Luca Zanetti
Department of Computer Science and Technology,
University of Cambridge
Email: luca.zanetti@cl.cam.ac.uk

December 8, 2019

Dear Members of the Hiring Committee,

Currently a Research Associate at the University of Cambridge, I am writing to apply for the position of Assistant Professor in Data Science. My research interests range over the fields of machine learning, algorithms, and discrete probability. I am particularly interested in designing efficient algorithms for massive data sets and large-scale networks. With my research I aim to contribute towards establishing the theoretical foundations of the emerging field of data science.

My PhD studies were mostly focused on the problem of graph clustering. Many data sets of practical interest can be naturally represented by a graph, and graph clustering is one of the main ways in which we can analyse and extract information from this data. More precisely, the goal of graph clustering is to partition a graph into clusters so that nodes belonging to the same cluster are more likely to be connected than nodes belonging to different clusters. In a co-authored COLT 2015 paper, which also appeared on the SIAM Journal on Computing 2017, I studied spectral clustering, probably the most popular algorithm for graph clustering. Despite its popularity and great practical performances, a rigorous analysis of spectral clustering had long been missing. My work provides the *first* theoretical justification of the performances of spectral clustering on a large family of graphs of practical relevance. In particular, I have established the following fact: whenever there is a large gap between two consecutive eigenvalues of the Laplacian operator of a graph, spectral clustering is able to uncover a meaningful clustering. Such a fact had long been known to practitioners, but it lacked a strong theoretical justification. This result accomplishes one of the main goals of my research: reducing the gap between the theory and practice of data science.

Another result from my PhD thesis, which has recently appeared in the ACM Transactions on Parallel Computing, deals with the problem of graph clustering in a distributed setting: the goal is to cluster a distributed network where each node is a computational unit and communication is allowed only between neighbouring nodes. With the ever increasing size and distributed nature of real world data sets, designing an efficient distributed graph clustering algorithm is of great practical relevance: my paper presents the first graph clustering algorithm with strong theoretical guarantees that works in such a highly distributed model of computation.

While my PhD studies have been focused on the problem of clustering in undirected graphs, many real world networks cannot be described by an undirected graph without a loss of information that might be crucial for certain applications. Moreover, networks are not always static entities, but develop and change over time. For this reason, I have recently been working on generalising and extending known techniques beyond undirected static graphs.

Some results in this direction have been collected in my co-authored ICALP 2019 paper. In this paper I investigate random walks, which are basic stochastic processes on graphs and have many applications in several different fields, data science and statistics included. Traditionally, most of the attention has been devoted to random walks on *static* graphs, graphs that do not change over time, while random walks on *dynamic* graphs, i.e., graphs that do change over time, have often been neglected. However, with the continuous rise in the importance of social networks, which are most faithfully modelled as dynamic graphs, the poor understanding of random walks on dynamic graphs is becoming increasingly problematic. My work significantly advances our knowledge on the topic. For example, I've shown that, whenever a technical condition is satisfied, random walks on dynamic graphs behave essentially the same as random walks on static graphs; as soon as this technical condition is not satisfied, many desirable properties associated with random walks on static graphs are lost.

In the near future I plan to continue extending our knowledge of graphs beyond the undirected static case, with a particular focus on algorithms for network analysis. As a case in point, I've recently submitted to publication a co-authored paper that deals with the problem of clustering in directed graphs. In this work I study a new notion of clusters in directed graphs and propose an algorithm that is able to discover patterns in real world data sets pertaining, for example, human migration or trade between countries.

With respect to teaching, I have been an Affiliated Lecturer at the University of Cambridge since October 2018. At Cambridge I have co-designed and taught the third year undergraduate course Probability and Computation, whose aim is to enrich students' knowledge of probability theory and showcase some of its applications to algorithm design. In the past two years I've also been responsible for a module on Graph Clustering for the Machine Learning course of the MPhil program in Advanced Computer Science. The aim of this module is to impart students the basics of graph clustering through lectures, and let students become acquainted with recent research advances through coursework. I've also been a supervisor and teaching assistant for courses in Algorithms, Complexity Theory, and Discrete Mathematics. Finally, I am currently co-supervising a Bachelor thesis on directed graph clustering and working with a PhD student on a project related to random walks on dynamic graphs.

In summary, I believe my research would nicely complement the research already being carried out in the department. It has strong connections with the research being conducted by the Data Science group, particularly by Prof. Milan Vojnovic. Moreover, my teaching experience and my background in Computer Science, which has a strong mathematical and foundational emphasis, make me comfortable teaching a broad variety of courses in the area of Computing, introductory courses in Probability, Statistics, Discrete Mathematics, and Linear Algebra, but also more advanced courses in Machine Learning, Data Science, and Algorithms.

I would be grateful for an opportunity to further discuss my application during an interview.

Yours sincerely,

Luca Zanetti