A Large Scale Synthetic Social Provenance Database

Mohamed Jehad Baeth, Mehmet S. Aktas
Computer Engineering,
Yıldız Technical University
Istanbul / Turkey
Email: mohamed.jehad.baeth@std.yildiz.edu.tr , mehmet@ce.yildiz.edu.tr

***Abstract*—Provenance about data derivations in social networks is commonly referred as social provenance, which helps in estimating data quality, tracking of resources, and understanding the ways of information diffusion in social networks. We observed several challenges related to provenance in the social network domain. First, provenance collection systems capture provenance on the fly; however, their collection mechanism may be faulty and have dropped provenance notifications. Hence, social provenance records may be partial, partitioned, or simply inaccurate. Although current provenance systems deliver a source of real provenance data, these systems do not provide a controlled provenance generation environment; and there are few that contain provenance with failures. Synthetic provenance databases are available in other domains, such as e-Science; but there is also a need for such a database in the social networking domain. To address these challenges, this study introduces a large-scale noisy synthetic social provenance database, which includes a high volume of large-size social provenance graphs. It also introduces metrics that can be used to capture such vital information as provenance for calculating data quality and user credibility.**

***Index Terms*— data quality, large-scale database, provenance, social networks, synthetic workflow simulation.**

I. INTRODUCTION

Social networks are described as online communities and groups of individuals communicating in a Web-based environment, in which their users can interact with each other by posting, commenting, or showing sentiment actions provided by the social network. In addition, social media have been used for gathering information about large-scale events, such as fires, earthquakes, and other disasters, all of which impact government and nongovernment organizations at the local, national, or even international level. Individuals also use social media to find reliable information about what is going on around them and thus are able to leverage new information as quickly as possible [1].

Social media deliver users a large-scale and easy-to-use platform that cannot be achieved using traditional media. Understanding information propagation in social media provides additional context, such as knowing the information originator and its transition modifications until the end of its life cycle. The normal social media user applies such knowledge to evaluate the trustworthiness and correctness of this information [2]. As in real life, the quality of information or objects created in social networks value is affected by its provenance.

We observed several challenges related to provenance in the social network domain. First, existing social networks do not provide any programming interface for accessing the provenance information of the data published therein; and there are no existing mechanisms for identifying and tracing data objects. Provenance collection systems capture provenance on the fly. However, their collection mechanisms may be faulty and have dropped provenance notifications. Hence, social provenance records may be partial, partitioned, or simply inaccurate. Incompleteness and inconsistency of provenance records, if they exist, are a challenge for analyzing provenance datasets [3], [4]. There is a need for a synthetically created social provenance database that is modeled on real social interactions and populated with failure patterns. Although synthetic provenance databases are available in other domains, such as e-Science, there is a need for such a database in the social networking domain as well. Second, social provenance records can grow large quickly because of the high number of participating actors. Although the number of services involved in e-Science workflows is in the order of hundreds, this number can grow to a scale in the order of thousands or millions of social interactions that take place on social media.

To address the abovementioned challenges, this study introduces a large-scale noisy synthetic social provenance database, which includes a high volume of large social provenance graphs. The study also introduces metrics that can be used to capture such vital information as provenance for calculating data quality and user credibility.

The remainder of this paper is organized as follows: Section 2 reviews related work. Section 3 discusses the social provenance metrics that are proposed to be included in social provenance datasets. Section 4 examines our methodological way of creating a synthetic social provenance dataset. Analysis of the generated provenance dataset is addressed in Section 5.

