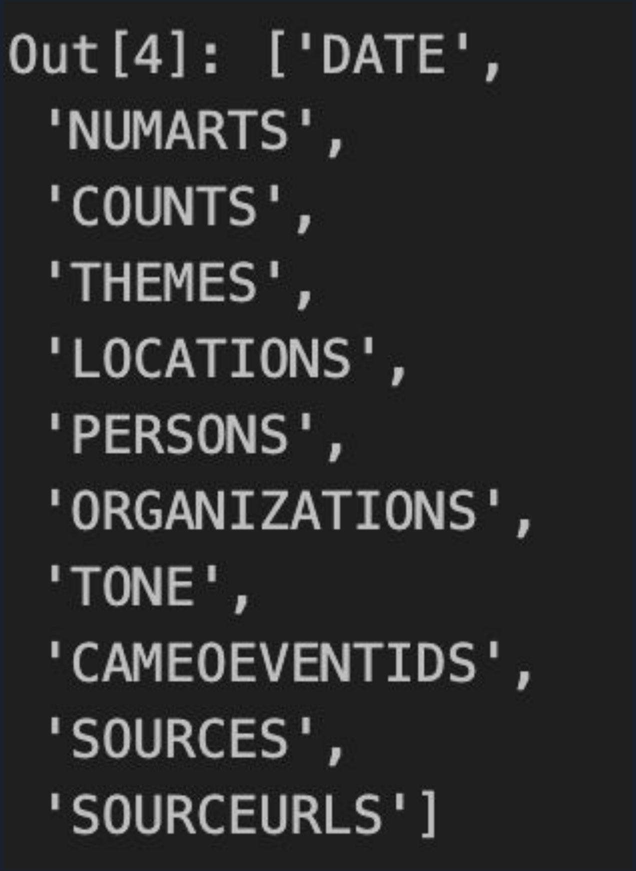
EMT-679 C Finals Project Presentation

-Ashwin Dhanasamy

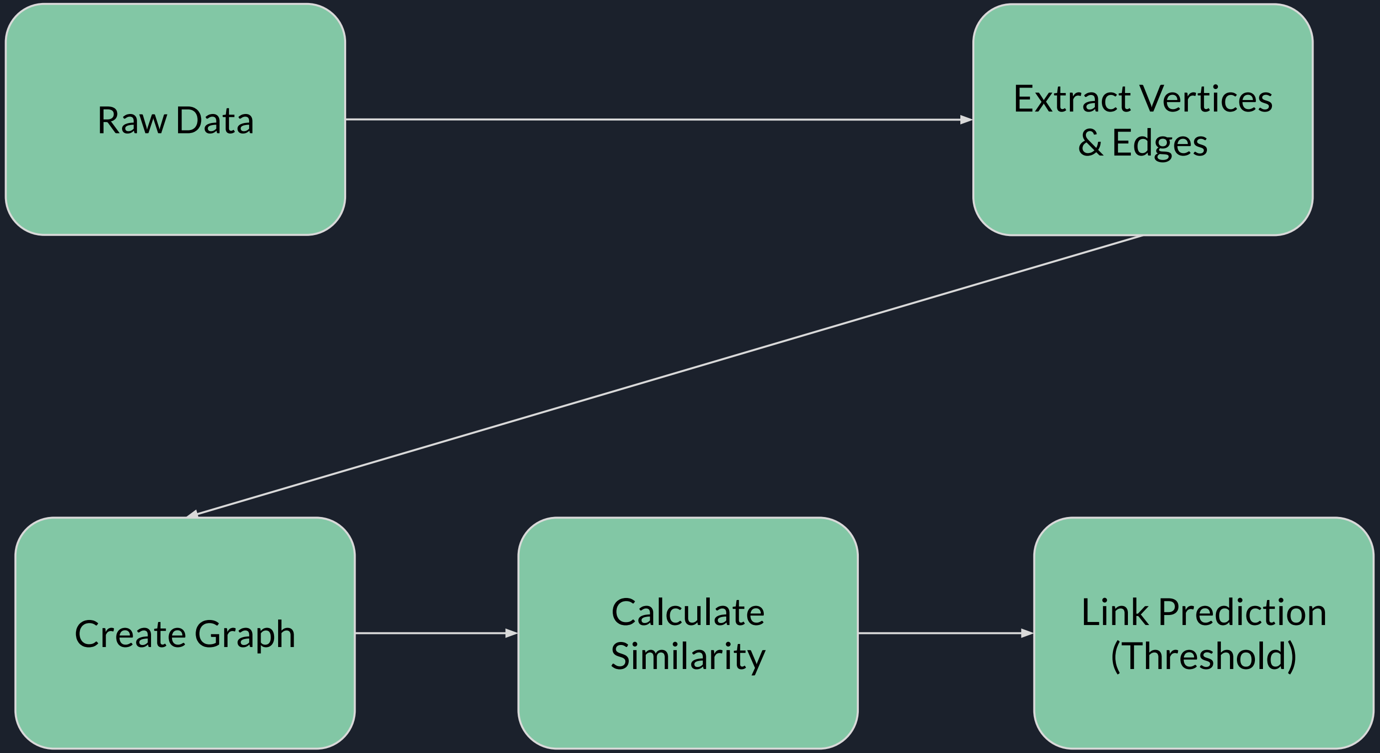
Objective

- Construct a knowledge graph containing **Persons** and **Organizations** as Vertices and **Themes** as Edges extracted from every GDELT event
- Perform link prediction between the vertices to find connections/relationships between unconnected vertices
