Gerência de Dados de Proveniência Distribuídos de Experimentos Científicos: um Mapeamento Sistemático*

Raama Costa Alves[1,2], Yuri Frota[1], Daniel de Oliveira[1]

[1]Instituto de Computação – Universidade Federal Fluminense (IC/UFF)

[2]Diretoria de Informática – Universidade do Estado do Rio de Janeiro (DINFO/UERJ)

raamacosta@id.uff.br, {yuri,danielcmo}@ic.uff.br

Abstract. *Scientific experiments based on simulations are composed of hundreds or thousands of program's calls that are benefited by high performance computing environments (HPC) to accelerate its execution. In such a context, managing provenance data in a distributed way becomes a huge challenge. Although there are many approaches to manage distributed data, there is not a de facto standard, becoming hard to correlate, to classify and to compare differences between these existing approaches. The main goal of this paper is to apply a systematic mapping about distributed provenance management and propose a taxonomy of this domain, classifying the existing approaches according to taxonomy classes.*

Resumo. *Experimentos científicos baseados em simulações são compostos de centenas ou milhares de invocações de programas que são beneficiados pelos ambientes de processamento de alto desempenho (PAD) para acelerar sua execução. Nesse contexto, a gerência dos dados de proveniência de forma distribuída torna-se um grande desafio. Apesar de existirem abordagens para gerência distribuída desses dados, não há um padrão de fato, tornando difícil correlacionar, classificar e comparar as várias abordagens existentes. O principal objetivo deste artigo é aplicar um mapeamento sistemático sobre a gerência de dados de proveniência distribuídos e propor uma taxonomia deste domínio, classificando as abordagens existentes de acordo com as classes da taxonomia.*

1. Introdução

A evolução da computação permitiu a popularização de um tipo de experimento baseado em simulações computacionais [de Oliveira et al. 2019], os chamados experimentos *in silico*, que geralmente estão diretamente associados à execução de uma série de artefatos de *software* (programas, serviços *Web*, bibliotecas, *etc*.) invocados em uma certa ordem. A execução desse tipo de experimento pode exigir muitos recursos computacionais e consumir/produzir um grande volume de dados. Assim, os ambientes de processamento de alto desempenho (PAD) passaram a ser utilizados. Nesse contexto, para garantir a capacidade de reprodução dos experimentos, os cientistas devem capturar e gerenciar os *Dados de Proveniência* [Freire et al. 2008].

*O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001, CNPq e FAPERJ. Raama Costa Alves agradece ao DINFO/UERJ pela liberação parcial de carga horária.

Os Dados de proveniência representam o histórico de um experimento e fornecem subsídio essencial para reproduzir e interpretar os resultados obtidos com a execução do experimento. À medida que os experimentos continuam a crescer em uma escala sem precedentes, as estratégias de gerência dos dados de proveniência devem ser revisitadas, de modo a aperfeiçoar seu uso e minimizar os custos. Em muitas das abordagens existentes, a execução dos experimentos ocorre de forma distribuída e os dados de proveniência coletados são transferidos de um recurso remoto para um repositório centralizado (*e.g.*, ProvStore[1] [Huynh and Moreau 2015]). Como esses dados também são importantes para a execução do experimento, torna-se fundamental que a captura, armazenamento e a consulta a tais dados seja eficiente. Uma das opções é armazenar e consultar esses dados de forma distribuída [Özsu and Valduriez 1991]. Apesar de uma série de *surveys* de proveniência já ter sido publicada [da Cruz et al. 2009, de Oliveira et al. 2018, Simmhan et al. 2005, Pimentel et al. 2019, Freire et al. 2008], nenhum deles trata a questão da gerência de dados de proveniência distribuídos.

O objetivo deste artigo é apresentar um mapeamento sistemático para melhorar o entendimento das abordagens existentes para gerência de dados de proveniência distribuídos. Além do mapeamento, propomos uma taxonomia mais ampla no que tange a gerência de proveniência distribuída. Esse artigo está organizado da seguinte forma: a Seção 2 descreve o protocolo do mapeamento sistemático. A Seção 3 apresenta a taxonomia proposta. Na Seção 4, mapeamos a taxonomia proposta nas abordagens selecionadas no mapeamento, e concluímos na Seção 5.

