Towards Multi-level Provenance Reconstruction of Information Diffusion on Social Media

Tom De Nies[‡] Io Taxidou[*] Anastasia Dimou[‡] Ruben Verborgh[‡]
Peter M. Fischer[*] Erik Mannens[‡] Rik Van de Walle[‡]

[‡]{tom.denies,anastasia.dimou,ruben.verborgh,erik.mannens,rik.vandewalle}@ugent.be
Ghent University - iMinds - Multimedia Lab, Belgium

[*]{taxidou,peter.fischer}@informatik.uni-freiburg.de
University of Freiburg, Germany

ABSTRACT

In order to assess the trustworthiness of information on social media, a consumer needs to understand where this information comes from, and which processes were involved in its creation. The entities, agents and activities involved in the creation of a piece of information are referred to as its provenance, which was standardized by W3C PROV. However, current social media APIs cannot always capture the full lineage of every message, leaving the consumer with incomplete or missing provenance, which is crucial for judging the trust it carries. Therefore in this paper, we propose an approach to reconstruct the provenance of messages on social media on multiple levels. To obtain a fine-grained level of provenance, we use an approach from prior work to reconstruct information cascades with high certainty, and map them to PROV using the PROV-SAID extension for social media. To obtain a coarse-grained level of provenance, we adapt our similarity-based, fuzzy provenance reconstruction approach – previously applied on news. We illustrate the power of the combination by providing the reconstructed provenance of a limited social media dataset gathered during the 2012 Olympics, for which we were able to reconstruct a significant amount of previously unidentified connections.

1. INTRODUCTION

Nowadays, information from social media is frequently analysed and processed for professional use. Examples include online journalism, rumor detection, and viral marketing [10]. In all these cases, it is important for the consumer to know the level of trust and relevance that the information carries. An important step in the process of determining trust of information is to expose its provenance [4]. To model provenance for information diffusion on social media, we specified PROV-SAID [14], an extension to the W3C PROV model. Using this model, the social and influence graphs can be represented in an interoperable way. However, automatically reconstructing the aforementioned graphs based on the APIs that most social media provide poses a challenge. Most current methods are designed to only model direct, high certainty influence edges, caused by *explicit* re-emission of messages (e.g., retweets) and combined with connections between users (social graph), in order to unveil who was influenced by whom. This approach does not consider the potentially large amount of *inexplicit* influences that are less certain, and thus more difficult to detect automatically (e.g., a user adapting another user's message, without explicitly referring to it). In this case, the provenance must be *reconstructed* somehow, unravelling the unobserved references that users are using but not giving credit to, and revealing their influencers.

In this paper, we combine a fine-grained, high-certainty approach, with a coarse-grained, less certain approach for provenance reconstruction. By doing this, we propose a multi-level approach for provenance reconstruction of information diffusion on social media. Our contributions in this paper are: 1) an approach for creation and integration of multi-level provenance; 2) a real-world application and evaluation of the PROV-SAID model; 3) a mapping in order to convert input data from social media into RDF; 4) a novel application of our previous work on similarity-based provenance reconstruction in the context of social media.

2. RELATED WORK

While *information diffusion* in social media has received a lot of attention, in particular its modeling [7], there is limited work on the reverse procedure, i.e. *information provenance*, which is the focus of this paper. We divide the state-of-the-art in this area in the following categories: (i) provenance through content similarity; (ii) provenance through social graph connections; (iii) provenance through user profile metadata.

(i) The work of [11] focuses on tracing news and quotes (referred to as *memes*) on the Web over time. The focus is on temporal patterns, mutations (alterations) that online phrases undergo and properties of the news' life cycle. A subsequent work using the same datasets and methods [13] shifts the focus on fine-grained content alterations. In [3], we reconstructed provenance of news articles automatically using semantic similarity. In this paper, we adapt this approach for social media and the PROV-SAID model.

(ii) Traditional information diffusion research includes tracing a piece of information back to its sources through social connections, revealing the concepts of influence and trust among the users involved. The work of [6] recovers information recipients sub-graphs given a small fraction of known recipients. In [8] unknown recipients are identified under the assumptions of degree and closeness propensity: nodes with a higher degree and closer to the sources are more likely to propagate information. [2] provides a provenance reconstruction method through social connections based on well established information diffusion models. Finally,

CIKM'15, October 19–23, 2015, Melbourne, Australia.

ACM 978-1-4503-3794-6/15/10 ...$15.00.
DOI: http://dx.doi.org/10.1145/2806416.2806642 .

in [15], we automatically reconstruct *information cascades* that show which paths information took, given a piece of information that propagates over a social graph. Information cascades are graphs that model how information is being diffused from user to user; in other words, our approach in [15] reconstructs the paths of users who propagate information back to the sources by finding intermediate influencers.

