Workshop Report: Provenance, Security, and Machine Learning

Adriane Chapman, James Cheney, and Paolo Missier

August 2020

1 Introduction

Recent proposals for securing systems against sophisticated attackers, such as Advanced Persistent Threats (APTs), include wholesale monitoring of system activity down to the level of individual system calls, sometimes in a causal, graph-based representation called provenance. While this level of detail should ensure malicious activity is recorded, and that the sources and effects of such activity can be understood after-the-fact, it is challenging to locate such activity against a background of high-volume, high-velocity activity, with high variability in the structure and meaning of data obtained from different sources. Automatic, reliable detection of realistic APT behaviour in provenance traces (with an acceptable false-positive rate) appears to be an open problem.

Until recently, progress in this area has been hindered by the absence of publicly-available datasets. In the US, a recently-concluded DARPA research program on Transparent Computing has produced publicly-available datasets including realistic APT behaviour in provenance traces recorded on a variety of mainstream operating systems. However, these datasets are not easy for the broader research community to reuse, due to large scale (each day of activity can require more than a gigabyte), and heterogeneity. Moreover, attacks are highly imbalanced (often consisting of under 0.01% of the data) and ground truth information that could be used for training is usually not available, rendering supervised machine learning techniques ineffective. Relevant techniques for semi-supervised or unsupervised machine learning or outlier/anomaly detection may require adaptation to deal with the large scale of the data or its complex structure.

This workshop aimed to bring together researchers with expertise in security, data management, and machine learning, each of which bear on this challenge. The workshop also involved participants in the DARPA Transparent Computing program who shared experience and understanding of the problem and the available datasets. Breakout sessions were organised to enable participants to contribute to a research vision and agenda for future work in this area, which forms the basis of this workshop report.

The event sought to bring together individuals who are interested in the following areas of data science and AI:

- Security researchers interested in provenance analysis or advanced persistent threats
- Data scientists interested in applying anomaly detection and unsupervised machine learning to security problems
- Database or distributed systems experts interested in supporting high-performance security analysis over complex information streams

The topics of this workshop cut across several challenge themes at the Turing, particularly Defence & Security. Participants from a range of backgrounds were welcome to apply to attend.

2 Program and Talk Abstracts

2.1 Agenda

- 09:45–10:20 Registration
- 10:20–10:30 Introduction and welcome
- 10:30–10:50 *Building a provenance-based IDS and the questions we ask ourselves*, Thomas Pasquier (University of Bristol)
- 10:50 - 11:10 *Tracking and analysis of causality at enterprise level*, Gabriela Ciocarlie (SRI International)

- 11:10– 11:30 *Visualization for provenance, security and machine learning*, Nick Holliman (Newcastle University)
- 11:30 – 11:50 Refreshment break
- 11:50 – 12:10 *Provenance, AI and proof in court*, Steven McGough (Newcastle University)
- 12:10 – 12:30 *Making Deep Neural Networks (DNNs) less susceptible to adversarial trickery*, Katy Warr (Roke Manor)
- 12:30 – 13:00 Panel discussion
- 13:00 – 13:45 Lunch
- 13:45 - 15:00 Break out sessions phase I
- 15:00 – 15:20 Refreshment break
- 15:20 - 15:45 Break out sessions reporting and refocusing
- 15:45 - 16:30 Break out sessions phase II
- 16:30 - 17:30 Discussion, report planning and the next steps

2.2 Talk Abstracts

- *Building a provenance-based IDS and the questions we ask ourselves*, Thomas Pasquier (University of Bristol)

  Provenance is the representation of a system execution as a directed acyclic graph. Whole-system provenance graph, representing the execution of an entire system from initialisation to shut down, can be comprised of millions of graph elements. It is believed that the use of such graphs can help build better intrusion detection systems. We have attempted to build full stack intrusion detection systems from kernel capture up to the data analysis. In the spirit of a constructive workshop, in this talk, I will present those attempts discussing our design decisions and the questions that we need to answer.

