Provenance for Longitudinal Analysis in Large Scale Networks

Andrei Stoica[1][0009−0003−1838−7856] and Mirela Riveni[2][0000−0002−4991−3455]

[1] University of Groningen
The Netherlands
andreistoica12@gmail.com
[2] Information Systems Group, University of Groningen
The Netherlands
m.riveni@rug.nl

Abstract. Concerns related to the veracity and originality of the content on social networks are at an ongoing rise. Considerable work has been done on information spreading, and tools have been built, while approaches with provenance-based analysis are rare. We are of the opinion that provenance-based analysis and visualization tools can make (mis-)information spreading analysis more efficient. Thus, we study provenance, and present a provenance pipeline for data analytics, where users are able to interact with multiple network analysis modules through a graphical user interface, and describe a proof-of-concept system. Although provenance visualization can suffice in capturing all the necessary metadata, integration with other network visualization modules suited to the same data enhanced our results analysis and conclusions. Having designed distinct provenance models, we captured and analysed lineage of information on community dynamics. We tested our proposed prototype with a real-world dataset comprising of more than 10 million filtered tweets, focused on COVID-19 vaccinations, and conducted an analysis on community dynamics with network science metrics and NLP.

Keywords: Large Scale Networks · Network Science · Data Analysis · Misinformation · NLP · Provenance.

Acknowledgment

We thank the Center for Information Technology of the University of Groningen for their support and for providing access to the Hábrók high performance computing cluster.

1 Introduction

The volume and complexity of the data that we are dealing with in every area of our social lives is only increasing. Thus, ways of structuring, analysing and visualizing data is of crucial importance for valuable insights when we deal with Big

Data in research. And, a critical challenge emerges: how do we guarantee the reliability and traceability of data, and how can we improve the ways of analysing and visualising data? This is where the notion of provenance comes into play. The authors in [14] have classified provenance systems into database-oriented, service-oriented and miscellaneous categories, depending on their application area. [3] introduced the concepts of "why-provenance", i.e. the process that generates the data, and [4] defined the "where-provenance" as the origin of data. [2] present an in-depth study on workflow provenance, in relation with workflows in scientific computing. Provenance is not studied enough in data science and its benefits in Big Data analytics research, and more specifically in social network analysis. We focus on case studies of social network analysis data, and specifically in analysing misinformation, from influential accounts, group structures around influential accounts, to network structure as parameters of influence in (mis-)information spreading. Furthermore, we analyse how provenance-based analysis can help in these types of analysis. Considerable work has been done on misinformation detection and spreading. However, work on provenance-based analysis lack, and we believe that this is an important approach for analysis and visualization of network data, as proven by the results of this work. This work focuses on the research questions posed in the following. A) How can we best model provenance data for efficient studies of large-scale social network data, such as misinformation spread analysis? B) What are the benefits that provenance-based analysis can bring to social-network data analytics? C) What are the components and how best to integrate them for a holistic social network analysis pipeline with provenance-based analysis and visualisation? We introduce a proof-of-concept prototype, which includes modules that receive network data as input in multiple formats, provide functionality of running various network-science metrics and clustering algorithms, and get the results in a file, in graph-based visual forms, in network view and in provenance-based model graphs. This paper is organized as follows. In Section II we discuss related work on provenance and its application areas. Section III describes the implementation of a module integrated in the data analytics pipeline, i.e., for opinion changes with NLP. In Section IV we discuss our framework for provenance-based data analysis and visualisation. Section V discusses our provenance models and experiments with our proof-of-concept prototype. In Section VI we conclude the paper including a discussion on future work.

2 Related Work

2.1 Provenance of data in different application areas

Provenance in social computing is discussed by the authors in [13], where they present provenance models for crowdsourcing for teams. The authors have conducted experiments with synthetic and real world data of software engineering teams that collaborate online, and have modeled provenance for task assignment and task execution. They also provide visualisation of their provenance models based on logs files. Our work is similar to this work in terms of granularities, as

we also present models considering individual and network level, but for a different application area. Our approach is specifically tailored to social networks data, and we also bring the novelty with providing a proof-of-concept prototype. Provenance is also explored for crowdsourcing, and is discussed in [16], [8]. Provenance for process mining is described in [20]. In the area of Web Data, the author in [7] proposes a provenance model suitable for systems that consume linked data.

