Big Data & Artificial Intelligence Case Competition

Alex Pegg, Nimisha Mundade, Christian Thormeyer, Massil Beguenane

Scotiabank

Team Profile

Massil Beguenane

Massil is a fourth-year Commerce student, specializing in Finance, majoring in Economics, and minoring in Statistics.

massil.beguenane@mail.utoronto.ca
Toronto, Canada

Nimisha Mundade

Nimisha is a fourth-year Economics student. She moved to Canada three years ago from Switzerland to pursue her post-secondary education.

nimisha.mundade@mail.utoronto.ca
Toronto, Canada

Christian Thormeyer

Christian is a third-year Business student, specializing in Management with a minor in Political Science.

christian.thormeyer@mail.utoronto.ca
Toronto, Canada

Alex Pegg

Alex Pegg is a fourth-year student, pursing a double major in Computer Science and Economics.

alex.pegg@mail.utoronto.ca
Toronto, Canada

The Problem in a Global Context

Who are the victims?

In 2019, approximately **24.9 million** people were detected as victims of human trafficking. This represents about **0.32% of the global population** who are being robbed of their human rights as a direct result of being trafficked for sexual exploitation, forced labor or other purposes.

Who are the perpetrators?

A human trafficker is anyone who knowingly contributes in the trafficking of people with the intent of exploiting a victim; this includes recruiters, transporters, and employers. Traffickers often belong to the **same ethnic group as their victims and over 35% of prosecuted traffickers are females.**

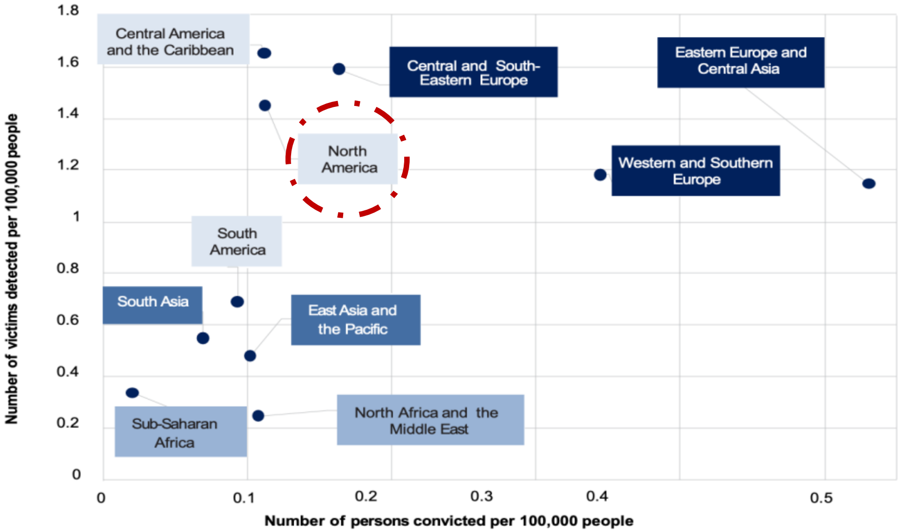
Why does this trade exist?

Slavery did not end with abolition in the 19th century, it only evolved and changed forms. Human trafficking is a **form of modern slavery** and one of the most lucrative criminal businesses in the world with estimated illegal profits of over **US$150 billion** annually.

How does this impact us?

Human trafficking is a crime against humanity and many of its victims are not just exploited but **tortured and killed.** Recently reports show that proceeds from human trafficking have been used to **enable armed conflicts and other criminal activities.**

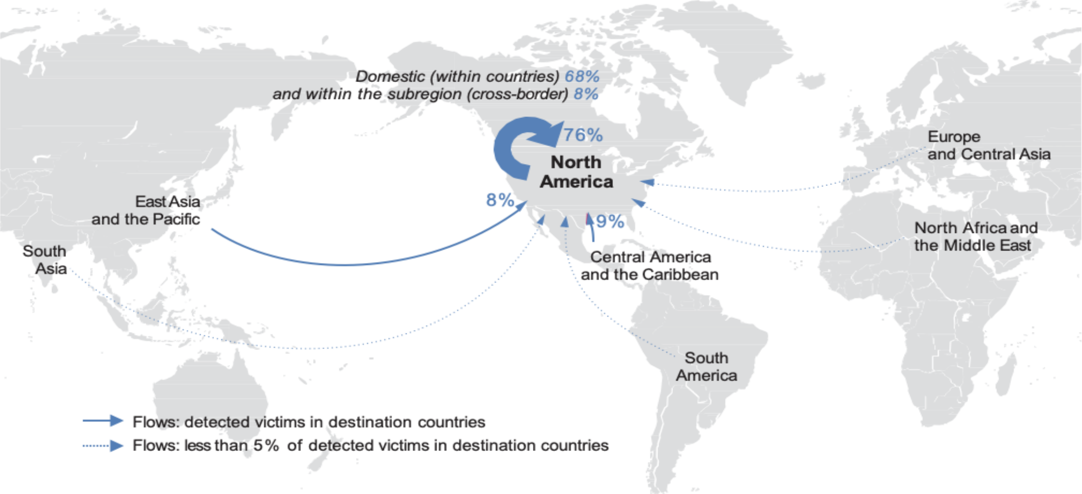
The Problem in the Canadian Context



Source: UNODC elaboration of national data.

1. Number of victims as a proportion of the population is one of the **highest in North America**
2. In 2016, **17,000** people were living in conditions of modern slavery in Canada
3. About **95%** of the trafficking victims in Canada were female and **72%** were under the age of 25

1. Most victims in Canada originate from **within the domestic borders**
2. At risk populations include **indigenous women and girls, LGBTQ+, immigrants, children in the welfare system and economically disadvantaged people**



