RESEARCH STATEMENT

Primal Pappachan

My research interests lie in the intersection of data management, access control policies, and privacy mechanisms. During my Ph.D., I have focused on building *policy-based privacy-by-design data management systems*. With advances in technology and arrival of new domains, the data collected and managed by data management systems increasingly contain newer forms of personally identifying information (PII). It is important that such PII is protected from unwanted access and inferences. For example, today's mobile and wearable devices collect information about movement, heart beats, sleep, and weight which might lead to inferences about users life style and their inclination to pathologies. This requirement of privacy has been made much more urgent by the recent introduction of stringent privacy laws such as General Data Protection Regulation (GDPR) and California Consumer Privacy Act (CCPA).

As a result of this changing landscape of technologies and privacy laws, today's data management systems face a lot of challenges in 1) meeting the privacy requirements mandated by regulations 2) enforcing privacy mechanisms efficiently in real time. Therefore, it is important to redesign these systems taking privacy into consideration. This is challenging as it introduces additional overheads to the different phases of data management. In my Ph.D., I have worked on several of those challenges with a focus on making privacy-by-design data management practical and scalable. In particular, I have worked on 1) Understanding privacy requirements of newer domains such as Internet of Things and coming up with policy models, 2) Building scalable systems for enforcement of user defined policies in static and streaming systems, 3) Enforcing policies while ensuring there are no disclosures of protected data items. As a graduate student, I have had the privilege of working on such challenges with researchers from diverse areas such as database systems, privacy, mobile computing, usable security, Semantic Web, distributed computing, and cyberphysical systems. I have gained a wider perspective of computer science research through such collaborations which has also led to new research projects.

Research activities

I was first inspired about the need for privacy-by-design systems during my Master's at University of Maryland, Baltimore County. I had the exciting opportunity of getting access to smart glasses technology (Google Glass through Google's developer program). After playing with it for few days, I recognized the importance of privacy in a device that is always on. I developed **FaceBlock** [5] which implements *privacy-aware pictures* for smart glasses technology. A privacy-aware picture is the resultant picture after enforcing the privacy requirements from users around the smart glasses user. Bystander's privacy requirements in the form of privacy policies were communicated to the smart glass by their smartphones running FaceBlock. Upon receiving these privacy policies, the FaceBlock running on the smart glass enforced those by obfuscating the users from their pictures. Smart glasses and Google Glass didn't take off, but the lessons learnt from this experience convinced me to pursue building privacy by design systems for my Ph.D. at University of California, Irvine (UCI).

Privacy Policies for Smart Buildings: With the advent of Internet of Things (IoT), almost every device around us is smart, always on, and collecting data about us. The need for privacy has therefore become more crucial in IoT settings. An example of IoT is today's smart buildings where facility managers require a mechanism to notify residents and visitors of data collection practices. Similarly, building residents might want to specify their privacy preferences regarding these data collection practices. I led the development of a policy language/model [3] which can be used to express these requirements. I presented this work at the CS Research showcase at UCI and won the best the best poster award out of 42 participants. I implemented this policy model in the IoT data management system called TIPPERS [1] which is deployed at UCI. The access

[1] https://tippers.ics.uci.edu/

control interfaces I developed for TIPPERS, assisted users in specifying privacy preferences on the data collection practices advertised by the building.

Scalable Fine Grained Access Control Policy Management for DBMS: Based on my work in IoT, it was clear that in these latest domains of Big Data, there could be a large number (more than 100K) of user-defined fine grained policies . When it comes to enforcement of these large number of policies, today's data management systems are not able to efficiently handle the large number of checks required at the time of answering queries. The traditional approaches used in DBMSs are specification of authorization views and query writing. These methods have high overheads which makes them unsuitable for real-time query processing under large policy loads. There is a need for scalable Fine Grained access Control (FGAC) for Database Management Systems (DBMS). I developed **Sieve** [7] which is a general purpose middleware for DBMSs that scales access control for real time query processing. Sieve does this by reducing number of checks to be performed by filtering irrelevant records and impertinent policies. Given a set of policies, Sieve uses them to generate a set of guarded expressions that are chosen carefully to exploit the best existing indexes, thus filtering the tuples against which policies are checked. Sieve also includes a policy evaluation operator which utilizes the context of a record (e.g., user/owner associated of record) and the query metadata (e.g., the person posing the query) to filter away policies which are not of interest to the tuple under consideration. By adaptively combining these two strategies based on a cost model, Sieve is able to significantly reduce overhead of policy checking. We tested Sieve in different IoT settings and compared its performance to baseline approaches of query rewriting on two different DBMSs. The results show that Sieve significantly reduces the overhead (2X-10X) of access control at query time. This work is the first to identify and propose a solution for scalable access control and has opened up an interesting research area.

