PRF and taking *l* least significant bits as output. Note that the parameter *l* should still be chosen large enough to ensure reasonably low probability of success for a random guess.

Complexity analysis:

Communication Cost: In this scheme, hash function 'h' is used to shorten a long bit-string. Therefore, communication cost is less than the scheme presented in [11]. But, still in this scheme, base station needs to know the IDs of all nodes. Therefore, each aggregator needs to append the IDs of its children that did not reply to the query in the aggregate, that generates high communication cost.

Computation Cost: To encrypt its data, a node performs one PRF invocation, one length matching hash, and one mod M addition. It also performs one extra addition for aggregation. The authors considered the cost of evaluating 'h' to be negligible in the calculation of overall computation cost for encryption. As a result, the cost of encryption is dominated by a single PRF invocation.
Security: This scheme allows aggregators to aggregate encrypted data of their children without having to decrypt. As a result, even if an aggregator is compromised, it cannot learn the data of its children, resulting in much stronger privacy. This scheme is vulnerable to message loss, the base station will obtain bogus aggregate with a single message loss.

2.2 Perturbation scheme

In Perturbation scheme, sensor nodes use encryption keys and private/public seeds generated by randomization techniques. Cluster based Private Data Aggregation (CPDA) [14] and Contie et al. scheme [16] are perturbation based protocols.

2.2.1 Cluster based Private Data Aggregation (CPDA) [14]: In CPDA, sensors are randomly grouped into clusters using a distributed protocol proposed in [14]. In each cluster, cluster leader is responsible for aggregating data received from the cluster members. To maintain data privacy, all sensors within a cluster share a common (non-private) knowledge of non-zero numbers, referred to as seeds, which are distinct with each other. Sensors in each cluster customize their private data into k-1 polynomial using shared seeds and random numbers (private), where k is the total number of nodes in a cluster. Then, each sensor encrypts its customized value by using a unique shared key between sensors. Sensor S_i keeps one share, and exchange remaining shares with (k-1) nodes in same cluster. Each sensor S_i assembles all the data including its own by using the additive property of polynomials and sends them to their respective cluster leader. Finally, cluster leader computes the aggregate value and forwards the derived sum of the cluster to the base station along the TAG routing tree [15]. Fig.1. illustrates the CPDA scheme step by step among the three nodes, where A is the cluster leader of this cluster and B and C are cluster members.

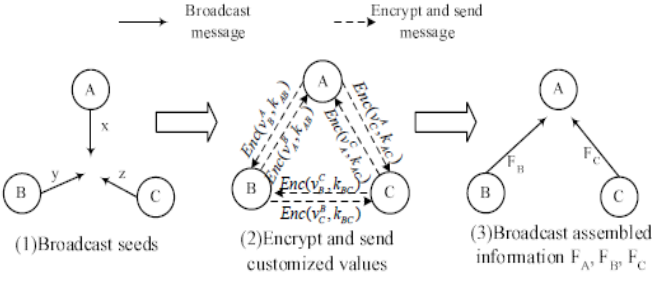


Fig. 1. CPDA scheme

Complexity analysis:

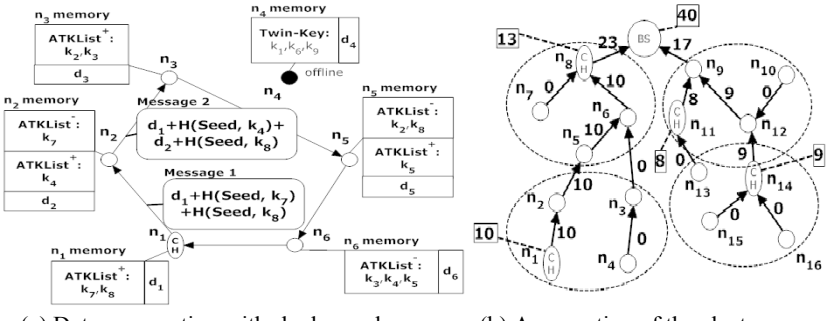
Communication Cost: Cluster formation step in CPDA has a complexity of $O(p_cN)$, where p_c is the probability of a node indepently becoming a cluster leader and N is the number of nodes in the network. In addition, exchanging of seeds within a cluster takes O(k) messages and exchanging their encrypted customized data within a cluster takes O(k) message. Where k is the number of nodes in cluster. This is done in each cluster. Hence, the CPDA approach suffers from the high communication overhead.
Node Failure: When cluster leader failure occurs, data aggregation within the cluster fails. When a cluster member sends data partially to a few nodes and then fails, it results a loss in accuracy.

