La statistique : un solide outil pour la détection des risques de fraude

Présentation dans le cadre des conférences de statistique de l'UdeM

Émilyne Lortie - 13 octobre 2023

REVENU QUÉBEC

Table des matières

Études

01 Mes études

Baccalauréat en mathématiques

- Obtenu à l'été 2020
- Orientation statistique
- Cours préférés : MAT1720 (Probabilités), STT1700 (Intro à la stat), STT2700 (Concepts et méthodes en stat), STT3700 (Inférence statistique), STT2105 (Statistique bayésienne), STT3781 (Laboratoire de statistique)

Maitrise en statistique

- Obtenue à l'automne 2021
- Option stage
  - Superviseurs de stage : Mylène Bédard (UdeM) et Benoit Masse (URCA, CHU Ste-Justine)

01 Mes études

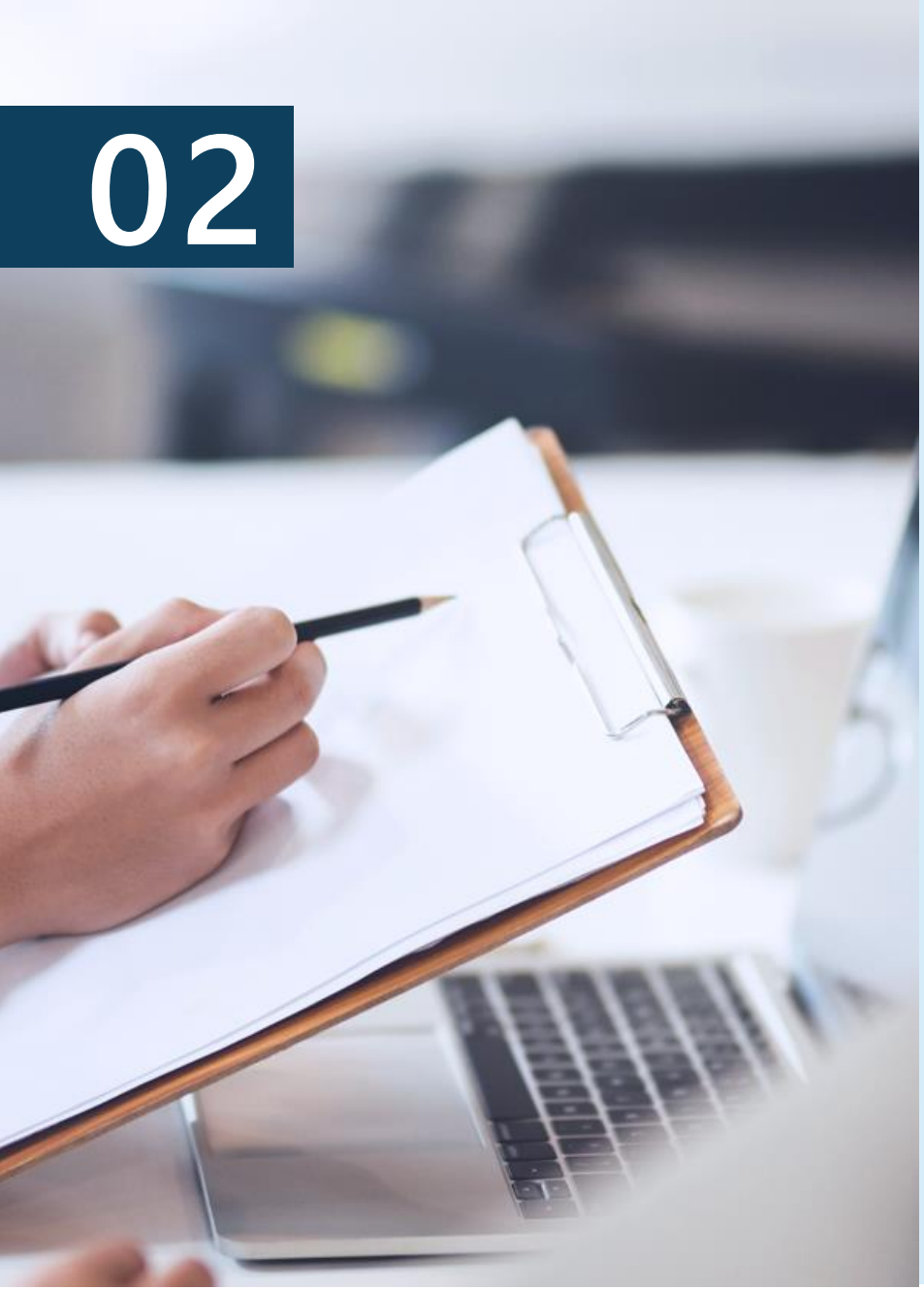
Maitrise en statistique (suite)

