Viewpoint | Personal Data and the Internet of Things

It is time to care about digital provenance.

Thomas Pasquier
University of Bristol

David Eyers
University of Otago

Jean Bacon
University of Cambridge

ABSTRACT

The Internet of Things promises a connected environment reacting to and addressing our every need, but based on the assumption that all of our movements and words can be recorded and analysed to achieve this end. Ubiquitous surveillance is also a precondition for most dystopian societies, both real and fictional. How our personal data is processed and consumed in an ever more connected world must imperatively be made transparent, and more effective technical solutions than those currently on offer, to manage personal data must urgently be investigated.

The need for greater transparency

We have all read market predictions describing billions of devices and the hundreds of billions of dollars in profit that the Internet of Things (IoT) promises.[1] Security and the challenges it represents [27] are often highlighted as major issues for IoT, alongside scalability and standardisation. In 2017, FBI Director James Comey warned, during a senate hearing, of the threat represented by a botnet taking control of devices owned by unsuspecting users. Such a botnet can seize control of devices ranging from connected dishwashers,[2] to smart home cameras and connected toys, not only using them as a platform to launch cyber attacks, but also potentially harvesting the data such devices collect.

In addition to concerns about cybersecurity, corporate usage of personal data has seen increased public scrutiny. A recent focus of concern has been connected home hubs (e.g., Amazon Alexa, Google Home).[3] Articles on the topic discussed whether conversations were being constantly recorded and if so, where those records went. Similarly, the University of Rennes faced a public backlash after revealing its plan to deploy smart-beds in its accommodation to detect *"abnormal"* usage patterns.[4] A clear question emerges from IoT-related fears, *"how and why is my data being used?"*

As concerns grow, legislators across the world are taking action in order to protect the public. For example, the recent EU General Data Protection Regulation (GDPR) which took effect in May 2018,[5] and the forthcoming ePrivacy Regulation[6] place strong responsibility on data controllers to protect personal data, and to notify users of security breaches. The EU commission defines a Data Controller as the party that determines the purposes for which, and the means by which, personal data is processed (*why* and *how* the data is processed). EU regulations further impose constraints on *where* EU citizens data can be processed and *what* type of data (i.e., "special category" data falls under more stringent constraints). The data controller must provide means for end users to determine whether their data is properly handled and means to effect their rights. Overall, there must be mechanisms to determine *what* data is processed, *how*, *why* and *where*.

Such concerns have drawn researchers to look at means to develop more accountable and transparent systems [10, 24]. The problem has also been clearly highlighted by the EU Data Protection Working Party: *"As a result of the need to provide pervasive services in an unobtrusive manner, users might in practice find themselves under third-party monitoring. This may result in situations where the user can lose all control on the dissemination of his/her data, depending on whether or not the collection and processing of this data will be made in a transparent manner or not."*

Indeed, modern computing systems contain many components that operate as black boxes; they accept inputs and generate outputs but do not disclose their internal working. Beyond privacy concerns, this also limits the ability to detect cyber-attacks, or more generally to understand cyber-behaviour. Because of these concerns DARPA, in the US, launched the *Transparent Computing* project[7] to explore means to build more transparent systems through the use of digital provenance with the particular aim of identifying advanced persistent threats. While DARPA's work is a good start, we believe that there is an urgent need to reach much further. In the rest of the article, we explore how provenance can be an answer to some IoT concerns and the challenges faced to deploy provenance techniques.

Digital Provenance

There is a growing clamour for more transparency, but straightforward, widespread technical solutions have yet to emerge. Typical software log records often prove insufficient to audit complex distributed systems as they fail to capture the complex causality relationships between events. Digital provenance [8] is an alternative means to record system events. Digital provenance is the record of information flow within a computer system in order to assess the origin of data (e.g., its quality or its validity).

The concept first emerged in the database research community as a means to explain the response to a given query [16]. Provenance research later expanded to address issues of scientific reproducibility, notably by providing mechanisms to reconstitute computational environments from formal records of scientific computations [23]. More recently, provenance has been explored within the cybersecurity community [25] as a means to explain intrusions [18] or more recently to detect them [14].

