Spam Campaign Detection, Analysis, and Investigation

By

Son Dinh, Taher Azeb, Francis Fortin, Djedjiga Mouheb and Mourad Debbabi

Presented At

The Digital Forensic Research Conference

DFRWS 2015 EU Dublin, Ireland (Mar 23rd- 26th)

SPAM CAMPAIGN

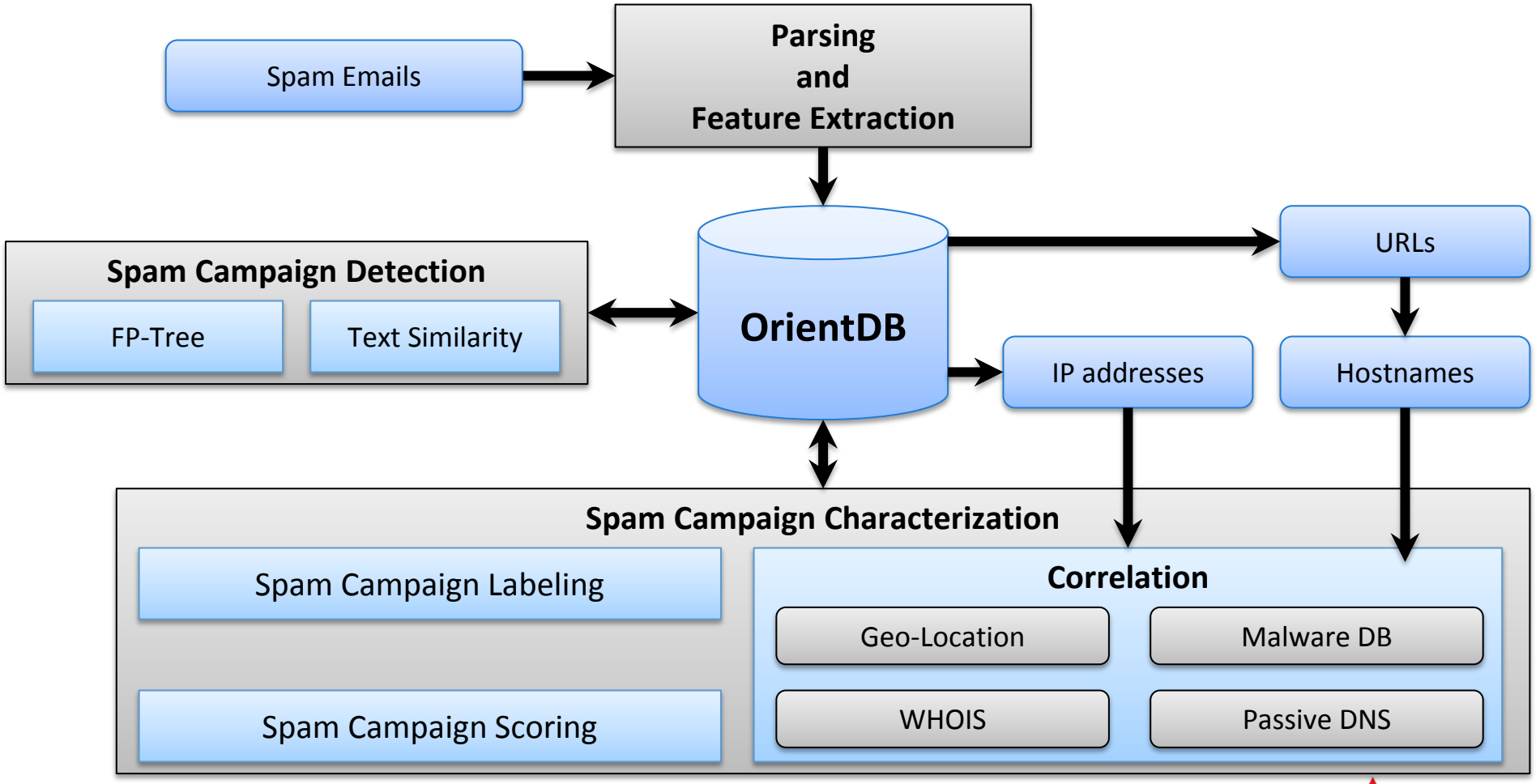
DETECTION, ANALYSIS AND INVESTIGATION

S. Dinh, T. Azab, F. Fortin, D. Mouheb, M. Debbabi

Outline

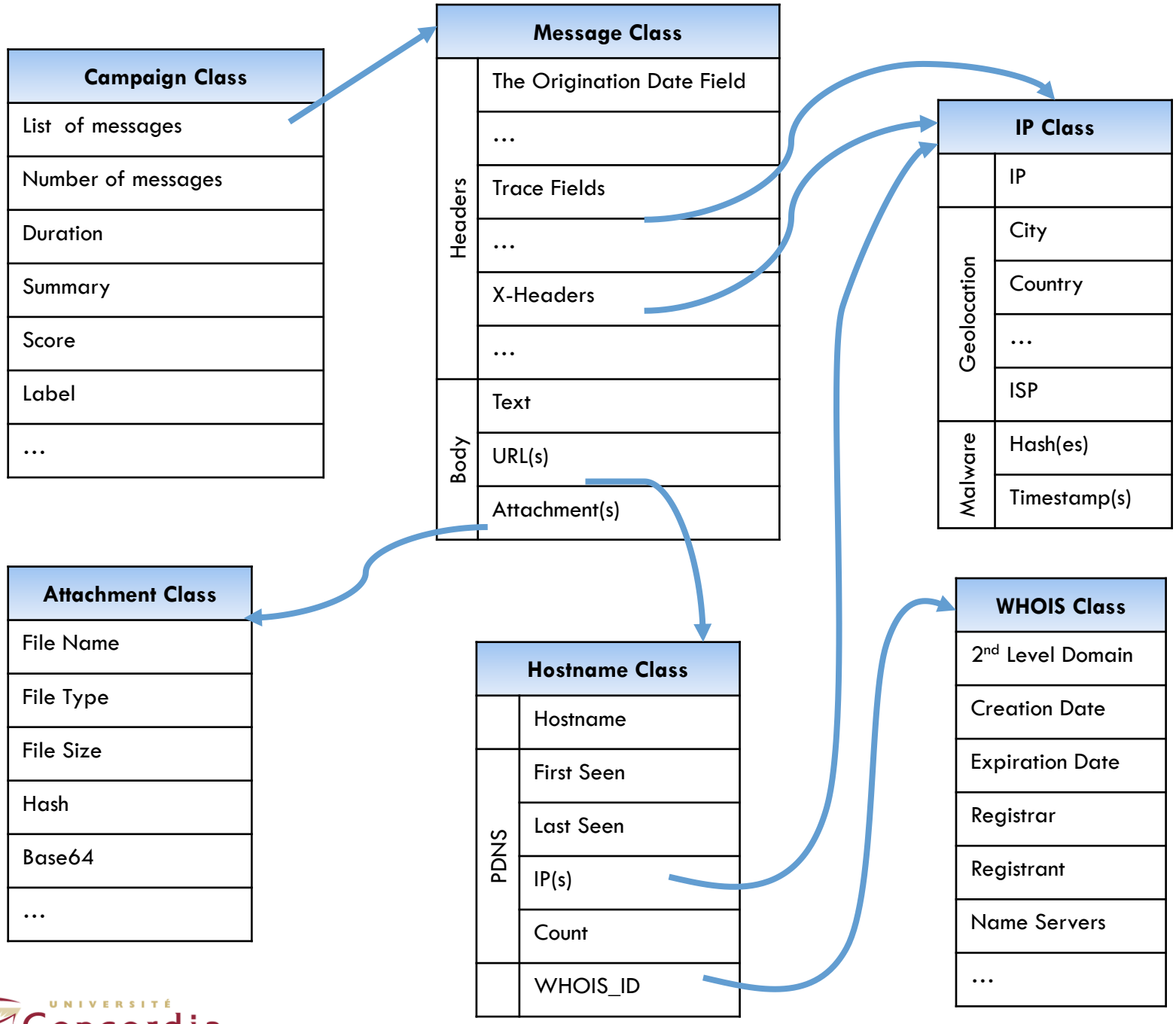
- Architecture
- Central Database
- Parsing and Feature Extraction
- Spam Campaign Detection
- Spam Campaign Characterization
- Results

