|  |
| --- |
| Le big data au service de populations fragilisées et de communautés éphémères. |
| Proposition détaillée. |
|  |
| PSC en partenariat avec LUNALOGIC. 2015/2016. |
|  |
| **Effectué par Weiwei Guo, Diego Souza, Henrique Fiuza, Mohamed amine Boulmakoul et Elghali Lalami.**  **Document réalisé grâce à Mr Philippe Laurier encadrant du PSC.** |
|  |
|  |

**Sommaire**

**I- Enjeu et motivation du travail, objectif final**

***1-Contexte et enjeux socioéconomiques***

***2-L’approche retenue et son caractère structurant sur le futur mode d’organisation des élèves***

***3-Les principaux objectifs poursuivis***

II- **Revue et analyse de l’état de l’art / des approches concurrentes ou alternatives**

**III- Objectifs intermédiaires, avec leur échéancier,**

**IV-Méthodes, organisation du travail, répartition des tâches**

**V-Identification des moyens auxquels le projet fera appel**

**VI-Contributions de partenaires internes (laboratoires, binets) et externes (entreprises organismes)**

**VII-Éventuels résultats préliminaires**

**VII-Références bibliographiques (publications, brevets...)**

**I- Enjeu et motivation du travail, objectif final**

Le *big data* au service de populations fragilisées et de communautés éphémères.

Le secteur de l’assurance vit une révolution où ce ne sont plus les personnes qui sont observées, mais les liens entre elles. Liens qui mettent en lumière des mutualités latentes, permettant de modifier le calcul du risque ; qui font ressortir des communautés éphémères ou dispersées mais rattachables à des classes de risques.

Serait ainsi concevables des formes novatrices d’assurance et de protection sociale, qui soient dynamiques. C'est-à-dire, à travers la collecte et l’analyse des données massives et hétérogènes de populations, découvrir des typages de graphes locaux à travers les caractéristiques structurelles des données, identifier des formes canoniques pour y associer un risque. Le projet expérimentera une plate-forme informatique (collecte de données, algorithme de fouille de graphe, classification, et filtrage collaboratif de recommandation) orientée sur des populations fragilisées.

***1-Contexte et enjeux socioéconomiques***

Dans un contexte persistant de montée de la précarité et d’accélération des aléas de la vie (paupérisation, difficulté de l’accès aux soins, chômage, déclassement), de plus en plus de personnes sont amenées à réduire, voire supprimer, leur protection sociale et les assurances obligatoires ou facultatives dont elles bénéficiaient antérieurement. De nos jours, le nombre de Français qui ne peuvent avoir accès à une couverture de santé est estimé à 4 millions, et plus de 15 millions de personnes ne disposent plus de moyens financiers suffisants pour se protéger correctement : accidents divers, problèmes de santé, dépendance, assurance des biens (habitation, automobile).

Les dispositifs sociaux s’avèrent complexes, insuffisamment transparents et génèrent des effets de seuils qui excluent une partie de ces populations situées au-dessus des minima de ressources donnant droit à toutes les protections sociales, mais au-dessous des revenus qui leur permettraient d’accéder à un bon niveau de protection. Par leur structure organisationnelle et par leur structure de coûts, ces dispositifs sociaux tout autant que les mutuelles et les assurances privées rencontrent une difficulté grandissante à répondre à ces populations en voie d’être désocialisées ou déracinées, et à maintenir une qualité de relation et d’information croisée qui permette d’y détecter l’émergence simultanée de besoins de protection dispersés ou éphémères.

* Enjeu social, de meilleure compréhension des fragilités humaines (à l’aide de supports de détection technologiques), afin de faciliter leur anticipation et leur traitement. Cette compréhension ayant utilité à se faire presque en temps réel (En tant qu’exemples simples mais touchant fréquemment les personnes fragiles, dépendantes ou seules : début d’une épidémie, difficultés financières liées à une échéance collective sur tout un territoire …), ce qui rejoint une des caractéristiques attendues du *big data*.
* Enjeu entrepreneurial, d’adaptation des modèles économiques et des outils déployés par les entreprises du secteur de l’assurance.