2. Protocolo do Mapeamento Sistemático

Nesta seção, apresentamos o protocolo utilizado para o mapeamento sistemático simplificado para identificar as abordagens de gerência de proveniência distribuída. De acordo com [Petersen et al. 2015], o principal objetivo de um mapeamento sistemático é produzir a **Visão Geral** de uma área de pesquisa. No contexto deste artigo, o mapeamento sistemático tem o objetivo de identificar abordagens que lidam com a **gerência de dados de proveniência distribuídos** e categorizá-las de acordo com seus objetivos e com a maneira como gerenciam os dados de proveniência. Assim, foram definidas uma questão principal de pesquisa e duas secundárias: (i) *QP1:* Quais são as abordagens existentes para "Gerência de Dados de Proveniência Distribuídos?", (ii) *QP1.1:* Onde tais abordagens foram publicadas?, e (iii) *QP1.2:* Quando essas abordagens foram publicadas?

Foram aplicados o ***forward*** e ***backward snowballing*** para descobrir abordagens relevantes [Wohlin 2014]. O método *snowballing* se inicia com um conjunto de artigos chamados de **sementes**. O *forward snowballing* acessa artigos que citam artigos do conjunto de sementes e os adiciona ao conjunto se eles cumprirem os critérios de inclusão predefinidos. O *backward snowballing* acessa os artigos citados pelos artigos do conjunto de sementes e os adiciona ao conjunto se eles cumprirem os mesmos critérios de inclusão. Esse processo é interrompido após um número finito de iterações. No mapeamento aqui apresentado foram utilizadas 3 iterações. No contexto deste artigo, definimos os critérios de inclusão como artigos revisados por pares em inglês com abordagens que gerenciam proveniência de forma distribuída. Excluímos abordagens focadas em gerência de proveniência de forma centralizada. Seguimos as diretrizes recomendadas

[1]https://openprovenance.org/store/

por [Wohlin 2014] para definir o conjunto de sementes do processo de *snowballing* (*i.e.,* usamos o *Google Scholar* para obter o conjunto de sementes em vez de escolher um repositório específico como IEEEXPlore ou *ACM Digital Library*). Pesquisamos "*Distributed Provenance Data*" no *Google Scholar* e selecionamos 11 artigos para o conjunto semente com base em nossos critérios de inclusão. Após as iterações, obtivemos 26 artigos. Esses trabalhos foram publicados em 6 periódicos distintos e 20 conferências/*workshops*. A Tabela 1 apresenta a citação do trabalho e sua classificação de acordo com a taxonomia proposta na Seção 3.

3. Taxonomia para a Proveniência em Ambientes Distribuídos

Uma taxonomia é uma estrutura que tem o objetivo de classificar conceitos e arranjá-los de uma forma hierárquica. Desse modo é possível proporcionar maior entendimento do domínio de conhecimento e comparar diferentes abordagens para gerência de dados de proveniência distribuídos [de Oliveira et al. 2010]. Nesta seção são apresentados os aspectos utilizados para elaborar as classes de taxonomia proposta. A classificação proposta é organizada de acordo com as características das abordagens em termos de **Captura**, **Tipo**, **Modelo de Fragmentação**, **Consulta** e **Armazenamento**, conforme apresentado na Figura 1.

A **Captura** dos dados de proveniência adotada pode ser dividida em três subcategorias: **Manual (MAN)**, **Automática (AUT)** ou **Semiautomática (SAU)**. A captura **Manual** refere-se à coleta de dados realizada por um determinado usuário durante a execução de um programa, *workflow* ou *script*. Esse tipo de captura é laboriosa e propensa a erro. Já a captura **Automática** está relacionada a capacidade de se obter dados de proveniência a partir do monitoramento da execução de um programa, *workflow* ou *script*. Em geral, a captura automática de dados de proveniência tende a coletar um grande volume de dados. Na captura **Semiautomática**, o usuário define quais dados deverão ser capturados de forma automática, *i.e.*, existe uma intervenção do usuário antes da execução do programa ou *workflow* para definir quais são os dados de interesse. Toda captura possui um **Nível**, que define a granularidade da captura. Podem ser capturados dados somente da aplicação executada, do sistema de *workflow* utilizado ou sobre tudo o que acontece no sistema operacional.