Architecture



Central Database

- Relational database management system (RDBMS)
- The number of spam emails exceeds 6 *trillion* in Q4 2014
  - McAfee Labs Threats Report (February 2015)
- OrientDB:
  - Flexibility (document database)
  - Interconnectivity (graph database)
  - Scalability
- Documents: spam campaigns, emails, IP addresses, domain names, attachments
- Connections and interconnections between spam campaigns, emails, IP addresses, domain names and attachments



Preprocessing

Parsing and Feature Extraction

- Parser
  - Parse and store standard header fields (RFC5322)
  - Unpack email bodies (single-part or multi-part)
  - Extract embedded URLs
- Feature Extractor
  - Content Type
    - text/*
    - multipart/*
    - image/*
    - application/*
    - etc.
  - Character Set
    - utf-8
    - iso-8859-1
    - windows 1250
    - shift-jis
    - koi8-r
    - etc.
  - Subject
    - Decoded to Unicode
  - URL Tokens
  - Attachment

Parsing and Feature Extraction

- Feature Extractor
  - Email Layout

- text layout
  - T for a paragraph with only text
  - U for a URL
  - N for an empty line
  - For example:

    TNTNUNN

- HTML layout
  - Top levels of the DOM tree
  - For example:

```
<html>
-<head>
-</head>
-<body>
--<p></p>
--<br />
--<div></div>
-</body>
<html>
```

- multipart layout
  - The structure of the multipart email
  - For example:

```
multipart/mixed
-multipart/alternative
--text/plain
--multipart/related
---text/html
---img/jpg
-application/pdf
```

Detection & Characterization

Spam Campaign Detection

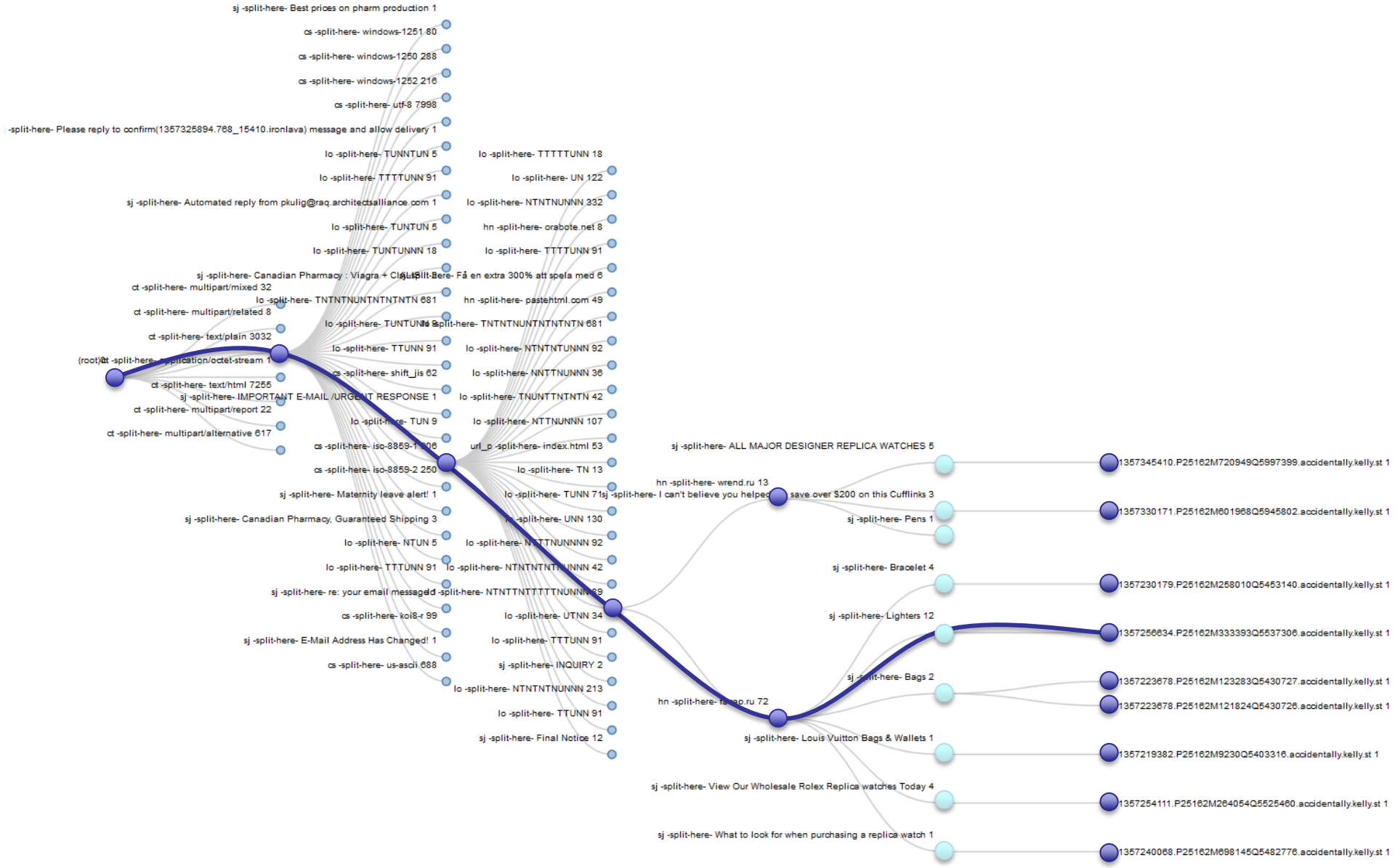
- Distance between two spam emails:
  - w-shingling and the Jaccard coefficient
  - Context Triggered Piecewise Hashing (CTPH)
  - Locality-Sensitive Hashing (LSH)
- Hierarchical, partitional, neural network-based, kernel-based clustering techniques
- Problems:
  - Scalability
  - Obfuscation techniques
    - Spam email templates
    - Randomly picked paragraphs from books or Wikipedia articles
    - Randomly generated subdomains and fast-flux service networks

Spam Campaign Detection

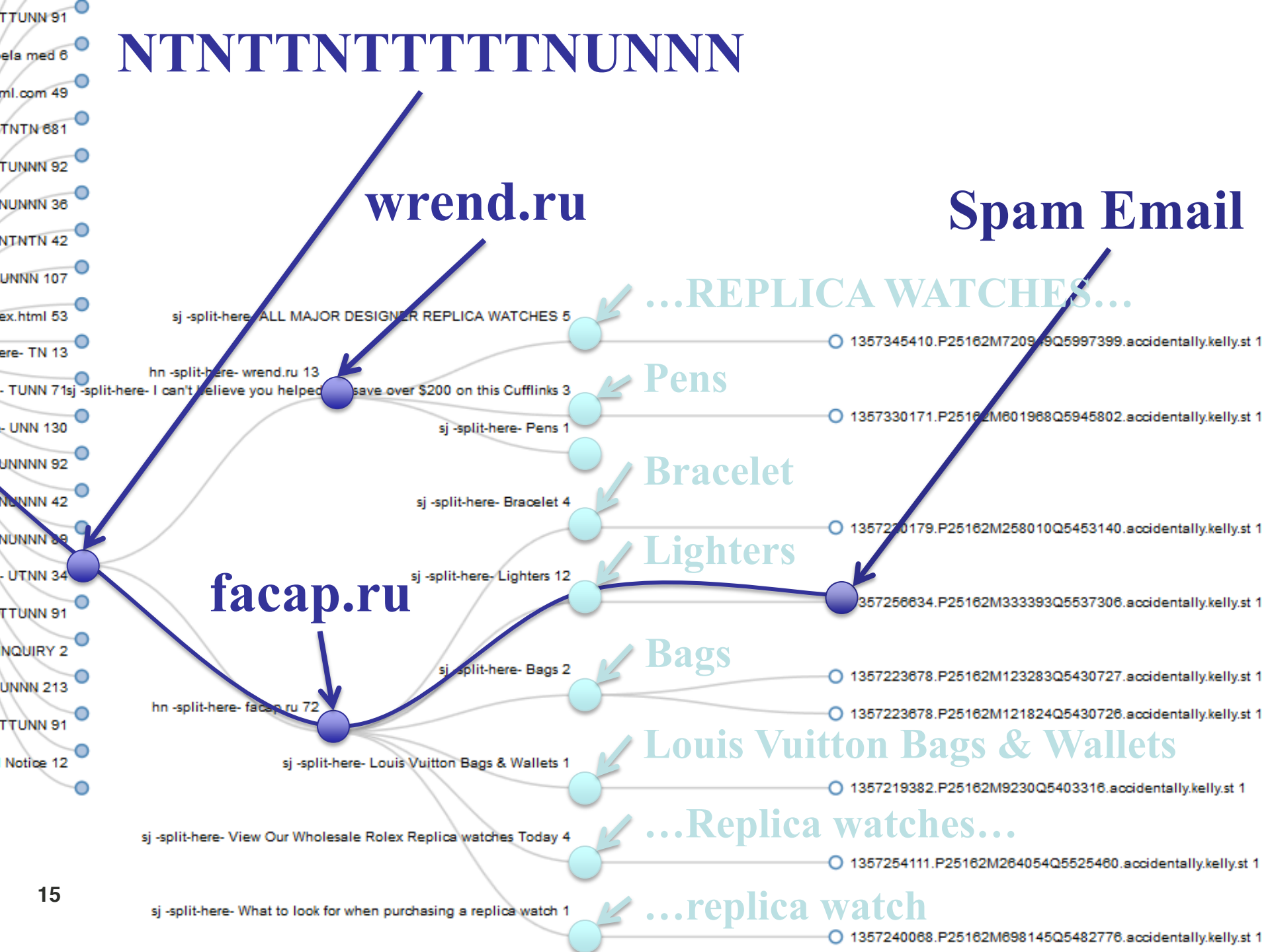
- Frequent-Pattern Tree (FP-Tree)
  - The more frequent a feature is, the more it is shared among spam emails
  - Less frequent features correspond to the obfuscated parts
  - Two scan of the dataset:
    - First scan to compute the number of occurrences for each feature
    - Second scan to insert feature vectors into the tree
  - The cost of inserting a feature vector fv into the FP-Tree is $O(|fv|)$, where $|fv|$ is the number of features in fv.
  - The **Content Type** feature is put at the beginning of each feature vector
  - The unique ID of each email is kept at the end of each feature vector
  - Embedded URLs are split into tokens