[1] https://www.forbes.com/sites/louiscolumbus/2017/12/10/2017-roundup-of-internet-of-things-forecasts/#94d9f6c1480e
[2] https://nvd.nist.gov/vuln/detail/CVE-2017-7240
[3] https://www.wired.com/2016/12/alexa-and-google-record-your-voice/
[4] http://www.lemonde.fr/pixels/article/2017/09/07/le-crous-de-rennes-annule-une-experimentation-de-lits-connectes-dans-une-cite-universitaire_5182434_4408996.html
[5] http://www.privacy-regulation.eu/en/index.htm
[6] https://ec.europa.eu/digital-single-market/en/proposal-eprivacy-regulation
[7] https://www.darpa.mil/program/transparent-computing

Provenance records are represented as a directed acyclic graph that shows causality relationships between the states of the objects that compose a complex system. As a consequence, it is compatible with automated mathematical reasoning. In such a graph, the vertices represent the state of transient and persistent data items, transformations applied to those states, and persons (legal or natural) responsible for data and transformations (generally referred to as entities, activities and agents respectively). The edges represent dependencies between these entities. The analysis of such a graph allows us to understand *where*, *when*, *how*, *by whom* and *why* data has been used [7, 9].

An outcome of research on provenance in the cybersecurity space is the understanding that the capture mechanism must provide guarantees of completeness (i.e., all events in the system can be seen), accuracy (i.e., the record is faithful to events) and a well-defined, trusted computing base (i.e., the threat model is clearly expressed) [22]. Otherwise, attacks on the system may be undetected, dissimulated by the attacker or misattributed. We argue that in a highly *ad hoc* and interoperable environment with mutually untrusted parties, the provenance used to empower end users with control and understanding over data usage requires similar properties.

Who to trust?

In the IoT environment the number of involved stakeholders has the potential to explode exponentially. Traditionally, a company managed its own server infrastructure, maybe with the help of a subcontractor. The cloud computing paradigm further increased complexity with the involvement of cloud service providers (sometimes stacked, e.g., Heroku PaaS on top of the Amazon IaaS cloud service), third party service providers (e.g., CloudMQTT) and other tenants sharing the infrastructure. The IoT further increases this complexity, with potentially *ad hoc* and unforeseen interactions between devices and services on top of the complex cloud and edge computing infrastructure most IoT services rely on.

One answer to this problem is to build applications in "silos" where the involved parties are known in advance, but as a side-effect locking-in devices and services to a single company (e.g., the competing smart-home offerings by leading technology companies). This is far from the IoT vision of a connected environment, but most existing products fall in this category. There are obviously major business considerations behind this model, and it should be noted that the EU GDPR mandates for some form of interoperability (although it is yet unclear how it should be interpreted [12]).

An alternative to such *"lock-in"* would be to make devices' consumption of data transparent and accountable. If data is exchanged across devices, the concerned user should be able to audit its usage. However, in an environment where arbitrary devices could interact (although it must be remembered that EU GDPR requires explicit and informed user consent), how can trust be established in the audit record? This requires an in-depth rethinking of how IoT platforms are designed, potentially exploring the security-by-design approach based on hardware roots of trust [13] to provide trusted digital enclaves in which behaviour can be audited and to encourage some form of *"accountability-by-design"* principles where transparency and the implementation of a trustworthy audit mechanism is a core concern in product design.

Such solutions have been explored in the provenance space, for example, by leveraging Software Guard Extensions (SGX) properties to provide a strong guarantee of the integrity of the provenance record [4]. Similarly, remote attestation techniques leveraging Trusted Platform Module (TPM) hardware have been proposed [6] to guarantee the integrity of the capture mechanism. However, how to provide such guarantees in an IoT environment, where such hardware features may not be available, is a relatively unexplored topic.

Where does the audit live?

The fully realised IoT vision is of vast distributed and decentralised systems. If we assume trustworthy provenance capture is achievable, the issue of guaranteeing that the provenance record can be audited remains. If you are to audit the processing of personal data, guarantees about the integrity and availability of the provenance record must exist. If you agreed to share your daily activity for research, the activities of insurance companies scraping your data for possible health risks must not be able to masquerade as benign research use, nor should data collection for political purposes be able to pass as harmless entertainment, as in the Cambridge Analytica scandal.[8] Similarly, the availability (durability) of the audit record must be guaranteed. There is no point to an audit record if it can simply be deleted.

Further, Moyer et al. evaluated the storage requirements of provenance when used for security purposes in relatively modest distributed systems [21]. In such a context, several thousands of graph elements can be generated per second and per machine, resulting in a graph containing billions of nodes to represent system execution over several months. It is unclear how some past research outcomes: e.g., detection of suspicious behaviour [2], privacy-aware provenance [11] or provenance integrity [15]; scale to very large graphs as such concerns were not evaluated. Similarly, while blockchain is heralded [19] as an integrity-preserving means to store provenance, it is unclear how well it could expand to such scale. Several options have been explored to reduce graph size, such as identifying and tracking only sensitive data objects [5] or performing property-preserving graph compression [17] however none has yet adequately addressed the scalability challenge.

