

A Graph Testing Framework for Provenance Network Analytics

Bernard Roper, Adriane Chapman, David Martin, Jeremy Morley

University of Southampton
b.a.roper@soton.ac.uk

Abstract. Provenance Network Analytics is a method of analyzing provenance that assesses a collection of provenance graphs by training a machine learning algorithm to make predictions about the characteristics of data artifacts based on their provenance graph metrics. The shape of a provenance graph can vary according to the modelling approach chosen by data analysts, and this is likely to affect the accuracy of machine learning algorithms, so we propose a framework for capturing provenance using semantic web technologies to allow use of multiple provenance models at runtime in order to test their effects.

Keywords: Graph, Network, Analytics.

1 Introduction

Provenance data describes the events, agents, resources and relationships that have led to the creation of a piece of data or thing and as such is naturally expressed as a graph. Provenance is used in a range of application domains, e.g. geospatial [1]–[3] and scientific experimentation [4]–[6]. Some of these applications generate large and complex graphs resulting in a volume of data that is beyond the scope of inspection and query. While some strategies exist [7]–[9] to simplify their representation for human usability, these techniques are typically made for an individual inspecting a single provenance graph to judge fitness for use of a specific artefact.

Provenance Network Analytics (PNA) is an approach proposed by Huynh et al [10], [11], which instead attempts to help users assess fitness for use for an artefact by assessing a collection of provenance graphs. In their work, they use a set of provenance specific network metrics [12] adapted from network theory [13]. These are used to summarize a dependency subgraph graph as a feature vector to train machine learning algorithms to predict characteristics of the data artefact for which the provenance has been expressed.

This technique is used in [10] to assess the quality of a map feature from CollabMap, a crowdsourced mapping initiative used for disaster relief planning. Using feature vectors from these provenance graphs, the authors trained a machine learning algorithm to predict user trust ratings with 95% accuracy. They have also tested this in other applications; identifying message types in a disaster response simulation game and identify-

ing owners of PROV-N documents, achieving a high degree of classification accuracy across these domains [11].

However, in [11] the authors note that the model chosen for the provenance could impact the quality of the ultimate machine learning model produced. We replicated this in [14], using PROV graphs generated from Open Street Map (OSM) history data, obtaining only 54% accuracy when attempting to predict the incidence of fix-me tags left by users to indicate issues with the data describing a map feature.

Fig. 1. Comparison of techniques between various approaches for machine learning over provenance graphs, and the ultimate accuracy, from [14]

	OSM [14]	Huynh et al [10]
Analysis goal	Predict prevalence of fix-me tags	Predict user trust ratings
Graph structure	6 relationships, 3 vertex types	3 relationships, 4 vertex types
Feature Vector	MFD, #vertices, #edges, diameter	MFD, #vertices, #edges, diameter
ML Technique used	decision tree classifier	decision tree classifier
Target Attributes	fix-me tag	Trusted/uncertain rating
Target flags ratio	50:50	50:50
Most Relevant metrics	diameter	#vertices, #edges
Data sets	Two geographic sets containing 30265 and 97393 features, adjusted to 298 and 1604	Three sets divided by data type: 5175 buildings, 4911 evacuation routes and 3043 route sets
Accuracy of results	54%	95%

The inability to replicate the classifier accuracy of [10] in [14] could have any number of reasons. While it could be argued that provenance is not useful for making predictions about the characteristics of data, the results obtained from the work by Huynh et al [10], [11] are sufficient to discount this. Alternatively, the specific characteristic (i.e. the fix-me tag) cannot be predicted by the provenance analytics method. While we cannot discount this entirely, it seems unlikely, as this characteristic is analogous to a user trust rating. Another possibility is that there are errors in the way the machine learning algorithms were used. This is of course possible and will be investigated further during this project. However, there are two important factors which bear deeper investigation:

- The network metrics chosen. It is apparent from the previous Provenance Network Analytics work [11] that these metrics have an impact on the machine learning accuracy and that this varies depending on the type of feature from which the provenance is derived.
- The shape of the extracted subgraphs, defined by the way the provenance is modelled and expressed by analysts. Huynh et al [11] found that the results from one of the applications they studied, although still useful, were significantly poorer than the other two applications. From visual inspection they noted that the shape of these graphs was quite distinctive and so parameterized their capture method to vary the shape of the graph. Doing so effected the classification accuracy.