***2-L’approche retenue et son caractère structurant sur le futur mode d’organisation des élèves***

Notre groupe sera amené à travailler sur ce qui pourrait être défini comme une chaîne comportant quatre maillons :

* des bases de données relatives à des personnes clientes de compagnies d’assurance.
* des outils informatiques d’analyse de ces données, susceptibles de faire ressortir des particularités au sein de ces populations (à savoir des liens réunissant des personnes simultanément touchées par de mêmes fragilités sociales, financières, psychologiques, etc.).
* une confrontation avec des experts du métier de l’assurance, pour une réflexion commune sur l’adaptation de leur actuelle offre, en fonction des hypothèses ressorties des précédentes analyses de données.
* une confrontation avec des professions confrontées aux fragilités humaines et sociales : psychologues, assistants sociaux, etc.

De ce fait, la mission revêtira un caractère interdisciplinaire : mathématiques, informatique, sciences sociales. L’un de ses intérêts consistera donc dans cette aptitude à embrasser l’ensemble d’une chaîne allant des outils techniques (bases de données, algorithmes) à des méthodes de gestion ou à l’interprétation de phénomènes sociaux.

***3-Les principaux objectifs poursuivis***

Le but du PSC est d’explorer la faisabilité, par l’exploitation des données numériques hétérogènes disponibles, d’identifier des groupes de population ayant en commun certains paramètres qui permettraient de les rattacher à des besoins spécifiques de couverture sociale. Il serait ainsi possible de proposer des couvertures sociales ou assurancielles à ces populations, qui soient dynamiques, adaptées et moins coûteuses.

Le 1er objectif scientifique est la preuve de concept des heuristiques d’apprentissage automatique (« *Big Data* »), appliquées aux calculs du risque. Il s'agit, à travers la collecte et l’analyse des données massives et hétérogènes des populations concernées, de découvrir des typages de graphes locaux à travers les caractéristiques structurelles des données, d'identifier des formes canoniques pour y associer un risque.

Les données permettront de constituer à chaque instant T (discrétisé) un “graphe marqué”, dont les nœuds/sommets sont les individus auxquels sont attachés un certain nombre de descripteurs/marques au temps T encapsulant les caractéristiques sociodémographiques X(T) de l’individu (e.g. âge, localisation) et les besoins en matière de prestations sociales Y(T) (e.g. couverture de sinistres, assurance maladie, assurance chômage) et dont les arcs/arêtes représentent les liens entre les membres du réseau. Une autre façon de représenter l’information disponible à l’instant T est de construire un graphe biparti, dont les sommets sont répartis en deux groupes : l’un étant constitué par les personnes présentes dans les bases de données, l’autre correspondant à l’ensemble des besoins possibles en matière de prestation sociale. Les liens décrivent alors les besoins des individus à chaque instant.

Diverses techniques permettent de regrouper les personnes aux comportements/besoins similaires (détection de communauté), comme le *clustering* spectral, la factorisation de matrices, l’optimisation de la modularité. Les difficultés peuvent provenir d’un nombre de liens observés très petit par rapport au nombre de liens possibles. Il convient alors de formuler des hypothèses structurelles (parcimonie, *sparsity*) sur la matrice d’adjacence de manière à réduire la dimension du problème. Par ailleurs, l’enjeu est de réaliser cette étape de réduction de la dimension de façon incrémentale, à chaque instant, et de produire une visualisation interprétable de l’évolution du graphe ainsi segmenté.

Le 2ème objectif scientifique est d’ordre social : identifier de nouveaux modèles économiques fiables de micro-assurance, et une garantie de pérennisation et de plus grande efficience du modèle de protection sociale, notamment pour les populations vulnérables.

De manière plus indirecte mais qui conservera une dimension pédagogique en informatique et en droit : le travail prévu sur des bases de données réelles touchant à la vie privée de populations d’assurés, il sera essentiel de mettre en œuvre des processus de respect de cette confidentialité. Il y aura donc sur ce point une découverte des processus usuels dans le monde des entreprises, au regard des prescriptions de la CNIL.

II- **Revue et analyse de l’état de l’art / des approches concurrentes ou alternatives**

1. **Concurrents, travaux antérieurs**

L’essentiel des avancées dans ce domaine vient des pays émergents et des États-Unis.

Dans les pays émergents, la nécessité politique de développer des systèmes d’assurance universelle a permis le développement de plates-formes d’interlocution entre les assurés, les centres de soins et les opérateurs mobiles qui offrent l’avantage après une analyse fine des comportements de parvenir à des taux de renouvellement des primes de micro-assurance qui sont de l’ordre de 95% des personnes assurées.

Aux États-Unis, le nouveau système d’assurance santé a conduit aux développements de nombreux services de prévention et de contrôle de la qualité et de l’efficience des soins basés sur l’observation fine des comportements et la collection de base de données de protocoles thérapeutiques.

