Privacy-Preserving Geo-distributed Graph Processing

M2R Internship 2017-2018

Supervised by : Tristan Allard (Univ. Rennes 1/Irisa) & Shadi Ibrahim (Inria/LS2N)

External collaborators : Amr El Abbadi (Univ. California Santa Barbara) & Amélie Chi Zhou (Univ. Shenzhen)

Contact : tristan.allard@irisa.fr and shadi.ibrahim@inria.fr

Keywords: graph processing, differential privacy, encryption, security, big data, distributed systems.

Graph processing is an emerging computation model for a wide range of applications, such as social network analysis [3], natural language processing [7] and web information retrieval [6]. Many graph applications, such as social networks, involve large sets of data spread in multiple geographically distributed (geo-distributed) datacenters (DCs). For example, Facebook receives terabytes of text, image and video data everyday from users around the world [1]. In order to provide reliable and low-latency services to the users, Facebook has built four geo-distributed DCs to maintain and manage those data. Thus, to run graph processing algorithms on top of those data, for example, studying the relationships between Facebook users using shortest path algorithms, answering user queries using pattern matching algorithms, or finding the most influential person based on centrality measures, it requires the coordination (e.g., data exchange) of multiple geo-distributed DCs. The geo-distributed DCs are usually located at different countries and it is very likely that the regulations and levels of privacy guarantees differ in those countries. For example, in 2015, Europe's highest court invalidated a longstanding data-sharing agreement between the EU and the U.S., which prevents many U.S. companies such as Facebook from freely transferring its European users' personal data between Europe and the U.S. [2]. This shows that it is important not to reveal data to other DCs during graph processing.

The goal of this internship is precisely to design and validate an algorithm that allows multiple coordinating DCs to compute a graph processing function such as, e.g., the pagerank centrality measure, while preserving the privacy of individual data. The algorithm will follow the widely-used GAS graph processing model proposed in PowerGraph [4], which iteratively executes user-defined vertex computations until convergence. From the side of privacy, the DCs will be considered honest-but-curious (they do not deviate from the algorithm but infer anything that can be inferred). Stronger attackers may be considered. The algorithm will be protected by satisfying the *edge differential privacy* model [5]. Loosely speaking, edge differential privacy hides the presence/absence of any single edge by adding noise to the information disclosed during the computation. The additional use of encryption scheme(s) can be considered. The security, efficiency, and quality of the algorithm will be validated theoretically and experimentally. The experiments will be carried out using large-scale scientific testbed (i.e., Grid'5000) or using public clouds (e.g., Microsoft Azure or Amazon EC2) over publicly available dataset (e.g., the Stanford Large Network Dataset Collection[1]). Depending on the completion of the work, this work could lead to the publication of a research article.

The internship will take place in Rennes. It is co-supervised by Tristan Allard (Druid team, Univ. Rennes 1 / Irisa) and Shadi Ibrahim (Ascola team, Inria / LS2N). Two external collaborators will participate to the work : Amr El Abbadi (Univ. California Santa Barbara) and Amélie Chi Zhou (Univ. Shenzhen). The ideal intern will be creative, autonomous, hard-working, and curious. Pursuing by a PhD thesis may be considered depending on the results of the internship.

[1] http://snap.stanford.edu./data/index.html

References

[1] Scaling the Facebook data warehouse to 300 PB. https://goo.gl/Eyv6o3, 2014.

[2] The court of justice declares that the commission's us safe harbour decision is invalid. https://goo.gl/vLg6aw, 2015.

[3] A. Ching, S. Edunov, M. Kabiljo, D. Logothetis, and S. Muthukrishnan. One trillion edges: Graph processing at facebook-scale. volume 8, pages 1804–1815, 2015.

[4] J. E. Gonzalez, Y. Low, H. Gu, D. Bickson, and C. Guestrin. Powergraph: Distributed graph-parallel computation on natural graphs. In *OSDI'12*, pages 17–30.

[5] V. Karwa, S. Raskhodnikova, A. Smith, and G. Yaroslavtsev. Private analysis of graph structure. *ACM Trans. Database Syst.*, 39(3):22:1–22:33, Oct. 2014.

[6] E. Minkov, W. W. Cohen, and A. Y. Ng. Contextual search and name disambiguation in email using graphs. In *SIGIR '06*, pages 27–34, 2006.

[7] L. Zhu, A. Galstyan, J. Cheng, and K. Lerman. Tripartite graph clustering for dynamic sentiment analysis on social media. In *SIGMOD '14*, pages 1531–1542.