The modelling of provenance is something of an art form, and characteristics of a provenance graph can vary depending on the application, use-case, and analysis requirements. E.g. nodes can be abstracted for reasons of confidentiality and data protection

[15], or granularity can be varied to manage computing resources [16]. These approaches to the expression of a graph decide its topological characteristics and are likely to influence the effectiveness of PNA. OSM history data presents a variety of ways in which provenance could be extracted to create provenance graphs whose form differs depending on the modelling approach chosen.

For example, **Table 1** shows two structurally different graphs of provenance for the same OSM map artefact. The accompanying table shows some graph theoretic measurements and values for MFD (maximum finite distance), a provenance specific measurement used in [10], [11]. The graphs are obviously different in appearance and produce a different set of measurement values.

Table 1. Two provenance graphs of an OSM map feature

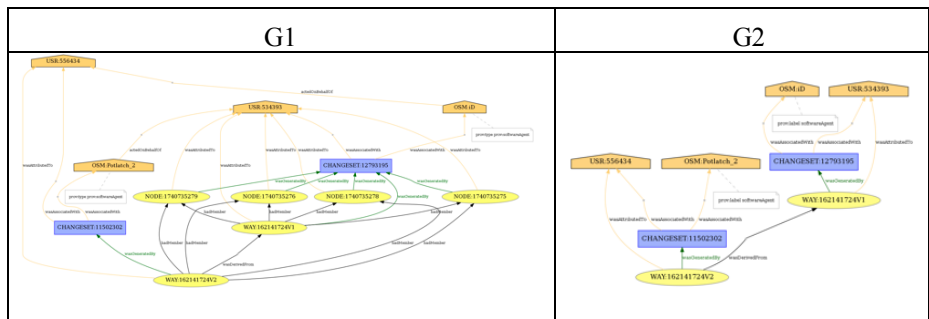


Table 2. Metrics from the graphs in **Table 1**

metrics	G1	G2	MFD	G1	G2
Nodes	12	8	entity-entity	1	1
Edges	27	9	entity-activity	2	2
Components	1	1	entity-agent	3	3
Diameter	3	5	activity-entity	0	0
			activity-activity	0	0
			activity-agent	2	1
			agent-entity	0	0
			agent-activity	0	8
			agent-agent	1	9

It is likely that different approaches to provenance modelling will result in variations in the accuracy of machine learning classifiers. To identify any effect, a framework for testing the PNA method using graphs built using a range of modelling approaches is needed. Our contributions in this work are the following:

- We create a provenance extraction framework that allows the shape of a provenance graph to be changed at runtime.
- We showcase the use of this framework on Open Street Maps, and show how an OSM XML history file can be parsed into a history representation that allows any number and shape of provenance graphs to be generated programmatically

2 A multi-model graph analysis framework.

The system proposed here is related to methods of ‘scraping’ provenance from log files generated by an application as part of its instrumentation, such as [17], [18]. The diagram in **Fig. 2** shows our process, which uses **OSM XML History Data**, which is in the same format as the OSM dataset but contains the state of each map artefact at any stage in its history, including timestamp, software used, external dataset derivations and an ID of the creator agent. Rather than scraping a specific expression of provenance from the data by parsing, **XSLT** is used to transform it into an **RDF Graph**. This is encoded using OWL and the **PROV-O** ontology, which are used to enrich the data set by entailing more triples to generate a comprehensive and universal provenance graph from which different PROV-DM representations can be extracted.