Dans le monde, il existe de nombreux projets de micro-assurance ou des assurances de santé à base communautaire, mais ils ne sont pas couplés à une exploitation des données personnelles disponibles. BIMA, une société de micro-assurance, offre en Afrique et en Inde des couvertures à des prix abordables, par le biais des réseaux mobiles. La Micro Insurance Academy, une ONG à New Delhi, a lancé en 2013 une initiative, qui rassemble les universités européennes et des organismes de financement pour développer des solutions de micro-assurance.

1. **Etat de l’art et des connaissances (Big Data)**

L’apprentissage automatique (*Machine-Learning*), à l'interface des mathématiques et de l'informatique, vise à la conception, l'analyse et la mise en œuvre des algorithmes permettant à une machine d'extraire des informations à partir de données, afin d'effectuer des tâches automatiquement aux fins de prédiction, aide à la décision ou de représentation de données. Combiné à des systèmes informatiques efficaces pour la collecte et la gestion de grands ensembles de données, l'apprentissage automatique permet des avancées significatives et rapides. Il constitue aujourd'hui un véritable corpus de méthodes algorithmiques et informatiques, capables de s'adapter à des données de nature différente à des fins diverses. Il est maintenant un champ de recherche décisif pour le développement des Sciences Humaines et Sociales.

Depuis 10 ans, la fouille de graphes (*graph mining*) a reçu une attention croissante, motivée par des domaines d'application tels que la compréhension à grande échelle du Web, la détection de communautés virtuelles dans les réseaux sociaux, les études en biologie et en génétique. Alors que le *data mining* (fouille de données) se concentre sur les valeurs des données, le *graph mining* (la fouille de graphes) accorde autant d'importance à la structure des données qu'à son contenu. La fouille de graphes couvre de nombreuses techniques pour l'identification des anomalies ou les tendances récurrentes dans un graphe, la découverte de groupes ou ensembles de nœuds qui partagent des modèles communs d'attributs et relations, l'extraction de motifs qui permettent de distinguer des catégories de graphes, et l'utilisation de ces modèles pour prédire la catégorie de nouveaux graphes. Il existe un grand nombre de travaux sur la classification de graphes. Une approche possible est de considérer une mesure permettant de quantifier la structure des communautés et de formuler la question de *clustering* comme un problème d'optimisation de complexité non polynomiale, pour lesquels d'assez bonnes solutions peuvent être obtenues de manière récursive ou en utilisant des méta-heuristiques adéquates. Plusieurs approches de classification de graphes existent, tels que celles basées sur la maximisation de la modularité ou sur la classification spectrale, qui fonctionnent bien de façon empirique, sont simples à mettre en œuvre et qui tirent avantage de la disponibilité des bibliothèques d'algèbre linéaire rapide. Le principe de la classification spectrale est de calculer les plus petits vecteurs propres de la matrice de Laplace du graphe normalisé et de regrouper les sommets en fonction de leur représentation dans l'espace propre correspondant.

**III- Objectifs intermédiaires, avec leur échéancier,**

Les objectifs listés ci-après ne sont au présent stade que des hypothèses, dépendantes de la découverte progressive du niveau qualitatif des bases de données et des outils informatiques prévus. Il est vraisemblable que quelques tâches ponctuelles de correction dans la mise en forme des données, ou de corrections informatiques de *bugs*, soient à réaliser. L’échéancier général étant :

* audit qualitatif des bases de données, mené en interne par la compagnie d’assurance qui les mettra à disposition : septembre et début octobre.
* Test et validation des algorithmes de fouille de données et de fouille de graphes : octobre-décembre.
* Travail d’analyse des bases de données : décembre à fin de la mission.
* Confrontation entre les hypothèses issues du travail d’analyse des données, et les dires d’experts issus du monde de l’assurance ou des métiers sociaux : janvier à fin de la mission.

Ce phasage laissant toutefois place à l’établissement de contacts transversaux entre les divers acteurs, dès le stade initial (octobre). En particulier pour favoriser une meilleure compréhension des réalités sociales destinées à transparaître à travers les travaux d’analyse technique.

**IV-Méthodes, organisation du travail, répartition des tâches**

Au sein du groupe, la répartition des tâches sera stabilisée durant les deux premiers mois de la mission, en fonction des compétences et affinités rapportées aux trois domaines de l’étude : mathématiques, informatique, sciences sociales.