2.2 Provenance in social networks

A methodology for misinformation detection on Twitter [3] data is presented in [1]. Provenance data is used to determine the original source of a piece of information. The study introduces both user- and content-related metrics based on social provenance attributes captured in a social provenance database and the fuzzy analytic hierarchy process algorithm determines the weights of these metrics in order to assess the credibility of information which circulates on the network. However, the authors in this paper work with synthetic data, while we use a real dataset, and in addition provide a framework. Tracking provenance data with retweets is presented in [10]. In this work we consider all types of reactions and also properties of accounts and tweets, details which give us more effective analysis results about the important indicators in information spread. The networks field of study with regard to provenance is not studied enough, especially with respect to possible provenance models specifically tailored to social networks interactions. This is where our work is important. We provide provenance models for modeling social network data from different perspectives and different granularity as can be seen from Section IV and describe an implemented proof of concept pipeline for social network analysis, where data is mapped to the PROV-O ontology, analysed, and visualised with different views.

2.3 Provenance Visualization

The authors in [19] have presented a visualization approach that takes graphs as input and outputs provenance-based results from various types of analysis such as graph comparison, summarising, and stream data. Another graph-based, provenance data visualisation tool named, Prov Viewer is presented by authors in [9]. Our proposed framework differs from these two tools in that it is a more holistic architecture proposal, on which users can also run different network-science based metrics and algorithms, and specify if they want to have them shown as network-based results or as provenance graphs.

[3] We use the term Twitter because the dataset we work with has been collected before its acquisition, it should be noted that when we refer to Twitter we refer to the platform before its acquisition and renaming to X.

3 Misinformation Spreading Analysis: Opinion changes with NLP

3.1 Module overview

One module which is integrated into the data analytics pipeline described throughout the paper is the detection of opinion changes with the help of natural language processing techniques. The dataset [5] we apply our algorithms on is structured in a similar fashion to how the official Twitter API generates the data. More precisely, Twitter generates a JSON tweet object. Some of the keys in the tweet dictionary had been refactored (either renamed, recomputed, removed or changed the type of) prior to us working with the dataset. After performing the changes, all recordings had been merged into 20 .csv files, counting for more than 10 million tweets, most of them corresponding to 20 days, starting around the beginning of March 2021, roughly one year after the COVID-19 pandemic gained momentum. To accomplish the desired results, we keep track of the situations where one user posted multiple reactions, i.e. replies, quotes, retweets, or any combination of them, to a source tweet. Opinion change is obviously a comprehensive task. In the context of social media, resources are even more limited. We only have access to a user's activity, which in most cases turns out to be reactive rather than proactive and not necessarily well argumented. Users can either react to certain source tweets by adding textual comments or they can like or retweet the original post. On top of that, users can interact by following or unfollowing each other, which may also signal approval or disapproval of certain users' views. Therefore, due to the subjective nature of opinion changes, especially within social media, detecting them is a delicate task. There is a plethora of NLP libraries which compute text sentiments. Our choice was SentiStrength [15]. The choice was made based on the thorough dictionary used to analyse the texts, as well as the possibility to compute a scale score, not only a binary score (negative or positive), meaning sentiment scores range from -4 (most negative) to 4 (most positive). This allowed us to analyse the intensity of the opinion changes further down the line. The SentiStrength algorithm is lexicon-based, meaning it makes use of a pre-defined sentiment lexicon. This is a collection of words that were assigned sentiment scores. Its documentation mentions the algorithm is designed to estimate the strength of positive and negative sentiment in short texts in English, even for informal language. The authors point out that it has human-level accuracy for short social web texts in English, except political texts. Nevertheless, there are limitations to the algorithm, such as a limited context sensitivity. Lexicon-based algorithms do not spot that words can have distinct meanings in different contexts. Also, sarcasm or irony pose serious problems. For example, the text *"Side effects: sore arm, and an overwhelming dread of having to go back to work tomorrow."* is an ironic text which was assigned the minimum score of -4, where in fact it reflects a positive reaction to a source tweet. The lexicon should be updated regularly in order to include new words, as well. One way to overcome these issues is to use machine learning algorithms. They use learning techniques, such as (deep) neural networks, support vector machines or logistic