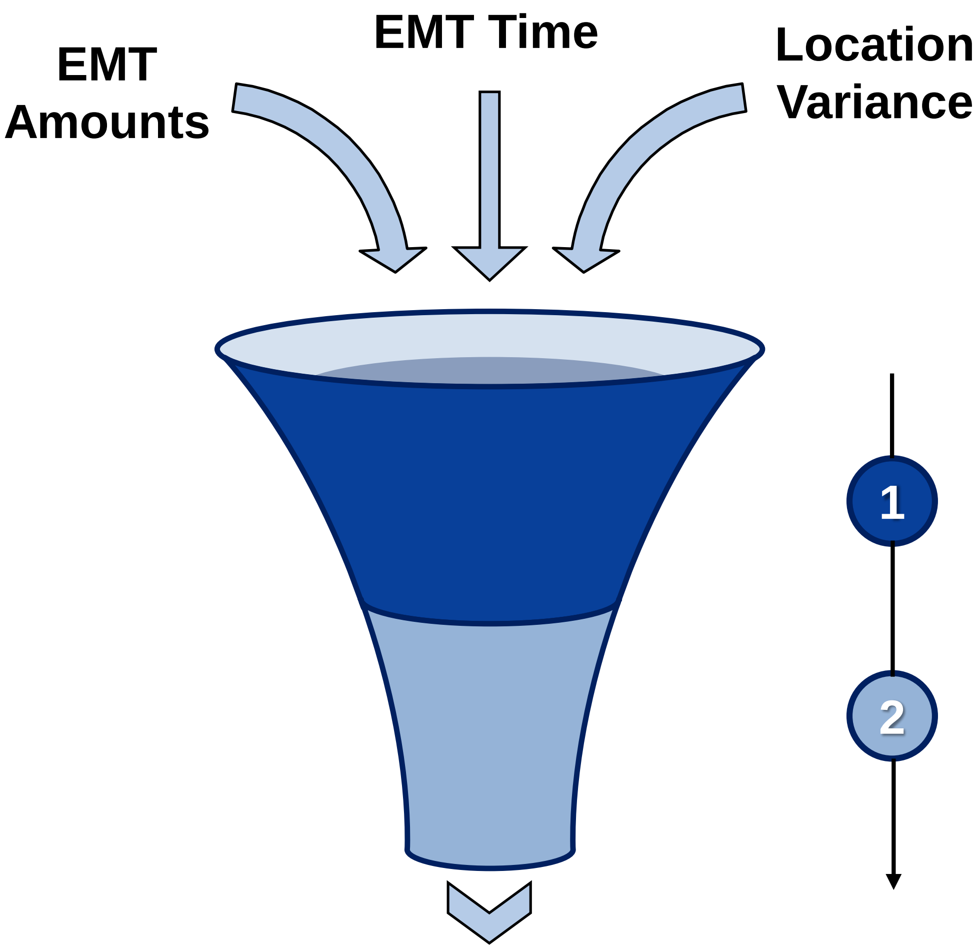
Source: UNODC elaboration of national data.

The lack of data and knowledge on how to interpret existing data to identify incidents related to human trafficking are the two key obstacles that prevent us from eliminating this issue.

Scotiabank

Our Solution

A two-step filtering system that reduces the number of false positives & classifies behaviour that is most likely human trafficking



1 **Filter 1 -** Finding ***Outliers*** Using Big Data

2 **Filter 2 -** Using ***HT Score*** *as a 2nd Filter*

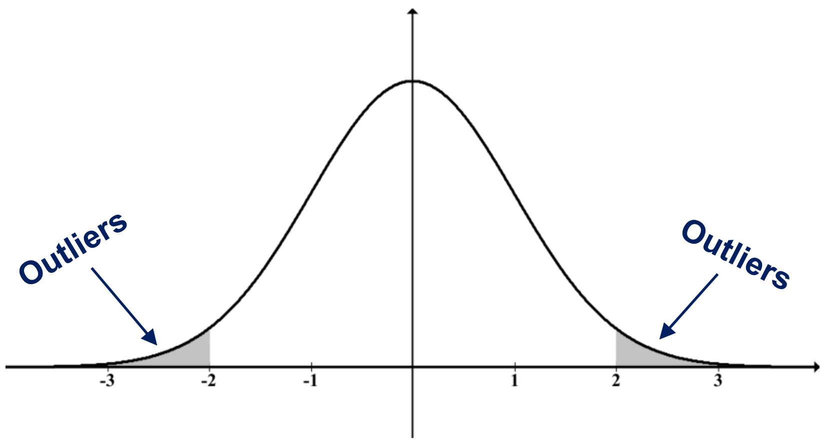
Determine and Classify Potential Behavior of Human Traffickers

Scotiabank

Filter 1 Assumptions

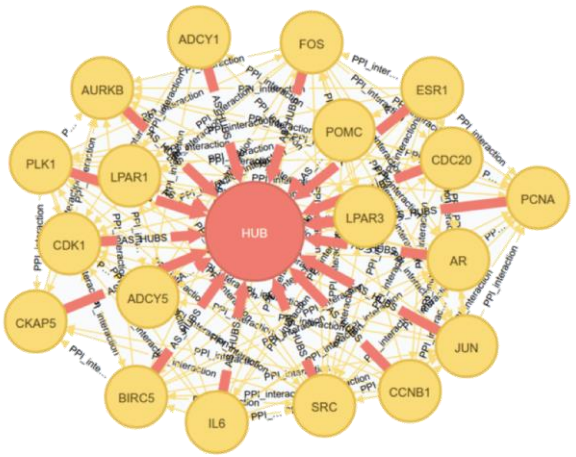
Outliers

1. **Money Laundering Activity is unusual behavior**
2. **General or regular behavior patterns in EMT can be modeled**
3. **Outliers may be flagged as suspicious**



Sophisticated Networks

1. **Money laundering activity is mostly sophisticated**
2. **Most legal businesses will not use EMT's as mean to do business**
3. **Highly connected clusters of EMT's can be flagged as suspicious**



We assume Money Laundering can be detected by finding outliers in the data

Filter 1 Assumptions

Regular Payments

1. **Money Laundering spending patterns are sophisticated**
2. **Similar amounts at regular intervals during the same timeframe**
3. **Note that some legitimate businesses use EMT's**

Regular Payments at similar time of day at equal intervals is potentially suspicious

Filter 1 Methodology

Accounts & Emails

1. **Average and Variance of all money flows**

2. **Average and Variance in Location**
   Calculate the center for an account/email, find the variance in the distance

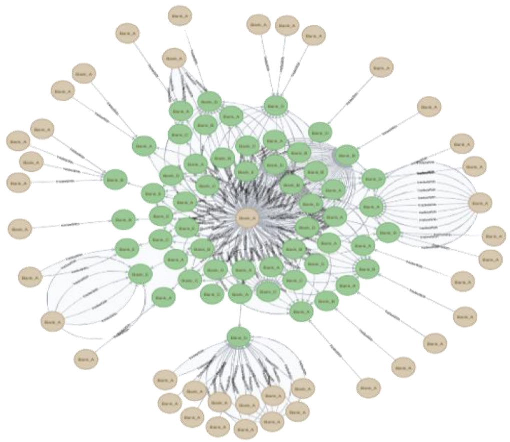
3. **Percentage of Late Transactions**
   Percent of total transactions that take place between 10 PM & 6 AM

	userID	timestamp	payeeEmail
0	044f9d391a27b59859fb3b274237671ff246bb9b69ae7d...	2018-06-01 19:02:54.026-0400	2fe5ce59f8fbac0fccaca974b9bb08ab4b2afe3d5703d1...
1	3ec92838d15518ea50355e7adfa01d470a7e49479c426d...	2018-06-01 05:37:11.480-0400	480e4ce89838a108880a7ba13475d9aa92e995476a86ea...
2	0f772a1e33aec1998c7a917e5cf67f30eeb485db693c57...	2018-06-01 00:44:01.691-0400	02fc4c55153e0f1645e8c3b24a0d225aa09a3e6806a652...

