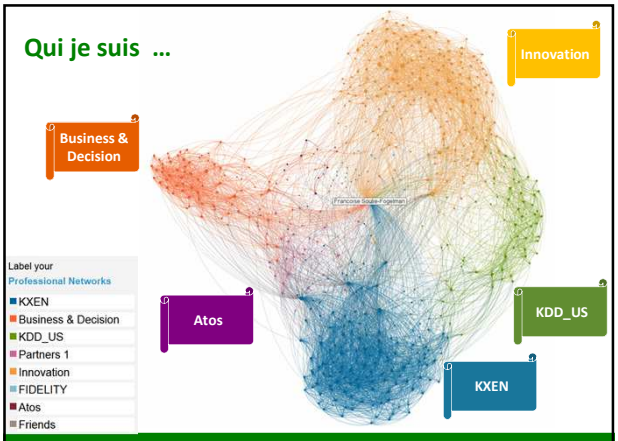
Utilisation des réseaux sociaux pour le data mining

Françoise Soulié Fogelman
francoise.soulie@outlook.com

CNAM
Séminaire de Statistique appliquée
Mercredi 13 novembre 2013

Qui je suis …



2

… et mes centres d'intérêt

Domaines de la communauté KDD http://www.kdd.org/kdd2013

The darker the color, the more recent the trend

3

Agenda

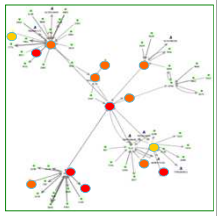
- Qu'est ce qu'un réseau social
- Analyse des réseaux sociaux & data mining
- Un exemple: la fraude à la carte bancaire

4

Qu'est ce qu'un réseau social

Les données en réseau

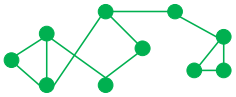
- Le Data mining repose sur une hypothèse fondamentale
  - Les observations sont i.i.d.
    - Indépendantes, identiquement distribuées
- Mais en pratique ce n'est souvent pas le cas
  - Les individus ne sont pas indépendants
    - Réseau de relations sociales
- L'analyse des réseaux sociaux – ou Social Network Analysis – vise à modéliser ces interactions



6

Qu'est ce qu'un réseau social ?

- **Un réseau social est un graphe**
  - Entités : les nœuds
  - Interactions : les liens entre nœuds
    - Les nœuds liés sont dits **amis** ou **voisins**
  - Les nœuds peuvent avoir des attributs (nom, tags …)
  - Les liens aussi (poids)
- **Tout réseau social est un graphe**
  - Avec des propriétés particulières
- **Mais tout graphe n'est pas un réseau social**
- **Il existe**
  - **Des réseaux sociaux dirigés / non dirigés**
  - **Des réseaux sociaux explicites / implicites (ou directs/indirects)**



7

Réseau social explicite

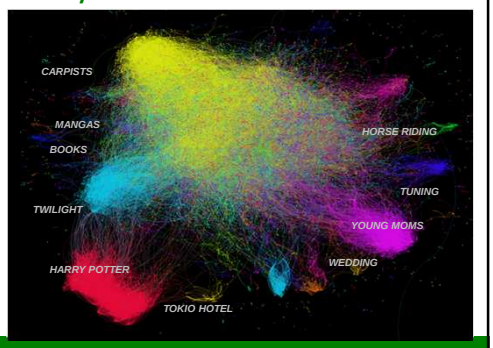
Sites sociaux : Facebook

http://www.facebook.com/notes/facebook-engineering/visualizing-friendships/469716398919

8

Réseau social explicite

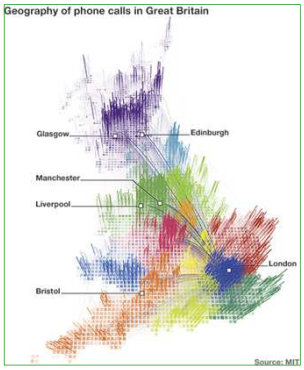
Sites sociaux : Skyrock http://fr.skyrock.com/



9

Réseau social explicite

Appels téléphoniques



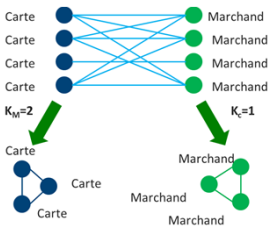
http://www.flickr.com/photos/squeaki/6181086582/

