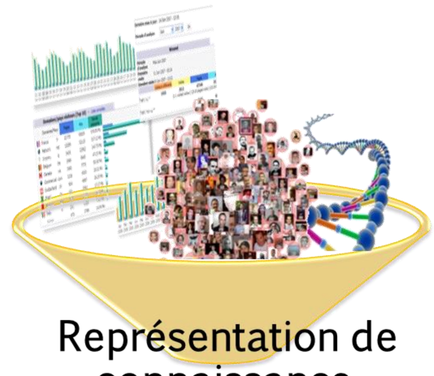
3. GRANDES MASSES DE DONNEES

De la connaissance à la décision
Exploiter de gros volumes de données par l'analyse sémantique



Parties prenantes

MINES Alès

LGI2P – équipe KID

Equipe

Sylvie Ranwez
Jacky Montmain
Michel Crampes
Gérard Dray
Stefan Janaqi
Michel Plantié
François Trousset
Nicolas Fiorini
Sébastien Harispe

Partenaires

Montpellier SupAgro

Universitat Rovira i Virgili

Instituts thématiques Inserm
Institut national de la santé et de la recherche médicale

Vincent Ranwez[1]

Montserrat Batet[2]

David Sánchez[2]

[1] Equipe DAVEM, SupAgro Montpellier,

[2] Université Rovira i Virgili de Tarragone, Espagne

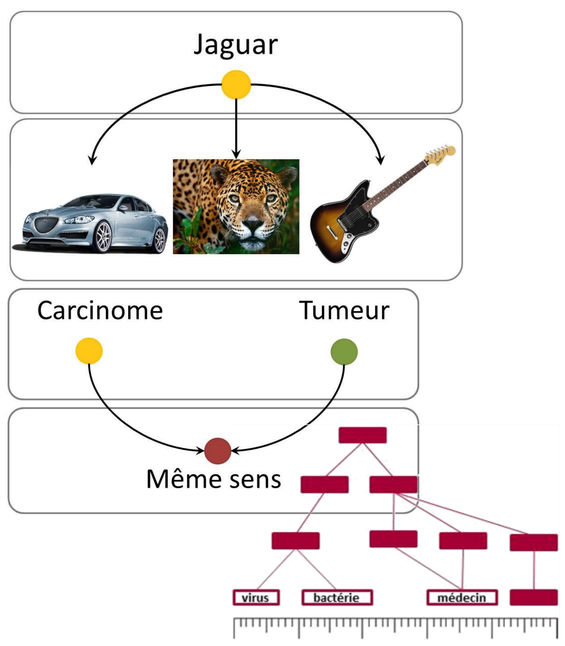
Sources d'information diverses, hétérogènes

Données non structurées

- Corpus de textes (publications, pages Web, réseaux sociaux)
- Ressources multimédia (images, sons, vidéos…)

Données structurées

- Données liées, massives et à caractère sémantique
- Représentation de la connaissance (ontologies), graphes RDF(S), OWL,…



Recherche d'information conceptuelle

Améliorer la pertinence des résultats

- Désambiguïser, généraliser, spécialiser les requêtes en utilisant une ontologie de domaine
- Assurer de meilleurs taux de précision et de rappel par des mesures sémantiques appropriées

Personnalisation, visualisation et interactivité

- Paramétrisation et personnalisation du système et de l'interface : pondération, reformulation, lentilles
- Visualisation globale des résultats sur une carte sémantique : affichage en 2D des résultats en fonction de leur degré de pertinence
- Justification des résultats

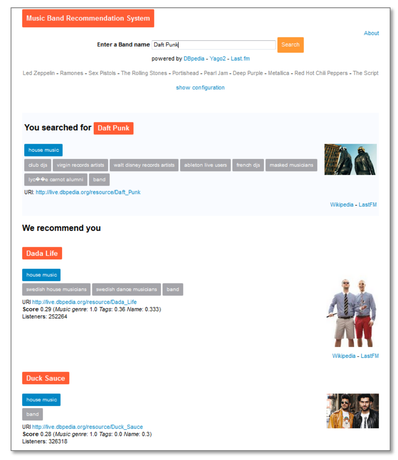
Développements logiciels

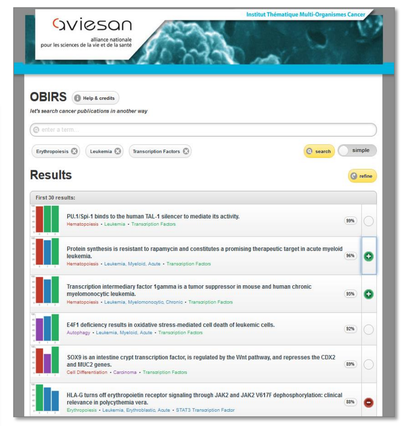
SML – *Semantic Measures Library*
Librairie logicielle dédiée au calcul de similarité sémantique
- Open-source, langage Java, exécutable en ligne de commande
- Traitement de gros volumes de données

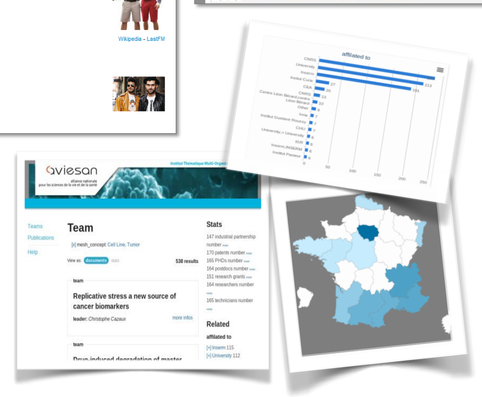
OBIRS – *Ontology Based Information Retrieval System*
Recherche d'information conceptuelle centrée sur l'utilisateur
- Application à la recherche de gènes indexés par la Gene Ontology
- Recherche de publications scientifiques biomédicales

Système de recommandation
Exploitation des mesures de similarité dans le contexte des données liées pour un système de recommandation

Kalitmo
Visualisation de données structurées pour la gestion de collectifs







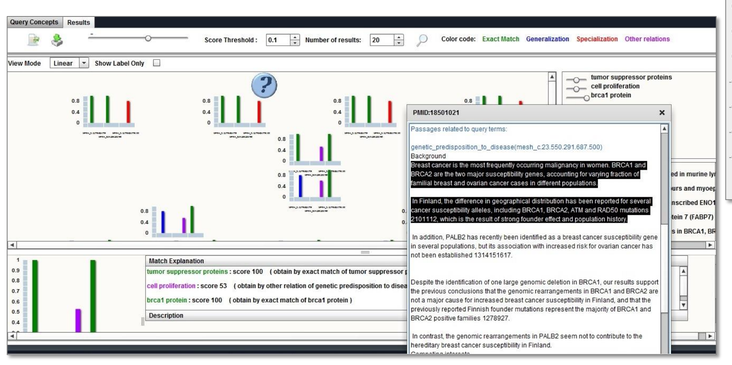
Extraction d'opinion

- Fouille de textes et apprentissage supervisé sur de larges corpus (Web) : TAL, segmentation, analyses statistiques
- Identification d'un lexique, détection de critères relatifs à un domaine
- Polarisation contextuelle, extraction conceptuelle et évaluation de critères

Développements logiciels

Synopsis
Outil de détection d'opinion, prise en compte d'une évolution temporelle

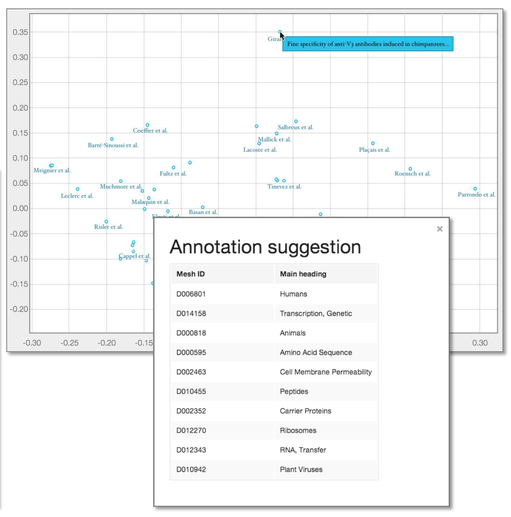
CoLexIR
Environnement hybride de RI associant deux approches conceptuelle et lexicale
Application à la fouille de corpus scientifiques avec mise en évidence des passages pertinents pour l'utilisateur



Indexation conceptuelle

Indexation par propagation

- Interface interactive d'assistance à l'indexation
- Annotation conceptuelle semi-automatique



Complexité" 26 et 27 mars 2014 colloque de l'Institut Mines-Télécom **"Numérique : Grande échelle &**

Contact sylvie.ranwez@mines-ales.fr Site web http://kidknowledge.wp.mines-telecom.fr

Enabling Large-scale Model Driven Engineering

Parties prenantes

Auteurs

Benelallam Amine[1]

Tisi Massimo[1]

Sunyé Gerson[2]

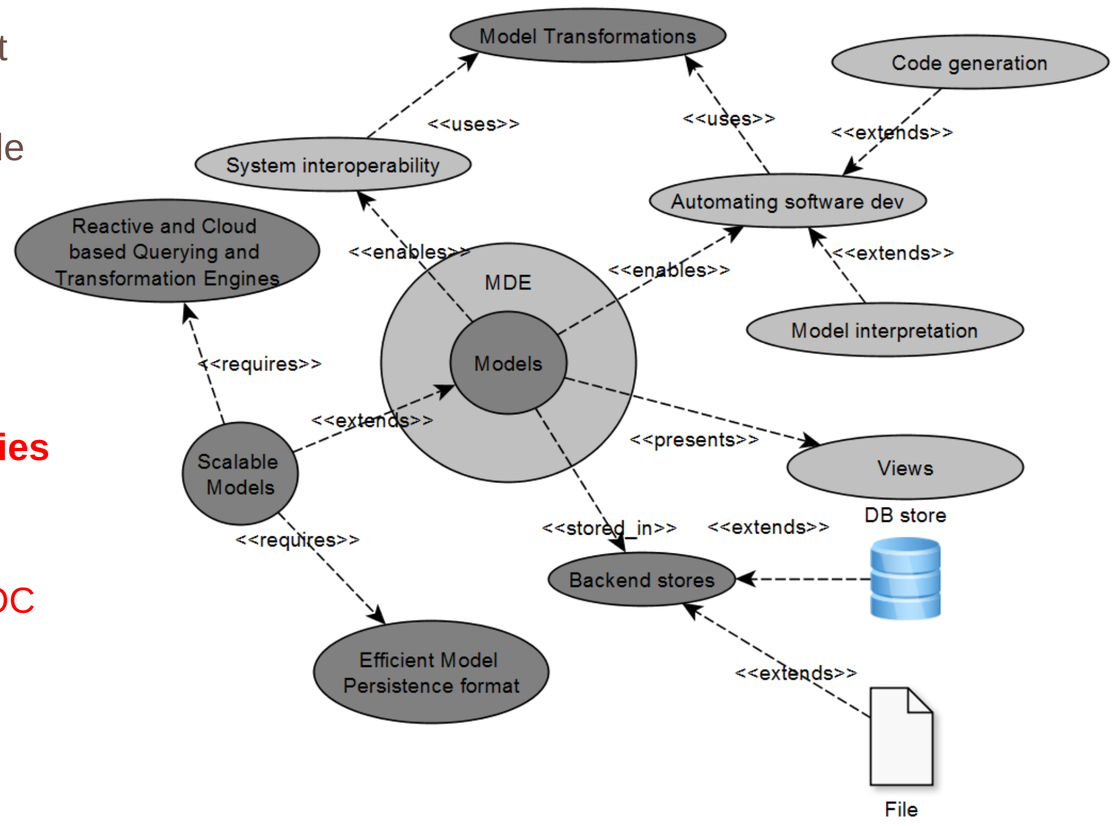
Gomez Llana Abel[1]

[1] AtlanMod, Mines Nantes, INRIA, LINA

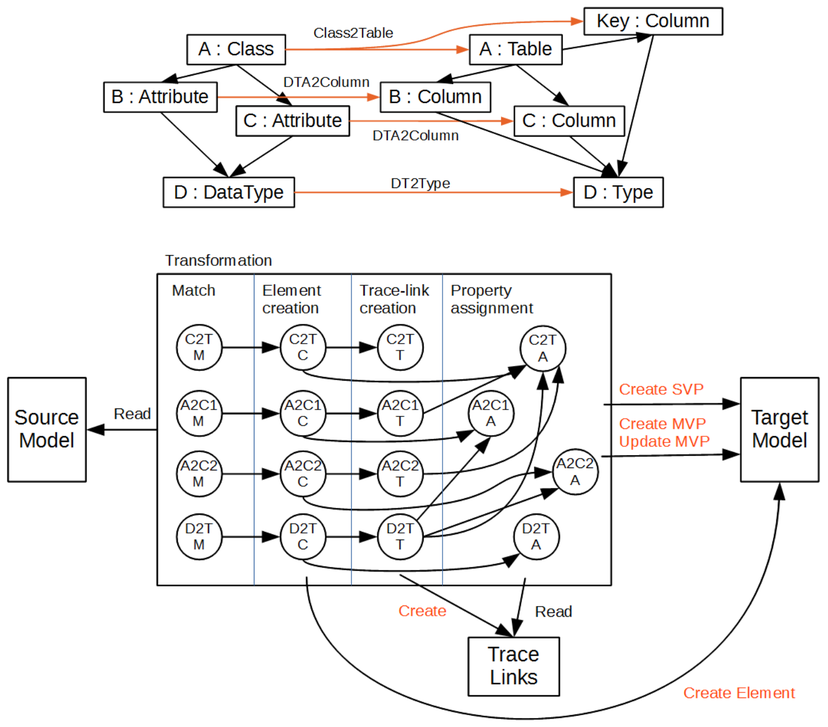
[2] AtlanMod, Université de Nantes, LINA

I- Model Driven Engineering

- Applying modeling to assist software development during the entire life-cycle
- Automating software development throughout code generation, validation, visualization etc.
- Large MDE community around Eclipse Modeling project
- Large set of tools in the Eclipse Modeling Project from industry and academics
- **The current generation of modeling technologies is stressed to its limits**
- **Examples from industry :**
  - Reverse-engineering systems with millions of LOC
  - Synchronizing views on building information models of several Gbs
- **Need of scalable MDE solutions for very large and complex systems**



II- Enabling scalability in Model Transformations (MT)

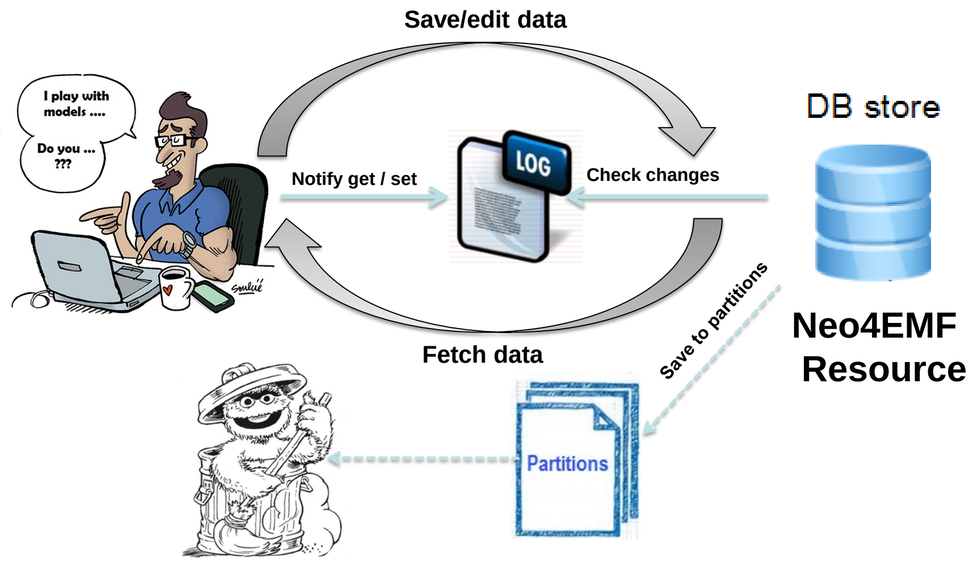


- The ATL transformation language:
  - is a model transformation language and toolkit developed in the AtlanMod research team
  - provides a parallel engine for faster transformations
  - enables change propagation and model synchronization using an incremental execution
  - reduces memory footprint and computation using lazy transformations
  - enables infinite transformations using lazy transformations

III- Enabling scalability in Model Persistence

Neo4EMF

- A model persistence framework (MPF) is a middleware that assists the storage of models
- The Neo4EMF model persistence framework:
  - provides a No-SQL backend using a graph database
  - enables loading large models using an on-demand loading mechanism
  - enables a lightweight first time loading by fetching objects not their data
  - involves a change (access) log to unload (save) models elements



Contact amine.benelallam@inria.fr

Site web http://www.emn.fr/z-info/atlanmod/index.php/User:Benelallam

Predicting personalized response to drug and environmental chemicals from genomic data with machine learning

[1,2] Erwan Scornet, [1] Elsa Bernard, [1] Yunlong Jiao, [1] Veronique Stoven, [1] Thomas Walter, [1] Jean-Philippe Vert

[1]Center For Computational Biology, Mines ParisTech ; INSERM U900 ; Institut Curie

[2]Paris VI

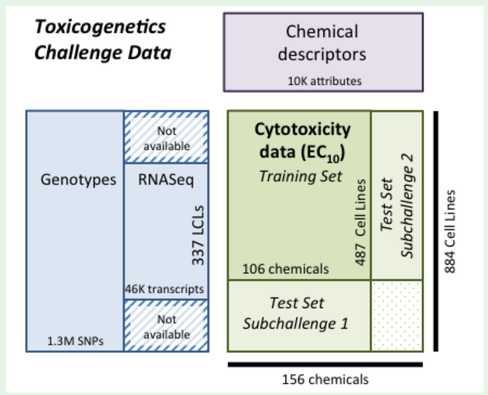
March 2014

Abstract : The response to drugs and environmental chemicals changes with people. Variation in genotype can explain different reaction to drug : while some treatment may be effective for some people, it can be useless and even harmful for others. The same mechanisms are at stake regarding the reaction to environmental agents, such as allergens. Recent advances in high-throughput sequencing open the way to personalized treatment based on genotype data. In this work, we predict the toxicity level of 106 chemicals for each patient using both information on patients and on chemicals.

Challenge of Toxicogenetic

- Provide personalized information about chemical toxicity for each patient
- Use genetic infomation of each patient to predict the chemical toxicity
- Use chemical information (substructures, compounds,...) to learn toxicity across chemicals

Data description



Data were available thanks to the DREAM challenge 8.0 which gathers :

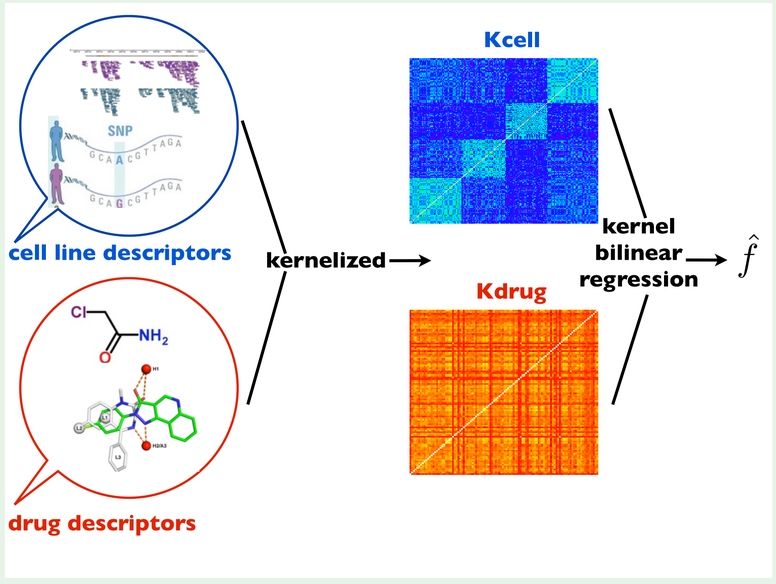
- Genotype data from the 1000 Genomes project (aiming at finding most genetics variants with frequency of 1%)
- RNA sequencing data from the Geuvadis Project (sequencing of some cell lines of the 1000 Genomes Project)
- Three covariates (sex, population and batch)

Toxicity $\longrightarrow$ The drug concentration that reduces the ATP synthesis of 10%.

Methods

To predict the toxicity values, we aim at

- Creating a measure of similarity between chemicals
- Creating a measure of similarity between genotypes



Let $\mathbf{x_i}$ be a vector of descriptors of cell line $\mathbf{i}$ and $\mathbf{y_j}$ be a vector of descriptors of chemical $\mathbf{j}$

- We model the toxicity $\mathbf{t_{ij}}$ of the chemical $\mathbf{j}$ on the cell line $\mathbf{i}$ by

$$\mathbf{t_{ij}} = \underbrace{\mathbf{f(x_i, y_j)}}_{\text{"bilinear"}} + \underbrace{\mathbf{b_j}}_{\text{offset}} + \underbrace{\epsilon_{\mathbf{ij}}}_{\text{noise}}$$

- We estimate $\mathbf{f}$ and $\mathbf{b}$ by penalized least-square regression

$$\min_{\mathbf{f}\in\mathcal{H},\mathbf{b}\in\mathbb{R}^p} \sum_{i=1}^{n}\sum_{j=1}^{p}\{\mathbf{t_{ij}} - \mathbf{f(x_i, y_j)} - \mathbf{b_j}\}^2 + \lambda\|\mathbf{f}\|^2_{\mathcal{H}}.$$

To specify $\|\mathbf{f}\|^2_{\mathcal{H}}$ we just have to choose two kernels :

$\longrightarrow$ $\mathbf{K_{cell}(x_l, x_{l'})}$ which measures the similarity between cell lines $\mathbf{x_l}$ and $\mathbf{x_{l'}}$.

$\longrightarrow$ $\mathbf{K_{drug}(y_m, y_{m'})}$ which measures the similarity between chemicals $\mathbf{y_m}$ and $\mathbf{y_{m'}}$.

- Then, the solution of the previous optimization problem is given by

$$\hat{\mathbf{f}}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n}\sum_{j=1}^{p} \hat{\alpha}_{\mathbf{ij}}\mathbf{K_{cell}(x, x_i)K_{drug}(y, y_j)}$$

where $\hat{\alpha}_{\mathbf{ij}}$ depend only on

$\longrightarrow$ $\mathbf{K_{cell}} = \left(\mathbf{K_{cell}(x_l, x_{l'})}\right)_{\mathbf{l,l'=1...,n}}$

$\longrightarrow$ $\mathbf{K_{drug}} = \left(\mathbf{K_{drug}(y_m, y_{m'})}\right)_{\mathbf{m,m'=1,...,p}}$

$\longrightarrow$ The toxicity matrix.

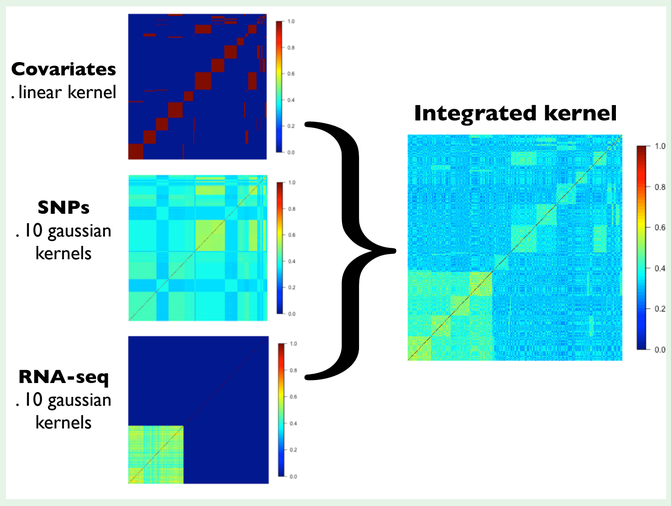
We also know that $\hat{\mathbf{b}}$ only depends on these three matrices.

References

1. B. Chem and al. *Journal of Chemical Information and Modeling* 49 :2044–2055, 2009.
2. Y. Yamanishi and al. *Journal of Chemical Information and Modeling* 51 :1183-1194, 2011.
3. P. Mahé and al. *Journal of Chemical Information and Modeling*, 45 :939–951, 2005.
4. T. Evgeniou and al. *Journal of Machine Learning Research* 6 :615-637, 2005.

What type of kernel do we choose ?

Kernel on cell lines :



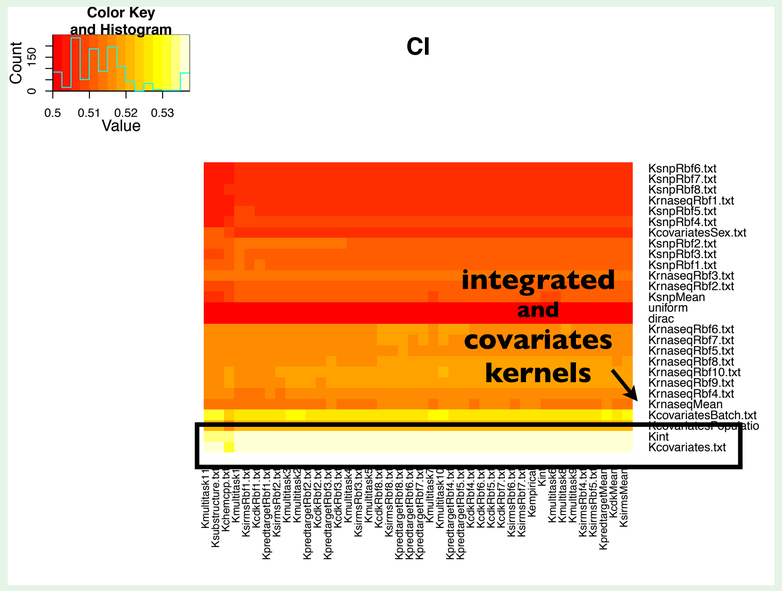
$\longrightarrow$ **29** kernels whose one that incorporate all informations (integrated kernel).

Kernel on chemicals :

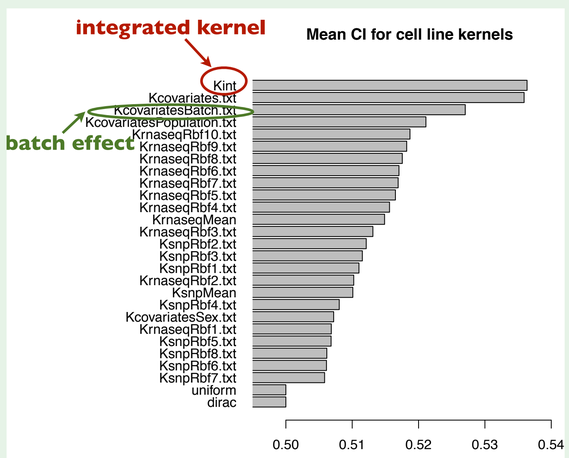
- Dirac kernel (build one model per chemical)
- Uniform kernel (build one single model)
- Kernel based on chemical features (using link to human protein, 2D substructures,...)
- Empirical kernel (evaluate the link between chemicals)
- An integrated kernel

Results

$\longrightarrow$ Evaluate the performance of each model by cross validation

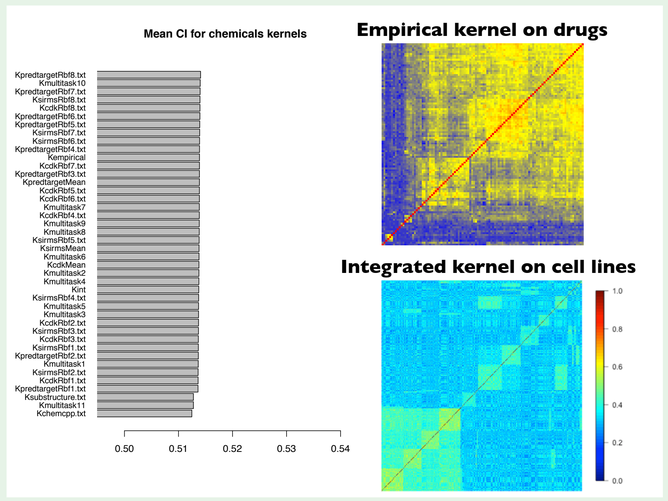


$\longrightarrow$ Integrated kernel and covariate kernel have the lowest concordance index.



- The integrated kernel has the highest prediction accuracy.
- The covariate kernel has a high prediction accuracy whereas it does not use any genomic information.
- This is due to a strong batch effect.

$\longrightarrow$ We choose the integrated kernel for cell lines.

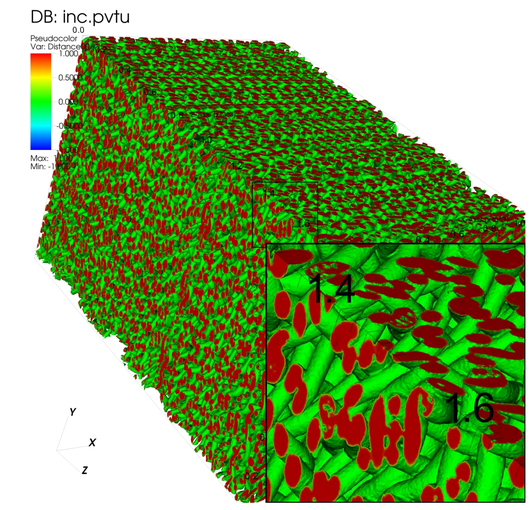


- All chemical kernels have the same accuracy : Dirac kernel performs as good as kernels using information on chemicals.
- Maybe the chemical descriptors do not fit this particular toxicogenetic problem.

$\longrightarrow$ We choose the empirical kernel for chemicals : since no kernel seems to be the best, we evaluate the link between chemicals instead of using some predetermined kernel.

$\longrightarrow$ This model ranked second out of 100 models submitted, in the DREAM challenge 8.0.

Simulation directe à partir de la datamasse des images 3D



1200x1200x1791
780 millions de nœuds
4000 cœurs de calcul

STRATEGIE GENERALE

Objectifs

- Construire des maillages de façon pertinente et efficace à partir d'imagerie 3D
- Compression de l'image, sans perte d'information, par manipulation au format maillage
- Utilisation de ces maillages dans différents domaines d'application

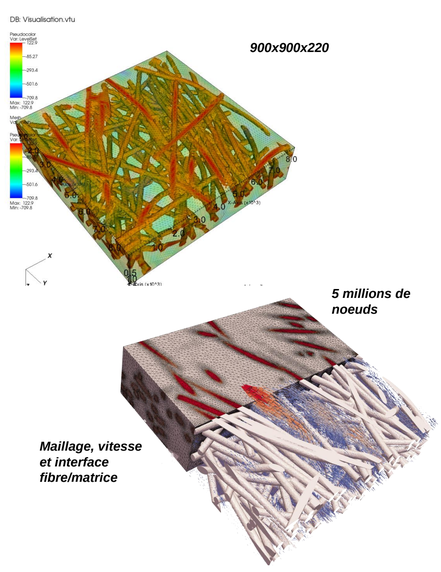
Résultats atteints / prototypes / démonstrateurs

- « Mailleur d'images »: génération avec un mailleur topologique et adaptation de maillage par minimisation de l'erreur d'interpolation de l'image sur le maillage, associée à une technique de réinitialisation de la valeur du voxel/pixel
- Au bout d'une année: génération de microstructures virtuelles, de géométries élancées en sous-sol, d'environnements urbains; calculs éléments finis et éléments frontière sur certaines applications
- Plateformes logicielles: CimLib, Morph-M, Neper, Zebulon, …

.