- Auxiliaire d'enseignement pour le cours STT1700 à l'automne 2020, hiver 2021 et automne 2021
- Cours préférés : STT6532 et STT6533 (Consultation statistique 1 & 2)
  - Prix et bourse ASSQ UdeM 2021

> Association des statisticiens et statisticiennes du Québec (ASSQ)
> Vous pouvez être membre gratuitement de l'ASSQ pendant vos études.
> Titre « stat. ASSQ », conférences, 5@7, bulletin Convergence, et plus encore.

- Club de lecture en statistique

02



Stage en biostatistique à l'URCA du CHU Sainte-Justine

02 Stage en biostatistique

- **Durée de 4 mois (mai 2021 à août 2021)**
- **Unité de recherche clinique appliquée (URCA) du Centre hospitalier universitaire de Sainte-Justine**
  - Maintenant incorporé dans le Centre de soutien à la gestion et analyse de données (CSGAD)
  - Directeur : Benoit Masse, PhD
  - Biostatisticien(ne)s, programmeurs et programmeuses, chargé(e)s de projet
  - Conception, développement, réalisation, analyse et publication d'études cliniques
- **Projet de stage et participation à des consultations avec des clients (statistiques descriptives)**

URCA Unité de recherche clinique appliquée

CHU Sainte-Justine
Le centre hospitalier universitaire mère-enfant

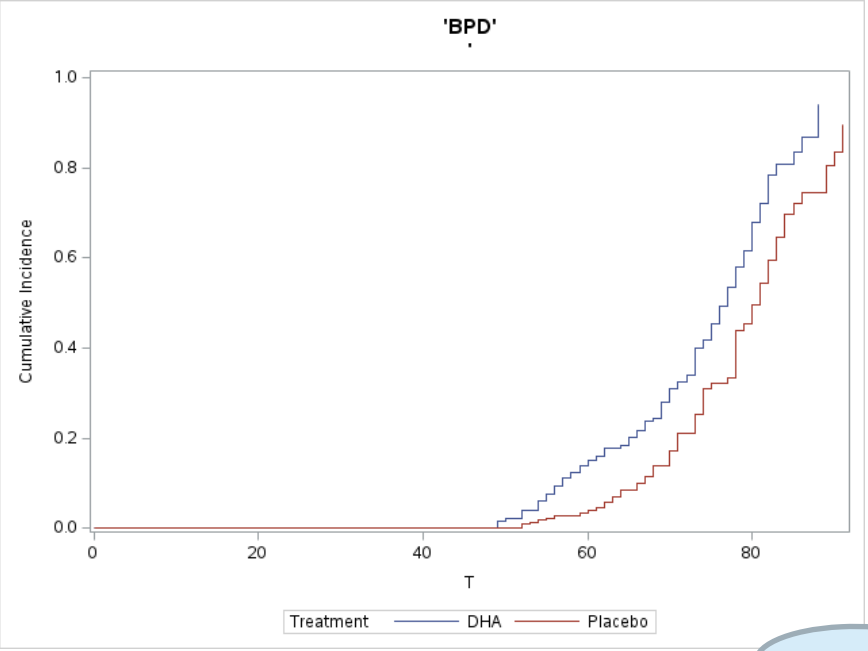
Université de Montréal

02 Projet de stage

- **Réanalyse de l'étude MOBYDIck :** analyse de survie en présence d'un risque compétitif
- **Qu'est-ce que l'étude MOBYDIck?**
  - Essai clinique randomisé (2015 à 2018)
  - Effet d'un supplément d'oméga 3 sur la survie sans dysplasie brochopulmonaire (BPD) chez les nourrissons prématurés allaités par la mère
- **But :** Modéliser le temps avant l'apparition de la BPD chez des nourrissons prématurés en tenant compte du risque de décès et de comparer le groupe traitement au groupe placebo

02 Projet de stage (suite)

- **Fonction d'incidence cumulée**
  - Permet d'estimer adéquatement la probabilité de survenue de la BPD au fil du temps, en tenant compte de la mortalité



