CRIM 30 ans
Votre longueur d'avance

Détection de menaces par classification de paquets de pare-feu en R

PRÉSENTÉ À R QUÉBEC
PAR LE CRIM
LE 26 MAI 2017

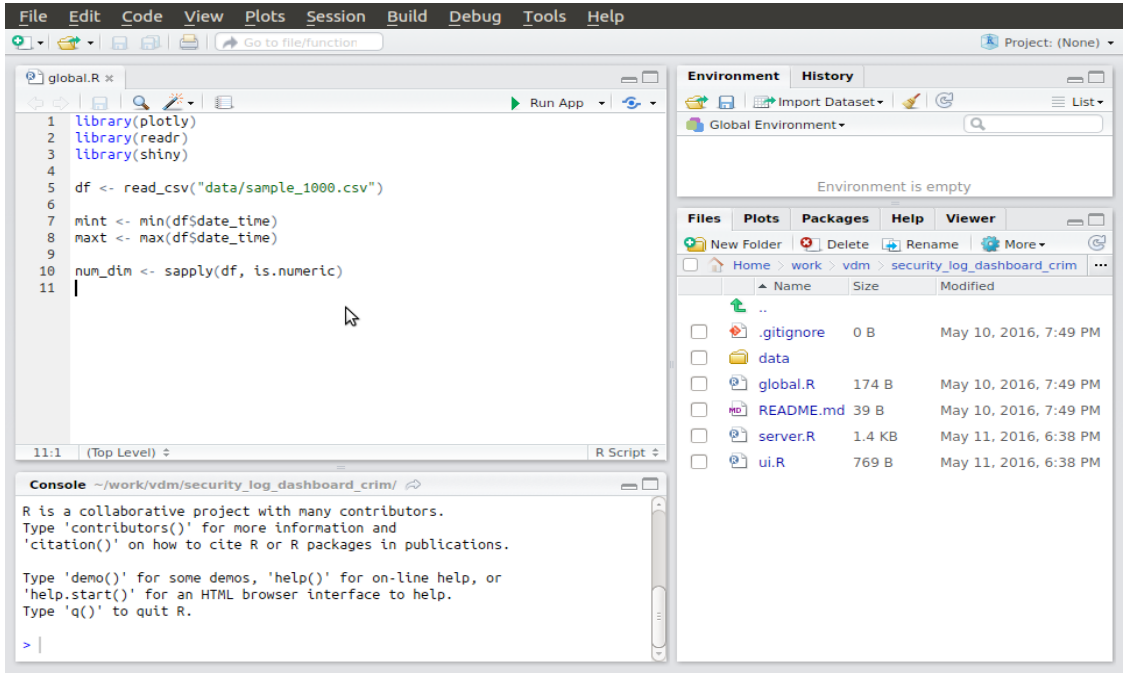
WWW.CRIM.CA

Principal partenaire financier
Économie, Science et Innovation Québec

TABLE DES MATIÈRES

RStudio

- **RStudio**
  - IDE, en local ou via un navigateur web avec un serveur R
  - R est devenu un langage de programmation généraliste
    - Prise en main facile pour des analyses statistiques
    - Création de visualisation de données de haute qualité
    - Applications web



Analyse exploratoire de données (Ex. AlienVault BD)

- **Lecture de données simplifié**

```
read.table()
read.csv()
read.delim()
download.file()
```

```
# URL for the AlienVault IP Reputation Database (OSSIM format)
# storing the URL in a variable makes it easier to modify later
# if it changes. NOTE: we are using a specific version of the data


avURL <-
  "http://reputation.alienvault.com/reputation.data"

# use relative path for the downloaded data
avRep <- "data/reputation.data"

# using an if{}-wrapped test with download.file() vs read.xxx()
# directly avoids having to re-download a 16MB file every time
# we run the script
if (file.access(avRep)) {
  download.file(avURL, avRep)
}
## trying URL "http://reputation.alienvault.com/reputation.data"
## Content type 'application/octet-stream' length 17668227 bytes
## opened URL
## ==================================================
## downloaded 16.8 Mb
```

  - La liste « AlienVault »
    - @IP s à risque

Analyse exploratoire de données (Ex. AlienVault BD)

- **Un premier regard sur les données**
  - – Assigner des noms de colonne significatifs
  - – Utilisation de fonctions intégrées pour avoir un aperçu sur la structure des données
  - – Avoir une vue sur les premières lignes, exp: `head()`
  - –Constater
    - R a déduit correctement que IP, Type, Country, et Locale sont de type catégorique ☺
    - R n'a pas reconnu que "Reliability" et "Risk" sont qualitatives, ☹

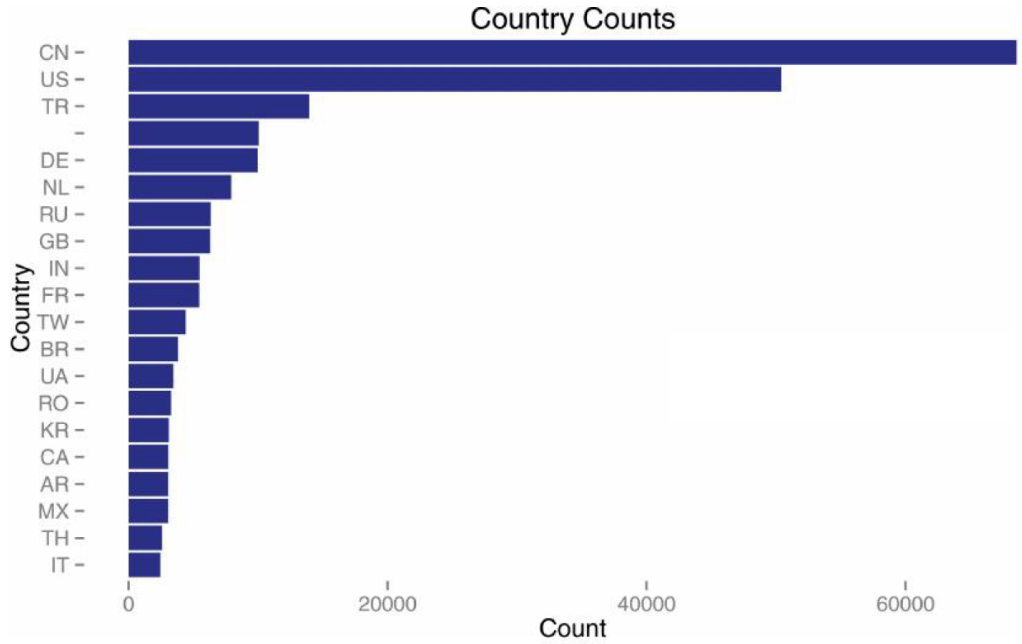
```
# read in the IP reputation db into a data frame
# this data file has no header, so set header=FALSE
av <- read.csv(avRep,sep="#", header=FALSE)

# assign more readable column names since we didn't pick
# any up from the header
colnames(av) <- c("IP", "Reliability", "Risk", "Type",
                  "Country", "Locale", "Coords", "x")
```

```
str(av) # get an overview of the data frame
## 'data.frame': 258626 obs. of  8 variables:
## $ IP : Factor w/ 258626 levels "1.0.232.167",..: 154069 154065
##    154066 171110 64223 197880 154052 154051 154050 56741 ...
## $ Reliability: int 4 4 4 6 4 4 4 4 4 6 ...
## $ Risk : int 2 2 2 3 5 2 2 2 2 3 ...
## $ Type : Factor w/ 34 levels "APT;Malware Domain",..: 25 25 25 31 25
##    25 25 25 25 31 ...
## $ Country : Factor w/ 153 levels "","A1","A2","AE",..: 34 34 34 143
##    141 143 34 34 34 1 ...
## $ Locale : Factor w/ 2573 levels "","Aachen","Aarhus",..: 2506 2506
##    2506 1 1374 2342 2506 2506 2506 1 ...
## $ Coords : Factor w/ 3140 levels "-0.139500007033,98.1859970093",..:
##    489 489 489 1426 2676 1384 489 489 489 489 ...
## $ x : Factor w/ 34 levels "11","11;12","11;2",..: 1 1 1 7 1 1 1 1 1
##    7 ...
```

```
head(av) # take a quick look at the first few rows of data
##               IP Reliability Risk          Type Country     Locale
## 1 222.76.212.189           4    2 Scanning Host      CN     Xiamen
## 2 222.76.212.185           4    2 Scanning Host      CN     Xiamen
## 3 222.76.212.186           4    2 Scanning Host      CN     Xiamen
## 4    5.34.246.67           6    3      Spamming      US
## 5  178.94.97.176           4    5 Scanning Host      UA     Merefa
## 6    66.2.49.232           4    2 Scanning Host      US Union City
##                       Coords  x
## 1   24.4797992706,118.08190155 11
## 2   24.4797992706,118.08190155 11
## 3   24.4797992706,118.08190155 11
## 4                 38.0,-97.0 12
## 5  49.8230018616,36.0507011414 11
## 6 37.5962982178,-122.065696716 11
```