MATERIAUX COMPOSITES

Perméabilité d'un milieu fibreux



- Arrangement 3D irrégulier obtenu par imagerie 3D [Orgéas et al, 3S-R], caractéristiques renfort: R=0.1mm, L=10mm, ϕ=0.83
- Images acquises par microtomographie-X (900x900x220 voxels), taille du voxel = 10 μm3
- Génération du maillage éléments finis (~5 millions de nœuds): interpolation directe de l'image et adaptation, sur 96 cœurs
- Calcul d'écoulement sur le système fibre- matrice (sur 96 cœurs) et homogénéisation pour obtention de la perméabilité

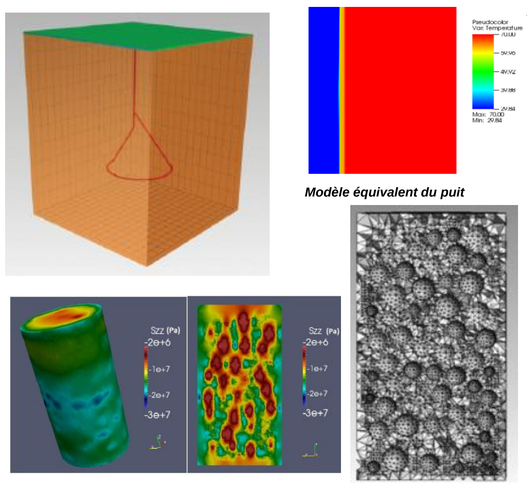
MICROSTRUCTURES METALLIQUES

Déformation d'un polycristal

- Image obtenue par tomographie de contraste à diffraction (DCT)
- Adaptation d'un algorithme de génération de cellules de Voronoï aux grains de l'image et maillage associé avec Neper
- Calculs de déformation sur l'échantillon avec Zebulon

Image obtenue par DCT

Maillage et cellules de Voronoï obtenues avec Neper



Compression simple d'un béton léger: distribution des contraintes calculées dans la méso-structure

GEOSCIENCES

Modélisation d'un échangeur géothermique

- Ecoulement d'eau dans un tube avec gainage (et couplage avec la température), validation CimLib/Fluent en 2D pour un modèle simple

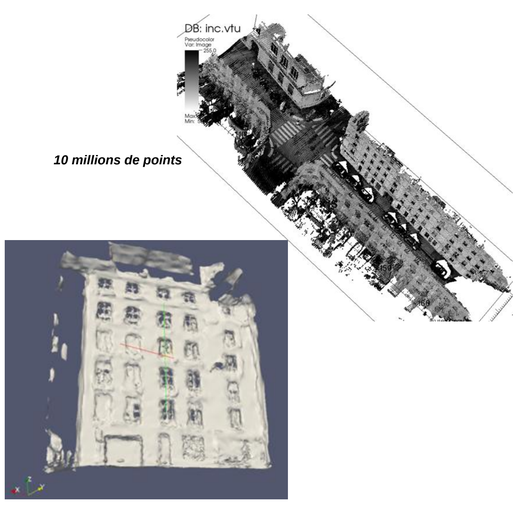
Modélisation de la de la méso-structure des bétons légers

- Compression d'un béton avec un calcul à l'échelle de la méso-structure pour obtenir le module de Young et proposer un mécanisme pour la rupture des bétons légers

MODELISATION DES ENVIRONMENTS URBAINS

Génération de maillages 3D à partir de nuages de points

- Rue de Paris, capture obtenue avec un Velodyne, 10 millions de points
- Compression de l'information par utilisation d'un maillage 3D surfacique, mais aussi 3D volumique (immersion dans une géométrie volumique)
- Applications: calculs sur ces environnements



Parties prenantes

MINES ParisTech
AGERE ECOLE DES MINES DE PARIS CENTRE DE ROBOTIQUE COOPERATIVE PERVASIVE
Cemef CENTRE DE MISE EN FORME DES MATERIAUX
CMM
GEOSCIENCES
CENTRE DES MATERIAUX PIERRE-MARIE FOURT
MINES Albi-Carmaux
Institut Clément Ader
MINES Alès
POLE MATERIAUX ET STRUCTURES DU GENIE CIVIL
MINES Douai
Technologie des Polymères et Composites Ingénierie Mécanique
SMS SCIENCES DES MATERIAUX ET DES STRUCTURES
Ecole Nationale Supérieure des Mines SAINT-ETIENNE

Auteurs

Luisa Silva
Rima Ghazal
Daniela Craciun
Jia-Xin Zhao
Min Quan Thai
Sébastien Nadler
et al.

Partenaires

INSTITUT Mines-Télécom
INSTITUT CARNOT M.I.N.E.S
GENCI GRAND EQUIPEMENT NATIONAL POUR LE CALCUL INTENSIF
PRACE

Numérique : Grande échelle & Complexité

Mars 2014

Contact luisa.silva@mines-paristech.fr
Site web www.cemef.mines-paristech.fr

MoGDIW, an integrated workflow for cell motility genes discovery in high-throughput time-lapse screening data

[1,2,3]Alice Schoenauer Sebag, [2]Céline Raulet-Tomkiewicz, [2]Robert Barouki, [1]Jean-Philippe Vert, [1]Thomas Walter

[1]Center For Computational Biology, Mines ParisTech ; INSERM U900 ; Institut Curie

[2]INSERM U747 ; Paris V

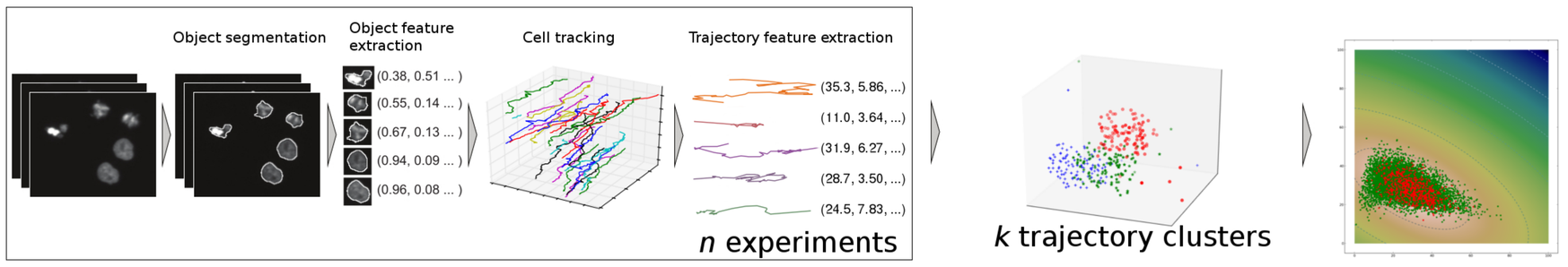
[3]Agro ParisTech

December 2013

Abstract : Cellular migration is a fundamental biological process. Progress in the fields of gene silencing and high-throughput (HT) microscopy has only recently made its study possible on a large scale. However, all existing HT migration screens measure motility at the level of cell population. Here, we present MoGDIW, a generic integrated workflow which addresses cell motility genes discovery in HT time-lapse screening data at single cell level. It is composed of cell tracking, cell trajectory mapping to an original feature space, migration pattern identification, and discriminant characterization of each experiment in terms of migratory behaviours. In comparison with an existing migration screen, MoGDIW application to a genome-wide time-lapse screen shows little overlap. However, its results are enriched in migration and adhesion-related (MAR) genes, and could be visually confirmed.

Motility **G**enes **D**iscovery **I**ntegrated **W**orkflow

Aim : quantitavely assess and compare single cell migration under different chemical perturbations in time-lapse microscopy data



Cell tracking

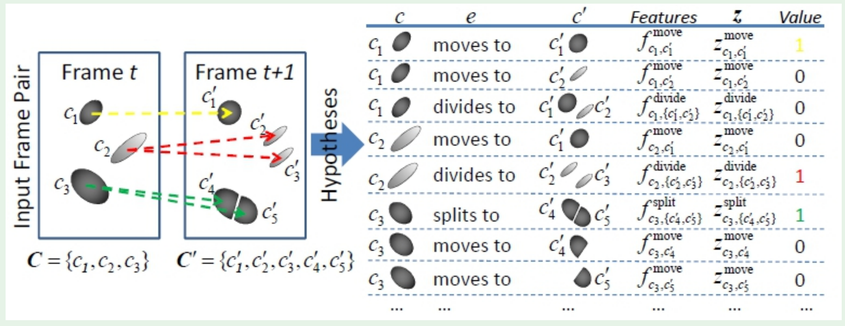
Challenges :

- High population density and high phenotypic variability
- Frequent absence of overlaps in consecutive frames (time interval : 30')
- Appearing, disappearing, merging and splitting objects
- Minimum use of prior knowledge
- *Usable for non-experts*

$\Longrightarrow$ Strategy : use of Machine Learning

Structured learning for cell tracking

- Idea (from [3]) :



- Prior knowledge : possible events **e** which can occur to an object between two frames

$$\{move,\ appear,\ disappear,\ split\ in\ 2\ or\ 3,\ merge\ at\ 2\ or\ 3\}$$

- Data : consecutive frames with already segmented objects
- Target : learn the model on annotated data and use it.
  - Model :

$$\hat{\mathbf{z}}(\mathbf{t}) = \arg\max_{\mathbf{z}}\ \mathbf{L}(\mathbf{x}(\mathbf{t}), \mathbf{z}; \mathbf{w}) = \sum_{\substack{\mathbf{e} \\ \mathbf{Obj_{i,t}} \\ \mathbf{Obj_{j,t+1}}}} < \mathbf{w^e}, \mathbf{f^e_{i,j}} > \mathbf{z^e_{i,j}}$$

$$\mathbf{st}\ \ \forall \mathbf{i} \in \{\mathbf{0}, \ldots, \mathbf{N(t)}\} \sum_{\substack{\mathbf{e} \\ \mathbf{Obj_{j,t+1}}}} \mathbf{z^e_{i,j}} = \mathbf{1}$$

$$\mathbf{and}\ \ \forall \mathbf{j} \in \{\mathbf{0}, \ldots, \mathbf{N(t+1)}\} \sum_{\substack{\mathbf{e} \\ \mathbf{Obj_{i,t}}}} \mathbf{z^e_{i,j}} = \mathbf{1}$$

  - Learning **w** : Support Vector Machine (SVM, algorithm : bundle method [6])

Results

- Training set : $\sim$ **32 000** links with 0.5% appear ; 0.5% disappear ; 1% merge ; 2% split

Algorithm (software)	Mean recall	Mean precision	Ref
Constrained nearest neighbour (Cell Cognition)	72.7%	62.8%	[1]
Linear assignement problem (Cell Profiler)	78.3%	73.0%	[2]
Structured learning	91.1%	91.5%	[3]

Providing a graphical user-interface for annotating videos

- Integration in Cell Cognition [1], an open source software platform for the analysis of live cell imaging data, with the IMBA, Vienna
- Extensions with regard to cell track annotation :
  - Generation of cell tracks rough estimation using a Nearest Neighbor tracker
  - GUI extension to support manual correction of the Nearest Neighbor trajectories

Experiment characterization in terms of migratory patterns

Trajectory mapping to an original feature space :

TABLE: Feature types and examples of corresponding features

Goal	Examples of feature
Track characterization	Diffusion coefficient, persistence, track evenness
Static quantification	Convex hull area, largest move, total path length
Dynamic quantification	Mean acceleration, mean instantaneous speed

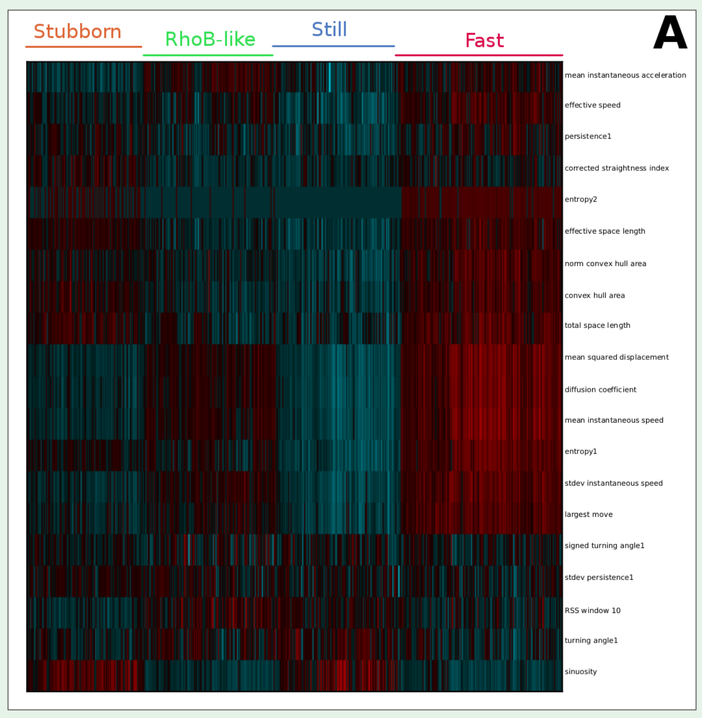
Experiment characterization :

1. Clustering of all trajectories across experiments
2. Characterization of an experiment by the cluster histogram of its trajectories
3. Pearson's χ^2 test for testing significant deviation from control histogram

Analysis of Mitocheck screen

- Data : MitoCheck data set [4], $\sim$ **200,000** videos of HeLa cells, produced by high-throughput live cell imaging, following selective down-regulation of all protein coding genes, one by one, by RNA interference (RNAi)
- Gene subset : 1,081 genes previously selected by [5] consisting of
  - phosphatases
  - kinases
  - migration and adhesion-related genes (MAR genes, a priori selected by the Geiger Laboratory, Weizmann Institute, Israël)
- Method : MoGDIW with K-means, k=4
- Results in comparison with [5] :
  - Small overlap of selected gene lists
  - No enrichement in known migration genes in either case
  - Enrichment of MoGDIW high-confidence list in MAR genes (61% vs 34% for [5])

FIGURE: A. Identification of MoGDIW cluster characteristics using single linkage hierarchical clustering ; B. Venn diagram comparing MoGDIW's and [5]'s high-confidence genes ; C. Trajectory examples



Perspectives

- Application to the whole Mitocheck dataset and biological validation of hit genes
- Application to newly generated Environmental Toxicology data to assess the consequences of chemical exposure on single cell migration

References

1. *Held et al.*, Nature Methods, 7(9) :747-54, 2010.
2. *Jaqaman, K. et al*, Nature Methods, 5(2008) :695-702.
3. *Lou et Hamprecht*, NIPS, 2011.
4. *Neumann et al*, Nature, 464 :721-727, 2010.
5. *Simpson et al*, Nat. Cell Biol., 10 :1027-1038, 2008.
6. *Tsochantaridis et al*, JMLR, 6 :1453-1484, 2005.

De l'amont vers l'aval

Modélisation et Simulation Numérique
Calcul Intensif et ses Applications

Adaptation de maillage anisotrope, calcul parallèle et capture d'interface

Applications au matériaux et à la mise en forme, mais pas seulement...

PARTENAIRES

Entre autres...

PUBLICATIONS

Quelques unes...

Digonnet, *App Math Model* (2000)
Bruchon, *Int Journ Num Meth Eng* (2009)
Hachem, *Journ of Comp Physics* (2010)
Coupez, *Journ of Comp Physics* (2011)
Ville, *Int Journ Num Meth Fluids* (2011)
Silva, *Int Journ Mat Form* (2012)
Bernacki, *Scripta Mat* (2011)
Tillier, *Int Journ Or Maxi Surg* (2012)
Carozzani, *Met Mat Trans* (2013)

REMERCIEMENTS

GENCI, PRACE, ANR

CONTACT

luisa.silva@mines-paristech.fr

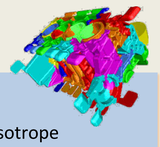
www.cemef.mines-paristech.fr

CEMEF MINES ParisTech
CS 10207
06904 Sophia Antipolis - France

CimLib

C++ et calcul scientifique

Caractéristiques phare



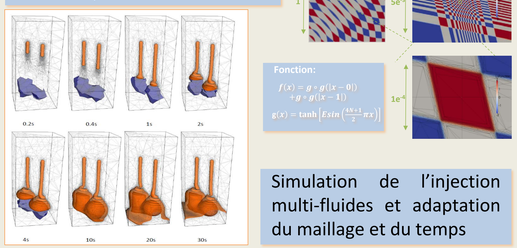
- une base éléments finis
- haut degré de parallélisme
- adaptation de maillage et du temps anisotrope
- méthodes de stabilisation performantes
- calcul de surfaces libres ou interfaces par des méthodes robustes
- validation automatique et mise à disposition sur un serveur de partage de projet
- utilisée pour le développement de logiciels industriels: Rem3D, Thost, Ximex, Forge, Thercast, Transweld

Adaptation anisotrope

Maillage et temps

Automatique, en parallèle et en dynamique

2h, 23 millions de nœuds, 10 itérations, 256 cœurs

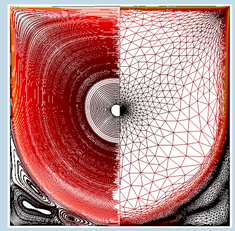


Simulation de l'injection multi-fluides et adaptation du maillage et du temps

Méthodes numériques avancées

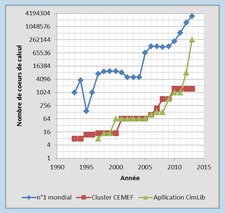
▶ Génération et adaptation de maillages anisotropes

- Adaptation anisotrope en dynamique et en parallèle (3D) basée sur une carte de métrique et sur des estimateurs d'erreurs anisotropes 3D sur le gradient des fonctions ou sur la distribution de la longueur des arêtes, avec contrôle du nombre d'éléments



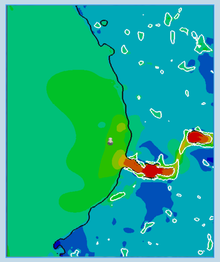
▶ Calculs massivement parallèles

- Haute performance des simulations grâce à l'exploitation efficace de la parallélisation de la gestion du maillage et de la résolution des systèmes linéaires
- Repartitionnement dynamique, solveurs itératifs parallèles et multigrilles
- Benchmarking et applications dans des supercalculateurs du Tiers1 (Genci) et Tiers0 (Prace)



▶ Ecoulements multiphasiques, interactions fluide-structure et interfaces mobiles

- Approche monolithique et méthode des volumes immergés
- Méthodes éléments finis stabilisées pour les écoulements du très petit (très visqueux) au très haut (peu visqueux) nombre de Reynolds, avec couplages thermiques et cinétiques
- Interactions thermomécaniques entre les phases, changement de phase
- Evolution des interfaces par des approches type level-set

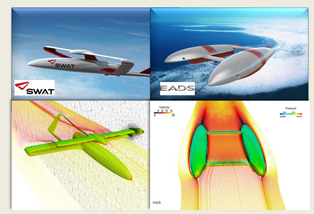


Aérodynamique et aérothermie

Ecoulements turbulents, transferts thermiques

Interactions avec les structures, fixes ou mobiles

Ecoulement d'air
Haut Re (1^{e8})

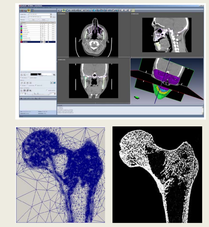


Drone léger à aile bouclée et dirigeable à 7000 m

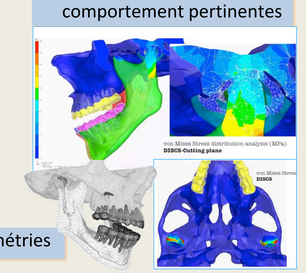
Simulation et le vivant

Biomécanique

Tissus mous ou structures osseuses



Calcul mécanique et lois de comportement pertinentes

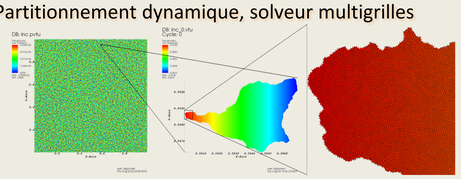


Reconstruction des géométries

Massivement parallèle

Supercalculateurs et supercalculs

Partitionnement dynamique, solveur multigrilles



Génération, adaptation de maillage

dim	# éléments (milliards)	# coeurs
2D	66,8	100 000
3D	81,4	100 000

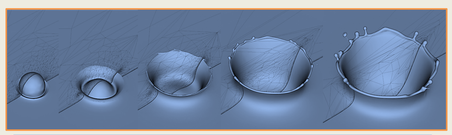
Résolution d'un très grand système linéaire

dim	# inconnues (milliards)	temps calcul	# coeurs
2D	100	319 s	262 144
3D	55	447 s	65 536

Ecoulement multiphasique

Interfaces implicites évolutives

Approche monolithique et méthode level-set



Evolutions complexes: chute d'une goutte de lait

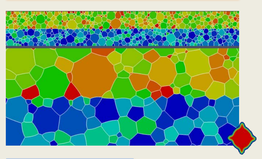
Injection assistée-eau: transport multi level-set

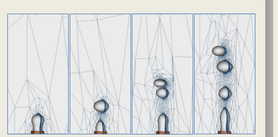
Physique numérique

Changements de phase

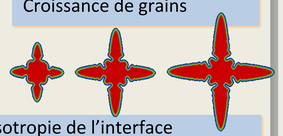
Germination et croissance

Ebullition
Croissance et ascension de bulles de vapeur





Recristallisation
Croissance de grains

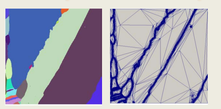


Solidification
Croissance dendritique et anisotropie de l'interface

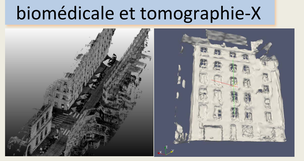
Simulation et le réel

Imagerie et Données Massives

3D volumique ou surfacique



Microstructures numériques, biomédicale et tomographie-X



Environnements urbains et nuages de points

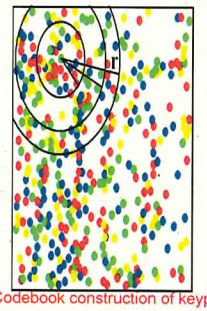
Visual textures as realizations of multivariate log-Gaussian Cox processes

Huu-Giao Nguyen, Ronan Fablet, Jean-Marc Boucher
Institut Telecom / Telecom Bretagne / LabSTICC (CS 83818 - 29238 Brest Cedex 3 – France)
{huu.nguyen, ronan.fablet, jm.boucher}@telecom-bretagne.eu

Abstract: In this paper, invariant texture characterization and recognition are addressed from the characterization of the spatial distribution of image. Visual keypoint sets in visual textures are here regarded as realizations of spatial point processes. We show that empirical second-order statistics considered in [4] relate to a non-parametric form of a log-Gaussian Cox model and investigate the relevance of parametric Cox models for texture recognition issues. Reported results validate the proposed descriptor compared to state-of-the-art approaches[1] with three datasets: UIUC, KTH-Tips, Brodatz.

Descriptive statistics of multivariate point processes

Local feature extraction



Codebook construction of keypoints[2]

A *spatial point process* **S** is defined as a locally finite random subset of a given bounded region **B** in $\mathbf{R}^2$. A realization of such a process is a spatial point pattern {s = $\{s_1,...,s_n\}$ of n points in **B**. Given realizations of apoint process, the moments of random variable are relevant descriptive statistics. The p^{th}-order moment is defined by:

$$\mu^{(p)}(B_1\times..\times B_n)=E\{N(B_1)..N(B_n)\}$$

For a marked spatial point process $\{(s_i,m_i)\}$ in *given bounded regions* B_i, where m_i is a mark associated to point s_i [3], the second-order moment is given by:

$$\alpha^2(B_1\times B_2)=E\sum_{s_1\in S}\sum_{(s_2\neq s_1)\in S}I_{B_1}(s_1)I_{B_2}(s_2)$$

Ripley's K function considers circular analyzing regions and resorts to the mean numbers of points of type **j** in a region of radius **r** centered at the points of type **i** :

$$K_{ij}(r)=(\lambda_i\lambda_j)^{-1}\alpha_{ij}^2(r)$$

$$=(\lambda_i\lambda_j)^{-1}E\sum_h\sum_{l\neq h}\delta_i(m_h)\delta_j(m_l)\mathbb{I}(\|s_h-s_l\|\leq r)$$

Log-Gaussian Cox processes

Cox processes $\{X_i\}$ with random intensity functions $\{Z_i\}$ are point processes such that $X_i|Z_i$ is a Poisson process, where $Z_i=\exp(Y_i)$. $\{Y_i\}$ is a multivariate Gaussian field on **S** characterized with mean function μ=EY(s) and covariance function $c_{ij}(r)=\mathrm{Cov}(Y_i(s_1),Y_j(s_2),r)$ [4].

• Intensity function: $\lambda_i=\exp(\mu_i+\sigma_i^2/2)$

• Pair correlation function: $g_{ij}(r)=\exp(\beta_{ij}(r))$

• Ripley's K function: $K_{ij}(R)=2\pi\int_0^R r\,g_{ij}(r)dr$

• The estimation of the pair correlation function is given by:

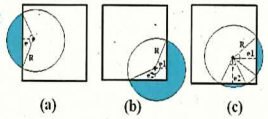
$$g_{ij}(r)=(2\pi r\lambda_i\lambda_j)^{-1}\sum_h\sum_{l\neq h}\delta_i(m_h)\delta_j(m_l)\xi(\|s_h-s_l\|,r)b_{s_h}$$

• Given a parameterization **L(β,r)**, namely Exponential, Hyperbolic or Cardinal sine, model parameters are estimated from the minimization of the following criterion:

$$\int_0^R\{\sigma_{ij}L(\beta_{ij},r)-c_{ij}(r)\}^2dr$$

The proposed descriptor is formed by: $(\lambda_i, \sigma_{ij}, \beta_{ij})$

Edge-effect corrections:



Feature dimension reduction: A codebook of keypoint pairs $u=M(s_i,s_j)$ from two categorized keypoint s_i,s_j is considered, such that:

$$g_u(r)=(2\pi r\lambda_u)^{-1}\sum_h\sum_{l\neq h}\delta_u(M(s_h,s_l))\xi(\|s_h-s_l\|,r)b_{s_h}$$

Scaling effects:The actual radius of image is estimated by a reference radius r_{ref} and scale factor φ_{ref} (the rate of average point densities per surface unit).

$$r_i=\frac{\varphi_{ref}}{\varphi_i}r_{ref}$$

Application to texture recognition

Parameter setting:

• Categories of visual keypoints k = {60,120,150}, pairs of keypoints k*=60.

• r_{ref}= min(w,h)/2 ln(x) where **w**(idth),**h**(ight) the size of image, x=(1:0.1:exp(1)].

• Covariance function **L(β,r)** : Gaussian function.

• Classifier: random forest.

Result:

	N_t	Gabor filter	Cooc matrix	BoK[4]	Keypoint cooc.[15]	Multi-Fractal [27]	Keypoint combination[28]	Descriptive stat.[21]	log-Gaussian Cox
UIUC	1	31.22±3.14	45.33±3.03	67.25±2.75	37.62±2.63	61.14±2.90	72.53±2.45	**76.66±1.65**	75.21±1.75
	10	57.37±1.93	70.67±1.72	8[illegible]2±1.45	54.14±1.72	83.68±1.65	93.7±1.15	94.23±0.78	**95.42±0.71**
	20	67.78±1.23	80.12±1.30	9[illegible]9±1.15	91.67±1.38	93.65±1.31	93.37±0.95	97.34±0.25	**97.84±0.32**
Brodatz	1	78.52±1.72	75.42±1.73	83[illegible]4±1.50	54.33±1.63	85.95±0.91	83.35±1.35	87.67±0.81	**88.81±0.92**
	3	85[illegible]4±1.41	83.22±1.04	92.73±0.91	93.17±0.87	92.41±0.73	94.34±0.45	95.67±0.33	**96.14±0.41**
KTH-Tips	5	62.32±2.41	62.83±2.42	64.42±2.81	71.67±2.63	72.63±2.45	78.7±2.35	**81.34±1.53**	80.17±2.15
	20	82.65±1.39	80.45±1.67	8[illegible]2±1.45	82.47±1.48	87.18±1.53	90.23±1.51	92.15±1.26	**92.42±1.11**
	40	89.87±0.88	88.95±0.77	90[illegible]9±0.65	91.15±1.05	91.33±0.97	94.35±0.87	95.05±0.41	**95.74±0.45**

Classification rates and standard deviations of proposed descriptor over 50 random selections compared to state-of-the-art approaches.

N_t	DoG+Sift	FH+Surf	FH+Brief	(Hes-Lap)+Daisy	(Har-Lap)+(Sift-Spin)
1	75.21±1.75	75.05±1.54	**75.43±1.71**	75.13±1.69	74.85±1.87
5	91.95±1.13	93.73±1.11	91.42±1.23	**92.13±1.19**	91.15±1.4[illegible]
10	95.42±0.71	95.15±0.91	95.22±0.85	**95.47±1.08**	95.23±[illegible].72
15	**96.87±0.65**	95.14±0.63	96.43±0.51	96.75±0.53	96.37±[illegible].6[illegible]
20	**97.84±0.32**	95.75±0.41	97.25±0.34	97.67±0.35	97.14±[illegible].37

Comparison performance of proposed model with the different detector-descriptor types on UIUC dataset.

References:

[1].J.Zhang et al. "Local features and kernels for classification of texture and object categories: a comprehensive study". IJCV, 73(2), p.213–238, 2007.
[2].G.Csurka et al. "Visual categorization with bags of keypoints". ECCV, p.1–22, 2004.
[3] H-G.Nguyen et al. "Spatial statistics of visual keypoints for texture recognition." ECCV, pp. 764–777,2010.
[4].J.Møller et al.. "Log-Gaussian cox processes," SJS, 25(3), p. 451–482, 1998.

RANDOM WALK MODELS FOR GEOMETRY-DRIVEN IMAGE SUPER-RESOLUTION

APPLICATION TO REMOTELY SENSED GEOPHYSICAL FIELDS AT OCEAN SURFACE

Ronan Fablet
Professor, Institut Mines-Télécom/Télécom Bretagne
Brest, France

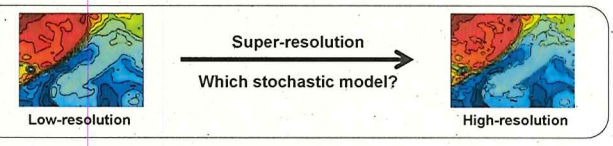
Co-authors

Brahim Boussidi
Institut Mines-Télécom/Télécom Bretagne
LabSTICC/TOMS
Brest, France

Emannuelle Autret, Bertrand Chapron
Ifremer/LOS (Space oceanography)
CERSAT
Brest, France

ABSTRACT

This paper addresses stochastic geometry-driven image models and their application to super-resolution issues for textured geophysical fields. Whereas most stochastic image models rely on some priors on the distribution of grey-level configurations (e.g., patch-based models, Markov priors, multiplicative cascades,...), we here focus on geometric priors. Regarding image level-lines as realizations of 2D random walks, we introduce a stochastic geometry-driven model for 2D textures and consider an application to image super-resolution. The targeted application is the stochastic interpolation of missing data in muti-sensor sea surface observation.