Using Accounts ID and Emails, We can start to build relationships and find outliers in the data

Filter 1 Methodology

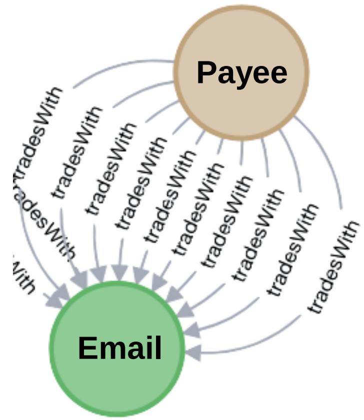
Clusters



Connected Network of Accounts

1. **Average and Variance of all money flows**
2. **Size of the Cluster**
   Total number of accounts and emails
3. **Average Connections Per Member**
   Total number of connections per account and email in the cluster

Specific Relationships



Account that pays the same person multiple times

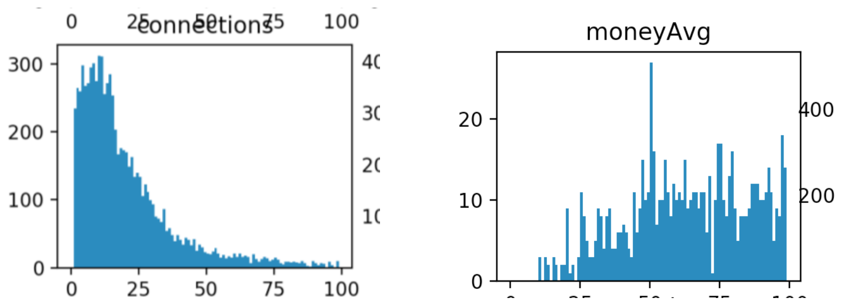
1. **Average and Variance of all money flows**
2. **Variance in Location**
   Find the variance in distance of all payments
3. **Percent of Late Transactions**
   Percent of transactions between 10 PM – 6 AM
4. **Average and Variance of Time Elapsed**
   Order the Payments by time and calculate the time between the EMTs

Scotiabank

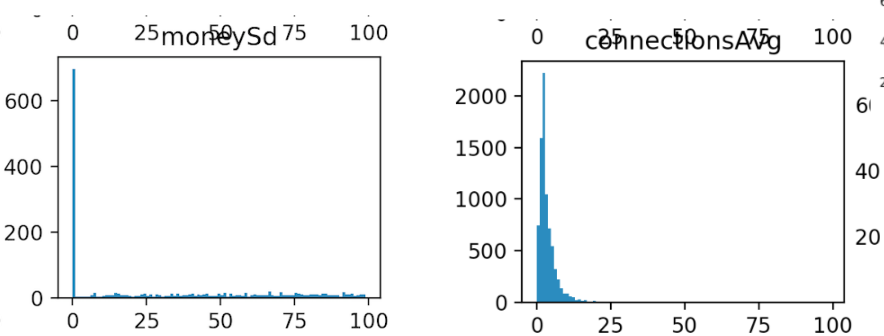
Filter 1 Methodology

The Distributions of Variables

Small samples: data appears to be normally distributed, truncated at 0.

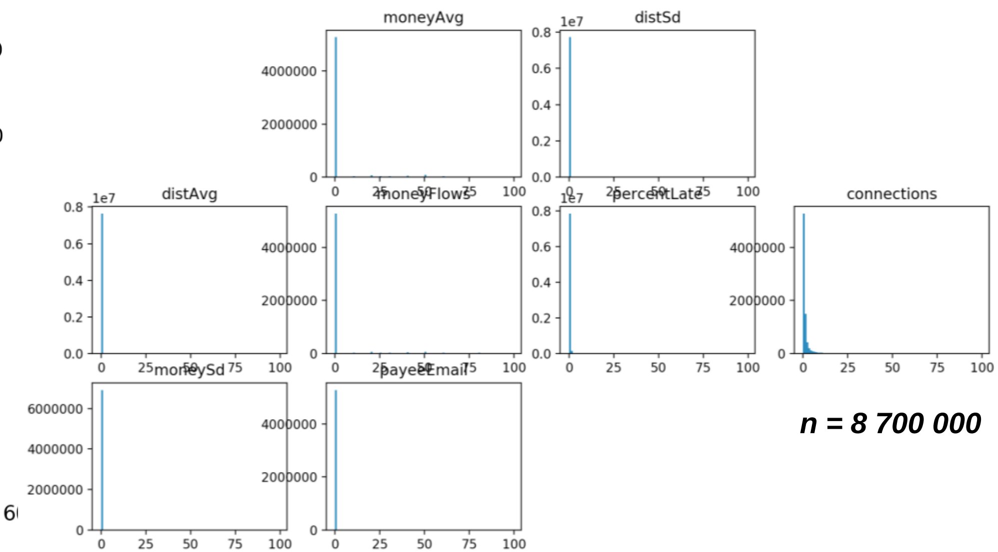


Standard deviations are distributed **chi-squared.**



***n* = 1 000**

Large samples: tight distributions about the mean



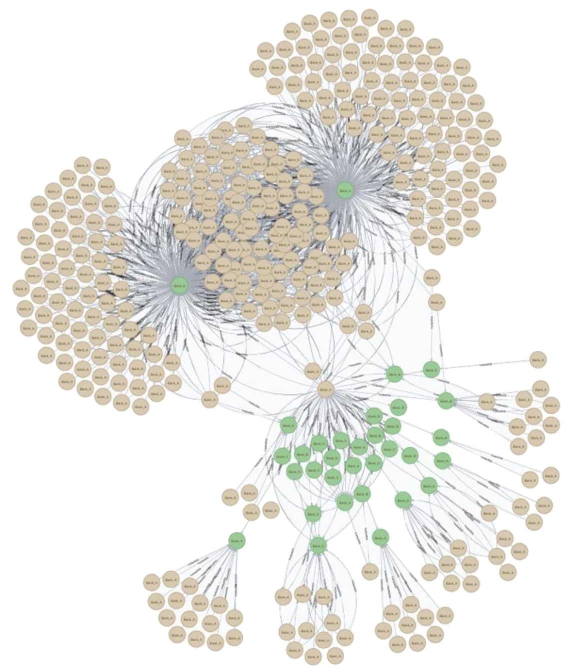
***n* = 8 700 000**

Assume all variables are normally distributed

Technology

Using Neo4j, we were able to visualize highly connected clusters that could represent potential criminal networks

Neo4j



A **Graphing Database**
Vertices = Accounts and Emails
Edges = EMT Relations

B **A Visualization tool for connections**
Visualize relationships and spatial data

C **Parameterization**
Help generate new variables

D **Community Detection Algorithm**
Detect Connected Clusters such as potential criminal networks

Technology

Community-Detection Algorithm

UnionFind

The Weakly Connected Components algorithm
Fully connected subgraphs.

```
neo4j> MATCH (net:network)
       RETURN net.nid, net.size
       ORDER BY net.size DESC
       LIMIT 5;
+-------------------+
| net.nid | net.size |
+-------------------+
| 1085650 | 2301367  |
| 1045727 | 119      |
| 3857614 | 105      |
| 3857179 | 92       |
| 3805945 | 90       |
+-------------------+
```