Analyse exploratoire de données (Ex. AlienVault BD)

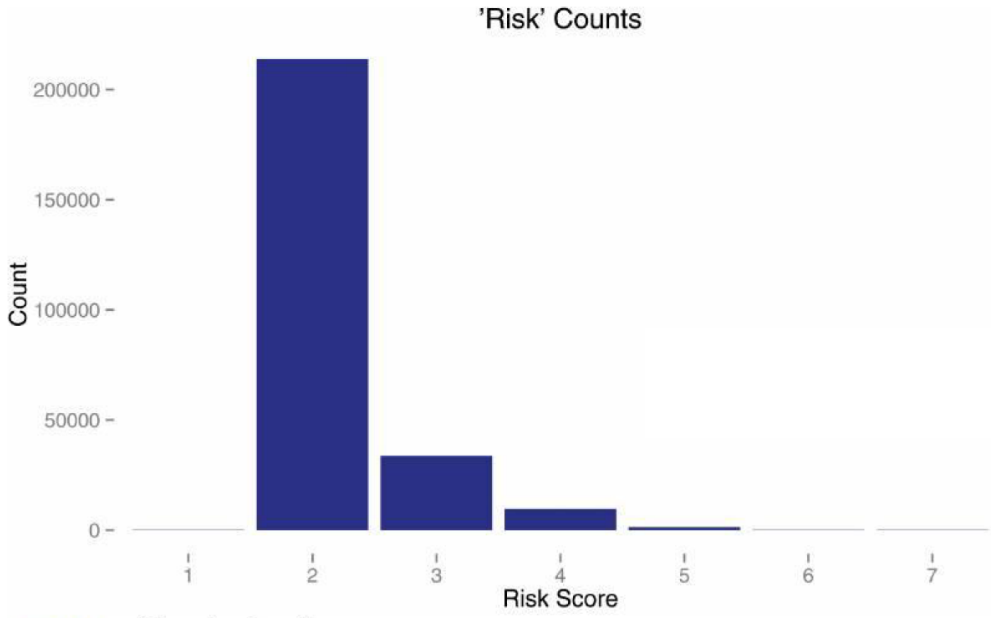
- **Un regard plus approfondi sur les données**



- **Chine et USA ensemble comptent pour 46% des nœuds malveillants de la liste, la Russie compte seulement pour 2.4%**
- **3% des @ips ne peuvent pas être géo-localisés**

Analyse exploratoire de données (AlienVault BD)

- **Un regard plus raffiné sur les données :**
  - **La variable ‘Risk’**

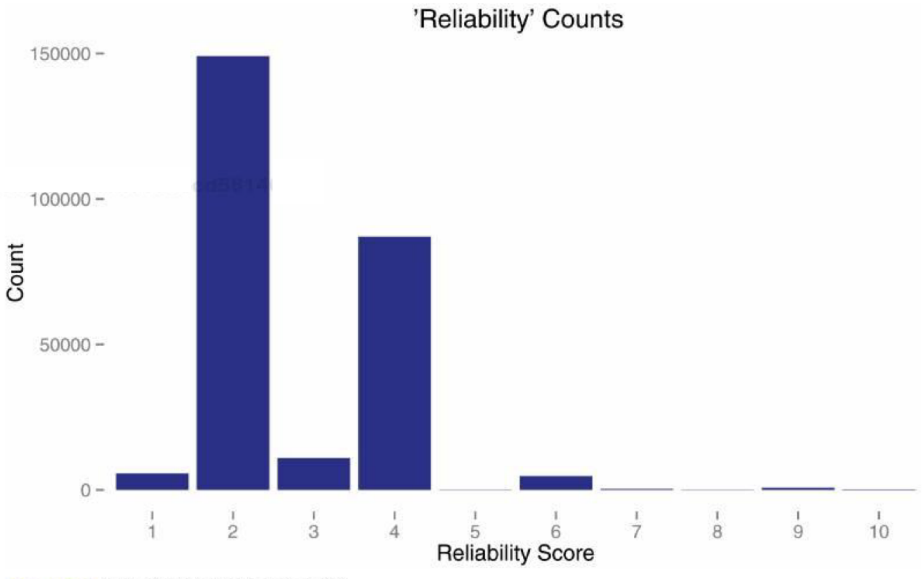


➢Le niveau de risque de la majorité des nœuds est négligeable,

➢ Pratiquement, il n’y a pas de IP catégorisés 1, 5, 6, ou 7 , et aucune adresse n’est présente dans la plage [8 à 10]

Analyse exploratoire de données (AlienVault BD)

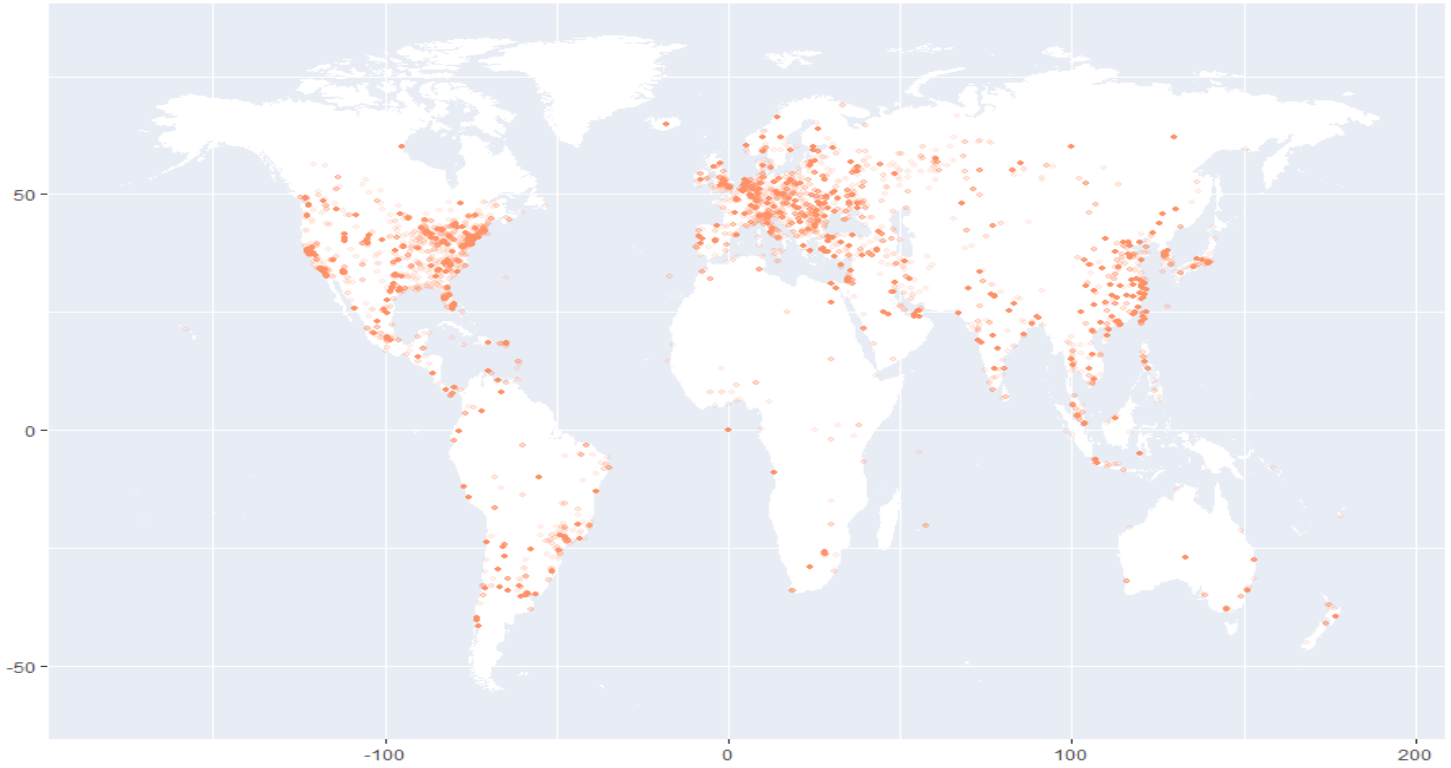
- **Un regard plus raffiné sur les données :**
  - **La variable ‘Reliability’**



