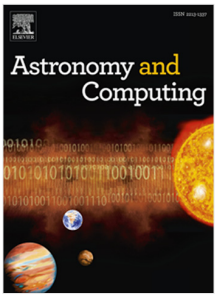
Accepted Manuscript



Enabling data science in the Gaia mission archive: The present-day mass function and age distribution

Daniel Tapiador, Ángel Berihuete, Luis M. Sarro, Francesc Julbe, Eduardo Huedo

PII: S2213-1337(16)30083-X
DOI: http://dx.doi.org/10.1016/j.ascom.2017.02.001
Reference: ASCOM 177

To appear in: *Astronomy and Computing*

Received date : 12 August 2016
Accepted date : 23 February 2017

Please cite this article as: Tapiador, D., et al., Enabling data science in the Gaia mission archive: The present-day mass function and age distribution. *Astronomy and Computing* (2017), http://dx.doi.org/10.1016/j.ascom.2017.02.001

Enabling Data Science in the Gaia Mission Archive: The Present-Day Mass Function and Age Distribution

Daniel Tapiador[a,*], Ángel Berihuete[b], Luis M. Sarro[c], Francesc Julbe[d], Eduardo Huedo[a]

[a]*Departamento de Arquitectura de Computadores y Automática, Facultad de Informática, Universidad Complutense de Madrid, Spain*
[b]*Departamento de Estadística e Investigación Operativa, Universidad de Cádiz, Spain*
[c]*Departamento de Inteligencia Artificial, UNED, Madrid, Spain*
[d]*Departament d'Astronomia i Meteorologia ICCUB-IEEC, Barcelona, Spain*

Abstract

Recent advances in large scale computing architectures enable new opportunities to extract value out of the vast amounts of data being currently generated. However, their successful adoption is not straightforward in areas like science, as there are still some barriers that need to be overcome. Those comprise (i) the existence of legacy code that needs to be ported, (ii) the lack of high-level and use case specific frameworks that facilitate a smoother transition, or (iii) the scarcity of profiles with the balanced skill sets between the technological and scientific domains.

The European Space Agency's Gaia mission will create the largest and most precise three dimensional chart of our galaxy (the Milky Way), providing unprecedented position, parallax and proper motion measurements for about one billion stars. The successful exploitation of this data archive will depend on the ability to offer the proper infrastructure upon which scientists will be able to do exploration and modeling with this huge data set.

In this paper, we present and contextualize these challenges by building two probabilistic models using Hierarchical Bayesian Modelling. These models represent a key challenge in astronomy and are of paramount importance for the Gaia mission itself. Moreover, we approach the implementation by leveraging a generic distributed processing engine through an existing software package for Markov chain Monte Carlo sampling. The two computationally intensive models are then validated with simulated data in different scenarios under specific restrictions, and their performance is assessed to prove their scalability. We argue that this approach will not only serve for the models in hand but also for exemplifying how to address similar problems in science, which may need to both scale to bigger data sets and reuse existing software as much as possible. This will lead to shorter time to science in massive data archives.

Keywords: Scalable Data Science, Hierarchical Bayesian Analysis, Present-Day Mass Function, Present-Day Age Distribution, emcee Ensemble sampler, Gaia mission

1. Introduction

The massive amounts of data that the world produces every day pose new challenges to modern societies in terms of how to leverage their inherent value. Social networks, instant messaging, video, smart devices and scientific missions are just mere examples of the vast number of sources generating data every second. As the world becomes more and more digitalized, new needs arise for organizing, archiving, sharing, analyzing, visualizing and protecting the ever-increasing data sets, so that we can truly develop into a data-driven economy that reduces inefficiencies and increases sustainability, creating new business opportunities on the way [1].

Traditional approaches for harnessing data are not suitable any more as they lack the means for scaling to the larger volumes today available in a timely and cost efficient manner. This has somehow changed with the advent of Internet companies like Google or Facebook, which have devised new ways of tackling these issues. However, the variety and complexity of the value chains in the private sector as well as the increasing demands and constraints in which the public one operates, needs an ongoing research that can yield newer strategies for dealing with data, facilitate the integration of providers and consumers of information, and guarantee a smooth and prompt transition when adopting these cutting-edge technological advances.

Scientific data output is no exception to this data deluge and is currently increasing at 30% every year [2]. Some studies [3] conclude that the usage of existing scientific data sets decline 17% per year, with 80% of them being simply unavailable after 20 years. Given that a lot of research endeavours are nowadays publicly funded, more and more presure is being allocated to them in order to get an optimum return on investment, not only from the specific project outcome perspective, but also from the potential synergies produced when leveraging the results of other existing undertakings (and those to come). This is particularly the case in astronomy and astrophysics, where the expo-

*Corresponding author
Email addresses: dtapiador@gmail.com (Daniel Tapiador), angel.berihuete@uca.es (Ángel Berihuete), lsb@dia.uned.es (Luis M. Sarro), fjulbe@am.ub.es (Francesc Julbe), ehuedo@fdi.ucm.es (Eduardo Huedo)