Huge Community Size:
2 out of 8 million

We are an average of *3.57* Facebook friends away from anyone in the world

Need an algorithm that creates clusters of highly cohesive members

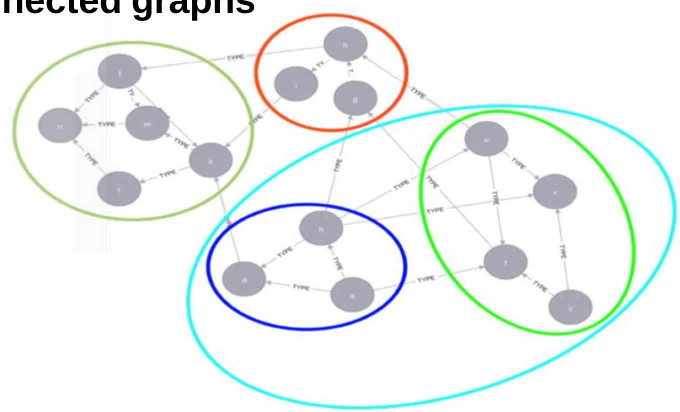
Louvian

Detects highly connected communities

```
+----------------------+
| net.nid | net.size   |
+----------------------+
| 1085650 | 2301367    |
| 1045727 | 119        |
| 3857614 | 105        |
| 3857179 | 92         |
| 3805945 | 90         |
+----------------------+
```

n.nid	n.size
20	57776
9	52398
58	45947
69	33520
77	29525

Distinguishes groups inside fully connected graphs



Scotiabank

Technology

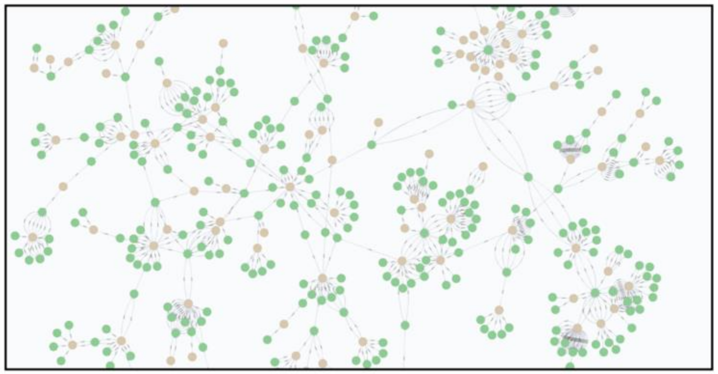
Community-Detection Algorithm

Which Edges and Weights?

A **using money_amount weight**

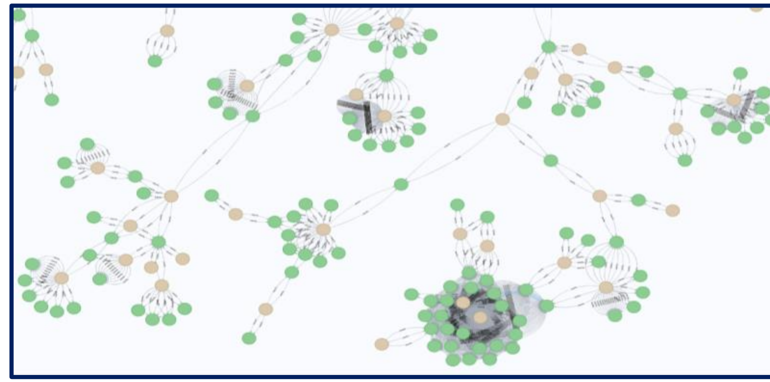
B **using number of connections**

Sparse Graph

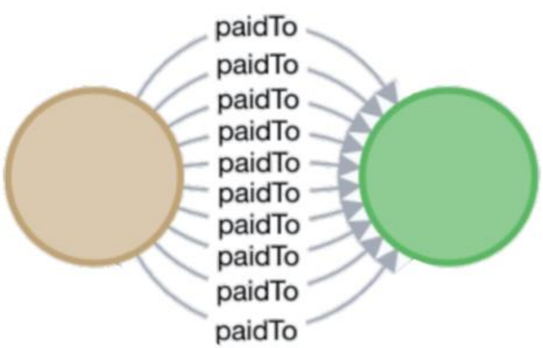


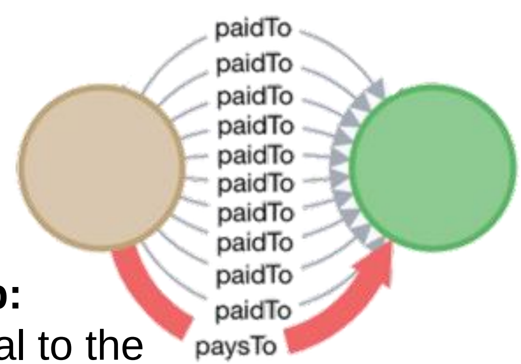
Random Cluster of Size 500

Dense Graph



Random Cluster of Size 500





New relationship:
Has attribute equal to the number of connections

Technology

Python

pandas for data reading and formatting values

```
f = pd.read_csv(filename)
f = f.dropna()
full = f.values
```

numpy for calculating mean, covariances and general matrix operations

```
mean = np.mean(X, axis=0)
cov = np.cov(X.T)
```