regressions to identify the relationship between text features and sentiment labels from labeled training data. However, the disadvantage is the lack of a large, up-to-date labeled training dataset, relevant to our topic, to train such models on. Moreover, machine learning algorithms can be much more computationally demanding that lexicon-based, albeit the execution of the latter was challenging in itself given our large dataset.

3.2 Text cleaning

Although the SentiStrength library is built to interpret relatively informal social media texts, there is a specific jargon that needs to be treated manually beforehand, so that the algorithm yields the best possible outcome. Therefore, the tweets suffered the following alterations: all words were converted to lowercase, new line characters were replaced by white spaces, tags were removed, URLs were removed, punctuation was removed, contractions were converted to full forms, emojis (emoticons, symbols and pictographs, flags, etc.) were removed, stopwords were removed. Lastly, we opted to create a custom, smaller list of stopwords that ought to be removed from tweet texts before being passed on to the sentiment analysis algorithm for processing. We studied an already built-in list of stopwords provided by another popular NLP module, namely **NLTK**, and it served general purposes, thus results for our particular case were worse than expected. This is due to the list being too extensive and a lot of words having an impact in computing the sentiment of a tweet, such as "not" or "all".

3.3 Detecting opinion changes

Before describing the process of detecting opinion changes, we need to define the design choices and assumptions we considered. First, a group of recordings may contain a considerable number of tweets and within the respective group, there may be more than one opinion change. Hence, for the sake of capturing the most relevant information, we opted to only track the largest opinion change within the group, which we refer to as the opinion change of the group, i.e. the difference in sentiment score between the largest score and the lowest score. Furthermore, even though more tweets' texts may generate the same maximum or minimum score, we only take the earliest occurrence of each (minimum and maximum scores) into account. Lastly, we base our analysis on the assumption that retweets represent a full and undisputed agreement with the original post and we artificially assigned the sentiment score of 4 (the maximum value of the SentiStrength range). Having established the ground rules we take into consideration, we describe the process of observing changes in opinions. For each group of tweets, we compute the tweets' sentiment scores. If there is a change within a group, i.e. there is at least one negative score and at least one positive score, then we identified an opinion change and we add it to a dictionary with the shape: keys representing a tuple of **(author_id, reference_id)**, and values representing a list of sentiment scores. Note that the chronological order of the tweets in each group was retained, meaning we can spot if we have a negative or positive opinion change within a group. Our analysis revolves around the senti-

ments of reactions as we follow the dynamics of interactions and whether these have an impact on people's interpretation of the respective topics.

4 Social Networks Data Analysis with Provenance

4.1 Background, Problem Statement and Methodology

This study defines and implements social network provenance models, specifically tailored to large-scale social network analysis, i.e., Twitter interactions. On top of that, it also provides a graphical user interface to further ease interactions with the created provenance models, as well as other types of network visualizations. Different provenance representations of lineage data in social networks can highlight trends, they can uncover important features of the entities and agents participating in such interactions. Both textual and visual provenance information can further be used to detect misinformation spread, predict user behaviour or classify tweet content, e.g., for opinion change and sentiment analysis. Provenance can be defined as the documented history of an entity's origin, transformations and interactions within a given context. A formalized way to create provenance representations is the PROV Data Model (PROV-DM) [17], and the provenance ontology (PROV-O) [18] standardised by W3C. The main elements of the data model are the entities, activities and agents [17]. They are also encoded in the PROV ontology (PROV-O) [18], which is meant to represent, exchange and integrate provenance information retrieved from different systems and from within distinct environments.