Plus particulièrement, la partie informatique sera traitée par Diego Souza et Henrique Fiuza, tandis que la partie mathématique sera traité par Weiwei Guo, Mohamed amine Boulmakoul et Lalami Elghali. Quant à la partie science social elle sera traitée par tout les membres du groupe.

Toutefois, une priorité sera donnée à la préservation du caractère interdisciplinaire de l’étude, c'est-à-dire au dialogue interne à l’équipe, et externe avec les partenaires. Le but étant l’établissement d’un mode de fonctionnement et de travail itératif, en particulier pour la bonne compréhension des défauts ou qualités respectifs des bases de données et des algorithmes (comprendre collectivement si telle éventuelle absence de résultat provient d’un défaut à corriger sur les bases de données ou sur les algorithmes, ou dans une mauvaise coordination des deux). De même, une logique itérative prévaudra lorsqu’il s’agira de confronter les observations nées du terrain en matière de fragilités humaines existantes, et les hypothèses de fragilités ressortant des travaux informatiques et mathématiques sur les bases de données.

**V-Identification des moyens auxquels le projet fera appel**

Les partenaires extérieurs mettront à disposition :

* des algorithmes issus de la recherche, et actuellement en phase expérimentale.
* Des bases de données de grande taille, relatives à des populations d’assurés.
* Une mise en relation avec des experts du monde de l’assurance, pour un dialogue autour des hypothèses nées de l’exploration des bases de données.

**VI-Contributions de partenaires internes (laboratoires, binets) et externes (entreprises organismes)**

Des liens seront développés avec quatre profils de partenaires externes principaux :

* Lunalogic, entreprise de services informatiques ayant développé des compétences sur le domaine du big data et de la fouille de données.
* Télécom ParisTech, sur les mêmes domaines.
* Une compagnie d’assurance de taille nationale, disposant de bases de données et ayant accepté le principe de leur mise à disposition dans le cadre de la présente étude. Les conditions de respect de la sécurité et de la confidentialité du contenu de ces bases, a fait l’objet d’un accord et de la mise au point d’une procédure, avec les deux autres partenaires ci-avant. Les données feront l’objet d’une pseudonymisation forte, mais neutre vis-à-vis des objectifs scientifiques poursuivis.
* Le MIIIR –institut de recherche en micro assurances- ainsi que la société Solidassur, également positionnée sur ce segment.

**VII-Éventuels résultats préliminaires**

Il n’est pas prévu de résultats préliminaires. Toutefois, un phasage est prévu, qui scindera en deux étapes la tâche globale : selon les premiers résultats des audits qualitatifs des bases de données mises à disposition de l’étude, il est envisageable que les premiers tests pour la validation des algorithmes soient effectués sur des bases de données artificielles. Ce, de manière à ne pas être affectés durant ce processus initial par d’éventuels biais et imperfections de ces bases de données.

En effet, le caractère innovant de la mission amènera à découvrir et jauger à la fois des outils informatiques et des bases de données, dont le potentiel et le caractère exploitable restent à mesurer. Aussi apparaît-il souhaitable de découpler cette mesure et validation, d’abord des outils informatiques sur des bases artificielles, puis des bases réelles à l’aide de ces mêmes outils.

**VII-Références bibliographiques (publications, brevets...)**

[1] L’accès aux soins des plus démunis 40 propositions pour un choc de solidarité, Aline Archimbaud Sénatrice de Seine-Saint-Denis, Rapport au Premier ministre, septembre 2013 - http://www.onpes.gouv.fr/IMG/pdf/rapport\_officiel-aline\_archimbaud-2.pdf

[2] Lutte contre la pauvreté : Plan pluriannuel contre la pauvreté et pour l'inclusion sociale la feuille de route du Gouvernement pour 2014, Gouvernements Jean-Marc Ayrault en fonction du 16 mai² 2012 au 31 mars 2014, Publié le 4/01/2014 http://www.gouvernement.fr/gouvernement/lutte-contre-la-pauvrete-la-feuille-de-route-du-gouvernement-pour-2014

[3] Refonder l'Assurance Maladie (2014), Brigitte Dormont, Jean Tirole, *et al*, CAE-Conseil d'Analyse Economique, Note CAE n°12 avril 2014.

[4] Efficient Eigen-updating for Spectral Graph Clustering. (2013). S. Clémençon, C. Dhanjal (UPMC) and R. Gaudel (Lille 3). To appear in Neurocomputing.

[5] Distributed Matrix Completion (2012). Christina Teflioudi, Faraz Makari, Rainer Gemulla. In the Proceedings of 2012 IEEE 12th International Conference on Data Mining.