scipy for calculating the cdf of the multivariate normal

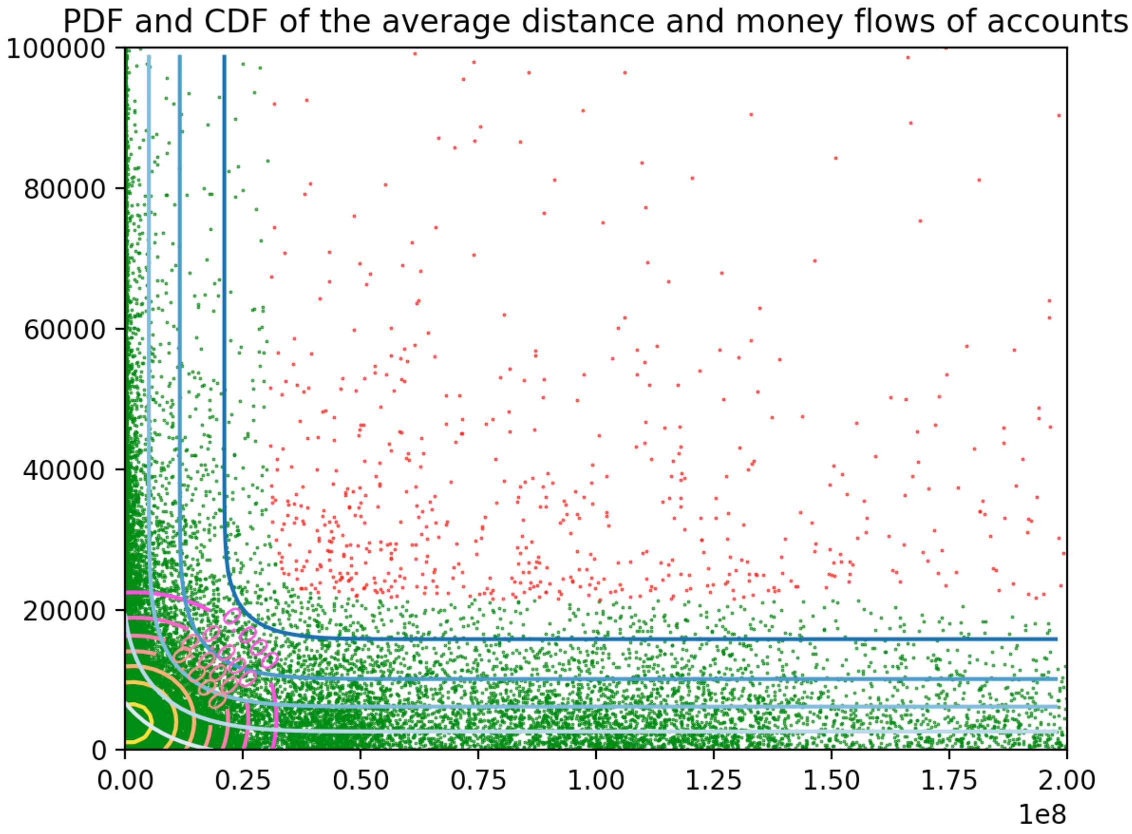
```
dist = multivariate_normal(mean=mean,
                           cov=cov,
                           allow_singular=True)
```

neo4j for running complex database queries

```
with driver.session() as session:
    paysTos = session.run("""
        MATCH (a)-[r:paysTo]->(e)
        RETURN a.userID AS userID, r, e.payeeEmail AS payeeEmail
    """)
```

Findings

PDF & CDF of a 2D Normal Distribution

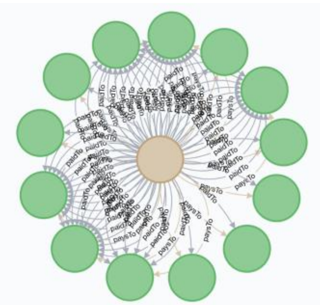


- Multivariate normal distribution PDF (yellow/pink) and CDF (blues) of 2 variables
- Red points are outliers with 5% probability, green are regular.
- Change the probability threshold to increase or decrease the classifications.
- Actual outlier detection will use higher dimensions

Scotiabank

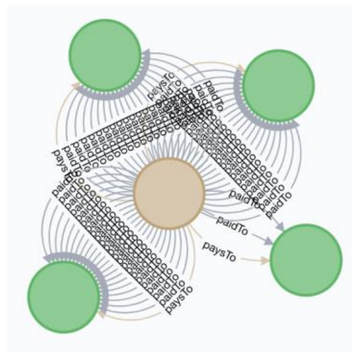
Findings

Findings - accounts



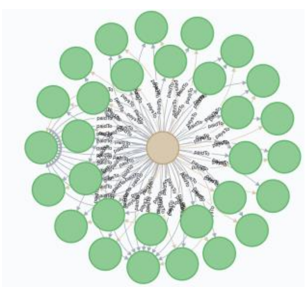
account hash:
1d1a9555214f4e9492d230caa
17bc3515768ef5b732f3ccffe5
0806ca6362631

probability: 0.59%



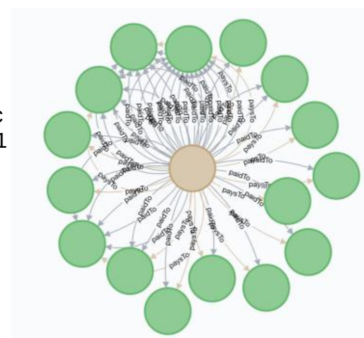
account hash:
fb17cd979aead85cc9b813d
52a1996342c9e909b4a5ef6
5223e9ce99fdbe2ed1

probability: 1.91%



account hash:
235cb0129e62c70eba9bfc3f56c
4a2c7071abe8bed135d32a3d11
ef30b25cf08

probability: 1.77%



account hash:
1ab775511e837b95908fe540a9
e006cf231f80e7e65b830dcd91c
7b77989848e

probability 0.68%

All accounts with less than 2.5% probability

- 50 – 70% late payments
- Average money flows of $72 450
- Average of 64 transfers
- Average 110 latitude/longitude units variance in locale

Scotiabank

Findings

Findings - accounts

account hash:
1d1a9555214f4e9492d230caa
17bc3515768ef5b732f3ccffe5
0806ca6362631

probability: 0.59%

account hash:
fb17cd979aead85cc9b813d
52a1996342c9e909b4a5ef6
5223e9ce99fdbe2ed1

probability: 1.91%

account hash:
235cb0129e62c70eba9bfc3f56c
4a2c7071abe8bed135d32a3d11
ef30b25cf08

probability: 1.77%

account hash:
1ab775511e837b95908fe540a9
e006cf231f80e7e65b830dcd91c
7b77989848e

probability 0.68%

All accounts with less than 2.5% probability

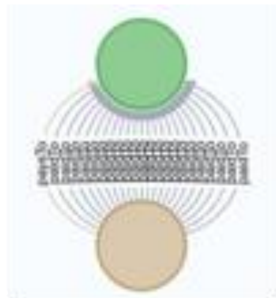
- 50 – 70% late payments
- Average money flows of $72 450
- Average of 64 transfers
- Average 110 latitude/longitude units variance in locale

Findings

Findings - emails

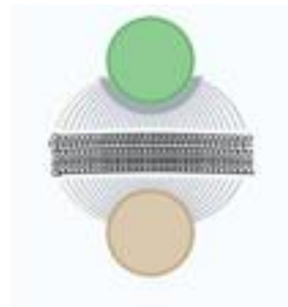
email hash:
35a9e259dd060044a257d0cdbf5
fa4eec16c043b3168728ecf7410f
fd8c47273

probability: 1.27%



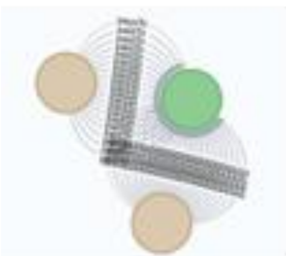
email hash:
e2ee14a7e9bc86f2f7e04cafe628
a1536fb047e66ec1660ce721c94
ccf75a171

probability: 1.43%



email hash:
6bd1080934f233cc5616a0c25be
feab18179df6c3f53093da1d1467
24818a11b

probability: 2.06%

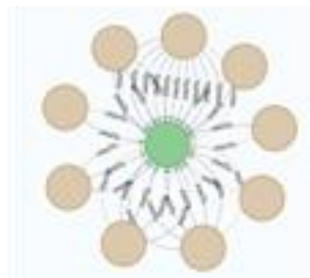


email hash:
6f6042c664c3c532d0efa64a677
9a27c300dc646a9e35285f1ccd5
2149d96e7e

probability: 2.46%

email hash:
d94622b650f2a5001d3d50220c2
33113440757b8b9d8ae1d40bb3
391b33f4725

probability: 1.91%



email hash:
f9ca58220a5796b4c07f2376024
3fbee24e415afc6e20f0ab11e5aa
6166b300a

probability: 2.46%

All emails with less than 2.5% probability

- 60 – 80% late payments
- Average money flows of $30 000
- Average of 37 transfers
- Average 106 latitude/longitude units variance in locale