How to communicate information?

Means must be developed to communicate about data usage, but also about the risks of inference from the data. Not only must the nature of the data be considered, but also other properties such as the frequency of capture [3]. For example, a 100 Hz smart meter reading can in some cases indicate what television channel is currently being watched; even a daily average reading could inform about occupancy. Here, it is important to be able to explore and represent the outcome of complex computational workflow [1].

Provenance visualisation has been an active research topic for over a decade, yet no fully satisfactory solution has been proposed. The simplest possible visualisation is to render the graph, however

[8] https://www.nytimes.com/2018/03/17/us/politics/cambridge-analytica-trump-campaign.html

beyond trivially simple graphs such a representation is too complex and dense to be easily understood, even by experts. We go further and suggest that educational background, socio-economic environment and culture may play a part in how interpretable such information is.

In order to make the accountability and transparency of IoT platforms effective, a better communication medium must be provided. An approach often taken is to analyse motifs in the graph to extract high-level abstractions (e.g., Missier et al. [20]), meaningful to the average end-user. In recent work [26], it was proposed to represent such a high-level abstractions as a comic strip.

We need to care about digital provenance

Building transparent and auditable systems may be one of the greatest software engineering challenges of the coming decade. As a consequence, digital provenance and its application to cybersecurity and the management of personal data has become a hot research topic. We have highlighted key active areas of research and their associated challenges. It is fundamental for industry practitioners to understand the threat posed by the black-box nature of the IoT, the potential solutions, and the challenges to a practical deployment of those solutions. Accountability-by-design must become a core objective of IoT platforms.

Thomas Pasquier (http://tjmp.org) is a Lecturer (Assistant Professor) at the University of Bristol in the Cyber Security Group, and a visitor at the University of Cambridge in the Department of Computer Science and Technology.
David Eyers (https://www.cs.otago.ac.nz/staff/David_Eyers) is an Associate Professor in the Department of Computer Science at the University of Otago.
Jean Bacon (http://www.cl.cam.ac.uk/~jmb25/) is Professor Emerita of Distributed Systems at the University of Cambridge.