The resulting RDF is added to a **Triple Store** created using the Apache JENA Java libraries. The PROV graphs for map features are obtained using **SPARQL** queries which return **RDF Graphs** as Apache JENA RDF model objects, which can be converted to network graph representations and feature vectors using the **JENA-JUNG Graph Analysis Library**. The **feature vectors** will be used to train a **Machine Learning** classifier.

We capture data with the PROV-DM elements that allow data enrichment by inference using the PROV-O ontology. **Fig. 3** shows the attribution and derivation relationships of an OSM map artefact. The relationships in bold show provenance that has been explicitly declared in the RDF produced by the XSLT transformation. The other relationships have been inferred by a reasoner using PROV-O.

Fig. 2. The framework process

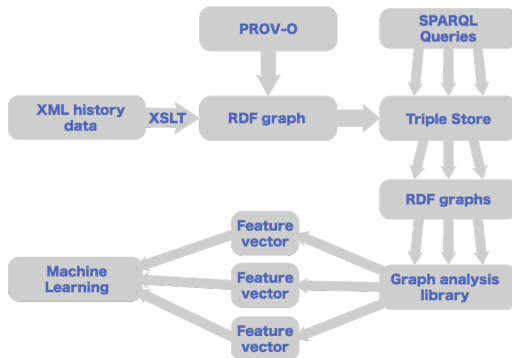
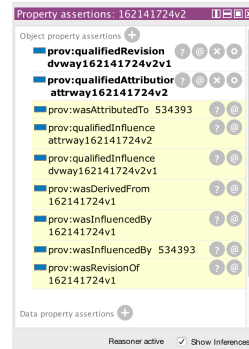


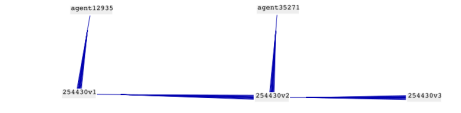

Fig. 3. Inferred triples in Protégé



We also use a qualified relations design pattern [19] for the provenance relationships, so that each edge is reified into an individual, linked with a qualified relation edge so that more triples can be inferred, creating the simpler wasAttributedTo and wasDerivedFrom relationship.

Once this process is complete, PROV graphs are then extracted using different SPARQL queries to the same set of PROV data as seen in **Fig. 4**

Fig. 4. Two SPARQL Queries with their resultant graphs

<pre> CONSTRUCT { ?version prov:wasDerivedFrom ?entity. ?entity prov:wasAttributedTo ?agent. } WHERE { ?entity provanalytics:versionOf osm:254430. ?version prov:wasDerivedFrom ?entity. ?entity prov:wasAttributedTo ?agent } </pre>	
<pre> CONSTRUCT { ?entity prov:qualifiedAttribution ?attr. ?attr prov:entity ?entity. ?attr prov:agent ?agent. ?version prov:qualifiedRevision ?rev. ?rev prov:entity ?entity. } WHERE { ?entity provanalytics:versionOf osm:254430. ?entity prov:qualifiedAttribution ?attr. ?attr prov:agent ?agent. ?version prov:qualifiedRevision ?rev. ?rev prov:entity ?entity. } </pre>	

This framework allows specification of PROV models using SPARQL. The example above shows two graphs produced by different SPARQL queries run over RDF data extracted from an OSM history file with axioms generated by a reasoner in Protégé [20]. Using feature vectors from results like these we train a ScikitLearn Decision Tree Classifier [21]. This provides a human readable output with information about the significance of the various graph metrics in the classification process, which can be used to help inform the design of other PROV models which can be extracted from the data using SPARQL.

3 Future work

Once this framework is completed we will create another XSLT module for use with Ordnance Survey history data and examine other target quality characteristics. We will also explore other machine learning techniques to see if classification accuracies can be improved and if so, whether the decision tree classifier can still be used alongside other algorithms to provide information about the role of the various metrics and different graph morphologies and what insights this might give us into the social worlds and processes of data creation.

Because we are using RDF in a triple store we will be able to update our Provenance dataset as the OSM history is updated. This dataset could be used to produce a provenance powered spatial representation of predicted data quality that updates over time.