10

Réseau social implicite

- **Les graphes bipartites**
  - Deux sortes de nœuds
    - url & tags de contenus, clients & produits, cartes de crédit & marchands par exemple
  - Un lien ne relie que 2 nœuds de sortes différentes
- **Projeter le graphe bipartite**
  - On obtient 2 graphes unipartites « implicites »
  - Pondération
    - Nombre de Cartes/Marchands communs
    - Support, confiance …

$$Supp(a \rightarrow u) = \frac{\#\ Marchands\_visités\_par\_a\_et\_u}{\#\ Marchands}$$



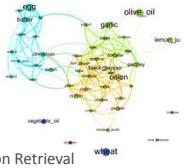
11

Analyse des réseaux sociaux & data mining

Analyse de réseaux sociaux

Social Network Analysis : beaucoup de domaines d'étude

- Par exemple (cours de Lada Adamic)
  https://www.coursera.org/course/sna, http://open.umich.edu/education/si/si508/fall2008
  - Week 01: Introduction to Networks
  - Week 02: Basic Network Metrics
  - Week 03: Centrality
  - Week 04: Prestige and Small World Networks
  - Week 05: Power Laws and Growing Networks
  - Week 06: Graph Traversal
  - Week 07: Community Structure
  - Week 08: Search
  - Week 09: Networks in the Web and Information Retrieval
  - Week 10: Information Diffusion
  - Week 11: Networks Over Time
  - Week 12: Network Resilience

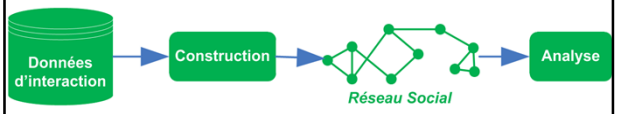


→ **Graph mining (Mining Networked data)**

13

Analyse de réseaux sociaux

- **Le processus**



- **L'analyse**
  - Descriptive
    - Visualisation du réseau, des communautés
    - Caractéristiques du réseau
    - Diffusion …
  - Prédictive
    - Extraction de « variables sociales »

Nœud
- **Caractéristiques structurelles**
  - Degré, communauté
- **Attributs**
  - Dans le cercle, la communauté

14

Analyse de réseaux sociaux

Pourquoi utiliser des variables sociales ?

- **Les variables sociales « résument » les interactions**
  - La non-indépendance est donc prise en compte
  - Il serait – bien sûr – préférable de disposer d'un cadre théorique complet
    - Data mining sur échantillon (non i).i.d.
- **Les modèles sont systématiquement meilleurs**
  - L'information apportée par les variables sociales est toujours très significative
    - Ce sont même quelques fois les variables les plus significatives

15

Les difficultés

- **Les données d'interaction sont massives**
  - Facebook (>1 Md), Twitter (>500 M), LinkedIn (>200 M)
  - 1 M d'entités → au pire 10^{12} liens
  - Comment tirer un échantillon aléatoire représentatif ?

→ **La scalabilité est cruciale**
- Les algorithmes doivent scaler (linéaire)

- **L'évolution au cours du temps est très rapide**

→ **Les algorithmes doivent s'exécuter rapidement**

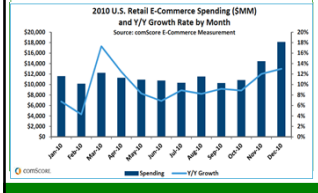
16

Un exemple: la fraude à la carte bancaire

La fraude à la carte bancaire sur Internet

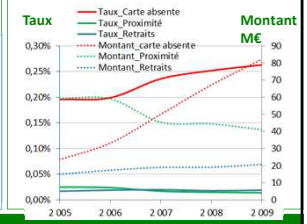
- **Le commerce en ligne augmente partout**
  - Exemple : US

http://www.comscore.com/Press_Events/Presentations_Whitepapers/2011/2010_US_Digital_Year_in_Review



- **Et donc la fraude aussi**
  - En taux & en montant
  - Réalisée par le grand banditisme
  - En France : Carte absente = poste / téléphone / en ligne

http://www.banque-france.fr/observatoire/telechar/gb/2009/rapport-annuel-OSCP-2009-gb-fraud-statistics-for-2008.pdf



18

La fraude à la carte bancaire sur Internet

L'analyse de la fraude a un double objectif couvert par 2 types d'analyse

- Détection : éviter les pertes financières
- Investigation : identifier les gangs responsables

- **Pour cela, on exploite les données disponibles**
  - Données de transactions
  - Données clients & Données produits
  - Données Banques & Données Marchands …
- **Projet ANR eFraudBox**
  - Avec Thales, Altic, GIE CB, LIP6 et LIPN