- *TRACE: Tracking and analysis of causality at enterprise level*, Gabriela Ciocarlie (SRI International)

  Tracking provenance and causality across an enterprise at scale and with flexible granularity requires a combination of technologies that are integrated in a holistic fashion. We present TRACE, a framework that combines host-level tracking techniques with a proven enterprise-wide tracking system. At the host level, TRACE performs static analysis to identify unit structures and inter-unit dependences, such that an output event is causally associated with the input events within the same unit. The enterprise-wide provenance tracking system builds upon the SPADE engine. The system collects provenance from individual entities such as hosts and constructs a distributed enterprise-wide causal graph.

- *Visualization for provenance, security and machine learning*, Nick Holliman (Newcastle University)

  Visualization methods are key to linking computational processes and human processes, they form the primary method for communicating data science results to augment human decision making. It is important to bear in mind human cognition has limits and that it is easy to overwhelm these limits with overly interactive, overly detailed visualizations. We argue we must take a principled approach to the science and engineering of visualization methods if we are to demonstrate their value.

  We will then outline tools and approaches relevant to visualization in provenance, security and machine learning, including record-based and graph-based visualization tools. These tools can help data scientists explore and understand the shape of data sets before undertaking detailed analysis.

  While there are many visualization toolsets available for technical specialists there are far fewer available for communicating to high level decision makers and the general public, people who don't have the technical training to interrogate statistical analyses. We will then summarise our own work on visual entropy and suggest how it could provide a visualization method for summarising complex outcomes to non-technical audiences. To illustrate this, we will present a live example using visual entropy glyphs to communicate machine learning classifier outcomes.

  We will conclude by suggesting some of the key challenges in visualization that relate to provenance, security and machine learning including:

  – Provenance visualization for non-technical users.

  - Visualization summarisation tools for large scale graphs and temporal data flows.
  - Uncertainty visualization and its impact on decision makers.

- *Provenance, AI and proof in court*, Steven McGough (Newcastle University)

  Much has been made in recent years about the use of AI in criminal cases. These range from the good stories of AI's able to make the same judgements as real judges through to shocking cases where AI has flagged up prejudice within the legal system such as bias in setting bail. In the CRITICAL project we are developing AI techniques to identify who are the 'bad guys' from cases in the Cloud. However, in order to use this evidence in court we need to do more than just say 'the computer says it was him'. In order to make a case in court we need to provide a chain of evidence from the original source data through to the final conclusion showing not only what the data was, but also what was applied to the data to come up with the final conclusions. In this talk I shall outline some of the major issues which need to be addressed with this process and the challenges which need to be overcome.

- *Making Deep Neural Networks (DNNs) less susceptible to adversarial trickery*, Katy Warr (Roke Manor)

  The Deep Neural Networks (DNNs) that play a pivotal role in modern day unstructured and complex data processing have proven susceptible to adversarial trickery. For image data, the popular press has referred to these tricks as 'Optical Illusions for AI'; subtle changes that are either ignored or imperceptible to humans but cause DNNs to misinterpret the data. In this presentation, Katy will explain why these attacks are possible, examine the risks that they might present, and explore potential mitigations.

3 Working Groups

3.1 Data

This breakout group focused on challenges associated with managing and processing (large amounts of) provenance data gathered for security purposes. As is the case in many current settings, there are so-called "big data" challenges resulting from unprecedented *volume*, *velocity* and *variety* of data being recorded by provenance systems. These challenges raise the need for more sophisticated systems based on incremental or streaming processing of data, distributed querying, etc. A further requirement is that data be retained for long periods (e.g. up to 10 years) to support audit or regulatory compliance policies. Many of these are general purpose data management needs that can often be met by standard systems, including scalable, open source database management systems, however, the particular issues of provenance may raise new issues not considered in other settings (e.g. to deal with complex property graph structures that are different from the social network graphs often used to benchmark large-scale databases).