- This can be useful to identify hidden Central nodes or relationships that may be of importance.

Dataset



```
Out[4]: ['DATE',
 'NUMARTS',
 'COUNTS',
 'THEMES',
 'LOCATIONS',
 'PERSONS',
 'ORGANIZATIONS',
 'TONE',
 'CAMEOEVENTIDS',
 'SOURCES',
 'SOURCEURLS']
```



Raw Data

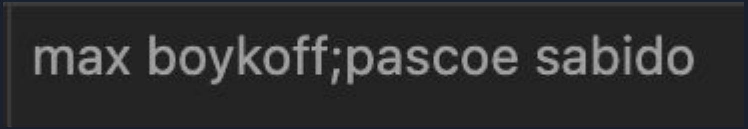
Extract Vertices & Edges

Create Graph

Calculate Similarity

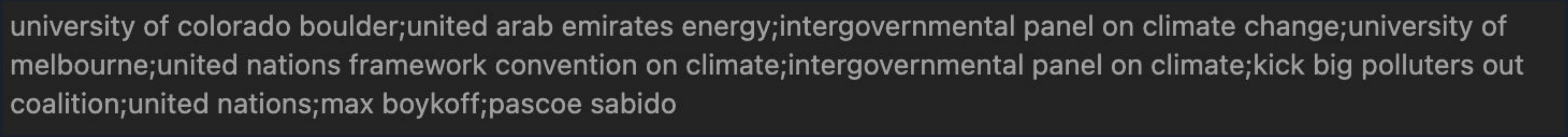
Link Prediction (Threshold)