II. RELATED WORK

Several provenance systems currently exist that function as a source of provenance data. However, these systems do not deliver a controlled provenance generation environment; and there are few examples of such systems that can generate provenance with failures [5]. On the other hand, there are several synthetic workloads that have been developed for many different purposes. Some were used in the area of distributed systems [6]–[8]; some were generated for use in the networking research area [9], [10]; and each was used to evaluate performance, as well as for benchmarking purposes,

Copyright (c) IARIA, 2017. ISBN: 978-1-61208-558-6

in their respective areas. Lately, there has been increasing interest in generating synthetic social workloads in the social network domain. Although social networks have high availability, sometimes the collection of social network data may not be feasible due to privacy concerns, where access to such data is restricted to analysts. Some of the introduced synthetic social network generators rely on samples of similar datasets, such as in [11], where a social media dataset can be cloned from an existing set of statistics. Another interesting example of a recent synthetic social network generator [12] simulates the LinkedIn social network. Here, the generation process had two stages. The first stage was the construction of the base network. The second was the addition of LinkedIn endorsements where users can publicly verify that people they know are qualified in the skill that they claim for themselves. However, neither of these simulated network data attempt to model failures. The unreliability of the protocol between the provenance tool and the application are discussed in [5], where a 10 GB database with several scientific workflows was generated using WORKEM, a workflow emulator tool [13]. The database was generated based on real-life e-Science workflows. This study used the Karma provenance capture and management system to manage the scientific provenance datasets, which were compatible with the Open Provenance Model (OPM). The use of such a simulated database in unmanaged workflows is discussed in [14].

To the best of our knowledge, there are no generated workloads and synthetic provenance datasets that have been developed specifically for social provenance research. This study introduces a large-scale noisy synthetic social provenance database, including a high volume of large-size social provenance graphs. It also introduces metrics that can be used to capture such vital information as provenance, which can be used for calculating data quality and user credibility in social networks.

III. SOCIAL PROVENANCE DATABASE REQUIREMENTS

Cheah *et al.* identified several requirements that must be met for a provenance database [5]: large scale, diversity, and realism. A provenance database should consist of a significant number of provenance records to support research at scale and should be drawn from varied workflows that have different characteristics in terms of size, breadth, and length. Also, the composition of workflows used to generate the provenance should have failure characteristics. In addition to abovementioned requirements, we added another requirement: usability. We argue that a provenance database should address not only the generic requirements, but also its domain-dependent requirements.

In this study, we generated a social provenance database that meets the abovementioned requirements as follows: We met the diversity requirement by generating three different types of social provenance, each representing a different scale of social interactions. The categories of social interactions that we used are 100, 1K, and 5K. For each type of social interaction, we created a hundred social-workflow execution traces. We met the realism requirement by producing the same dataset with a 10 percent rate of notification failure and a 10 percent execution failure rate. (Cheah *et al.* generated a noisy 10 GB provenance database with failure characteristics [5] for scientific datasets. Their study included failure characteristics for both provenance-notification failures and workflow-execution failures. Note that we do not consider the latter, since a social workflow is not dependent on a specific workflow. Finally, we met the usability requirement by taking into account the major research problems in the social network domain. Here, we are particularly motivated by research problems that have been investigated by the PRONALIZ project, a Turkish National Science Foundation–funded research project [15]. PRONALIZ investigates the use of provenance in social media to develop methodologies for detection of information pollution and violation of copyrights. We created a publicly accessible Web page for this database and made it available for download at [16]. Throughout the experience of using social media, it can be inferred that its users face two major problems. One is the determination of data authenticity and quality. It is challenging to rate the reliability of a source in a user-generated content platform, where sources might propagate false information, causing the spread of a polluted material. Thus, it would be difficult to determine the actual quality of data and how much weight the data should be given. The second problem is the uncertainty of data visibility due to the dynamic nature of content shared on social media, in which changes can occur on the platform's privacy settings or at the user level by applying more restrictive privacy measures. These policies determine copyrights on a user's shared data. User data, which are intended to be disseminated in a friend circle, may be spread via resharing within the social network. Users are not aware of who can see their data or apply a process to the data. Thus, problems like violation of copyrights can arise. To create a social provenance database that can be used by researchers to address these problems, we identified a number of metrics.

To obtain a better understanding of metrics and an improved definition of the credibility or trustworthiness of an information source, we first need to present our social network provenance model, which we believe can be used as a generic model for provenance representation on all existing social networks.