The working group discussed several additional broad themes.

Policies for data and its provenance Retaining large amounts of data over long periods has a cost. Often, databases have some retention policy regarding how long certain information must be kept or when it must be discarded. Without such a policy, the default is often to keep everything just in case it is needed later. However, doing so is not only costly, but also creates technical debt because it may be difficult to navigate or comprehend the retained data.

In a security context, a major challenge is that it is difficult to anticipate what patterns are exhibited in provenance records prior to an attack. While it may be easy to detect attacks retroactively based on known patterns, future attacks may not match these past attack "signatures" just as with pattern-based anti-virus signatures. In the settings where provenance-based security is needed, attackers may gain access to a system in order to monitor it or consolidate their access but wait for months or years before performing hostile actions. For example ransomware attacks may infiltrate a system and then wait a year before taking action.

This means that discarding data (e.g. for cost reasons) or abstracting it (e.g. to remove detail that doesn't appear necessary for attack detection) always carries a risk that evidence of an attack may be irretrievably lost. Thus, either completely lossless storage/compression techniques should be used, or retention/deletion policies should be based on some foundations that ensure that no data needed to detect an attack is discarded. But since it is hard to formally define attacks, it seems difficult to determine in advance what information is safe to discard.

A related issue is the need for high-level agreements ("data level agreements") regarding accountability for cloud data services. When sensitive data is stored on cloud servers, it is important to have a clear provenance record regarding what has been done to the data, where it has been stored and so on, including situations in which systems or services have been compromised. Current service level agreements focus instead on backup policies and availability, but do not consider integrity or accountability concerns that provenance is intended

to address. Changing practices in this area seems difficult due to the lack of incentives, and may need to be led by regulation or actions by major players, based on proof-of-concept research prototypes.

Data usage policies *Data trusts* are being proposed to manage how certain datasets can be used when they cannot just be made open. This is a particular issue with research datasets where there is a privacy dimension to the data. Usage policies are currently set at the level of the company that retains the data, whose motivation may be to maximise extracting value from its (users') data rather than maximise benefits to the users or respect their privacy. It would be desirable to give more control to users themselves to allow or deny access to their data. Provenance may be useful both for assessing whether such data is suitable for a given purpose, and for users (or systems acting on their behalf) to assess whether the data has been used according to the users' preferences. Currently, however, there is no regulation mandating that this level of control be available nor are there widely used systems capable of enforcing such data usage policies. Finally, in this setting the meta-level issue of safeguarding the integrity of the provenance record itself is an issue — the so-called provenance-of-provenance.

Data drift and reproducibility *Data drift* is a key concern with any data analysis. The issue is that if the behaviour of the system generating some data varies over time, then training data obtained at one time may lose predictive value. This issue can of course arise in provenance security scenarios in which we train a machine learning classifier or anomaly detector on "normal" data, but over time what is "normal" changes resulting in a high error rate. On the other hand, retaining adequate provenance records of datasets used for training (for any application) would be a way to help mitigate data drift, by at least making it possible to recognise when a training dataset might be out of date or otherwise non-representative of new incoming data.

In security settings, the problem of data drift is exacerbated by the fact that curated, balanced, annotated training data is often unavailable, even for research settings; in real settings, an unsupervised approach is indispensable since attacks are never labeled in advance and are typically a very small fraction of the whole dataset. This means, however, that it may be even harder to tell when data drift has occurred. It may be necessary to identify, or adapt existing, data drift detection/mitigation techniques for unsupervised anomaly detection.

3.2 Provenance

In addition to using provenance to identify Advanced Persistent Threats (APTs) as in the DARPA Transparent Computing project, another use of provenance was identified as being of high interest, related to security, and providing additional examples of many of the concerns also evident in APTs. Provenance can be used to provide assurance. For instance, at a personal level, can an individual be assured that personal data is being used as agreed? In a similar, industrial example, provenance can be used to assure customers that the cloud providers are behaving according to contract.