Scotiabank

Findings

Findings - emails

All emails with less than 2.5% probability

- 60 – 80% late payments
- Average money flows of $30 000
- Average of 37 transfers
- Average 106 latitude/longitude units variance in locale

email hash:
35a9e259dd060044a257d0cdbf5
fa4eec16c043b3168728ecf7410f
fd8c47273

probability: 1.27%

email hash:
e2ee14a7e9bc86f2f7e04cafe628
a1536fb047e66ec1660ce721c94
ccf75a171

probability: 1.43%

email hash:
6bd1080934f233cc5616a0c25be
feab18179df6c3f53093da1d1467
24818a11b

probability: 2.06%

email hash:
6f6042c664c3c532d0efa64a677
9a27c300dc646a9e35285f1ccd5
2149d96e7e

probability: 2.46%

email hash:
d94622b650f2a5001d3d50220c2
33113440757b8b9d8ae1d40bb3
391b33f4725

probability: 1.91%

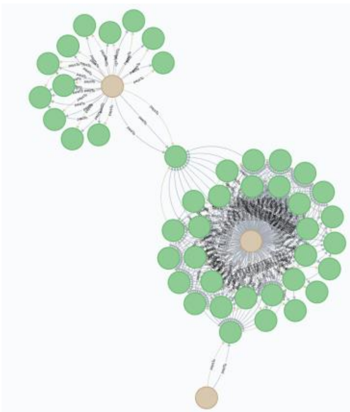
email hash:
f9ca58220a5796b4c07f2376024
3fbee24e415afc6e20f0ab11e5aa
6166b300a

probability: 2.46%

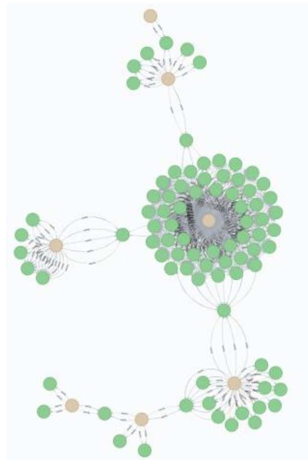
Scotiabank

Findings

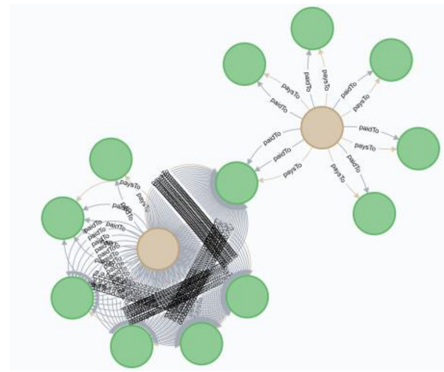
Findings - clusters



probability: 6.26%



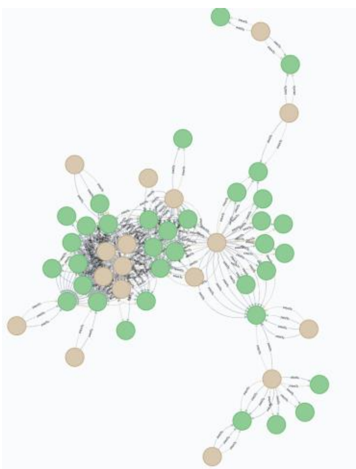
probability: 8.77%



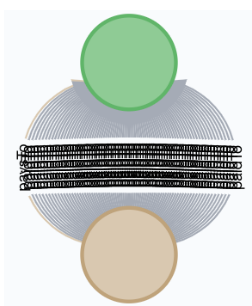
probability: 6.82%

All networks with less than 10% probability

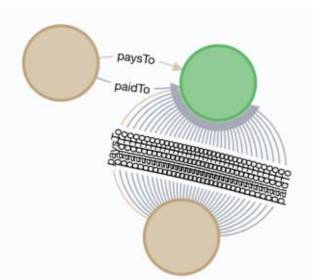
- Average money flows of $420 149
- Average of 14 transfers per member
- Average of 36 members
- Average of $3 249 per transaction



probability: 6.48%



probability: 0.95%



probability: 8.55%

probability: 3.30%

Scotiabank

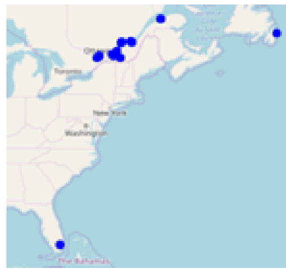
Findings

Findings - clusters

Montreal

probability: 6.26%

Quebec/
Ottawa

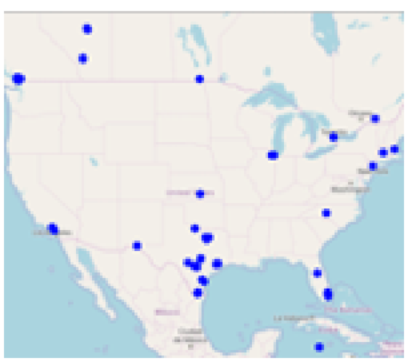


probability: 8.77%

Kingston

probability: 6.82%

North America

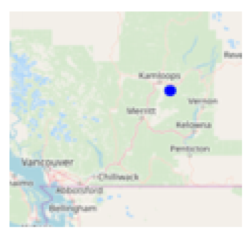


probability: 6.48%

Quebec

probability:
0.95%

Vancouver



probability: 8.55%

Montreal/
Ottawa

probability: 3.30%

All networks with less than 10% probability

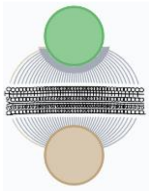
- Average money flows of $420 149
- Average of 14 transfers per member
- Average of 36 members
- Average of $3 249 per transaction

Scotiabank

Findings

Using data from these large networks, we can use the 2nd Filter to detect human trafficking