Users in social networks tend to provide numerous pieces of information about themselves, which varies from one social network to another. For example, a Twitter user has a dedicated area for only his or her bio, location, personal website URL, and date of birth, whereas a Facebook user can provide much more information, such as personal interests, political affiliation, books read, movies watched, educational background, and schools attended. Table 1 shows some of these attributes or types of information and the percentage of users who have added this information to their Facebook profiles and left it public for everyone to see, according to [17].

Copyright (c) IARIA, 2017. ISBN: 978-1-61208-558-6

TABLE I. LIST OF ATTRIBUTES AND PERCENTAGE OF USERS WHO REVEAL THEM ON FACEBOOK.

Attribute	Percentage
Current City	30.17
Gender	81.77
Relationship Status	26.24
Education and Word	25.13
Email	1.32
Interested in	18.66
Music	45.77
Movies	27.92
Activities	18.74
Television	33.30

The availability of such information plays an important role in the creation of social network provenance metrics. The metrics used in generated social workflows are as follows:

A. User Information Provenance Availability Measure

The availability of a user's personal information indicates trustworthiness of this user as for Social network user getting information from another well-known user lends credibility to this information. The availability function, as defined by [17], objectively quantifies progress in obtaining a user's personal attribute values. The availability function describes how much user provenance metadata are available for the statement of interest, in that it allows a user to perform a simple comparison of search strategies employed to obtain provenance attributes. It also allows prioritizing attributes by giving each a specific weight, where the sum of the weights of all attributes is 1; and an attribute with a weight of 0 will have no effect on the outcome of the measure.

B. User Information Provenance Legitimacy Measure

Finding a user provenance attribute might provide some insight; however, a certainty measure of those attributes is needed to indicate validity of found attributes. This can be made by matching found attribute values with attributes found in other sources. The legitimacy function is computed by averaging the number of independent social media sites used to verify the attribute and is proposed to quantify whether or not the provenance attribute values found are valid [17].

C. User Information Provenance Social Popularity Measure (Prestige Centrality)

Typically, a high-profile social network user, who might represent a celebrity or an important individual, has a large number of followers. In other words, a famous user enjoys high popularity, indicated by having many ties with others. In the case of an undirected graph, which is the situation in some social networks, such as Facebook, this metric can instead be represented by centrality, where an actor with a high degree of importance maintains numerous contacts with other network users. A central user occupies a structural position (network location) that serves as a source or conduit for larger volumes of information exchange and other resource transactions with other actors. This can be measured by simply calculating the summation of each actor's number of degrees in a nondirected graph and then normalizing it by dividing it by the maximum number of degrees allowed by the social network.

D. Information Provenance Social Impact Measure

The importance of a piece of information may be inferred by the number of social activities associated with it. For example, a tweet with a high number of Favor, Retweet, and Reply operations may reflect the controversial nature of that information.

Thus, we calculate data proximity in the context of a user's relationships by measuring the social interactions of users who are not directly connected to the subject, divided by the total number of interactions on a piece of information, and dividing the set of all directly not connected users who have performed a social action on a piece of information posted by a user to the set of all unique users who have performed a social action

E. Information Prominence or Proximity Prestige

Thus, we calculate data proximity in the context of a user's relations by measuring the social interactions of users who are not directly connected to the subject, divided by the total number of interactions on a piece of information, and dividing the set of all directly not connected users who have performed a social action on a piece of information posted by a user by the set of all unique users who have performed a social action.

F. The Impact of a Post on a User's Prestige

An increase in the number of followers in response to a post on a social network might provide an indication of the importance of these data. For example, on Twitter a nonprestigious user may gain a very large number of followers by posting valuable information or introducing a piece of information. This should show the impact of the information published on the prestige of its publisher. Table 2, below, shows different categorizations of the presented metrics.