- **Test de Gray**
  - Pour comparer les deux traitements (Omega-3 vs placebo) sur la base de leur fonction d'incidence cumulée
  - Vérifie l'hypothèse nulle selon laquelle les fonctions d'incidence cumulée sont identiques pour tous les groupes de traitement

Tableau 4.2. Tests de Gray pour la BPD

Variable de temps	χ^2_1	Valeur-p
T	15,670	<0,001

Une différence significative entre les deux courbes d'incidence cumulée est détectée au niveau 5%

T : Temps de survie sans la BPD

02 Projet de stage (suite)

- **Modèles – Analyse de survie : 2 modèles testés**

Modèle de risques spécifiques à la cause	Modèle de Fine et Gray

  - But : estimer l'effet de l'intervention sur le taux instantané d'occurrence de la maladie
  - Mesure utilisée pour comparer les modèles : Taux de risques (Hazard ratios)

$$HR = \frac{chance\ qu'un\ évènement\ se\ produise\ dans\ le\ groupe\ traitement}{chance\ qu'un\ évènement\ se\ produise\ dans\ le\ groupe\ contrôle}$$

02 Cours utiles

- STT1700 – Introduction à la statistique
- STT2700 – Concepts et méthodes en statistique
- STT3700 – Inférence statistique
- STT3260 – Modèles de survie
- STT3510 – Biostatistique

03

Services aux étudiants – Conseils Carrières

03 Services aux étudiants – Conseils Carrière

- **Services aux étudiants (SAÉ) : accueil et intégration, santé, soutien aux études, …**
- **Conseils Carrière**
  - Pour les étudiants et les diplômés (< 2 ans)
  - Conseillères et conseillers en développement de carrière
  - Planifier une recherche de stage ou d'emploi
  - Parfaire votre CV, lettre de présentation et dossier de candidature
  - Préparer aux entrevues
  - Alerte-emploi

04

Premier emploi

04 Premier emploi

- **Processus assez long : environ 2 mois (octobre 2021 à décembre 2021)**
  - Entrevue et 2 tests écrits
- **Professionnelle en exploitation et en analyse de données (janvier 2022)**
- **Direction générale des entreprises**
  - Direction de l'évaluation du risque et de la gestion de l'information
    - Service de l'évaluation du risque, de la modélisation et de l'échantillonnage statistique
- **Nouvelle équipe en détection des risques de fraude**

Inscriptions sans réelle activité commerciale	Fausse facturation	Usurpation d'identité
Cryptomonnaie	Commerce électronique	Autres risques émergents

04 Premier emploi

- **Mes tâches :**
  - Modélisation (Détection d'anomalie, modèles prédictifs, segmentation, …)
  - Extraction de données dans l'Environnement informationnel de Revenu Québec
  - Développement de fiches de documentation
  - Recherche et exploration (Lecture d'articles scientifiques, documentation de nouvelles méthodes, systèmes ou algorithmes, …)
  - Comité des bonnes pratiques en programmation (développement de formations, développement de modules réutilisables SAS, …)
  - Accompagnement des nouveaux employés
  - Formation personnelle (DataCamp, Formations SAS, Guilde d'IA et Valorisation des données,…)
- **Langages de programmation utilisés :** PL/SQL, SAS, R et Python