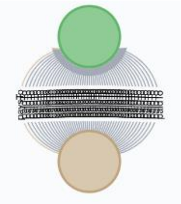
Findings – specific relationships



account hash:
a1c3a96f713576e8574b82d9c7bc951
9e2ef165dc44dda0ccb5cf24b26d595cf

email hash:
35a9e259dd060044a257d0cdbf5fa4ee
c16c043b3168728ecf7410ffd8c47273

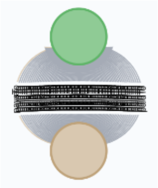
probability: 0.000196%



account hash:
6a2bbf6de61b1d820dcd09b0be4c5d1bb
78a3d3cbe219cc55d62bceacf0f93ab

email hash:
b820c32a62c9de58f544070a101208f82
677f0c38cc9c7642ee21a7390622998

probability: 0.000245%

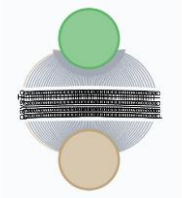


account hash:
be071d2017c20b420e640eaa02ee07ec
4c9087b0e6b05e4cc71b0d9e935c59e4

email hash:
c551e1c63224642f029b63f36e50f1bd75
fee4e1470b1fa7f56ba1a3681c1054

probability: 0.000052%

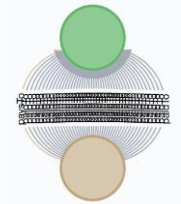
All relations with less than 0.0000025% probability



account hash:
29a4c50cb31763caa141d4b05a4fa288f
1dd5b1bac966bc0332c4a34c02bb61a

email hash:
f83acb632a5a1a00eb0ff7f7bad2ea36ffc
e0b61277e5980657ca22bbd4bbff9

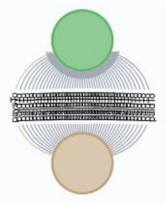
probability: 0.000039%



account hash:
b0677261368c9f3c2f988d74522c3d5249
44a50fc1b17e7963d5bc9ad5e39d89

email hash:
117c5ad60673bd61f0132201ea0f8bbf66
ff56dea6a9e31adb153cb6ca55da5a

probability: 0.000168%



account hash:
69b0cb86a858107a9d77dad52b55be281
20895b11cbc0db334182f9b7f6990b5

email hash:
60b0b17e4c72ae475a9ecdc823ddfb27f6
69d795d21c7c5851d10e326d2ebab1

probability: 0.000043%

- Average of 54 transfers
- 85-100% to late transfers

Scotiabank

Filter 2 Methodology

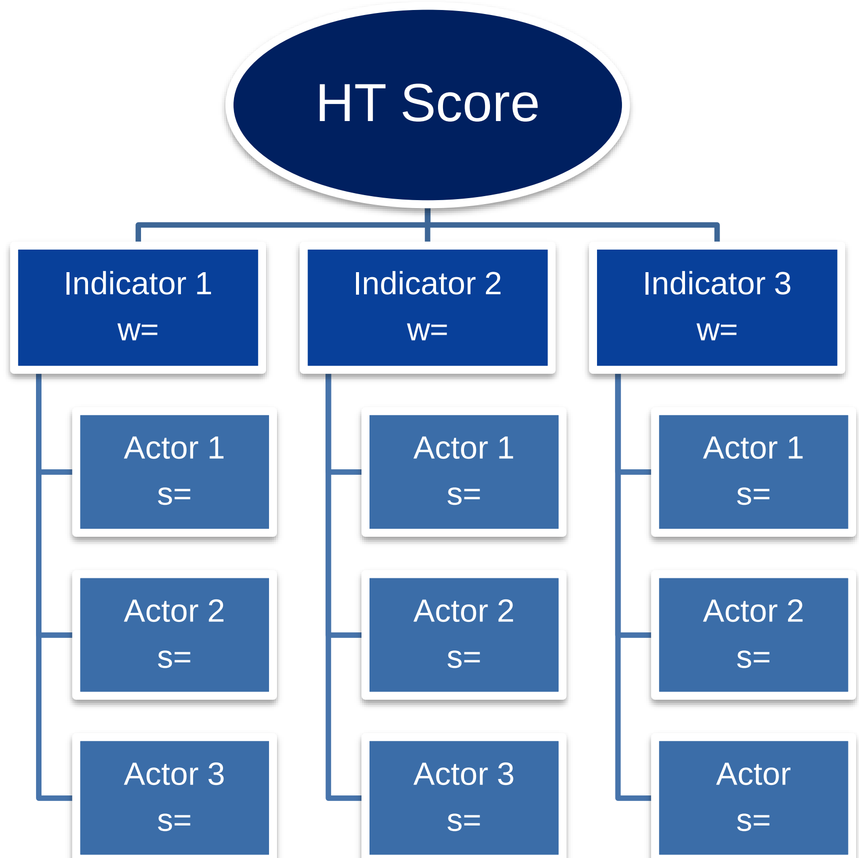
Human Trafficking (HT) Score

- Method of risk scoring

- Actor score (s) is a number value that represents an actor's performance under a specific indicator. Number values are between 0 (low) and 10 (high).

- The weight for each indicator (w) is a predetermined value that measures the correlation between the indicator and human trafficking. Assigned weights are between 0 and 1 and the sum of all indicators is 1.

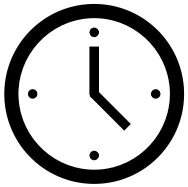
- The HT score is calculated by multiplying an individual's actor score for each indicator (s) by the weight for that particular indicator (w) and adding the products.



Filter 2 Indicator Examples

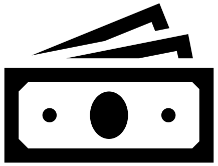
PercentLate

The percentage of total payments that an actor is making between 10pm and 6am; a higher percent would translate to a higher score (s).



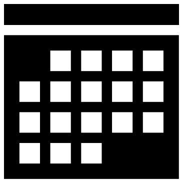
MoneyFlows

The total dollar amount of e-transfers sent by an actor; a higher relative amount would translate to a higher score (s).



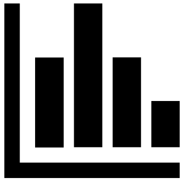
ElapsedAvg

The average elapsed time between e-transfers sent by an actor; a lower average time would translate to a higher score (s).



ElapsedSd

The standard deviation of the elapsed time between e-transfers; a lower standard deviation would translate to a higher score (s).



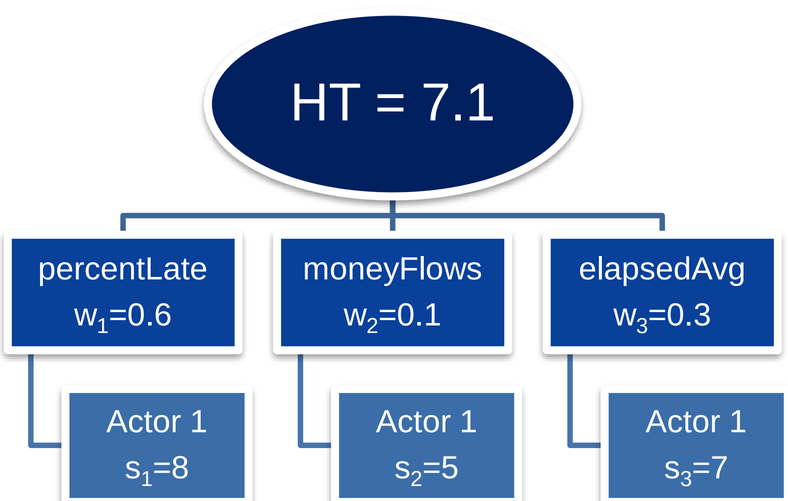
Case Example

Example

- Example on HT Score Calculation on a specific outlier (Actor) found during the Filter 1 process.