Discussion around detection of Advanced Persistent Threat (APT) activity and provenance for assurance led to the identification of requirements and gaps in provenance that need to be considered in order to make provenance useful, including: provenance content; provenance quality metrics; provenance modelling; privacy requirements.

Provenance Content Concerns The examples of provenance captured within the DARPA Transparent Computing project and at the level of the OS raised concerns about determinism, and the ability to produce "the same" provenance for the same executions. There is a tension on the representation and abstraction of these non-deterministic executions and the future usage of provenance. Because the OS is non-deterministic, and the state of the system is so complex it is never the same, the provenance captured will never be deterministic.

From an attack detection standpoint, the provenance must accurately reflect this non-determinism. If the provenance is abstracted to be deterministic, any attacks that utilise a non-deterministic approach would be missed. On the other hand, with a non-deterministic system, it is impossible to perform a deterministic analysis. For use within the judicial system, or for reproducible science, a deterministic analysis of the provenance is needed.

In order to bridge this gap, it is important to explore approaches to capture provenance that support both of these uses. A possible approach is to move deterministic analysis to the application level instead of the OS-layer. This obviously raises possible problems in mapping between layers, completeness and integrity. It was noted that within the Avionics industry, many non-deterministic real time systems have a certification of determinacy, and exploration of these systems may facilitate understanding in this area.

Provenance Quality Metrics In order to help judge whether the provenance captured is appropriate for a given use, several possible measurements were put forward. These include:

1. Accuracy: The provenance actually represents what happened.
2. Completeness: The provenance contains all activities and entities and agents in a given time period.
3. Integrity: The provenance provided has not been modified.
4. Granularity: The level to which distinct entities and activities are distinguished within the provenance record.
5. Confidentiality: The level of information exposure or protection within a provenance record.
6. Availability: The access and usage restrictions on the provenance information.

It is possible to have provenance that satisfies some but not all of these metrics. For instance, it is possible to have accurate but incomplete provenance. Additionally, full completeness is impossible. Instead, guarantees on the level of completeness will allow determination if the provenance is fit for purpose, e.g. the completeness level that can actually be used to detect a given attack. In a similar vein, it is possible to be complete, but not granular. The level of granularity required effectively becomes a choice of information gain vs utility and performance. The actual granularity needs is based on analysis for which provenance is being captured. For instance, in the Transparent Computing program, granularity was needed to disentangle relationships, particularly causality relationships vs correlation. However, by providing too much granularity, it may become impossible to reassemble the provenance in a meaningful way.

Moreover, the measurement of accuracy is also dependant upon the provenance modelling performed. If there are multiple semantic interpretations of the same events, and those interpretations have different structures, they can all be accurate, according to their model, but wildly divergent from each other.

Measuring the integrity of the provenance also requires guarantees that formally state that the provenance you have collected is actually correct. If you have strong guarantees on the system performance, the need to capture provenance is reduced. However, despite two major contributions towards the provenance and integrity problem, more work in this area is needed. It is essential to have to have mechanisms to "monitor the monitor"; that is, if provenance is to be used to detect threats, it is essential to know that it is functioning properly. Within this space, achieving integrity of the capture is possible, but keeping the data integrity is still an ongoing problem. Techniques suggested so far do not scale at either the point of signing or verifying.

Expansion of, and organisation of, repositories of provenance samples would benefit the community. The repositories cannot be a mere "dumping ground" of provenance datasets. Instead, the provenance needs to be accompanied by information about: the purpose for which it was collected; how it was collected; the set of systems, and usage of those systems the traces relate to, and the quality metrics identified above for those traces.

