Probabilistic Approaches for Big Data Analysis

Main advisor: Philippe Nain (Inria, HDR), Maestro team, will hire the candidate
Coadvisor: Fabrice Huet (I3S), Scale team

Summary

We believe that probability is going to play a major role in the future approaches to distributed Big Data Analysis for two main reasons. The first one is the possibility to reduce the size of the input data each machine has to process through randomization. The second one is related to the fact that many data sets are intrinsically probabilistic for a variety of reasons.

The purpose of this thesis is to investigate the potential of probabilistic approaches for distributed Big Data Analysis with a special focus on large graphs processing with a large number of commodity machines.

Detailed description

Traditional approaches to large data processing relied on parallel and distributed algorithms, but the involved data scale poses challenges to deterministic algorithms. For this reason, an increasing interest on randomized distributed algorithms has raised in the research community [Calafiore09,Ishii10]. Indeed, as stated in the preface to the now classic Motwani and Raghavan's book “Randomized Algorithms”, “for many applications a randomized algorithm is the simplest algorithm available, or the fastest, or both.”

In current approaches randomization has been used with the main goal to reduce the number of operations of an algorithm. For many data sets, only linear or sublinear algorithms can be of a practical use and only randomized algorithms can achieve sublinear bounds [Avrachenkov12].

However the number of operation is only one of the factors contributing to the total computation time. Access to disks and data transfer across different machines are often the bottlenecks. This is even more true when the computing infrastructure is made by a large number of commodity machines as it is in our reference scenario: even sublinear computation costs may be of no use if data cannot be efficiently stored in the memory or a large amount of data has to be transferred across the machines. How to efficiently divide the dataset across the machines in order to avoid such performance losses is of the maximum importance and particularly challenging. We believe that randomization can offer a practical solution also to this problem. An example of such approach is in [Neglia13] where the authors use an edge sampling procedure on the complete Twitter graph (more than 400GB), in order to produce some much smaller subsamples (at least one order of magnitude) that can be efficiently transferred and processed in parallel. A starting point of the thesis is to investigate what graph metrics can be calculated in a distributed way by relying on random edge or node sampling. The long term objective is to generalize the idea to different random sketches of the graph, in the same spirit of what has been successfully done for matrix analysis [Mahoney10].

While in randomized algorithms the input data are usually considered deterministic, a new challenge in Big Data Analysis comes from the fact that the input is itself probabilistic. This may be due to a variety of reasons: noise intrinsically due to a measurement process, noise artificially added to guarantee some privacy [Dwork07] or ambiguity of queries in natural languages. In the last seven years there has been a new research trend on the probabilistic databases [Dalvi07], whose focus is on the results of queries. This literature mainly considers centralized approaches and we plan to study the new challenges raising in a distributed environment. Usually distributed process of data relies on deterministic algorithms to efficiently split the data stream across the different machines available [Pellegrino13]. When considering probabilistic data, this process becomes much more challenging. We believe that it is very difficult to perform this task without taking explicitly into account the probabilistic nature of the data.

For this reason the distribution problem is going to be the first problem the student will address in this research axis.
The theoretical analysis will only be one contribution if this thesis, but an equally important output will be a software framework for the analysis of large data graphs. This framework could be of practical use also for other teams in SophiaTech campus working on similar data.

Interdisciplinarity and novelty

The research goals are at the crossing of the scientific interests and backgrounds of the two advisors: Philippe Nain is an expert on probability and stochastic processes, Fabrice Huet is an expert on distributed systems for large scale data processing. Their joint backgrounds will help the student to design algorithms with guaranteed theoretical performance and efficiently implement them in a real environment. If this subject is funded, it would be a completely new cooperation between the two teams Philippe and Fabrice belong to. Moreover there has been no ongoing research activity between these teams since 2009.

Candidate

We have already selected an excellent candidate for this topic. We attach to the present application his cv together with his university marks.
We highlight a few points:

- he was one of the best students at Karazin Kharkiv National University, Ukraine, from which we have hired two other very good PhD students,
- since October 2013 he is working in Maestro team with a pre-PhD contract, and he is doing a very good job studying the performance evaluation of BitTorrent-enhanced satellite missions,
- he had a previous experience in research with two papers already published at two international conferences.

References

[Calafiore09] G. C. Calafiore, *Distributed randomized algorithms for probabilistic performance analysis*, Systems & Control Letters, Volume 58, Issue 3, March 2009, Pages 202-212
[Ishii10] Hideaki Ishii, Roberto Tempo, *Distributed Randomized Algorithms for the PageRank Computation,* IEEE Transactions on Automatic Control, 55: 1987-2002, 2010
[Mahoney10] Michael W. Mahoney, *Randomized algorithms for matrices and data*, NOW Foundations and Trends in Machine Learning, Vol 3, Issue 2, 2010, pp 123-224
[Akbarinia13a] R. Akbarinia, P. Valduriez, and G. Verger, *Efficient Evaluation of SUM Queries over Probabilistic Data*, IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 4, pp. 764-775, April 2013
[Akbarinia13b] R. Akbarinia and Florent Masseglia, *Fast and Exact Mining of Probabilistic Data Streams Reza*, Machine Learning and Knowledge Discovery in Databases, Lecture Notes in Computer Science Volume 8188, 2013, pp 493-508
[Avrachenkoc12] K. Avrachenkov, N. Litvak, M. Sokol and D. Towsley, *Quick detection of nodes with Large Degrees*, 9th International Workshop on Algorithms and Models for the Web Graph, WAW 2012.
[Dalvi07] N. Dalvi and D. Suciu. *Efficient query evaluation on probabilistic databases*. The VLDB Journal, 16:523–544, October 2007.
[Dwork07] C. Dwork, *Differential privacy,* in Automata, Languages and Programming, ser. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2006, vol. 4052, pp. 1–12.
[Pellegrino13] L. Pellegrino, F. Huet, F. Baude and A. Alshabani, *A Distributed Publish/Subscribe System for RDF Data*, Data Management in Cloud, Grid and P2P Systems (2013) 39-50

[Neglia13] G. Neglia, X. Ye, M. Gabielkov, A. Legout, *How to Network in Online Social Networks*, Inria Research Report 8423, December 2013.