TABLE II. LIST OF SOCIAL PROVENANCE ATTRIBUTES CAPTURED IN THE SOCIAL PROVENANCE DATABASE

Metric	Graph Type		Perspective		Time Dependent
	Directed	Non-Directed	Data in the Center	User in the Center	
Verifiability	X	X		X	
Popularity	Prestige	Centrality		X	
Availability	X	X		X	
Social Impact	X	X	X		
Prestige	X		X		X
Artifact Impact	X		X	X	X

IV. GENERATION OF THE SYNTHETIC DATASET

Normally, a scientific workflow describes the accomplishment of a scientific objective process, which is expressed by the task being done and its dependencies. Typically, scientific workflow tasks are computational steps for scientific simulations or data analysis steps [18]. On the other hand, a social workflow is always bound to run on a

social network. Its operations and data are defined by the social network itself. In turn, each social network names the social operations and data formats differently. To obtain a dataset with controllable characteristics that capture the nature of information propagation on social media, we created a fully synthetic dataset imitating Twitter. This synthetic dataset was designed to meet criteria that may not be achievable when collecting data from a Twitter live feed due to users' privacy settings and availability of different types of personal information, which can impose real issues when evaluating to-be-developed misinformation-detection algorithms. We choose to use W3C's PROV for provenance and metadata modeling rather than its predecessor Open Provenance Model OPM. In this study, we introduce a set of properties that can be used to map the social operations to PROV-O entities. Table 3 lists these properties along with their explanations. Fig. 1 shows how we mapped Social Provenance attributes to each PROV-O entity.

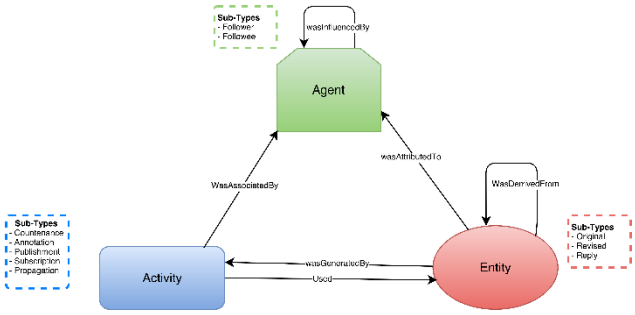

Figure 1. PROV-O Specification based Provenance Nodes and social provenance sub-types.

TABLE III. TERMINOLOGY IN THE PROPOSED SOCIAL NETWORK PROVENANCE MODEL

Sub-Type (Properties)	Explanation	Equivalence in Social Networks
Countenance	To support or approve a statement or an entity or its content	Like(v), Favor(v)
Annotation	to remark, make an observation or make criticism	Reply, Comment
Publishment	To issue textual or graphical materials for public distribution	Post(v), tweet(v)
Subscription	To follow or watch the movement or course/progress of something or someone	Follow, get notified
Propagation	To reproduce transmit, spread or disseminate.	Share, Retweet
Follower	A person who follows another and becomes a subscriber to his/her feed of tweets.	Follower, Liker
Followee	A person who is being tracked on a social media website or application.	User
Original	The blog or post in its state at time of creation by its original creator	Tweet(n), Post(n)
Revised	Reconsider and alter (something) in the light of further evidence.	Retweet, Shared post

Twitter is described to be the largest data source openly accessible to everyone through its stream and search API. Thus, it is the source of much recent research. Currently, many tools have been developed based on mining the large amount of data for information such as tracking earthquakes, world health and the spread of communicable diseases, or even providing real-time information during crises by extracting information from users' Twitter feeds. In short, Twitter is currently used to mobilize emotionally and physically. Social workflows represent an abstract view of the various social patterns observed on Twitter. It can be understood, visualized, and represented in different formats; thus, analysis of it may also be conducted.

A simple workflow normally represents tweets of users who have no intention of engaging or creating a general topic by not using a hashtag. Such tweets usually tend to generate minimal engagement limited to the user's followers. However, high-prestige users with very large numbers of followers can stimulate many interactions and create a widespread impression. On the other hand, we define a composite social workflow as a group of separate workflows, where all users are using a unified topic. Generally, in such events, the majority of the participating users employ a global hashtag or the directed mention of a celebrity's official Twitter account. An example of such social interactions is solidarity and debate where normally an opinion-based community is polarized [19]. Users' interaction dynamics and patterns were observed and analyzed in different social events than belongs to different topic [20]. The study shows different characteristics of the collected social workflows observed from real Twitter data. The possible numbers of user engagements and social interactions in our generated social workflows were derived from these observations, as shown in Table 4. We generated workflows for each of the described categories, in which each workflow is executed four times with different failure-generation modules.

