Défi Mastodons 2016
Troisième appel à projets interdisciplinaire
La qualité des données dans le Big Data

Identification

Civilité et Nom du porteur du projet (CNRS)	M. CHEPELIANSKII Dimitri (SHEPELYANSKY Dima)
Titre long (max 150 caractères)	Applications de la matrice Google pour les réseaux directionnels et Big Data
Acronyme	APLIGOOGLE

Résumé du projet : (6 lignes maxi)
Les échanges de nos sociétés modernes peuvent être vus comme un riche flux de données et métadonnées se propageant au travers de réseaux dirigés complexes. Étudier ceux-ci est tout particulièrement important pour l'analyse et la compréhension du savoir humain, du commerce international, de la biologie des systèmes et des télécommunications mobiles. Dans ce projet interdisciplinaire, nous appliquerons aux Big data les outils avancés issus de la physique théorique et des mathématiques (chaînes de Markov, matrice Google, matrices aléatoires, ...)

Exposé scientifique du projet :

Vision scientifique et originalité du projet Au cours des dernières décennies, nos sociétés modernes ont produit une immense quantité de données provenant de sources disparates. Dans la majorité des cas, ces sources produisent des données dont la structure est plus riche que celle d'un simple réseau d'ordre 2 à savoir une collection de nœuds reliés entre eux par des liens dirigés portant éventuellement un poids statistique. Les données que nous allons considérer dans ce projet contiennent des informations différentes des métadonnées associées aux nœuds et aux liens. Nous allons nous intéresser à une modélisation multi-fonctionnelle de ces réseaux qui seront ainsi représentés par des tenseurs d'ordre au moins égal 3. On envisagera, par exemple, une caractérisation des échanges commerciaux internationaux par une étude tensorielle liant l'exportateur, l'importateur et les divers produits échangés; de même, on caractérisera un document wiki par son éditeur, son contenu, ses commentateurs, son édition linguistique; dans le domaine des communications mobiles on peut caractériser également un terminal par l'ensemble des applications mobiles installées, utilisées, sa connectivité radio ou encore le profile de son utilisateur (mobilité, consommation des données, …); en biologie des systèmes deux protéines interagiront suivant une multitude d'interactions possibles, responsables ou non de la progression d'un cancer; … Toutes ces données comportent de plus une dimension temporelle permettant d'apprécier l'évolution de ces réseaux multifonctionnels et éventuellement d'en prévoir le comportement futur.

Ce projet s'inscrit dans le perpétuel challenge de l'analyse et la valorisation des immenses et toujours croissantes quantités de données et fournira une compréhension approfondie du flux d'information multifonctionnel au travers des réseaux tensoriels en modélisant leur évolution temporelle. Les principales applications porteront sur les différentes éditions linguistiques de Wikipedia, les échanges issus du commerce international, les réseaux de télécommunications mobiles, et les réseaux dits "omiques" issus de la biologie des systèmes.

La faisabilité de ce projet repose sur l'expérience des partenaires, sur leur connaissance des différents domaines scientifiques impliqués, et sur l'accès aux différentes bases de données: réseaux sociaux, éditions Wikipedia, échanges économiques internationaux (bases de données de l'Organisation Mondiale du Commerce [OMC], de l'Organisation de Coopération et de Développement Économiques [OCDE], et de l'Organisation des Nations Unies [ONU COMTRADE]), et des bases de données biologiques notamment celles de l'Institut Curie.

Axes de recherche et verrous scientifiques Dans ce projet interdisciplinaire, nous appliquerons aux Big data les outils avancés issus de la physique théorique et des mathématiques: chaînes de Markov, opérateurs de Perron-Frobenius, matrice Google, graphes aléatoires, théorie des matrices aléatoires et chaos quantiques, méthodes spectrales. Ces outils mathématiques avancés sont requis pour traiter de manière efficace les gigantesques ensembles de données collectées. Ce projet est construit principalement sur ces méthodes, algorithmes, et outils récemment développés par les partenaires

du consortium. Ce projet permettra aussi de développer et de proposer des méthodes de construction de ces graphes multi-fonctionnels puisque les données traitées proviennent toutes de sources technologiques potentiellement tronquées ou inconsistantes.