Technology stack: The dataset we use is based on the ID's provided by authors in [6], which we hydrated to get the tweet contents that we were interested in. As far as the set of technologies is concerned, throughout the implementation of this project, various ones are used. The provenance models are implemented using the ProvToolbox Java library [11, 12]. With respect to the actual information contained in the provenance representations, it is retrieved from the original dataset, using Python, due to its build-in capabilities to perform fast, vectorized operations on large datasets. Another implementation of this study presents an interactive software solution which integrates a number of social network analysis modules, namely provenance visualizations, network graphs and opinion changes analysis. They comprise the logic behind a GUI with the following architecture: the frontend component is a React application. There are two backend applications, a Spring Boot one for the provenance Java module and a Flask one for the Python-based network generation algorithm.

4.2 Provenance Modeling

Model 1 - Individual tweets: The first model we design aims to cover information about individual tweets. Therefore, at the center of the representation there is a source/original tweet. One can choose any number of tweets from the subnetwork of reactions associated with the original tweet and display them. Relevant characteristics are chosen for every type involved in the model, i.e.

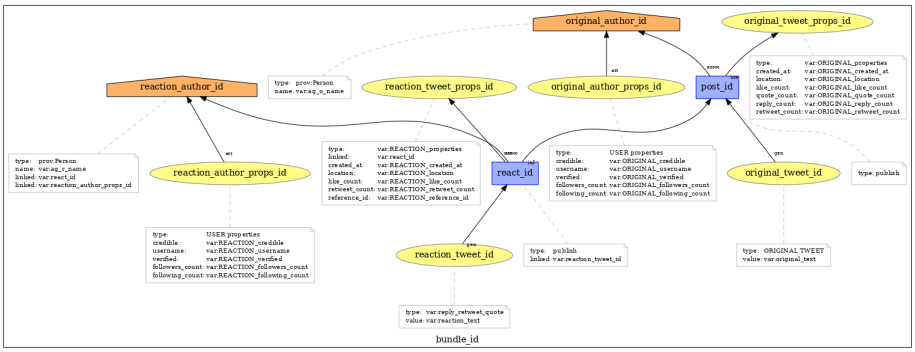

Fig. 1. Model 1. Original tweet and a reaction

original/reaction tweet and properties entities, post or react activities, author agents and properties. It aims to present a detailed overview of atomic participants within a social network. The complete model, displaying a reaction, along with the original tweet, is shown in Figure 1. The figure is a SVG representation of provenance information created with the ProvToolbox library.

Model 2 - Aggregate statistics about reactions over time: Model 2 aims to capture a broader picture than Model 1, it places a source tweet at the heart of the representation and displays aggregate statistics about reactions to the source tweet occurring after certain amounts of time. Thus, it is designed to track time dynamics as well. The main particularity of Model 2 is the switch of focus towards groups of reactions, instead of individual ones. This enables us to get a better understanding of the general trends within social networks, and not only limiting the study to the behaviour and properties of the most popular tweets. The graphical representation of Model 2 is given in Figure 2.

Model 3 - All tweets at a given time-point: Unlike Models 1 and 2, Model 3 takes into consideration all tweets at a given time-point, and it captures time dynamics as well. The representation's aim is to outline essential information on as many textual tweets as possible, regardless of the type (i.e., original tweets, replies and quotes).

4.3 Experiments, Results and Visualisation

All provenance experiments are conducted on the dataset described in Section III. A few notable results were seen when using the provenance pipeline to analyse our dataset. An instance of Model 2 was captured, in a way that the tweet with the most reactions was selected, and its subnetwork studied. The chosen subnetwork includes all reactions, i.e. all replies, quotes and retweets to the original tweet, posted at a time difference indicated by the time interval. In order to draw a conclusion on how fast and how many users respond to a tweet, the interval between the first and last reactions was considered, and split into a few shorter ones. For each interval, the reactions were filtered. Having a subset of reactions, the required aggregate statistics could be computed. For