Spam Campaign Detection

- Spam Campaign Identification
  - Traverse the tree
  - Conditions:
    - The number of children >= `min_num_children`
    - The average count of children >= `freq_threshold`
    - The path to root must contain **one** feature type not in `n_obf_features`
    - The number of leaves of the sub-tree >= `min_num_messages`
  - If a node satisfies the conditions:
    - The leaves of the sub-tree are spam emails from the same campaign
    - The path from the root to this node contains the common features
    - The sub-tree is then removed from the tree

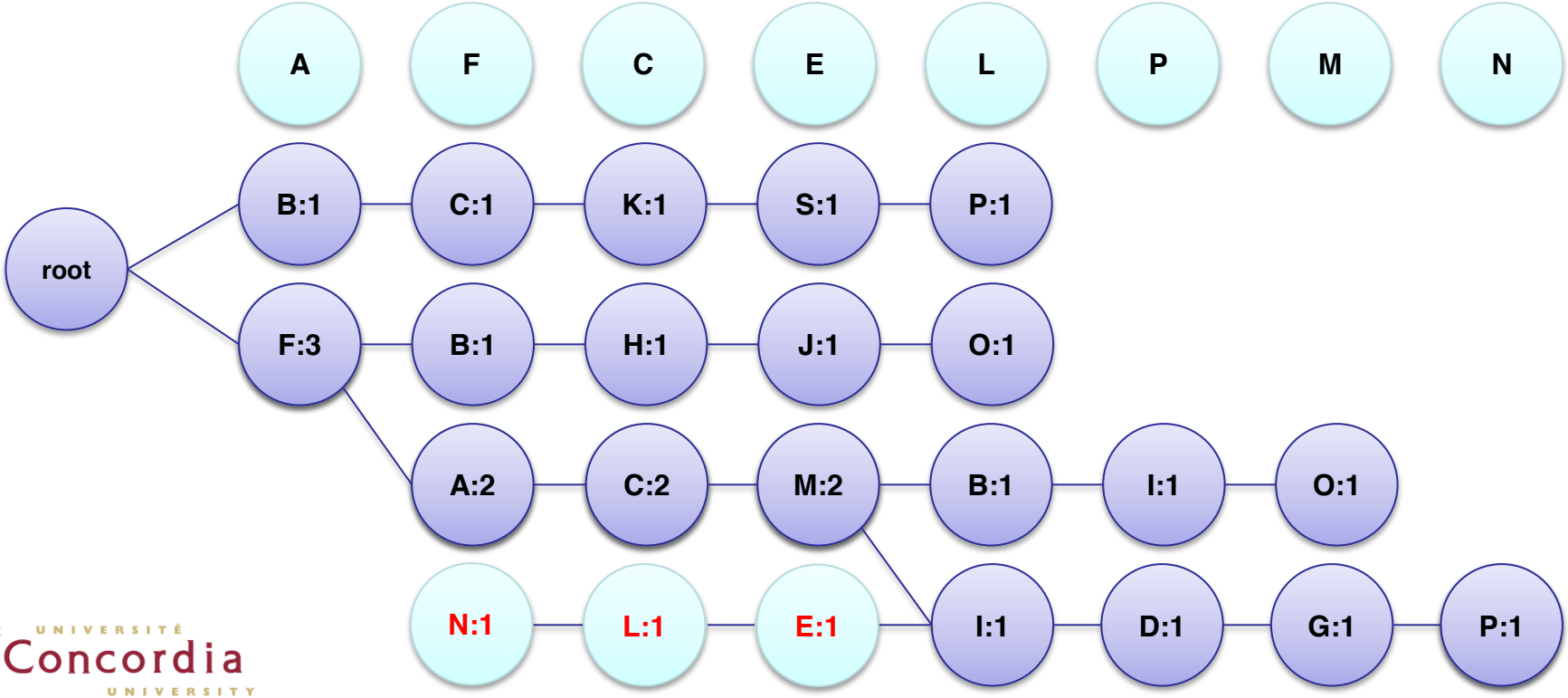



text/plain
root
iso-8859-1
NTNTTNTTTTTTNUNNN
(root)
ct -split-here- application/octet-stream 1
ct -split-here- multipart/alternative 617
ct -split-here- multipart/mixed 32
ct -split-here- multipart/related 8
ct -split-here- multipart/report 22
ct -split-here- text/html 7255
ct -split-here- text/plain 3032
cs -split-here- iso-8859-1 306
cs -split-here- iso-8859-2 256
cs -split-here- koi8-r 99
cs -split-here- shift_jis 62
hn -split-here- pastehtml.com 49
lo -split-here- NNTTNUNNN 36
lo -split-here- NTNTNTUNNN 42
lo -split-here- NTNTNTUNNN 92
lo -split-here- NTTNUNNN 107
lo -split-here- NTTTNUNNNN 92
lo -split-here- TN 13
lo -split-here- TNUNTTNTNTN 42
lo -split-here- TTTTUNN 91
lo -split-here- TUNN 71sj -sp
lo -split-here- TUNTUNNN 18
lo -split-here- TTTUNN 91
lo -split-here- TNTNTNUNTNTNTNTN 681
lo -split-here- TTUNN 91
lo -split-here- TUNTUNN 8
lo -split-here- TN 9
lo -split-here- UNN 130
lo -split-here- UTNN 34
sj -split-here- Canadian Pharmacy : Viagra + Cialis
sj -split-here- Få en extra 300% att spela med 6
sj -split-here- Maternity leave alert! 1
sj -split-here- Canadian Pharmacy, Guaranteed Shipping 3
sj -split-here- re: your email messageId
id -split-here- NTNTTNTTTTTTNUNNN 89