[6] Exact Matrix Completion via Convex Optimization (2008). Emmanuel J. Candès and Benjamin Recht. Foundations of Computational Mathematics. Vol 9. pp. 717--772. 2009.

[7] Factoring nonnegative matrices with linear programs. Victor Bittorf, Benjamin Recht, Christopher Re, and Joel A. Tropp. Advances in Neural Information Processing Systems. 2012.

[8] Parallel Stochastic Gradient Algorithms for Large-Scale Matrix Completion. Benjamin Recht and Christopher Re. Mathematical Programming Computation Vol 5, No. 2, pp. 201-226. 2013.

[9] Approaches to Recommendation (2013). S. Clémençon, C. Dhanjal. Internal Report Institut Mines Télécom/ Télécom ParisTech.

[10] Visual Mining of Epidemic Networks. S. Clémençon, H. de Arazoza, V.C. Tran and F. Rossi (2011). In Proceedings of the International Work Conference on Artifical Neural Networks, IWANN 2011 Torremolinos (Spain).

[11] Le Filtrage collaboratif vu comme un problème de consensus d’ordonnancements. S. Clémençon, R. Gaudel (Telecom ParisTech) (2011). In Proceedings of CAp‘2011, Chambéry (France).

[12] Efficient Eigen-updating for Spectral Graph Clustering, 2013, S. Clémençon et al. to appear in Neurocomputing.

[13] La polarité dans la politique de sécurité, (Mars 2013), Liliana Rosero, Thèse de doctorat à Institut Mines Télécom/ Télécom ParisTech sous la direction de Michel Riguidel.

[14] Nomadisme informatique, nomadisme des identités ? (2012), Revue La Jaune et la Rouge n°675, P. Laurier et C. Dufetrelle.

[15] Rétablir l’accès à l’assurance des personnes avec l’Eco-assurance (2014). Revue Actes, n°447, G Peltzer.

[16] A Distributed Algorithm for Large-Scale Generalized Matching (2013). F. M. Manshadi, B. Awerbuch, R. Gemulla, R. Khandekar, J. Mestre, M. Sozio. In 39th International conference on Very Large Databases (VLDB).

[17] Social content matching in MapReduce (2011). Gianmarco De Francisci Morales, Aristides Gionis, Mauro Sozio. In 37th International conference on Very Large Databases (VLDB).

[18] Parallel Graph-Based Semi-Supervised Learning (2012). J. Bilmes and A. Subramanya. In Scaling Up Machine-Learning: Parallel and Distributed Approaches, pp 240-261. Editors R. Bekkerman, M. Bilenko and J. Langford.

[19] Hierarchical Clustering for Graph Visualization (2001). S. Clémençon, H. de Arazoza, V.C. Tran and F. Rossi (2011). In Proceedings of the European Symposium on Artifical Neural Networks, ESANN 2011 Bruges (Belgium).

[20] Large-scale spectral clustering with MapReduce and MPI (2012). W.Y. Chen, Y. Song, H. Bai, C.J. Lin and E.Y. Chang. In Scaling Up Machine-Learning: Parallel and Distributed Approaches, pp 240-261. Editors R. Bekkerman, M. Bilenko and J. Langford.

[21] Ressources numérisées : le risque d'un nouveau féodalisme (2011), Revue La Jaune et la Rouge n°662, M. Riguidel et P. Laurier.

[22] Penser et mettre en oeuvre l’anonymisation de données médicales - Expérimentation à l’Hôtel-Dieu et dans une compagnie aérienne (2012), Revue Technologies et innovations hospitalières, n°15, p 34-35, P. Laurier et P. Jouniaux.

[23] Final recommendations report on future global research challenges in ICT trust and security, 2013, Riguidel, M. et. al., “BIC: Building International Cooperation for Trustworthy ICT”, FP7.

[24] Assurance, prévention, prédiction... dans l’univers du *big data*, octobre 2012, F. Ewald, Institut Montparnasse, coll. Recherches.

[25] La société du calcul, automne 2013, F.X. Albouy, Revue Risques 95.

[26] On-Line Learning Gossip Algorithm in Multi-Agent Systems with Local Decision Rules, S. Clémençon *et al.* to appear in the Proceedings of the 2013 IEEE International Conference on Big Data, Santa Clara.

[27] Maximal Deviations of Incomplete U-statistics with Applications to Empirical Risk Sampling, S. Clémençon, *et al.* to appear in the Proceedings of the SIAM International Conference on Data-Mining SDM13, Austin.