Structural Analysis of Whole-System Provenance Graphs

Jyothish Soman, Thomas Bytheway, Lucian Carata, Nikilesh D. Balakrishnan,
Ripduman Sohan, and Robert N. M. Watson

Computer Laboratory, University of Cambridge, UK
{firstname.lastname}@cl.cam.ac.uk

Abstract. System based provenance generates traces captured from various systems, a representation method for inferring these traces is a graph. These graphs are not well understood, and current work focuses on their extraction and processing, without a thorough characterization being in place. This paper studies the topology of such graphs. We analyze multiple Whole-system-Provenance graphs and present that they have hubs-and-authorities model of graphs as well as a power law distribution. Our observations allow for a novel understanding of the structure of Whole-system-Provenance graphs.

1 Introduction

Provenance has become a topic of relevance lately with the advent of multiple systems which augment the existing ones using provenance [1, 6]. There are multiple WSP capturing systems which can generate detailed data regarding the interactions happening at the machine level, these are naturally representable as graphs. For such provenance systems, the structure and evolution of the graph are both relevant in the design and optimisation of the storage, analysis and synthetic data generators. This has equivalence in other domains such as web-graphs, social networks, road-networks etc. A large volume of work present in literature support this [7, 2]. For example, the work in [7] presents the analysis of the spread of disease in a human-interaction network. This draws parallels with security related WSP research.

With this aim, we present an analysis of provenance graphs. Process traces are taken from a set of running machines, and the structure of the graphs so generated are studied. The results are used to present that the graphs are similar in structure to a well studied class of graphs namely, Power-law graphs. Additionally, we are able to show that they are similar to a Hubs-and-Authorities model of graphs [3] In the rest of the paper, nodes, edges and degree are used to only discuss the graph properties. Degree represents the total number of edges, both incoming and outgoing from a given node.

OPUS and PVM: OPUS [1] is a user space provenance system designed for tracking provenance on a system. For this work, it was used to provide a graph representation of the various interactions on a tracked machine, both implicit and explicit in the form of a graph. In OPUS, a process is considered the active agent, and the changes made by it are tracked. This includes interactions of the process with files and the I/O systems and its communication with other processes using pipes and sockets. OPUS uses Provenance-Versioning-Model (PVM) to handle state changes in entities, this is done with the intention of reducing the number of false dependencies, and de-densifying the resulting graph.

2 Graph types in WSP

In the current context, two set of graph models are relevant, namely power-law graphs and the Hubs-and-Authorities graph model. In this section, we would discuss them in further detail.

Power-law graphs: Power-law graphs have a degree distribution of the form $n(k) = Ax^k$, where $n(k)$ is the number of nodes with a degree k, and A and x are constants. Such graphs also tend to have a tail in the degree distribution. Examples include human-interaction networks such as social-networks, IMDB actors graphs, web-graphs, email graphs, citation graphs and recommendation-networks, to technological graphs such as Autonomous system graphs, web-graphs etc. [4]. Such graphs are also common in time-evolving graphs where interactions (edges) and elements (nodes) are added to an already existing graph.

In machines, such a process is possible as a limited number of processes and files have a higher probability of addition of incoming edges to them. Longer standing processes would accumulate both incoming and outgoing edges. Additionally, files and libraries would have multiple processes linking to them over time. Hence, the number of edges they accumulate would increase, which is in line with the preferential attachment theory.

Hubs and Authorities: In Hubs-and-Authorities (HaA) model of Kleinberg [3], each node can be either a hub or an authority. Hubs are nodes which connect high relevance nodes, and authorities are representative of the immediate neighbourhood. An authority on the other hand would be specialised nodes, and would only have information regarding a specific issue. HaA model also allows for a class of nodes which cannot be classified as either and are disregarded. The equivalent in WSP would be processes being authorities, and files, sockets and other mediums by which they share state being hubs.

The power-law graphs and HaA model together can be used to describe the graphs present in WSP . In Section 4, we would present graphs and their structural and property similarities to these two models.

3 Setup

For the generation of the test graphs, single machine-level traces using DTrace running on FREEBSD were taken from multiple machines running varied applications. A total of 9 traces are captured on active machines running multiple processes. OPUS was used to summarise the graphs and to merge nodes from the traces to form a more cohesive view of the traces.

4 Results

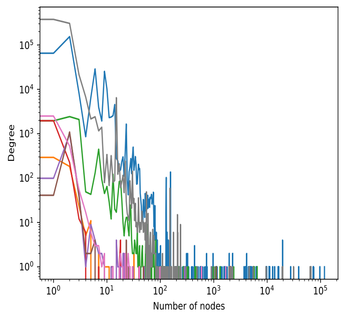

Fig. 1: Degree distribution.