Axe 1- Réseaux multi-produit du commerce mondial: Le porteur du projet a obtenu l'accès à la base de données COMTRADE des Nations Unies listant les échanges commerciaux internationaux d'environ 5000 produits entre 227 pays des Nations Unies. La taille de la matrice de Google que l'on peut construire à l'aide de ces données est de l'ordre du million [1.1] et les données collectées permettent de retracer son évolution sur les 50 dernières années. Les flux commerciaux sont classés à l'aide des algorithmes PageRank et CheiRank développés pour le World Wide Web et pour d'autres réseaux dirigés de grande échelle. Pour le commerce mondial ces algorithmes traitent tous les pays du monde avec la même équité, et ce, indépendamment de leur richesse propre. Avec une telle équité de traitement, cette méthode place tout de même le groupe des pays les plus industrialisés au sommet du classement [1.1]. De plus, elle met en exergue de nouvelles propriétés et possibilités d'analyse en comparaison avec la méthode de classement usuelle de type importation/exportation. Ici, le vecteur de l'algorithme CheiRank (algorithme PageRank sur le réseau avec liens directionnels inversés [1.1]) caractérise naturellement les flux d'échange « sortants » d'un pays. L'algorithme PageRank permet l'analyse et le classement des flux « entrants ». Ces flux « sortants » (resp. « entrants ») sont clairement reliés aux flux des exportations (resp. des importations). Ainsi, l'analyse bi-dimensionnelle CheiRank/PageRank s'avère naturellement efficace pour représenter les échanges commerciaux internationaux. La taille de la matrice Google associée au commerce international multi-produits est de l'ordre du million. Son analyse nécessite ainsi l'utilisation de la méthode d'Arnoldi pour déterminer l'ensemble de ses vecteurs et valeurs propres. Nous avons montré que l'analyse des vecteurs propres permet d'extraire les communautés cachées de groupes commerciaux, tout comme c'est le cas pour la matrice Google de Wikipedia [1.1]. Dans cet axe du projet, nous souhaitons comparer les données COMTRADE des Nations Unies aux données de l'OMC (Genève), et notamment mener une étude sur la stabilité structurelle de ce réseau. Les données de l'OMC comportent les interactions entre plusieurs secteurs de production, ce qui est absent des données COMTRADE [1.1]. La construction de la matrice de Google multi-produits sera clé pour l'étude de la stabilité structurelle. En effet, cette étude cherche à déterminer l'influence de la variation du prix de produits spécifiques (e.g. du gaz, du pétrole, etc.) sur les échanges multi-produits. Il est théoriquement possible de prendre en compte directement la variation des données lors de la création de la matrice de Google, mais il reste à l'appliquer sur les données brutes de l'OMC. Ces données devront être correctement filtrées, re-calées et débruitées pour modéliser des réseaux multi-fonctionnels exploitables, et ainsi obtenir des résultats solides.

Axe 2- Réseaux multilingues Wikipedia: L'édition Wikipédia associée à une langue donnée constitue un réseau dirigé de citations hyperliens entre les articles de cette même édition. L'analyse de l'édition anglaise à l'aide de la matrice de Google a été effectuée par le porteur du projet en 2010. Cependant, il est important d'analyser les différents points de vue culturels encodés dans les différentes éditions. Cette tâche a débuté avec l'analyse de 24 éditions Wikipedia [1.1,2.1], où chaque édition est indépendante l'une de l'autre. Nous prévoyons de construire un seul réseau global Wikipedia regroupant les 24 (voire plus) éditions linguistiques en exploitant les hyperliens reliant ces éditions. On peut mettre en relation ces réseaux multilingues avec les réseaux d'échanges commerciaux de l'axe 1. En effet, les langues jouent ici le rôle des biens échangés. Ainsi, les méthodes et modèles défini dans l'axe 1 seront directement exploitables pour l'analyse du réseau Wikipedia global. Cette analyse nous permettra de représenter la structuration du savoir humain via le prisme des différentes cultures en étudiant l'intrication de celles-ci. Nous notons que les classements des personnages historiques utilisant la source Wikipedia [1.1] continuent d'attirer beaucoup d'attention comme en témoigne l'activité des groupes de Stony Brook (http://ww.whoisbigger.com) et du MIT (http://pantheon.media.mit.edu). Ces résultats ont été remarqués par The Guardian, The Indepedent, Le Figaro, ... (voir http://www.quantware.ups-tlse.fr/FETNADINE/press.htm). Le spectre et les états propres de la matrice Google du réseau global multilingue Wikipedia seront analysés par la méthode d'Arnoldi pour extraire les communautés cachées et les différences culturelles, étendant les résultats déjà obtenus pour l'édition anglaise de Wikipedia. La taille du réseau global multilingue Wikipedia est estimé à environ 17 millions de nœuds. Cette taille reste tout de même en deçà de la taille maximale jusqu'ici étudiée de 41 millions pour le réseau Twitter 2009 étudié par le groupe Quantware [1.1].