Enforcing access control policies under data dependencies: In Sieve, policies are specified on data and the enforcement only depends upon controlling the sharing of data tuples directly touched upon by the policy. Typically in data management systems, data is stored and access control policies are specified at different levels - base data tables, semantic views. Therefore it is important to enforce access control policies at these different levels to protect the privacy of the user. I developed a new model of policy enforcement that takes into account the various data dependencies that exists between data while enforcing policies. The data dependencies are relationships between attributes and attribute values stored in the same or different relations. Examples of these are key constraints, association rules, and inverse of view definition functions. Users specify policies on semantic views and the goal of enforcement is make sure that any tuple that is denied by a user is not shared directly or indirectly through data dependencies. In a variation of the problem setting, I also allow users to trade off policy compliance for query answering correctness. This work [4] is currently in progress and will improve the privacy protection offered by Sieve further.

Policy driven privacy preserving data streams: In the previous works, I have focused on building policy-based privacy-aware systems for data stored in DBMSs. In IoT settings, the management of data streams from sensors to service providers is done by data controllers who are also responsible for making sure that these streams are privacy regulation compliant. For example, GDPR mandates that data controllers state data collection and usage purposes (Purpose limitation), and allow users to opt-out of sharing as they prefer (Right to object). I have worked on **PE-IoT** (Privacy Enhanced-Internet Of Things) [2] which intercepts data flows, enforces user policies on them, and transforms them into privacy-enhanced data flows by applying Privacy Enhancing Technologies (PETs) such as encryption, perturbation, and anonymization. It introduces a new abstraction of data sharing, *data products*, which effectively combines policy and mechanism for generating privacy enabled data flows. I also developed a temporal policy model for PE-IoT with the goal of satisfying Purpose limitation and Right to object requirements of GDPR. This model allows data controllers to define a data product and specify the purposes for which it is shared. It also provides a mechanism for data subjects to opt-in/opt-out of a defined data product. The temporal semantics of the policy can be used to retroactively opt-in/opt-out from specific data products. In addition, a data product defined using this policy groups together different data flows using a hierarchy of tags. This reduces the overhead for data controllers in applying PETs which makes the data flows privacy preserving. It also reduces the burden of frequent participation in privacy intervention for data subjects. Retroactive policy semantics provides a flexible and powerful method for data subjects to control sharing of their future and historical data.

Societal Computing: In addition to *policy-based privacy-by design data management systems*, I have also worked on couple of other projects with the goal of using technology for the betterment of society. As my M.S.

thesis, I developed **Rafiki** [6] with the goal of assisting Community Health Workers in underdeveloped areas by inferring possible diseases and treatments by representing the diseases, their symptoms, and patient context in OWL ontologies and by reasoning over this model. During my Ph.D. I closely worked with an independent group of undergraduate students and advised them in developing **ZotBins**[2] [1] - a system that consists of smart bins fitted with sensors, a waste recognition framework, and a variety of user applications - with goal of assisting communities towards achieving Zero Waste.

Future Research

I am excited by the potential of my research in building *policy-based privacy by design data management systems* and intend to continue my work in different directions.

Policy based data management systems: In the short term, I plan to further explore the opportunities for co-optimizing query processing and policy enforcement in DBMS. There is still room for significant improvement if data management systems consider policies as first class citizens. DBMSs can then use this knowledge to build policy-based indices to intertwine query optimization and policy enforcement more closely. Further on, DBMS can also determine unreachable parts of the database based on policies and move these away from caches. Similar to the saying of "One size does not fit all" in database community, "one policy or access control model does not fit for every scenario". Thus, there is a need to develop policy models supporting different data models (e.g., non-relational, array based), different database technologies (e.g., polystores) and new domains (e.g., intelligent transportation systems, smart water management). In many of these scenarios, multiple entities might be participating in data exchanges across different forms of networks such as cloud or fog based. Therefore, it would be important to support multiparty distributed access control across such systems where data is manipulated and policies need to be enforced at different places.