- **Quantifie la précision de la catégorie du ‘Risk’ de chaque nœuds,**
  - ➢Les valeurs sont généralement de niveau 2 et 4
  - ➢Entre les deux, une précision de niveau 3 est insignifiante

Analyse exploratoire de données

- **Mapping des nœuds malveillants vers leurs sites physiques**



Package 'ggplot2'

December 30, 2016

Version 2.2.1

Title Create Elegant Data Visualisations Using the Grammar of Graphics

Description A system for 'declaratively' creating graphics, based on "The Grammar of Graphics". You provide the data, tell 'ggplot2' how to map variables to aesthetics, what graphical primitives to use, and it takes care of the details.

Depends R (>= 3.1)

Imports digest, grid, gtable (>= 0.1.1), MASS, plyr (>= 1.7.1), reshape2, scales (>= 0.4.1), stats, tibble, lazyeval

Suggests covr, ggplot2movies, hexbin, Hmisc, lattice, mapproj, maps, maptools, mgcv, multcomp, nlme, testthat (>= 0.11.0), quantreg, knitr, rpart, rmarkdown, svglite

Enhances sp

License GPL-2 | file LICENSE

Analyse exploratoire de données

- **Mapping des nœuds malveillants vers leurs sites physiques**

Package 'maps'

August 29, 2016

Title Draw Geographical Maps
Version 3.1.1
Date 2016-07-19
Author Original S code by Richard A. Becker and Allan R. Wilks. R version by Ray Brownrigg. Enhancements by Thomas P Minka and Alex Deckmyn.
Description Display of maps. Projection code and larger maps are in separate packages ('mapproj' and 'mapdata').
Depends R (>= 2.14.0)
Imports graphics, utils
LazyData yes
Suggests mapproj (>= 1.2-0), mapdata (>= 2.2-4), sp, maptools
License GPL-2

Analyse exploratoire de données : Utilitaires R

- **Utilitaires de manipulations d'@IP: 'iptools'**

  - **Ip_to_hostname**

```
ip_to_hostname("162.243.111.4")
[[1]]
[1] "dds.ec"
```

  - **ip_to_numeric**

```
#Convert your local, internal IP to its numeric representation.
ip_to_numeric("192.168.0.1")
#[1] 3232235521

#And back
numeric_to_ip(3232235521)
```

  - **range_boundaries**

```
range_boundaries("172.18.0.0/28")
##   minimum_ip  maximum_ip min_numeric max_numeric         range
## 1 172.18.0.0 172.18.0.15  2886860800  2886860815 172.18.0.0/28
```

Package 'iptools'

April 4, 2016

Type Package

Title Manipulate, Validate and Resolve 'IP' Addresses

Version 0.4.0

Date 2016-04-04

Author Bob Rudis <bob@rudis.net> [aut, cre],
Oliver Keyes <ironholds@gmail.com> [aut],
Tim Smith [ctb]

Maintainer Bob Rudis <bob@rudis.net>

Description A toolkit for manipulating, validating and testing 'IP' addresses and ranges, along with datasets relating to 'IP' addresses. Tools are also provided to map 'IPv4' blocks to country codes. While it primarily has support for the 'IPv4' address space, more extensive 'IPv6' support is intended.

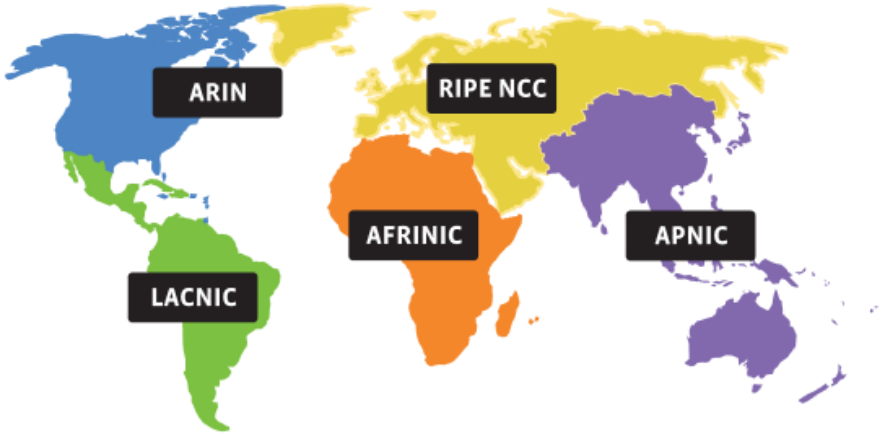
License MIT + file LICENSE

Analyse exploratoire de données : Recoupement de données

- **La structure de données “dataframe” R facilite l’intégration et le recoupement de nouvelles données pour une analyse plus complète:**
  - **Incorporation de nouvelles données**
  - **‘Parsing’ et transformation de nouvelles données en dataframe**
  - **Validation du contenus et de la structure des ces données**
  - **Extraction de l’information pertinente de la nouvelle source de données**
  - **Ajout de nouvelles variables au dataframe existant**
  - **Conduire de nouvelles analyses**

Analyse exploratoire de données : Recoupement

- **Répartition des nœuds malicieux**
  - **Intégration des données de l'IANA (l'organisation responsable de gérer les adresses Internet)**