Provenance modelling The provenance model chosen impacts both the content and quality metrics as discussed above. Moreover, in a large, distributed system, in which many stakeholders exist, it is imperative that the semantics of the provenance are very clearly stated so that future analysis can be performed. Within the DARPA Transparent Computing, a common data model (CDM) existed. However, because each provider used the terms slightly differently, the provenance was not transferable across providers, and was not easy to analyse. In order to support future analysis, interoperability, expectation management and semantic equivalence mappings are essential. Within the DARPA Transparent Computing program, each of the provenance systems were operating at a different OS levels, and it was difficult to get semantic equivalence across those concepts. Instead of integrated provenance, the result is multiple sources of provenance, that can be partially ordered across paths, assuming multi systems reporting and accurate clocks.

Directly related to the provenance modelling questions is the "capture location". The possibilities for where to capture provenance information range from kernel space to user space. The choice of capture location influences:

1. Intrusiveness. The ability to observe and log provenance information easily depends on how the underlying program is designed. In some cases, this is non-intrusive log-monitoring. In others, an intrusive modification of the underlying program.
2. Richness of information capturable. Depending on where the provenance capture occurs, the information available or "visible" is different.

3. Cost. While libraries that facilitate provenance capture exist, the software engineering must still occur to place those calls. Depending on the intrusiveness of the capture, as discussed above, the number, heterogeneity, ownership, openness and richness of the required provenance, it can be costly to capture enough provenance to create complete or usable provenance information. In addition to the cost of creation, the cost of system maintenance must also be considered.

Protection Provenance can contain personal information (e.g. subscriber id in a smart meter), as well as general networking traffic which contains information that can identify individuals and their information. It is necessary to protect information found within the provenance records while maintaining the utility of the provenance. While there is some work already done in this area, other avenues should be explored.

For instance, can computation and analysis be done over the secure provenance? Some systems are moving provenance via encrypted channels, but the data access and analytics over it is performed after decryption. This is problematic not only because private information can be extracted after decryption, but also because it takes too long to decrypt. Most work on querying encrypted databases has been on relational databases, and little on graph databases. More work in this area would facilitate keeping provenance around for use in detecting APTs and for assurance.

Finally, a cryptology formalism for provenance could benefit the community to facility integrity and privacy of collected provenance information. Starting with a hardware root of trust to guarantee the integrity of a given system, Multiparty Communication (MPC) and homomorphic or polymorphic encryption should be considered.

3.3 Machine learning

The *Machine learning* WG focused on the connection between provenance and machine learning, seeking to explore new and unanticipated roles for provenance in the context of machine learning. The opposite, namely the question of how to apply ML techniques to provenance analysis, was out of the scope of this discussion.

This breakout group covered a number of different topics, initially considering a standard data-to-knowledge pipeline consisting of:

1. Data acquisition and selection from multiple sources
2. Quality assurance and data pre-processing / preparation
3. Model Learning
4. Model validation
5. Model Predictions and their interpretability

The breakout session did not spend much time discussing specific details of learning techniques that could be used in the third step, other than to observe that in security settings, unsupervised techniques are called for since future attack behaviour may not be similar to past observations. Instead, the breakout group mostly focused on challenges in the pre-learning stages (1,2) or post-learning stages (4,5).

Data acquisition and selection from multiple sources Some of the issues with data acquisition in the DARPA Transparent Computing program are due to data coming dynamically from multiple sources, without any annotations and no defined semantics. That is, different provenance recording systems sometimes represented similar behaviour in different ways, making it difficult to develop general techniques for recognising known "suspicious" patterns (e.g. network access to download and write file, then execute file) or to generalise from notions of "normal" behavior learned from one dataset to another. In this situation, the only viable approach has been to try and detect anomalies by means of unsupervised learning models. This, however, proved to be noisy, leading to the need for interpretability or explainability of results, as discussed later; for example, in an operating system there are a few very special system processes such as `init` whose behaviour is distinctive, and thus anomalous, but which are (hopefully) not part of an attack.

