Social **Data Provenance**

Zyad Al-Azazi - Mir Wise Khan

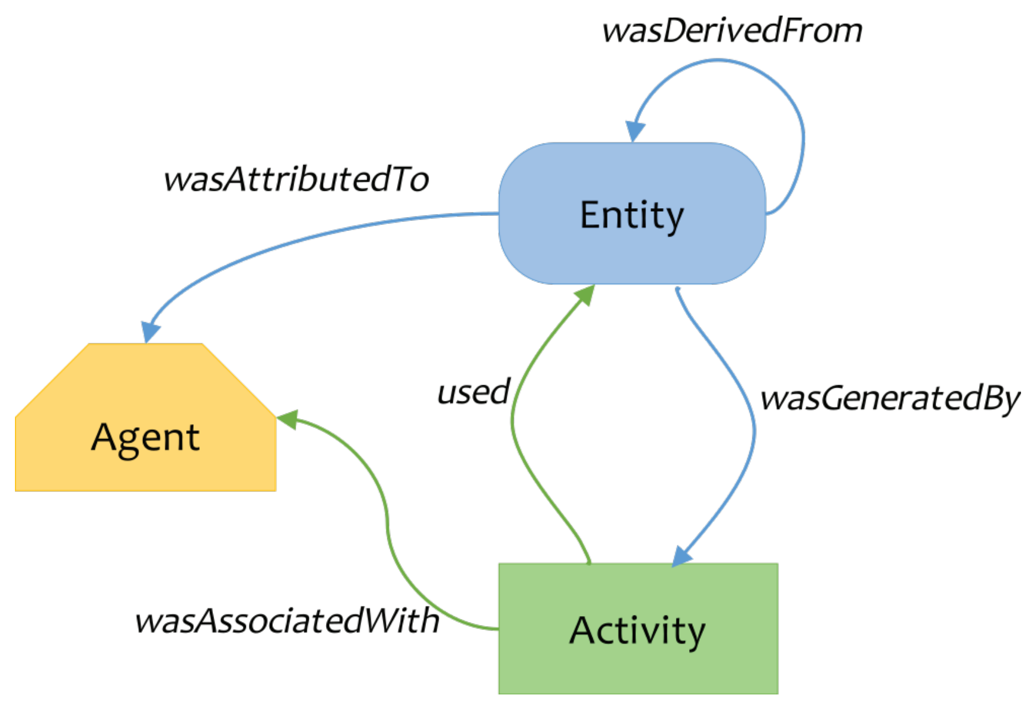
EBISS 2023 - Barcelona

Background & Motivation

With the increasing volume of user-generated data on social media, tracing the data transferred on social media platforms and evaluating its trustworthiness has become a very complicated task with serious implications on our communities. The process by which a piece of data origins are traced is called **Data Provenance**. A term that is used to refer to information that traces how a certain state of data was obtained with respect to its origins, ownership, transformations and movement. It serves as an important process for evaluating trustworthiness of data, reproducing a certain data state as well as providing means for transparency and interpretability.

A very specific branch of data provenance that is primarily concerned with capturing and representing the origins, derivation processes, relationships and history of data on social networking sites is referred to as **Social Data Provenance**. This branch concentrates more on evaluating the trust of shared data, understanding information dissimentation patterns and determining ownership.