PERSONS



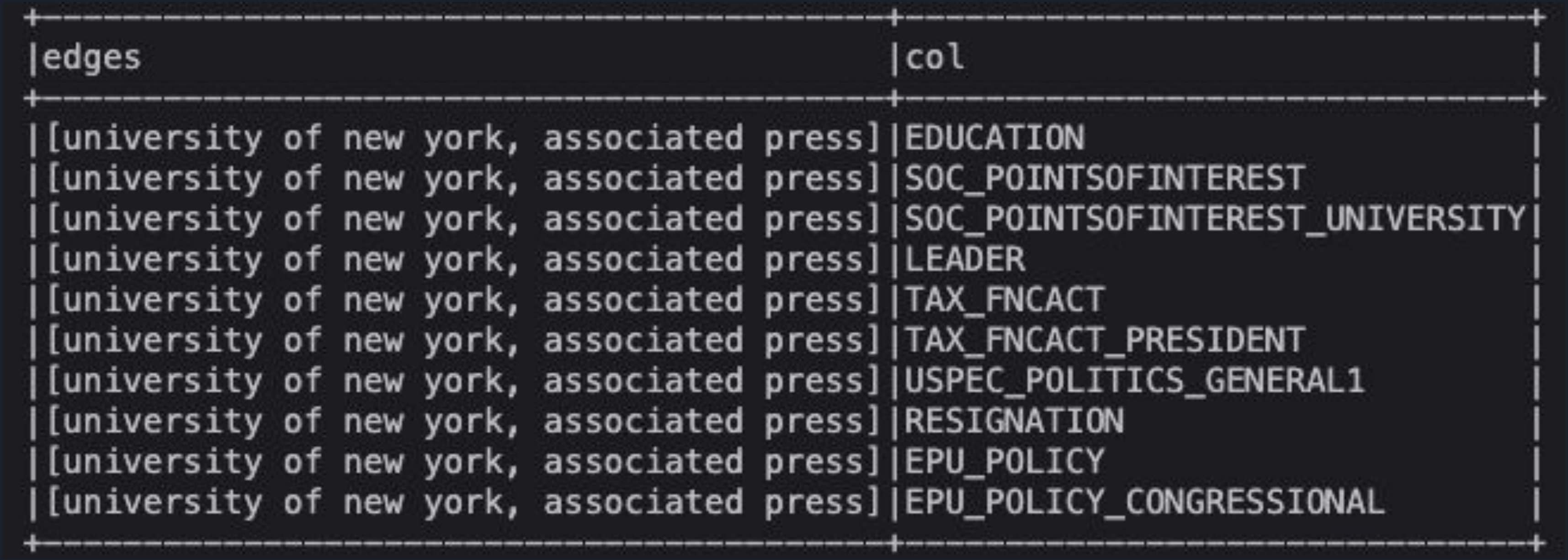
max boykoff;pascoe sabido

Merged PERSONS AND ORGANIZATIONS



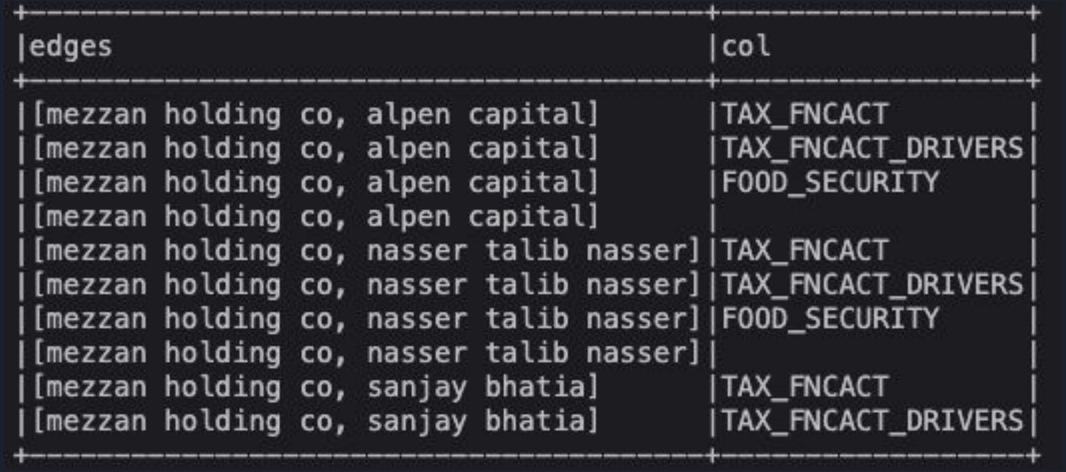
university of colorado boulder;united arab emirates energy;intergovernmental panel on climate change;university of melbourne;united nations framework convention on climate;intergovernmental panel on climate;kick big polluters out coalition;united nations;max boykoff;pascoe sabido