Axe 3- Réseaux mobiles dirigés: Les utilisateurs des incontournables réseaux de communications mobiles engendrent un réseaux dirigé complexe, à la fois vaste et fortement dynamique. Une grande part du trafic généré est véhiculé par les réseaux 3G/4G actuellement. Dans 2 ans, le volume mondial des échanges annuels doit atteindre 190 exabytes. Pour

transporter une telle masse de donnée, les futurs réseaux 5G ont pour ambition de fournir des communications de 1Gbps à des dizaines d'utilisateurs situés au même étage d'un immeuble de bureaux, cela en reliant simultanément des centaines voire des milliers de machines et objets communicants. Le design de la 5G est l'un des principal challenge dans le domaine des TIC pour les 5 prochaines années. Pour faire évoluer les systèmes actuels à ce nouvel objectif, des algorithmes adaptatifs et innovants doivent être élaborés. Pour cela, il est fondamental de comprendre la nature des échanges existant dans ces réseaux mobiles. Ceci permettra de définir des algorithmes de communication adaptatifs (e.g. algorithmes de délestage distribués, de pré-chargement, de communications opportunistes, …). Les premiers développements ont débuté dans le contexte du projet CHIST-ERA MACACO (http://macaco.inria.fr) dans le but de définir des algorithmes distribués exploitant la corrélation entre le contexte et le contenu des communications. Dans ce but, des données (anonymisées) traçant les contenus et le contexte d'utilisation d'une cinquantaine de smartphones ont été collectées sur une période d'un an, et ce toutes les cinq minutes. Ces données listent, entre autres, les différentes applications lancées, et notamment celles qui émettent ou chargent des données. Les différents réseaux (mobile, WiFi, bluetooth, ..) sont aussi tracés. Dans cet axe du projet, nous proposons d'extraire les propriétés fondamentales de ces traces collectées par le projet MACACO en utilisant les concepts du chaos quantique et l'analyse de la matrice Google. Nous chercherons à caractériser les ensembles d'applications générant de fortes interactions, en les liant notamment à une analyse des réseaux sociaux sous-jacents. Les méthodes développés dans les axes précédents seront exploitées à cet effet. Nous chercherons aussi à caractériser la dynamicité des échanges en analysant l'impact des variations de contexte sur le volume des contenus émis ou reçus par les applications.

Axe 4- Réseaux "omiques" multifonctionnels: Le groupe "Biologie Computationnelle des Systèmes du Cancer" (https://sysbio.curie.fr/) de l'Institut Curie possède une longue expérience de l'analyse de larges graphes représentant les réseaux biologiques impliqués dans la progression des cancers avec comme but une meilleure interprétation des profiles moléculaires des tumeurs, un meilleur diagnostic et un meilleur diagnostic basé sur la connaissance de l'interaction entre les molécules. Le groupe possède une solide expertise de l'utilisation de bases de données publiques de réseaux biologiques et a aussi développé ses propres bases de données comme l'Atlas of Cancer Signaling Network [4.1] représentant un vaste réseau biologique impliqué dans la progression des tumeurs. Le groupe a développé plusieurs méthodes mathématiques et numériques pour rechercher et visualiser les données "omiques" au sein de grands réseaux biologiques. Ces méthodes sont basées sur la technologie Google Maps [4.2], sur l'analyse de données "omiques" à l'aide de réseaux biologiques à grandes échelles [4.3,4.4,4.5,4.6], sur la modélisation mathématique des réseaux biologiques impliqués dans le cancer [4.7,4.8,4.4]. Le groupe possède également une longue expérience dans le développement de méthodes avancées pour la réduction dimensionnelle de données "omiques" linéaires et non linéaires en utilisant les réseaux biologiques que ce soit pour le calcul ou pour l'interprétation des résultats [4.4,4.9,4.5]. Dans ce projet, nous projetons d'améliorer les approches mathématiques existantes pour l'analyse des données "omiques" en utilisant notamment les propriétés de "scalabilité' des réseaux biologiques (à savoir la capacité à résoudre un problème de décomposition en valeurs propres pour les Laplaciens associés à des très grands graphes contenant des dizaine de milliers de nœuds). Nous chercherons également à développer des nouvelles approches numériques pour sonder la structure des réseaux biologiques et pour y voir des motifs particuliers associés aux données "omiques" (clusters, corrélations, composantes). De même, en lien avec les 3 autres axes, nous construirons la matrice de Google associé aux réseaux "omiques" qui permettra d'identifier, à l'instar des échanges commerciaux et du réseaux Wikipedia, les "communautés cachées" de protéines responsables de certains cancers.