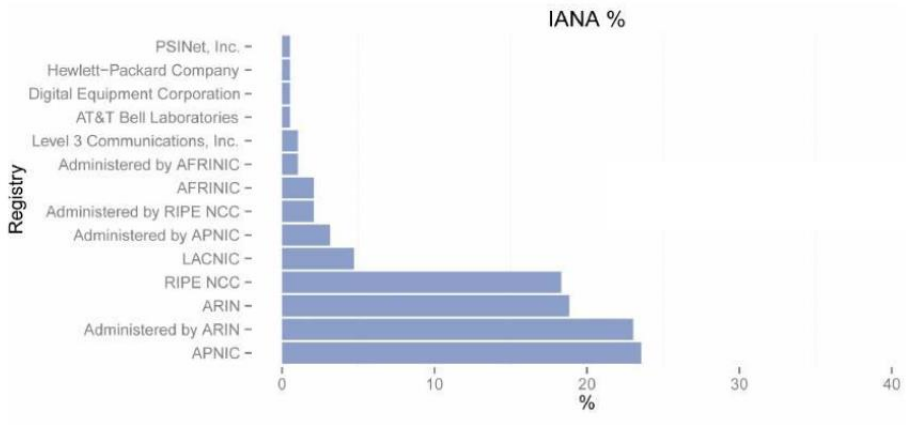


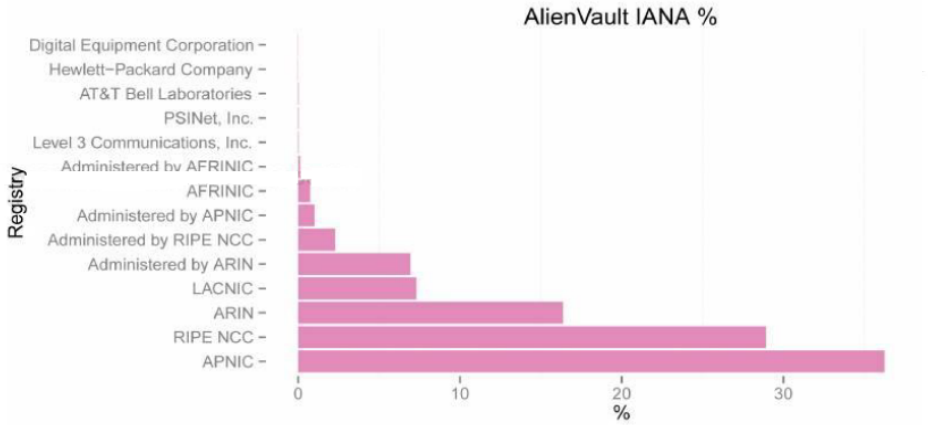
  - **Un niveau plus élevé que le regroupement pas pays**

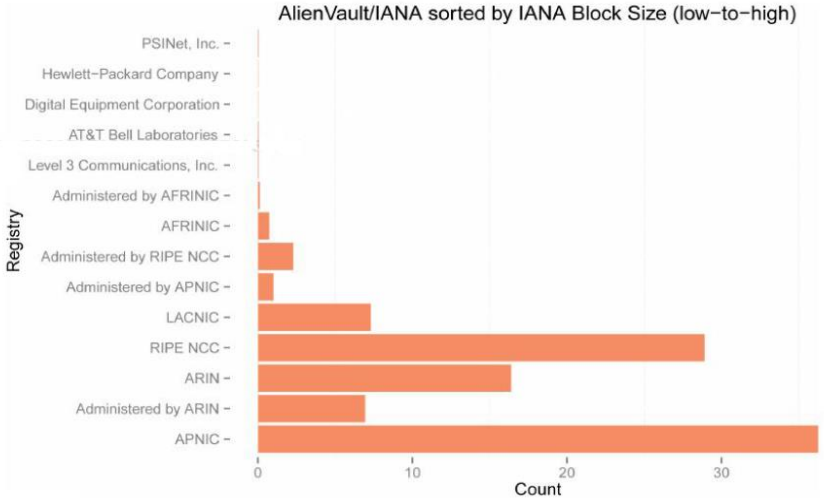
```
# R code to incporporate IANA IPv4 allocations
# retrieve IANA prefix list
ianaURL <- "http://www.iana.org/assignments/ipv4-address-space/ipv4-
address-space.csv"
ianaData <- "data/ipv4-address-space.csv"
if (file.access(ianaData)) {
  download.file(ianaURL, ianaData)
}

# read in the IANA table
iana <- read.csv(ianaData)

# clean up the iana prefix since it uses the old/BSD-
# number formatting (i.e. allows leading zeroes and
# we do not need to know the CIDR component.
iana$Prefix <- sub("^(00|0)", "", iana$Prefix, perl=TRUE)
iana$Prefix <- sub("/8$", "", iana$Prefix, perl=TRUE)

# define function to strip 'n' characters from a string
# (character vector) and return the shortened string.
# note that this function is 'vectorized' (you can pass it a single
# string or a vector of them)
rstrip <- function(x, n){
  substr(x, 1, nchar(x)-n)
}

# extract just the prefix from the AlienVault list
av.IP.prefix <- rstrip(str_extract(as.character(av.df$IP),
                                   "^([0-9]+)\\."), 1)
# there are faster ways than 'sapply()' but we wanted you to
# see the general "apply" pattern in action as you will use it
# quite a bit throughout your work in R
av.df$Designation <- sapply(av.IP.prefix, function(ip) {
  iana[iana$Prefix == ip, ]$Designation
})
```

```
##       Administered by AFRINIC         Administered by APNIC
##                           322                          2615
##          Administered by ARIN      Administered by RIPE NCC
##                         17974                          5893
##                       AFRINIC                         APNIC
##                          1896                         93776
##                          ARIN          AT&T Bell Laboratories
##                         42358                            24
## Digital Equipment Corporation       Hewlett-Packard Company
```

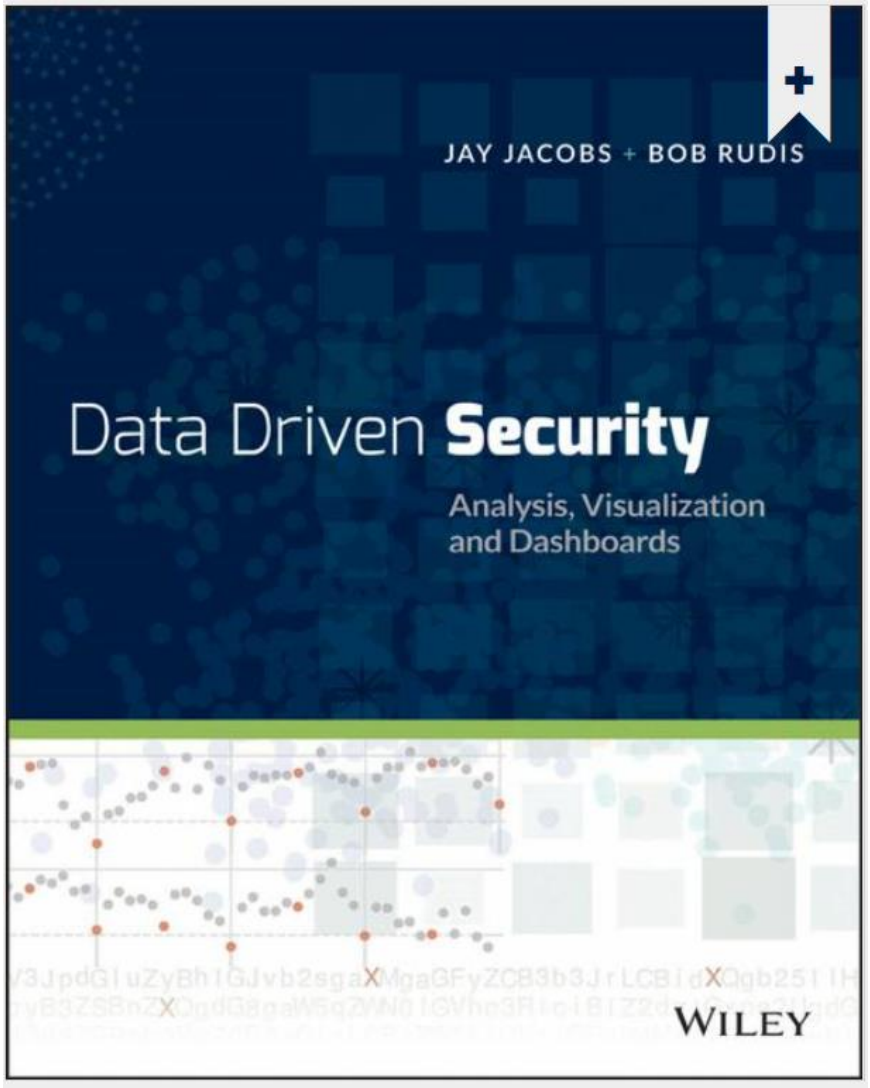
Analyse exploratoire de données : Recoupements







Le Livre :


JAY JACOBS + BOB RUDIS
Data Driven Security
Analysis, Visualization and Dashboards
WILEY