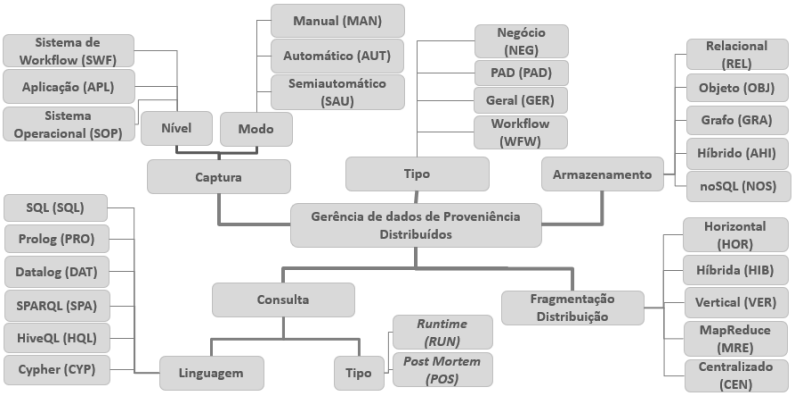

Figura 1. Taxonomia de Gerência de Dados de Proveniência Distribuídos.

A classe **Tipo** se destina a classificar os tipos de sistemas que gerenciam os dados de proveniência distribuídos. As principais categorias adotadas são: **Workflow (SWF)**

(que se refere aos Sistemas de *Workflow* como o SciCumulus [de Oliveira et al. 2012]), **PAD (PAD)** (que se referem a ferramentas de coleta de proveniência em ambientes PAD, como o LPS [Dai et al. 2017]), de **Negócio (NEG)** (como o proposto em [Dalpra et al. 2015]) e **Geral (GER)** que está relacionada aos sistemas de gerência de proveniência de uso geral como a Matrioshka [da Cruz et al. 2011]).

Considerando o grande volume de dados de proveniência armazenados, há uma preocupação quanto ao processamento de consultas sobre esses dados. Uma das abordagens mais utilizadas para resolver o problema de processamento sobre um grande volume de dados é a **Fragmentação** [Özsu and Valduriez 1991]. A classe **Fragmentação** se refere às alternativas de fragmentação de dados de proveniência existentes. Se considerarmos dados representados em relações (tabelas), os mesmos comumente estão representados sem fragmentação (**centralizado (CEN)**). Entretanto, temos as alternativas clássicas de fragmentação discutidas por [Özsu and Valduriez 1991] que podem ser aplicadas: **Horizontal (HOR)**, **Vertical (VER)** e a **Híbrida (HIB)**, que consiste na aplicação da fragmentação o horizontal sobre a vertical ou vice-versa. Existe ainda a fragmentação dos dados aplicada em arcabouços **MapReduce (MRE)**, em que o balanceamento de carga e a localidade dos dados influenciam diretamente na fragmentação e distribuição dos mesmos [Oliveira et al. 2015].

A classe **Consulta** refere-se às características utilizadas na consulta aos dados de proveniência já capturados. Essa classe pode ser especializada em duas sub-categorias: **Tipo** e **Linguagem**. A subclasse **Tipo**, destina-se a classificar o modo como as consultas de proveniência podem ser submetidas e divide-se em duas classes: ***Post-Mortem (POS)*** e ***Runtime (RUN)***. A subclasse ***Post-Mortem*** refere-se às consultas que são processadas após o experimento ter sido executado. Já a subclasse ***Runtime***, concerne às consultas que podem ser realizadas durante a execução do experimento. A subclasse **Linguagem** agrupa as abordagens de acordo com a linguagem de consulta utilizada. Essa categoria possui as seguintes subclasses: **SQL (SQL)**, **Prolog (PRO)**, **DATALOG (DAT)**, **SPARQL (SPA)**, **HiveQL (HQL)** e **Cypher (CYP)**. É importante ressaltar que a classe *Linguagem* pode ser estendida para novas linguagens.

A classe **Armazenamento** destina-se a agrupar as abordagens de gerência de dados de proveniência distribuídos de acordo com o modo de armazenamento. A subclasse **Relacional (REL)** agrupa as abordagens que utilizam SGBDs relacionais. **Objeto (OBJ)** reúne as abordagens em que a unidade básica de armazenamento é um objeto. A subclasse **Grafo (GRA)** incorpora as abordagens onde os dados de proveniência são representados como grafos. A subclasse **NoSQL (NOS)** engloba as abordagens que utilizam modelos de dados colunares, orientados a documento e chave-valor. E finalmente a subclasse **Híbrido (AHI)** abrange as abordagens que possuem a flexibilidade de utilizar mais de um tipo de representação.