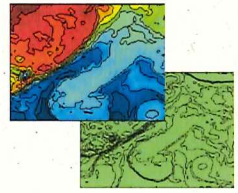


Image level-lines as realizations of 2D random walks

Why level-lines?

Image level-lines provide a contrast-invariant image representation [5] and fully characterize the geometry of an image.

Correlated random walk model [1]:

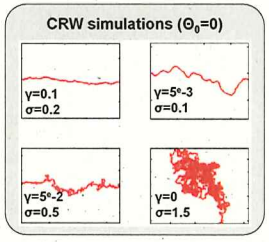
$$d\theta(s) = -\gamma\,(\theta(s) - \theta_0) + \sigma dW(s) \quad (1)$$

Directional drift — Brownian process

Parameters γ and σ control the regularity of the random walk, in terms of regularity along the walk and of oscillation around the directional drift. The CRW model is associated with a Fokker-Plank representation, from which one can derive the stationary statistics of the random direction θ and turning angle δθ [1].

$$p(\theta) \propto exp\left(-\frac{\gamma}{\sigma^2}(\theta - \theta_0)^2\right)$$

$$p(\delta\theta) \propto exp\left(-\alpha\frac{\gamma}{\sigma}\delta\theta^2\right) \quad \text{with} \quad \alpha = \Delta s\,(1 - exp(-\gamma\Delta s))$$



Stochastic geometry-driven texture model

Stochastic geometry field model as a generalization of (1) for orientation fields:

$$\nabla_\theta\theta = -\gamma\,(\theta - \theta_0) + \sigma\nabla_\theta W \quad (2)$$

Reference orientation — 2D Brownian sheet

Where ∇_θ is the gradient in direction θ

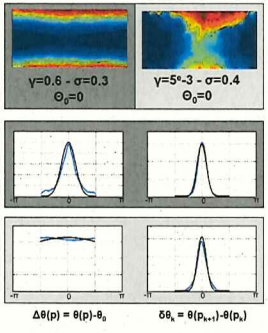


Image I such that its level-lines are everywhere tangent to random vector field u_θ:

$$\langle \nabla I(p), u_\theta(p) \rangle = 0, \forall p \quad (3)$$

Implementation:

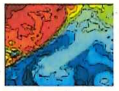
1)Simulate an orientation field from a numerical integration of the stochastic equation,

2)Solve for (3) as a variational minimization (cf. [3]).

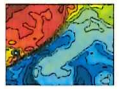
Application to texture-based super-resolution of geophysical fields

Problem statement

I_{HR}: High-resolution image
NxM grid



I_{LR}: Low-resolution image
N/KxM/K grid (here, K=2⁴)



Stochastic super-resolution model

Given a low-resolution image I_{LR}, sample a high-resolution image such that:

$$\begin{cases} \nabla_\theta\theta(p) = -\gamma(p)\,(\theta(p) - \theta_{LR}(p))\,dp + \sigma(p)\nabla_\theta W(p) \\ \tilde{I} = \arg\min_I \int \| \langle \nabla I(p), u_\theta(p) \rangle \| dp \quad \text{subject to } I_{LR} = \mathcal{P}\left[\tilde{I}\right] \end{cases}$$

with W a Brownian sheet and θ_{LR} the orientation field of the low-resolution image (i.e., the angle of the local tangent to the level-lines). $\mathcal{P}$ is an orthogonal projection operator (here a wavelet operator) such that:

$$I_{LR} = \mathcal{P}\,(I_{HR}) \quad \text{and} \quad \mathcal{P}\,[I_{HR} - \mathcal{P}(I_{HR})]$$

The low-resolution gradient drives the high-resolution geometrical variability

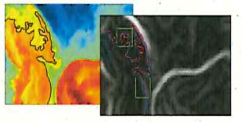
Parameter fields γ and σ are set according to the following observation :

- Large gradients in I_{LR} result in more regular level-lines in I_{HR};
- Conversely, weak gradients in I_{LR} involve regular level-lines in I_{HR}.

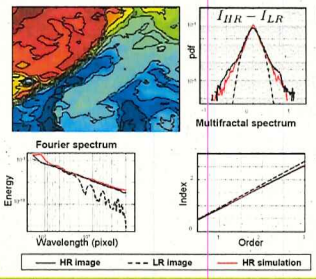
$$\gamma(p) = \gamma_0 \|\nabla I_{LR}(p)\|^\nu \quad \sigma(p) = \sigma_0 \|\nabla I_{LR}(p)\|^{-\beta}$$

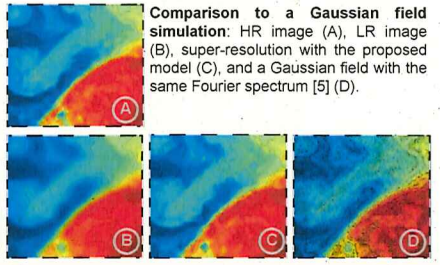
(Here, we set empirically γ_0=2.14, ν=0.54, σ_0=0.14, β= 0.13)

Level-lines vs. $\|\nabla I_{LR}\|$



Simulation for sea surface temperature observations





Comparison to a Gaussian field simulation: HR image (A), LR image (B), super-resolution with the proposed model (C), and a Gaussian field with the same Fourier spectrum [5] (D).

References

[1] Codling, E.A., Plank, M.J., Benhamou, S. Random walk models in biology. J. R. Soc. Interface, 5:813-834, 2008.

[2] Fablet, R., Pujolle, S., Chessel, A., Benzinou, A., Cao F. 2D Image-based reconstruction of shape deformation of biological structures using a level-set representation. CVIU, 111(3):295-306, 2008.

[3] Freeman, W.T., Liu, C. Markov Random Fields for Super-resolution and Texture Synthesis. In A. Blake, P. Kohli, and C. Rother, eds., Advances in Markov Random Fields for Vision and Image Processing, Chapter 10. MIT Press, 2011.

[4] Galerne, B., Gousseau, Y., Morel., J., 2011. Random phase tex- tures : Theory and synthesis. IEEE Transactions on Image Processing. 20(1), 257–267.

[5] Monasse, P., Guichard, F. Fast Computation of a Contrast Invariant Image Representation. IEEE TIP, 9:860- 872, 2000.

Contact ronan.fabet@telecom-bretagne.eu Webpage: perso.telecom-bretagne.eu/ronanfablet

New Decoding Strategy For Underdetermined MIMO Transmission Using Sparse Decomposition

Yasser Fadlallah, Abdeldjalil Aïssa-El-Bey, Karine Amis, Dominique Pastor and Ramesh Pyndiah
Telecom Bretagne, Brest, France

Parties prenantes

Auteurs

Yasser Fadlallah

Abdeldjalil Aïssa-El-Bey

Karine Amis

Dominique Pastor

Ramesh Pyndiah

Partenaires

Introduction

- Maximum Likelihood joint detection enables to detect at once the symbols transmitted in the same time intervals.
- ML detector selects the closest point to the received signal in the receive constellation.
- ML is optimal for medium to high SNR values.
- Computation cost increases exponentially with the signal dimension.
- Alternative solution such as sphere decoder keeps an exponential increase of the computation cost.
- **Goal:** find out an efficient detection problem of the transmitted symbols with moderate computation cost.

System Model

- MIMO flat fading channel.
- Perfect knowledge of the channel matrix at the receiver.
- Transmit symbols belong to a finite alphabet constellation.
- Received signal is defined as

Sparse Decomposition

- Let $Q=\{q_1,...,q_M\}$ the finite alphabet transmit constellation. Let $\mathbf{q}=[q_1,...,q_M]$ the vector space in which the finite alphabet vector can be cast, and $\mathbf{B}_q$ the decomposition matrix defined as

$$\mathbf{B}_q=\begin{pmatrix}\mathbf{q} & \mathbf{0}_M & \cdots & \mathbf{0}_M\\ \mathbf{0}_M & \mathbf{q} & \ddots & \vdots\\ \vdots & \ddots & \ddots & \mathbf{0}_M\\ \mathbf{0}_M & \cdots & \mathbf{0}_M & \mathbf{q}\end{pmatrix}$$

- The transmitted symbol vector $\mathbf{x}$ can be rewritten after symbol decomposition on $\mathbf{q}$ as : $\mathbf{x}=\mathbf{B}_q\,\mathbf{s}$
- The received signal is reformulated as: $\mathbf{y}=\mathbf{H}\mathbf{B}_q\,\mathbf{s}+\mathbf{z}.$

New MIMO Detector based on ℓ_1-norm minimization

- The new problem is the decoding of the binary source vector $\mathbf{S}$. To this end, we propose to solve it using the following minimization problem

$$\arg\min_{s\in R^{NM\times 1}}\|\mathbf{s}\|_0,\quad \text{subject to}\ \mathbf{s}\in\left\{\left\|\mathbf{y}-\mathbf{H}\mathbf{B}_q\mathbf{s}\right\|<\varepsilon,\ \text{and}\ \mathbf{B}_1\mathbf{s}=1_N\right\}$$

where ε is a constant defined later, and $\mathbf{B}_1=\begin{pmatrix}1_M & \mathbf{0}_M & \cdots & \mathbf{0}_M\\ \mathbf{0}_M & 1_M & \ddots & \vdots\\ \vdots & \ddots & \ddots & \mathbf{0}_M\\ \mathbf{0}_M & \cdots & \mathbf{0}_M & 1_M\end{pmatrix}$

- In the literature of sparse reconstruction, the ℓ_0-norm can be relaxed by the ℓ_1 -norm minimization, and the problem is resolved using

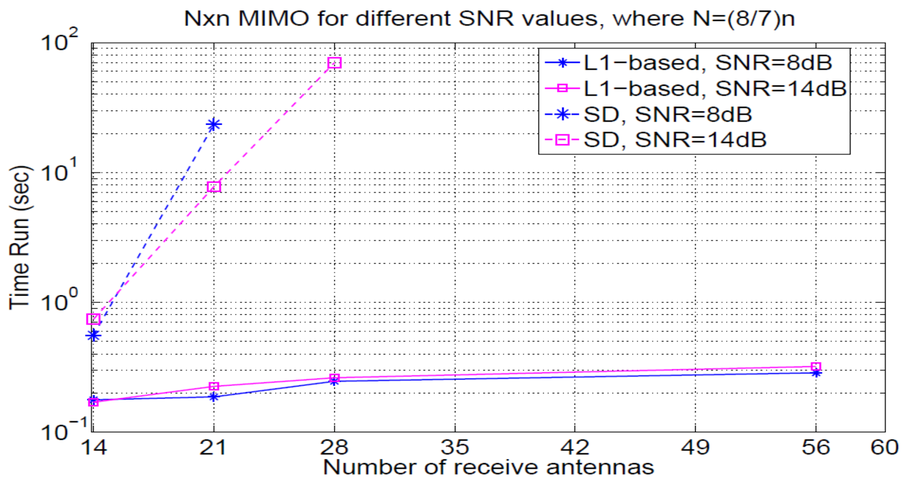
$$\arg\min_{\mathbf{s}\in R^{NM\times 1}}\|\mathbf{s}\|_1,\quad \text{subject to}\ \mathbf{s}\in\left\{\left\|\mathbf{y}-\mathbf{H}\mathbf{B}_q\mathbf{s}\right\|<\varepsilon,\ \text{and}\ \mathbf{B}_1\mathbf{s}=1_N\right\}$$

Applications

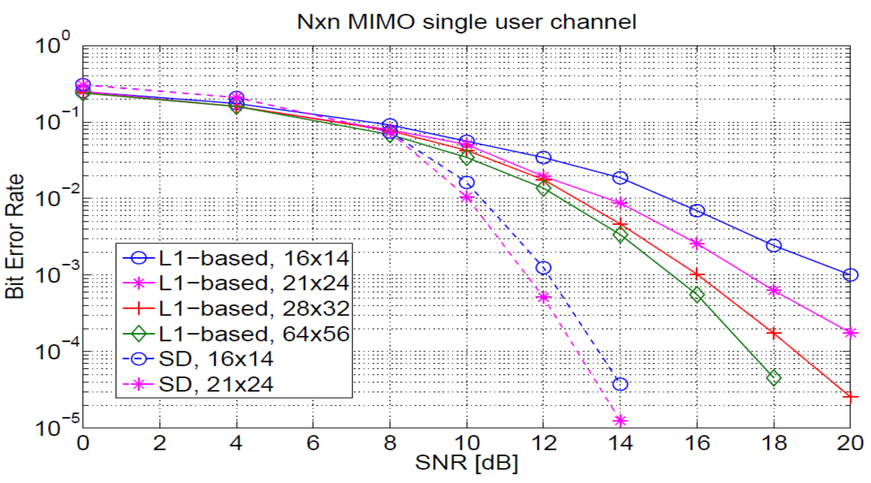
- Large MIMO systems
  - In a noiseless channel, the equivalence between the ℓ_0-norm and the ℓ_1-norm hold for large dimensions of ***s***.
- MIMO frequency selective channel
  - The received signal can be written as

$$\begin{pmatrix}\mathbf{y}(1)\\ \vdots\\ \mathbf{y}(T_f+L)\end{pmatrix}=\begin{pmatrix}\mathbf{H}_0^T & \cdots & \mathbf{H}_L^T & \mathbf{0}^T_{n\times N}\cdots\mathbf{0}^T_{n\times N}\\ \mathbf{0}^T_{n\times N} & \ddots & \ddots & \ddots\ddots\vdots\\ \vdots & \ddots & \ddots & \ddots\ddots\vdots\\ \mathbf{0}^T_{n\times N} & \cdots & \mathbf{0}^T_{n\times N} & \mathbf{H}_0^T\cdots\mathbf{H}_L^T\end{pmatrix}\begin{pmatrix}\mathbf{x}(1)\\ \vdots\\ \mathbf{x}(T_f)\end{pmatrix}+\begin{pmatrix}\mathbf{z}(1)\\ \vdots\\ \mathbf{z}(T_f+L)\end{pmatrix}$$

Simulations Results



- We assume 4-QAM modulation.
- The computational complexity keeps almost invariant with the system dimensions and the SNR level, whereas the SD time-run increases exponentially with these two factors



Conclusions

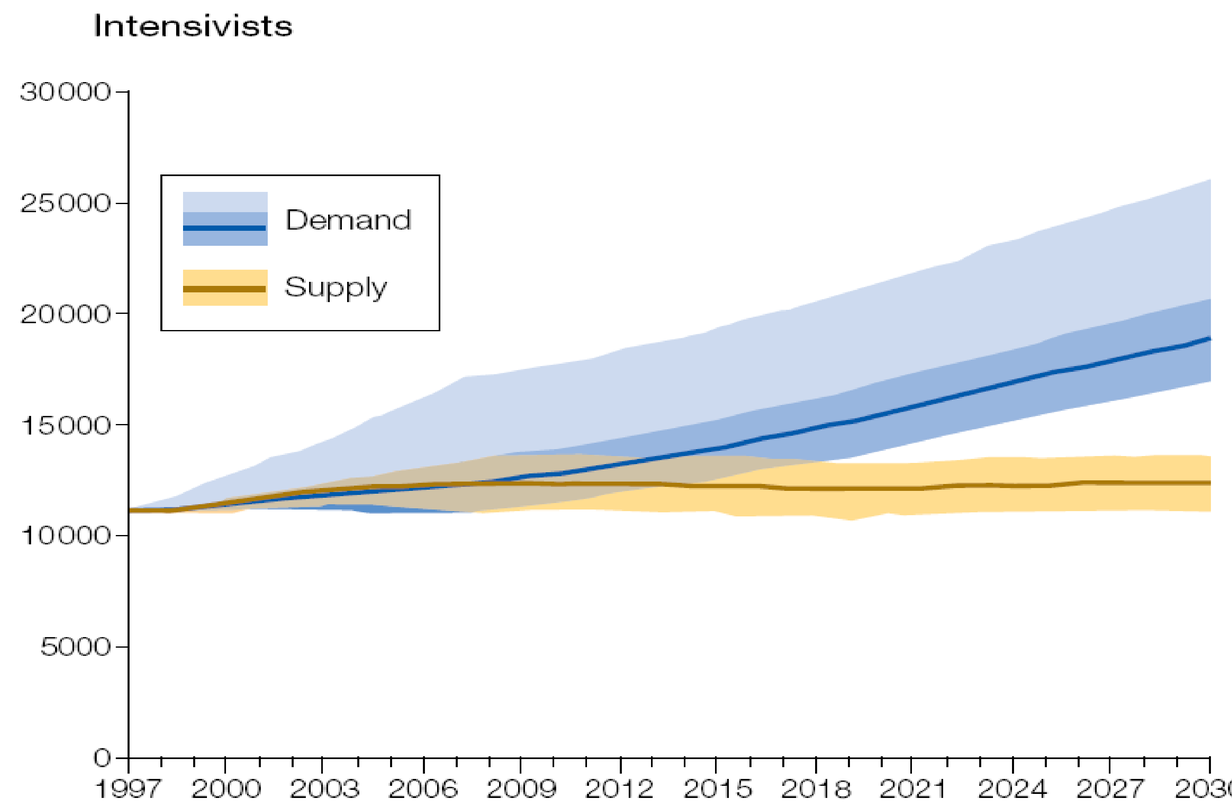
- Proposition of a new detection method for determined and underdetermined MIMO systems, based on sparse decomposition of the signal belonging to a finite constellation.
- The proposed detection method is solvable in polynomial time, and uses iterative algorithm such as primal dual interior point method.

Colloque « Numérique : Grande échelle & Complexité »
26-27 Mars 2014

Contact abdeldjalil.aissaelbey@telecom-bretagne.eu Site web www.telecom-bretagne.eu

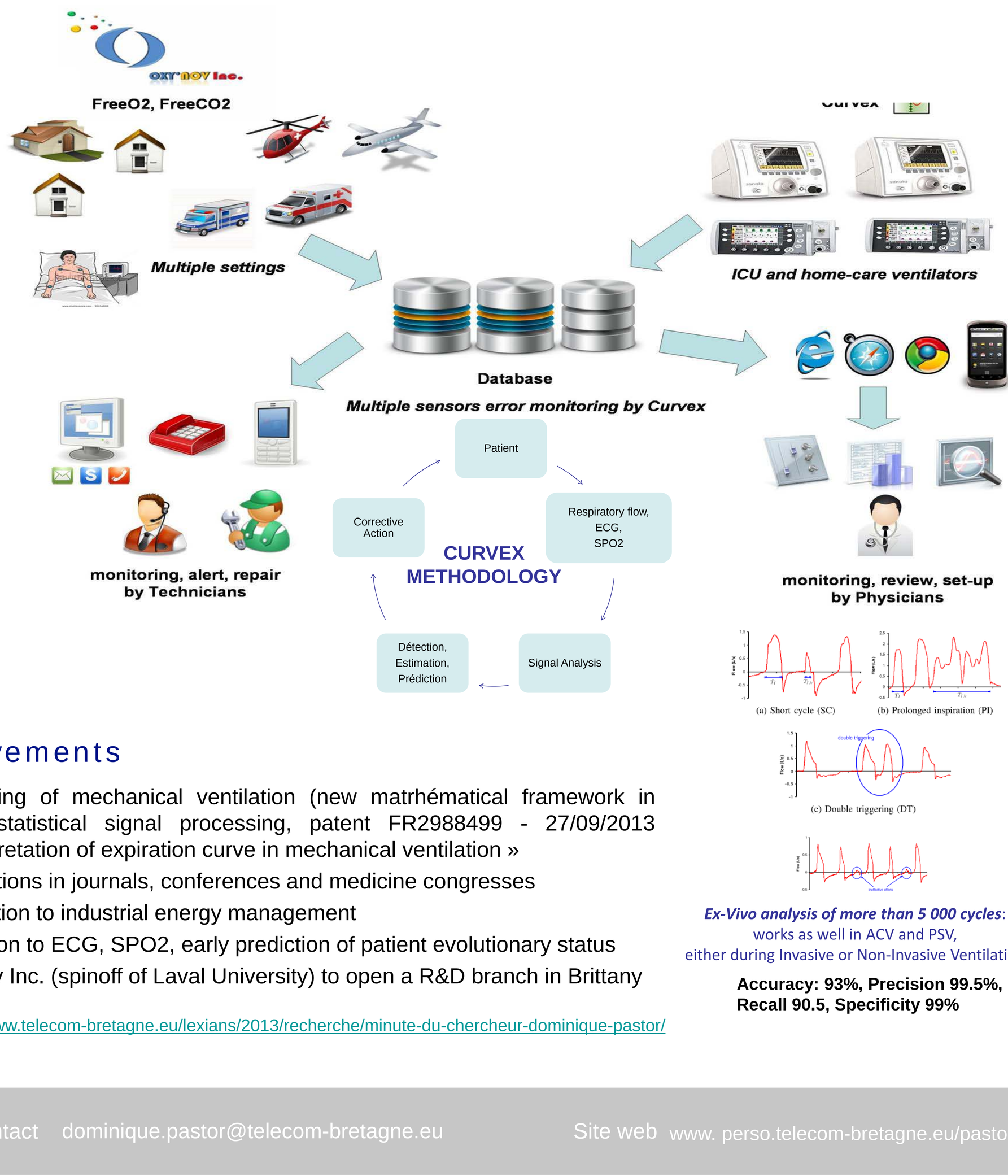
Plateforme CURVEX de surveillance des signaux physiologiques en médecine d'urgence et réanimation

Hospital System staffing under pressure



Source: Angus JAMA 2000

The CURVEX solution



Auteurs


Q.-T. Nguyen (TB/Lab-STICC)


S. Cherif (TB/Lab-STICC)


D. Pastor (TB/Lab-STICC)


E. L'Her (UBO/CHU/LATIM)

Partenaires

Achievements

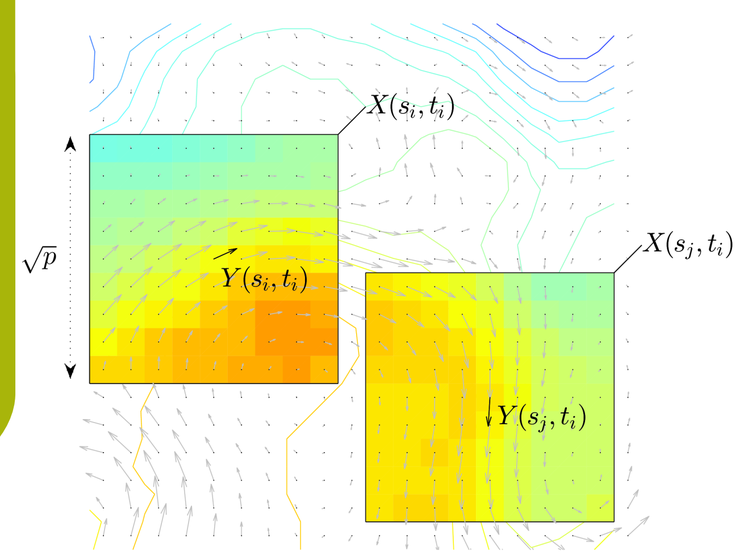
- Monitoring of mechanical ventilation (new matrhématical framework in robust statistical signal processing, patent FR2988499 - 27/09/2013 « interpretation of expiration curve in mechanical ventilation »
- Publications in journals, conferences and medicine congresses
- Application to industrial energy management
- Extension to ECG, SPO2, early prediction of patient evolutionary status
- Oxy'nov Inc. (spinoff of Laval University) to open a R&D branch in Brittany

http://www.telecom-bretagne.eu/lexians/2013/recherche/minute-du-chercheur-dominique-pastor/

Ex-Vivo analysis of more than 5 000 cycles:
works as well in ACV and PSV,
either during Invasive or Non-Invasive Ventilation

Accuracy: 93%, Precision 99.5%, Recall 90.5, Specificity 99%

Une approche statistique pour la caractérisation et le suivi des dynamiques superficielles des océans à partir d'images satellitaires



Auteurs

Pierre Tandeo
Sileye Ba
Ronan Fablet
Institut Mines-Telecom
Telecom Bretagne
LabSTICC – TOMS
Brest, France

Bertrand Chapron
Emmanuelle Autret
Ifremer
Laboratoire d'océanographie spatiale
CERSAT
Brest, France

Partenaires

Ifremer

Lab-STICC

Remerciements

AVISO

Remote Sensing Systems
www.remss.com

Références

[1] Isern-Fontanet et al. Potential use of microwave sea surface temperatures for the estimation of ocean currents, GRL, vol. 33, pp. L24608, 2006

[2] Lapeyre et al. Dynamics of the upper oceanic layers in terms of surface quasigeostrophy theory, JPO, vol. 36, pp. 165--176, 2006

[3] De Sarbo et al. A maximum likelihood methodology for clusterwise linear regression, Journal of Classification, vol. 5, pp. 249--282, 1988

Résumé

Les mesures satellitaires de courants (U,V) et de température de l'eau de surface (SST), fournissent une information sur les dynamiques de l'océan. Certaines études (cf. **[1]** et **[2]**) ont montré que les champs de température peuvent être considérés, dans certaines situations, comme des traceurs actifs de la dynamique de surface. Dans ce cas, de fortes corrélations existent entre les variations locales de SST et les courants (U,V). Existe-t-il d'autres relations entre la température et les courants ? Quand et où la SST peut-elle être considérée comme un traceur actif ou passif ? Dans cette étude, nous mettons en place une méthode statistique et explorons un historique d'observations satellitaires pour identifier et suivre des modes dynamiques cachés.

Méthode

- K fonctions de transfert cachées entre :
  - **Y** → courant (U,V) en un point
  - **X** → température SST au voisinage (patch)

$$\mathbf{Y}(s_i,t_i) = \sum_{k=1}^{K} \mathbf{H}_k(\mathbf{X}(s_i,t_i))$$

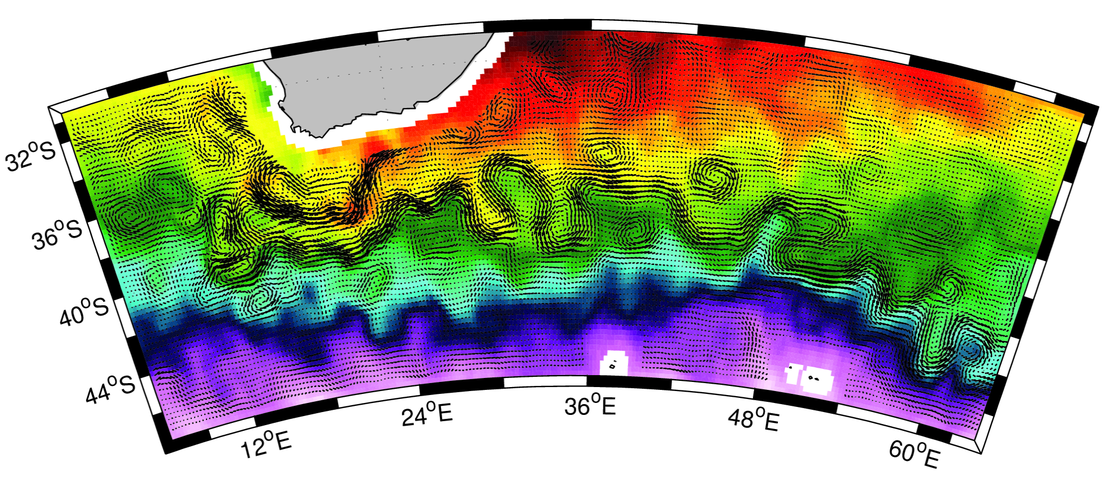
- Identification des K fonctions de transfert à partir d'un modèle de régressions linéaires latentes (cf. **[3]**) :

$$p\left(\mathbf{Y}|\mathbf{X},\boldsymbol{\theta}\right) = \sum_{k=1}^{K} \lambda_k \mathcal{N}_k\left(\mathbf{Y};\mathbf{X}\boldsymbol{\beta}_k,\boldsymbol{\Sigma}_k\right)$$

- Estimation des paramètres par l'algorithme EM
- Suivi des modes dynamiques à partir des cartes de probabilités *a posteriori*

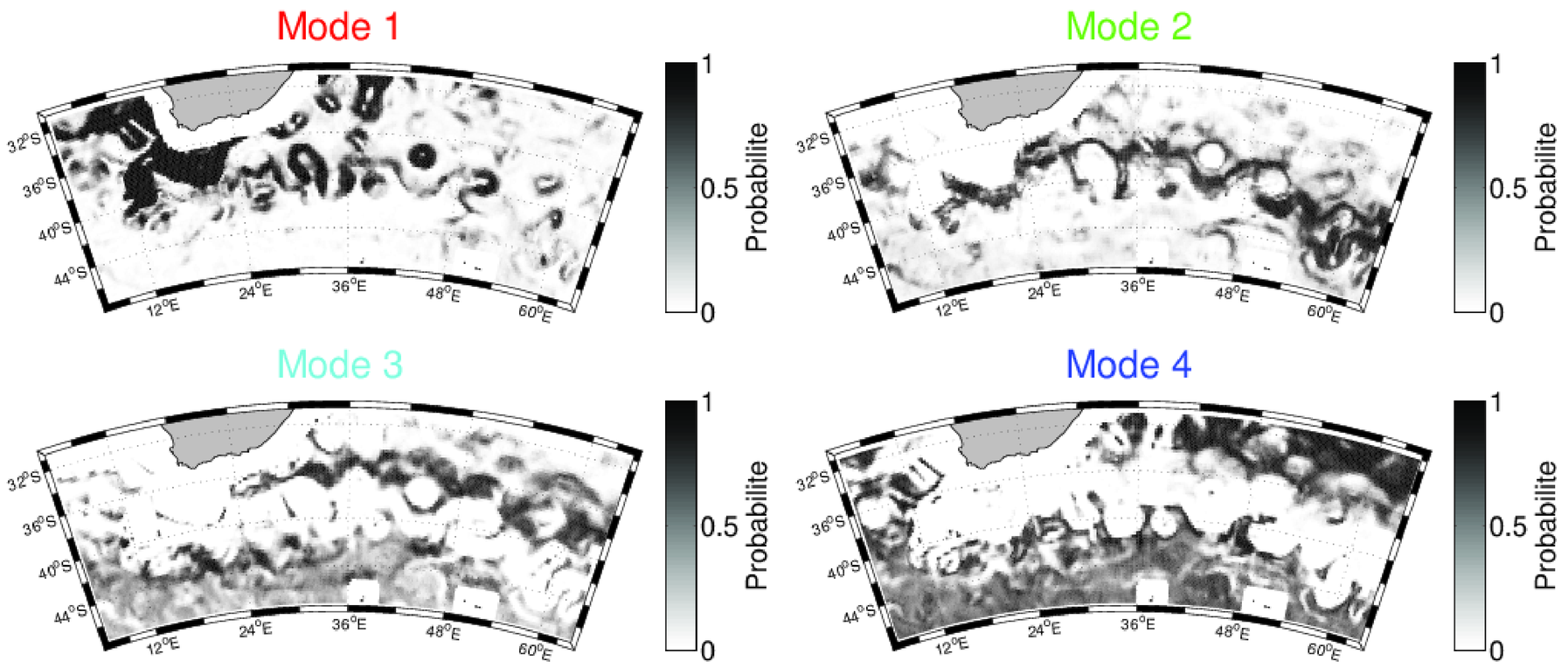
Données

- Température SST :
  Produit interpolé RSS
  Jounalier au 1/4°
- Courant (U,V) :
  Produit interpolé AVISO
  Journalière au 1/3°



Résultats

	Mode 1	Mode 2	Mode 3	Mode 4
Courant (intensité)	fort (principal Aiguilles + tourbillons)	fort (segondaire Aiguilles)	faible	faible
Courant (sens)	Nord-Sud	Ouest-Est	Est-Ouest	Ouest-Est
Température	élevée	élevée	moyenne	froide
SQG-like		√		√



Conclusion

La méthode proposée permet un apprentissage à l'aveugle, sans *a priori* physique, de fonctions de transfert cachées entre la SST et (U,V). Ces relations cachées correspondent à différents modes dynamiques dont certains s'apparentent à la théorie SQG. A partir du calcul des probabilités *a posteriori*, nous pouvons suivre l'évolution spatio-temporelle de ces modes dynamiques.