References

1. P. Yue, M. Zhang, X. Guo, and Z. Tan, ‘Granularity of geospatial data provenance’, in *2014 IEEE Geoscience and Remote Sensing Symposium*, 2014, pp. 4492–4495.
2. J. Maso, B. Pross, Y. Gil, and G. Closa, Eds., ‘Testbed 10 Provenence Engineering Report’. OGC, 14-Jul-2014.

3. P. Yue, J. Gong, L. Di, L. He, and Y. Wei, 'Semantic provenance registration and discovery using geospatial catalogue service', in *Proceedings 2nd International Workshop on the Role of Semantic Web in Provenance Management, Shanghai, China, 2010*, pp. 23–28.
4. W. Oliveira, L. M. Ambrósio, R. Braga, V. Ströele, J. M. David, and F. Campos, 'A Framework for Provenance Analysis and Visualization', *Procedia Computer Science*, vol. 108, pp. 1592–1601, 2017.
5. U. Acar, P. Buneman, and J. Cheney, 'A graph model of data and workflow provenance', p. 10.
6. S. Miles, P. Groth, M. Branco, and L. Moreau, 'The requirements of recording and using provenance in e-Science experiments', p. 15.
7. S. Davidson *et al.*, 'Provenance in Scientific Workflow Systems', p. 7, 2007.
8. L. Moreau, 'Aggregation by Provenance Types: A Technique for Summarising Provenance Graphs', *Electronic Proceedings in Theoretical Computer Science*, vol. 181, pp. 129–144, Apr. 2015.
9. P. Macko and M. Seltzer, 'Provenance Map Orbiter: Interactive Exploration of Large Provenance Graphs', p. 6.
10. T. D. Huynh, M. Ebden, M. Venanzi, S. D. Ramchurn, S. Roberts, and L. Moreau, 'Interpretation of crowdsourced activities using provenance network analysis', in *First AAAI Conference on Human Computation and Crowdsourcing*, 2013.
11. T. D. Huynh, M. Ebden, J. Fischer, S. Roberts, and L. Moreau, 'Provenance Network Analytics: An approach to data analytics using data provenance', *Data Mining and Knowledge Discovery*, Feb. 2018.
12. M. Ebden, T. Huynh, L. Moreau, S. Ramchurn, and S. Roberts, 'Network analysis on provenance graphs from a crowdsourcing application', *Provenance and Annotation of Data and Processes*, pp. 168–182, 2012.
13. M. E. J. Newman, *Networks: an introduction*. Oxford ; New York: Oxford University Press, 2010.
14. B. Roper, 'Investigating the Role of Data Provenance in Assessing Variations in the Quality of Open Street Map Data', MSc, University of Southampton, 2017.
15. P. Missier, J. Bryans, C. Gamble, V. Curcin, and R. Danger, 'ProvAbs: Model, Policy, and Tooling for Abstracting PROV Graphs', in *Provenance and Annotation of Data and Processes*, 2014, pp. 3–15.
16. T. Pasquier *et al.*, 'Practical Whole-System Provenance Capture', *arXiv:1711.05296 [cs]*, pp. 405–418, 2017.
17. T. De Nies *et al.*, 'Git2PROV: Exposing Version Control System Content as W3C PROV', in *Poster and Demo Proceedings of the 12th International Semantic Web Conference*, 2013, vol. 1035, pp. 125–128.
18. D. Ghoshal and B. Plale, 'Provenance from log files: a BigData problem', 2013, p. 290.
19. L. Moreau and P. Groth, *Provenance: An Introduction to Prov*. Morgan & Claypool Publishers, 2013.
20. 'protégé'. [Online]. Available: <https://protege.stanford.edu/>. [Accessed: 07-Apr-2018].
21. F. Pedregosa *et al.*, 'Scikit-learn: Machine learning in Python', *Journal of Machine Learning Research*, vol. 12, no. Oct, pp. 2825–2830, 2011.