05

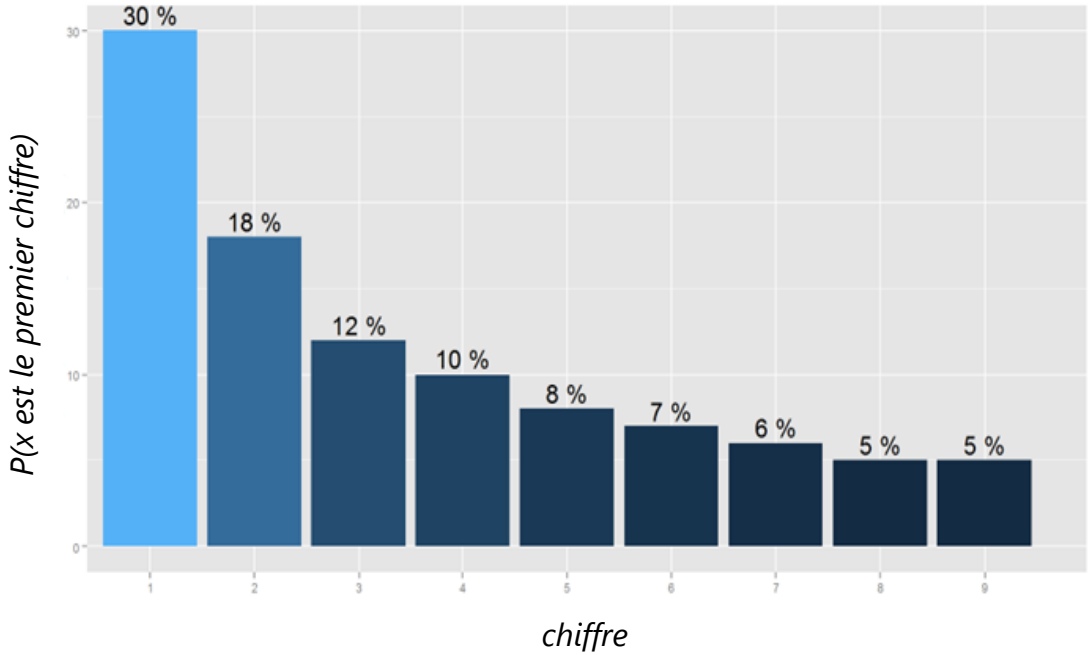
Projet Loi de Benford

05 Projet Loi de Benford

- **Loi de Benford**
  - Loi de probabilités discrètes
  - Concerne le ième chiffre d'un nombre
    - 1er chiffre d'un nombre : $\{1, \dots, 9\}$
    - 2e chiffre d'un nombre : $\{0,1, \dots, 9\}$
    - 3e chiffre d'un nombre : $\{0,1, \dots, 9\}$
    - etc !

$$\boldsymbol{P(X = i) = log_{10}\left(i + \frac{1}{i}\right), \forall\, i \in \{1, \dots, 9\}}$$

chiffre	0	1	2	3	4	5	6	7	8	9
1er	NC	30,1%	17,6%	12,5%	9,7%	7,9%	6,7%	5,8%	5,1%	4,6%
2e	12,0%	11,4%	10,9%	10,4%	10,0%	9,7%	9,3%	9,0%	8,8%	8,5%
3e	10,2%	10,1%	10,1%	10,1%	10,0%	10,0%	9,9%	9,9%	9,9%	9,8%