Perspectives

L'utilisation d'autres traceurs actifs tels que la salinité (SSH) ou la couleur de l'eau (Chl-a) ainsi que des données à haute résolution spatiale permettrait un rafinement du modèle. De plus, le suivi des probabilités *a posteriori* par un modèle stochastique et l'utilisation des K fonctions de transfert permet d'envisager une estimation des courants de surface à partir de la température.

GRETSI
Septembre 2013

Contact pierre.tandeo@telecom-bretagne.eu

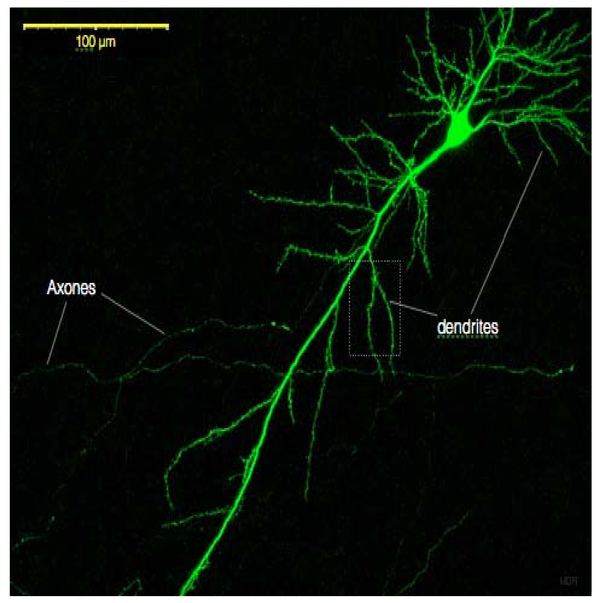
Site web http://tandeo.wordpress.com

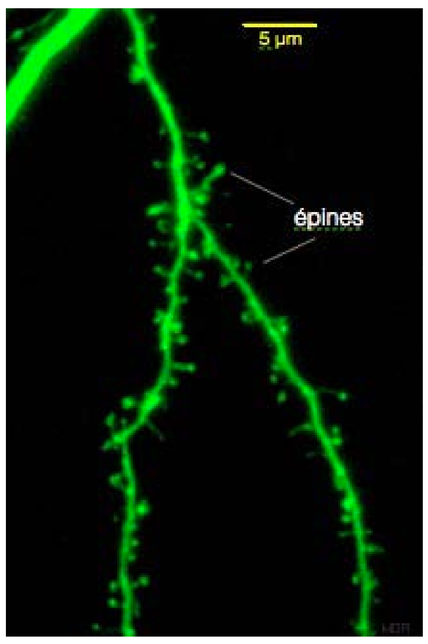
Nouveau modèle de neurone formel biologiquement plausible

DE NOUVEAUX RÔLES POUR UNE STRUCTURE ANCIENNEMENT CONNUE

Les clusters d'épines discrètes agissent comme des unités de calcul

- Les épines dendritiques ont été identifiées anatomiquement par Ramon Y. Cajal en 1911 qui les a qualifiées de "*espinas*" en raison de leur ressemblance avec des épines sur les tiges des fleurs.
- L'idée que les dendrites ne sont que des câbles passifs qui relaient les signaux entrants sur le corps de la cellule ne tient plus.
- Des études récentes révèlent que les sections dendritiques contenant des clusters d'épines discrètes agissent comme unités de calcul (Blom H, Rönnlund D, Scott L, Westin L, Widengren J, et al., *Nature*, 2013) .
- Cette clusterisation est influencée par les entrées sensorielles (Frost N. A., Shroff H., Kong H., Betzig E., Blanpied T. A., *Neuron*, Vol. 67, Issue 1, 15 July 2010).





Plasticité structurale et apprentissage

- Les épines dendritiques subissent en réalité d'une part des changements de leur forme et d'autre part un turn-over permanent (elles apparaissent et disparaissent)
- Lamprecht and Le Doux (2004) proposent une revue des mécanismes de plasticité structurelle associées aux épines dendritiques du neurone post-synaptique. Ces mécanismes sont généralement associées au renforcement et à la stabilisation de l'apprentissage synaptique.

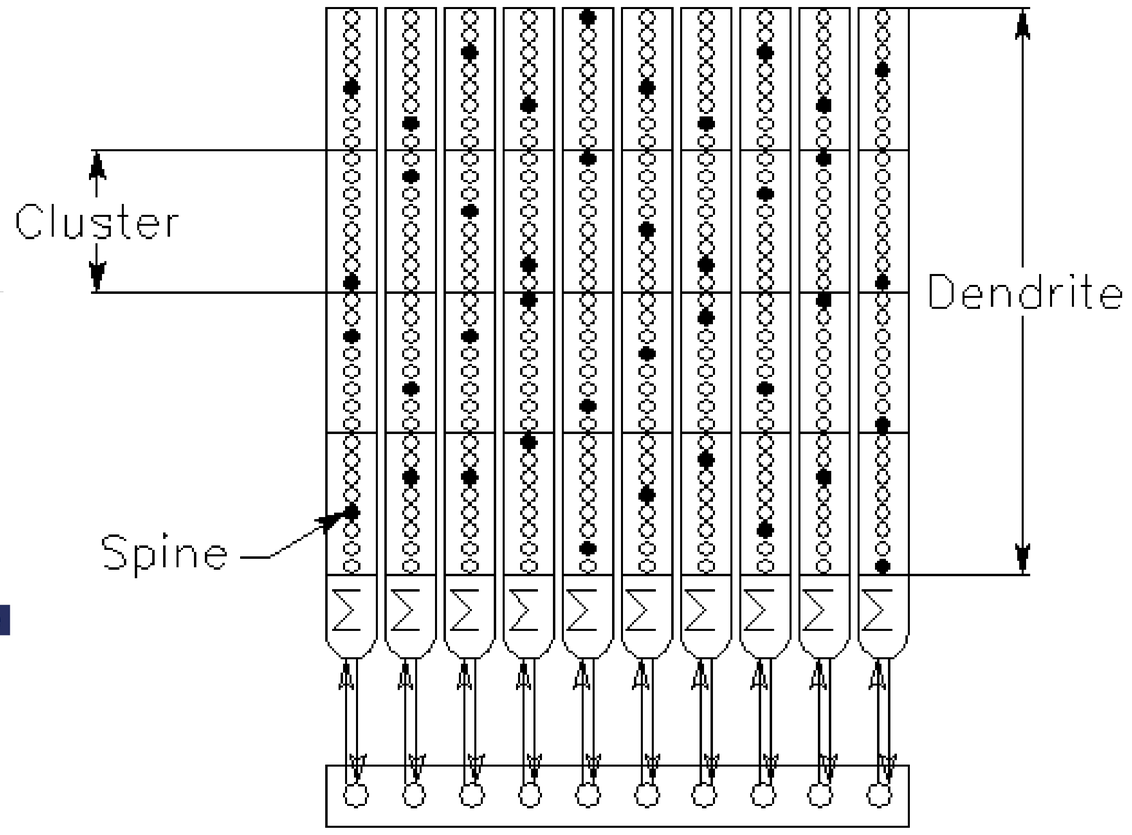
Auteurs

Ehsan Sedgh Gooya

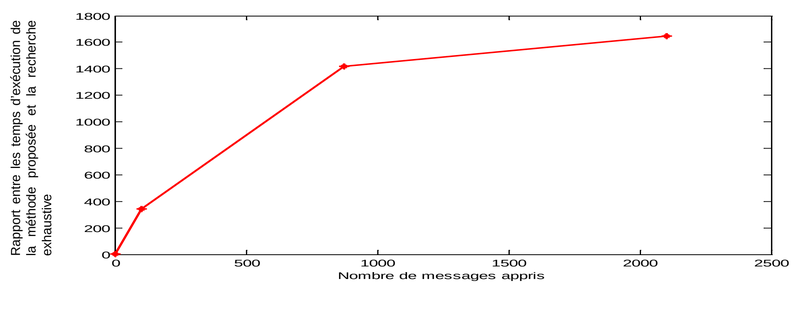
Dominique Pastor

Partenaires

MODELE FORMEL ET PROPRIETES



- Comme les réseaux à cliques de Berrou et Gripon, ce modèle n'emploie que des connexions binaires (pas de poids synaptiques)
- La taille de la mémoire (le stockage) n'est pas un réel problème. Par contre, il faut savoir retrouver très vite et sans erreur un message, ce que permet le neurone formel proposé
- La formalisation mathématique montre que ce neurone ne commet aucune erreur: il reconnaît tout ou une partie des messages qu'il a appris et ne reconnaît aucun message ou partie de message qu'il n'aurait pas appris
- Il est aussi performant qu'une recherche exhaustive, mais pour un coût de calcul largement inférieur.



- Les calculs de cette simulation sont faits en utilisant une programmation Matlab et des signaux générés aléatoirement
- Pour cette simulation, le dendrite est composé de 4 clusters et 8 épines par clusters. Chaque dendrite est associée de manière univoque à un message appris.
- Aucune itération de décodage
- On apprend en ajoutant du matériel, sans altérer ce qui a déjà été appris et en gardant les mêmes capacités de discrimination

PERSPECTIVES

- Introduction de la plasticité à l'aide d'un paramètre de tolérance afin de traiter le cas de messages distordus et/ou bruités
- Structure distribuée et/ou hiérarchique (réseau de neurones)
- Concept d'apprentissage attentif
- Vers des machines auto-apprenantes

26-27 mars 2014 Colloque IMT "Numérique, grande échalle et complexité"

Contact dominique.pastor@telecom-bretagne.eu
ehsan.sedghgooya@telecom-bretagne.eu

Site web www.perso.telecom-bretagne.eu/pastor/

Suivi d'objets multiples dans des scènes vidéo complexes

Parties prenantes

TELECOM Lille

Auteurs

Christelle Garnier
Mehdi Oulad Améziane
François Septier
Yves Delignon
Emmanuel Duflos

Partenaires

LAGIS
UMR CNRS 8219

CENTRALE LILLE

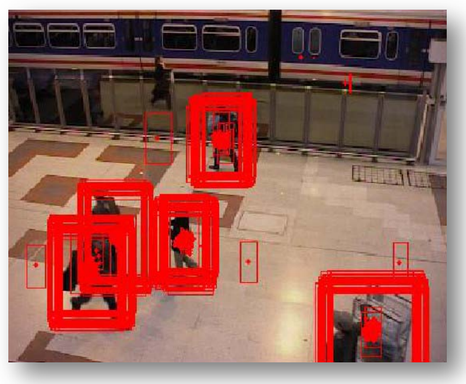
CONTEXTE

- Très grand volume de données vidéo à analyser dans de nombreux domaines d'application : vidéosurveillance, robotique, véhicules autonomes, interactions homme-machine, imagerie du vivant…
- Besoin de systèmes autonomes et intelligents, capables d'extraire automatiquement les informations utiles et de les interpréter.

OBJECTIF, PROBLEMATIQUE ET METHODES

Développer des modèles et des méthodes pour détecter et suivre de façon automatique, fiable et robuste des objets multiples dans des séquences vidéo issues d'une ou de plusieurs caméras.

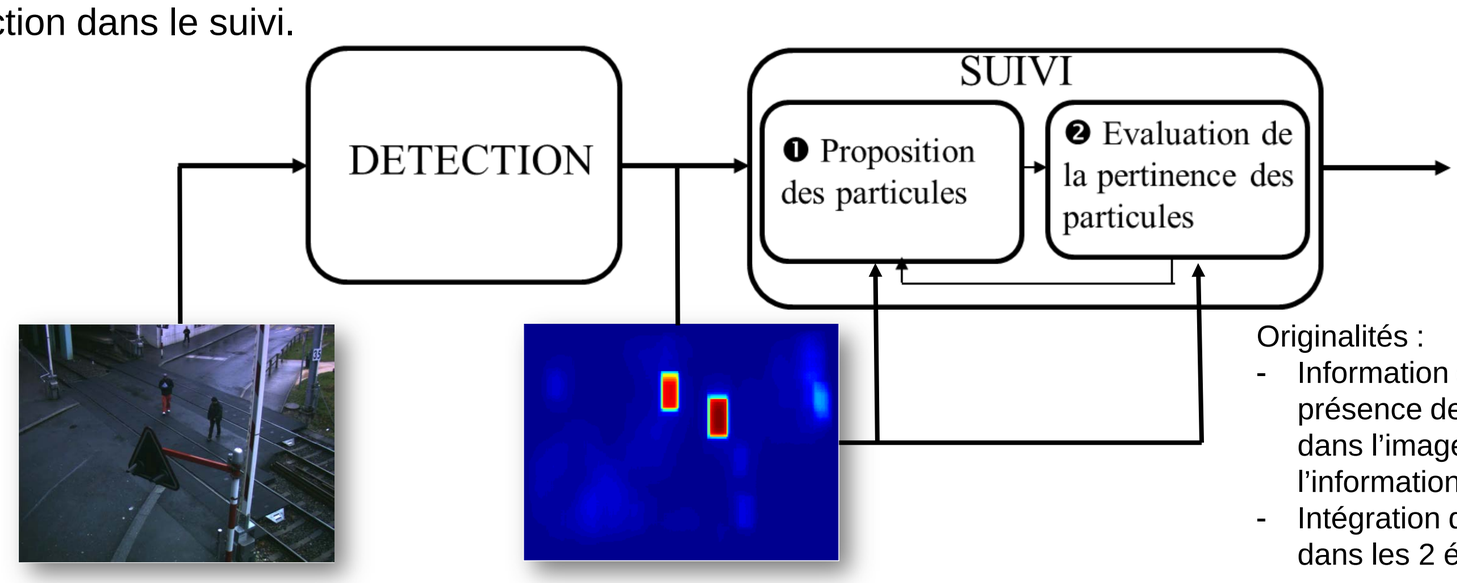
- **Principales difficultés** : nombre croissant d'objets (grande dimension), apparitions / disparitions d'objets, bruit, fausses détections, complexité de l'environnement, non-stationnarités (variations de l'environnement, du mouvement et de l'apparence des objets), occlusions…
- **Problème d'estimation séquentielle** : déterminer le nombre d'objets et leurs paramètres caractéristiques au cours du temps.
- **Outils méthodologiques** :
  Méthodes séquentielles de Monte-Carlo (filtrage particulaire),
  Méthodes de Monte-Carlo par chaînes de Markov (MCMC).



Exploration de l'espace d'état par les particules

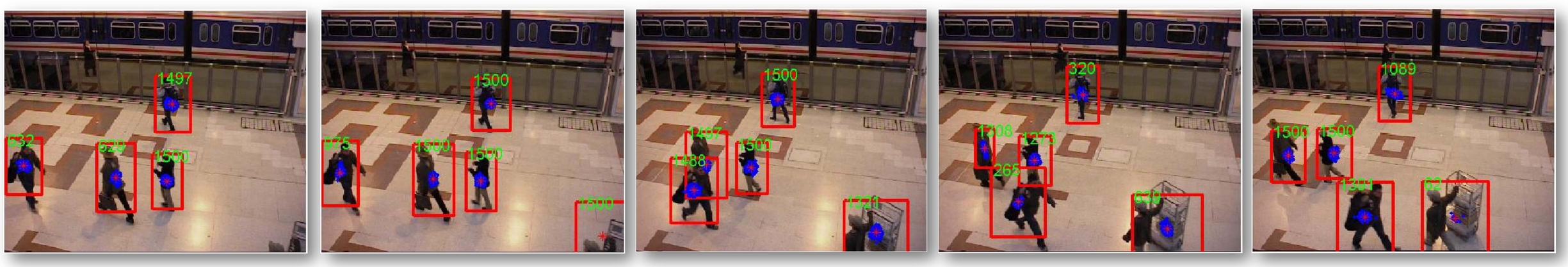
TRAVAUX ACTUELS

- **Optimisation des liens entre détection et suivi** : intégration de la sortie « soft » d'une méthode de détection dans le suivi.



Originalités :
- Information de probabilité sur la présence des objets d'intérêt dans l'image (plus riche que l'information binaire usuelle),
- Intégration de cette information dans les 2 étapes de suivi.

- **Gestion des variations de mouvement et d'apparence des objets** : nouveaux modèles dynamiques.
- **Amélioration du suivi en grande dimension** :
  - Versions séquentielles des méthodes MCMC plus performantes que les filtres particulaires lorsque la dimension augmente,
  - Lois de proposition plus efficaces pour explorer l'espace d'état et guider rapidement les algorithmes vers les zones à forte vraisemblance.



Suivi multi-objets : (1) suivi de 4 personnes, (2) détection automatique d'une 5ème personne, (3) 1ère occlusion partielle, (4) 2ème occlusion partielle, (5) suivi après les occlusions

Colloque Numérique Institut Mines Telecom
Mars 2014

Contact christelle.garnier@telecom-lille.fr
Site web www.splab.telecom-lille1.eu

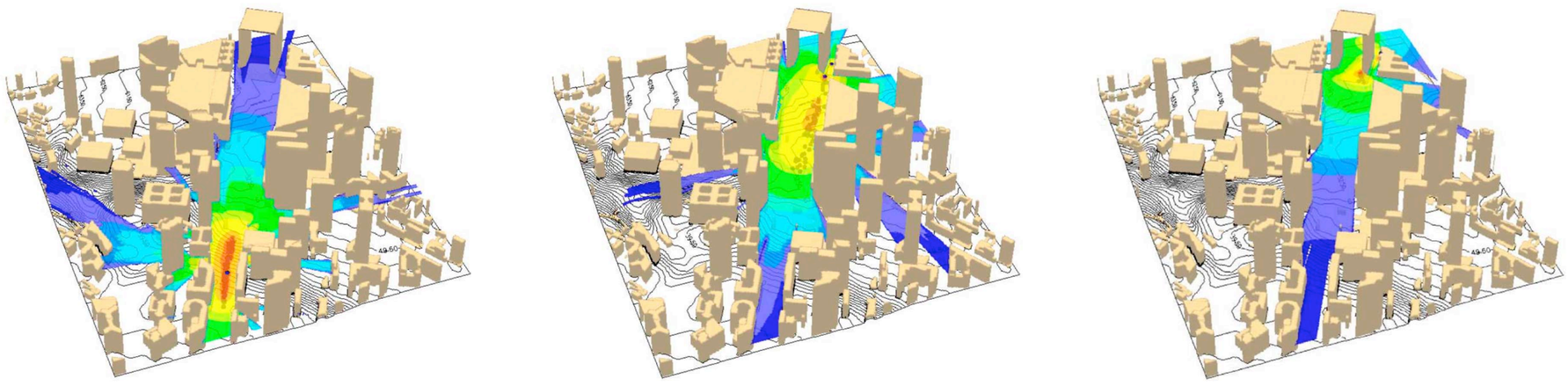
Localisation de sources de pollution depuis un réseau de capteurs

Contexte

- Menace grandissante de rejets délibérés ou accidentels d'agents nucléaires, radioactifs, biologiques ou chimiques (NRBC) ayant des conséquences dramatiques pour la population et l'environnement
- Mise en place d'un réseau mondial de capteurs dans le cadre du Traité d'Interdiction Complète des Essais Nucléaires (TICEN)

Objectifs

Développer une méthode de détection et de localisation de sources de rejets polluants depuis des mesures bruitées de concentration issues de multiples capteurs



Simulation d'un rejet d'agent toxique au coeur du quartier de la Défense à Paris

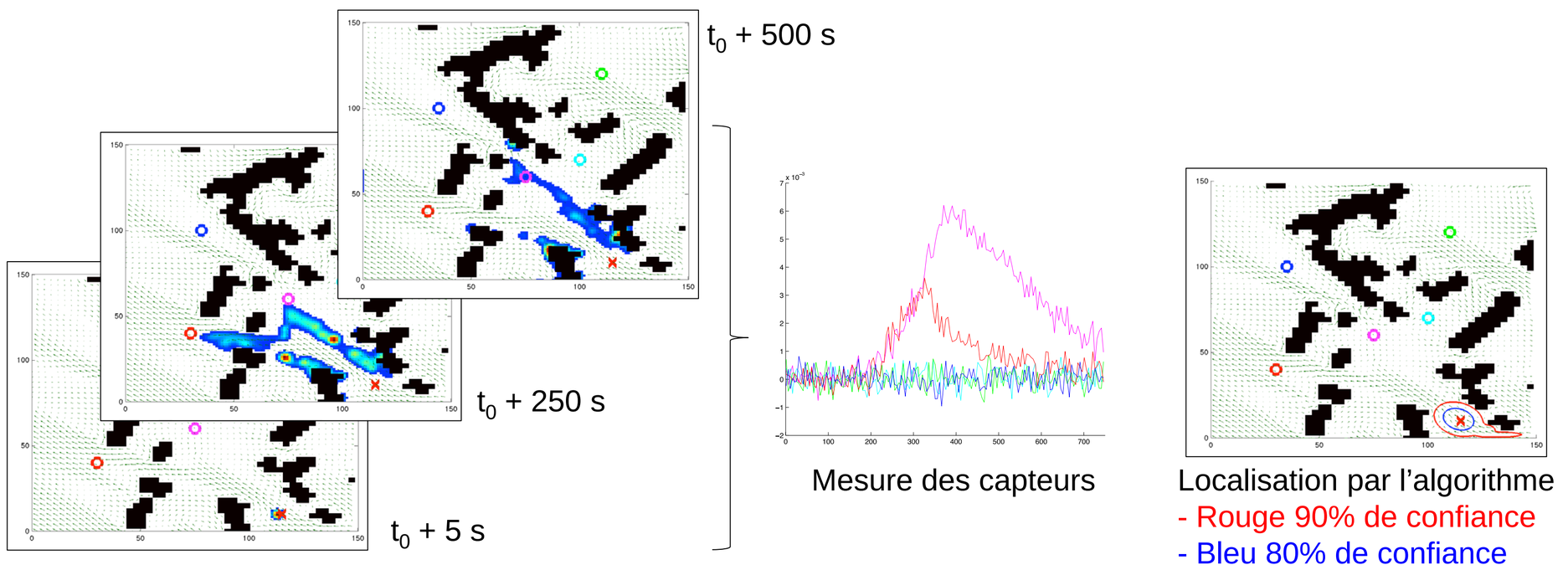
Problématiques

- Complexité des modèles météorologiques nécessaires pour la simulation réaliste de la dispersion atmosphérique d'agents toxiques.
- Imperfection et inhomogénéité des capteurs utilisés.
- Besoin d'une solution rapide et fiable afin de minimiser les conséquences d'un rejet.

Travaux actuels

Développement d'une méthode statistique de type Monte-Carlo permettant contrairement aux approches existantes de:

- Fournir un intervalle de confiance sur l'estimation fournie à l'utilisateur.
- Exploiter un modèle de dispersion atmosphérique de flux de turbulences complexes par des modèles Lagrangien (Parallel Micro-SWIFT-SPRAY, PMSS)
- Converger plus rapidement vers les zones de rejet les plus probables grâce à l'utilisation de techniques adaptatives (algorithmes PMC, AMIS).



Scénario d'un rejet à t_0 d'une source localisée à (115,10) dans un environnement avec bâtiments (noir) et 5 capteurs (cercles)

Auteurs

Harizo Rajoana (Aria Tech.)
François Septier
Yves Delignon
Patrick Armand (CEA)

Partenaires

UMR CNRS 8219

Contact françois.septier@telecom-lille.fr

Site web www.telecom-lille.fr/people/septier

Colloque Numérique Institut Mines Telecom

Mars 2014

Riemannian Geometry for 3D Human Video Retrieval

Rim Slama*, Hazem Wannous*, Mohamed Daoudi+;*
+ Institut TELECOM/TELECOM Lille 1, France.
* Laboratoire d'Informatique Fondamentale de Lille (UMR Lille1/CNRS 8022), Université de Lille 1, France
{rim.slama, hazem.wannous, mohamed.daoudi}@telecom-lille.fr

Context and Issues

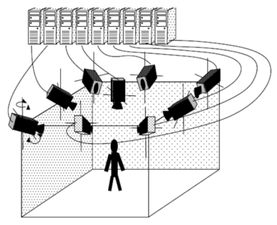
- Long sequence of 3D videos : massive amounts of data

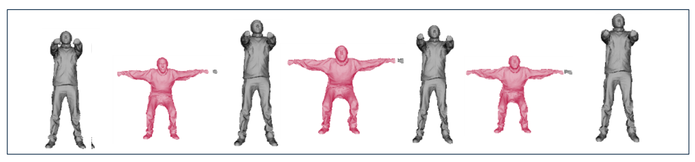
☠ browsing and searching for relevant information quickly become difficult

➔Need for 3D video segmentation system

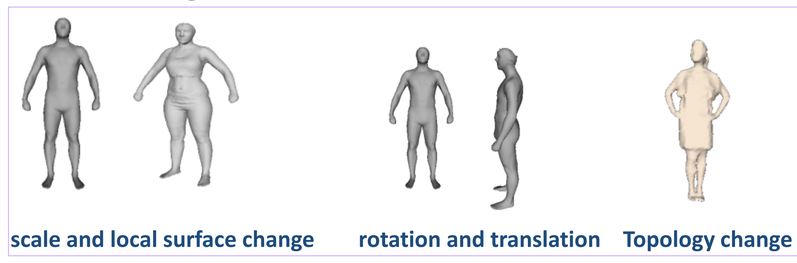
➔Pose/Motion retrieval

➔Video summarization



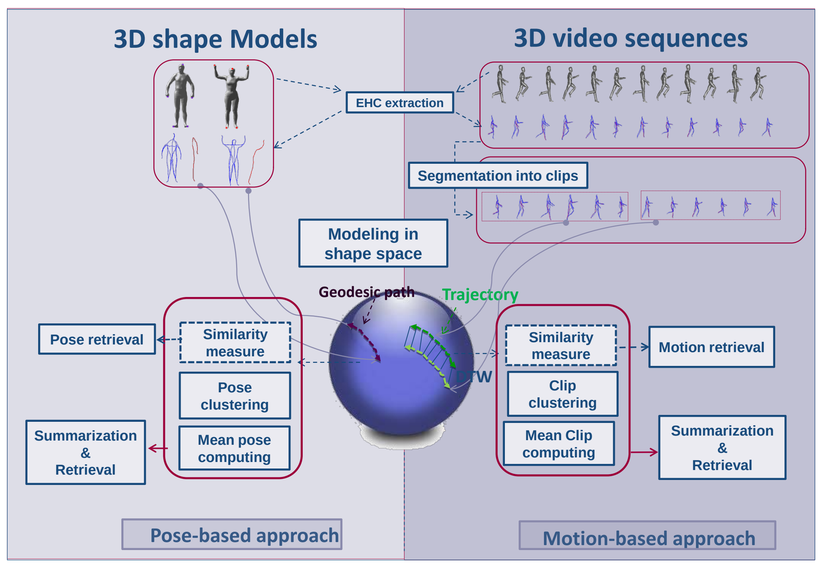


- Challenges:

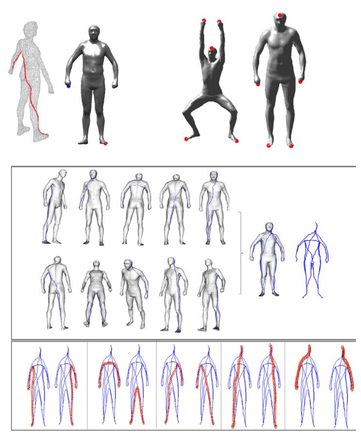


Approach: Riemannian Geometry

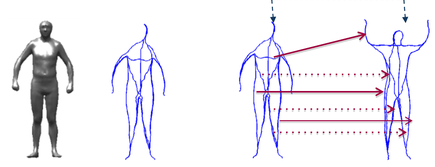
Overview



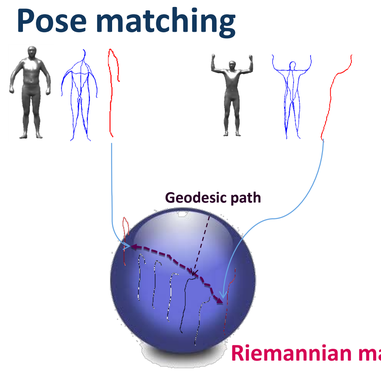
Extremal Human Curve extraction



EHC comparison



Pose matching

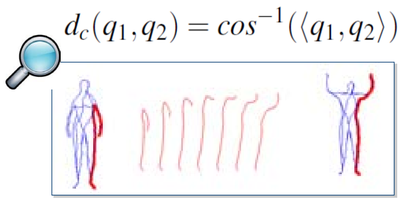


$$\beta \xrightarrow{\text{SRVF}} q : \mathbb{S}^1 \to \mathbb{R}^3, \quad q(t) \doteq \frac{\dot{\beta}(t)}{\sqrt{\|\dot{\beta}(t)\|}}$$

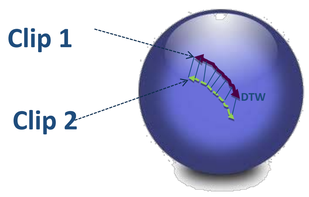
$$\mathcal{C} = \{q : I \to \mathbb{R}^3, \|q\| = 1\} \subset \mathbb{L}^2(I, \mathbb{R}^3)$$

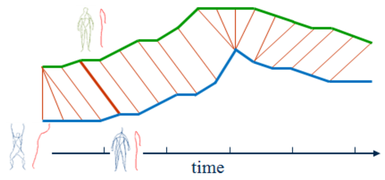
Geodesic distance between $q_1, q_2 \in \mathcal{C}$

$$d_c(q_1, q_2) = cos^{-1}(\langle q_1, q_2 \rangle)$$



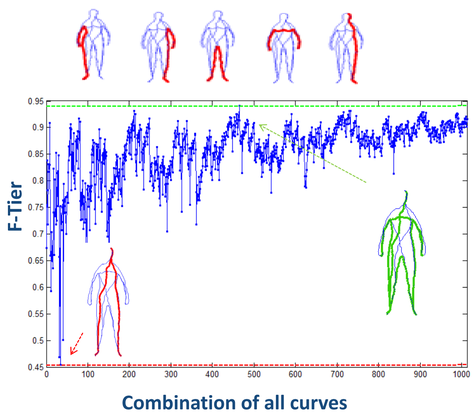
Clip matching



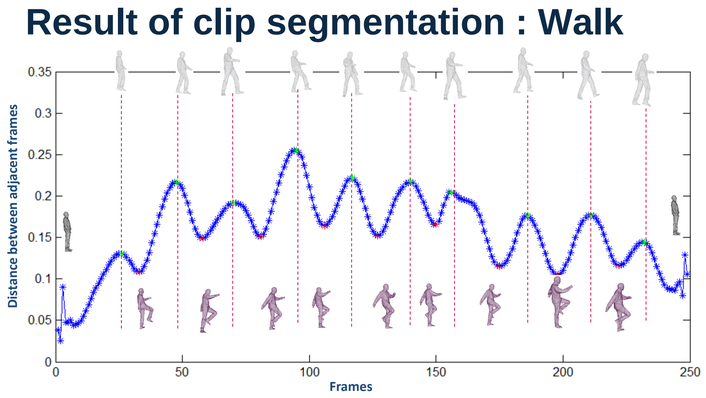


Results and Conclusion

Selected curves



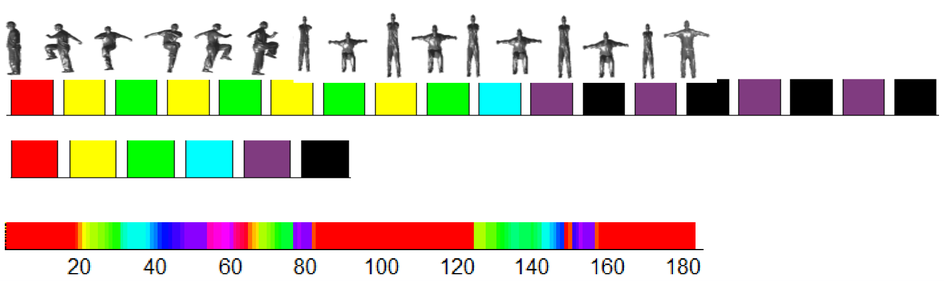
Result of clip segmentation : Walk



Conclusion

- Five curves are sufficient to represent at best the body pose
- Velocity curve is used to segment the long sequences into clips
- Clip matching using DTW on Riemannian manifold gives 93.44% of second tier rate and allows being invariant to speed
- Summarization by clustering is exploited in content-based motion retrieval

Videoclustering



Selected publications

- R. Slama, H. Wannous, M. Daoudi, 3D human motion analysis framework for shape similarity and retrieval. Image Vision Computing Journal 32(2): 131-154 (2014)

Office D106N, Telecom Lille1 -
Rue G. Marconi, Cité Scientifique
Villeneuve d'Ascq, 59650, FRANCE.
rim.slama@telecom-lille1.eu
(+33) (0)3 20 33 55 17
(+33) (0)3 20 33 55 99

Rim Slama, PhD. student,

Statistical Shape Analysis of a Large 3D Faces Dataset

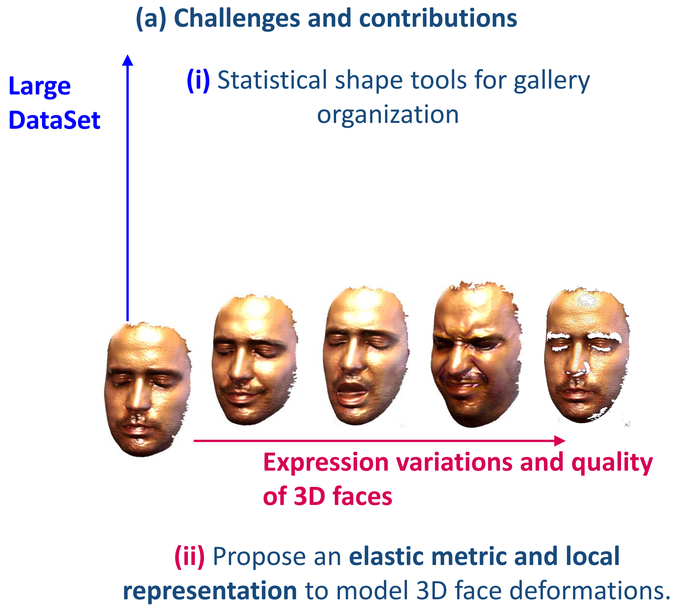
Hassen Drira*, Boulbaba Ben Amor*, Mohamed Daoudi* and Anuj Srivastava[T]

* Institut Mines-Télécom/Télécom Lille, France.
[T] Departement of Statistics, Florida State University, Tallahassee, FL 32306, USA.