2.2.2 Privacy-preserving Robust Data Aggregation [16] : This scheme has following key elements: First, establishment of twin-keys for different pairs of sensors in the network, which is an anonymous process that prevents each node in a pair from deriving the identity of the other node (twin-node). Second, for each aggregation phase, sensor node uses an anonymous liveness announcement protocol to declare the liveness of each twin-key, so that each node becomes aware of whether a twin-key it possesses will be used by the anonymous twin-node. Finally, during the aggregation phase, each node encrypts its own value by adding shadow values computed from the alive twin-keys it holds. As a result, the contribution of the shadow values for each twin-key will cancel out each other and the correct aggregated result is finally obtained. This scheme consists of three steps: local cluster formation, twin-key establishment and data aggregation. In the local cluster formation step, nodes are grouped into several clusters using a cluster algorithm proposed in [17]. Each cluster forms a different logical Hamiltonian circuit, and each pair of neighboring nodes in the circuit shares a pair-wise key. In the twin-key establishment step, it is assumed that each node contains a pre-deployed key-ring of K symmetric key, using the set-up procedure of Eschenauer and Gligors protocol [18]: the K keys are randomly chosen from a larger key-pool of size P. Each node n_i anonymously checks which ones of its K keys are also shared with other nodes in the same cluster and establishes a number of twin keys with the other nodes. In particular, a node n_i establishes a twin key with another node (twin-node) in the cluster when n_i is aware there is a node sharing a key with it. Data aggregation step is further divided into two parts: First, each cluster computes the the aggregated value of its nodes, together with a twin-key liveness announcement procedure. During this phase an aggregate is routed twice along the Hamiltonian circuit. Each node adds its own sensed value to the aggregate. At the same time, for each alive twin-key it adds (or removes, in accordance with the liveness announcement) a corresponding shadow value. The cluster head obtains the correct aggregate for the cluster. The liveness announcement guarantees that any shadow value, computed from a twin-key that is added in the aggregation by one node will be removed by another node that shares the same twin-key. Second, by using a tree structure [15], the cluster heads contribute to the aggregate with the cluster. Finally, the base station receives the aggregated results of the cluster heads. Fig.2. show the data aggregation with shadow values and aggregation of the cluster aggregates, respectively.

Complexity analysis:

Communication and Computation Cost : In this scheme, an aggregate is routed twice along the Hamiltonian circuit and each node is required to test each of its pre-distributed keys to find out the required twin-keys shared with other nodes. Each node has to send out the hash values corresponding to its twin-keys that have not yet been declared alive by other nodes. Furthermore, each node has to compute the hash for each of its pre-distributed keys and to encrypt each message it sends out. For each agreed twin-key, k, each node has to compute two hash values. One hash is computed for the verification of the liveness announcement of k. The other hash is computed for the k's corresponding shadow value added in the aggregated value. Therefore, the computation of each node in the worst case is 2A hash computations, considering A agreed twin-keys. This scheme suffers from high computation and communication cost.

Security: This scheme preserves data privacy from external and internal adversaries. It is resilient to eavesdropping attack due to the pair-wise encryption between nodes. In this scheme, each node value is protected by one or more shadow values, hence it is resilient to node compromise attacks. The secrecy of the shadow value, in turn, is protected by the secrecy of the twin- keys. To compromise the privacy of non-captured node, n_i, the attacker has to obtain the keys used to generate the shadow values that n_i uses to protect its own privacy.



(a) Data aggregation with shadow values (b) Aggregation of the cluster aggregates

Fig. 2. Conti et al. Scheme

2.3 Shuffling Scheme

In shuffling scheme, each sensor node slices its sensed data randomly into a certain number (say, n) of pieces, and one piece is kept on itself, the remaining n-1 pieces are securely distributed to n-1 neighbor nodes. After the data pieces are received from the neighboring nodes, all the sensors decrypt the data by using their shared keys, sums up all the received data slices, and sends the sum to their parent node. SMART[14] and iPDA[19] are shuffling based protocols.

2.3.1 Slice-Mix-AggRegaTe (SMART) [14]: This scheme guarantees data privacy through data "slicing and assembling". In SMART, base station runs key distribution

scheme [18] to generate and distribute key ring in all nodes. The key ring for sensor S_i is represented by k_i and K is the union set of all key rings. SMART scheme consists three steps: In the first step, each sensor S_i (i = 1,2,...,N), randomly select J neighbor nodes in h hops (J is a design parameter). Sensor S_i then slices its private data into J pieces. One piece is kept at sensor S_i itself, the remaining J-1 pieces are encrypted and sent to nodes in the randomly selected set of S_i. In the second step, on receiving data pieces from the neighbors, all sensors decrypt the data by using their shared keys and sum all the received data slices. In the third step, when a node receives all data slices, it forwards a message of the sum to its parent, which in turn forwards the message to the base station using tree-based routing protocol [15]. Eventually the aggregation reaches the base station.