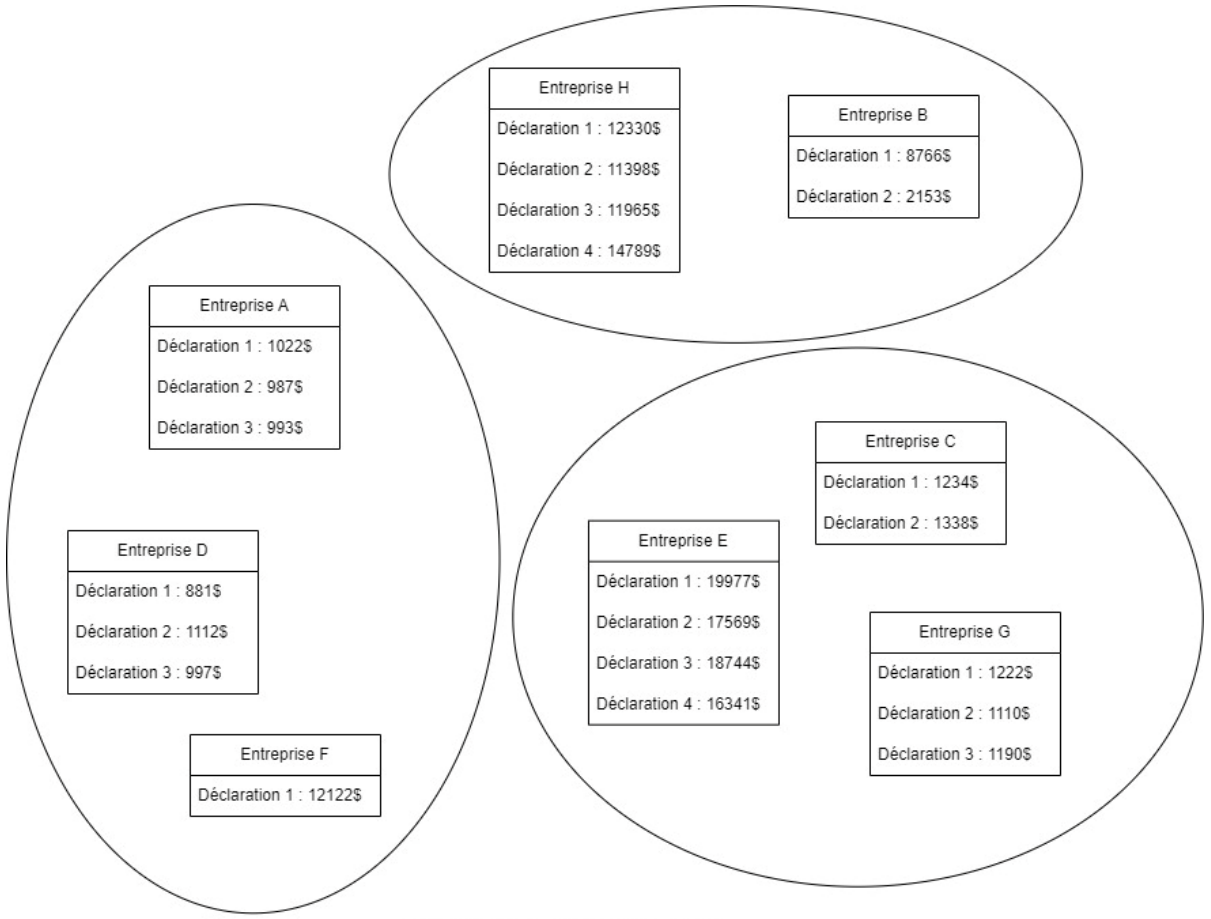
05 Projet Loi de Benford

- **Qu'est-ce que la loi de Benford ?**
  - C'est la loi des nombres anormaux
  - Les chiffres qui constituent certaines catégories de nombres apparaissent dans la nature avec des fréquences bien précises

Exemples	Contre-exemples
Indicateurs boursiers, données démographiques, …	Tirage de loto, # de téléphones, …

  - La loi est invariante de l'échelle du nombre
    - Ex. changement de devise
- **Applications**
  - Détection de la fraude fiscale et comptable
  - Détection de la fraude électorale
  - Détection de la fraude scientifique

05 Projet Loi de Benford



Les données de ce diagramme sont factices

- **But :** identifier des entreprises québécoises qui déclarent des montants fabriqués (déclarations en taxe)
  - Vérifier l'adéquation des montants déclarés par les entreprises à la loi de Benford
- **Comment ?**
  - Segmentation stratifiée pour découper les entreprises
  - Au travers d'un segment, on calcule les proportions observées

$F1D$	$P_{observée}$
1	0,293
2	0,177
3	0,131
4	0,097
5	0,080
6	0,068
7	0,060
8	0,049
9	0,045

$F2D$	$P_{observée}$
0	0,122
1	0,113
2	0,110
3	0,103
4	0,099
5	0,094
6	0,092
7	0,091
8	0,088
9	0,088

$F2D$	$P_{observée}$
0	0,101
1	0,103
2	0,102
3	0,102
4	0,100
5	0,098
6	0,097
7	0,098
8	0,099
9	0,099

$L2D$	$P_{observée}$
00	0,009
01	0,009
02	0,010
...	
98	0,011
99	0,010

Les données de ces tableaux sont factices

05 Projet Loi de Benford

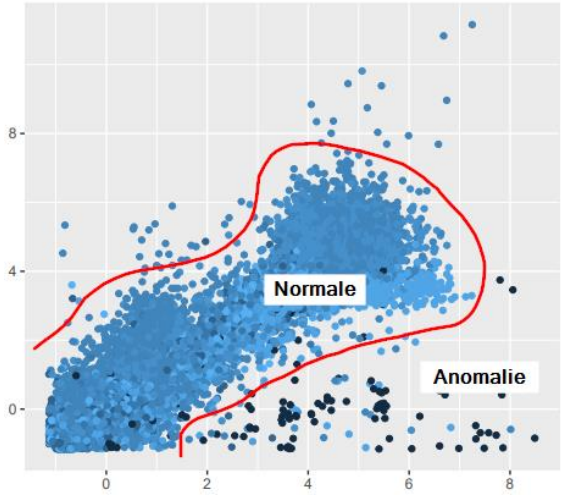
- **Cote finale pour l'entreprise j, qui englobe toutes ses n_j déclarations**

$$Cote_j = \sum_{i=1}^{n_j} cote_{ij} \quad \text{où } n_j \text{ est le nombre de périodes de l'entreprise } j$$

  - Basée sur une addition de χ^2 qui nous proviennent de tests de proportions et tests du χ^2
  - Degrés de liberté dépend du nombre de périodes de déclarations n_j
  - Approche paramétrique : toutes les statistiques menaient à des p-valeurs très significatives!