4. Classificação das Abordagens usando a Taxonomia Proposta

Nesta seção, apresentamos de forma resumida as principais abordagens que tratam da gerência de dados de proveniência distribuídos, de acordo com a taxonomia proposta na Seção 3. A Tabela 1 apresenta as abordagens selecionadas após o *snowballing*. Podemos observar que nenhuma das abordagens classificadas fornece todas as funcionalidades e características apresentadas na taxonomia proposta. Assim, os usuários devem anali-

sar suas necessidades e verificar na classificação qual a abordagem é a mais adequada para seu cenário. É interessante observar que a maioria das abordagens foca na captura *automática* da proveniência, o que eventualmente faz com que dados fora do interesse do usuário sejam armazenados (que provavelmente não serão consultados), além de necessitar de uma maior capacidade de armazenamento. Avaliar o *trade-off* entre capturar muitos dados e o custo de armazenamento é um desafio, conforme discutido por [Suriarachchi and Plale 2016].

Além disso, grande parte das abordagens se baseia numa abordagem centralizada dos dados que pode limitar a otimização das consultas. Além disso, não há evidências sobre o uso de replicação dos dados de proveniência como alternativa para redução no tempo de processamento de consultas. É interessante observar também que grande parte das abordagens utiliza *scripts* próprios para consultar os dados distribuídos. O uso de linguagens como Prolog pode ser interessante no cenário distribuído, pois permite inferências que seriam complicadas de serem implementadas em outras linguagens. Em relação ao armazenamento, grande parte das soluções utiliza bancos de dados relacionais. Com o advento de novas tecnologias como os SGBDs *Polystore* [Duggan et al. 2015], novas possibilidades de armazenamento e distribuição dos dados entrarão em voga.

Tabela 1. Classificação das abordagens de acordo com a taxonomia proposta.

Modelos	Categorias																										
	Captura						Tipo				Fragmentação/Distribuição					Consulta								Armazenamento				
	Modo			Nível												Tipo		Linguagem										
	MAN	AUT	SAU	SOP	APL	SWF	WFW	PAD	GER	NEG	HOR	VER	CEN	MRE	HIB	RUN	POS	SQL	PRO	DAT	SPA	HQL	CYP	REL	OBJ	GRA	AHI	NOS
[Heinis and Alonso 2008]			x			x	x						x				x	x						x				
[Abraham et al. 2010]			x			x	x							x			x				x							x
[Liu et al. 2010]		x			x				x		x					x		x						x				
[Park et al. 2011]		x				x	x							x			x					x						x
[Zhou et al. 2011]		x		x					x			x				x			x					x				
[Ikeda et al. 2011]	x					x	x							x		x						x							x
[da Cruz et al. 2012]		x			x					x		x					x	x											x
[Gehani and Tariq 2012]			x	x					x				x			x		x					x					x	
[Malik et al. 2013]	x			x					x				x			x		x						x					
[Zhao et al. 2013]		x		x				x						x		x						x							x
[Dai et al. 2014]			x		x			x					x			x									x				
[Hammad and Wu 2014]	x				x					x			x			x		x						x					
[Bates et al. 2015]		x		x					x				x			x		x						x					
[Rundle 2016]		x			x				x				x			x		x						x					
[Li et al. 2016]		x		x			x							x			x					x							x
[Ma et al. 2016]	x				x				x				x			x					x			x					
[Xie et al. 2016]			x	x					x					x		x						x							x
[Pineda-Morales et al. 2016]		x				x	x								x	x		x									x		
[Dai et al. 2017]			x		x			x					x			x							x			x			
[Aniello et al. 2017]		x			x				x					x		x						x							x
[Zhang et al. 2017]		x		x					x				x			x				x				x					
[Niu et al. 2017]		x			x				x				x			x		x						x					
[Arab et al. 2018]			x		x				x			x				x		x						x					
[Kelbert and Pretschner 2018]		x			x				x						x	x								x					
[Zawoad et al. 2018]			x		x				x		x					x		x						x					

5. Conclusão

Neste artigo, apresentamos um mapeamento sistemático e uma taxonomia para gerência de dados de proveniência distribuídos de experimentos *in silico*. Acreditamos que será útil para a comunidade científica avaliar e comparar diferentes abordagens de gerência de proveniência distribuída. Ao classificar tais abordagens usando a taxonomia proposta, usuários (em geral cientistas) podem avaliar quais atendem às suas necessidades para executar experimentos científicos em nuvens, *clusters*, etc com garantia de captura e consulta aos dados de proveniência.