Created pairs and exploded pairs and themes from each row



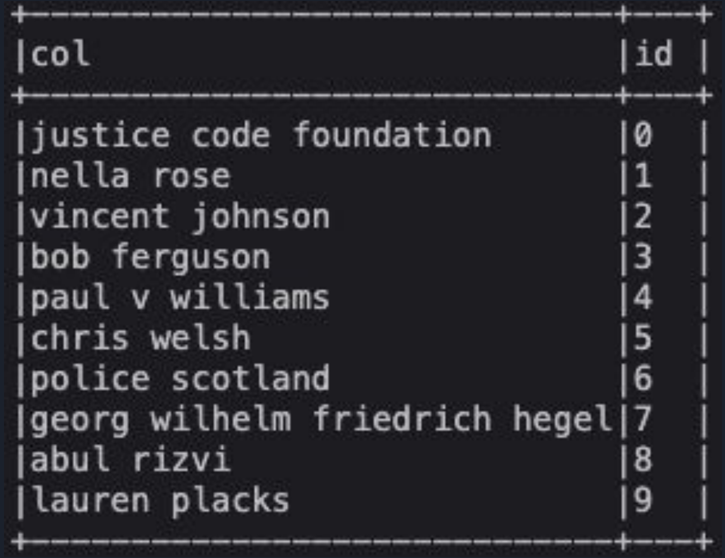
```
+------------------------------------------+-------------------------------+
|edges                                     |col                            |
+------------------------------------------+-------------------------------+
|[university of new york, associated press]|EDUCATION                      |
|[university of new york, associated press]|SOC_POINTSOFINTEREST           |
|[university of new york, associated press]|SOC_POINTSOFINTEREST_UNIVERSITY|
|[university of new york, associated press]|LEADER                         |
|[university of new york, associated press]|TAX_FNCACT                     |
|[university of new york, associated press]|TAX_FNCACT_PRESIDENT           |
|[university of new york, associated press]|USPEC_POLITICS_GENERAL1        |
|[university of new york, associated press]|RESIGNATION                    |
|[university of new york, associated press]|EPU_POLICY                     |
|[university of new york, associated press]|EPU_POLICY_CONGRESSIONAL       |
+------------------------------------------+-------------------------------+
```

EDGES



```
+----------------------------------------+------------------+
|edges                                   |col               |
+----------------------------------------+------------------+
|[mezzan holding co, alpen capital]      |TAX_FNCACT        |
|[mezzan holding co, alpen capital]      |TAX_FNCACT_DRIVERS|
|[mezzan holding co, alpen capital]      |FOOD_SECURITY     |
|[mezzan holding co, alpen capital]      |                  |
|[mezzan holding co, nasser talib nasser]|TAX_FNCACT        |
|[mezzan holding co, nasser talib nasser]|TAX_FNCACT_DRIVERS|
|[mezzan holding co, nasser talib nasser]|FOOD_SECURITY     |
|[mezzan holding co, nasser talib nasser]|                  |
|[mezzan holding co, sanjay bhatia]      |TAX_FNCACT        |
|[mezzan holding co, sanjay bhatia]      |TAX_FNCACT_DRIVERS|
+----------------------------------------+------------------+
```

VERTICES



```
+-----------------------------+----+
|col                          |id  |
+-----------------------------+----+
|justice code foundation      |0   |
|nella rose                   |1   |
|vincent johnson              |2   |
|bob ferguson                 |3   |
|paul v williams              |4   |
|chris welsh                  |5   |
|police scotland              |6   |
|georg wilhelm friedrich hegel|7   |
|abul rizvi                   |8   |
|lauren placks                |9   |
+-----------------------------+----+
```

N2, 3 worker nodes, 2 core, 8GB memory

Time Period - Number of rows from GDELT - Vertices - Edges

1 day - 58787 rows, 58454 vertices, 138433197 edges

10 days - 1143122 rows , 1371732 vertices, 2979300126 edges

21 days - 2028986 rows, 2090435 vertices, 5512139825 edges

30 days - 2992727 rows, 2830023 vertices , 8138436246 edges

CNGF

$$\text{Similarity}^{\text{CNGF}}(u,v)=\sum_{z\in\Gamma(u)\cap\Gamma(v)}\frac{|\phi(z)|}{\log d_z},$$