Quality assurance and data pre-processing / preparation A related issue is data poisoning or pollution, which may affect datasets derived from provenance-based security scenarios just as in general. Provenance techniques may offer potential solutions here, but they can also create new problems. For example, when there is a risk that data be affected by deliberate pollution (a form of explicit bias), can provenance *about the data sources and their content* be used to track this type of bias? Conversely, is it possible to improve data quality using machine learning techniques, e.g. using anomaly detection or clustering to help clean the data?

The working group speculated that it would be interesting to collect provenance through the entire pipeline and use it (either through hand-written analysis or using some kind of machine learning) to generate automated data quality assessment across data types. A key challenge in this setting, as with provenance-based security generally, would be how to generate suitable ground truth annotations, either for training a supervised quality assurance technique or to evaluate unsupervised methods.

Model validation The question of the future roles of humans in AI-assisted decision making is increasingly important in many areas, including the digitisation of evidence and judicial processes as covered in one of the plenary talks. Model outcomes should be validated with help from human assessment. In a security setting, it seems appropriate to perform data triage: that is, seek to eliminate false negatives (and avoid missing any real attacks) and prioritise recall, while minimising the false positive rate (that is, minimising the effort taken to investigate anomalous behaviour that turns out to be benign). Some work was mentioned that has already been done towards *active learning* for improving the effectiveness of basic unsupervised anomaly detection techniques: given a proposed ranking of anomalous objects, a human user evaluates each one starting at the top. Feedback that a result is an actual attack results in similar objects being given a slightly higher anomaly score, while false positives (benign but anomalous activity) result in reweighting so that similar (and hopefully benign) activity is given a lower anomaly score in the future. Of course, this approach does rely on the human analysts' decisions being correct; this problem can be mitigated by having several different experts evaluate the same data.

Interpretation of model predictions Explainability of machine learning models and predictions has become recognised as an important general problem, and this is also the case when applied to provenance. On the other hand, provenance records about the training data might form part of the information available to explain or account for the behaviour of a model trained on the data. The provenance collection mechanism needs to be appropriate for each type of explanations. For instance there is system-level vs app-level provenance, with different issues associated with collection, different resulting abstraction levels, and consequently meeting different requirements.

Model interpretability and intelligibility raise several inter-related questions, which could be investigated in future research:

- How can the provenance associated with the pre-processing pipeline support trust and accountability in the model?
- Can we use explanations from intelligible models to annotate new data?
- Can we use provenance to assess the quality of explanations?
- Can privacy preserving ML be viewed as flip side of model interpretability?
- can provenance be used to track the sources of adversarial examples?

One of the issues with using provenance to provide explanations that extend to the entire pipeline (see above) is privacy: is there a need for provenance abstraction in this context and what abstraction mechanisms are appropriate?

3.4 Global challenges

The purpose of the *global challenges* breakout session was to discuss cross-cutting problems for research in provenance, security, and machine learning that may not fit neatly into the other breakout sessions. A specific challenge mentioned in the workshop plan was "What needs to be done for suitable datasets/challenges to be available and useful to the broader security/ML/DB communities?" This was based on one of the organisers' experience with the DARPA Transparent Computing program (TC), which produced some publicly available datasets which were (at the time of the workshop) not widely known and lacked documentation that would make it easy for people not involved in the DARPA TC program to reuse the data for their own research.

Accessible, reusable datasets The breakout group discussed some aspects of this data and challenges associated with making it accessible and reusable. The data was made available in a format called Common Data Model (CDM) which was defined by agreement among different projects in TC, including six different provenance recording systems and three projects focusing on data analysis. The latter three projects faced the challenge of *integrating* the results of the six provenance recording systems (covering Windows, Android, BSD, and Linux operating systems) and performing some kind of analysis on it to find attacks. This was difficult

in part because, even though CDM specified some aspects of how the provenance data was to be represented, there were still variations among different systems regarding both what information was recorded at all, and how "the same" information was recorded by different systems. As a concrete example, some systems would record only high-level information about processes, files, users and events, while others would also record detailed information about the run-time behaviour of processes, including details of file reads and writes down to the byte level. As another example, different systems recorded information about the command-line initialisation and parameters of processes in different ways. A still further complication was that some systems would create multiple records for the same entity, leading to the need for deduplication.