- **Détection d'anomalies**
  - Forêt d'isolation ⇒ score entre 0 et 1
  - Percentiles

- **Langages de programmation utilisés:**
  - SAS et PL/SQL
  - R



05 Cours utiles

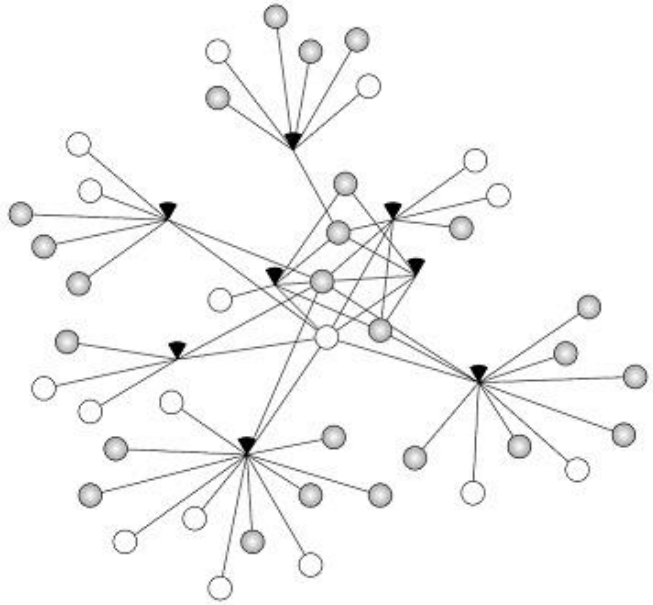
- STT1700 – Introduction à la statistique
- STT2700 – Concepts et méthodes en statistique
- STT3700 – Inférence statistique
- STT3790 – Apprentissage statistique

06

Projet GOTHCA!

06 Projet GOTCHA!

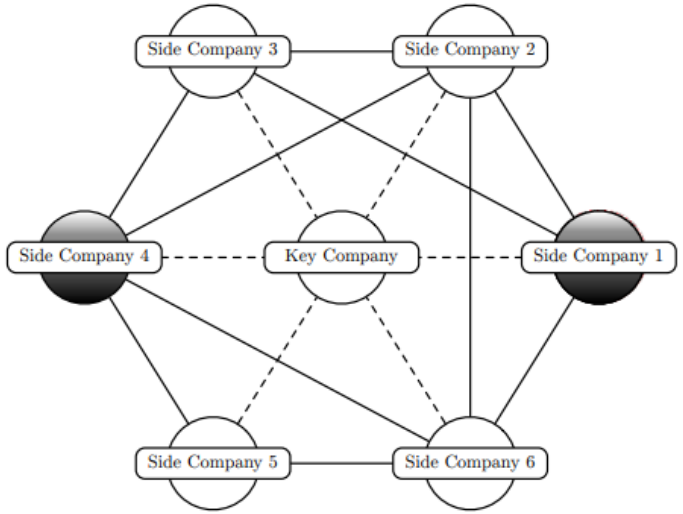
- **Approche**
  - Inspirée de travaux réalisés par l'administration fiscale belge
  - Basée sur l'algorithme PageRank de Google
  - Propagation du risque de non-conformité fiscale via les ressources partagées entre les entreprises