TABLE IV. GENERATED SOCIAL WORKFLOWS USERS' POOL AND NUMBER OF SOCIAL INTERACTIONS

Users Pool	Number of Social Interactions	Number of generated workflows
10	10	100
10	100	100
100	100	400
1000	1000	500
5000	5000	500
5000	10000	100

A. *Database Generation Framework*

The four components used in the creation of the provenance database were WorkflowGen, WorkflowSim, ProvToolbox, and the Komadu provenance repository. Fig. 2 shows an overview of the framework.

Copyright (c) IARIA, 2017. ISBN: 978-1-61208-558-6

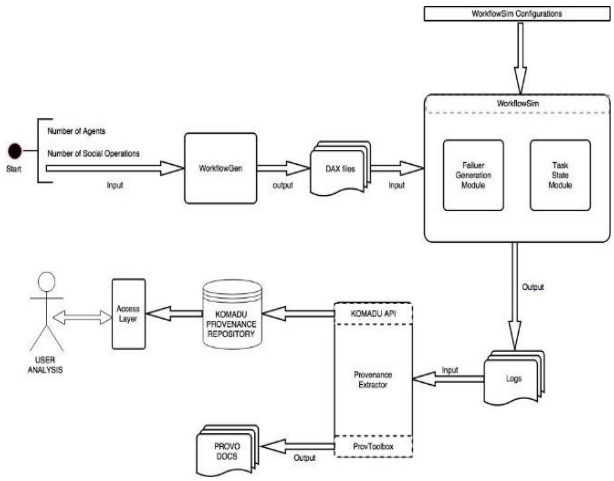

Figure 2. Social Provenance Dataset Generation Framework

Komadu [21] is a standalone provenance capture and visualization system for capturing, representing, and manipulating provenance. It uses the W3C PROV standard [22] considered to be the successor to the Karma [23] provenance capture system.

WorkflowSim is an open-source workflow simulator. It models workflows using a DAG model and supports implementations of some popular dynamic and static workflow schedulers and task-clustering algorithms [24]. WorkflowSim also has failure modeling that supports two failure types on both the job and task levels. Failure rates generated by WorkflowSim are modifiable according to user preference [24]. WorkflowGen, on the other hand, is a tool developed by the same team for the purpose of creating custom DAX workflows to facilitate evaluation of workflow algorithms and systems on a range of workflow sizes, thus creating realistic synthetic workflows resembling those used in the real world similar to the ones gathered from Twitter [25]. We used WorkflowSim as a simulation environment to execute the DAX files generated by WorkflowGen. DAX files represents the abstract description of a single workflow in XML format. The provenance recorded from the logs of the simulation were generated using ProvToolBox and put into Komadu [26].

B. Generated Workflows

The client responsible of the generation of random tweet data consider that any social scenario, no matter how many users are engaged in it or how many social activities has been made upon it, if visualized will be shaped as a multiforked sequential graph. First, the client keeps track of entities linked to the main workflow created either by retweeting or replying. In addition, the client considers only social activities that may be executed on a tweet in that context: Tweet, Like, Retweet, and Reply. The client also creates a pool of agents, where each agent has its own set of popularity, availability, and verifiability values. Finally, the client considers that every social operation is affected by the last social operation made on the same entity. Clients start by creating an initial activity representing a tweet operation, which leads to the creation of the original tweet entity. From that point, the client randomly invokes social operations until the wanted number of operations is reached. The following table shows the Prov-O representation of relationships between entities, agents, and activities created at every iteration, depending on the social operation type.