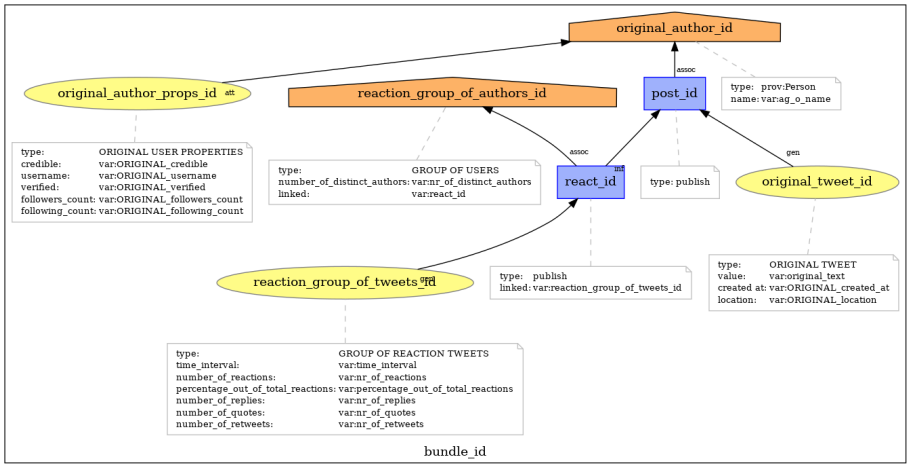

Fig. 2. Model 2. Original tweet, and aggregate reactions

example, the number of reactions per type, the percentage of reactions posted in each respective interval out of all reactions to that specific original tweet and the number of distinct users who posted the reactions of the selected time interval need the whole subset to be computed. Around 5% of the reactions were posted in that time, namely 2380. During the first three hours, roughly half of the reactions occur. The number is impressive, as one can consider this interval to be relatively short. After three hours, but within the first day after the moment the source tweet is posted, the percentage of reactions is equally impressive, almost half of the reactions are released. After the first day, reactions are sparse, only around 6-7%. It is clear to see that the vast majority of reactions occur right after a source tweet is posted. Interest on news or opinions in social networks plummets in a very short time. Figure 3 displays an instance of Model 3. Tweets

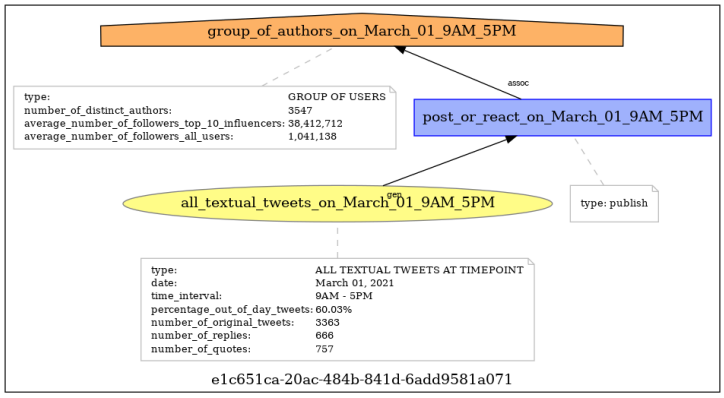

Fig. 3. Provenance graph with results for Model 3: March 1 - business hours

posted on March 1, 2021, which is a regular non-holiday Monday, were selected, during the interval 9AM - 5PM, denoted as the working hours. All timestamps have been converted to their local timezone.

5 Proof-of-concept Prototype for Provenance-based Data Analytics

Although provenance visualizations can suffice in capturing all the necessary metadata to prove a point, integration with other visualization modules suited to the same data available can only enhance the conclusions. Therefore, a complete pipeline was created, where users are able to interact with multiple social network analysis modules through a graphical user interface (GUI). In an attempt to offer an alternative perspective view on the data, the proposed implementation includes the option to construct network graphs, using the same filters as for the provenance visualizations. This can be observed in another snapshot of the front-end application in Figure 4. Within our GUI, users can depict