Complexity analysis:

Communication Cost: Each sensor node randomly selects a set of J nodes within h hops. Furthermore, Data pieces are sent out appropriately by each sensor node. Hence, this scheme suffers from the high communication overhead. Furthermore, randomly selected set of J nodes are not necessary in immediate communication range. Hence lot of power consumption is expended for communication.

Node Failure: Sensor node's failure results to loss in accuracy of the final result.

2.3.2 Integrity-Protecting Private Data Aggregation (iPDA) [19]: In this scheme, data privacy is achieved through data "slicing and assembling" scheme discussed in [14]; and data integrity is achieved through redundancy by constructing disjoint aggregation trees. Fig.3. shows two disjoint aggregation trees which are separately rooted at base station. Each sensor node sends its reading to both aggregation trees. The disjoint aggregation trees perform data aggregation individually. Hence, base station can detect data pollution attacks by comparing aggregation results along the disjoint aggregation trees. If the aggregation results received along both trees are same, the base station will accept the result. Otherwise, the base station knows that there exist either data pollution attacks or node failures, or both; hence base station reject it. In iPDA, each sensor node hides its individual data by slicing the data and sending encrypted data slices to different neighboring nodes, then the aggregators aggregate the received data and sends aggregated results to the base station. This scheme suffers same complexity issues as described earlier in the complexity analysis of SMART protocol. In addition this scheme has a high communication overhead due to the slicing technique and each sensor node has to send its reading to both aggregation trees. In this scheme, base station can detect data pollution attack, but does not propose any solution to protect from the attack.

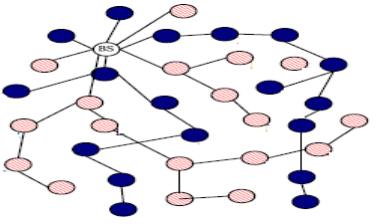


Fig. 3. iPDA scheme (Two disjoint aggregation trees)

3 Criteria, Challenges and Requirements to design privacy preserving data aggregation protocols

The following criteria summarize the required characteristics of privacy preserving data aggregation protocols for WSNs.

Data privacy: Data privacy ensures that the original data of any sensor node S_i should not be revealed to an adversary, or any other trusted participating nodes in the network.

Energy efficiency: Sensor's lifetime is strictly dependent on its power resource, thus the protocol should intelligently use power for preserving their energy. Privacy protection in WSNs consumes additional power, that can not be avoided. A good and efficient protocol should keep that additional communication overhead, computation cost, memory and payload size as small as possible.

Data Accuracy: Data may be lost due to node failure or wireless link during transmission, hence the accuracy of the final aggregated result can be affected. In the applications of WSNs where aggregated results are used to make some critical decision, the accuracy of final aggregated result at the base station is very important with the restriction that private data of sensor node is not disclosed. Therefore, an appropriate method is required to determine an accurate aggregated result in WSNs.

Fault tolerance: Sensor nodes are prone to failure due to lack of energy, hardware failure, and malicious attack. Wireless sensor network must be robust against failure of sensor node and the network functionality must be maintained. New nodes can be added into the network to compensate for failure nodes. A good protocol should allow node addition during data aggregation for maintain network functionality.

4 Conclusion and future work

In this paper, privacy preserving data aggregation protocols for WSNs have been analyzed and classified them according to schemes adopted to satisfy data privacy. Requirements and challenges for designing privacy preserving data aggregation protocols have been identified. We believe that our critical analysis of the existing protocols will provide new researchers guidelines to improve the existing protocols and to design new energy efficient privacy preserving data aggregation protocols for WSNs. The existing privacy-preserving data aggregation protocols have used differ-

ent schemes to achieve data privacy, such as privacy homomorphism, perturbation, and shuffling. Each type of the above schemes has some advantages and limitations which are summarized as:

(1) The protocols based on privacy homomorphism allow aggregation directly on encrypted data. Therefore, it minimizes possibility of attack at aggregator node. However, these schemes can only apply to some query based aggregation functions(e.g, sum, average, etc); hence they limits our ability to perform aggregation in network. Therefore, an efficient privacy homomorphic scheme is required that could support all aggregation functions.

The protocols based on perturbation scheme maintain privacy by exchanging seeds (non-private numbers) and exchanging of their encrypted customized data within a cluster, which results in high communication and computation overhead. Therefore, an optimal method is required to minimize communication overhead while maintaining privacy.

(3) The protocols based on shuffling guarantees data privacy through data `slicing and assembling' with randomly selected J nodes within h hops. Selection of nodes is done only once (i.e., initially). Since sensors are prone to failure due to lack of energy during data aggregation, that results in loss of data and loss in coverage. So, to maintain accuracy and privacy of the remaining data, each node in the J-list of the failed node, after recognizing the failure should broadcast the failure message to their corresponding J-list nodes to discard the failed node data and continue the operation with remaining nodes. Such a broadcast takes O(JN) messages.