Este artigo destaca que, apesar do grande interesse sobre o tema, ele ainda é um campo aberto. Novas soluções para gerência de dados de proveniência distribuídos se encontram disponíveis e muitas outras estão sendo desenvolvidas. É fundamental que os usuários possam escolher a melhor abordagem para seus experimentos. O uso da taxonomia e seu vocabulário comum pode facilitar os usuários a encontrar características comuns das abordagens existentes e ajudá-los a escolher a mais adequada.

Referências

Abraham, J., Brazier, P., Chebotko, A., Navarro, J., and Piazza, A. (2010). Distributed storage and querying techniques for a semantic web of scientific workflow provenance. In *IEEE Services*, pages 178–185.

Aniello, L., Baldoni, R., Gaetani, E., Lombardi, F., Margheri, A., and Sassone, V. (2017). A prototype evaluation of a tamper-resistant high performance blockchain-based transaction log for a distributed database. In *2017 EDCC*, pages 151–154.

Arab, B. S., Gawlick, D., Krishnaswamy, V., Radhakrishnan, V., and Glavic, B. (2018). Using reenactment to retroactively capture provenance for transactions. *IEEE Trans. on Know. and Data Eng.*, 30(3):599–612.

Bates, A., Tian, D. J., Butler, K. R., and Moyer, T. (2015). Trustworthy whole-system provenance for the linux kernel. In *USENIX Security*, pages 319–334, Washington, D.C. USENIX Association.

da Cruz, S. M. S., Campos, M. L. M., and Mattoso, M. (2009). Towards a taxonomy of provenance in scientific workflow management systems. In *2009 IEEE Services, Los Angeles, CA, USA*, pages 259–266. IEEE Computer Society.

da Cruz, S. M. S., Manhães, L. M. B., Costa, M., and Zavaleta, J. (2012). Analysing e-business applications with business provenance. In *DCNET/ICE-B/OPTICS*.

da Cruz, S. M. S., Silva, C. E. P., de Oliveira, D., Campos, M. L. M., and Mattoso, M. (2011). Capturing distributed provenance metadata from cloud-based scientific workflows. *JIDM*, 2(1):43–50.

Dai, D., Chen, Y., Carns, P., Jenkins, J., and Ross, R. (2017). Lightweight provenance service for high-performance computing. In *2017 26th International Conference on Parallel Architectures and Compilation Techniques (PACT)*, pages 117–129.

Dalpra, H. L. O., Costa, G. C. B., Sirqueira, T. F. M., Braga, R. M. M., Campos, F., Werner, C. M. L., and David, J. M. N. (2015). Using ontology and data provenance to

improve software processes. In *ONTOBRAS), São Paulo, Brazil*, volume 1442. CEUR-WS.org.

de Oliveira, D., Baião, F. A., and Mattoso, M. (2010). *Towards a Taxonomy for Cloud Computing from an e-Science Perspective*, pages 47–62. Springer London, London.

de Oliveira, D., Ocaña, K. A. C. S., Baião, F. A., and Mattoso, M. (2012). A provenance-based adaptive scheduling heuristic for parallel scientific workflows in clouds. *J. Grid Comput.*, 10(3):521–552.

de Oliveira, D. C. M., Liu, J., and Pacitti, E. (2019). *Data-Intensive Workflow Management: For Clouds and Data-Intensive and Scalable Computing Environments*. Synthesis Lectures on Data Management. Morgan & Claypool Publishers.

de Oliveira, W. M., de Oliveira, D., and Braganholo, V. (2018). Provenance analytics for workflow-based computational experiments: A survey. *ACM Comput. Surv.*, 51(3):53:1–53:25.

Duggan, J., Elmore, A. J., Stonebraker, M., Balazinska, M., Howe, B., Kepner, J., Madden, S., Maier, D., Mattson, T., and Zdonik, S. B. (2015). The bigdawg polystore system. *SIGMOD Rec.*, 44(2):11–16.

Freire, J., Koop, D., Santos, E., and Silva, C. T. (2008). Provenance for Computational Tasks: A Survey. *Computing in Science & Engineering*, 10(3):11–21.

Gehani, A. and Tariq, D. (2012). Spade: Support for provenance auditing in distributed environments. In *Middleware 2012*, pages 101–120. Springer Berlin Heidelberg.