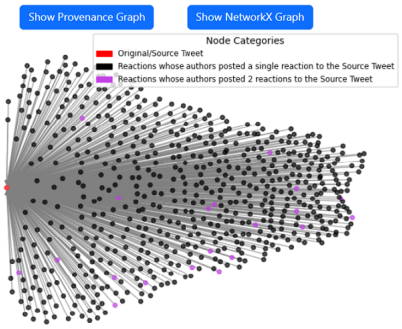

Fig. 4. GUI - reaction network graph, after applying filters

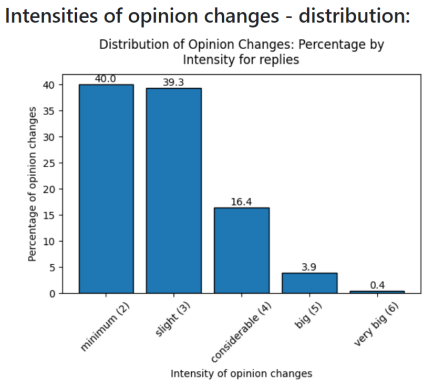

Fig. 5. GUI - intensity distribution of opinion changes for replies subnetwork

distributions of opinion changes within Twitter interactions, obtained using the Matplotlib Python module. The user input is the textual reaction types(s) to be included in the analysis, i.e. any combination of replies and/or quotes. The operation is performed by the algorithm described in Section III. An example representation of the GUI, where a user selects a subnetwork of replies, is given in Figure 5. Our prototype implementation and the experiments we conducted show that provenance data, if modeled properly, can give valuable insights in (mis-)information spread analysis, and the important properties of accounts and tweet types influential in spreading misinformation. Stakeholders can get different detailed views on multiple properties of networks, and provenance can help in better network property analysis when applied together with the well known network-science metrics and algorithms. In general, from our experiments, and

related work, we can conclude that provenance is of benefit to network analysis as it provides detailed and varied views of data, we thus take this opportunity to make a call for provenance to be studied wider.

6 Conclusion and Future Work

Provenance modeling and analysis in this study aim to offer a data analysis mechanism and framework, tailored to large scale network data, with an example from Twitter interactions. With the integration of provenance, along with network graphs and opinion change distributions, described in Chapter IV, trends or prominent features of the entities or agents, i.e. tweets or authors, participating in such interactions, can be effectively analysed. More specifically, they refer to the lineage of data, community detection and dynamics, optimal posting times for maximum exposure and common reaction time intervals. The pipeline can be of use to both researchers who appreciate alternative views on data or practitioners who would like to develop machine-learning algorithms to predict user behaviour within social networks or classify posts based on their documented lineage. Albeit not the sole focus of this research, as part of the application of the provenance module, we found from the real-life dataset a tendency for an almost immediate response to new information released through the network. Our future work includes testing our pipeline with experiments on eco-chambers on Mastodon, we will analyse group dynamics on this platform and see what insights we can get with provenance visualization with the purpose of further proving the generality of the prototype for graph-based large network datasets.

References

1. Baeth, M.J., Aktas, M.S.: Detecting misinformation in social networks using provenance data. Concurrency and Computation: Practice and Experience **31**(3), e4793 (2019). https://doi.org/https://doi.org/10.1002/cpe.4793
2. Barga, R.S., Digiampietri, L.A.: Automatic generation of workflow provenance. In: Proceedings of the International Provenance and Annotation Workshop (IPAW). pp. 1–9 (2006)
3. Buneman, P., Khanna, S., Tan, W.C.: Data provenance: Some basic issues. In: FST TCS 2000: Proceedings of the 20th Conference on Foundations of Software Technology and Theoretical Computer Science. pp. 87–93. Springer-Verlag, London, UK (2000)
4. Buneman, P., Khanna, S., Tan, W.C.: Why and where: A characterization of data provenance. Lecture Notes in Computer Science **1973**, 316–330 (2001)
5. DeVerna, M.R., Pierri, F., Truong, B.T., Bollenbacher, J., Axelrod, D., Loynes, N., Torres-Lugo, C., Yang, K.C., Menczer, F., Bryden, J.: Covaxxy: A collection of english-language twitter posts about covid-19 vaccines. Proceedings of the International AAAI Conference on Web and Social Media **15**(1), 992–999 (May 2021). https://doi.org/10.1609/icwsm.v15i1.18122
6. DeVerna, M.R., Pierri, F., Truong, B.T., Bollenbacher, J., Axelrod, D., Loynes, N., Torres-Lugo, C., Yang, K.C., Menczer, F., Bryden, J.: Covaxxy: A collection

