**Δημόσια Παρουσίαση της Διδακτορικής Διατριβής της**

**κας.  Κλεισαρχάκη Σοφίας**

Την Δευτέρα 28 Νοεμβρίου 2016 και ώρα 14:00, τοπική ώρα Γαλλίας, στο αμφιθέατρο του κτηρίου IMAG, 700 Avenue Centrale, 38401 Saint-Martin-d’Hères του Université Grenoble Alpes θα γίνει η δημόσια παρουσίαση και υποστήριξη της Διδακτορικής Διατριβής της υποψήφιας διδάκτορας του Τμήματος Επιστήμης Υπολογιστών κας. Kλεισαρχάκη Σοφίας με θέμα:

**“Difference Analysis in Big Data: Exploration, Explanation, Evolution”**

**ABSTRACT**

Variability in Big Data refers to data whose meaning changes continuously. For instance, data derived from social platforms and from monitoring applications, exhibits great variability. This variability is essentially the result of changes in the underlying data distributions of attributes of interest, such as user opinions/ratings, computer network measurements, etc. *Difference Analysis* aims to study variability in Big Data. To achieve that goal, data scientists need: (a) measures to compare data in various dimensions such as age for users or topic for network traffic, and (b) efficient algorithms to detect changes in massive data. In this thesis, we identify and study three novel analytical tasks to capture data variability: *Difference Exploration, Difference Explanation* and *Difference Evolution*.

Difference Exploration is concerned with extracting the opinion of different user segments (e.g., on a movie rating website). We propose appropriate measures for comparing user opinions in the form of rating distributions, and efficient algorithms that, given an opinion of interest in the form of a rating histogram, discover agreeing and disagreeing populations. Difference Explanation tackles the question of providing a succinct explanation of differences between two datasets of interest (e.g., buying habits of two sets of customers). We propose scoring functions designed to rank explanations, and algorithms that guarantee explanation conciseness and informativeness. Finally, Difference Evolution tracks change in an input dataset over time and summarizes change at multiple time granularities. We propose a query-based approach that uses similarity measures to compare consecutive clusters over time. Our indexes and algorithms for Difference Evolution are designed to capture different data arrival rates (e.g., low, high) and different types of change (e.g., sudden, incremental). The utility and scalability of all our algorithms relies on hierarchies inherent in data (e.g., time, demographic).

We run extensive experiments on real and synthetic datasets to validate the usefulness of the three analytical tasks and the scalability of our algorithms. We show that Difference Exploration guides end-users and data scientists in uncovering the opinion of different user segments in a scalable way. Difference Explanation reveals the need to parsimoniously summarize differences between two datasets and shows that parsimony can be achieved by exploiting hierarchy in data. Finally, our study on Difference Evolution provides strong evidence that a query-based approach is well-suited to tracking change in datasets with varying arrival rates and at multiple time granularities. Similarly, we show that different clustering approaches can be used to capture different types of change.

Supervisor Uoc: Professor Vassilis Christophides

Supervisor Grenoble Alpes: Dr Sihem Amer-Yahia