# Fouille de données distribués : requêtes skyline et spanners de graphes géométriques

**Responsable(s) :** Nicolas Hanusse et Sofian Maabout

lieu : LaBRI

téléphone : 05.40.00.26.04

e-mail : `hanusse, maabout@labri.fr`

équipe(s) : CombAlgo & MaBioVis

**Mots-clés :** Fouille de données, Algorithmique de graphes, Spanner, Skyline, optimisation de requêtes

## Description du sujet

Le but de ce sujet est l'étude et la proposition d'algorithmes de requêtes multi-dimensionnelles de type "skyline" en base de données. Informellement, un graphe est construit à partir de relations entre les données et toute requête correspond à une navigation dans ce graphe. Malheureusement, les propositions de littérature manipulent de trés gros graphes, difficilement maintenable en mémoire et pour lesquels les temps de requêtes ne sont pas garantis.

Etant donné un ensemble de points $\mathcal{P}$ dans un espace multidimensionnel, la requête skyline, notée $\mathcal{S}(\mathcal{P})$ retourne le sous-ensemble de $\mathcal{P}$ qui est constitué de l'ensemble des points qui ne sont *dominés* par aucun autre point : On dit que $p_1$ domine $p_2$ si (1) pour chaque dimension, la coordonnée de $p_1$ est inférieure ou égale à celle de $p_2$ et (2) pour au moins une dimension, la coordonnée de $p_1$ est strictement inférieure à celle de $p_2$. Par exemple, on cherche les restaurants les moins loins et les moins chers par rapport à un point donné. L'ensemble des restaurants représente l'ensemble $\mathcal{P}$ dont les coordonées sont $(Distance, Prix)$.

L'originalité de l'approche proposée dans cette étude est l'usage de spanners. Les spanners représentent des résumés de graphe permettant de garantir une certaine qualité lors du calcul de distance. Le travail proposé consiste à étudier l'apport de ces derniers pour le calcul des requêtes skyline.

Le mémoire commencera par une étude de bibliographique puis par la conception d'un algorithme basé sur les spanners. Selon la description de l'algorithme obtenu, il pourra être envisagé de réaliser une expérimentation sur des jeux de données réels.

## Références bibliographiques

– PREFER : A System for the Efficient Execution of Multi-parametric Ranked Queries International Conference on Management of Data (SIGMOD) , 2001 Vagelis Hristidis , Nick Koudas , Yannis Papakonstantinou

– Algorithms and Analyses for Maximal Vector Computation. VLDB Journal. Vol 16, Number 1, January 2007, pages 5-28. Parke Godfrey, Ryan Shipley, Jarek Gryz,
– Dominant Graph : An Efficient Indexing Structure to Answer Top-K Queries, in Proceedings of the 24th International Conference on Data Engineering (ICDE), 2008. Lei Zou, Lei Chen
– Geometric Spanner Networks, Narasimhan, Giri ; Smid, Michiel, 2007, Cambridge University Press