PROV Data Model



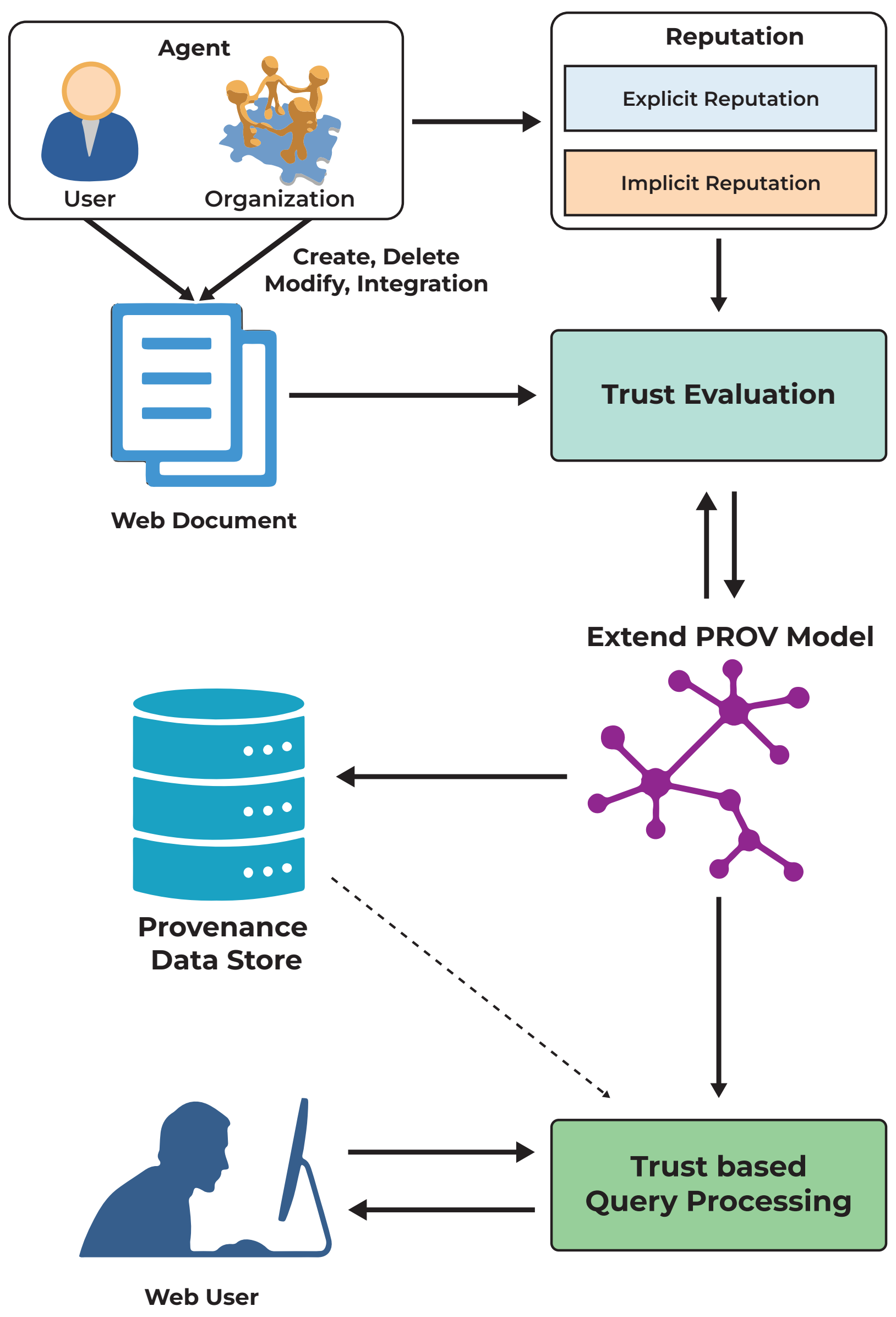
Provenance as of NOW

TRUST EVALUATION SYSTEMS

Trust Evaluation of Multimedia Documents on Social Semantic Web

This approach utilises knowledge graphs by extending the existing PROV model by adding subclasses to the main classes. The trust scheme is automated and it incorporates both explicit and implicit reputations.

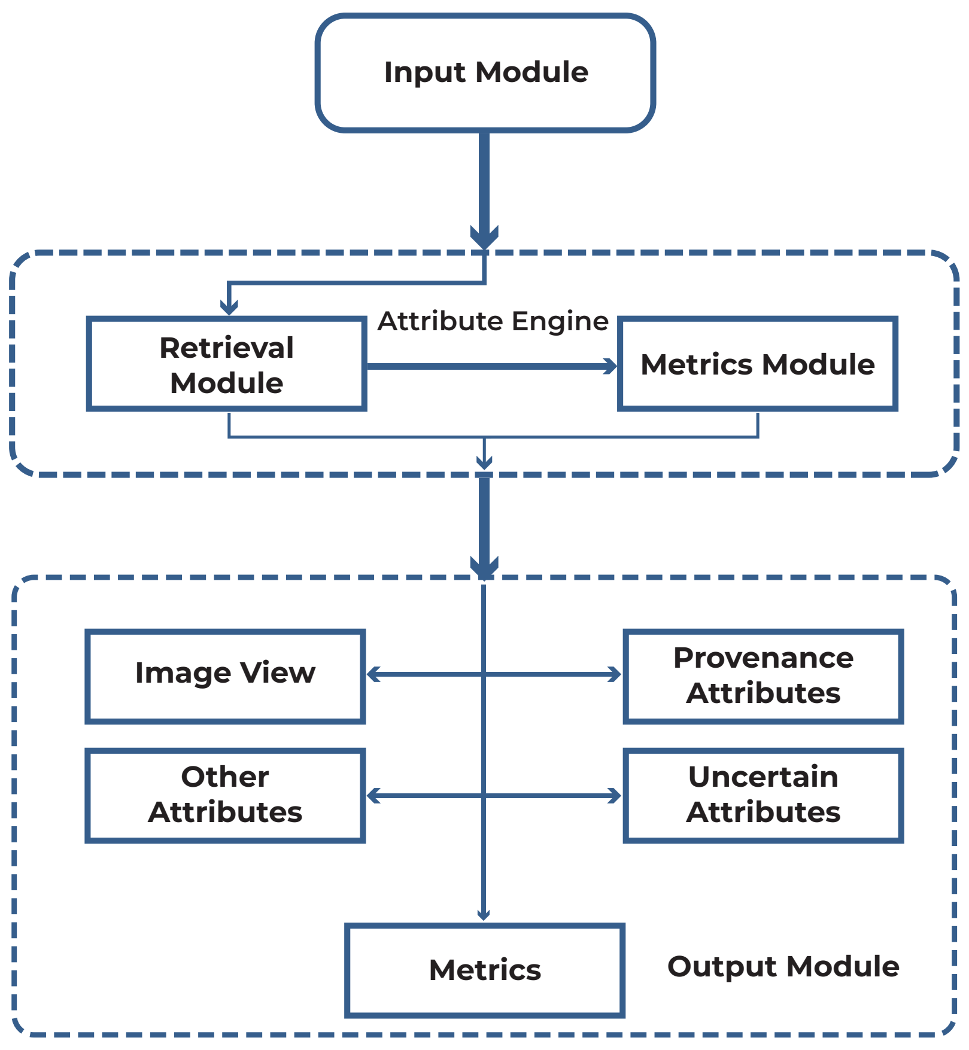
Evaluation Metrics: A comparitive approach to existing results.