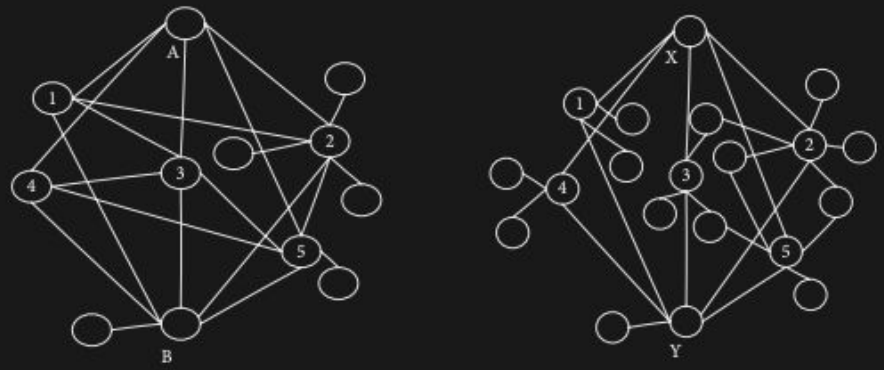


Figure 1: Two social network graphs with the same node degree.

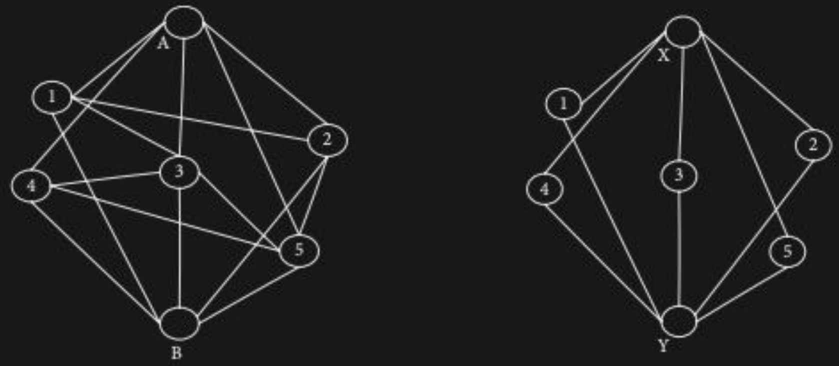
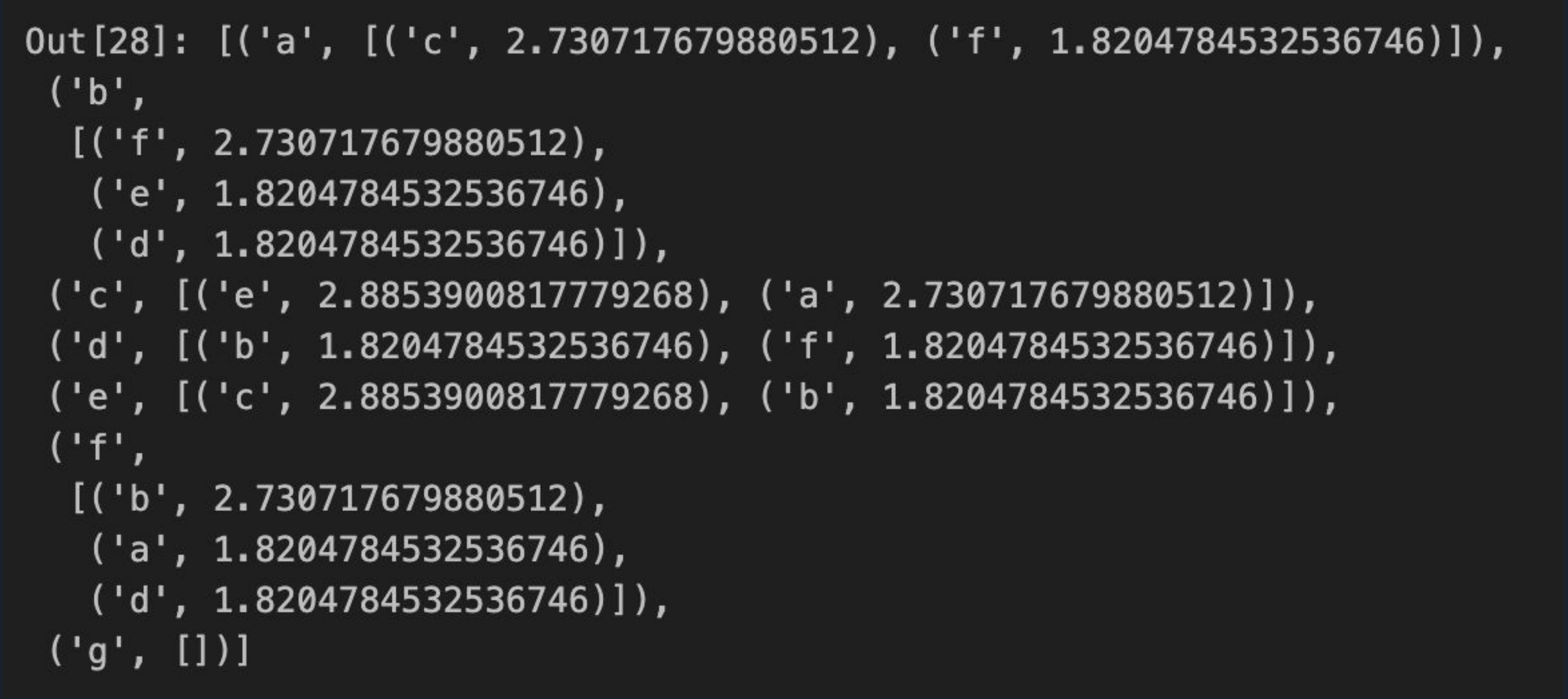