Hammad, R. and Wu, C. (2014). Provenance as a service: A data-centric approach for real-time monitoring. In *2014 IEEE International Congress on Big Data*, pages 258–265.

Heinis, T. and Alonso, G. (2008). Efficient lineage tracking for scientific workflows. *Proceedings of the ACM SIGMOD International Conference on Management of Data*.

Huynh, T. D. and Moreau, L. (2015). Provstore: A public provenance repository. In Ludäscher, B. and Plale, B., editors, *Provenance and Annotation of Data and Processes*, pages 275–277, Cham. Springer International Publishing.

Ikeda, R., Park, H., and Widom, J. (2011). Provenance for generalized map and reduce workflows. In *CIDR, Asilomar, CA, USA*, pages 273–283. www.cidrdb.org.

Kelbert, F. and Pretschner, A. (2018). Data usage control for distributed systems. *ACM Transactions on Privacy and Security (TOPS)*, 21.

Li, X., Xu, X., and Malik, T. (2016). Interactive provenance summaries for reproducible science. In *2016 IEEE 12th International Conference on e-Science (e-Science)*, pages 355–360.

Liu, M., Taylor, N., Zhou, W., Ives, Z., and Loo, B. (2010). Maintaining recursive views of regions and connectivity in networks. *IEEE Trans. on Knowl. and Data Eng.*, 22:1126–1141.

Ma, T., Wang, H., Cao, J., Yong, J., and Zhao, Y. (2016). Access control management with provenance in healthcare environments. In *IEEE (CSCWD)*, pages 545–550.

Malik, T., Gehani, A., Tariq, D., Zaffar, Fareed", e. Q., Bai, Q., Giugni, S., Williamson, D., and Taylor, J. (2013). *Sketching Distributed Data Provenance*, pages 85–107. Springer.

Niu, X., Kapoor, R., Glavic, B., Gawlick, D., Liu, Z. H., Krishnaswamy, V., and Radhakrishnan, V. (2017). Provenance-aware query optimization. In *IEEE (ICDE)*, pages 473–484.

Oliveira, D., Boeres, C., Fausti, A., and Porto, F. (2015). Avaliação da localidade de dados intermediários na execução paralela de workflows big data. In *Brazilian Simposium on Databases*.

Özsu, M. T. and Valduriez, P. (1991). *Principles of Distributed Database Systems*. Springer.

Petersen, K., Vakkalanka, S., and Kuzniarz, L. (2015). Guidelines for conducting systematic mapping studies in software engineering: An update. *Information & Software Technology*, 64:1–18.

Pimentel, J. F., Freire, J., Murta, L., and Braganholo, V. (2019). A survey on collecting, managing, and analyzing provenance from scripts. *ACM Comput. Surv.*, 52(3):47:1–47:38.

Pineda-Morales, L., Liu, J., Costan, A., Pacitti, E., Antoniu, G., Valduriez, P., and Mattoso, M. (2016). Managing hot metadata for scientific workflows on multisite clouds. In *2016 IEEE International Conference on Big Data (Big Data)*, pages 390–397.

Simmhan, Y., Plale, B., and Gannon, D. (2005). A survey of data provenance in e-science. *SIGMOD Rec.*, 34(3):31–36.

Suriarachchi, I. and Plale, B. (2016). Crossing analytics systems: A case for integrated provenance in data lakes. In *IEEE e-Science, Baltimore, USA*, pages 349–354. IEEE Computer Society.

Wohlin, C. (2014). Guidelines for snowballing in systematic literature studies and a replication in software engineering. EASE '14. ACM.

Xie, Y., Feng, D., Tan, Z., and Zhou, J. (2016). Unifying intrusion detection and forensic analysis via provenance awareness. *Future Generation Computer Systems*, 61:26 – 36.

Zawoad, S., Hasan, R., and Islam, K. (2018). Secprov: Trustworthy and efficient provenance management in the cloud. In *IEEE INFOCOM*, pages 1241–1249.

Zhang, Y., O'Neill, A., Sherr, M., and Zhou, W. (2017). Privacy-preserving network provenance. *Proc. VLDB Endow.*, 10(11):1550–1561.

Zhao, D., Shou, C., Maliky, T., and Raicu, I. (2013). Distributed data provenance for large-scale data-intensive computing. In *IEEE (CLUSTER)*, pages 1–8.

Zhou, W., Ding, L., Haeberlen, A., Ives, Z., and Loo, B. (2011). Tap: Time-aware provenance for distributed systems.