of english-language twitter posts about covid-19 vaccines. In: Proceedings of the International AAAI Conference on Web and Social Media. vol. 15, pp. 992–999 (2021)
7. Hartig, O.: Provenance information in the web of data (2009)
8. Huynh, T.D., Ebden, M., Venanzi, M., Ramchurn, S.D., Roberts, S.J., Moreau, L.: Interpretation of crowdsourced activities using provenance network analysis. In: Hartman, B., Horvitz, E. (eds.) Proceedings of the First AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2013, November 7-9, 2013, Palm Springs, CA, USA. pp. 78–85. AAAI (2013). https://doi.org/10.1609/HCOMP.V1I1.13067, https://doi.org/10.1609/hcomp.v1i1.13067
9. Kohwalter, T.C., de Oliveira, T.N., Freire, J., Clua, E., Murta, L.: Prov viewer: A graph-based visualization tool for interactive exploration of provenance data. In: Mattoso, M., Glavic, B. (eds.) Provenance and Annotation of Data and Processes - 6th International Provenance and Annotation Workshop, IPAW 2016, McLean, VA, USA, June 7-8, 2016, Proceedings. Lecture Notes in Computer Science, vol. 9672, pp. 71–82. Springer (2016). https://doi.org/10.1007/978-3-319-40593-3_6, https://doi.org/10.1007/978-3-319-40593-3_6
10. Migliorini, S., Gambini, M., Quintarelli, E., Belussi, A.: Tracking social provenance in chains of retweets. Knowledge and Information Systems pp. 1–28 (2023)
11. Moreau, L.: ProvToolbox. https://lucmoreau.github.io/ProvToolbox/, accessed: 2023-06-10
12. Moreau, L.: ProvToolbox GitHub Repository. https://github.com/lucmoreau/ProvToolbox, accessed: 2023-06-10
13. Riveni, M., Nguyen, T., Aktas, M.S., Dustdar, S.: Application of provenance in social computing: A case study. Concurr. Comput. Pract. Exp. **31**(3) (2019). https://doi.org/10.1002/CPE.4894, https://doi.org/10.1002/cpe.4894
14. Simmhan, Y.L., Plale, B., Gannon, D.: A survey of data provenance in e-science. SIGMOD Record **34**(3), 31–36 (Sep 2005). https://doi.org/10.1145/1084805.1084812
15. Thelwall, M., Buckley, K., Paltoglou, G.: Sentiment strength detection for the social web. J. Assoc. Inf. Sci. Technol. **63**(1), 163–173 (2012). https://doi.org/10.1002/ASI.21662, https://doi.org/10.1002/asi.21662
16. Willett, W., Ginosar, S., Steinitz, A., Hartmann, B., Agrawala, M.: Identifying redundancy and exposing provenance in crowdsourced data analysis. IEEE Transactions on Visualization and Computer Graphics **19**(12), 2198–2206 (2013). https://doi.org/10.1109/TVCG.2013.164
17. World Wide Web Consortium: PROV-DM Specification. https://www.w3.org/TR/2013/REC-prov-dm-20130430/ (2013), accessed: 2023-05-23
18. World Wide Web Consortium: PROV-O Specification. https://www.w3.org/TR/2013/REC-prov-o-20130430/ (2013), accessed: 2023-05-26
19. Yazici, I.M., Aktas, M.S.: A novel visualization approach for data provenance. Concurr. Comput. Pract. Exp. **34**(9) (2022). https://doi.org/10.1002/CPE.6523, https://doi.org/10.1002/cpe.6523
20. Zerbato, F., Burattin, A., Völzer, H., Becker, P.N., Boscaini, E., Weber, B.: Supporting provenance and data awareness in exploratory process mining. In: International Conference on Advanced Information Systems Engineering. pp. 454–470. Springer (2023)