19

Évaluation des performances de détection

On utilise deux indicateurs

		Prévu		
		P	N	
Réel	P	VP	FN	F
	N	FP	VN	Non F
		A	Non A	

- **Couverture (ou Rappel)**
  - C'est le taux de cas de fraude identifiés
  - On veut peu de Faux Négatifs
    - Ces fraudeurs ne seront pas investigués

$$Couv(s) = {}^{VP}/_{F}$$

- **Pertinence (ou Précision)**
  - C'est le taux d'alertes réellement frauduleuses
  - On veut peu de Faux Positifs
    - Ces dossiers seront investigués pour rien

$$Pert(s) = {}^{VP}/_{A}$$

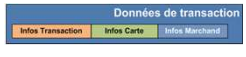
Difficultés

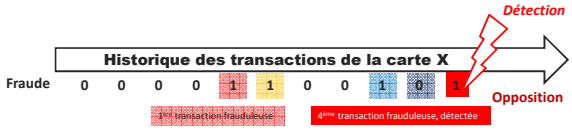
- Le taux de fraude **est très faible**
- Le taux d'alertes **doit être très faible**
- Et les volumétries **sont très fortes**

20

Le processus de détection

- **Données de transactions**
  - Incluant l'information de fraude (si elle est disponible)

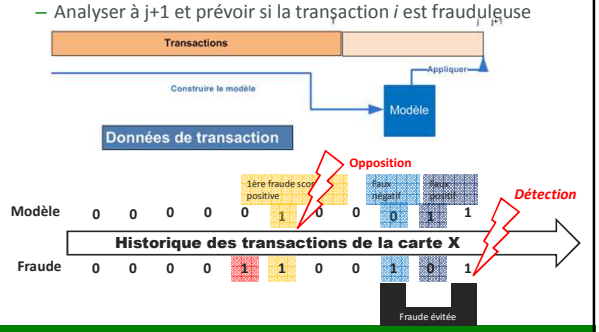




21

Le processus de détection

- **Construire un modèle prédictif**
  - Analyser à j+1 et prévoir si la transaction *i* est frauduleuse



22

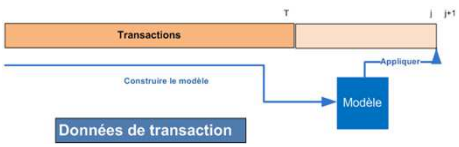
Les difficultés

- **Les volumes sont massifs**
  - Plus de 300 M de transactions par an en France
  - Plus de 40 M de cartes bancaires en France
  - Le commerce électronique est mondial
    - Plusieurs centaines de milliers de marchands dans le monde
- **La fraude change rapidement**
  - Un modèle doit être produit
    - Tous les mois / ans ?
    - Sur un grand volume (1 mois / 1 an de transactions ?)
  - Un modèle doit être appliqué
    - À chaque transaction
- **On a donc des contraintes fortes de temps de calcul**
  - Mais « *More data usually beats better algorithms* » ?

23

Construire un modèle de détection

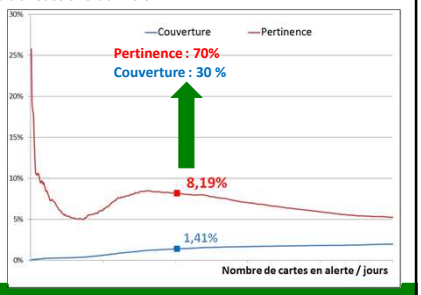
- **Sur un mois (par exemple)**
  - 30 M de transactions
  - 3% de fraude (Fia-Net, 2010)
- **Deux problèmes pour les techniques de data mining**
  - Nombre de transactions
  - Classe Fraude très sous-représentée



24

Modèle baseline

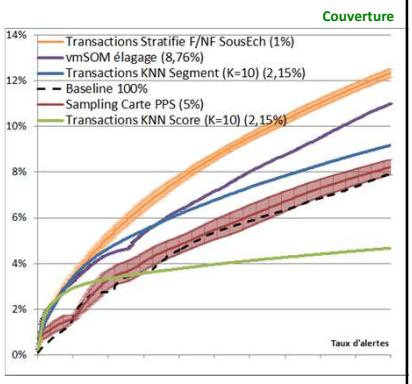
- **Modèle entraîné sur avril et testé en mai**
  - Avec KXEN InfiniteInsight™ 6.0
    - Sur toutes les transactions du mois
    - AUC
- **Loin du but !**