(iii) Lastly, provenance can be derived through user profile metadata, attributing relevance and trust to the information emitted according to the characteristics of the contributor. The work of [9] implements a tool for collecting such user information from different media sites, while not providing any information on the provenance paths and sources.

The work in this paper combines concepts from (i) and (ii) in order to reveal provenance paths, by extending and adapting the solutions proposed in [15] and [3]. Finally, the results are modeled and combined in an interoperable way using the PROV-SAID model [14], which extends the W3C PROV model [12]. PROV-SAID provides a rich description of provenance with regard to information diffusion concepts such as: *direct* and *indirect derivations*, *copied* and *modified messages*, and influence types such as *follow relationships* and *interaction influences*.

3. METHOD

As highlighted in the related work and illustrated in Figure 1, we reveal provenance paths on two levels: (1) *low-level (fine-grained)*, based on structure as in [15] and (2) *high-level (coarse-grained)*, based on content similarity as in [3]. These methods are then combined using *PROV-SAID* [14]. In order to convert the XML-based influence graph of [15] into PROV-SAID, we use the *RML mapping language* [5]. RML is used in combination with a processor to convert proprietary data – such as XML – to RDF. In our case the data is converted to PROV-O, which is the ontology that expresses the PROV Data Model. Note that by using RML, we ensure that any input can be converted to PROV-O, rendering our method interoperable and reusable in many applications.

3.1 Low-level, fine-grained provenance

To obtain low-level provenance, we build upon on our previous work [15], that reconstructs the so-called *information cascades* found in social media. Diffusion paths are reconstructed according to who is influenced by whom given messages that propagate over a social graph, with the assumption that users propagate identical messages (e.g., by retweeting) and identify possible influencers. When applied to the Twitter dataset described in Section 4, the reconstructed information cascades comprise of retweets, where users give credit only to the initial source of a message, not the intermediate source that exposes the message to them. In other words, it remains unclear which paths information took from the sources to the recipients. Therefore, the algorithm as described in [15] leverages the social graph in order to reconstruct the intermediate diffusion paths and find influencers, given the assumptions that information flows over the social graph and users are influenced by their connections in order to propagate a piece of information.

The algorithm outputs *edges*, directed from a *tweet A* to a *tweet B*. For each tweet, we have access to the *tweet-id*, *timestamp* and *userid*. When we map this to PROV-SAID using RML, we obtain the following PROV-O sub-graph for each edge:

```
status:tweetA_id  a  prov-said:Message ;
  prov:wasAttributedTo  user:tweetA_userid ;
  prov:wasGeneratedBy  _:emit-tweetA_id ;
  prov:generatedAtTime  tweetA_time .
status:tweetB_id a prov-said:CopiedMessage;
  prov:wasAttributedTo  user:tweetB_userid ;
  prov:wasQuotedFrom  status:tweetA_id ;
  prov:wasGeneratedBy  _:emit-tweetB_id ;
  prov:generatedAtTime  tweetB_time .

user:tweetA_userid  a  prov:Agent .
user:tweetB_userid  a  prov:Agent ;
  prov:wasInfluencedBy user:tweetA_userid ;
  prov:qualifiedInfluence  [
    a prov-said:FollowRelationship ;
    prov:agent user:tweetA_userid . ] ;
  prov:qualifiedInfluence  [
    a prov-said:InteractionInfluenceRelationship;
    prov:agent user:tweetA_userid . ] ;

_:emit-tweetA_id a prov-said:EmitMessage .
_:emit-tweetB_id a prov-said:EmitMessage ;
  prov:used  status:tweetA_id .
```

Note that the prefixes `status:` and `user:` refer to `https://twitter.com/statuses/` and `https://twitter.com/intent/user?user_id=`, respectively, and that the prefixes `prov:` and `prov-said:` refer to their respective namespaces. This representation of the information cascades as provenance is now suitable to be merged with other interoperable provenance, such as the high-level provenance described in Section 3.2.