Web-based Tool for User Provenance Data

This tool is focused more on the collection of user provenance data from different sources on the internet to facilitate the task of a user to subjectively assess the validity of information they counter.

Evaluation Metrics: Information Availability, Information Legitimacy & Retreival Speed.

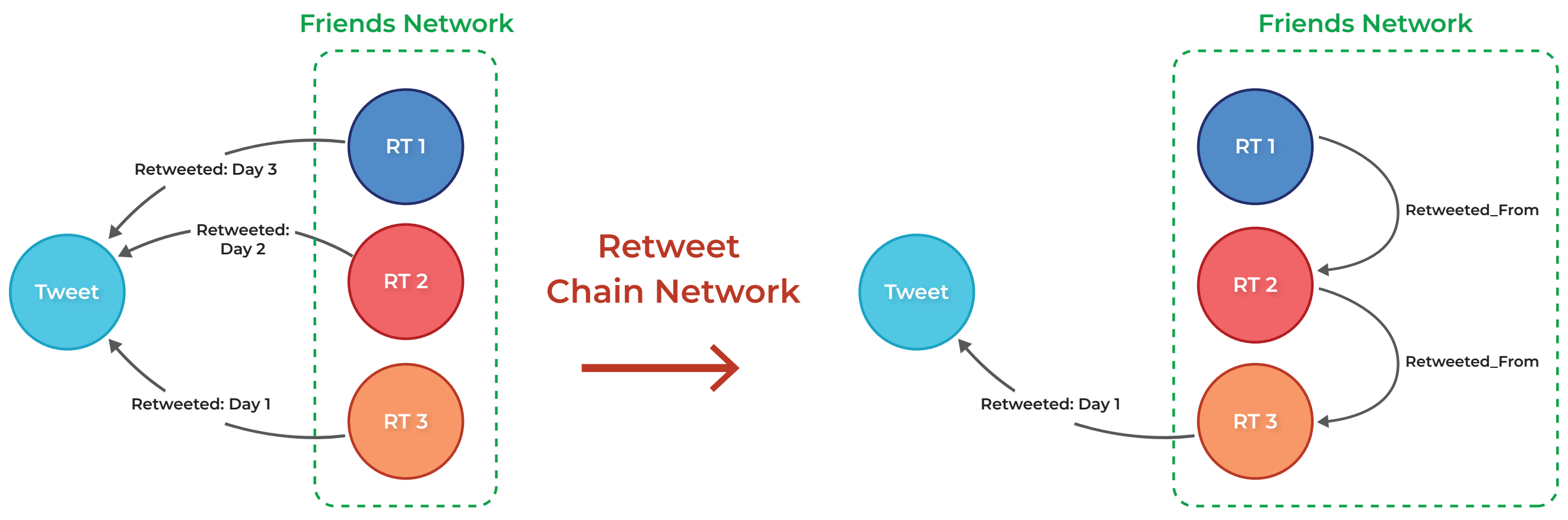


TWEET TRACKING SYSTEMS

Tracking Social Provenance in Chains of Retweets

A formally-defined approach that focuses on implicit provenance and utilizes Provenance Constraint Networks to approximate the ownership percentage of tweets and predict retweet chains in terms of how content reaches users before it is retweeted.