DÉTECTION D'ANOMALIES : Conventionnelle vs Intelligente

- **Approche conventionnelle**
  - **Monitorer les données du Pare –feu (ex. Tableaux de Bord)**

- **Approche intelligente**
  - **Exploration visuelle**
    - **Recoupement (Pare –feu et méta données)**
      - ➢ Liste noire (BD. AlienVault)
        - Charger les @IPs destination
        - Filtrer les @IPs qui ne sont pas dans la liste noire
        - Les nœuds restant avec une précision élevée sont classés à haut risque
      - ➢ Représentations graphiques
  - **Exploration intelligente**
    - **Forage de données**
    - **Apprentissage**

DÉTECTION D'ANOMALIES : Conventionnelle vs Intelligente

- **Pourquoi l'approche intelligente ???**
  - **Disponibilité d'un grand volume de données générées par l'infrastructure cybernétique**
  - **Augmentation du nombre de tentatives criminelles d'accès aux données**
  - **Caractère mutatif des procédés de violation**

- **Comment ?**
  - **Grouper et recouper les données**
  - **Extraire les comportements et les entités par**
    - **Statistiques**
    - **Apprentissage et reconnaissance de motifs**

Pare-feu du CRIM

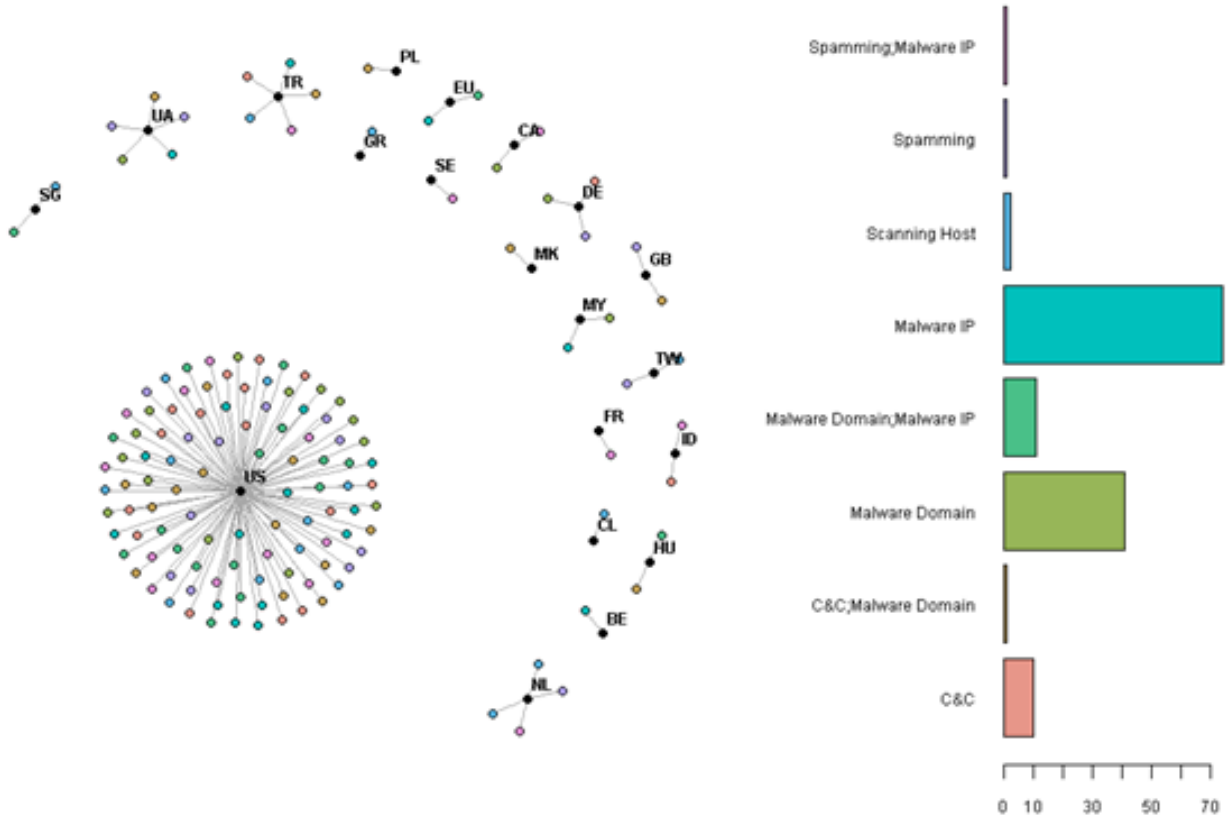
- FORTIGATE a 114 variables

Filter

lineID	reason	dstport	sessionid	type	remport	virusid	hostname	icmpcode	action	PuppetAgent	dstip
4516	NA	25	NA	traffic	NA	NA		NA		NA	132.217.11.74
4517	NA	55905	1284062905	traffic	NA	NA		NA		NA	132.217.151.12
4518	NA	53	1284062925	traffic	NA	NA		NA		NA	10.20.10.110
4519	NA	NA	1284121711	traffic	NA	NA		NA		NA	10.30.10.22
4520	NA	55905	1284062906	traffic	NA	NA		NA		NA	132.217.151.12
4521	NA	53	1284062922	traffic	NA	NA		NA		NA	10.20.10.110
4522	NA	NA	1284121710	traffic	NA	NA		NA		NA	10.1.1.99
4523	NA	80	NA	traffic	NA	NA		NA		NA	132.217.123.41
4524	NA	5900	1284151833	traffic	NA	NA		NA		NA	132.217.151.131
4525	NA	55905	1284151845	utm	NA	NA		NA	pass	NA	132.217.151.12
4526	NA	81	NA	traffic	NA	NA		NA		NA	132.217.124.113
4527	NA	53	1284062961	traffic	NA	NA		NA		NA	10.20.10.110
4528	NA	53	1283870343	traffic	NA	NA		NA		NA	10.20.10.110
4529	NA	53	1284047715	traffic	NA	NA		NA		NA	10.20.10.110
4530	NA	53	1284062956	traffic	NA	NA		NA		NA	10.20.10.110
4531	NA	443	1284117216	traffic	NA	NA		NA		NA	10.20.30.20
4532	NA	55905	1283991840	traffic	NA	NA		NA		NA	132.217.151.12
4533	NA	28392	1284062897	traffic	NA	NA		NA		NA	89.211.137.72
4534	NA	53	1284062986	traffic	NA	NA		NA		NA	10.20.10.110
4535	NA	28392	1284062896	traffic	NA	NA		NA		NA	89.211.137.72
4536	NA	55905	1284151874	utm	NA	NA		NA	pass	NA	132.217.151.12
4537	NA	55905	1284062984	traffic	NA	NA		NA		NA	132.217.151.12
4538	NA	53	1284063004	traffic	NA	NA		NA		NA	10.20.10.110
4539	NA	1000	1284062027	traffic	NA	NA		NA		NA	192.0.1.0
4540	NA	53	1283969881	traffic	NA	NA		NA		NA	10.20.10.110
4541	NA	55905	1284062985	traffic	NA	NA		NA		NA	132.217.151.12
4542	NA	53	1284003557	traffic	NA	NA		NA		NA	10.20.10.110
4543	NA	53	1284028651	traffic	NA	NA		NA		NA	10.20.10.110
4544	NA	55905	1284062999	traffic	NA	NA		NA		NA	132.217.151.12