TABLE V. PROV-O REPRESENTATION OF SOCIAL OPERATIONS AND ENTITIES

Social Operation	Prov-O representation
Post	Generation(tweet_activity, main_tweet) Attribution(main_tweet, agent1) Association(tweet_activity, main_tweet)
Like	Association(new_agent, like_activity) Usage(like_activity, tweet_x)
Retweet	Association(new_agent, retweet_activity) Generation(retweet_activity, new_tweet) Usage(retweet_activity, tweet_x) Attribution(new_tweet, new_agent) Derivation (new_tweet, tweet_x)
Reply	Association(new_agent, reply_activity) Generation(reply_activity, new_tweet) Usage(reply_activity, tweet_x) Attribution(new_tweet, new_agent)

We generated 1600 workflows with 100, 1000, 5000, and 10000 social operations and 500 workflows for every category except for the 10K we generated 100 workflows. The workflows were generated with different sizes of agent pools, ranging from 10 to 5000 agents, and then executed in the following forms:

- Social workflows with complete successful runs.
- Social workflows with simulation execution faults generated using WorkflowSim's fault-generation module, which represents missing notifications coming from the social network to specific actions.
- Social workflows with provenance collection faults, in which some of the provenance data extracted are dropped. This type of fault represents errors that might happen during provenance ingestion into the data repository. The dropped provenance data are selected randomly during workflow simulation at a 10 percent rate.
- Social workflows with faults on both execution and provenance collection levels.

We observed 6400 workflow executions. Fig. 3 shows the distribution of workflows by execution case. In total, we had 1936 successfully executed workflow provenances, 1246 workflows with execution failures, 1917 workflow execution provenances with 10 percent notification drops, and 1283 workflow execution provenances with both failure types. The final size of the dataset is around 10 GB of *.provn* provenance files.

Copyright (c) IARIA, 2017. ISBN: 978-1-61208-558-6

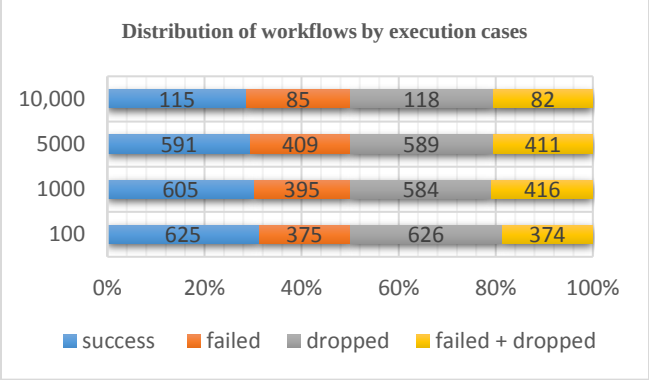

Figure 3. Distribution of Workflows by execution cases

Our observations of individual faulty runs also show that the larger a workflow, the higher the failure rate and the dropped notification rate. The following figures below provide samples from all kinds of generated provenance data of all types of social workflows. Fig. 4 shows the visualization of 10 successful social operations workflow runs.

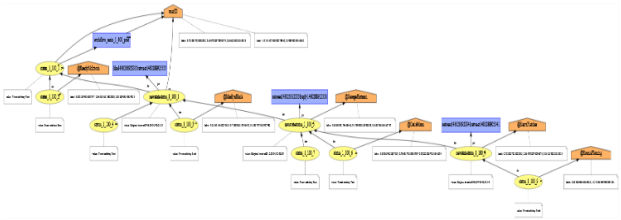

Figure 4. provenance visualization of a successful workflow run

Fig. 5 shows the provenance visualization of 10 social operations workflows with provenance collection failures. It may be observed that some of the relations are missing within the provenance visualization presented in Fig. 2.

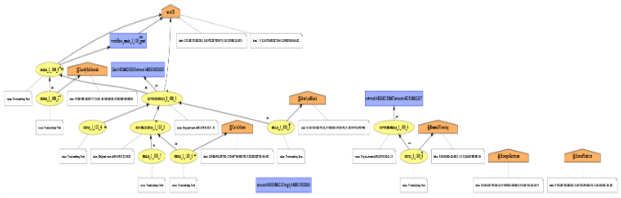

Figure 5. provenance visualization of a workflow execution with provenance collection 10% error rate

Fig. 6 shows the provenance visualization of the same 10 social operations workflow executions with errors on both the notification collection level and provenance ingestion level. Missing activities and missing dangling entities are both observed in the visualization below.