3.2 High-level, coarse-grained provenance

To obtain high-level provenance, we consider what is missing from the dataset generated in Section 3.1. Since the approach in [15] only relies on relationships exposed through a social media API, it does not consider all messages that were copied or revised without this being tracked by the social media software (e.g., when a user copy-pastes a message instead of retweeting it). To reconstruct this kind of information diffusion, we adapt our approach introduced in [3] to be usable with social media content. The core assumption of this approach is: *"if two messages are highly similar, there is a high probability that they share some provenance"*. The adapted approach consists of the following steps:

1. remove all tracked copied messages from every information cascade as generated in Section 3.1, keeping only the root messages;
2. index this reduced dataset using a feature model and semantic similarity function (e.g., TF-IDF and the cosine similarity), and compute the full similarity matrix of all messages;
3. apply a similarity-based clustering algorithm such as SimClus [1] to divide the dataset into (possibly overlapping) clusters of messages that all have a similarity to each other higher than a predetermined threshold;
4. for each cluster:
   - identify the oldest message as the root message of that cluster;
   - connect all other messages to the root message:
     - if the message is identical to the root message, using a `prov:wasQuotedFrom` relationship;
     - if the message is not identical to the root message, using a `prov:wasRevisionOf` relationship.

The expected result of this approach is that the vast majority of messages will be clustered as a singleton, meaning that no new relationships are introduced. Nonetheless, for those messages that do get clustered together, we know that they exhibit a high similarity. We use their temporal information to estimate their provenance relationship, thereby enriching the dataset and exposing previously hidden knowledge about the information diffusion.

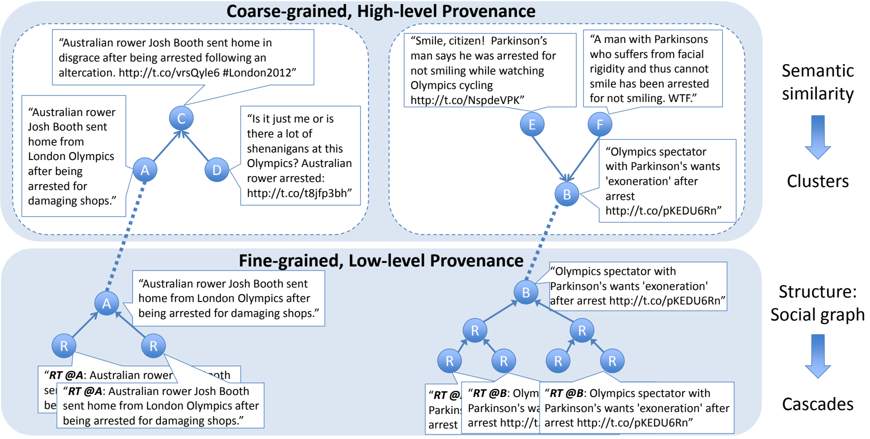
Coarse-grained, High-level Provenance
"Australian rower Josh Booth sent home in disgrace after being arrested following an altercation. http://t.co/vrsQyle6 #London2012"
"Australian rower Josh Booth sent home from London Olympics after being arrested for damaging shops."
C
A
D
"Is it just me or is there a lot of shenanigans at this Olympics? Australian rower arrested: http://t.co/t8jfp3bh"
"Smile, citizen! Parkinson's man says he was arrested for not smiling while watching Olympics cycling http://t.co/NspdeVPK"
"A man with Parkinsons who suffers from facial rigidity and thus cannot smile has been arrested for not smiling. WTF."
E
F
B
"Olympics spectator with Parkinson's wants 'exoneration' after arrest http://t.co/pKEDU6Rn"
Semantic similarity
Clusters
Fine-grained, Low-level Provenance
A
"Australian rower Josh Booth sent home from London Olympics after being arrested for damaging shops."
R
R
"RT @A: Australian rower Josh Booth sent home from London Olympics after being arrested for damaging shops."
B
"Olympics spectator with Parkinson's wants 'exoneration' after arrest http://t.co/pKEDU6Rn"
R
R
R
R
R
R
"RT @B: Olympics spectator with Parkinson's wants 'exoneration' after arrest http://t.co/pKEDU6Rn"
Structure: Social graph
Cascades

Figure 1: Overview of the integrated, multi-level provenance. The arrows for the low-level refer to `prov:wasQuotedFrom` for all copied messages (retweets); for the high-level they refer to `prov:wasRevisionOf` for all modified messages.

When we integrate this result in the next step, we are effectively reconnecting entire information cascades, whose connection was lost to the social media API. Note that due to the calculation of the full similarity matrix, this approach will have an quadratic complexity w.r.t. the number of messages considered, so it should always be applied on a pre-filtered dataset (e.g., a search result).

3.3 Integration of Multi-level Provenance

Because both algorithms output interoperable PROV, the integration of the two aforementioned levels of provenance consists of simply merging the two sets of RDF statements. However, it is important to understand the new structure this will give to the data. We clarify how the data is enriched by the combination of the two reconstructed provenance sets using Figure 1.