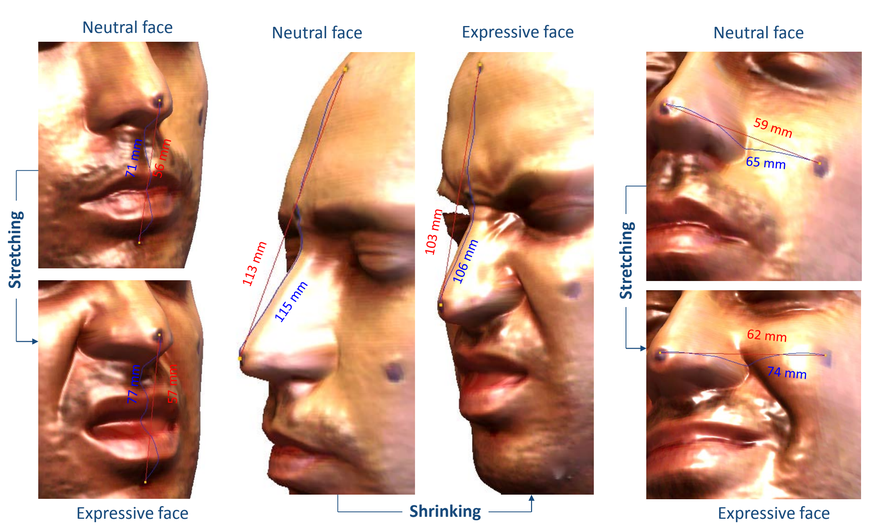
hassen.drira@telecom-lille.fr

MOTIVATIONS

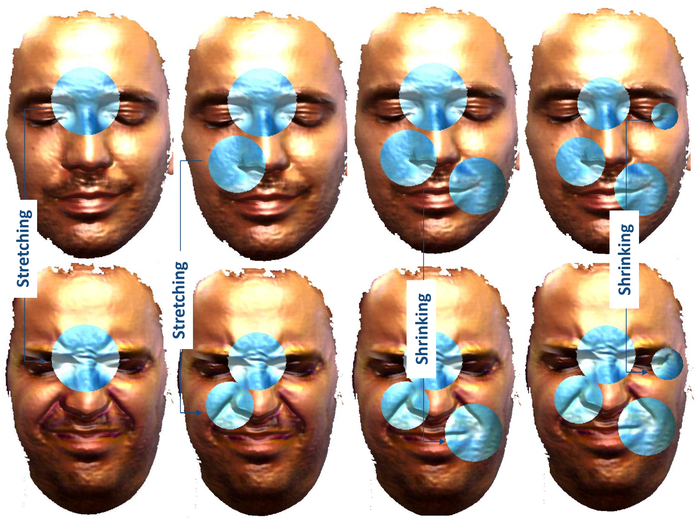
(a) Challenges and contributions



(b) Need for elastic representation of facial surfaces

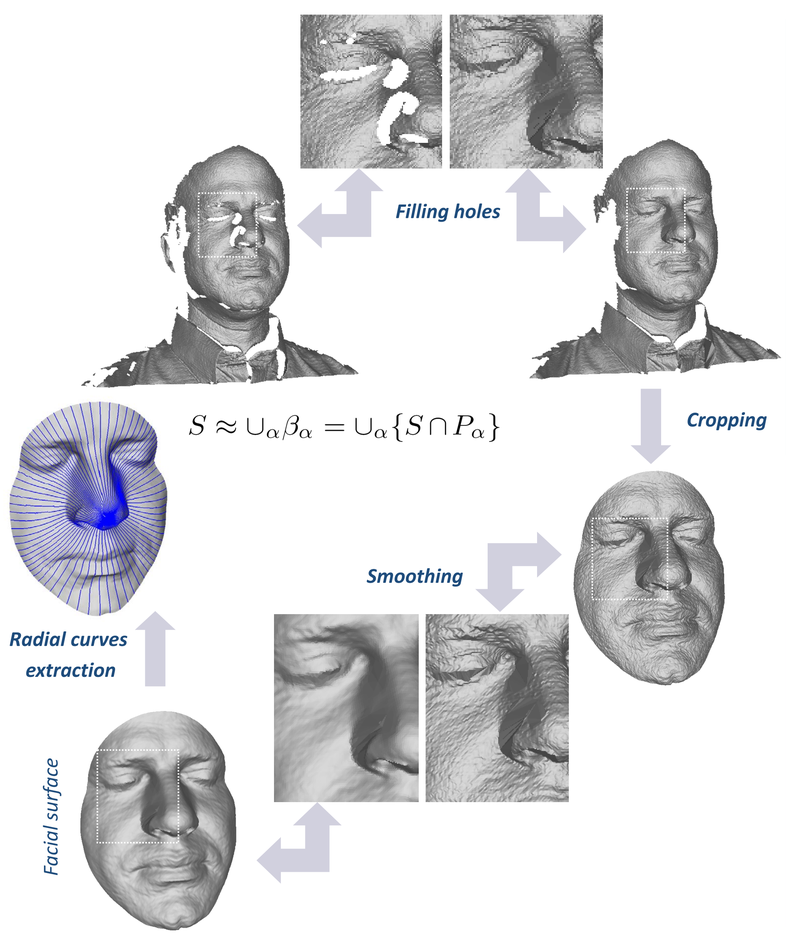


(c) Facial patches Stretching/Shrinking (elastic deformations)

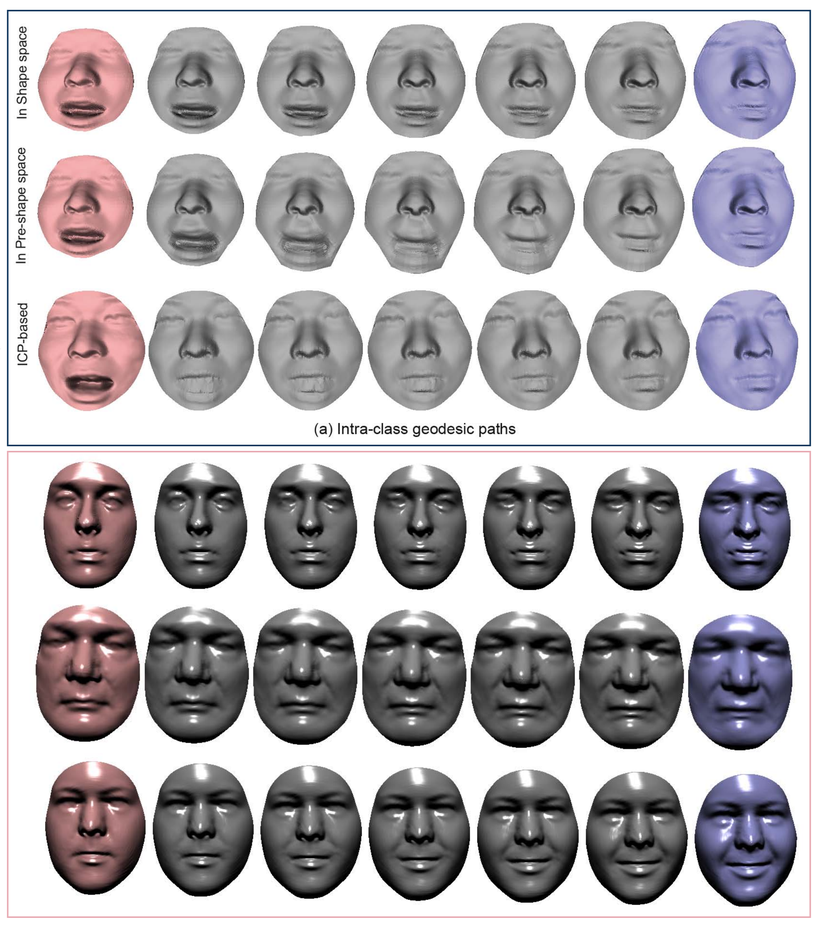


STATISTICAL SHAPE ANALYSIS of 3D FACE

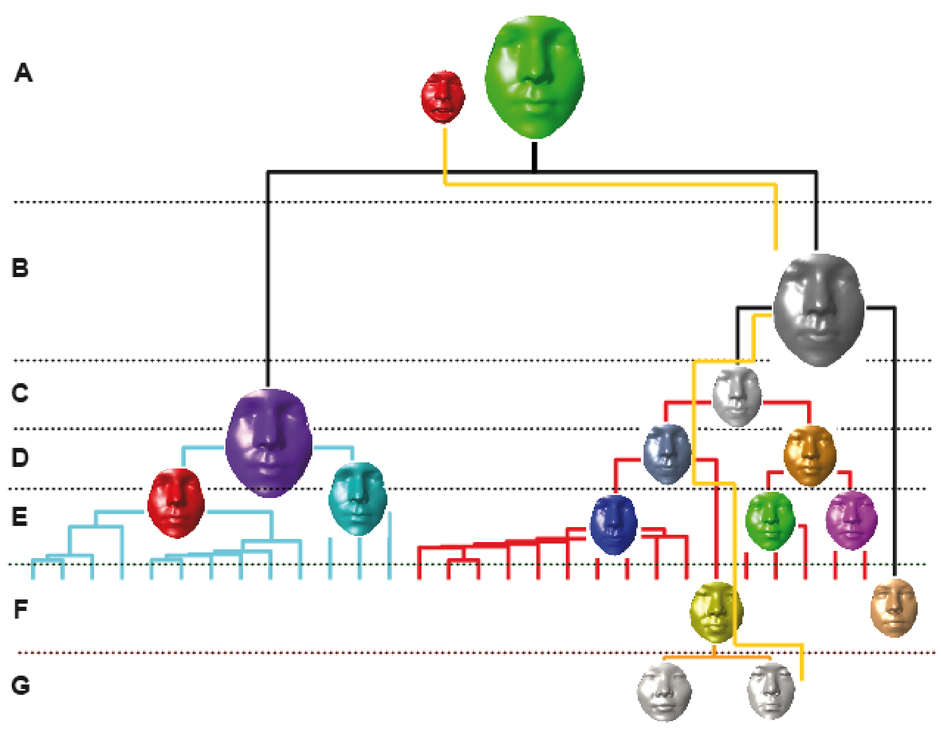
(a) Preprocessing of 3D scans and facial representation



(b) Examples of geodesics



(c) Gallery organization: Hierarchical retrieval of 3D faces Gallery



EXPERIMENTAL RESULTS

(a) Experimental protocol

- FRGCv2 datset: 4007 facial scans
- 466 subjects
- Gallery: neutral faces of 466 subjects
- Probe: remaining

(b) Experimental results

Comparison of rank-1 scores on the FRGCv2 dataset with the state-of-the-art results.

Spreeuwers [29]	Wang et al. [32]	Haar et al. [31]	Berretti et al. [2]	Queirolo et al. [26]	Faltemier et al. [9]	Kakadiaris et al. [13]	Our approach
99%	98.3%	97%	94.1%	98.4%	97.2%	97%	97%

Comparison of verification rates at FAR=0.1% on the FRGCv2 dataset with state-of-the-art results (the ROC III mask and the All vs. All scenario).

Approaches	Kakadiaris et al. [13]	Faltemier et al. [9]	Berretti et al. [2]	Queirolo et al. [26]	Spreeuwers [29]	Wang et al. [32]	Our approach
ROC III	97%	94.8%	-	96.6%	94.6%	98.4%	97.14%
All vs. All	-	93.2%	81.2%	96.5%	94.6%	98.13%	93.96%

Selected publications

1. Hassen Drira, Boulbaba Ben Amor, Anuj Srivastava, Mohamed Daoudi, Rim Slama: 3D Face Recognition under Expressions, Occlusions, and Pose Variations. IEEE Trans. Pattern Anal. Mach. Intell. 35(9): 2270-2283 (2013).

2. Hassen Drira, Boulbaba Ben Amor, Anuj Srivastava, Mohamed Daoudi: A Riemannian analysis of 3D nose shapes for partial human biometrics. ICCV 2009: 2050-2057

SUPPORTED BY ANR

Laboratoire d'Informatique Fondamentale de Lille
CNRS/Lille1 UMR 8022

Improved bandit algorithms: Go Bayesian!

Emilie Kaufmann, Telecom ParisTech & CNRS

In a nutshell

What is the performance of Bayesian bandit algorithms from a frequentist point of view? Bayes-UCB and Thompson Sampling appear to outperform frequentist algorithms on their own ground, which is supported by optimal regret bound for the Bernoulli case.

Bayesian vs. Frequentist Model for MAB

K independent arms. Arm a depends on parameter θ_a and has expectation μ_a; optimal arm is $a^* = \text{argmax } \mu_a$ and $\mu^* = \mu_{a^*}$ is the highest expectation of reward associated.

Two probabilistic modelings

Frequentist :	**Bayesian :**
• $\theta_1, \dots, \theta_K$ unknown parameters	• $\theta_a \overset{i.i.d.}{\sim} \pi_a$
• $(Y_{a,t})_t$ is i.i.d. with distribution ν_{θ_a}	• $(Y_{a,t})_t$ is i.i.d. conditionally to θ_a with distribution ν_{θ_a}

At time t, arm A_t is chosen and reward $X_t = Y_{A_t,t}$ is observed

Two measures of performance

• Minimize (classic) regret $R_n(\theta) = \mathbb{E}_\theta \left[\sum_{t=1}^{n} \theta^* - \theta_{A_t} \right]$	• Minimize "Bayesian" regret $R_n = \int R_n(\theta) d\pi(\theta)$

Optimal algorithms

• *Asymptotically optimal algorithms* satisfy, for $a : \mu_a < \mu^*$, $\limsup_{n\to\infty} \frac{\mathbb{E}_\theta[N_a(n)]}{\log(n)} \leq \frac{1}{\text{KL}(\nu_{\theta_a}, \nu_{\theta^*})}$ They are optimal in the sense of Lai and Robbins' lower bound (1985) on the number of draw of a sub-optimal arm	• An index policy inspired by that of Gittins (1979) adapted to non-discounted rewards minimizes Bayesian regret → no frequentist guarantees

⇒ Our goal: Design algorithms inspired by the Bayesian modeling that are asymptotically optimal in the frequentist setting.

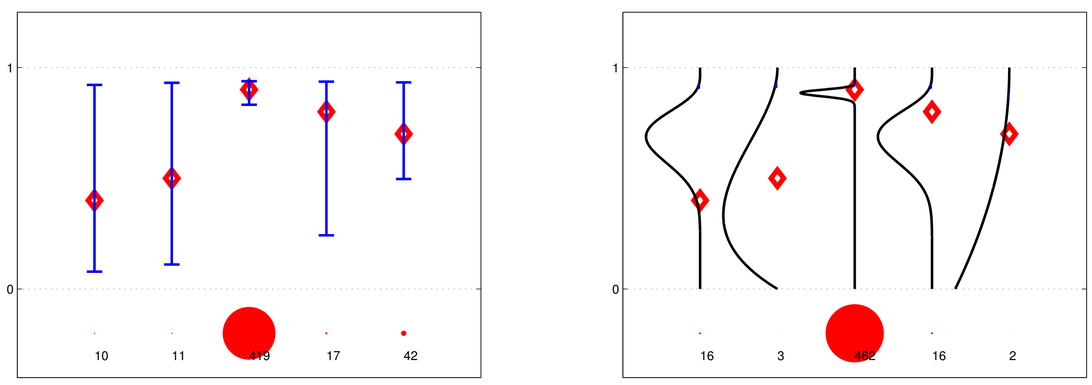
Bayesian versus frequentist algorithms

Some quantities that naturally arise in the Bayesian modeling are

- $\Pi_t = (\pi_1^t, \dots, \pi_K^t)$ the current posterior over $(\theta_1, ..., \theta_K)$
- $\Lambda_t = (\lambda_1^t, \dots, \lambda_K^t)$ the current posterior over the means $(\mu_1, ..., \mu_K)$

Successful algorithms inspired by the frequentist modeling use

- Upper Confidence Bound for the empirical mean... (UCB)
- ... built using KL-divergence (KL-UCB, asymptotically optimal)



Whereas a *Bayesian algorithm* uses Π_t to determine action A_t.

Bayes-UCB and Thompson Sampling

Bayes-UCB algorithm chooses $A_t = \underset{a=1...K}{\text{argmax}}\ q_a(t)$, with

$$q_a(t) = Q\left(1 - \frac{1}{t(\log t)^c}, \lambda_a^t\right)$$

Thompson Sampling is a *randomized* algorithm:

$$\begin{cases} \forall a \in \{1..K\}, \quad \theta_a(t) \sim \lambda_a^t \\ A_t = \text{argmax}_a\ \theta_a(t) \end{cases}$$

Parameters: c (in practice, take $c = 0$), initial prior Π_0

Bayes-UCB: Theoretical elements

ν_{θ_a} is the Bernoulli distribution $\mathcal{B}(\theta_a)$, π_a^0 the (conjugate) prior $\mathcal{U}([0,1])$

- **Bayes-UCB is asymptotically optimal for Bernoulli bandits**

Theorem 1 *Let $\epsilon > 0$; for the Bayes-UCB algorithm with parameter $c \geq 5$, the number of draws of a sub-optimal arm a is such that :*

$$\mathbb{E}_\theta[N_a(n)] \leq \frac{1+\epsilon}{KL\left(\mathcal{B}(\theta_a), \mathcal{B}(\theta^*)\right)} \log(n) + o_{\epsilon,c}\left(\log(n)\right)$$

- **Bayes-UCB is very close to a frequentist algorithm**

The Bayes-UCB index $q_a(t)$ is closely related to the one used by the KL-UCB algorithm (Cappé et al. 2013): $\tilde{u}_j(t) \leq q_j(t) \leq u_j(t)$ with:

$$u_a(t) = \underset{x > \frac{S_a(t)}{N_a(t)}}{\text{argmax}} \left\{ d\left(\frac{S_a(t)}{N_a(t)}, x\right) \leq \frac{\log(t) + c\log(\log(n))}{N_a(t)} \right\}$$

$$\tilde{u}_a(t) = \underset{x > \frac{S_a(t)}{N_a(t)+1}}{\text{argmax}} \left\{ d\left(\frac{S_a(t)}{N_a(t)+1}, x\right) \leq \frac{\log\left(\frac{t}{N_a(t)+2}\right) + c\log(\log(n))}{(N_a(t)+1)} \right\}$$

where $d(x,y) = KL\left(\mathcal{B}(x), \mathcal{B}(y)\right) = x\log\frac{x}{y} + (1-x)\log\frac{1-x}{1-y}$

Bayes-UCB appears to build automatically confidence intervals based on Kullback-Leibler divergence, that are adapted to the geometry of the problem in this specific case.

Thompson Sampling: Theoretical elements

- **TS is asymptotically optimal for Bernoulli bandits**

Theorem 2 *Let $\epsilon > 0$. With b defined below, for every sub-optimal arm a, there exists a constant $N(b, \epsilon, \theta_a, \theta^*)$ such that*

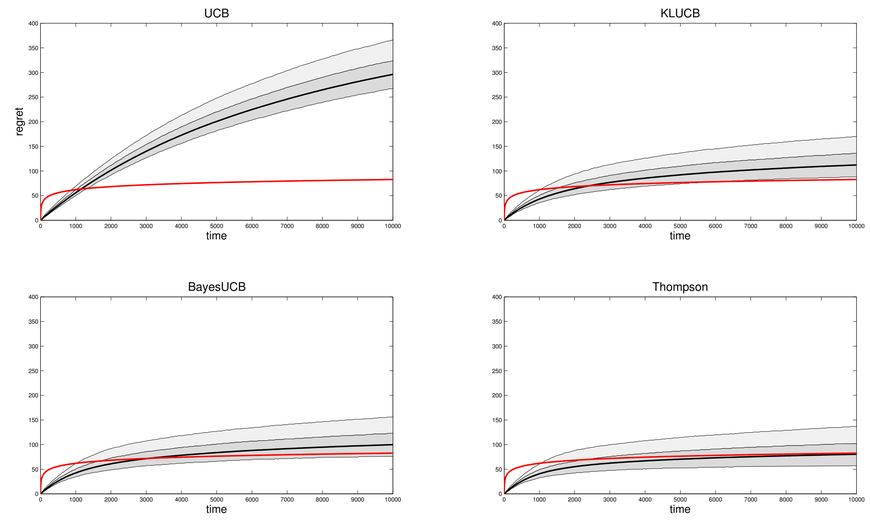
$$\mathbb{E}[N_a(n)] \leq (1+\epsilon)\frac{\ln(n) + \ln\ln(n)}{d(\theta_a, \theta^*)} + N(b, \epsilon, \theta_a, \theta^*) + 5 + 2C_b.$$

Proof Bottleneck: For some constants $b = b(\mu) \in (0,1)$ and $C_b < \infty$,

$$\sum_{t=1}^{\infty} \mathbb{P}\left(N_1(t) \leq t^b\right) \leq C_b.$$

Numerical experiments and beyond

- **Bayesian algorithms are practically as efficient as optimal frequentist algorithms** or even better!



Regret of the various algorithms as a function of time. The red curve show the lower bound, the black bold curve the mean regret and the dark and light shaded the central 99% and the upper 0.05%

- **They are easier to implement:** KL-UCB solves an optimization problem whereas Thompson Sampling only produces one sample!
- **They are easy to generalize:** general models where sampling from a posterior distribution is possible (using MCMC), sparse linear bandit, contextual bandit model...

References

[1] O. Cappé, A. Garivier, O. Maillard, R. Munos, G. Stoltz Kullback-leibler Upper Confidence Bounds for optimal sequential allocation *To appear in Annals of Statistics*, 2013

[2] E. Kaufmann, O. Cappé, A. Garivier, Bayesian Upper Confidence Bounds for bandit problems *AISTATS*, 2012

[3] E. Kaufmann, N. Korda, R. Munos Thompson Sampling: an asymptotically optimal finite-time analysis *ALT*, 2012

Numerical cost for time series prediction via aggregation

Andrés Sánchez Pérez
andres.sanchez-perez@telecom-paristech.fr

François Roueff
francois.roueff@telecom-paristech.fr

1. Summary

The aggregation technique provides an estimator with well-established and excellent theoretical properties that applies for a wide family of times series which includes the AR(d). However the numerical computation of this estimator relies on a Markov chain Monte Carlo method whose performances should be evaluated.

2. Stable AR(d) process

The $\boldsymbol{AR}(d)$ is the stationary solution of :

$$X_t \quad = \quad \sum_{j=1}^{d} \theta_j X_{t-j} + \sigma \xi_t \ , \tag{1}$$

where the innovations ξ_t are i.i.d. with $\mathbb{E}\xi_t = 0$.

Denote by $s_d(1) = \left\{ (\theta_1, \ldots, \theta_d) \ : \ \boldsymbol{\theta}(z) = 1 - \sum_k \theta_k z^k \neq 0 \text{ for } |z| < 1 \right\}$.

We assume that the parameter $\bar{\boldsymbol{\theta}} = (\theta_1, \ldots, \theta_d) \in s_d(1)$ and that (ξ_t) have compact support $\Rightarrow \exists \mathcal{B} \in \mathbb{R}_+^* : X_t \in [-\mathcal{B}, \mathcal{B}] \, \forall t$.

3. Forecasters

Let $(X_1, \ldots, X_n)$ observed values from this ***stationary time series*** $X = (X_t)_{t \in \mathbb{Z}}$. Consider a family of predictors $\{f_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta\}$. For any $\boldsymbol{\theta} \in \Theta$, $f_{\boldsymbol{\theta}}$ is a function from which we obtain :

$$\hat{X}_t^{\boldsymbol{\theta}} \quad = \quad f_{\boldsymbol{\theta}}\left(X_{t-1}, \ldots, X_{t-d}\right) \ , \tag{2}$$

a possible forecasting of X_t according to $\boldsymbol{\theta}$.
Let ℓ be a loss function; we define the prediction risk as
$R(\boldsymbol{\theta}) \quad = \quad \mathbb{E}\left[\ell\left(\hat{X}_t^{\boldsymbol{\theta}}, X_t\right)\right]$ and the empirical version of the risk as
$r_n\left(\boldsymbol{\theta}; X_1, \ldots, X_n\right) = \dfrac{1}{n-d} \sum_{t=d+1}^{n} \ell\left(\hat{X}_t^{\boldsymbol{\theta}}, X_t\right)$.

4. Gibbs estimator

For a $\lambda > 0$ (temperature parameter), we define the ***Gibbs estimator*** as the expectation of a r.v. drawn under the Gibbs measure $\pi\{-\lambda r_n\}$:

$$\hat{\boldsymbol{\theta}}_{\lambda,n} \quad = \quad \pi\{-\lambda r_n\}[\mathrm{Id}] = \int_{\Theta} \boldsymbol{\theta} \pi\{-\lambda r_n(\cdot)\}(\mathrm{d}\boldsymbol{\theta}) \ , \tag{3}$$

where $\nu[h] = \int h \mathrm{d}\nu$ and $\nu\{h\}(\mathrm{d}\boldsymbol{\theta}) = \dfrac{\exp(h(\boldsymbol{\theta}))}{\nu[\exp(h)]} \nu(\mathrm{d}\boldsymbol{\theta})$.

5. A theoretical result

Theorem 1 *In the context of the AR(d), for a bounded $\Theta \subset \mathbb{R}^p$, a uniform* ***prior*** *π yields that $\exists$ a constant $\mathcal{E} : \forall \epsilon > 0$, with probability at least $1 - \epsilon$,*

$$R\left(\hat{\boldsymbol{\theta}}_{\sqrt{n},n}\right) \quad \leq \quad \inf_{\boldsymbol{\theta} \in \Theta} R(\boldsymbol{\theta}) + \mathcal{E} \frac{\log^2(n)}{\sqrt{n}} + \frac{2}{\sqrt{n}} \log\left(\frac{1}{\epsilon}\right) \ . \tag{4}$$

6. Numerical approximation [2]

The ***Metropolis-Hastings algorithm*** generates a Markov chain $\boldsymbol{\Phi} = \{\boldsymbol{\Phi}_i\}_{i \geq 0}$ with the target distribution ρ as a unique invariant measure, based on another Markov chain which serves as a proposal. We tested :

- The *independent Hastings algorithm* where the proposal is i.i.d. with density q such that $\dfrac{q(y)}{\rho(y)} \geq \beta, \forall y \in \Theta$ for some $\beta > 0$.
- The *Metropolis-Hastings algorithm* where the proposal is a Markov chain with conditional density kernel q on $\bar{\Theta} \times \bar{\Theta}$ such that $\beta = \inf_{x \in \bar{\Theta}, y \in \bar{\Theta}} \dfrac{\rho(y)}{\rho(x)} \inf_{x \in \bar{\Theta}, y \in \bar{\Theta}} q(x, y) > 0$.

$\bar{\boldsymbol{\theta}}_m = \dfrac{1}{m} \sum_{i=0}^{m-1} \boldsymbol{\Phi}_i$ is a numerical estimate of $\int x \, \rho(x) \, \mathrm{d}x$.