To ameliorate this situation, one of the TC analysis projects, ADAPT, developed a refined data model called ADM (ADAPT Data Model) and ingester that attempted to integrated the different CDM data sources and put them into a more uniform ADM format, applying deduplication along the way. The ingester would also load the data into a standard Neo4j graph database where it can be queried much more easily than in the native CDM form. With hindsight, further work to standardise not only how the data was stored but also standardise the meanings of different components in the data model might have saved effort here.

Ground truth Another obstacle to using the TC datasets for security research is the lack of "ground truth" annotations signposting the parts of the data corresponding to attacker activity. Such information was provided to TC participants but only in human-readable form, such as markdown documents showing interactive sessions during which attacks were performed, or spreadsheets listing the sources, targets and other attributes of objects in the data that were involved in an attack (such as key processes, files, or IP addresses). Some projects, such as ADAPT, manually went through some of the datasets to annotate processes, files, network connections, and events that were apparently part of the attack; however, this was done on an ad hoc basis and there was no independent verification or confirmation of the resulting ground truth data. This experience highlighted the importance of planning to create reusable datasets that include the kind of information that will make them useful for further research; since this was not done, another interesting challenge is whether it is possible to use the human-readable ground truth information that is available to automatically extract ground truth annotations after-the-fact, perhaps re-purposing techniques from natural language processing.

Additional issues Additional issues discussed included the problem of generalising from a laboratory environment (such as the TC datasets, which were recorded using random background activity generators instead of on real systems with real user activity) to real settings; the difficulty of dealing with different levels of granularity or abstraction in different datasets (e.g. the process level vs. byte level), and the related problem of aligning high-level (and human-accessible) attack information with low-level logs; the problem of data drift; and the general issue of privacy raised by any wholesale monitoring, particularly post-GDPR.

4 Open Problems, Challenges, and Next Steps

The workshop concluded with plenary sessions to exchange the main topics of discussion int he breakout groups, and to discuss next steps including the writing of this report. A summary of the open problems (highlighting common themes mentioned in more than one breakout group) is as follows:

- Data drift detection and mitigation techniques for provenance.
- Repository of provenance records with supporting information on reason for collection, underlying system and execution expectations and quality metrics for the community to utilise.
- Provenance protection, particularly analysis on encrypted information, homomorphic and polymorphic encryption of provenance.
- Analysis of cost of provenance capture in real systems, and methods to mitigate.
- Exploring applications of provenance to data quality assurance / cleaning for machine learning, e.g. to protect against or provide retroactive accountability in the presence of data poisoning or bias attacks.

5 Conclusions

Provenance-based security is an emerging topic which has attracted interest from both the academic security community and industry (with several projects in the DARPA Transparent Computing program consisting of joint academic and industry partnerships). However, as explored in the workshop, the initial wave of

enthusiasm for the possibilities of this approach needs to be channeled towards overcoming certain challenges that are likely to hinder progress or make it difficult to compare and evaluate solutions. In particular, making the existing publicly-available datasets resulting from efforts such as Transparent Computing useful to a wide audience would likely be very beneficial, since there is a community of researchers interested in using this data to test their ideas (including several workshop participants) who are unsure how to do so. Furthermore, the workshop highlighted several challenges arising from the interaction between provenance, security and machine learning that would benefit from further research, and although a one-day workshop was clearly not sufficient to fully scope and begin to solve these problems, the workshop participants, coming from different communities, had the opportunity to make new connections and begin to form collaborations in this area.