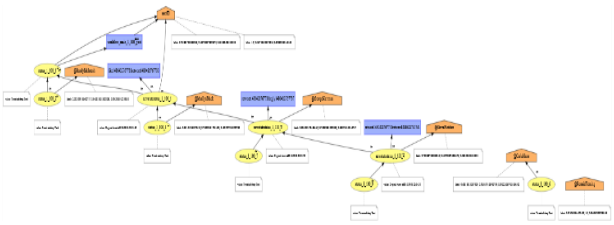

Figure 6. provenance visualization of a workflow execution with both provenance collection error and notification failure error

The social provenance database was developed to serve as a test platform for development of failure-resilient misinformation-detection algorithms.

V. CONCLUSION

In this paper, we have shown the need for a large-scale simulated social provenance database. Taking Twitter as an example, we introduced a large-scale noisy synthetic social provenance database, in which we used various social provenance metrics and attributes to capture vital information for calculating data quality and user credibility. The introduced provenance database consists of social workflows of different-size and different-breadth workflows, each created with randomly generated social interaction scenarios utilizing WorkflowSim and WorkflowGen tools. It also has failure characteristics that represent both notification drop failures and provenance collection failures to simulate real-life provenance capture. We created a publicly accessible website at [15] to make the dataset available for research that deals with large-size and high-volume provenance graphs that are downloadable directly as XML files and are accessible through a Komadu repository query interface. We are now using the provenance database to study social provenance quality and to develop misinformation and copyright violation detection algorithms.

ACKNOWLEDGMENT

This study is part of the PRONALIZ project supported by TUBITAK's (3501) National Young Researchers Career Development Program (Project No: 114E781, Project Title: Provenance Use in Social Media Software to Develop Methodologies for Detection of Information Pollution and Violation of Copyrights).

REFRENCES

[1] S. Ranganath, P. Gundecha, and H. Liu, "A tool for assisting provenance search in social media," *Proc. 22nd ACM Int. Conf. Conf. Inf. Knowl. Manag. - CIKM '13*, pp. 2517–2520, 2013.

[2] I. Taxidou, T. De Nies, and R. Verborgh, "Modeling Information Diffusion in Social Media as Provenance with W3C PROV," In Proceedings of the 24th International Conference on World Wide Web (WWW '15 Companion), pp. 819–824, 2015.

[3] P. Chen, B. Plale, and M. S. Aktas, "Temporal representation for scientific data provenance," in *2012 IEEE 8th International Conference on E-Science, e-Science 2012*, 2012.

[4] P. Chen, B. Plale, and M. S. Aktas, "Temporal representation for mining scientific data provenance,"

Futur. Gener. Comput. Syst., vol. 36, pp. 363–378, 2014.

[5] Y. W. Cheah, B. Plale, J. Kendall-Morwick, D. Leake, and L. Ramakrishnan, "A noisy 10GB provenance database," *Lect. Notes Bus. Inf. Process.*, vol. 100 LNBIP, no. PART 2, pp. 370–381, 2012.

[6] K. Sreenivasan and A. J. Kleinman, "On the construction of a representative synthetic workload," *Commun. ACM*, vol. 17, no. 3, pp. 127–133, 1974.

[7] P. Mehra and B. Wah, "Synthetic workload generation for load-balancing experiments," *IEEE Parallel Distrib. Technol.*, vol. 3, no. 3, pp. 4–19, 1995.

[8] R. Bodnarchuk and R. Bunt, "A synthetic workload model for a distributed system file server," *Proc. 1991 ACM SIGMETRICS Conf. Meas. Model. Comput. Syst. - SIGMETRICS '91*, pp. 50–59, 1991.