Figure 2: The extracted subgraph contains the prediction nodes and their common neighbor.

Sample Output



```
Out[28]: [('a', [('c', 2.730717679880512), ('f', 1.8204784532536746)]),
 ('b',
  [('f', 2.730717679880512),
   ('e', 1.8204784532536746),
   ('d', 1.8204784532536746)]),
 ('c', [('e', 2.8853900817779268), ('a', 2.730717679880512)]),
 ('d', [('b', 1.8204784532536746), ('f', 1.8204784532536746)]),
 ('e', [('c', 2.8853900817779268), ('b', 1.8204784532536746)]),
 ('f',
  [('b', 2.730717679880512),
   ('a', 1.8204784532536746),
   ('d', 1.8204784532536746)]),
 ('g', [])]
```

CHALLENGES FACED

- Installing the graphframes package to be used on the Data Proc cluster
  - Tried:
    - ssh into master node (VM) and installed the jar file
    - Tried to upload a custom script to be run when starting a cluster
  - Didn’t try:
    - Creating a custom dataproc image
    - Submitting a Spark Job

Databricks Free Trial - Trial Run

-The DBU's used after running the algorithm for vertex 0 was 11 and I was incurring charges on AWS to tune of $6.5 . At this point I interrupted the notebook to avoid being charged more than I could afford. Thus, my solution is still incomplete.

-The (interrupted) function spawned more than a 1500 jobs, each with only 1 stage per job(for 1 vertex). I think it can thus be inferred that my implemented solution is not parallelizable and does not effectively utilize sparks capabilities. Further analysis is needed to more effectively describe the problem.

FUTURE SCOPE

- Form a graph and perform link prediction
- Form graph for particular or group of persons, organizations, countries and theme (eg KILL, ARMED_CONFLICT, PANDEMIC etc ..) and perform link prediction on specific graphs to find hidden intricacies specific to a particular scope.