Theorem 2 *Note by* $\mathrm{diam}(\Theta) = \sup_{x,y \in \Theta} \|x - y\|$ *and define :*

$$M(\alpha, \gamma, \epsilon) = \frac{(2-\gamma)\,\mathrm{diam}(\Theta)}{2\alpha^2 \epsilon \gamma} + \frac{1}{2} \sqrt{\left(\frac{(2-\gamma)\,\mathrm{diam}(\Theta)}{\alpha^2 \epsilon \gamma}\right)^2 + \frac{4\mathrm{diam}(\Theta)}{\alpha^2 \epsilon \gamma}} \ , \tag{5}$$

In the two previous cases, $\forall m \geq M(\alpha, \beta, \epsilon)$, with probability at least $1 - \epsilon$,

$$\left| \bar{\boldsymbol{\theta}}_m - \int x \, \rho(x) \, \mathrm{d}x \right| \quad \leq \quad \alpha \, . \tag{6}$$

7. Application to the AR(d) process

Since $s_d(1) \subseteq B_d\left(2^d - 1\right)$, the prior π can be defined on $\Theta = s_d(1)$ or $B_d\left(2^d - 1\right)$. These two possibilities are combined with two different proposals in the Metropolis-Hasting algorithm.

- **Uniform prior on** $B_d\left(2^d - 1\right)$
  - *Uniform proposal :* $\beta_{\lambda,n} = \exp\left(-\lambda \mathcal{B}^2 \left(1 + \sqrt{d}\left(2^d - 1\right)\right)^2\right)$.
  - *Constrained random walk with Gaussian increment :*

$$\beta_{\lambda,n} = \left(\frac{n}{2\pi}\right)^{\frac{d}{2}} \exp\left(-2\left(2^d - 1\right)\left(\lambda 2^{d+1} \mathcal{B}^2 + \left(2^d - 1\right) n\right)\right) \ .$$

- **Pushforward measure on** $s_d(1)$

  A map from the reciprocal roots of $\boldsymbol{\theta}(z)$ into the coefficients $\theta_1, \ldots \theta_d$ and a measure on the first ones allow to define a prior on $s_d(1)$.
  - *Uniform proposal :* $\beta_{\lambda,n} = \exp\left(-\lambda \mathcal{B}^2 \left(1 + \sqrt{d}\left(2^d - 1\right)\right)^2\right)$.
  - *Constrained random walk with Gaussian increment :*

$$\beta_{\lambda,n} = \left(\frac{n}{2\pi}\right)^{1+p+2\lfloor \frac{d}{2} \rfloor} \exp\left(-2\left(\lambda 2^{d+1}\left(2^d - 1\right)\mathcal{B}^2 + \left(2^{1+d+2\lfloor \frac{d}{2} \rfloor} - 1\right)^2 n\right)\right).$$

Theorem 3 *$\exists$ a constant $\mathcal{F}$ such that $\forall m \geq M\left(\dfrac{\log(n)}{n}, \beta_{\sqrt{n},n}, \epsilon\right)$, with M defined as in Theorem 2, with probability at least $(1 - \epsilon)^2$,*

$$R\left(\bar{\boldsymbol{\theta}}_{\sqrt{n},n,m}\right) \quad \leq \quad \inf_{\boldsymbol{\theta} \in \Theta} R(\boldsymbol{\theta}) + \mathcal{F} \frac{\log^2(n)}{\sqrt{n}} + \frac{2}{\sqrt{n}} \log\left(\frac{1}{\epsilon}\right) \ . \tag{7}$$

This result remains true for the whole family of Causal Bernoulli Shifts (CBS) processes (see [1] for the definition).

8. Numerical results

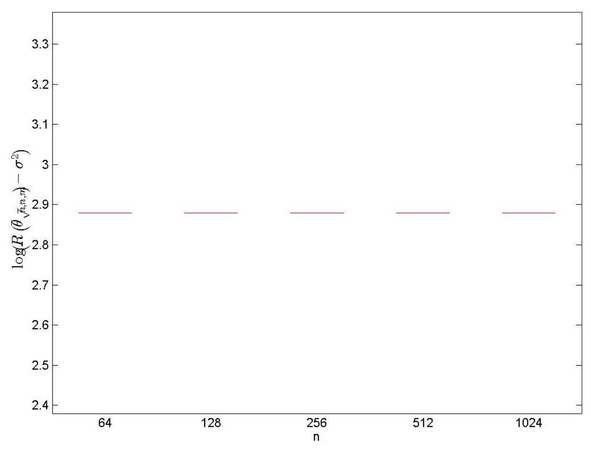


Figure 1: Uniform proposal, $d = 8$, $\Theta = B_8\left(2^8 - 1\right)$

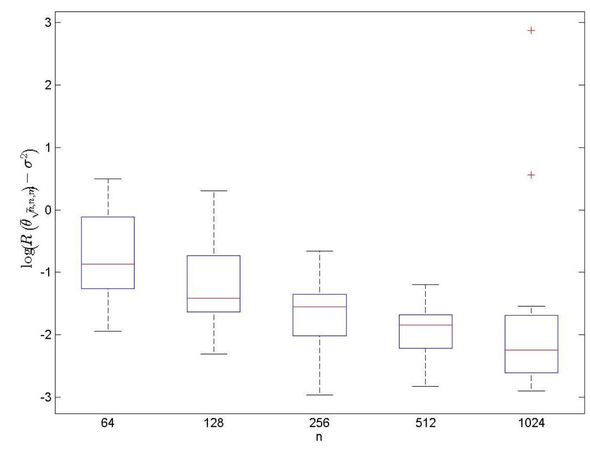


Figure 2: Gaussian proposal, $d = 8$, $\Theta = s_8(1)$

Figure 2 shows good results in contrast to Figure 1. However, using (5) and the obtained expressions of β yield to the following equivalence for the minimal number of iterations m guarantying a correct prediction error :

$$m \geq C_1(d) \frac{\log^2 n}{2n^2 \epsilon} \exp\left(C_2(d) \sqrt{n}\right) \ ,$$

where C_1 and C_2 are positive functions. Hence, to guaranty the Formula (7) we need a very high number of iterations, which can lead to a prohibitive computational cost.

9. References

References

[1] Pierre Alquier and Olivier Wintenberger. *Model selection for weakly dependent time series forecasting.* Bernoulli, 18(3) : 883-913, 2012.

[2] Krzysztof Łatuszyński and Wojciech Niemiro. *Rigorous confidence bounds for MCMC under a geometric drift condition.* J. Complexity, 27(1) : 23-38, 2011.

A NEW BROADCAST BASED DISTRIBUTED AVERAGING ALGORITHM OVER WIRELESS SENSOR NETWORKS

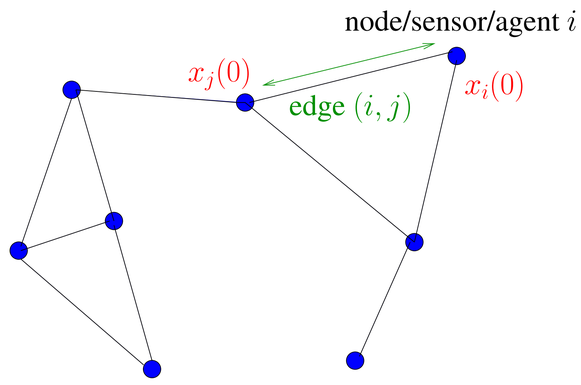
F. IUTZELER[(1)], P. CIBLAT [(1)] and W. HACHEM[(2)]

[(1)] Institut MINES-TELECOM ; TELECOM ParisTech ; CNRS LTCI | [(2)] CNRS LTCI ; TELECOM ParisTech

ABSTRACT

We study the distributed computation of the average value of initial measurements in a Wireless Sensor Network. Unlike existing works, we take benefit of the broadcast nature of the wireless channel to speed up the convergence speed

MODEL



- N nodes/sensors/agents
- $x_i(0)$ measurement at node i at time 0
- **Problem:** at each node, we want to compute

$$x_{\text{ave}} = \frac{1}{N}\sum_{i=1}^{N} x_i(0)$$

without fusion center and so with only local communications.

- **Applications:** practical measurements of temperature, gas pressure,
- At each time t, the sensors' updates can be rewritten in matrix form

$$\mathbf{x}(t+1) = \mathbf{K}(t)\mathbf{x}(t)$$

where $\mathbf{K}(t)$ is the update matrix for time t

Goal:

- We want $\mathbf{x}(t)$ to converge to the average consensus $x_{\text{ave}}\mathbf{1}$
- We want to wake up only one node at each time
- We want to use the broadcast nature of the wireless links

AVERAGING ALGORITHMS

Fundamental properties to ensure convergence to the average consensus:

	Property		Mathematical implication
1.	**Consensus preservation** $\mathbf{x}(t) = c\mathbf{1} \Rightarrow \mathbf{x}(t+1) = c\mathbf{1}$	$\Rightarrow$	**Row stochasticity** $\mathbf{K1} = \mathbf{1}$
2.	**Consensus value** $\mathbf{x}(t) = c\mathbf{1} \Rightarrow \mathbf{x}(t) = x_{\text{ave}}\mathbf{1}$	$\Rightarrow$	**Column stochasticity** $\mathbf{1}^T\mathbf{K} = \mathbf{1}^T$
3.	**Convergence** $\mathbb{E}\left[\|\mathbf{x}(t) - c\mathbf{1}\|^2\right] \to 0$	$\Rightarrow$	**Spectral radius** $\rho(\mathbb{E}[\mathbf{K}] - 1/N\mathbf{1}\mathbf{1}^T) < 1$

Consequences:

- Double-stochasticity needs feedback, so **no broadcast.**
- If only column-stochasticity, $\exists\ \mathbf{v}(t)$ non-negative vectors s.t.

$$\mathbf{K}(t)\cdots\mathbf{K}(2)\mathbf{K}(1) \sim \mathbf{v}(t)\mathbf{1}^{\top} \quad \text{and} \quad \mathbf{x}(t) \sim (Nx_{\text{ave}})\mathbf{v}(t)$$

Proposed solution:

- An other variable updated with the same matrix has to be considered to know $\mathbf{v}(t)$. Then $\mathbf{v}(t)$ can be removed by division.

$\Rightarrow$ **Variables do not converge to consensus, but the quotient does.**

PROPOSED ALGORITHM: *BWGossip*

Algorithm based on the *Sum-Weight* framework [Kempe2003, Bénézit2011] where sensors have **two** local variables jointly updated:

– a sum variable $\mathbf{s}(t)$
– a weight variable $\mathbf{w}(t)$

$$\begin{cases} \mathbf{s}(t+1) = \mathbf{K}(t)\mathbf{s}(t) \\ \mathbf{w}(t+1) = \mathbf{K}(t)\mathbf{w}(t) \end{cases} \quad \text{and} \quad \begin{matrix} \mathbf{s}(0) = \mathbf{x}(0) \\ \mathbf{w}(0) = \mathbf{1} \end{matrix}$$

BWGOSSIP

Assuming that at time t, the sensor i wakes up

▶ Sensor i broadcasts $\left(\frac{s_i(t)}{|\mathcal{N}_i|+1}; \frac{w_i(t)}{|\mathcal{N}_i|+1}\right)$

▶ At sensors in the neighborhood $\mathcal{N}_i$, we have:

$$\begin{cases} s_j(t+1) = s_j(t) + \frac{s_i(t)}{|\mathcal{N}_i|+1} \\ w_j(t+1) = w_j(t) + \frac{w_i(t)}{|\mathcal{N}_i|+1} \end{cases}, \forall j \in \mathcal{N}_i$$

▶ At sensor i, we have :

$$\begin{cases} s_i(t+1) = \frac{s_i(t)}{|\mathcal{N}_i|+1} \\ w_i(t+1) = \frac{w_i(t)}{|\mathcal{N}_i|+1} \end{cases}$$

▶ All other sensors stay idle.

RESULTS

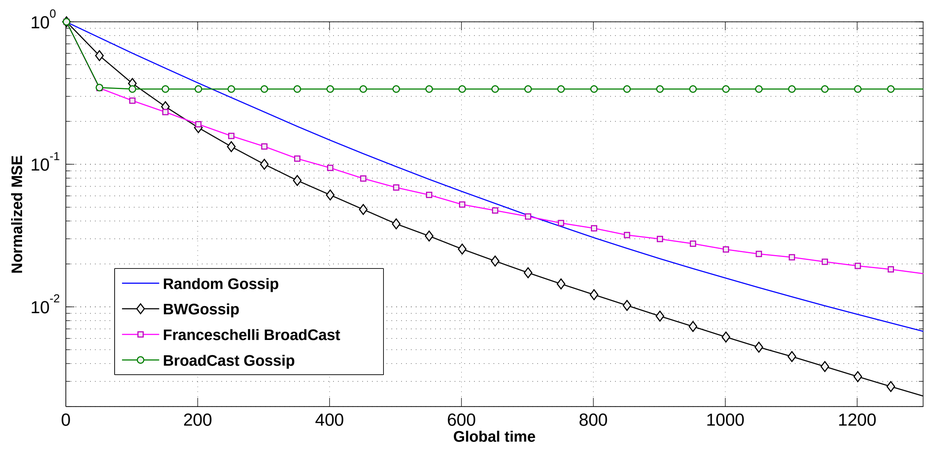
Theorem 1 CONVERGENCE

BWGossip *converges to the average consensus almost surely.*

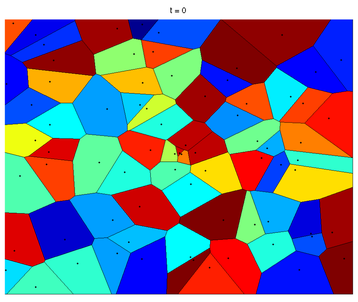
Theorem 2 CONVERGENCE SPEED **(main result)**

For the BWGossip *algorithm, we have:*

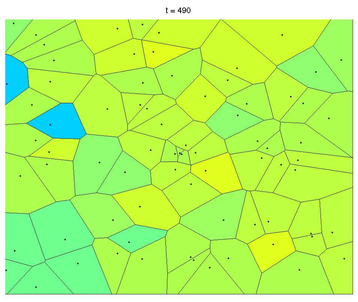
$$\forall \epsilon > 0, \quad \|\mathbf{x}(t) - x_{\text{ave}}\mathbf{1}\|_2^2 = \mathcal{O}_P\left((\Gamma+\epsilon)^t\right)$$
$$\textit{with } \Gamma = \rho\left((\mathbf{I}-\mathbf{J}) \otimes (\mathbf{I}-\mathbf{J}) . \mathbb{E}\left[\mathbf{K}\otimes\mathbf{K}\right]\right) < 1$$



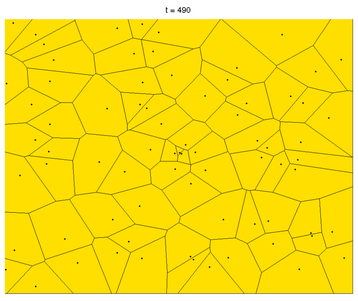
- ***BWGossip* outperforms the existing algorithms.**



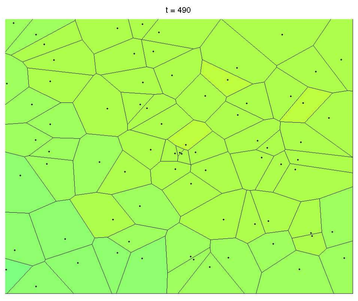
INITIALIZATION



RANDOM GOSSIP



BROADCAST GOSSIP

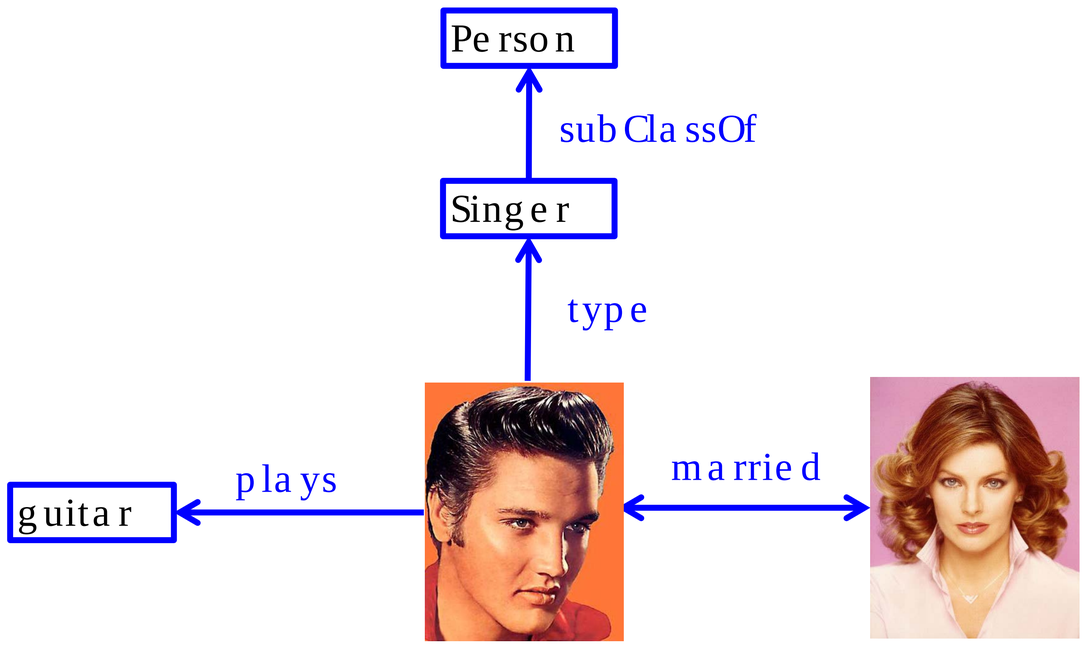


BW GOSSIP

The Independent Otto Hahn Research Group Ontologies

Fabian M. Suchanek

Ontology Construction



Ontologies serve, e.g., for disambiguation, translation, and question answering. We develop YAGO, the largest public ontology with a quality guarantee. YAGO is built automatically from Wikipedia and other sources.

http://yago-knowledge.org

YAGO – A Core of Semantic Knowledge
Fabian M. Suchanek, Gjergji Kasneci, Gerhard Weikum (WWW 2007)
+ follow-up publications in 2008, 2011, 2012, 2013

Rule Mining

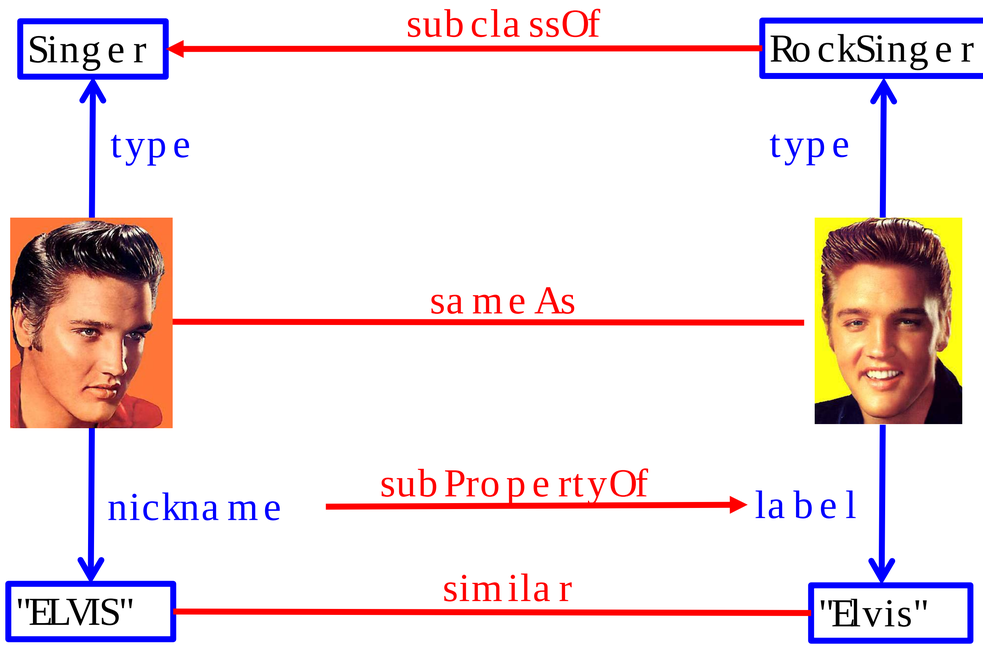
We develop techniques to mine rules, correlations, and schemas from an ontology. These serve to propose missing links, detect inconsistencies, reveal correlations, and make sense out of data. The semantics, incompleteness, and the scale of the data are different from classical settings.

Some rules we mined on YAGO are:

```
hasChild(x,y) & hasChild(z,y) => married(x,z)
wonAward(x,LeibnizPreis) => livesIn(x,Germany)
acadAdvisor(x,y) & almaMater(y,z) => worksAt(x,z)
```

AMIE: Association Rule Mining under Incomplete Evidence
Luis Galárraga, C. Teflioudi, K. Hose, F. M. Suchanek (WWW 2013)

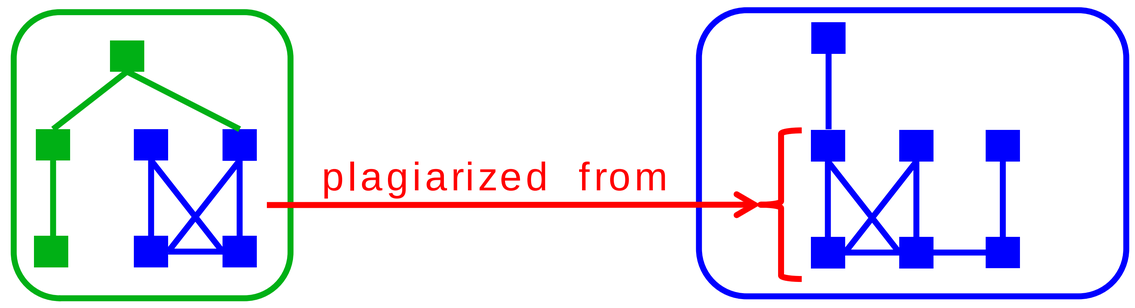
Ontology Matching



Many ontologies on the Semantic Web contain information about the same entities. To make use of complementary information, one has to determine which entities, classes, literals, and properties correspond. We develop statistical, logical, and probabilistic models and algorithms for this purpose.

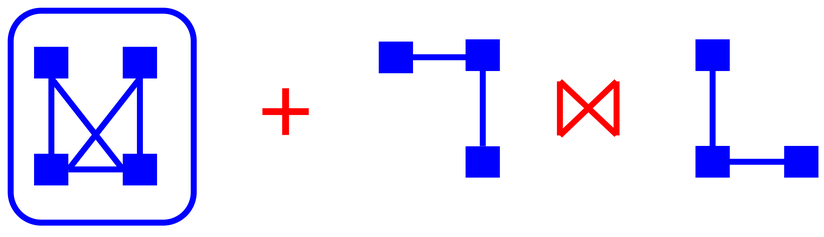
PARIS: Probabilistic Alignment of Relations, Instances and Schema
Fabian M. Suchanek, S. Abiteboul, P. Senellart (VLDB 2012)

Provenance



Statements can be illegally copied over from another ontology. We develop techniques to trace plagiarisms.

Watermarking for Ontologies (ISWC 2011)
Fabian M. Suchanek, David Gross-Amblard, Serge Abiteboul



We develop logic-based models to integrate Web services into ontologies.

SUSIE: Search using Services and Information Extraction
Nicoleta Preda, F. M. Suchanek, W. Yuan, G. Weikum (ICDE 2013)

MAX-PLANCK-GESELLSCHAFT

max planck institut informatik

20th international world wide web conference
28th march - 1st april 2011
hicc hyderabad, india

Casting a Web of Trust over Wikipedia: an Interaction-based Approach

Silviu Maniu, Talel Abdessalem, Bogdan Cautis; Télécom ParisTech – CNRS LTCI, Paris, France, {firstname.lastname@telecom-paristech.fr}

Our goal

Uncover a signed network over Wikipedia contributors from their interactions.

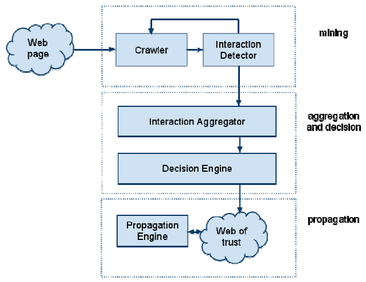
Signed link = a representation of the degree of trust/similarity or distrust/dissimilarity between two users

Several signed networks are already present in social media: Epinions (trust/distrust tags), Slashdot (friend/foe), Wikipedia Elections (support/oppose votes)

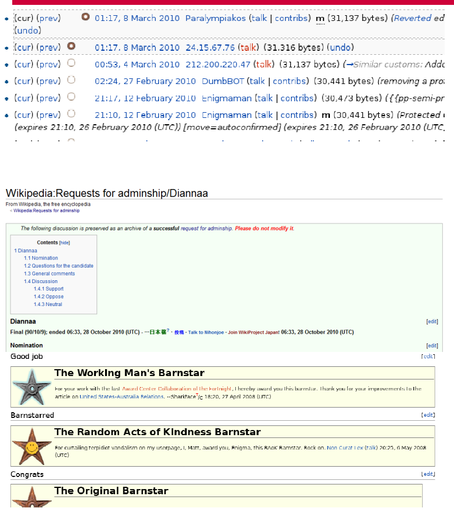
Motivation:

- Social applications can be enhanced by knowing such signed links (social search systems, recommender systems, trust and reputation, etc.)

General Architecture



Interactions in Wikipedia



Article editing:

- on text content (inserting, deleting and replacing text between the contributors)
- on the article revisions (reverting/discarding a version of an article and restoring another)

Adminship election:

- Contributors participate in so called requests for adminship, elections in which contributors can:
  - Support the candidate
  - Oppose the candidate

Interactions on user pages:

- Contributors can give each other prizes called barnstars (generally for good behavior)

Deciding link signs

+	−	−	+	−	+	−	+
insert	replace	delete	restore	revert	support	oppose	barnstars
operations on article text			operations on article revisions		adminship elections (RFAs)		user pages

The interaction vector = aggregation of all interactions for each pair of users

The decision process: -1 (negative) or +1 (positive) link

1. Annotate the atomic interactions with signs (positive or negative, as shown above)
2. Each interaction votes with its weight (measure) for the sign of the higher-order type
3. All types vote for the final link sign

WikiSigned - the resulting network

- 71,770 nodes and 463,312 edges, of which 85.93% positive

Dataset

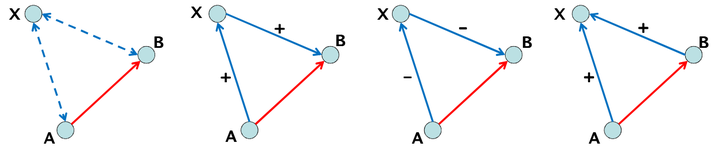
From the politics domain of the English Wikipedia:

- we extracted 320 articles, for a total of 442,297 revisions submitted by 105,177 contributors
- resulting in 800,057 total interactions, in which participate 42,631 adminship votes and 2,913 barnstars

Validation

Does this network represent a *plausible configuration of link signs*?

First approach: comparing our network with three existing explicit signed networks (Epinions, Slashdot, Wikipedia Elections)



the general case t_1 t_6 t_9

Using the concept of link triads and the predictions of two social theories: balance and status, for measuring:

The global properties of WikiSigned

- Our network has similar triad distribution as the explicit networks
- And it has the global structure of a network in which status theory holds (only one contradiction for triad signs)

The local properties of WikiSigned

- It can self-predict its link signs with 0.822 accuracy (AUC of 0.899)
- Good accuracy also in cross training-predicting (training on the row data and predicting on the column data)

	Epinions	Slashdot	Elections	WikiSigned
Epinions	0.906	0.905	0.787	**0.727**
Slashdot	0.929	0.806	0.792	**0.732**
Elections	0.922	0.895	0.814	**0.733**
WikiSigned	**0.889**	**0.844**	**0.784**	**0.822**

Second approach: application-level validation

- Predicting the importance or quality of articles by using the knowledge of link composition (number of positive and negative links) in training predictive models
- This knowledge of link signs helps the prediction when we predict the article importance

feature	Quality	Importance
Contributors	0.683	0.566
Contributors + normal links	0.740	0.779
Social links	0.807	0.750

Maui, Hawaii
CIKM 2012
Oct 28 - Nov 2, 2012

primates

MINES TELECOM
INSTITUT Mines-Télécom

A Privacy Management System for Social Networks

Imen BEN DHIA, Talel ABDESSALEM, Mauro SOZIO
Institut Mines-Telecom, Telecom ParisTech, CNRS LTCI, Paris, France
first.last@telecom-paristech.fr
http://dbweb.enst.fr/

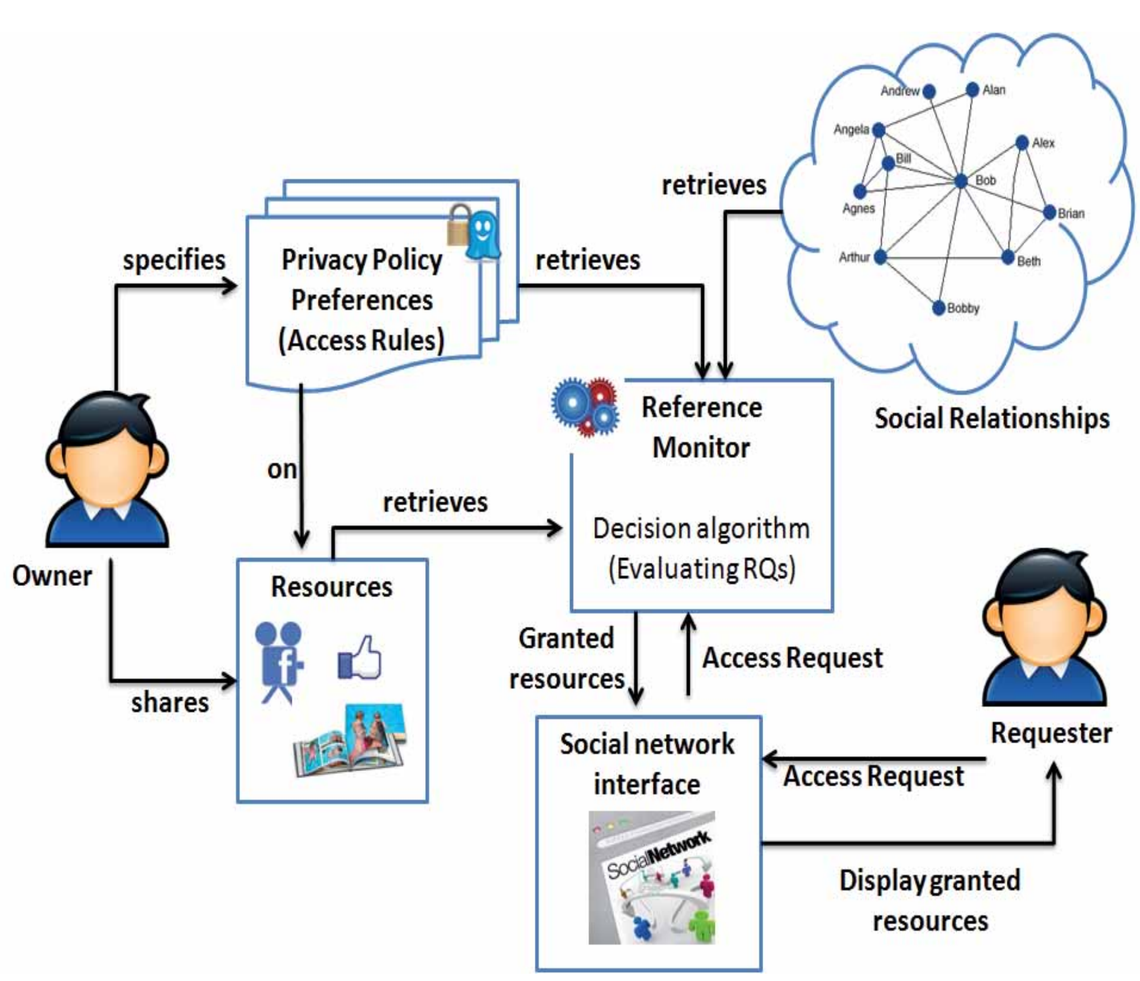
Context and Problem

- Social network users :
  - have different kinds of relationships (friend, colleague, etc.)
  - share content (personal data, photos, videos, etc.)
- Social networks usually grow quickly in terms of number of users, relationships established and pieces of shared information.