REFERENCES

[1] Umut Acar, Peter Buneman, James Cheney, Jan Van Den Bussche, Natalia Kwasnikowska, and Stijn Vansummeren. 2010. A graph model of data and workflow provenance.
[2] M David Allen, Adriane Chapman, Len Seligman, and Barbara Blaustein. 2011. Provenance for collaboration: Detecting suspicious behaviors and assessing trust in information. In *Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom), 2011 7th International Conference on*. IEEE, 342–351.
[3] Yousef Amar, Hamed Haddadi, and Richard Mortier. 2018. An Information-Theoretic Approach to Time-Series Data Privacy. In *Proceedings of the 1st Workshop on Privacy by Design in Distributed Systems*. ACM, 3.
[4] Nikilesh Balakrishnan, Lucian Carata, Thomas Bytheway, Ripduman Sohan, and Andy Hopper. 2017. Non-repudiable disk I/O in untrusted kernels. In *Proceedings of the 8th Asia-Pacific Workshop on Systems*. ACM, 24.
[5] Adam Bates, KR Butler, and Thomas Moyer. 2015. Take only what you need: leveraging mandatory access control policy to reduce provenance storage costs. In *Conference on Theory and Practice of Provenance. USENIX*. 7–7.
[6] Adam M Bates, Dave Tian, Kevin RB Butler, and Thomas Moyer. 2015. Trustworthy Whole-System Provenance for the Linux Kernel. In *USENIX Security Symposium*. 319–334.
[7] Peter Buneman, Sanjeev Khanna, and Tan Wang-Chiew. 2001. Why and where: A characterization of data provenance. In *International conference on database theory*. Springer, 316–330.
[8] Lucian Carata, Sherif Akoush, Nikilesh Balakrishnan, Thomas Bytheway, Ripduman Sohan, Margo Seltzer, and Andy Hopper. 2014. A primer on provenance. *Commun. ACM* 57, 5 (2014), 52–60.
[9] James Cheney, Laura Chiticariu, Wang-Chiew Tan, et al. 2009. Provenance in databases: Why, how, and where. *Foundations and Trends® in Databases* 1, 4 (2009), 379–474.
[10] Andy Crabtree, Tom Lodge, James Colley, Chris Greenhalgh, Kevin Glover, Hamed Haddadi, Yousef Amar, Richard Mortier, Qi Li, John Moore, et al. 2018. Building accountability into the Internet of Things: the IoT Databox model. *Journal of Reliable Intelligent Environments* (2018).
[11] Susan B Davidson, Sanjeev Khanna, Tova Milo, Debmalya Panigrahi, and Sudeepa Roy. 2011. Provenance views for module privacy. In *Proceedings of the thirtieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. ACM, 175–186.
[12] Paul De Hert, Vagelis Papakonstantinou, Gianclaudio Malgieri, Laurent Beslay, and Ignacio Sanchez. 2017. The right to data portability in the GDPR: Towards user-centric interoperability of digital services. *Computer Law & Security Review* (2017).
[13] Karim Eldefrawy, Gene Tsudik, Aurélien Francillon, and Daniele Perito. 2012. SMART: Secure and Minimal Architecture for (Establishing Dynamic) Root of Trust. In *Network and Distributed System Security Symposium*, Vol. 12. 1–15.
[14] Xueyuan Han, Thomas Pasquier, Tanvi Ranjan, Mark Goldstein, and Margo Seltzer. 2017. FRAPpuccino: Fault-detection through Runtime Analysis of Provenance. In *Workshop on Hot Topics in Cloud Computing (HotCloud'17)*. USENIX, USENIX.
[15] Ragib Hasan, Radu Sion, and Marianne Winslett. 2009. The Case of the Fake Picasso: Preventing History Forgery with Secure Provenance. In *FAST*, Vol. 9. 1–14.
[16] Melanie Herschel, Ralf Diestelkämper, and Houssem Ben Lahmar. 2017. A survey on provenance: What for? What form? What from? *The VLDB Journal—The International Journal on Very Large Data Bases* 26, 6 (2017), 881–906.
[17] Md Nahid Hossain, Junao Wang, R Sekar, and Scott D Stoller. Dependence-Preserving Data Compaction for Scalable Forensic Analysis. In *USENIX Security Symposium*.
[18] Samuel T King and Peter M Chen. 2003. Backtracking intrusions. *ACM SIGOPS Operating Systems Review* 37, 5 (2003), 223–236.
[19] Xueping Liang, Sachin Shetty, Deepak Tosh, Charles Kamhoua, Kevin Kwiat, and Laurent Njilla. 2017. Provchain: A blockchain-based data provenance architecture in cloud environment with enhanced privacy and availability. In *International Symposium on Cluster, Cloud and Grid Computing*. IEEE/ACM, 468–477.
[20] Paolo Missier, Jeremy Bryans, Carl Gamble, Vasa Curcin, and Roxana Danger. 2014. ProvAbs: model, policy, and tooling for abstracting PROV graphs. In *International Provenance and Annotation Workshop*. Springer, 3–15.
[21] Thomas Moyer and Vijay Gadepally. 2016. High-throughput ingest of data provenance records into Accumulo. In *High Performance Extreme Computing Conference (HPEC), 2016 IEEE*. IEEE, 1–6.
[22] Thomas Pasquier, Xueyuan Han, Thomas Moyer, Adam Bates, Olivier Hermant, David Eyers, Jean Bacon, and Seltzer Margo. 2018. Runtime Analysis of Whole-System Provenance. In *Conference on Computer and Communications Security (CCS'18)*. ACM.
[23] Thomas Pasquier, Matthew K Lau, Ana Trisovic, Emery R Boose, Ben Couturier, Mercè Crosas, Aaron M Ellison, Valerie Gibson, Chris R Jones, and Margo Seltzer. 2017. If these data could talk. *Scientific Data* 4 (2017), sdata2017114.
[24] Thomas Pasquier, Jatinder Singh, Julia Powles, David Eyers, Margo Seltzer, and Jean Bacon. 2018. Data provenance to audit compliance with privacy policy in the Internet of Things. *Personal and Ubiquitous Computing* (2018), 333–344.
[25] Devin J Pohly, Stephen McLaughlin, Patrick McDaniel, and Kevin Butler. 2012. Hi-Fi: collecting high-fidelity whole-system provenance. In *Proceedings of the 28th Annual Computer Security Applications Conference*. ACM, 259–268.
[26] Andreas Schreiber and Regina Struminski. 2017. Tracing personal data using comics. In *International Conference on Universal Access in Human-Computer Interaction*. Springer, 444–455.
[27] Jatinder Singh, Thomas Pasquier, Jean Bacon, Hajoon Ko, and David Eyers. 2016. Twenty security considerations for cloud-supported Internet of Things. *IEEE Internet of Things Journal* 3, 3 (2016), 269–284.