Disciplines impliquées, participants et acquis scientifiques du consortium Ce projet réunit un consortium interdisciplinaire de chercheurs en physique théorique, en informatique, et en biologie. **Laboratoire de Physique Théorique de Toulouse:** Le porteur du projet est Dmitrii Chepelianskii (Nom scientifique : ***Dima Shepelyansky***), toutes les informations sur http://www.quantware.ups-tlse.fr/dima, DR1 CNRS, au LPT UMR5152 à Toulouse. Il est l'auteur de 226 publications avec selon Goolge scholar 9200 citations, H-index 46. Il a coordonné le projet EC FET NADINE (mai 2012 à Avril 2015) consacré aux réseaux dirigés (http://www.quantware.ups-tlse.fr/FETNADINE) et noté excellent par l'EC FET. Il a par le passé coordonné d'autres projets EC FET. ***Klaus Frahm***, professeur de physique théorique à l'Université Paul Sabatier de Toulouse, est spécialiste de la théorie des matrices aléatoires et de la méthode de diagonalisation de Arnoldi pour les matrices de rang élevé. Les nombreuses publications sur le sujet sont recensées dans l'article de revue [1.1] L.Ermann, K.M.Frahm and D.L.Shepelyansky, "Google matrix analysis of directed networks", Rev. Mod. Phys. 87, 1261 (2015). **Institut de Recherche en Informations de Toulouse:** ***Katia Jaffrès-Runser***, maître de conférences à l'INP Toulouse (IRIT UMR5055 / pour plus d'information