NTNTTNTTTTTTNUNNN
wrend.ru
Spam Email
facap.ru
TTUNN 91
eela med 6
ml.com 49
TNTN 681
TUNNN 92
NUNNN 36
NTNTN 42
UNNN 107
ex.html 53
ere- TN 13
- TUNN 71
- UNN 130
UNNNN 92
NUNNN 42
NUNNN 89
- UTNN 34
TTUNN 91
NQUIRY 2
UNNN 213
TTUNN 91
Notice 12
sj -split-here- ALL MAJOR DESIGNER REPLICA WATCHES 5
hn -split-here- wrend.ru 13
sj -split-here- I can't believe you helped save over $200 on this Cufflinks 3
sj -split-here- Pens 1
sj -split-here- Bracelet 4
sj -split-here- Lighters 12
sj -split-here- Bags 2
hn -split-here- facap.ru 72
sj -split-here- Louis Vuitton Bags & Wallets 1
sj -split-here- View Our Wholesale Rolex Replica watches Today 4
sj -split-here- What to look for when purchasing a replica watch 1
…REPLICA WATCHES…
Pens
Bracelet
Lighters
Bags
Louis Vuitton Bags & Wallets
…Replica watches…
…replica watch
1357345410.P25162M720949Q5997399.accidentally.kelly.st 1
1357330171.P25162M601968Q5945802.accidentally.kelly.st 1
1357220179.P25162M258010Q5453140.accidentally.kelly.st 1
1357256634.P25162M333393Q5537306.accidentally.kelly.st 1
1357223678.P25162M123283Q5430727.accidentally.kelly.st 1
1357223678.P25162M121824Q5430726.accidentally.kelly.st 1
1357219382.P25162M9230Q5403316.accidentally.kelly.st 1
1357254111.P25162M264054Q5525460.accidentally.kelly.st 1
1357240068.P25162M698145Q5482776.accidentally.kelly.st 1

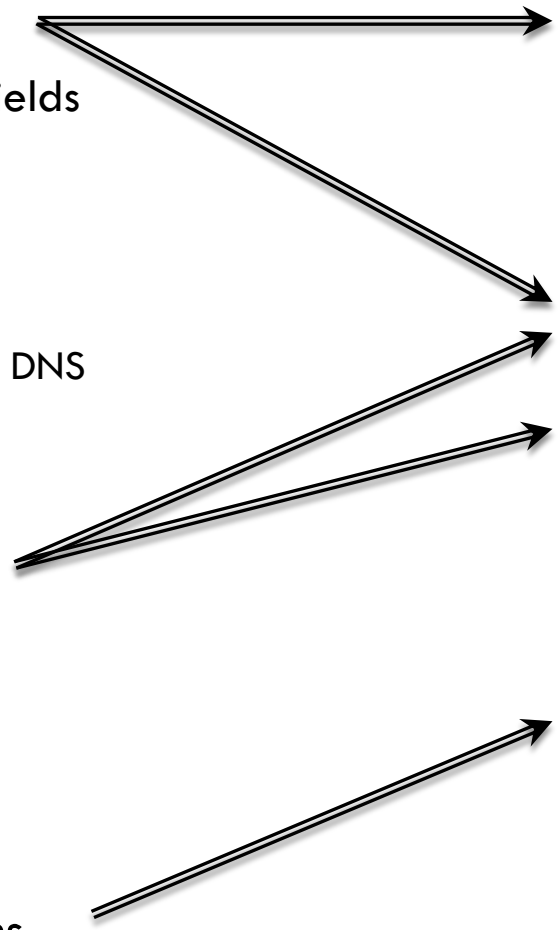
Spam Campaign Detection

- Incremental Frequent-Pattern Tree
  - Spam campaigns may last for a long period of time and therefore should be identified in their early stage
  - Feature vectors are extracted as soon as spam emails arrive and are inserted into the FP-Tree from the root level



Spam Campaign Characterization

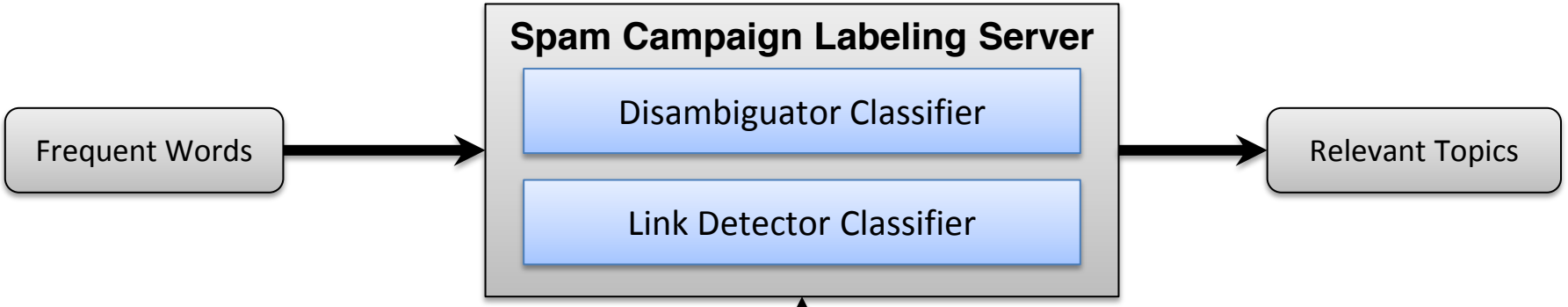
- IP addresses
  - From header fields
    - Received
    - X-*
  - From URLs
    - using passive DNS
- Hostnames
  - From URLs
- 2nd-level domains



- Geo-Location
  - City, Country
  - ISP
  - Organization
- Malware Database
- Passive DNS
  - Timeslot (First/Last seen)
  - IP address(es)
  - Count
- WHOIS
  - Name servers
  - Domain status
  - Creation date
  - Expiration date
  - Registrar & registrant

Spam Campaign Characterization

- Spam Campaign Labeling



Disambiguator Classifier

- Solves the issue of ambiguous terms
- Chooses the most suitable Wikipedia article for each term

Link Detector Classifier

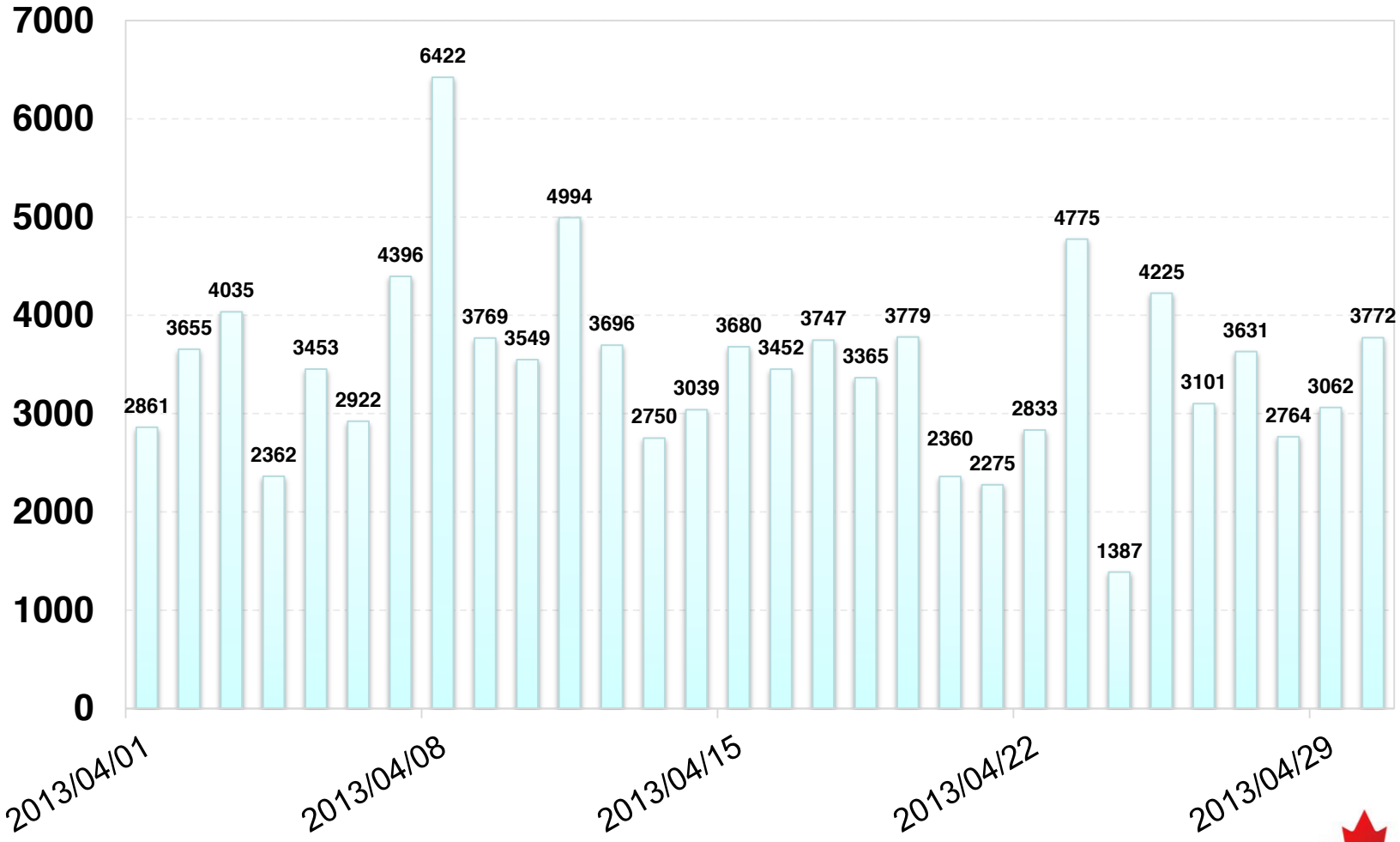
- Selects sufficiently relevant associations between terms and Wikipedia articles to retain