- The HT score will be a number that falls between 1 - 10; the closer to 10 the higher the chance that the behavior is classified as human trafficking

- The HT's score true strength lies in its ability to rank all actors from most probable to least probable of indicating human trafficking behavior which gives the AML teams the ability to focus on the most probable cases first.



$$\mathbf{HT = (s_1)(w_1) + (s_2)(w_2) + (s_3)(w_3) + \ldots + (s_n)(w_n)}$$

$$\mathbf{HT = (8)(0.6)+(5)(0.1)+(7)(0.3)}$$

<u>HT Score = 7.1</u>

Scotiabank

Limitations & Opportunities

Cannot Link UserID to Email

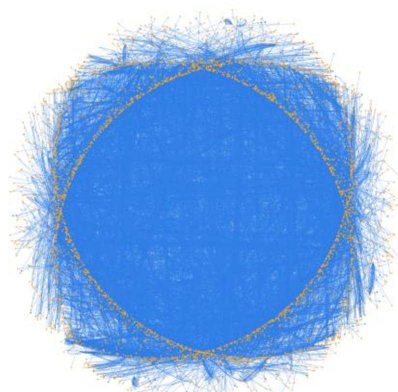
- Cannot link a sender UserID to the associated email, so chains of transactions are limited to one sender and one receiver

Supplementary Databases

- Do not have access to or cannot find relevant supplementary databases

Hardware Limitations

- Performing graph operations on this large dataset was very slow, even using a virtual computer with 32gb ram.



Data is Untargeted

- Traditional machine learning techniques of modelling require targets. An unsupervised task is much harder, especially for something as specific as human trafficking. Examples of EMTs from human trafficking may help create more accurate models.

```
0 rows available after 101212 ms, consumed after another 0 ms
```


Scotiabank

Conclusion

Using the two-step filtering system, we were able to find outliers that could represent money laundering or other illicit activities

Filter 1
Identifying Outliers

1. **Using Neo4J, we we're able to detect highly connected clusters of EMTs**
2. **Modelling accounts with regular payments & Sophisticated networks allows us to find outliers in the data**
3. **Visualize the connections and spatial amps in Neo4J to see if outliers seem suspicious, and tune hyper parameters to classify truly suspicious accounts**

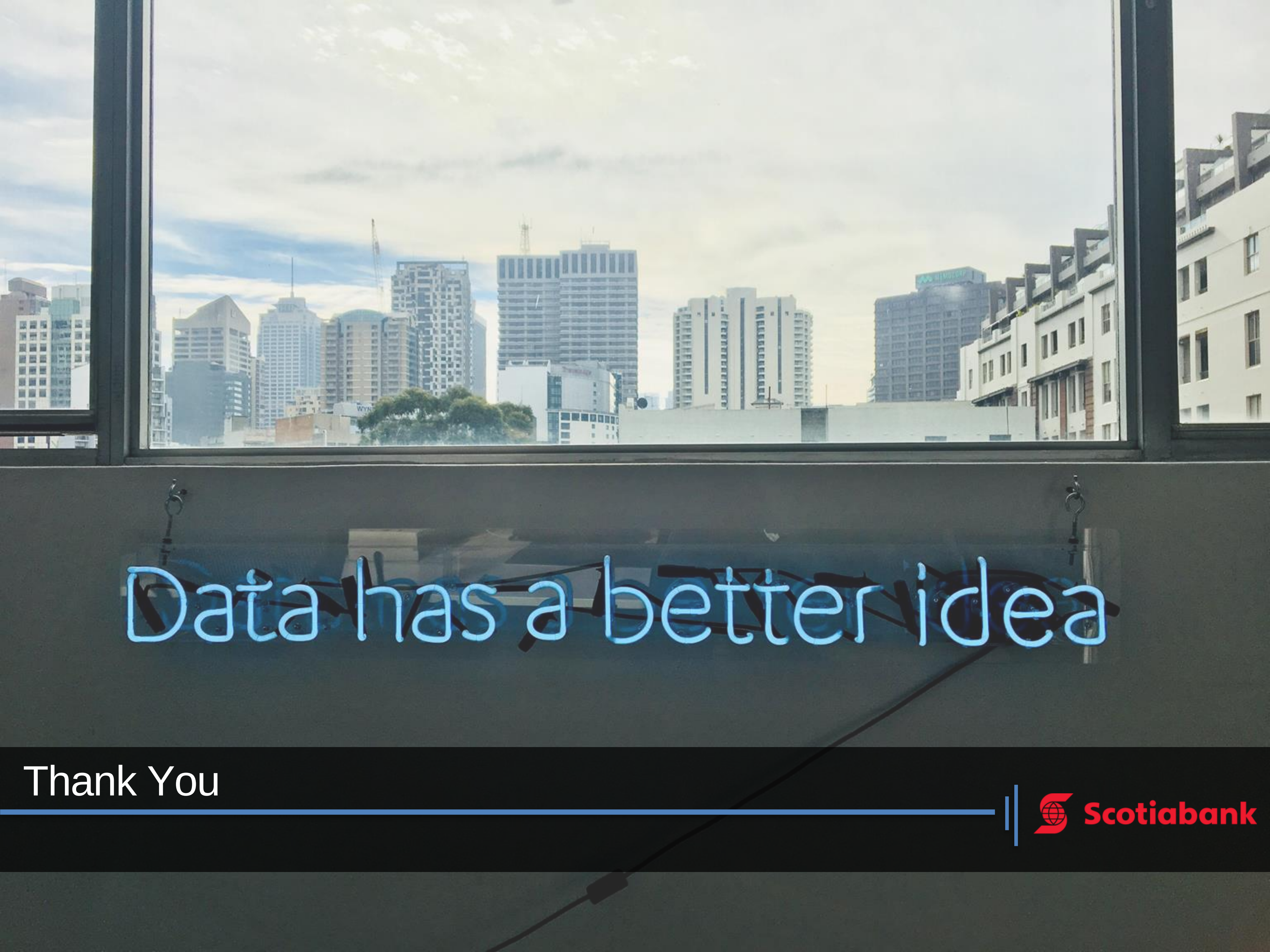
Filter 2
Human Trafficker (HT) Score

1. **The outlier data will then pass through the HT Scoring Calculator which uses indicators to detect potential human trafficking activity**
2. **The higher the score, the higher the probability that the account is linked to human trafficking activity**
3. **Our findings demonstrate that a score above 7 is very likely to be considered human trafficking activity**

Limitations

1. **Chains of transactions are limited to one sender and one receiver**
2. **The amount of data can prove difficult to process due to hardware capabilities**
3. **The data is untargeted which makes it more difficult to use Machine Learning**

Our outlier data is processed in order to specifically detect and rank potential Human Trafficking Activity



Thank You