Ph.D. offer - Institut 3IA Côte d'Azur
Generative deep learning for network analysis

Advisor and location:

- Team: Maasai project-team, Université Côte d'Azur & Inria
- Advisor: Charles Bouveyron (Université Côte d'Azur & Inria)
- Co-advisor: Pierre Latouche (Université Paris Descartes)
- Localization: Maasai project-team, Inria Sophia-Antipolis, 2004 Route des Lucioles, 06902 Valbonne

Context and project: In all aspects of everyday life, there is a massive digitalization of systems that is increasingly important. One of the consequences of this phenomenon is the massive production of data, especially textual data. For example, social and communication networks allow users to interact through text. It is therefore of strong interest to be able to analyze those networks using information on both the network structure and the text contents.

As the texts are sometimes very long and/or heterogeneous, the statistical analysis proves to be a very useful instrument to provide a synthetic vision of the textual corpora. The goals are multiple: select the most representative words in each text, estimate the probability of extraction of a word in relation to a topic of conversation, detect the feelings of the author (eg favorable or unfavorable opinion), etc. One of the best-known approaches used for unsupervised text classification is latent Dirichlet allocation (LDA, Blei et al., 2003). On the other hand, this method does not take into account the order of the words and, from a generative point of view, it can produce texts that have no sense (the words are drawn by chance in a dictionary). Other more recent methods are related to natural language processing (NLP), an area of artificial intelligence that aims at making computers capable of understanding and manipulating human language. For statistical NLP, deep-learning neural networks are used for textual analysis and prediction (see, for example, Collobert and Weston, 2008).
In the context of network analysis, the LDA model has been extended recently by Bouveyron et al. (2017) to allow the clustering of networks with textual edges. The stochastic topic block model (STBM) model is a probabilistic model for networks with textual edges allowing to discover meaningful clusters of vertices that are coherent from both the network interactions and the text contents. A classification variational expectation-maximization (C-VEM) algorithm is proposed to perform model inference. STBM has also been adapted to the dynamic framework through a non homogeneous Poisson process.

The purpose of this Ph.D position within the Institut 3IA Côte d'Azur will be, first, to make the state of the art on deep learning techniques for the NLP and text analysis. Second, the Ph.D. candidate will expand the STBM model based on one of these techniques. One way would be to use auto-encoders to reduce the size of corpora and select the most important topics. Finally, the question of extending the deep-based approach to the case of dynamic network should be considered.

Expected skills: The candidate should have a graduate degree (Master 2 degree). Him/her scholar background should include:

- statistical/machine learning, statistical inference, clustering, classification
- deep learning, variational auto-encoder, back-propagation,
- knowledge of R, Python and/or C++.

Application: Application files should contain a resumé, an application letter and grade records of the 2 last years. Applications should be sent by email to charles.bouveyron@univ-cotedazur.fr and pierre.latouche@math.cnrs.fr.

References:

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. the Journal of machine Learning research, 3:993–1022, 2003.
R. Collobert and J. Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. ICML, p. 160–167, 2008.
C. Bouveyron, P. Latouche and R. Zreik, The Stochastic Topic Block Model for the Clustering of Networks with Textual Edges, Statistics and Computing, vol. 28(1), pp. 11-31, 2017