Spam Campaign Characterization

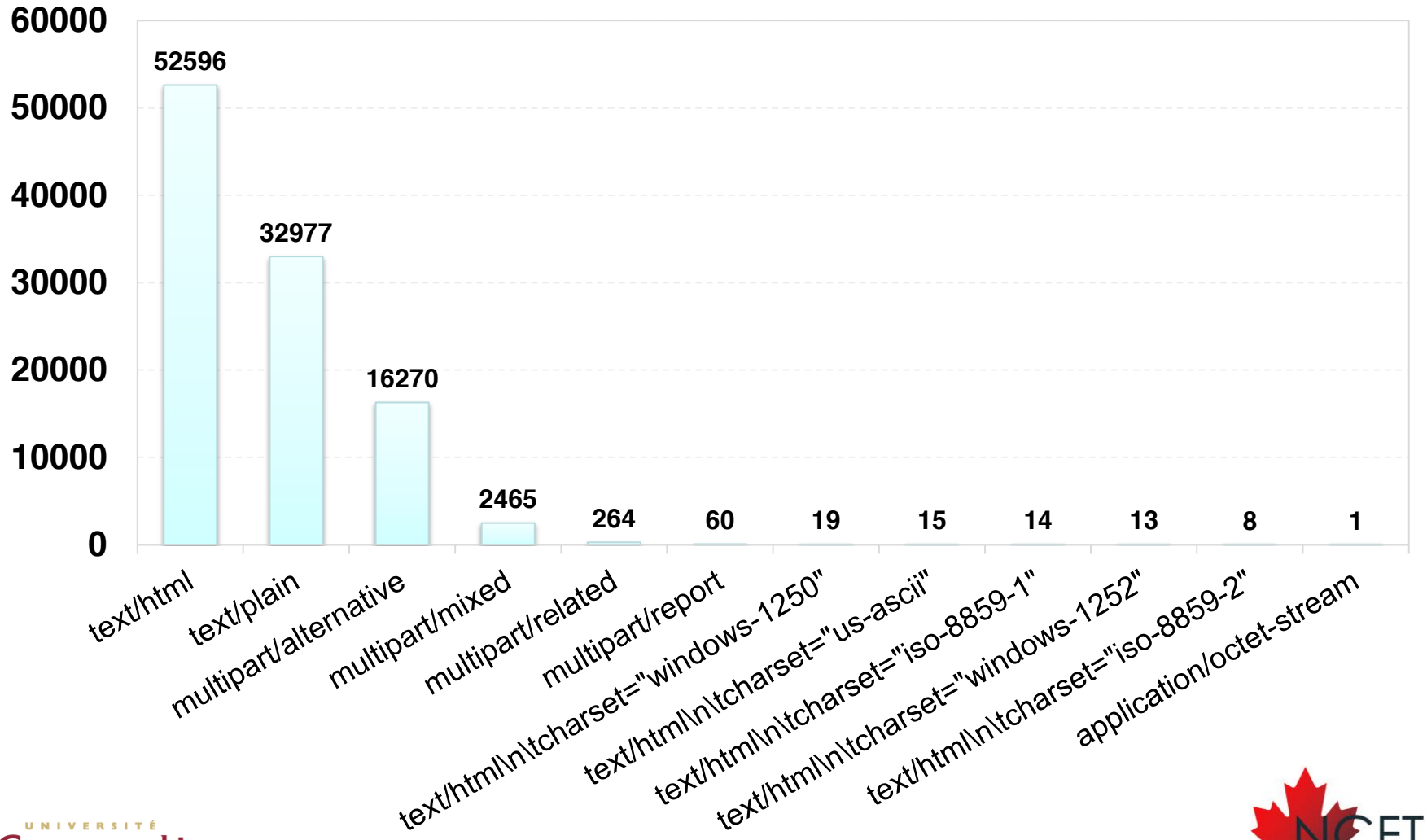
- Spam Campaign Scoring
  - An investigator may only need to pursue a particular objective
  - Each spam campaign is assigned a score computed using:
    - The number of spam emails inside a campaign
    - The number of IP addresses in Canada
    - The number of domain names that are resolved to IP addresses in Canada
    - The number of ".ca" TLDs that appear in the from field
    - The number of ".ca" TLDs that appear in the embedded URLs
    - The number of Canadian city names that appear in the content
    - The number of appearances of the string "Canada" in the content
    - The number of IP addresses that are associated with malware
    - The number of IP addresses that belong to a specific IP range
  - Each criterion has a customizable weight
  - The criteria have been verified by a law enforcement official