Each level of provenance differs in precision and granularity. The fine-grained, low-level provenance is very detailed, and was constructed with high certainty, since it consists solely of copied messages exposed by a social media API (in our case: the Twitter API). The coarse-grained, high-level provenance, however, was constructed in a much less certain way, relying on semantic similarity to reconstruct connections that were lost to the social media API. The two levels enrich each other, providing previously unidentifiable connections between messages for data consumers (e.g., social media analysts) to explore.

4. EVALUATION AND DISCUSSION

As a preliminary evaluation, we tested our approach on a dataset gathered using the Twitter Streaming API during the 2012 Olympics. We chose Twitter because it provides trace information for copied messages (retweets). The dataset was collected by following the keywords 'Olympics2012' and 'London2012'. We limited the dataset by only considering tweets with a certain keyword, in our case: *'arrest'*. This simulates a realistic scenario where a social media analyst first searches for a broad keyword (e.g., a trending topic), and then investigates the information diffusion paths among the results. Complementary, we desired to avoid messages carrying not important information, for example: "I am watching the Olympics". This way, we include relevant events that attract attention both by individual users and mass media, while yielding information cascades by being retweeted. The final dataset consists of 9047 tweets, of which 5174 are copied messages (retweets), and 3873 are original messages according to the Twitter API.

4.1 Low-level Provenance Reconstruction

We identified 31 cascades using the low-level reconstruction approach from Section 3.1, resulting in a skewed distribution from 5 to 1771 recorded retweets with the root tweet contained in the dataset (out of the total of 5174 retweets). This approach has already been thoroughly evaluated in [15], so we can safely assume that the identified cascades are correct.

4.2 High-level Provenance Reconstruction

Using the approach described in Section 3.2, we clustered the 3873 original messages from the dataset based on their semantic similarity. More specifically, we used the TF-IDF approach from traditional information retrieval to model all messages as vectors, and computed their similarity using the cosine similarity. We then executed the SimClus algorithm described in [1]. Essentially, SimClus divides the set of messages into clusters of messages that all exhibit a similarity higher than a predefined threshold to their respective cluster centre. To use the clusters to reconstruct provenance as described in Section 3.2, the major challenge lies in identifying the optimal similarity threshold. The threshold must be high enough to ensure that only messages that actually share provenance get clustered together, while it must also be low enough to avoid that too many messages are clustered as singletons, which would result in missed connections. Ideally, the optimal threshold would be found empirically by analysing the precision and recall of the provenance reconstruction approach, as it was done for news in [3]. For this paper we do not have access to a ground truth as the authors of [3] did. However, we can investigate the influence of the similarity threshold on the number of clusters and their size, which at least gives us an idea of its behaviour.

As illustrated by Figure 2, the total number of clusters is approximately proportionate to the similarity threshold. This means that if we use a low threshold, we will have a small number of relatively large clusters. On the other hand, if we use a high threshold, we can expect a high number of smaller clusters. When the threshold is set to 1, only identical messages will be clustered together, and therefore only retweets – no modified tweets – missed by the Twitter API will be identified. This is further confirmed by our observations of the number of clusters per cluster size, as illustrated by Figure 3. Here, we see that for the lower thresholds (0.3 and 0.5), the cluster size varies highly, whereas there are less different cluster sizes for the threshold 0.7. These observations are an indication that for the

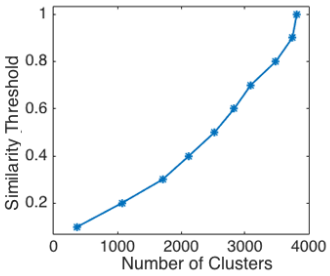

Figure 2: Total number of clusters for each similarity threshold.

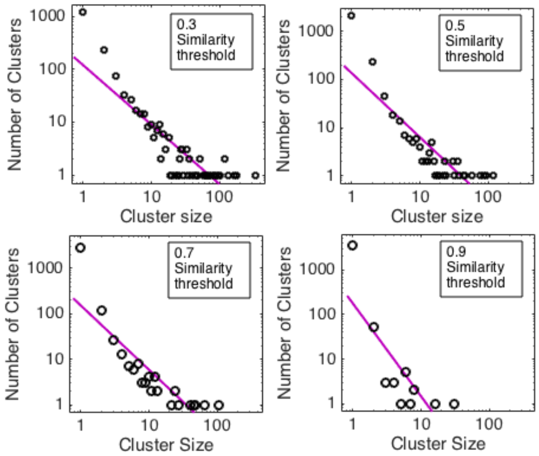

Figure 3: Distribution of the number of clusters per cluster size.

lower thresholds, many clusters are incorrectly merged, which will affect the precision of the reconstruction. On the other hand, we see that if the threshold is set too high (e.g., 0.9), that the larger clusters are split, resulting in missed provenance relationships – and thus affecting the recall. In all cases above 0.3, we see that the number of singletons does not vary significantly, which means that messages that do not belong together will most likely not be clustered together, regardless of the similarity threshold. While it is too early to make a definite decision regarding the optimal threshold without a content-based evaluation, these results lead us to expect that the optimum will be somewhere in the vicinity of 0.7. Using this threshold (0.7), we generated a set of 3094 clusters, and used the 206 non-singletons to reconstruct 879 provenance relationships (32 quotations and 847 revisions).