25

Échantillonner

- **Comparaison de méthodes**
  - Stabilité
  - Performance
- **Échantillon stratifié sous-échantillonné**
  - Simple
  - Rapide
  - Performant
  - Stable



26

Création de nouvelles variables (1)

- **Pour améliorer les performances de détection, on génère des variables supplémentaires …**
- **Profils**
  - Carte
  - Marchand
- **Agrégats glissants**
  - Jour, semaine, mois
  - Nombre / montant
    - Transactions, fraudes …
  - Moyenne, taux, déviation

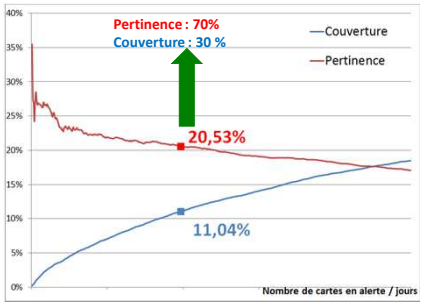
→ **> 700 variables**

Variables	Nombre
Initiales (transactions)	37
Agrégats Carte	304
Agrégats Marchand	370
Score Carte & Agrégats Score Carte	17
Score Marchand & Agrégats Score Marchand	17
Total	745

27

Création de nouvelles variables (1) – Résultats

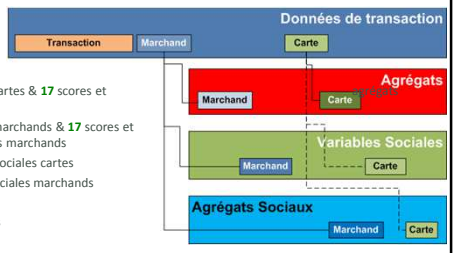
- **Sur un échantillon à 1%**
- **Rappel baseline**
  - **8,19%**
  - **1,41%**
- **Mieux, mais encore loin !**



28

Création de nouvelles variables (2)

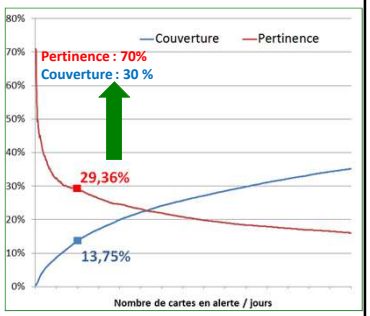
- **Ajouter des variables sociales**
- **On a donc**
  - **304** agrégats cartes & **17** scores et scores cartes
  - **370** agrégats marchands & **17** scores et agrégats scores marchands
  - **140** variables sociales cartes
  - **41** variables sociales marchands
  - **> 800** variables



29

Création de nouvelles variables (2) – Résultats

- **Beaucoup mieux, mais pas encore assez**
  - Pertinence x 3
- **Rappels**
  - Baseline
    - **8,19%**
    - **1,41%**
  - Avec agrégats
    - **10,53%**
    - **11,04%**



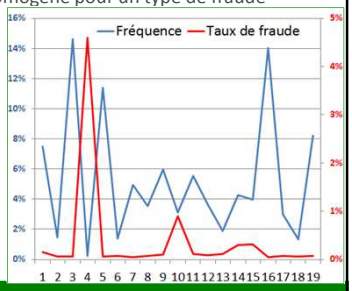
30

Segmentation

- **Il y a beaucoup de types de fraude**
  - Faire une segmentation cartes, avec les agrégats cartes
  - Chaque segment est homogène pour un type de fraude