Results

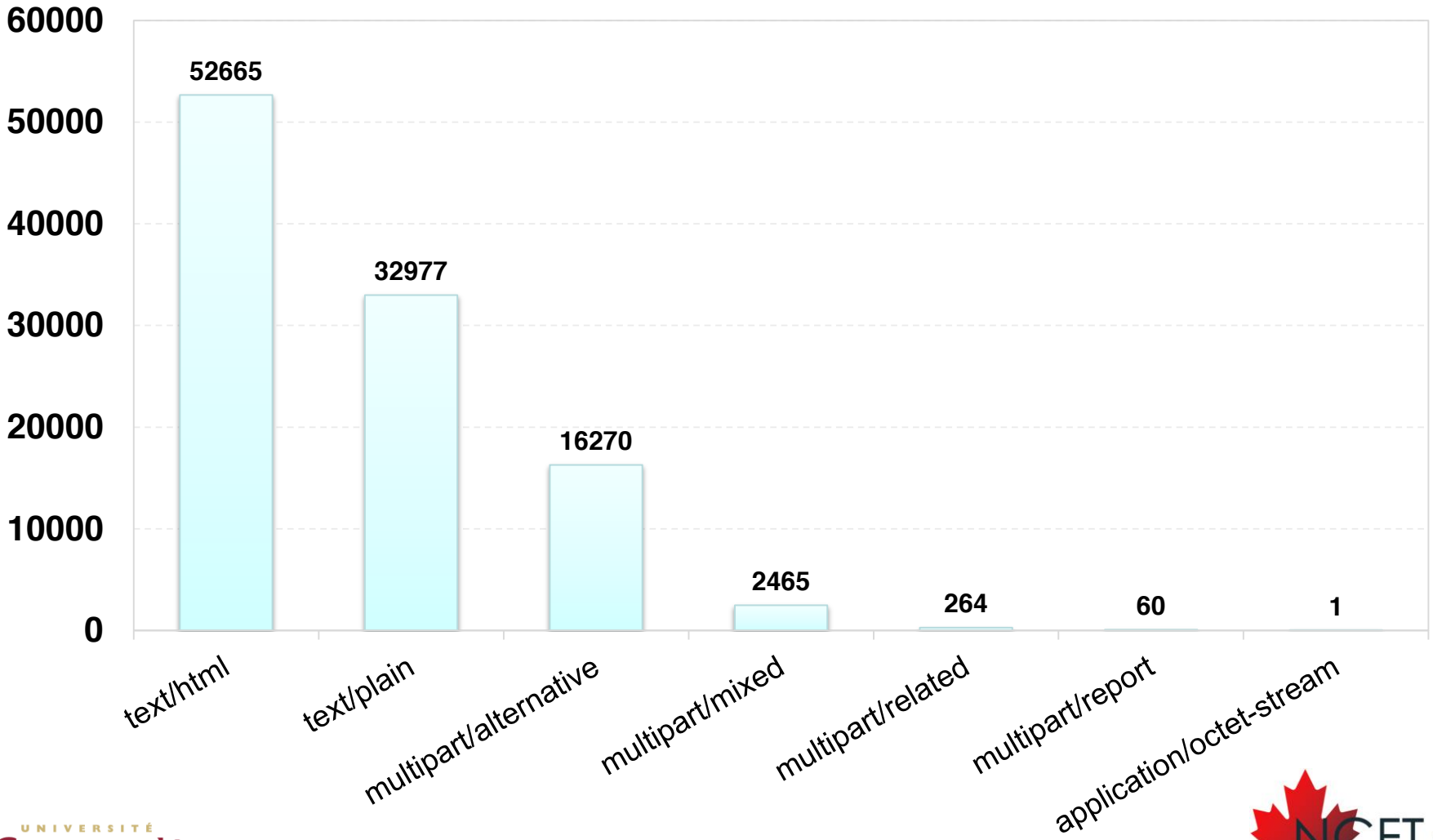
Spam Data

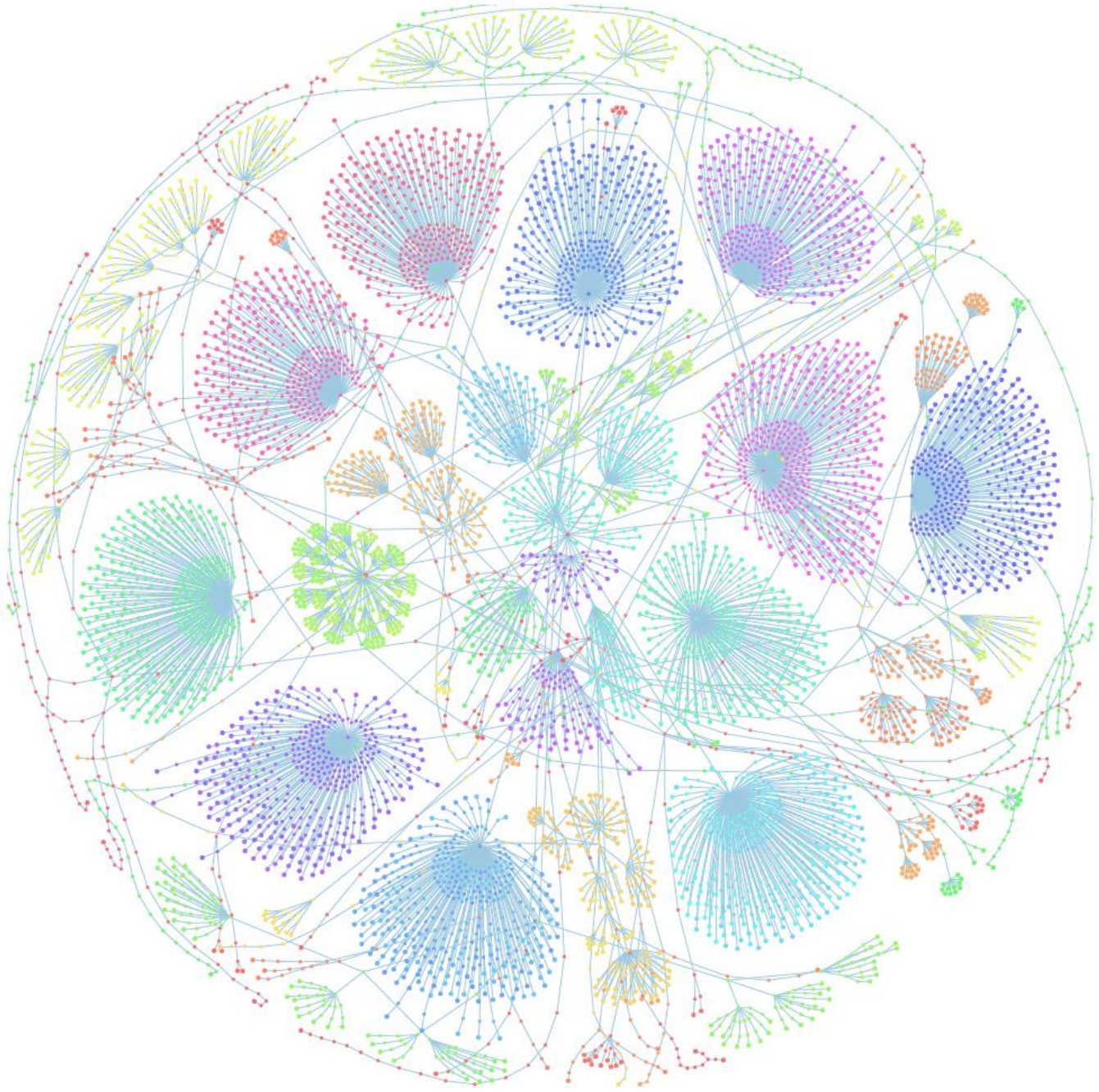

7000
6000
5000
4000
3000
2000
1000
0
2861
3655
4035
2362
3453
2922
4396
6422
3769
3549
4994
3696
2750
3039
3680
3452
3747
3365
3779
2360
2275
2833
4775
1387
4225
3101
3631
2764
3062
3772
2013/04/01
2013/04/08
2013/04/15
2013/04/22
2013/04/29