DÉTECTION D'ANOMALIES : A - Exploration visuelle

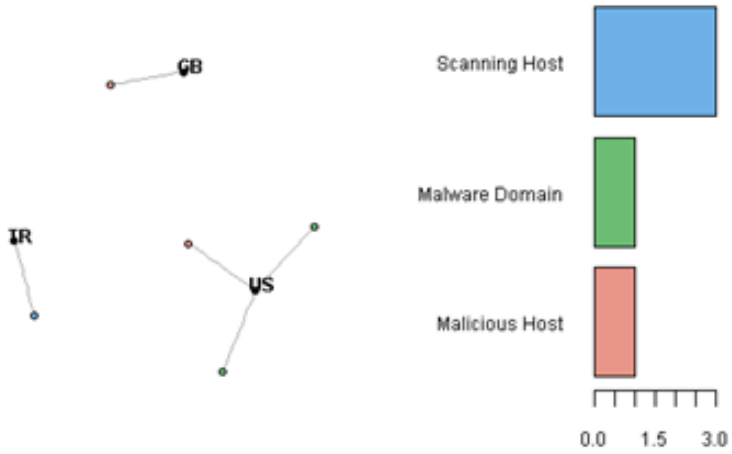
- **Visualizing Firewall data**



Graph of malicious destination traffic by country on a database of 24H of logs from volunteers (Reliability >=6)

DÉTECTION D'ANOMALIES : A - Exploration visuelle

- **Visualizing Firewall data**



Graph of malicious destination traffic by country on CRIM logs (Reliability>=2)

DÉTECTION D'ANOMALIES : A - Exploration visuelle

Visualisation interactive de données de Pare-Feu: « l'explorateur de menaces »

DÉTECTION D'ANOMALIES - B : Exploration du flux des logs

- Sélection de variables (12)

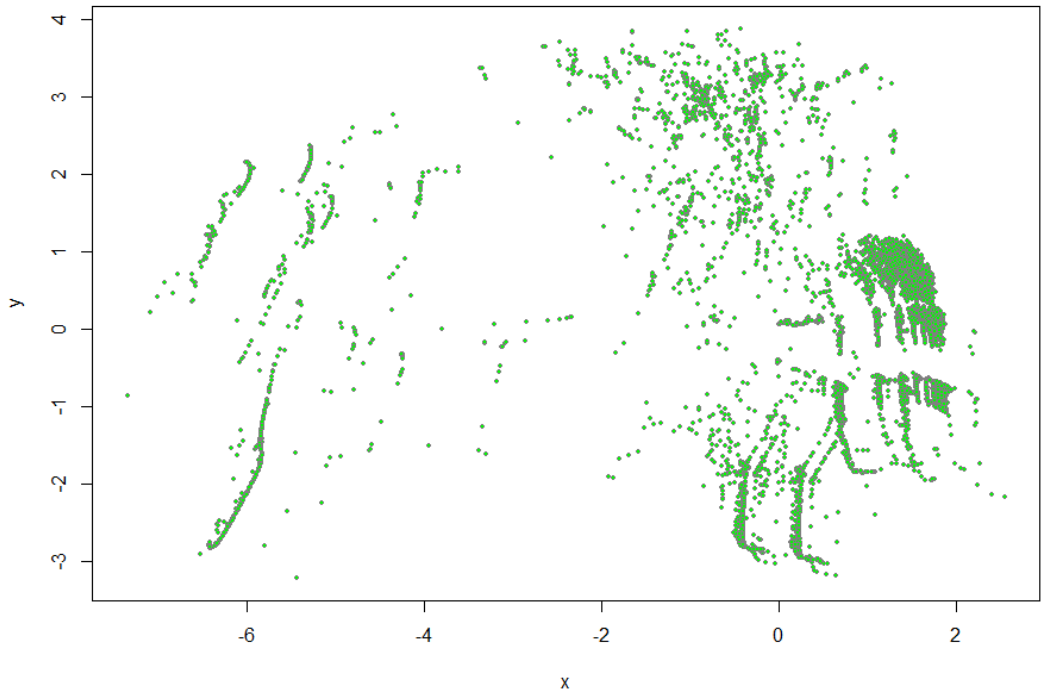
```
names
[1] "dstip"          "srcip"          "sentbyte"       "sentpkt"        "duration"       "dstport"
[7] "proto"          "srcport"        "service"        "sentbyteBypkt"  "count_dest_conn" "count_src_conn"
```

Basic features
Source IP
Source Port
Destination IP
Destination Port
Protocol
Duration
Packets Sent
Bytes per packet Sent

Derived-Connection based features	
Count-**dest**-conn	Number of flows to unique **destination IP** addresses inside the network in the last N flows from the same source
Count-**src**-conn	Number of flows from unique **source** IP addresses inside the network in the last N flows to the same destination

DÉTECTION D'ANOMALIES - B : Exploration du flux des logs

- **Positionnement multidimensionnel (Multidimensional Scaling : MDS)**
  - **Un cas d'analyse multivariée**
    - ➢ Exploite les dissimilarités dans les données
    - ➢ Permet une visualisation d'information en 2D (toute en préservant au max les distances dans l'espace originale)



DÉTECTION D'ANOMALIES - B : Exploration du flux des logs

- **Local Outlier Factor (LOF)**
  - **Une technique d'identification d'*outlier* par densité locale**
  - **Compare la densité locale d'un point à celle de son voisinage**

Package 'DMwR'

February 19, 2015

Type Package

Title Functions and data for ``Data Mining with R"

Version 0.4.1

Depends R(>= 2.10), methods, graphics, lattice (>= 0.18-3), grid (>= 2.10.1)

Imports xts (>= 0.6-7), quantmod (>= 0.3-8), zoo (>= 1.6-4), abind (>= 1.1-0), rpart (>= 3.1-46), class (>= 7.3-1), ROCR (>= 1.0)

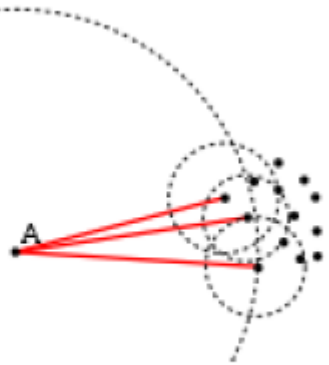
Date 2013-08-08

Author Luis Torgo

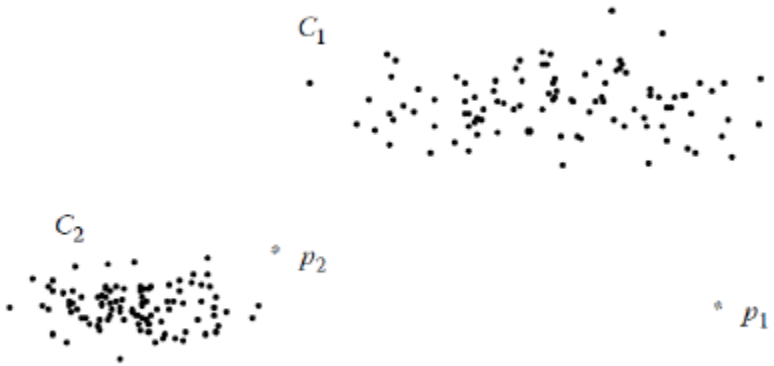
Maintainer Luis Torgo <ltorgo@dcc.fc.up.pt>

Description This package includes functions and data accompanying the book ``Data Mining with R, learning with case studies" by Luis Torgo, CRC Press 2010.