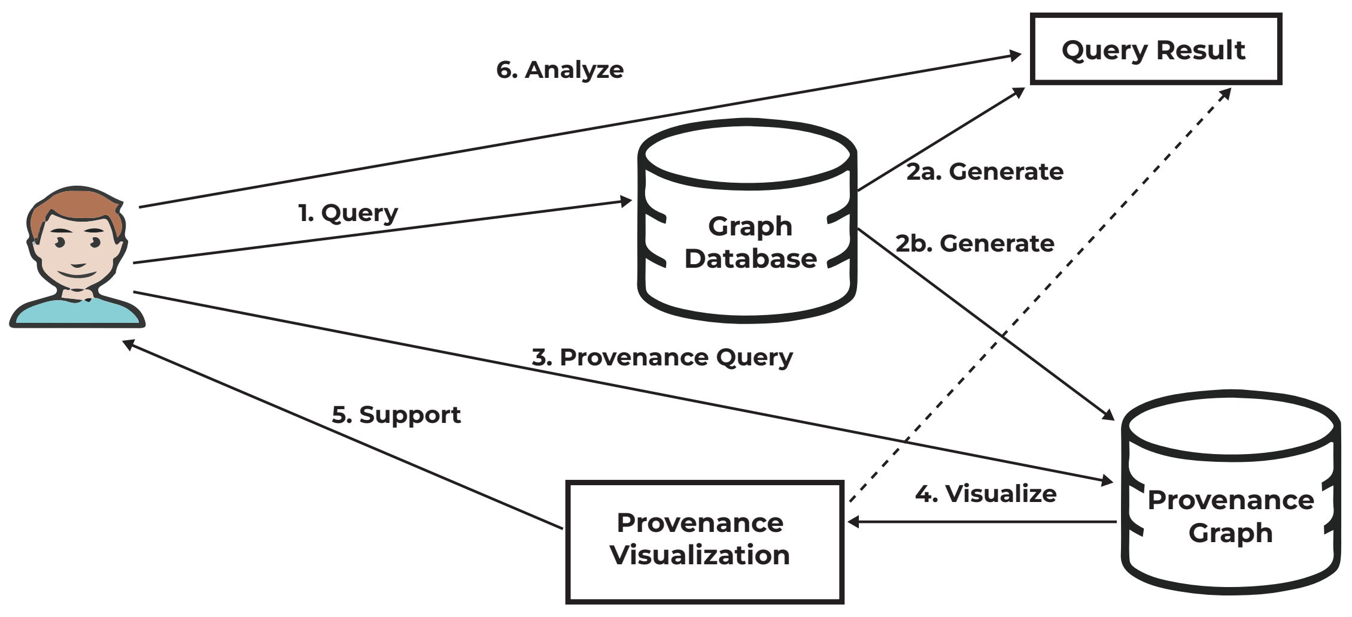
Evaluation Metrics: Average Constraint Size, Sparse Node Incidence & Retweet Source Incidence.



Zero-Information Loss Graph DB Architecture

This approach relies on property graphs to store the main data, in addition to the provenance data being stored in provenance graphs. It offers querying and visualization of the provenance graphs whose data is generated on tuple-level.

Evaluation Metrics: Data querying performance and provenance graph generation performance.



Conclusion

- Approaches are primarily driven by use cases and that dictates the metrics used in evaluating the approaches as well.
- Time sensitivity and versioning requirements demand for highly scalable solutions.
- Future research could further focus on standardizing the performance metrics for social data provenance approaches.

References

[1] Y. Tas, M. Baeth, and M. Aktas, “An approach to standalone provenance systems for big social provenance data,” 08 2016, pp. 9–16.

[2] A. Rani, N. Goyal, and S. K. Gadia, “Social data provenance framework based on zero-information loss graph database,” Social Network Analysis and Mining, vol. 12, no. 72, pp. 1–11, 2022. [Online].

[3] P. Gundecha, S. Ranganath, Z. Feng, and H. Liu, “A tool for collecting provenance data in social media,” 08 2013, pp. 1462–1465.

[4] K. Bok, S. Yoon, and J. Yoo, “Trust evaluation of multimedia documents based on extended provenance model in social semantic web,”
Multimedia Tools and Applications, vol. 78, 10 2019.

[5] S. Migliorini, M. Gambini, E. Quintarelli, and A. Belussi, “Tracking social provenance in chains of retweets,” Knowledge and Information Systems, pp. 1–28, 05 2023.