Raw MIME Types



Correct MIME Types

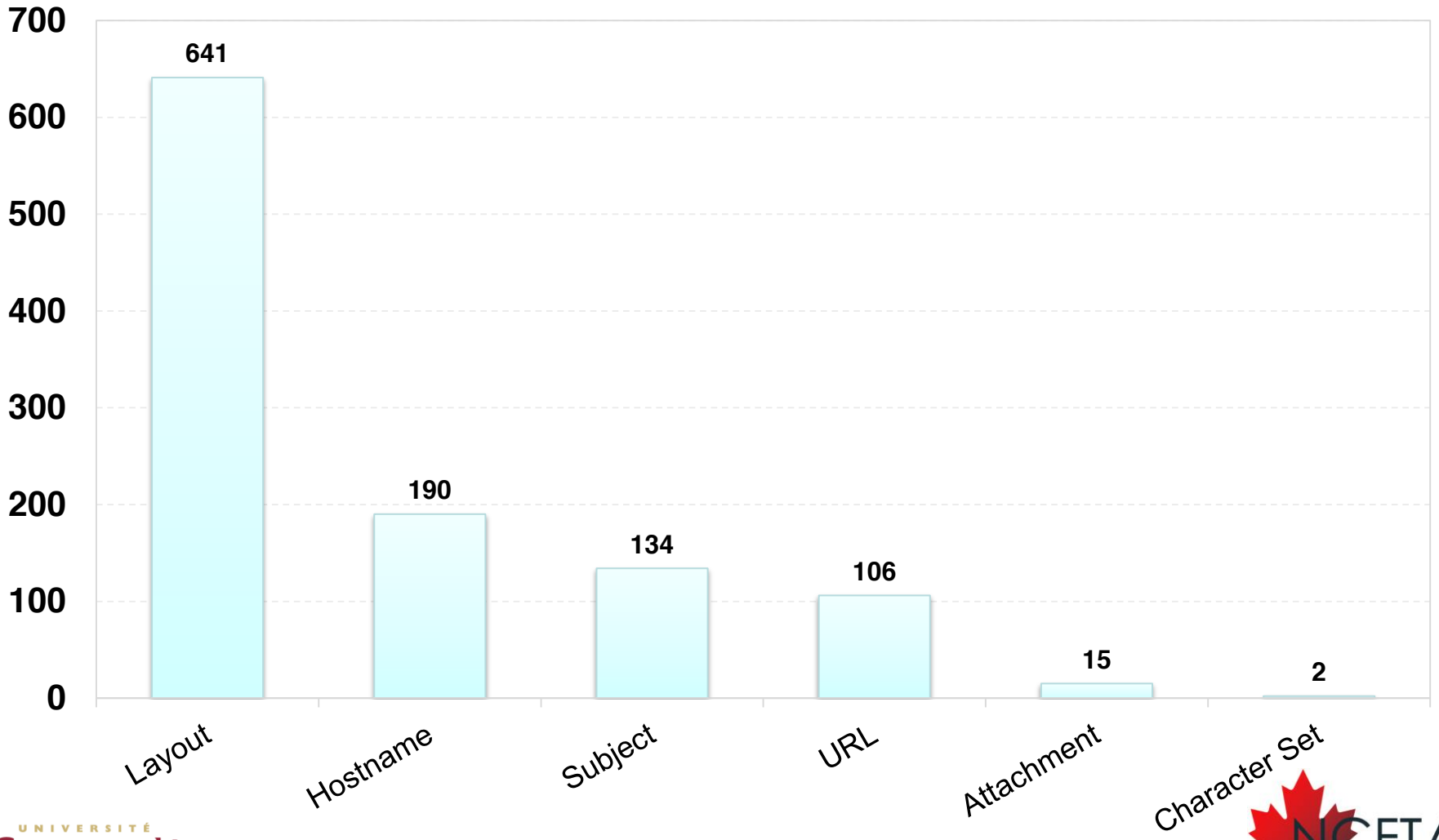




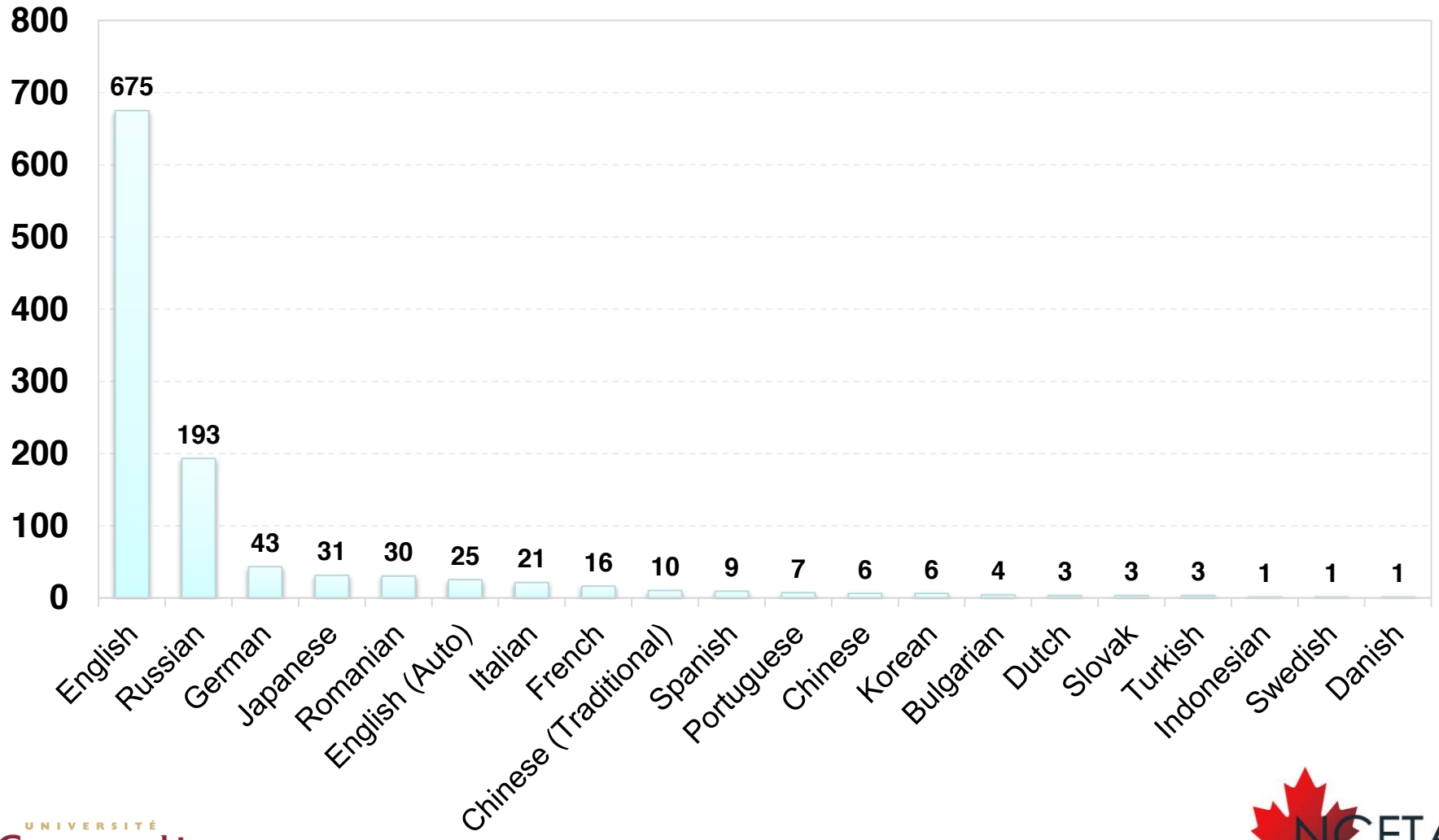
UNIVERSITÉ Concordia UNIVERSITY

NCFTA
National Cyber-Forensics and Training Alliance CANADA

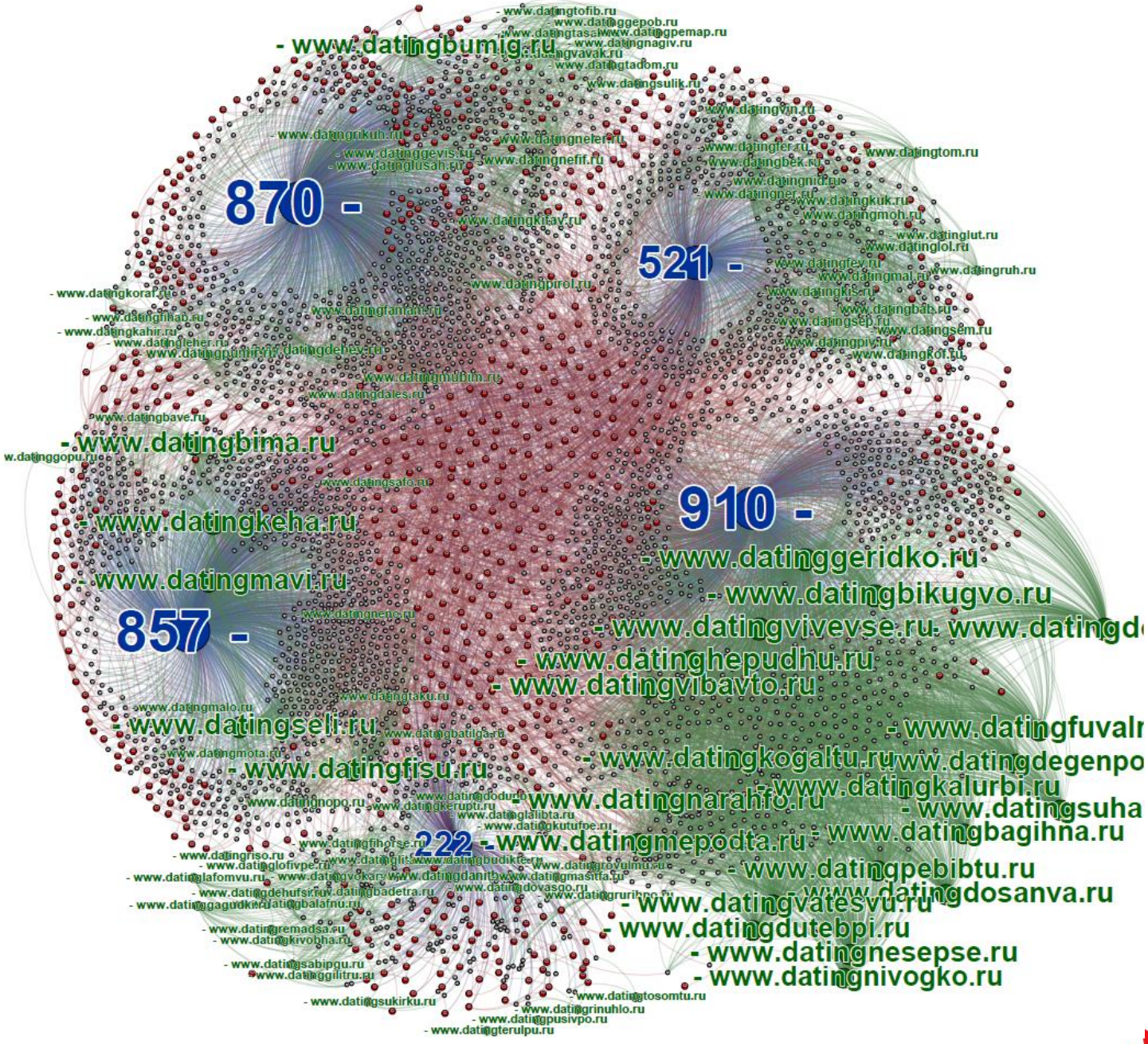
Decisive Features

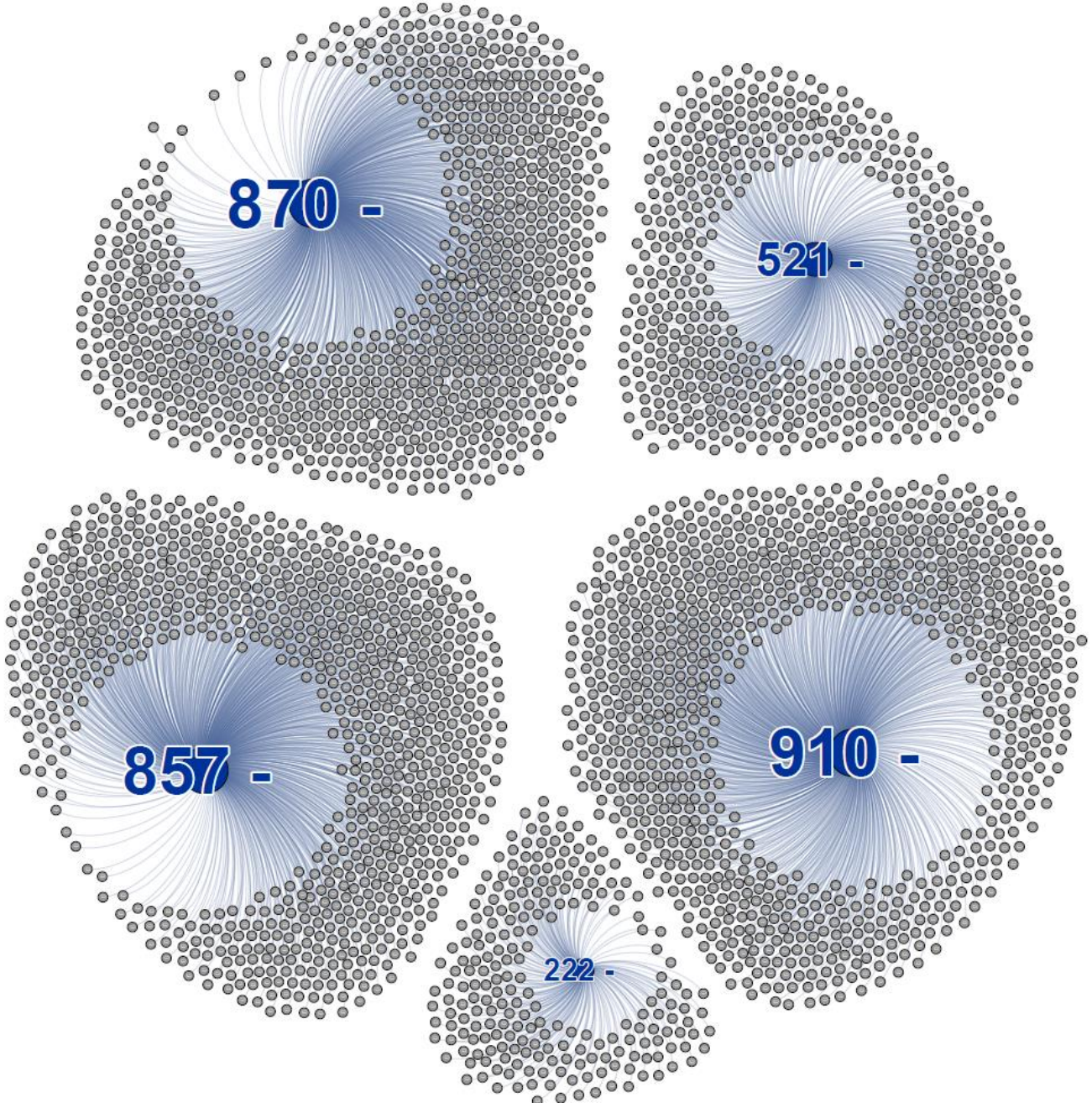


Spam Campaign Languages



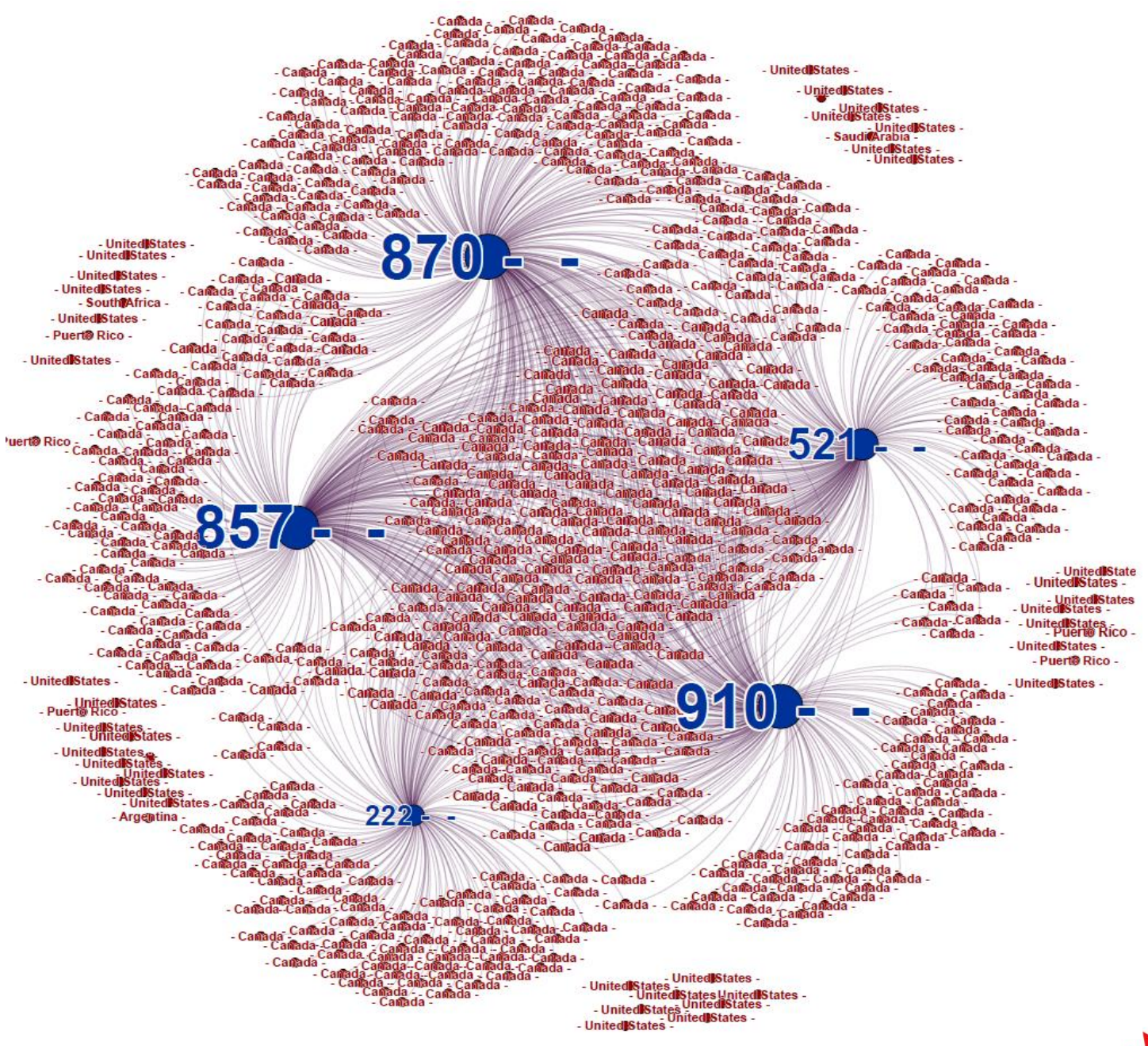
Spam Campaign Topics





UNIVERSITÉ Concordia UNIVERSITY

NCFTA
National Cyber-Forensics and Training Alliance CANADA


870
857
521
910
222

Conclusions

- Spam:
  - Still a serious threat itself
  - An instrument for other cyber attacks
- A spam campaign detection, analysis and investigation system that:
  - Identifies spam campaigns (on-the-fly)
  - Aggregates different data sources to expose campaign characteristics
  - Labels spam campaigns
  - Scores spam campaigns
- Future work:
  - Improve spam campaign labeling (more suitable for spam contents)
  - Spam campaign categorization (threat-based)

Acknowledgements

- Concordia University
  - Concordia Institute for Information Systems Engineering (CIISE)
- National Cyber-Forensics and Training Alliance (NCFTA) Canada
- Canadian Radio-television Telecommunications Commission (CRTC)