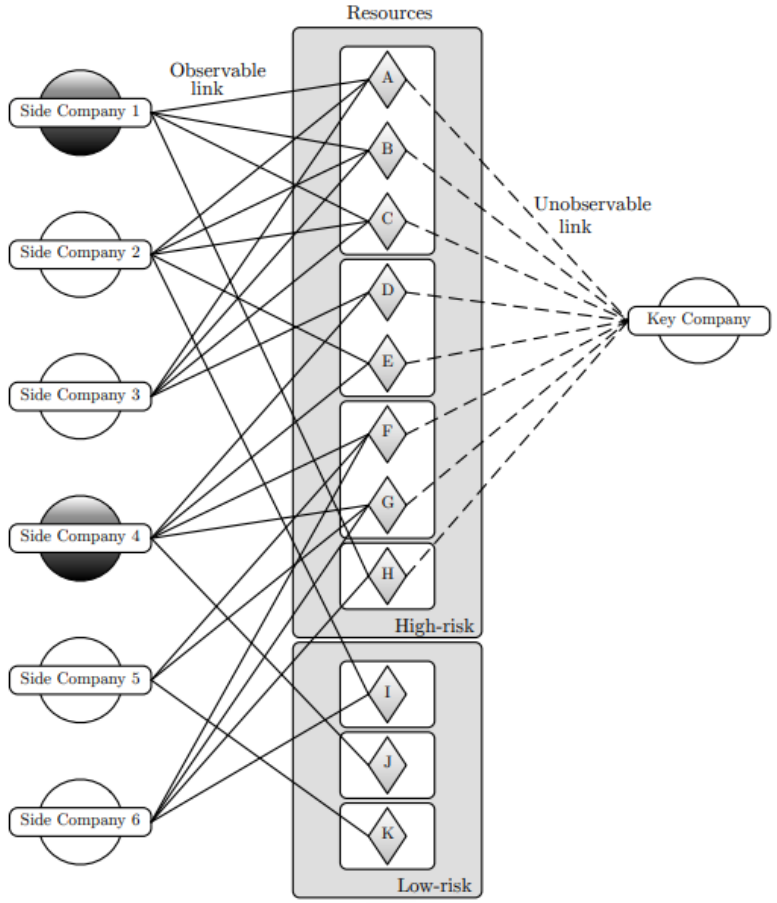


GOTCHA! Network-based Fraud Detection for Social Security Fraud

Dr. Véronique Van Vlasselaer
Department of Decision Sciences and Information Management, KU Leuven, Leuven, Belgium,
Veronique.VanVlasselaer@kuleuven.be

06 Projet GOTCHA!





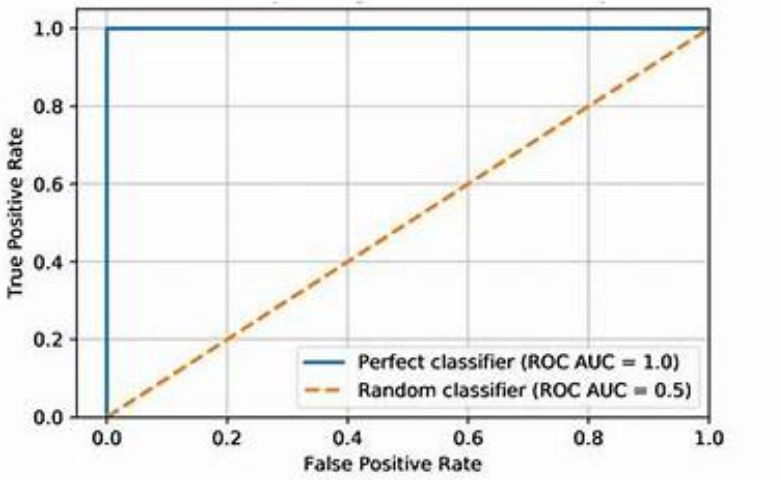
06 Projet GOTCHA!

- **Projet inter-DG à RQ**
  - Centre de recherche de Revenu Québec (Direction générale de l'innovation et de l'administration)
  - Direction de l'évaluation du risque et de gestion de l'information (Direction générale des entreprises)

- **Collaboration avec l'ARPI (atelier de résolution de problèmes industriels), une initiative du CRM et d'IVADO**
  - Solution pour les attributs (variables) basés sur les cliques
    - Aller au-delà des quadrangles
  - Guilt-by-Constellation: Fraud Detection by Suspicious Clique Memberships

06 Projet GOTCHA!

- **Modélisation statistique**
  - Division *stratifiée* en population d'entrainement et de validation
  - Problème de balancement des données : peu de cas positifs connus
    - SMOTE/SMOTENC
  - Modèles : forêt aléatoire et régression logistique
    - Problème de sur-apprentissage



- **Langages de programmation utilisés**
  - SAS et PL/SQL
  - R et Julia
  - R et Python

06 Cours utiles

- STT3790 – Apprentissage statistique
- STT3795 – Fondements théoriques en science des données
- STT3260 – Modèles de survie

Questions?