ProvDIVE: PROV Derivation Inspection and Visual Exploration

Sven Lieber [1], Io Taxidou[2], Peter M. Fischer[2], Tom De Nies[1], Erik Mannens[1]

[1] Ghent University - imec - IDLab, Belgium
[2] University of Freiburg, Germany
{sven.lieber, tom.denies, erik.mannens}@ugent.be
{taxidou,peter.fischer}@cs.uni-freiburg.de

Motivation In a previous work, we presented a method to reconstruct *PROV* derivations from short social media messages [1]. This method can capture a wide range of information spreading (and thus influence) among users, from explicit attribution like quoting to implicit means like content similarity. When applying this method to real-life datasets containing several million messages (e.g., a popular event), we are creating derivations in the same order of magnitude. To assess the provenance, it is useful to manually inspect the overall structure, the individual derivations and the users involved. Such tasks can be supported well by visualization techniques, yet thousands to millions of nodes are notoriously difficult to visualize [4].

Our Approach To the best of our knowledge, no provenance visualization tool exists that provides means to inspect the full range from a high-level abstraction (to provide an overview over the large volume of derivations) to a detailed view on the context of individual derivations (to provide a detailed understanding of an individual derivation). We aim to visualize such derivations in the following dimensions: 1) filtering of nodes and edges according to their attributes (user, time, type) 2) (partially) summarizing the graph for visualizing the structure 3) marking and filtering the (sub-)graph according to the classification of the derivation graph structures 4) presenting contextual information on individual derivations.

Poster Description This poster aims to gather early feedback regarding the proposed visualization methods of our social media use-case. To tackle the first dimension of filtering redundant messages, we leverage the ideas of Seltzer and Macko [4] for node selection. Not all messages might be interesting for the user, e.g. only social media messages within a specific timespan, from a specific author or containing specific keywords might be relevant. Likewise, we can filter derivation edges based on aspects like attribution type or confidence/strength as well as entire subgraphs/components on aspects like topics or sizes.

For the second dimension of visualizing large graphs of derivations, we need to employ some graph summarization techniques. Summarizing provenance graphs of large conversations is a well-known problem, and different solutions have been proposed [2–4]. The local clustering method by Macko et al. [3] uses influence metrics and thus fits our social media use-case. Messages are clustered among their derivation ancestors, based on different influence metrics, until a certain threshold is reached. Each cluster, visualized as a node, represents then all its members. To indicate the importance of each cluster, the node size can be proportional to the amount of members. Depending on the graph, the local clustering summarization might not be a viable solution, so that, e.g., node aggregation based on the node derivation history [2] may be more effective. Most datasets also contain many small conversations which require different approaches to summarize. One obvious way to abstract these is visualizing one node for each chain or tree, or more effectively, one node representing all derivation chains and trees with similar size or content.

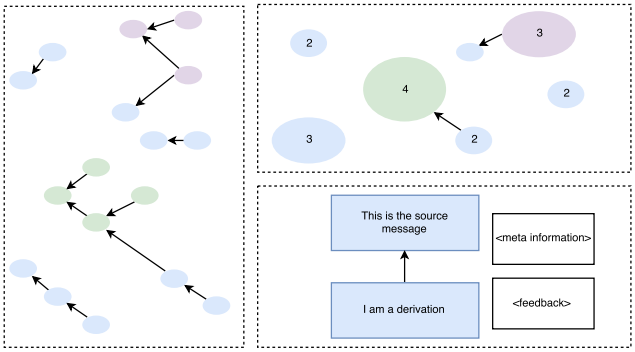

Figure 1: A whole derivation graph left (Each node respresents one message). A possible summarization on the top right (The node sizes and numbers indicate the amount of messages). A sketch of a possible manual inspection on the bottom right, which displays meta information to assist the user in verifying the derivation and a feedback panel.

Different shapes of derivation (sub-)graphs such as simple chains, trees with multiple branches or non-tree like structures (for example, Figure 1 on the left) imply different types of diffusion and are therefore interesting for social media researchers. As a result, filtering (sub-)graphs based on structural information as well as marking (sub-)graphs/summaries with such information provides useful insights.

At the level of individual derivations the context should not only show both the source and the derived messages, but also additional information, encoded as derivation properties. This information can be anything which assists the user in verifying the derivation (e.g. relationship of the authors). Towards pure visualization, the detailed derivation view can be extended to include a feedback area, where the user can include assessment (Fig. 1 bottom right).

Conclusion Reconstructing *PROV* derivations from real-life data generates massive datasets that require specialized visualizations. We consider four complementary dimensions to visualize the reconstructed derivations: filtering, summarizing, structural classification and contextual information on individual derivations.

References

[1] T. De Nies et al. Towards Multi-level Provenance Reconstruction of Information Diffusion on Social Media. *CIKM*, 2015. .

[2] X. Li and X. Xu. Interactive Provenance Summaries for Reproducible Science. *IEEE 12th International Conference on e-Science*, 2016.

[3] P. Macko et al. Local Clustering in Provenance Graphs (Extended Version). 2013.

[4] M. Seltzer and P. Macko. Provenance Map Orbiter: Interactive Exploration of Large Provenance Graphs. *TaPP'11*, 2011.