Protocol should support node addition for maintaining sensing coverage during data aggregation. This is not addressed in any existing privacy preserving data aggregation protocol.

Our ongoing work is to propose an energy efficient Privacy Preserving data Aggregation (EPPA) for WSNs where a secure key management along with shuffling technique will be adopted that will provide strong security and energy efficient system. In our scheme, nodes maintain privacy through `slicing and assembling' with their siblings within the immediate transmission range. In our work, we propose a new tree construction protocol, which we refer to as Secure Coverage Tree (SCT) protocol and a tree-reconstruction scheme which make it resilient to node failure. We implement our proposed EPPA protocol on top of the Secure Coverage Tree (SCT) protocol. We strongly believe that our proposed scheme will performs better than the existing protocols in terms of communication overhead, security, and fault tolerance.

References

1. Shio Kumar Singh, M. P. Singh, D. K. Singh, "Applications, Classifications, and Selections of Energy-Efficient Routing Protocols for Wireless Sensor Networks", International Journal of Advanced Engineering Sciences and Technologies, 2010.
2. J. Hill, R. Szewczyk, A. Woo, S. Hollar, and D. C. K. Pister, "System architecture directions for networked sensors", In Proceedings of the 9th International Conference on Architectural Support for Programming Languages and Operating Systems, 2000.

3. Kiran Maraiya, Kamal Kant, Nitin Gupta, "Wireless Sensor Network: A Review on Data Aggregation", International Journal of Scientific and Engineering Research, 2011.
4. S. Madden, R. Szewczyk, M. Franklin, and D. Culler, "Supporting aggregate queries over ad-hoc sensor networks", In Workshop on Mobile Computing and Systems Applications, 2002.
5. David Wagner, "Resilient aggregation in sensor networks", In Proceedings of the 2nd ACM workshop on Security of adhoc and sensor networks, 2004.
6. C. Intanagonwiwat, D. Estrin, R. Govindan, and J. Heidemann, "Impact of Network Density on Data aggregation in Wireless Sensor Networks", In Proceedings of the 22nd International Conference on Distributed Computing Systems, 2002.
7. N.Li, N.Zhang, S.K.Das, B.Thuraisingham, "Privacy-preserving in wireless sensor networks: A state-of-the-art survey", Ad Hoc Networks, 2009.
8. Rabindra Bista and Jae-Woo Chang, ``Privacy-Preserving Data Aggregation Protocols for Wireless Sensor Networks: A Survey", Sensors, 2010.
9. D.Westhoff, J. Girao, M. Acharya, Concealed data aggregation for reverse multicast traffic in sensor networks: encryption key distribution and routing adaptation, IEEE Transactions on Mobile Computing, 2006.
10. J. Domingo-Ferrer, "A Provably Secure Additive and Multiplicative Privacy Homomorphism", Proc. Information Security Conf., pp. 471-483, Oct. 2002.
11. C.Castelluccia, E.Mykletun, G.Tsudik, "Efficient Aggregation of Encrypted Data in WSN", in 'MobiQuitous', IEEE computer Society, 2005.
12. Claude Castellucia, Aldar C-F. Chan, Einar Mykletun, and Gene Tsudnik, "Efficient and Provably Secure Aggregation of encrypted data in WSNs", ACM transactions on Sensor Networks, vol.5, no.3, May 2009.
13. O. Goldreich, S. Goldwasser, and S. Micali, "How to construct random functions", Journal of the Association for Computing Machinay, ACM, 1986.
14. W.He, X.Liu, H.Nguyen, K.Nahrstedt, T.Abdelzaher, ``PDA: privacy-preserving data aggregation in wireless sensor networks", IEEE INFOCOM, 2007.
15. S. Madden, M.J. Franklin, J.M. Hellerstein, "TAG: A Tiny AGgregation Service for Ad-Hoc Sensor Networks", OSDI, 2002.
16. M.Conti, L.Zhang, S.Roy, R.D.Pietro, S.Jajodia, L.V.Mancini, "Privacy-preserving robust data aggregation in WSNs", Secur. Commun. Netw. 2009.
17. H. Choi, S.Zhu, and T. F. La Porta, ``SET: Detecting Node Clones in Sensor Networks", In Proceedings of IEEE 3rd International Conference on Security and Privacy in Communication Networks, 2007.
18. L. Eschenauer and V. D. Gligor, ``A key-management scheme for distributed sensor networks", in Proceedings of the 9th ACM Conference on Computer and Communications Security, November 2002.
19. W.He, H.Nguyen, X.Liu, K.Nahrstedt, T.Abdelzaher, ``iPDA: An integrity-protecting private data aggregation scheme for wireless sensor networks", In Proceedings of IEEE Military Communication Conference, MILCOM, San Diego, CA, USA, November, 2008.