Problem: *Social network users have difficulty with specifying which information should be shared with whom.*

Solution : Enable users to specify their privacy policy preferences in a more flexible and efficient way than existing privacy management systems.

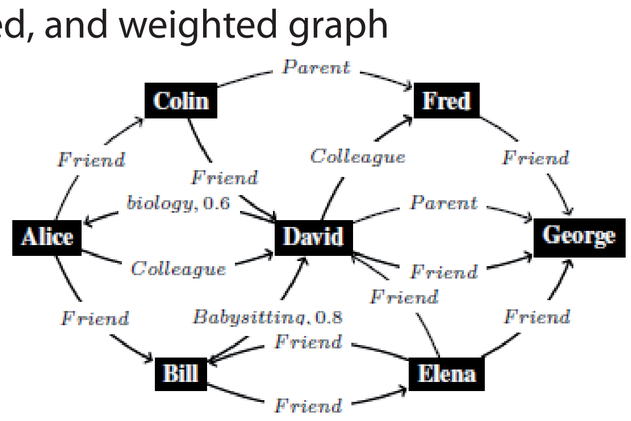
System Overview



Access Control Model

- **Social network model :**
  - Directed, edge-labeled, and weighted graph

(V, E, Σ, φ)



A Social Network Subgraph

- **Access rules specification based on reachability constraints :**
  - Semantics of the links, Links direction, Indirect relationships, Distance, User properties, Trust.

- **Access Rule (AR) :** Specification of the profiles of authorized users to access a given resource.

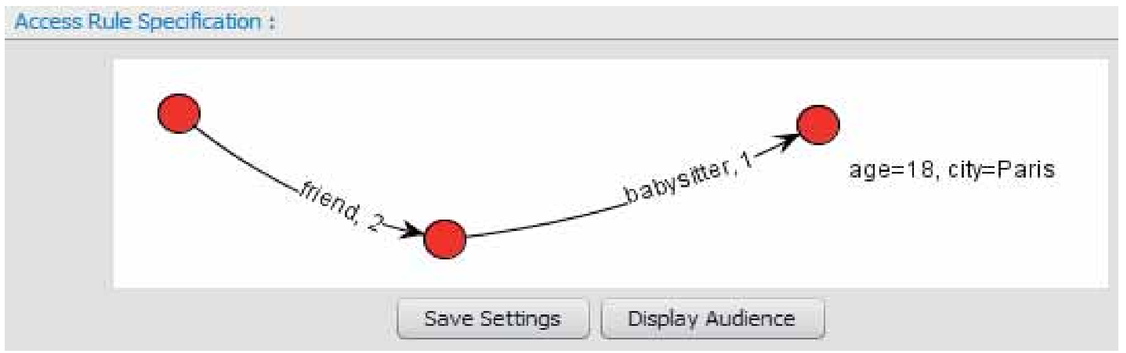
$$AR = (o, p, t_{min})$$

$$p = \{s_i\} \qquad s_i = (r, dir, l, C)$$

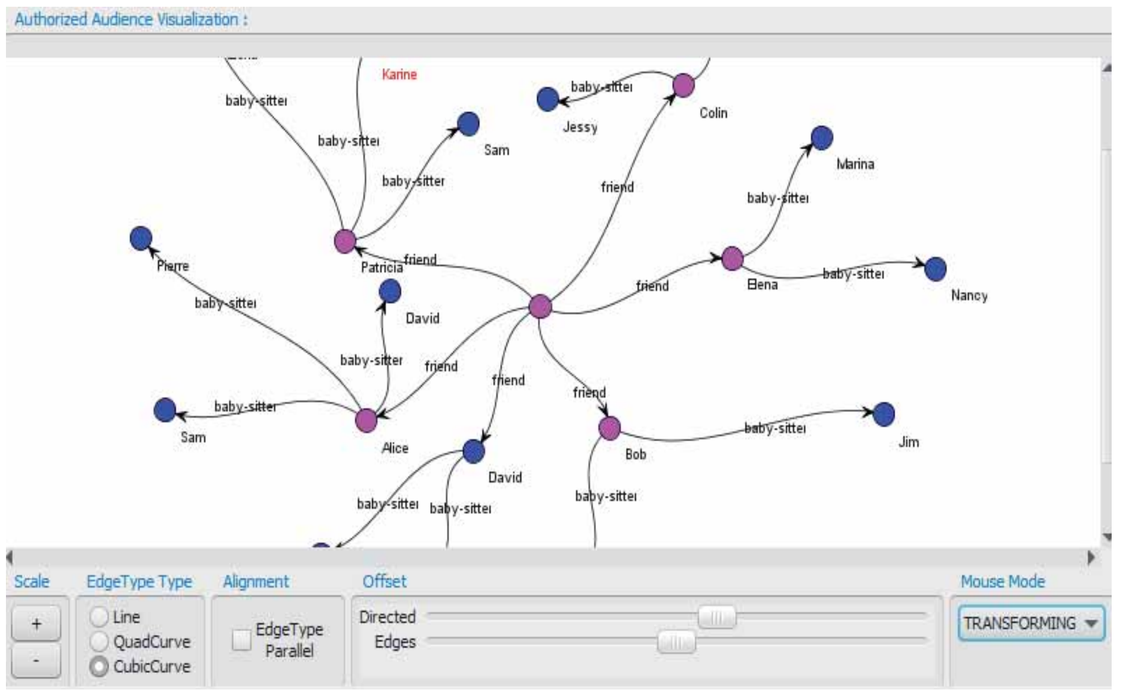
- **Example :**
  - p = Friend $^{+}$ [1, 2][city = Paris]/BabySitter $^{+}$ [1][city = Paris]
  - t_{min} = 0.8

Demonstration

- **Dataset :** LiveJournal social network, ~5 Million users, ~80 Million edges
- **Access Rule Specification :**



- **Authorized Audience Visualization :**



A Probabilistic XML Merging Tool

Talel Abdessalem
Télécom ParisTech
Paris, France

Mouhamadou Lamine BA
Université Cheikh Anta DIOP
Dakar, Senegal

Pierre Senellart
Télécom ParisTech
Paris, France

http://dbweb.enst.fr/

What this tool aim at...

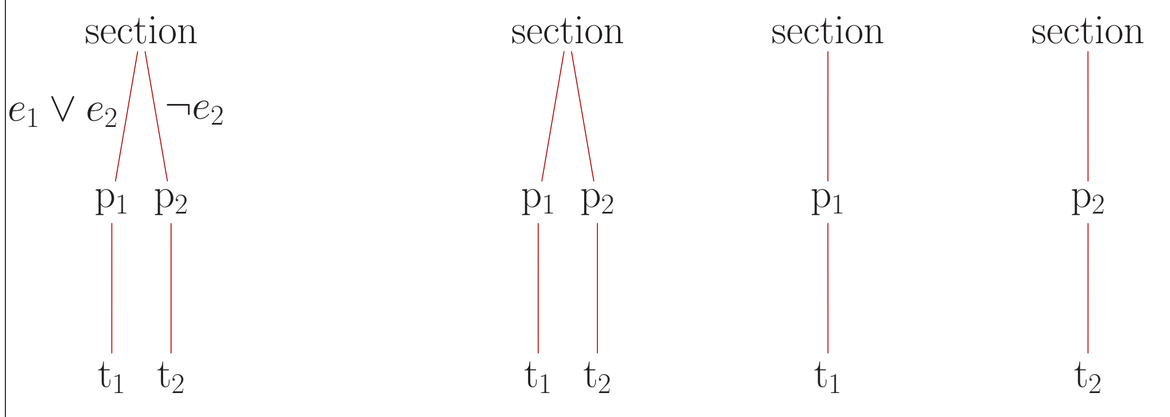
– Representing the outcome of semi-structured documents integration as a probabilistic tree

– Evaluating the uncertainty (modeled as probability values) of the result of the merge

– Querying the probabilistic repository with a subset of the XPath query language

Application domain: Wikipedia revisions
The tool enables merging the revisions of a given Wikiepdia page with:
→ an efficient evaluation of the uncertainty of the obtained result
→ an automatic management of conflicts.

Probabilistic Documents



$\Pr(e_1) = 0.7$
$\Pr(e_2) = 0.6$
P-Document

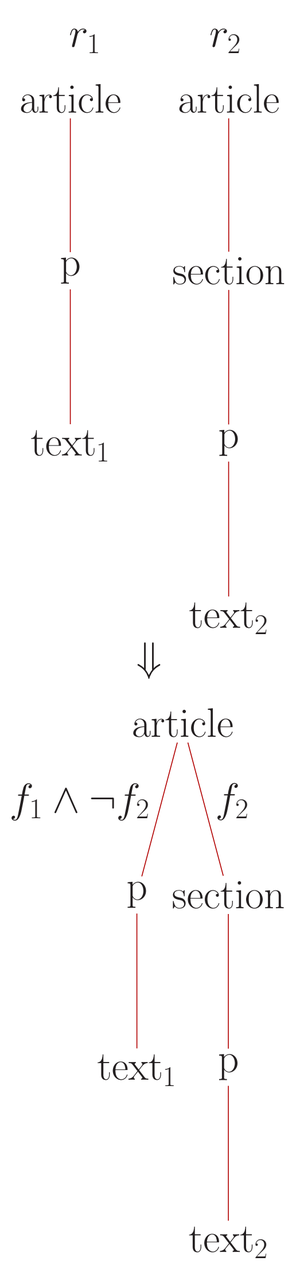
0.28 | 0.6 | 0.12

Corresponding possible documents and their probabilities

Merging of Wikipedia revisions

– A two-way tree merging technique for P-Documents

– Two steps: Matching of Revisions and Merging Matches



The result of the merge process

1. Matching of Revisions

Input: two revisions r_{k-1} and r_k and their associated event formula.
Output:
– **Deleted nodes x: $x \in r_{k-1}$ and x has no match in r_k.**

– **Added nodes x: $x \in r_k$ and x has no match in r_{k-1}.**

– **Matched couples (x, y): $x \in r_{k-1}$ and $y \in r_k$ match.**

2. Merging Matches

– **Deleted nodes:**

$$fie_{\text{new}}(x) = fie_{\text{old}}(x) \wedge (\neg f_k)$$

– **Matched couple:**

$$fie_{\text{new}}(x) = fie_{\text{old}}(x)$$

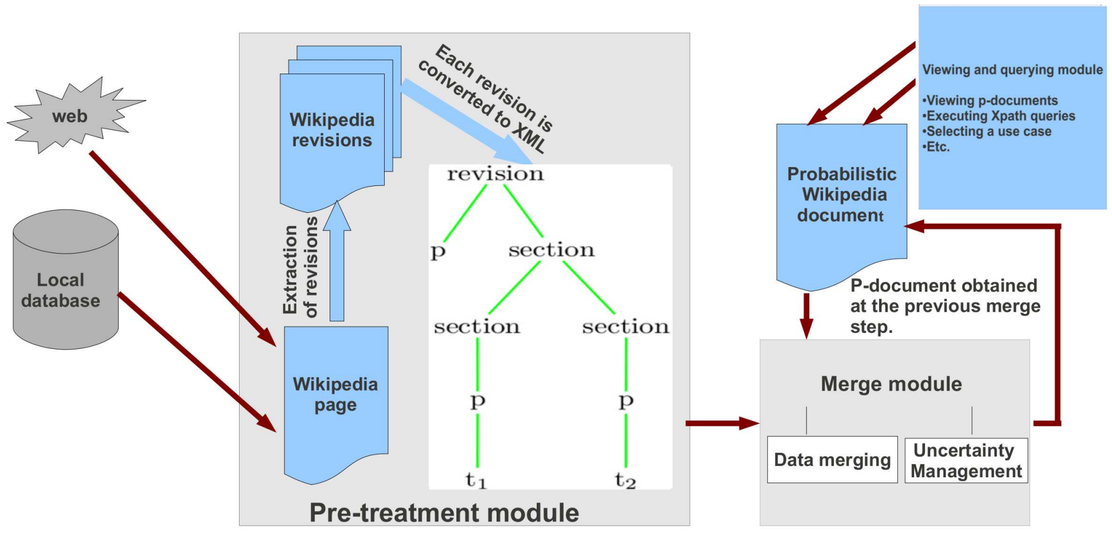
– **For added nodes:**

$$fie_{\text{new}}(x) = f_k$$

or

$$fie_{\text{new}}(x) = fie_{\text{old}}(x) \vee f_k$$

Architecture of the system



Description of the system

– System for managing Wikipedia documents.

Features

– A keyword-based search engine for Wikipedia pages
– Extracting the revisions of a given page
– Selecting the list of revisions to merge
– Building one's own Wikipedia article
– Displaying the result of the merge
– Demonstrating a certain number of use cases
– Using a subset of XPath query language

On-Line Learning Gossip Algorithm in Multi-Agents Systems with Local Decision Rules

Parties prenantes

MINISTÈRE DE LA DÉFENSE DGA

TELECOM ParisTech

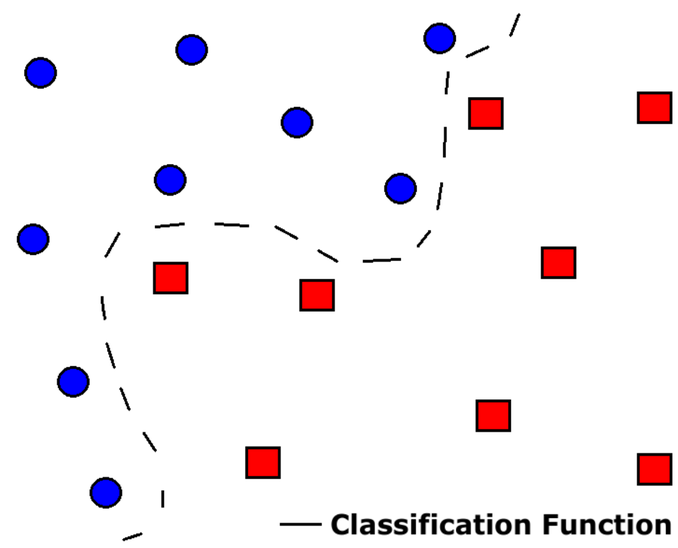
Auteurs

S. Clémençon, P. Bianchi, G. Morral et J. Jakubowicz

BACKGROUND

Motivation and applications

- Problem : investigate the **binary classification**
- Context : processing **BigData** for **statistical learning**
- Solution : implement in an **on-line** and **distributed** fashion



Learning problem

Given a training data set $(\boldsymbol{X},\boldsymbol{Y}) = (X_i, Y_i)_{i\geq 1}$, find the best prediction rule $sign(H^*)$ such the classifier function $H(X, \boldsymbol{\theta})$:

$\boldsymbol{\theta}^* = \min_{\boldsymbol{\theta}} R_\varphi(H(X, \boldsymbol{\theta}))$ → *minimizes the Risk function* R_φ

Particular case :

i) quadratic cost φ : $R_\varphi(H(X, \boldsymbol{\theta})) = ½\, E[\, (1- \boldsymbol{Y}H(X, \boldsymbol{\theta}))^2 \,]$

ii) mixture of experts : $H(X, \boldsymbol{\theta}) = \Sigma_j \theta_j\ h_j(X)$

PROPOSED DISTRIBUTED LEARNING

On-line Learning Gossip Algorithm (OLGA)

- A **distributed stochastic gradient descent** approach where the estimated parameter sequence $(\boldsymbol{\theta}_n)_{n\geq 1}$ is performed in 2 steps :

  [Gossip step] At iteration n, each agent i transmits $X_{n,i}$ to all randomly selected neighbours j with *probabiliy p* and obtain $h_j(X_{n,i}, \theta_{n-1,j})$

  [Local descent step] each agent i update its estimated parameter $\theta_{n,i}$ as follows :

$$\theta_{n,i} = \theta_{n-1,i} + \gamma_n \nabla_i h_i(X_{n,i}, \theta_{n-1,i})(Y_{n,i} - Y'^{(V)}_{n,i})$$

where

- $(\gamma_n)_{n\geq 1}$ is a decreasing step size sequence
- the r.v. $Y'^{(V)}_{n,i}$ *is an unbiased estimate of the global decision* $H(X_{n,i}, \theta_{n-1,j})$ *given by*

$$Y'^{(V)}_{n,i} = h_i(X_{n,i}, \theta_{n-1,i}) + 1/p\ \Sigma_j \delta^j_{n,i} h_j(X_{n,i}, \theta_{n-1,j})$$ *and* $(\delta^j_{n,i})$ are independent Bernoulli r.v 's $B(p)$

Asymptotic behaviour

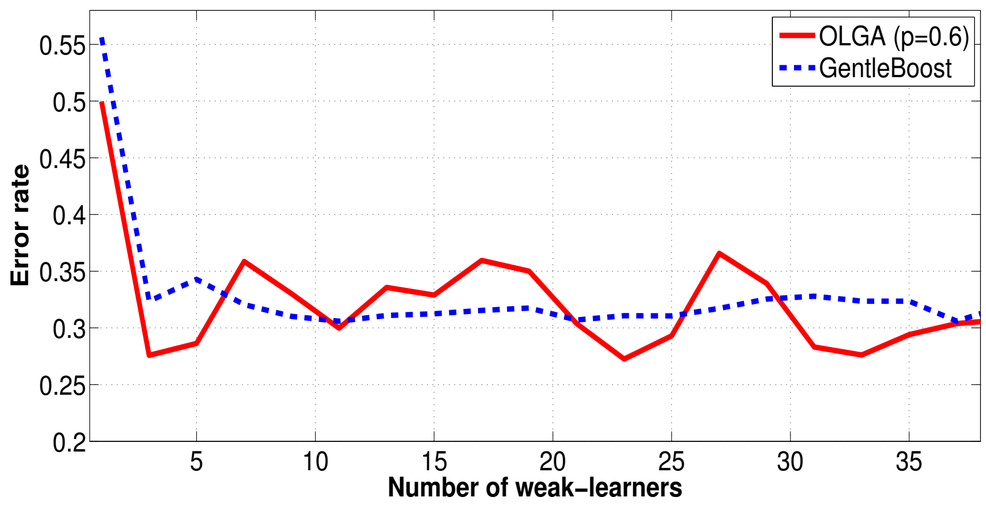
Theoretical results (under suitable assumptions)

- ✓ Consistency : $(\boldsymbol{\theta}_n)_{n\geq 1}$ **convergence *a.s.*** to the set of stationary points of R_φ
- ✓ Conditional Central Limit Theorem : qualify the error variance excess

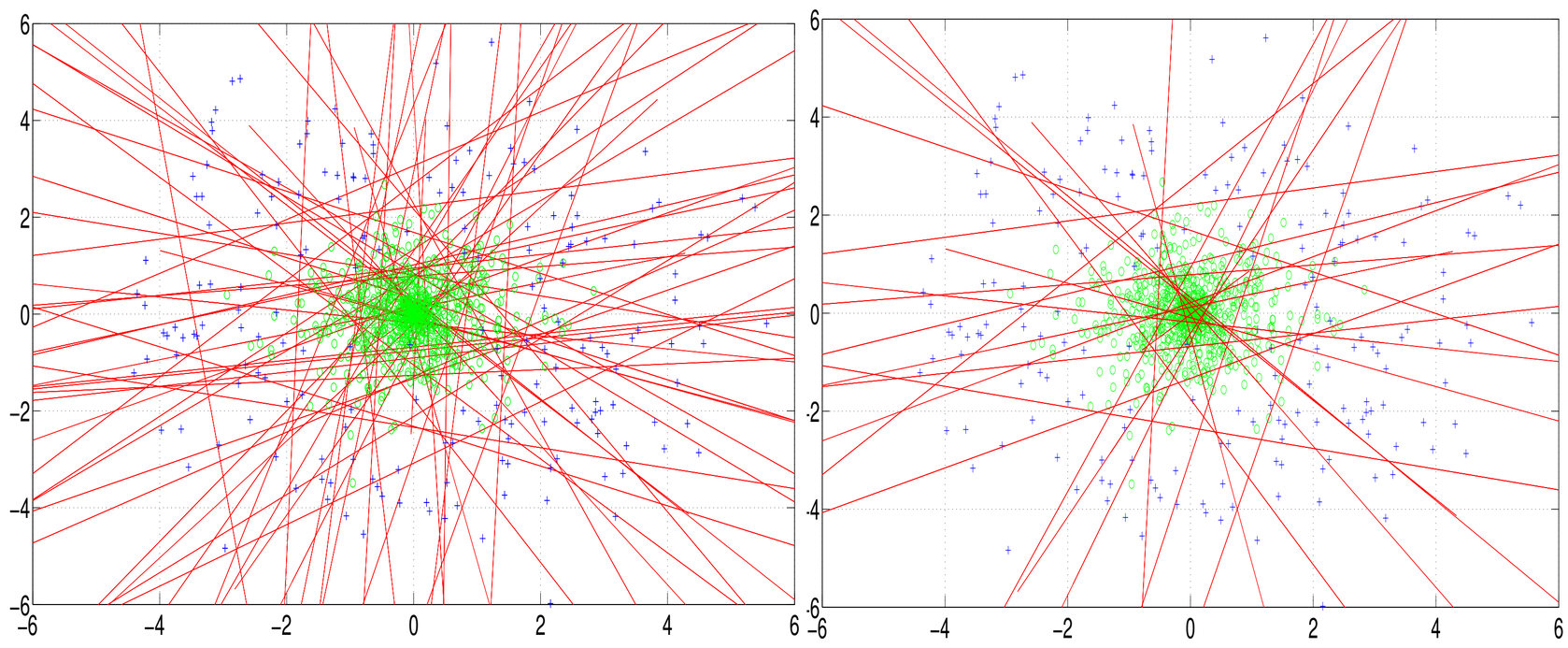
$$\gamma_n^{-1/2} (\theta_n - \theta^*) \rightarrow N(0 , \Sigma(\Gamma^*))$$

Γ^* : error in a centralized case **+ error excess induced by sparsification**

- ✓ The average network throughput is **reduced** by a factor **(1-p)**



Performance comparison between the centralized GentleBoost and OLGA for a benchmark dataset



Result classification with OLGA (-) on a simulated binary dataset (+ and o) using weak classifiers

Left : OLGA
Right : OLGA with **agent selection**

at each iteration n, each agent i of V declares idle or **active** under a suitable criterion → time-varying agents set $\boldsymbol{V_n}$

✓ Reduce redudancy classifiers and keep the relevants

26-27 Mars 2014 3ème colloque de l'Institut Mines-Télécom, « Numérique : Grande éc

Contact stephan.clemencon@telecom-paristech.fr

Presented at the IEEE Internation Conference on BigData 2013

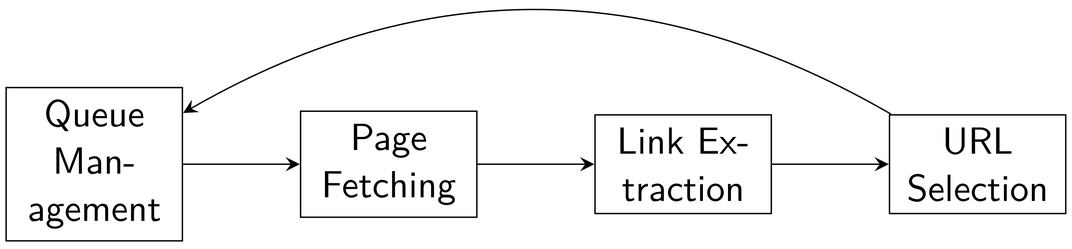
Demonstrating Intelligent Crawling and Archiving of Web Applications

Muhammad Faheem
Institut Mines–Télécom
Télécom ParisTech; CNRS LTCI
Paris, France
muhammad.faheem@telecom-paristech.fr

Pierre Senellart
Télécom ParisTech
& The University of Hong Kong
Hong Kong
pierre.senellart@telecom-paristech.fr

Traditional crawler

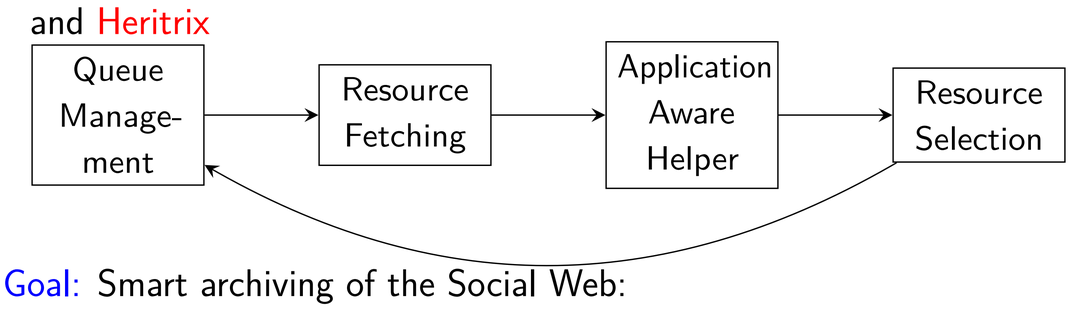
Traditional crawling: independent of the nature of the sites and their content management system



$\Longrightarrow$ Many HTTP requests, no guarantee of content quality

Application-aware helper

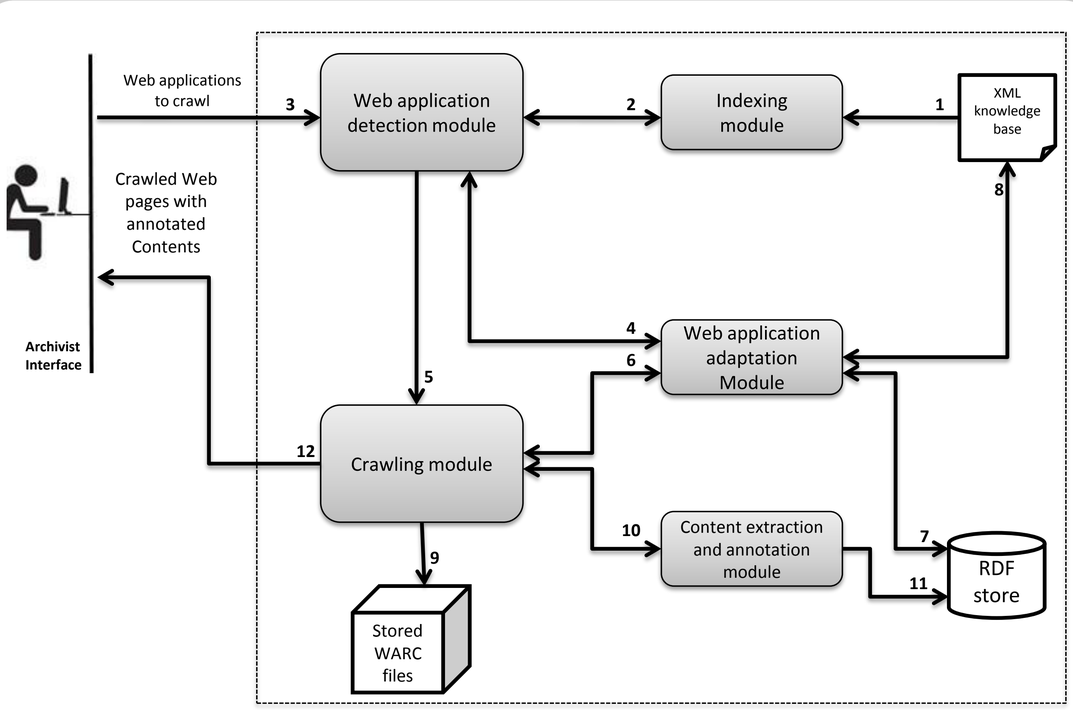
- Different crawling techniques for different Web sites
- Detect the type of Web application, kind of Web pages inside this Web application, and decide crawling actions accordingly
- Directly targets useful content-rich areas, avoids archive redundancy, and enriches the archive with semantic description of the content
- Implemented in 2 Web crawlers: Internet Memory Foundation crawler and Heritrix



Goal: Smart archiving of the Social Web:

1. Performing intelligent Crawling
2. Archiving Web objects

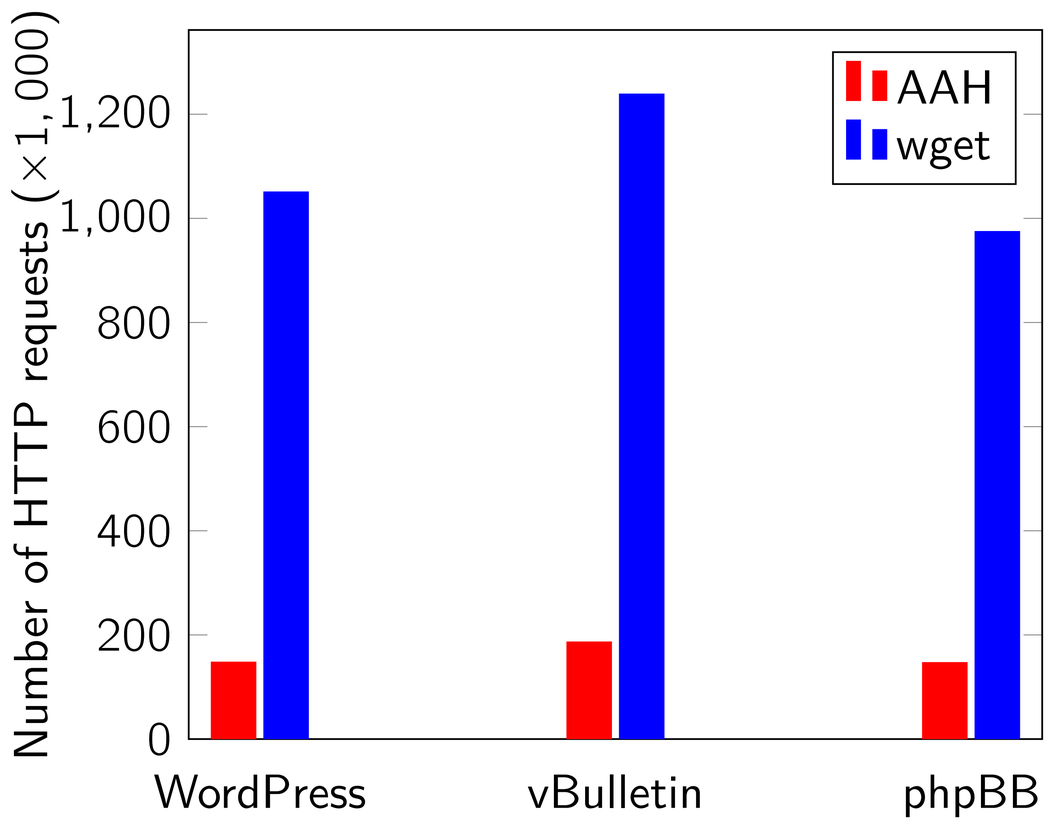
Architecture



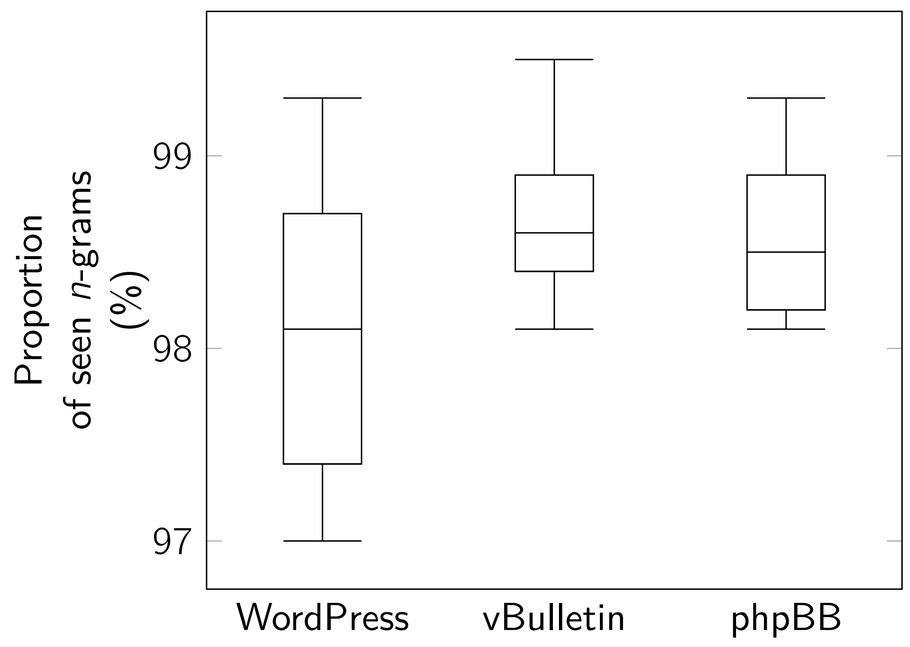
Methodology

- Knowledge base of known Web application types, algorithms for flexible and adaptive matching of Web applications to these types
  - Declarative, XML-based format
  - Integrated with YFilter for efficient indexing of KB.
- Type detected using URL patterns, HTTP metadata, textual content, XPath patterns, etc. E.g., vBulletin Web forum: `contains(//script/@src,'vbulletin_global.js')`
- Different crawling actions for different kinds of Web pages under a specific Web application
- Crawling action: not just a list of URLs; can be any action that uses REST API, complicated interaction with AJAX-based application, and extracts semantic Web objects

Crawl efficiency



Crawl effectiveness



CrowdMiner: Mining association rules from the crowd

Yael Amsterdamer Yael Grossman Tova Milo Pierre Senellart

TEL AVIV UNIVERSITY

TELECOM ParisTech

Introduction

- ***Crowd data sourcing*** collects data from the crowd, often by asking questions
- We want to learn about new domains from the crowd
  - E.g., health-related habits in some population
- Data is not recorded anywhere
- The contents of the domain are unknown
  - Discover what is **interesting** about this domain

What should we ask the crowd?

