Social Web-scale Provenance in the Cloud

Yogesh Simmhan[1] and Karthik Gomadam[2]
[1] Microsoft Research, yoges@microsoft.com
[2] University of Southern California, gomadam@usc.edu

Abstract. The lower barrier to entry for users to create and share resources through applications like Facebook and Twitter, and the commoditization of social Web data has heightened issues of privacy, attribution, and copyright. These make it important to track the provenance of social Web data. We outline and discuss key engineering, privacy, and monetization challenges in collecting and analyzing provenance of social Web resources.

Keywords: Provenance, social web, scalability, privacy, Cloud.

1 Introduction

The pervasiveness of social networks as an intrinsic part of users' online presence allows easy sharing of information among peers and the public at large. Social network services such as Facebook allow sharing of *free form* comments, *semi-structured* hash-tags, and *resources* like images with other users. External applications can query these user relationships and content through APIs, and publish feeds.

The ease of such portals that cause their success also masks privacy issues that have unintended consequences for the users [1,4], ranging from plagiarism to misrepresentation. Tracking the provenance of shared social network resources becomes crucial, yet challenging, given the *ad hoc* nature of the sharing model and the scalability needed [6]. For e.g., an indie artist who shares a soundtrack on Facebook may like to find out which of her friends or friends of friends (FOAFs) downloaded the music or published "similar" albums on finding remixed versions of her work [2].

Provenance in social networks pose additional challenges to those in workflows and databases, viz. (1) Identifying resources, relationships, and semantics from unstructured information, (2) Online scaling with social network size, frequency of feed updates, and popularity, (3) Ensuring privacy of aggregated provenance, and (4) Incentivizing service use and revenue given the expectations of free online services.

Provenance information for resources on the social web can be characterized as:

1. **Resource provenance**: This traces the creation, publishing, reuse, and deletion of social data artifacts like media and documents identified by URI/URLs.
2. **Social provenance**: This describes social operators such as "Like", "Comment", and "Share" applied to resources to track activities beyond its creation and reuse. Recording social relationships like Friends and FOAFs, over time, is also needed.
3. **System provenance**: This includes access statistics, download history and site metrics of the resource automatically and passively tracked by the social network.

1.1 Engineering and Scalability Challenges

A provenance system for the social Web involves: (1) *Integration* of provenance from social networks, and (2) *Subscription* to query provenance features of interest.

Architecture: Our proposed architecture employs a publish-subscribe model for aggregating social network feeds from users based on *Pubsubhubbub* [7]. The aggregator identifies resource entities and relationships in the feed using unstructured and structured content – a challenging research problem, and integrates it with prior provenance that is enhanced with specific resource metadata pulled from the network.

Provenance is accessed through user queries performed either on the feeds in near-realtime, or on the aggregated provenance and metadata. The former standing queries [9] have timeliness but restrict query attributes. The latter provide richer query terms but is performed once or triggered on a schedule. The queries may be as broad as requesting all updates to a resource or use heuristics to identify similar resources in a FOAF network. The query results can themselves be pushed as feeds.

Scaling: Pubsubhubbub is a scalable protocol for publishing feeds and query results. However, storage and query over the aggregated provenance has to scale too. *Cloud computing* provides a model for scaling the aggregator and querying hubs. SQL Azure [8] databases hosted in Virtual Machines (VMs) can store provenance metadata and scale on demand as the number of users increase. The metadata is partitioned across VM instances based on tight linkages in the friend network to ensure metadata locality – trading better query performance within closely linked friends for costlier access to distant friends or the public. Using a carousel approach that batches scheduled queries across users and scans tables can also achieve load balancing. Scaling with the rate of feeds and the number of queries is also key as service or resource popularity increases (e.g. a leaked music video). On demand scale-out by Clouds combined with dynamic repartitioning of stored provenance can address this.

1.2 Privacy & Monetization

Tracking provenance benefits privacy preservation and in determining its compromise. Awareness of who is *actually* viewing your resource can help detect incorrect privacy configurations, and bridge perceived and actual privacy. One challenge is to collect the provenance *transparently* with user opt-in, rather than giving the sense of yet another privacy invasion and harvest of personal information.

Another aspect of privacy is the *social granularity* of collected provenance. The provenance service can be a private service for groups of friends who sign up and track resources they publish. The group can even own the hosted provenance service and data in the Cloud using their own account – paying for the *private provenance service* and Cloud resources and ensuring no third party mines it. The diminished cost per user for the Cloud service as more users join is an incentive for FOAFs to join.

The above approach can create disconnected provenance repositories for each user group. These can be linked by exporting provenance through standards like Open Provenance Model [5] and those evolving in the W3C Provenance Incubator [3]. Else, third parties can provide a more connected, shared service across users in exchange for payment, or for a free, Ad supported model that mines accumulated provenance.

1.4 Related Work

Social networks provide users with system provenance on resource creation and use to popularize their access. Tools studied to reuse data in social networks and blogs enhance the metadata of republished resources with semantic annotations on their provenance to support accountability and enforce usage right policies [2]. Others have combined provenance of user assertions with their social network links to gauge the trust rating of the assertion, used in movie recommendation systems [1]. The W3C Provenance and Social Web Incubator groups have identified requirements of provenance for social networks and surveyed technologies that can address them [3].

2 Conclusion

The commoditization of social Web data increases concerns about privacy and resource sharing in social networks that can be addressed through tracking provenance of social Web resources. The issues around such provenance has a different quality from provenance collection for workflows and databases, both due to the fungible nature of the data and the scales involved. Our article highlights some of these issues and proposes an architecture for addressing this in part. These form the basis for further investigation into this important and emerging area of research.

Acknowledgments. The authors thank members of the W3C Provenance Incubator and Social Web Incubator groups for discussions that motivated some of these issues.

References

1. Combining Provenance with Trust in Social Networks for Semantic Web Content Filtering. Jennifer Golbeck, *International Provenance and Annotation Workshop (IPAW)*, LNCS Volume 4145/2006.
2. Data Republishing on the Social Semantic Web. Claudia Wagner and Enrico Motta, *Workshop on Trust and Privacy on the Social and Semantic Web*, 2009.
3. Requirements for Provenance on the Web. James Cheney, Yolanda Gil, Paul Groth (Editor), and Simon Miles, Working Report by the *W3C Provenance Incubator Group*, April 9, 2010.
4. Trust and privacy concern within social networking sites. Catherine Dwyer, Starr Roxanne Hiltz and Katia Passerini, *Americas Conference on Information Systems*, 2007.
5. The Open Provenance Model - Core Specification (v1.1). Luc Moreau, Ben Clifford, Juliana Freire, Yolanda Gil, Paul Groth, Joe Futrelle, Natalia Kwasnikowska, Simon Miles, Paolo Missier, Jim Myers, Yogesh Simmhan, Eric Stephan, and Jan Van den Bussche, *Future Generation Computer Systems*, 2010. (In Press)
6. The Foundations for Provenance on the Web. Luc Moreau, *Foundations and Trends in Web Science*, 2009. (Submitted)
7. Pubsubhubbub: A simple, open, web-hook-based pubsub protocol. http://code.google.com/p/pubsubhubbub/, (4 Jun, 2010).
8. Microsoft SQL Azure. http://www.microsoft.com/windowsazure/sqlazure/, (4 Jun 2010).
9. Data Stream Query Processing. Nick Koudas and Divesh Srivastava, *International Conference on Data Engineering (ICDE)*, 2005.