In other words, when we integrate this high-level provenance with the 31 cascades discovered by the low-level provenance reconstruction, we effectively introduce 879 new connections that were previously unidentified. This creates much larger graphs for the consumers of the provenance data to analyse, and provides an enriched view on the information diffusion process. The entire reconstructed provenance graph can be downloaded at `http://semweb.mmlab.be/ns/prov-said/cikm2015.ttl`

5. CONCLUSION AND FUTURE WORK

We proposed a method to reconstruct and integrate provenance on two levels of granularity: low-level through information cascades, and high-level through similarity-based clustering. This method augments the provenance of messages on social media, especially when there is external influence not deriving from one single source (in our case: Twitter) or for copied messages that do not give credit to their initial contributors. In these cases, an obvious influencer is not exposed by the social media API. Such messages do not produce large cascades resulting in low-level provenance, but are clustered together in the high-level provenance reconstruction of our approach.

For future work, we will extensively evaluate our approach on diverse datasets and combined data from different social media. Additionally, we will improve our method by applying more suitable metrics of message similarity for micropost text.

6. REFERENCES

[1] M. Al Hasan, S. Salem, and M. J. Zaki. Simclus: an effective algorithm for clustering with a lower bound on similarity. *Knowledge and information systems*, 28(3):665–685, 2011.

[2] G. Barbier, Z. Feng, P. Gundecha, and H. Liu. Provenance data in social media. *Synthesis Lectures on Data Mining and Knowledge Discovery*, 4(1):1–84, 2013.

[3] T. De Nies, S. Coppens, D. Van Deursen, E. Mannens, and R. Van de Walle. Automatic discovery of high-level provenance using semantic similarity. In *Provenance and Annotation of Data and Processes*, pages 97–110. Springer, 2012.

[4] T. De Nies, S. Coppens, R. Verborgh, M. Vander Sande, E. Mannens, R. Van de Walle, D. Michaelides, and L. Moreau. Easy access to provenance: an essential step towards trust on the web. In *COMPSACW*, 2013.

[5] A. Dimou, M. Vander Sande, P. Colpaert, R. Verborgh, E. Mannens, and R. Van de Walle. RML: a generic language for integrated RDF mappings of heterogeneous data. In *Proceedings of the 7th Workshop on Linked Data on the Web (LDOW2014), Seoul, Korea*, 2014.

[6] Z. Feng, P. Gundecha, and H. Liu. Recovering information recipients in social media via provenance. In *ASONAM*, pages 706–711, 2013.

[7] A. Guille, H. Hacid, C. Favre, and D. A. Zighed. Information diffusion in online social networks: A survey. *ACM SIGMOD Record*, 42(2):17–28, 2013.

[8] P. Gundecha, Z. Feng, and H. Liu. Seeking provenance of information using social media. In *CIKM*, 2013.

[9] P. Gundecha, S. Ranganath, Z. Feng, and H. Liu. A tool for collecting provenance data in social media. In *KDD*, pages 1462–1465, 2013.

[10] J. Leskovec, L. A. Adamic, and B. A. Huberman. The dynamics of viral marketing. *ACM TWEB*, 1(1):5, 2007.

[11] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. In *KDD*, pages 497–506, 2009.

[12] L. Moreau, P. Missier (Eds.), and W3C Provenance Working Group. PROV-DM: The PROV Data Model. W3C, 2013.

[13] M. P. Simmons, L. A. Adamic, and E. Adar. Memes Online: Extracted, subtracted, injected, and recollected. *ICWSM*, 11:17–21, 2011.

[14] I. Taxidou, T. De Nies, R. Verborgh, P. M. Fischer, E. Mannens, and R. Van de Walle. Modeling information diffusion in social media as provenance with W3C PROV. In *Proceedings of the 24th International Conference on WWW Companion*, pages 819–824, 2015.

[15] I. Taxidou and P. M. Fischer. Online analysis of information diffusion in twitter. In *Proceedings of the 23rd International Conference on WWW Companion*, pages 1313–1318, 2014.