Data mining for the crowd?

- The discovery of data patterns in databases is done by **data mining.**
- Not suitable for our case
  - People do not remember enough details!

For example, it is unrealistic to expect people to remember every activity they did in the past, everything they have eaten, etc.

- They are far more likely to remember **personally prominent patterns**

"I drink red wine about once a week"

The model

We learn *association rules* of the form $a,b \rightarrow c,d$

- E.g. , *"heartburn"* $\rightarrow$ *"baking soda", "lemon"*

The answers contain

- **Rule support** – frequency of a,b,c,d
- **Rule confidence** – frequency of c,d given a,b
- **Items** (for an open question)

- **Significant rules** – average user support and confidence exceed fixed thresholds
- Users treated as random samples

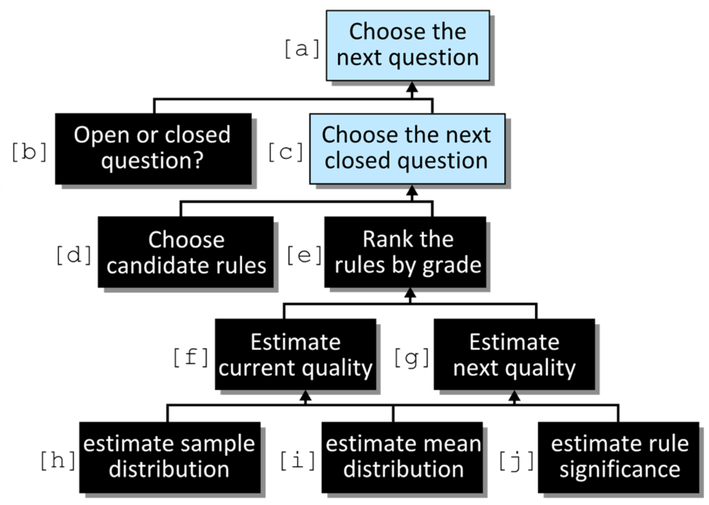
Our approach

- Use **personal summaries** to learn about **general trends**
- Treat individual answers as samples
- Combine two types of questions
  - **Open questions** *"Complete: When I feel* tired*, I usually* go for a walk*"*
  - **Closed questions** *"When you have a heartburn, do you take baking soda and lemon?"*
- Easier for users to answer
- Help digging deeper into their memories

We develop a system prototype *CrowdMiner* that interactively decides what to ask in order to discover significant data patterns

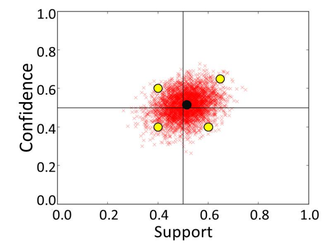
Choosing the Questions

A hierarchy of components that allow estimating the effect of the next question and choosing accordingly



Error Estimations

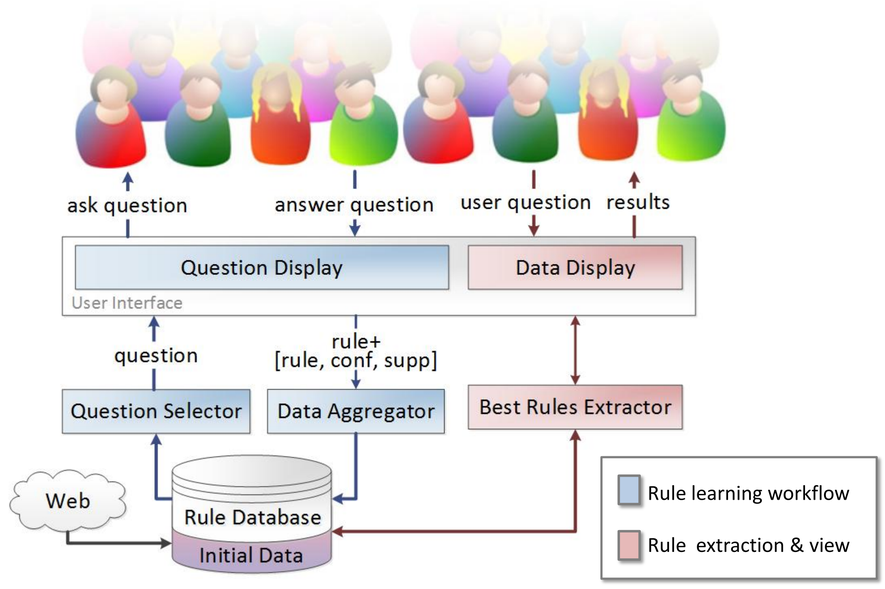
- Not all the users can be asked about every rule
- We want to estimate the probability of making an error – given the current knowledge
  - We learn a distribution of the answer support and confidence
  - **Significance estimation** – by the position of >0.5 of the distribution mass
  - **Error probability** – for the true mean to be on the other side of the thresholds
- The next question is the one expected to minimize the overall error



Well-Being Portal

- Learn about the **health habits** of others – by browsing the portal
  - Sports activities, eating habits, natural treatments
  - …
- Portal users are occasionally prompted with **questions**
  - About their personal habits
  - Computed by our algorithm
- User **answers** are processed to deduce rules (associations) between well-being concepts in the portal
- The portal allows browsing the learned rules

System Architecture



Opinion Dynamics and Community Detection

Institutions

INSTITUT Mines-Télécom

TELECOM SudParis

s@movar UMR 5157

Authors

Julio Cesar Louzada Pinto

Tijani Chahed

Jérémie Jakubowicz

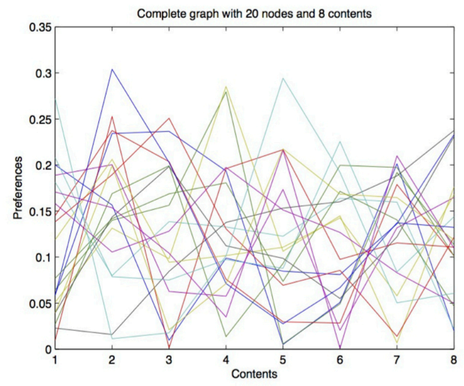
Partners

FONDATION TELECOM

INSTITUT CARNOT Télécom & Société numérique

Objectives

- Develop a stochastic opinion dynamics model with multiple contents and study the asymptotic behavior in simple cases.
- Develop a community detection (graph clustering) algorithm.



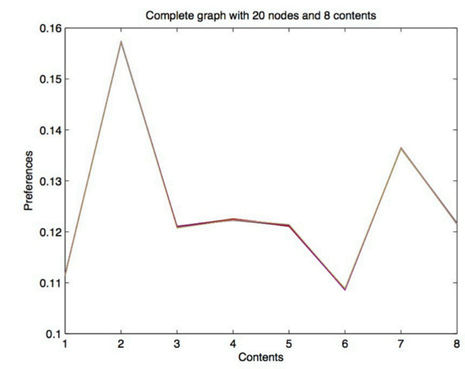
Model

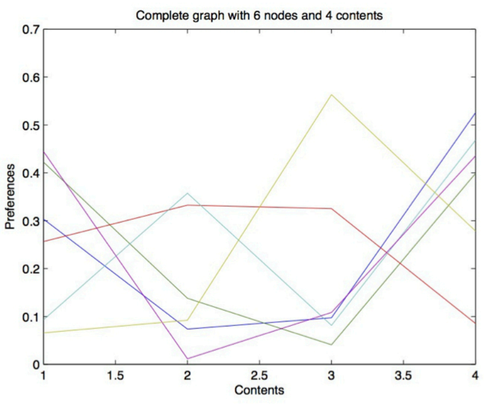
- *N* agents communicate about *K* contents via a graph *G = (V,E)* with inward adjacency matrix *A*.
- Agent *i* has score $X_t^{i,k}$ for content *k* at time *t.*
- *Preferences* $P_t^{i,k}$ are normalized scores, i.e., $P_t^{i,k} \propto X_t^{i,k}$.
- Agents update their scores linearly as

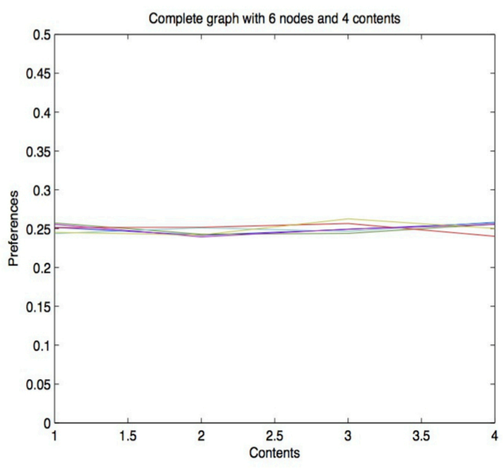
$$X_{t+1}^{i,k} = X_t^{i,k} + \sum_{j \to i} 1_{u_{t+1}^j = k}$$

with $P\left(u_{t+1}^j = k \mid F_t\right) = \left(f\left(P_t\right)\right)_{jk}$ the probability of agent *j* broadcasting content *k.*

- Function $f : \left(\Delta_K\right)^N \to \left(\Delta_K\right)^N$ models the way agents choose the contents to broadcast, where $\left(\Delta_K\right)^N$ is the set of the *NxK* stochastic matrices.

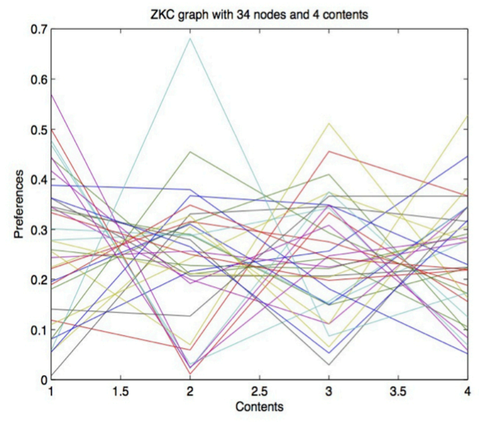


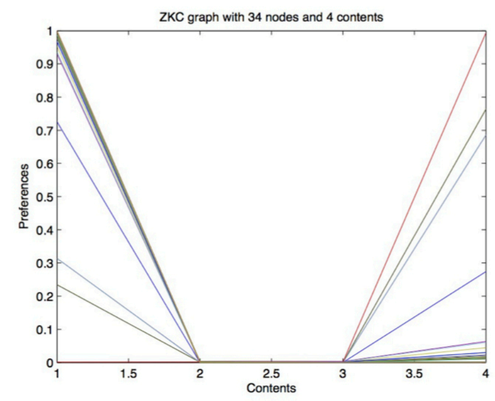




Types of function f

- f(x) = x – identity function.
- $\left(f\left(x\right)\right)_{ik} = \dfrac{e^{\beta x_{ik}}}{\sum_l e^{\beta x_{il}}}$ - soft-max function.
- $\beta \ll 1 \to \left(f\left(x\right)\right)_{ik} \sim \dfrac{1}{K}$
- $\beta \gg 1 \to \left(f\left(x\right)\right)_{ik} \sim 1_{k = argmax_j x_{ij}}$



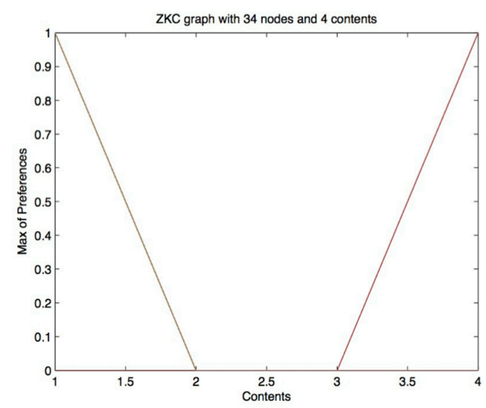


Results

- For f(x) = x: there exists a random variable $P_\infty \in \left(\Delta_K\right)^N$ such that $P_t \to P_\infty$ almost surely.

If *G* is strongly connected, then there exists a random variable $\pi \in \Delta_K$ such that $P_t \to 1\pi^T$ almost surely .

- For soft-max exponential with $\beta \ll 1$: If $inf_i \sum_j A_{ij} > 0$, then there exists a $\beta_{min} > 0$ such that for all $\beta \in \left[0, \beta_{min}\right]$ we have that $P_t \to \dfrac{11^T}{K}$ almost surely.
- If $\beta \gg 1$, then we expect the graph to be clusterized in communities which broadcast the same content; the one with the maximum preference. This creates an algorithm for community detection.
- Publication: A stochastic opinion dynamics model with multiple contents, CDC, Firenze, December 2013.



Journée Futur et Ruptures

January 2014

Contact julio.louzada_pinto@telecom-sudparis.eu Site web www-public.it-sudparis.eu/~louzada

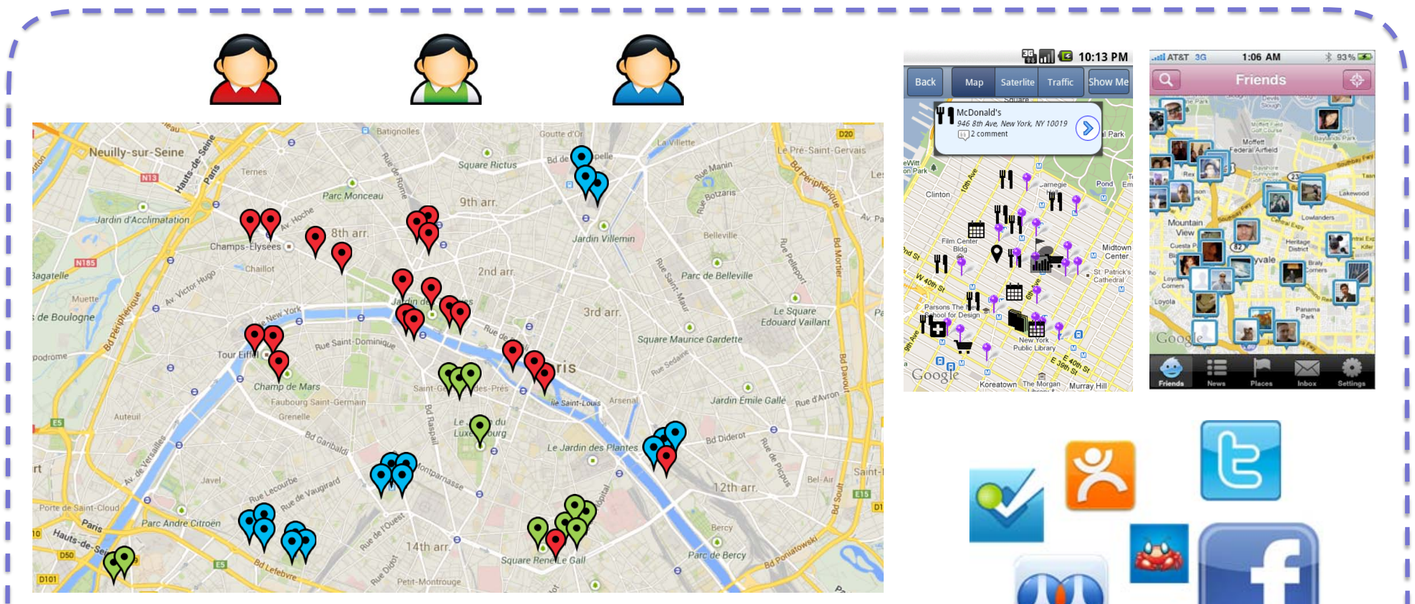
From Big Social Media Data to Service Personalization

Authors : Daqing Zhang, Dingqi Yang, Zhu Wang, Zhiyong Yu

Location Based Social Network Data Analytics

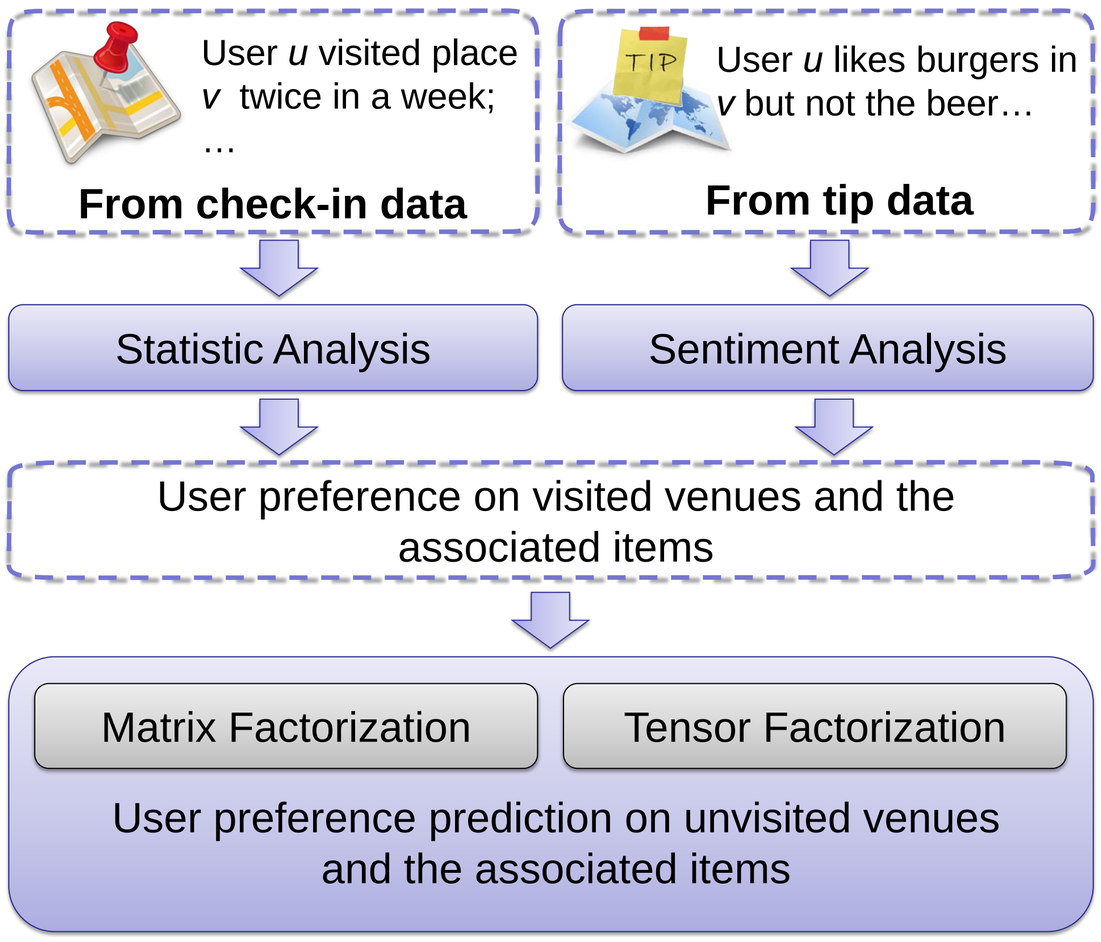
- With the increasing popularity of location based social networks, users generated significant volume of heterogeneous social media, e.g.,
  - Texts
  - Photos
  - Videos
  - Presences
  - …

- These digital footprints massively contain users' fine-grained preference
- Understanding this user preference can enable ubiquitous, personalized location based services.



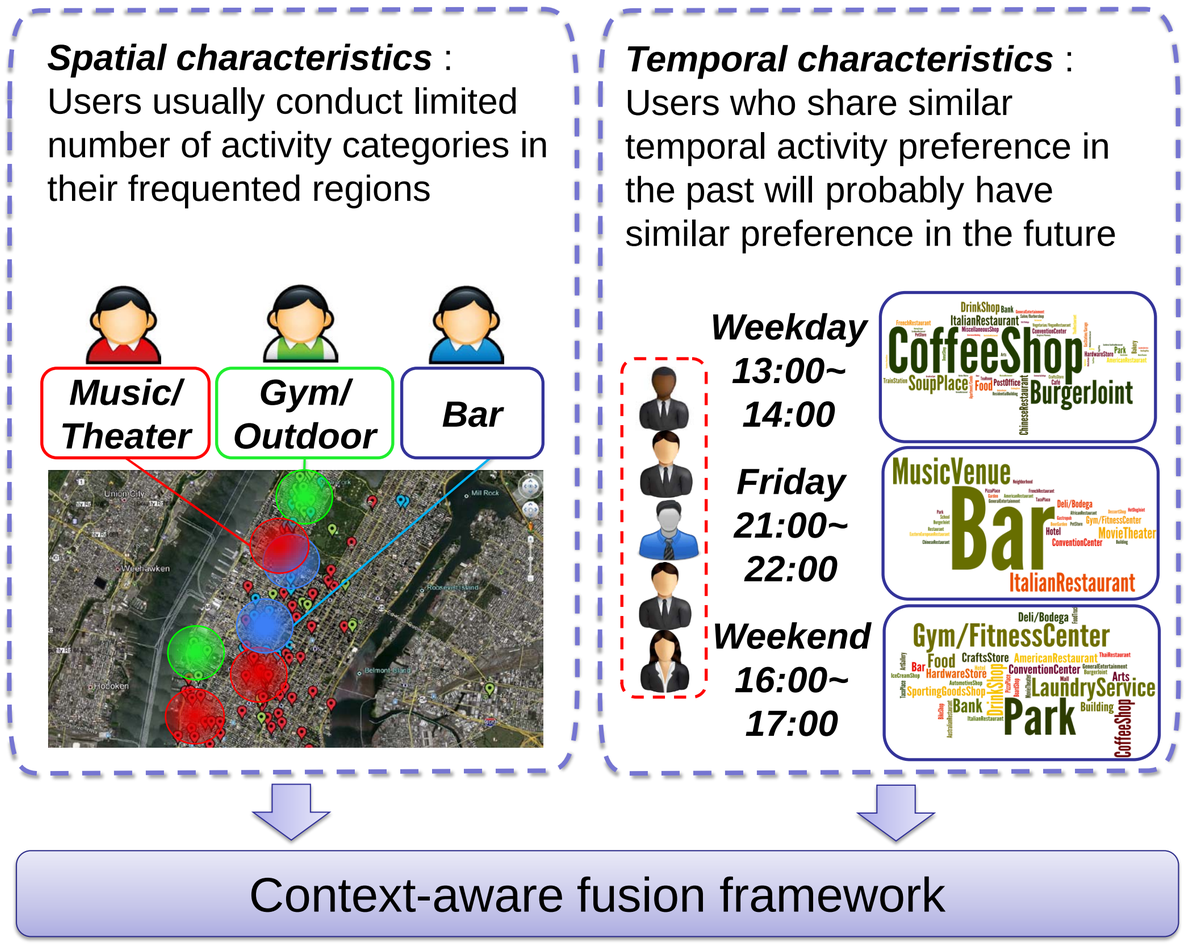
Preference Awareness

- Extracting fine-grained user preference on venues from heterogeneous data.
- Predicting user preference on unvisited venues.



Context Awareness

- Studying and modeling spatial temporal characteristics of user activity.
- Inferring a user's interest according to his current context.



An Example of Personalized Location Based Services

Contact Daqing.Zhang@telecom-sudparis.eu
Site web http://www-public.it-sudparis.eu/~zhang_da/

A Context-Risk-Aware Recommender System

Institut Mines-Télécom

TELECOM SudParis

s@movar UMR 5157

cnrs

Authors

Djallel Bouneffouf, Amel Bouzeghoub and Alda Gançarski

Context

s@movar UMR 5157 — TELECOM SudParis — cnrs — Mines-Télécom Institut

Nomalys

Objectives and Issues

Motivations

- Provide personalized and context-aware recommendations in mobile environments
- Consider content dynamicity and user's situations risk level

Key Challenges

- Infer higher-level goals from low-level observed operations
- Handle cold start and sparseness effect
  - Requires a large amount of information in order to make accurate recommendations
- Exploration vs. exploitation tradeoff
  - How to sacrifice a short term small reward to privilege larger rewards in the long term?
  - How to associate the situations risk level to the exploration/exploitation tradeoff?

Key Words & Key Technologies

- **Context-Aware Recommender Systems** (CARS) combine characteristics from context-aware systems and recommender systems in order to provide personalized recommendations to users in ubiquitous environments.
- **Machine learning** algorithms can be used to learn models and predict documents
  - **Reinforcement learning** is learning what to do: how to map situations to actions
- In probability theory, the **multi-armed bandit problem** models an agent that simultaneously attempts to acquire new knowledge and optimize her decisions based on existing knowledge
  - In each round, a learner takes an **action** (or **arm**) and in return receives a numerical **reward**
  - The goal is to optimize action-selection policy to maximise the total reward received
  - The learner needs to **explore** (try) the different actions and **exploit** the seemingly most rewarding arms
  - In practice, the learner has access to **contextual information** in each round to infer which action leads to the highest rewards

Contributions

Models

- Modeling the context-aware recommender system as a bandit algorithm
- Modeling user, context, situation and risk

Algorithms

- A new semi-uniform strategy: contextual-epsilon-greedy strategy
  - Combining content-based filtering and reinforcement learning
- An algorithm R-UCB
  - Computes the probability of exploration by using the situation risk level $R(S^t)$
  - Three methods of risk computing
    - Using situations similarity (R_m)
    - Using situations concepts (R_c)
    - Using a Gaussian distribution (R_v)

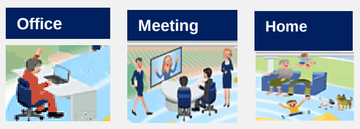
General Approach

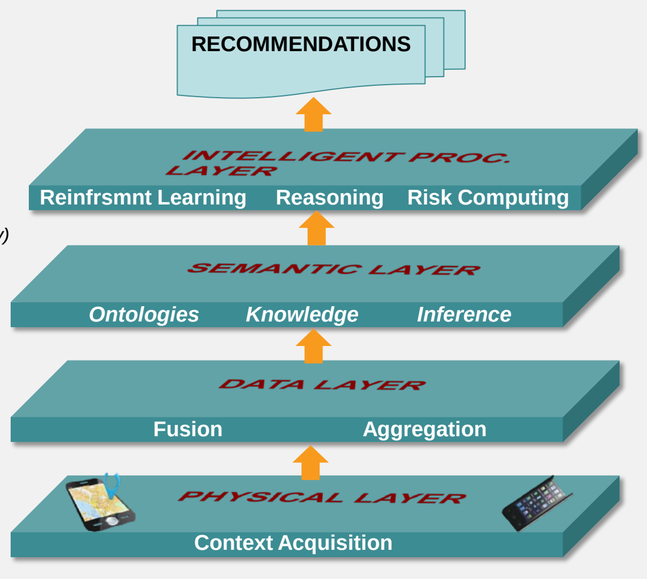
High level user's situation

$O_{Location}.Evry$, $O_{Temps}.afternoon(awd)$, $O_{sociale}.Company)$

Low level user's situation

("2.86, 1.34 ", Mon Oct-3-13 :10 :00-2013, "Arval")





Evaluation Results

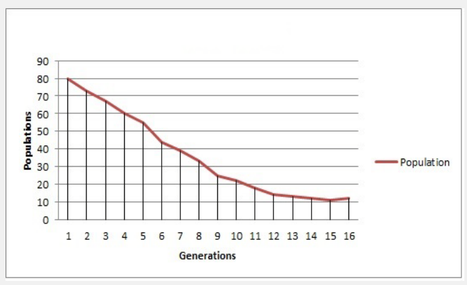
Data Set and Parameters

Nomalys Data Set

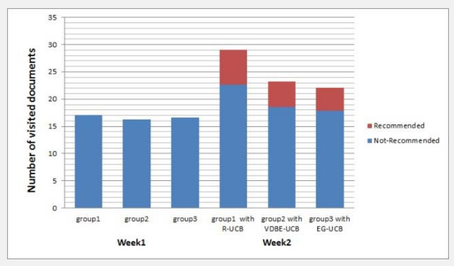
- 356 738 situations
- 5 518 566 navigations data
- 3500 users

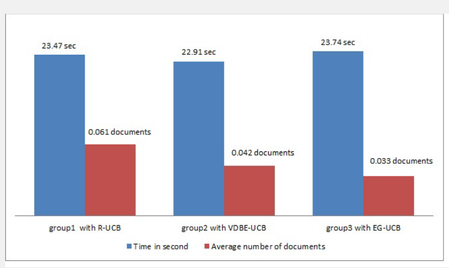
Genetic Algorithm

- Population: 80 Chromosomes
- Chromosomes: ϵ_{min} , ϵ_{max}, Threshold
- Results: $\epsilon_{min} \in [0.05; 0.13]$, $\epsilon_{max} \in [0.47; 0.56]$, Threshold $\in [0.7; 0.82]$



Online Evaluation





Offline Evaluation