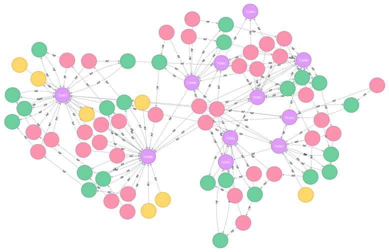

Fig. 2: Graph showing all interactions across 100 connected elements.

The analysis in this section deals with the graph structure, in terms of the degree distribution and general structure of the graph.

Figure 2 shows a small graph from which dangling nodes and nodes with no outgoing edges are removed. In this graph, the community property of the graph is better visible with two clear groups visible. This graph is similar in structure to the dolphin social network [5]. In the dolphin social network, there are two large groups, one of which is densely connected, and the other one is relatively sparse. In Figure 2, the purple nodes are processes, red and green are files, and yellow represents sockets. Processes are central to the graph, with files providing bridging connection between the nodes. This is a recurring structure in the graphs studied in this work. From this, we can extrapolate that similar interactions will be present in traces captured from other machines as well.

Figure 1 shows the degree distribution across all the 9 traces. It can be seen that the power law distribution is followed by all the graphs. Do note that versioning causes nodes to version, taking along all the active connections to the next node. Such high degree nodes suggest that certain nodes are able to have active connections to a large subset of the system. From a system stability and security perspective, these are high value nodes, and would need to be stable for a large duration of time. Additionally, such nodes would also add pressure on the storage system, as it would be continously adding edges, and a graph storage engine which does not coalesce storage for the edges of such nodes, would cause large parts of th storage to be read multiple times.

Effect of graph structure on storage and analytics: For storage and caching, the presence of high degree nodes can cause significant cache trashing given OPUS like versioning. Access to the properties of the nodes connected to such a high value node would require multiple accesses to the underlying storage. These nodes were added to the system earlier and would have varying lifetimes and hence likely to be stored in different parts of the storage. Thus, a conventional caching mechanism can cause significant cache trashing, as any such access can lead to a large number of cache-evictions.

5 Conclusions

This paper presents that WSP show a power-law distribution, with processes forming the hubs and the other elements connecting the processes forming the authorities. The lifetime of a node in the WSP graph is limited and the effects of such transient nature is also shown. Given the hub-authorities model, the lifetime does not affect the power-law model as long-lasting processes accumulate more edges, and system-critical read-only files do the same. This presents opportunities for not just storage engine, but also to caching, and analysis methods associated with the system.

6 Acknowledgements

This work is part of the CADETS Project sponsored by the Defense Advanced Research Projects Agency (DARPA) and the Air Force Research Laboratory (AFRL), under contract FA8650-15-C-7558. The views, opinions, and/or findings contained in this paper are those of the authors and should not be interpreted as representing the official views or policies, either expressed or implied, of the Department of Defense or the U.S. Government.

References

1. Balakrishnan, N., Bytheway, T., Sohan, R., Hopper, A.: OPUS: A lightweight system for observational provenance in user space. In: Presented as part of the 5th USENIX Workshop on the Theory and Practice of Provenance. USENIX, Lombard, IL (2013), https://www.usenix.org/conference/tapp13/opus-lightweight-system-observational-provenance-user-space
2. Buldyrev, S.V., Parshani, R., Paul, G., Stanley, H.E., Havlin, S.: Catastrophic cascade of failures in interdependent networks. Nature **464**(7291), 1025 (2010)
3. Kleinberg, J.M.: Hubs, authorities, and communities. ACM Comput. Surv. **31**(4es) (Dec 1999). https://doi.org/10.1145/345966.345982, http://doi.acm.org/10.1145/345966.345982
4. Leskovec, J., Kleinberg, J., Faloutsos, C.: Graphs over time: densification laws, shrinking diameters and possible explanations. In: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining. pp. 177–187. ACM (2005)
5. Lusseau, D., Schneider, K., Boisseau, O.J., Haase, P., Slooten, E., Dawson, S.M.: The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations. Behavioral Ecology and Sociobiology **54**(4), 396–405 (Sep 2003). https://doi.org/10.1007/s00265-003-0651-y, https://doi.org/10.1007/s00265-003-0651-y
6. Muniswamy-Reddy, K.K., Holland, D.A., Braun, U., Seltzer, M.: Provenance-aware storage systems. In: Proceedings of the Annual Conference on USENIX '06 Annual Technical Conference. pp. 4–4. ATEC '06, USENIX Association, Berkeley, CA, USA (2006), http://dl.acm.org/citation.cfm?id=1267359.1267363
7. Newman, M.E.: Spread of epidemic disease on networks. Physical review E **66**(1), 016128 (2002)