License GPL (>= 2)



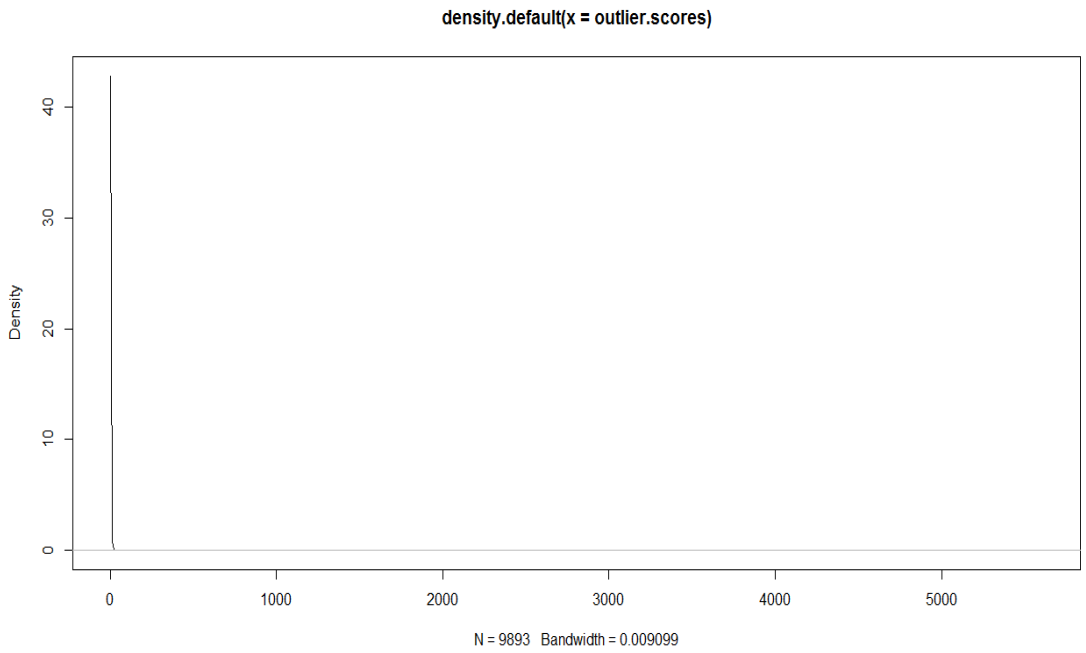
Idée de base de l'approche « Local Outlier Factor »



. Cas difficile pour les méthodes k-NN basées sur la notion de distance [Dua, 2011]

DÉTECTION D'ANOMALIES - B : Exploration du flux des logs

- **Local Outlier Factor (LOF)**
  - **Appliquée aux données du Pare-Feu du CRIM**

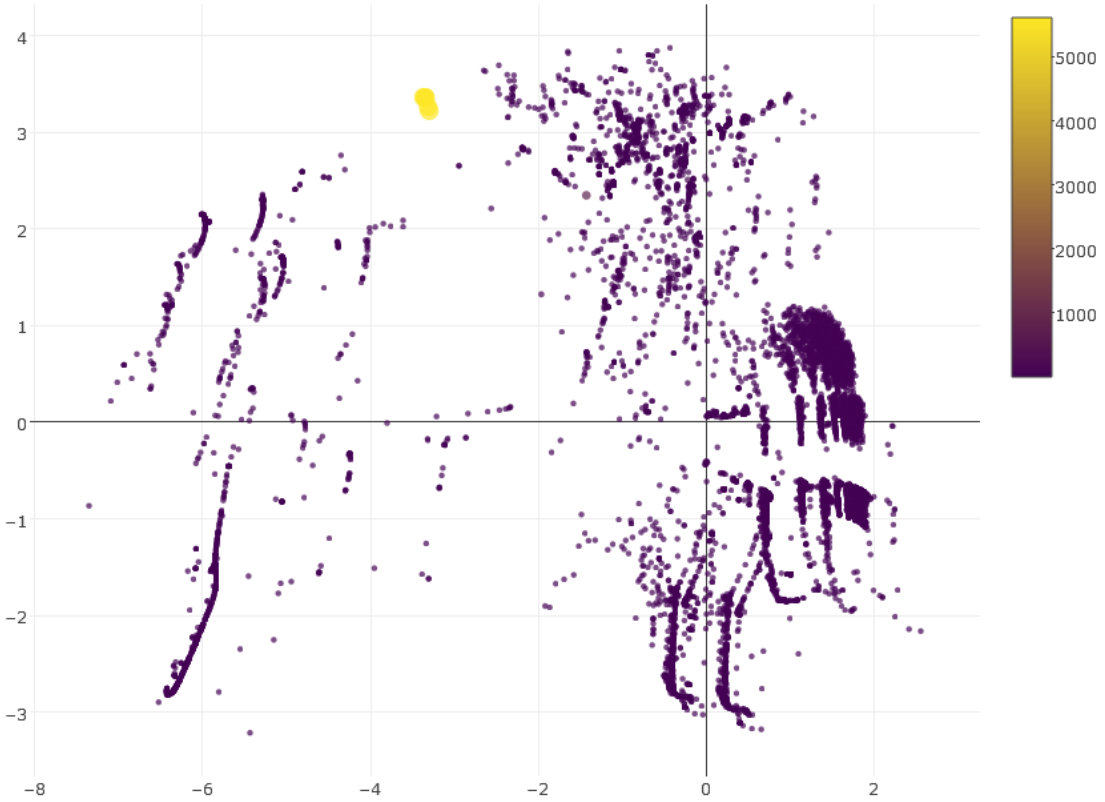


Exemple du Firewall du CRIM
(petite bande passante : 0.009)

id	Score
1	5611.272748
2	5611.272748
3	5607.653066
4	5605.809993
5	5605.743529
6	774.107524
7	164.051679
8	83.850971
9	46.777675
10	23.841999
11	10.047024
12	8.324195
13	7.352677
14	7.301954
15	7.128473
16	6.544991
17	5.846355
18	5.572101
19	5.562847
20	5.353736
21	5.320684
22	5.289731
23	5.044917
24	4.884915
25	4.863728
26	4.833249
27	4.763915
28	4.753362
...	...
...	...
9891	0.9541243
9892	0.9528790
9893	0.9528790

Tab. Scores par ordre décroissant du détecteur d'anomalie (en «rouge» anomalie potentielle)

DÉTECTION D'ANOMALIES - B : Exploration du flux des logs



Sortie de l'algorithme de détection sur le flux réseau du CRIM

DÉTECTION D'ANOMALIES - B : Exploration du flux des logs

- **À investiguer**

```
>
>
>
> investiguer2.lignes
            dstip          srcip sentbyte sentpkt duration dstport proto srcport service sentbyteBypkt count_dest_conn count_src_conn
2048 108.168.151.6   10.30.90.128      164       3   101468      80     6   60876    HTTP       54.66667               1              1
2055 108.168.151.6 132.217.254.98   588558   10126   101468      80     6   60876    HTTP       58.12344               1              1
4306  23.98.49.206 132.217.151.12  4366901    3038        6     443     6   51561   HTTPS    1437.42627               1              1
4704  23.98.49.206 132.217.151.12  4365459    3037        4     443     6   51562   HTTPS    1437.42476               1              1
6758  23.98.49.206 132.217.151.12  4365407    3036        9     443     6   51563   HTTPS    1437.88109               1              1
8464  23.98.49.206 132.217.151.12  4369733    3039        6     443     6   51567   HTTPS    1437.88516               1              1
8569  23.98.49.206 132.217.151.12  4369733    3039        8     443     6   51564   HTTPS    1437.88516               1              1
>
```

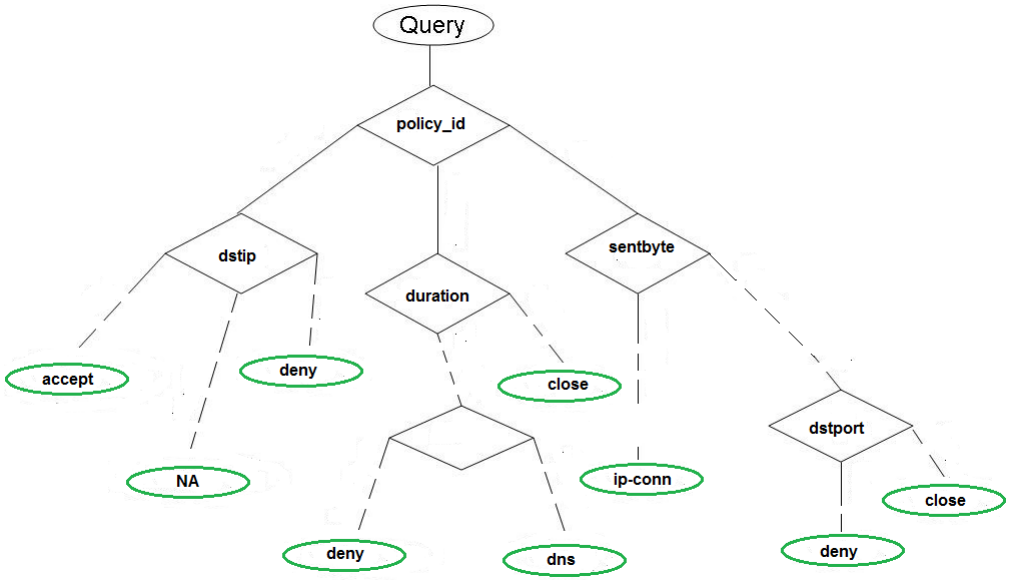
Cohérence des règles (politique)