http://www.irit.fr/~Katia.Jaffres), s'intéresse tout particulièrement à la modélisation et à l'évaluation des performances de réseaux sans-fil avec forte contraintes de qualité de services. Elle a travaillé avec des processus et modèles de Markov destinés à dériver des enveloppes de performance multi-critères pour caractériser la qualité de réseaux de communication sans-fil multi-sauts [2.1] ; elle s'intéresse aussi à la prise en compte des interactions, du contexte et du contenu des échanges pour la conception de réseaux de communication dynamiques pro-actifs [2.2][2.3]. Elle est responsable pour l'INPT du projet CHIST-ERA MACACO. Autres participants : **Samer el Zant** (*PhD IDEX Toulouse*), **Tao Peng** (*Post-doc CHIST-ERA MACACO*). Publications sur le sujet : [2.1] K. Jaffrès-Runser, *et al.*, "A Cross-layer Framework for Multiobjective Performance Evaluation of Wireless Ad Hoc Networks", in Elsevier Ad Hoc Networks, v.11, no. 8 (2013); [2.2] M. Schurgot, C. Comaniciu and K. Jaffrès-Runser, "Beyond Traditional DTN Routing: Social Networks for Opportunistic Communication", in IEEE Comm. Mag., v. 50, n. 7 (2012) [2.3]. P. Olmo Vaz de Melo, A. Viana, M. Fiore, K. Jaffrès-Runser, F. Le Moüel and A. A. F. Loureiro, "RECAST: Telling Apart Social and Random Relationships in Dynamic Networks", Elsevier PEVA, v.87(2015). **Institut UTINAM:** ***José Lages,*** maître de conférence en physique théorique à l'Université de Franche-comté, il travaille à l'Institut UTINAM (UMR6213 http://perso.utinam.cnrs.fr/~lages/). Récemment, avec le porteur de ce projet, il a proposé un classement mondial des universités en sondant les matrices Google de 24 éditions linguistiques différentes de Wikipedia [3.1]. Ce classement a suscité un vif intérêt médiatique (94 articles de presse recensés à ce jour dans Le Monde, MIT Technoly Review, Times Higher Education, … http://perso.utinam.cnrs.fr/~lages/datasets/WRWU/press/Press.html). Autres participants: ***Françoise Gazelle***, IR CNRS en charge de la plate-forme informatique, ***Sékou Diakité***, IE, ***Bernard Debray***, IR CNRS en charge des bases de données. Exemple de publication : [3.1] J. Lages, A. Patt, D. L. Shepelyansky, "Wikipedia Ranking of World Universities" arXiv:1511.09021 (http://perso.utinam.cnrs.fr/~lages/datasets/WRWU/). **Institut Curie:** ***Andrei Zinovyev*** (http://www.ihes.fr/~zinovyev/), master en physique théorique, doctorat en informatique (2001), HDR en biologie (2014). Il est le coordinateur scientifique du groupe "Biologie des Systèmes Computationnelle du Cancer" de l'Institut Marie Curie depuis 2005 (https://sysbio.curie.fr/). Ses recherches actuelles portent sur l'analyse et la modélisation mathématique de grands réseaux biologiques utiles à la recherche sur le cancer afin d'améliorer le diagnostic et le pronostic, sur les méthodes de réduction dimensionnelle de données, et sur l'analyse des données multi-niveaux "omiques" pour la recherche sur le cancer. Il est le co-auteur de 3 livres et de plus de 70 publications. Autres participants: ***Inna Kuperstein*** (CDD - Institut Curie), ***Laurence Calzone*** (permanent - Institut Curie), ***Urszula Czerwinska*** (doctorante). Quelques publications du groupe: [4.1] Kuperstein I, *et al.*, "Atlas of Cancer Signalling Network: a systems biology resource for integrative analysis of cancer data with Google Maps". 2015. Oncogenesis 4:e160; [4.2] Bonnet E, *et al.*, "NaviCell Web Service for network-based data visualization". 2015. Nucleic Acids Res. 43(W1):W560-5; [4.3] Kuperstein I, *et al.*, "The shortest path is not the one you know: application of biological network resources in precision oncology research". 2015. Mutagenesis 30(2):191-204; [4.4] Barillot E., *et al.*, Chapman & Hall, CRC Mathematical & Computational Biology, 2012, 452p.; [4.5] Rapaport F., *et al.*, "Classification of microarray data using gene networks." 2007. BMC Bioinformatics 8:35; [4.6] Czerwinska U, *et al.*, "DeDaL: Cytoscape 3 app for producing and morphing data-driven and structure-driven network layouts". 2015. BMC Syst Biol. 14;9:46; [4.7] Kuperstein I, *et al.* , "Network biology elucidates metastatic colon cancer mechanisms". 2015. Cell Cycle 14(14):2189-90; [4.8] Cohen PAD, *et al.*, "Mathematical Modelling of Molecular Pathways Enabling Tumour Cell Invasion and Migration". 2015. PLoS Computational Biology 11(11):e1004571; [4.9] Biton A., *et al.*, "Independent component analysis uncovers the landscape of the bladder tumor transcriptome and reveals insights into luminal and basal subtypes". 2014. Cell Reports 9(4), 1235-1245.

5 mots clés : matrice Google, réseaux complexes dirigés, chaînes de Markov, réseaux mobiles, réseaux omiques

Décrire, par poste de dépense (missions, fonctionnement…), la demande financière pour l'année 2016.

Projet idéalement prévu sur 3 ans. La demande financière pour 2016 est la suivante:
Missions et fonctionnement : 60keuros - réunion de travail à Toulouse ; - Invitation de chercheurs sur courte période ; - publication dans des revues open access à fort impact
Equipement : 20keuros - Achat de machines de calcul dotées d'une grande mémoire vive.

Visa du directeur de l'unité du porteur du projet