Translation of the Regulatory requirements into System level design choices: As stricter regulations are coming up in different parts of the world, this translation remains an open challenge. For example, the Right to be Forgotten requirement mandates that users should be able to ask data controllers to delete their data and provide proof of such deletion. However, due to data distribution and redundancy in Big data systems, ensuring that data is completely destroyed is extremely challenging. As data moves from one place to the other, compliance of retention policies and their verification has to be repeated again and again. Secure deletion schemes which utilize different cryptographic techniques have been proposed but none of them have been integrated into data management systems. This becomes even more challenging with complex data processing pipelines which utilize Machine Learning algorithms. In this setup, the contribution of each individual pieces of data becomes fuzzy in the deep layers of processing. Similarly, for verifiable compliance and Data Protection Impact Assessments (GDPR Article 35), current models of policy enforcement, require a trusted centralized entity. This becomes challenging in the aforementioned newer domains and therefore a verification method which relies on tamper proof logs is required in distributed settings with no trusted entities. In cases of potentially high-risk processing activities, data controllers can use this to study the impact of privacy policies on individual's data.

Combining Privacy Policy and Privacy Mechanism: In the long term, I would like to explore the advantages of combining these traditionally disparate fields. For example, in Differential Privacy, which provides bounds on privacy leakage with an unknown adversary, an open problem is how to appropriately set the noise factor. The policies specified by users could be potentially used to compute with the appropriate value. There are many challenges to be addressed to do such a combination meaningfully and efficiently. Translation of user specified policies into the parameters of an enforcement mechanism is an exciting avenue to explore. Similarly, building bridges between policy requirements and guarantees of the mechanism is a compelling problem to solve. Combining these two separate areas of research can also spur improvements in implementation of both. Understanding the impact on the privacy guarantees of mechanisms when policies are dynamically updated. Privacy and security systems are only as strong as its weakest component and in today's systems more often than not these are the users who are the least informed on privacy. Through combination of policy and mechanism it becomes possible to build **Explainable Security and Privacy** where users can understand, appropriately trust, and effectively manage the privacy by design systems.

[2] https://zotbins.github.io/

Privacy is a socio-technical challenge and requires an interdisciplinary effort from various perspectives such as computer science theory, social sciences, cybersecurity systems, and user applications. I hope to work with relevant stakeholders to better my own understanding of privacy in their own domains and thus fulfilling my goal of building *policy-based privacy-by-design data management systems* for serving as many challenging scenarios as possible.

References

[1] J. Cao, J. Chong, M. Lafreniere, O. Yang, P. Pappachan, S. Mehrotra, and N. Venkatasubramanian. The ZotBins solution to waste management using Internet of Things. In J. Nakazawa and P. Huang, editors, *18th Int. Conf. on Embedded Networked SensorSystems (SenSys)*, 2020.

[2] S. Ghayyur, P. Pappachan, G. Wang, S. Mehrotra, and N. Venkatasubramanian. Designing privacy preserving data sharing middleware for internet of things. In *3rd Workshop on Data: Acquisition To Analysis (DATA) co-located with SenSys*, pages 1–6, 2020.

[3] P. Pappachan, M. Degeling, R. Yus, A. Das, S. Bhagavatula, W. Melicher, P. E. Naeini, S. Zhang, L. Bauer, A. Kobsa, S. Mehrotra, N. M. Sadeh, and N. Venkatasubramanian. Towards privacy-aware smart buildings: Capturing, communicating, and enforcing privacy policies and preferences. In *37th Int. Conf. on Distributed Computing Systems Workshops, (ICDCS Workshops)*, 2017.

[4] P. Pappachan and S. Mehrotra. Enforcing access control policies under data constraints. *(In Progress)*.

[5] P. Pappachan, R. Yus, P. K. Das, T. Finin, E. Mena, and A. Joshi. A semantic context-aware privacy model for FaceBlock. In *2nd Workshop on Society, Privacy and the Semantic Web - Policy and Technology (PrivOn) co-located with ISWC*, 2014.

[6] P. Pappachan, R. Yus, A. Joshi, and T. Finin. Rafiki: A semantic and collaborative approach to community health-care in underserved areas. In *10th Int. Conf. on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom)*, 2014.

[7] P. Pappachan, R. Yus, S. Mehrotra, and J. Freytag. Sieve: A middleware approach to scalable access control for database management systems. *Proc. VLDB Endow.*, 13(11), 2020.