Entrée
"dstip","srcip","sentbyte","sentpkt","duration","dstport","proto","srcport","service","sentbyteBypkt","count_dest_conn", 'count_src_conn", "policy_id"
Sortie
'status'

Règles

[1] 344 427 405 104 15001 44 0 200 40 11 2000 9 371 287 204
[16] 48 15026 15810 100 377 18 1204 15500 435 110 15830 433 41 5004 1006
[31] 393 6012 400 15200 421 15802 6204 15817 6016 2101 15812 14 412 385 15850
[46] 472 2110 2402 343 15822 105 300 386 2507 384 6200 2105 475 60 436
[61] 68

Classes:
1. accept
2.NA
3.close
4.deny
5.ip-conn
6.dns



DÉTECTION D'ANOMALIES - C : Classification de paquets

```
xgb <-xgboost(data = data.matrix(data_train),
              label = data.matrix(data_lab),
              eta = 0.001,
              max_depth = 15,
              nround=25,
              subsample = 0.5,
              colsample_bytree = 0.5,
              seed = 1,
              eval_metric = "merror",
              objective = "multi:softmax",
              num_class = length(unique(res)),
              nthread = 3
)
```

Package 'xgboost'

January 5, 2017

Type Package

Title Extreme Gradient Boosting

Version 0.6-4

Date 2017-01-04

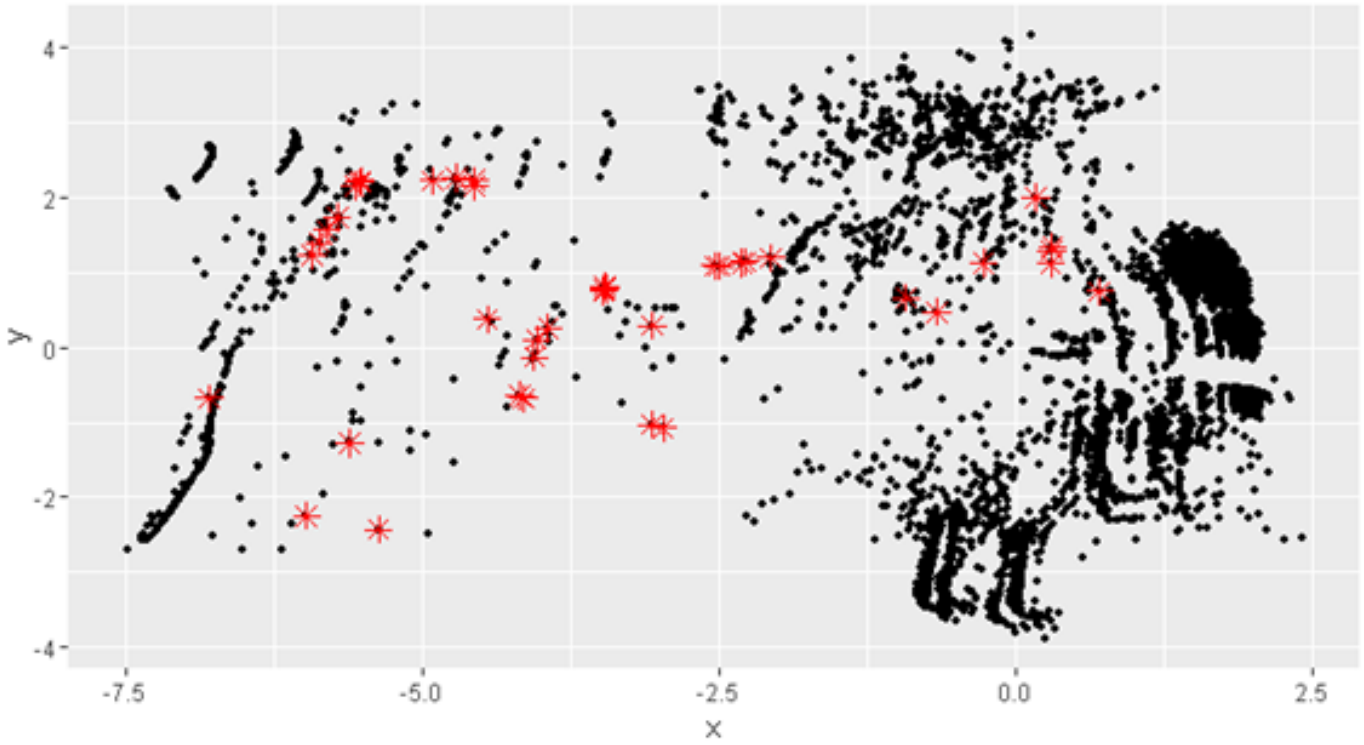
Author Tianqi Chen <tianqi.tchen@gmail.com>, Tong He <hetong007@gmail.com>, Michael Benesty <michael@benesty.fr>, Vadim Khotilovich <khotilovich@gmail.com>, Yuan Tang <terrytangyuan@gmail.com>

Maintainer Tong He <hetong007@gmail.com>

Description Extreme Gradient Boosting, which is an efficient implementation of the gradient boosting framework from Chen & Guestrin (2016) <doi:10.1145/2939672.2939785>. This package is its R interface. The package includes efficient linear model solver and tree learning algorithms. The package can automatically do parallel computation on a single machine which could be more than 10 times faster than existing gradient boosting packages. It supports various objective functions, including regression, classification and ranking. The package is made to be extensible, so that users are also allowed to define their own objectives easily.

License Apache License (== 2.0) | file LICENSE

Cohérence des règles (politique)



Sortie de l'algorithme de vérification de cohérence des règles de classification de packets sur le flux réseau du CRIM (En rouge, les lignes potentiellement non cohérentes).

Conclusion

- **Modules de profilage du trafic réseau**
  - **Visualisation des données**
    - **Meta données**
    - **outils de décisions +/- naif**
  - **Exploration du flux des connexions**
    - **Scanner toutes les lignes**
    - **LoF : Densité locale et distance de joignabilité**
    - **Seuillage pour détection de menaces**
  - **Cohérence des règles**
    - **Respecter la politique définie du parefeu**
    - **Classificateur xgboost**

Équipe Développement et technologies Internet
CRIM – Centre de recherche informatique de Montréal

Mohamed.dahmane@crim.ca
Tél. : 514 840-1235 poste 6976

Suivez-nous :

Dialoguez avec

Suivez-nous

@CRIM_ca

wwwCRIMca

Le CRIM est un centre de recherche appliquée en TI qui développe, en mode collaboratif avec ses clients et partenaires, des technologies innovatrices et du savoir-faire de pointe, et les transfère aux entreprises et aux organismes québécois afin de les rendre plus productifs et plus compétitifs localement et mondialement. Le CRIM dispose de quatre équipes de recherche en TI de calibre mondial et œuvre principalement dans les domaines des interactions et interfaces personne-système, de l'analytique avancée et de la science et technologie du logiciel. Détenteur d'une certification ISO 9001:2008, son action s'inscrit dans les politiques et stratégies pilotées par le ministère de l'Économie, de la Science et de l'Innovation, son principal partenaire financier.