→ **19 segments**

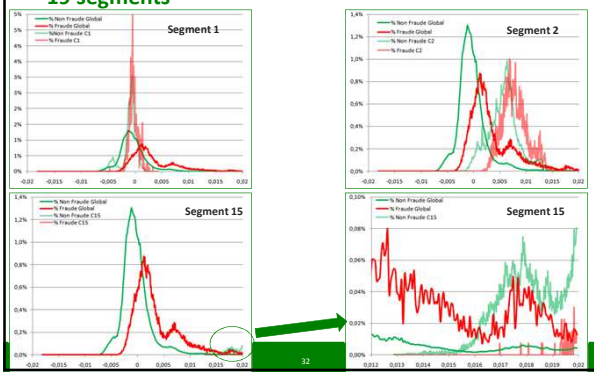
  - Différents types de fraude



31

Segmentation

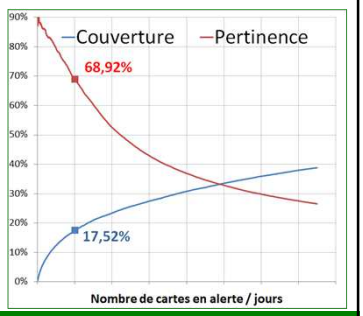
- **19 segments**



32

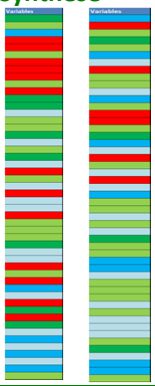
Segmentation

- **Faire un modèle par segment**
- **Rappels**
  - Baseline
    - **8,19%**
    - **1,41%**
  - Avec agrégats
    - **10,53%**
    - **11,04%**
  - Cible
    - **Pertinence : 70%**
    - **Couverture : 30 %**



33

Synthèse



Modèle	Couverture	Pertinence
Baseline	1,40%	8,18%
Baseline + Agg	9,13%	19,00%
Baseline + Agg + Agg Sociaux	9,09%	40,58%
Seg 19	5,09%	28,21%
Seg 19 + Ag.	7,38%	28,82%
Seg 19 + Agg + Agg Sociaux	16,46%	60,89%

- **Modèles**
  - Baseline ou segmenté
- **Importance des variables**
  - **Variables initiales**
  - **Agrégats marchands**
  - **Agrégats cartes**
  - **Variables sociales**
    - **Vert clair** : réseau Cartes
    - **Vert foncé** : réseau marchands

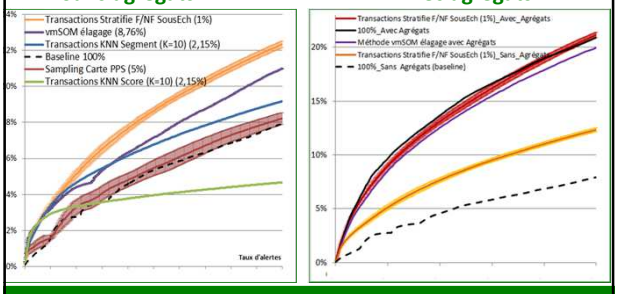
34

Echantillonnage

- **Avec agrégats, mieux vaut éviter d'échantillonner**

Couverture

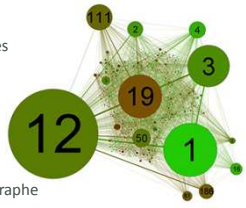
Sans agrégats **Avec agrégats**



35

L'investigation de la fraude

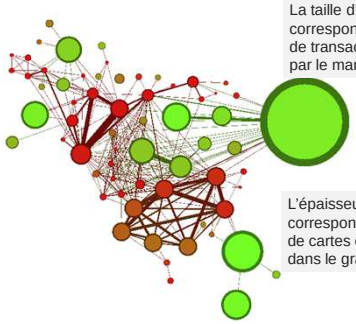
- **On construit le réseau bipartite Cartes-Marchands**
  - À partir du fichier des transactions acceptées d'un mois donné
    - On récupère la liste de toutes les cartes qui ont été **fraudées**
    - Pour chacune de ces cartes on extrait l'intégralité de ses transactions
- **On projette côté Marchands**
  - On obtient à la fois des marchands fraudés et non fraudés
  - Reliés entre eux quand ils ont des cartes en commun
- **On détecte les communautés**
  - Les groupes de marchands plus connectés entre eux qu'avec le reste du graphe
    - Ils sont visités par les mêmes cartes



36

L'investigation de la fraude

- **Zoom sur une communauté marchand**



La taille d'un nœud correspond au nombre de transactions traitées par le marchand

L'épaisseur des liens correspond au nombre de cartes en commun dans le graphe bipartite

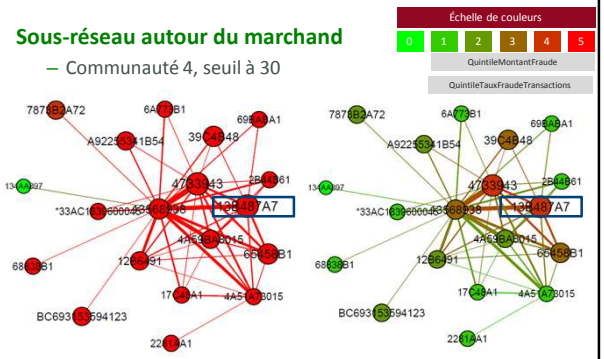
Rouge : fraude
Vert : pas fraude

37

La fraude à la carte bancaire sur Internet

Sous-réseau autour du marchand

– Communauté 4, seuil à 30



38

Questions ?