[9] S. Antonatos, K. G. Anagnostakis, and E. P. Markatos, "Generating realistic workloads for network intrusion detection systems," *ACM SIGSOFT Softw. Eng. Notes*, vol. 29, no. 1, p. 207, 2004.

[10] B. D. Noble, M. Satyanarayanan, G. T. Nguyen, and R. H. Katz, "Trace-based mobile network emulation," *Proc. ACM SIGCOMM '97 Conf. Appl. Technol. Archit. Protoc. Comput. Commun. - SIGCOMM '97*, pp. 51–61, 1997.

[11] A. M. Ali, H. Alvari, A. Hajibagheri, K. Lakkaraju, and G. Sukthankar, "Synthetic Generators for Cloning Social Network Data," *BioMedCom*, pp. 1–9, 2014.

[12] H. Pérez-Rosés and F. Sebé, "Synthetic generation of social network data with endorsements," *J. Simul.*, vol. 9, no. 4, pp. 279–286, 2014.

[13] L. Ramakrishnan, D. Gannon, and B. Plale, "WORKEM: Representing and emulating distributed scientific workflow execution state," *CCGrid 2010 - 10th IEEE/ACM Int. Conf. Clust. Cloud, Grid Comput.*, pp. 283–292, 2010.

[14] M. S. Aktas, B. Plale, D. Leake, and N. K. Mukhi, "Unmanaged workflows: Their provenance and use," *Stud. Comput. Intell.*, vol. 426, pp. 59–81, 2013.

[15] "Pronaliz Project." [Online]. Available: https://sites.google.com/view/pronaliz/home, retrieved: 15/02/2017.

[16] "synthetic social provenance database." [Online]. Available: https://sites.google.com/view/pronaliz/datasets, retrieved: 28/03/2017.

[17] G. Barbier, Z. Feng, P. Gundecha, and H. Liu, *Provenance Data in Social Media*, vol. 4, no. 1, pp. 1–84, 2013.

[18] I. Wassink et al. "Analysing scientific workflows: Why workflows not only connect web services," in *SERVICES 2009 - 5th 2009 World Congress on Services*, 2009, no. PART 1, pp. 314–321.

[19] M. Transfeld and I. Werenfels, "#Hashtagsolidarities: Twitter debates and networks in the MENA region," no. March, pp. 1–62, 2016.

[20] E. Del Val, M. Rebollo, and V. Botti, "Does the type of event influence how user interactions evolve on twitter?," *PLoS One*, vol. 10, no. 5, pp. 1–32, 2015.

[21] I. Suriarachchi, Q. Zhou, and B. Plale, "Komadu: A Capture and Visualization System for Scientific Data Provenance," *J. Open Res. Softw.*, vol. 3, no. 1, p. e4, 2014.

[22] W3C, "The PROV Data Model.," 2016. [Online]. Available: https://www.w3.org/TR/prov-dm/, retrieved: 15/02/2017.

[23] B. Cao, B. Plale, G. Subramanian, E. Robertson, and Y. Simmhan, "Provenance information model of Karma version 3," in *SERVICES 2009 - 5th 2009 World Congress on Services*, 2009, no. PART 1, pp. 348–351.

[24] W. Chen, M. Rey, and M. Rey, "WorkflowSim : A Toolkit for Simulating Scientific Workflows in Distributed Environments," *8th IEEE Int. Conf. eScience 2012 (eScience 2012)*, pp. 1–8, 2012.

[25] R. Ferreira, W. Chen, R. Ferreira, W. Chen, G. Juve, K. Vahi, and E. Deelman, "Community Resources for Enabling Research in Distributed Scientific Workflows Community Resources for Enabling Research in Distributed Scientific Workflows," no. October, 2014.

[26] L. Moreau, "ProvToolbox: Java library to create and convert W3C PROV data model representations." [Online]. Available: http://lucmoreau.github.io/ProvToolbox/ , retrieved: 15/02/2017.

Copyright (c) IARIA, 2017. ISBN: 978-1-61208-558-6