Suspicious events from experimental querying

This document will have experimental queries that detect suspicious events on the research project. The purpose of the document and the future meeting is to verify if the events spotted are suspicious and need further exploration. On the other hand, it is also chance to verify that the database is constructed correctly and none of these events are associated with a database disfunction.

The following experiments follow a stream of algorithms to detect communities within the data and later identify influential nodes within each cluster of community.
For our community detection part, we are using the algorithm "**Connected Components**", I have specifically chosen this algorithm because I have the recreation of the source code in java. In addition to this, as a centrality algorithm, I am using the algorithm "**Degree Centrality**", I specifically chosen this algorithm because the graph I am working with is unweighted and undirected. This means that all the relations are of equal weight and following the law of parsimony, I have chosen a simpler method to achieve the same results.

The pipeline of results is applied in Java through my source code, however the visualisation screenshots projected in this document are form the NEuler implementation. The centrality algorithm is only applied in the Java source code, and I have manually matched the abnormal behaviours identified in the knowledge graphs.
The Java results for the Connected Components algorithm provides identical results with the NEuler implementation.

Troubleshooting neo4j NOTE:

This note is for the individuals that have viewed the previous version of the document. After continuing my experiments, I have noticed that the results shown from my java implementation, NEuler and the neo4j browser did not match. This was because the browser was adding extra nodes and relations in the graph that were not considered in the Java code and NEuler procedures.

In NEuler there is a parameter called "***Rows to show***" which is the equivalent parameter in neo4j browser called "***:param limit(??)***". This parameter states to the browser how many rows from the resultant dataset it should project. Combining a large number with this parameter together with the neo4j browser setting parameter "***Connect nodes***", the browser was forced to add extra relations in the visualisation to connect the extra unneeded nodes in the knowledge graph.

For these reasons, the knowledge constructed below, only project the rows of interest and communities of larger than size 1.

Experiment goals and motivations:

The goal of this document and the various experiments that have taken place during this process is to identify suspicious structures that could be classified as audit fraud. Once these events are identified, the motivation is to understand how and why the algorithms can identify these events with the goal to further modify the original java source code and increase the information that we gain.

Experiment [1]

Parameters applied with the Algorithm

- Algorithm: Connected Components
- Label: ANY
- Relationship: ANNUAL_CHARACTERISE
- Community Node Limit: 18 (because largest cluster is of size 18)
- Rows to show: 61 (because we only had 61 communities of size >1)

From the above graph we can preview communities that form relations between nodes ***Areas*** and ***AnnualRisks***, the nodes are connected using the relation ***ANNUAL_CHARACTERISE***. At first glance, we can observe the communities within the red boundaries are significantly larger than the rest. With, all the larger clusters are owned by company **MARVEL**. We conclude that MARVEL has a larger number of risks associated with them when comparing with the other companies.

⇨ Further exploration needs to be applied here to verify if the larger number of risks are associated with fraudulent activities.

The above insight can be shown by the NEuler text results (the below screenshot).

Community	Size	Nodes
3830	18	Marvell Technology Group Ltd ×17, 1
3839	14	Marvell Technology Group Ltd ×13, 1
3826	12	Marvell Technology Group Ltd ×11, 1
3824	10	Marvell Technology Group Ltd ×9, 1
3846	9	Marvell Technology Group Ltd ×8, 1
3840	9	Marvell Technology Group Ltd ×8, 1
3838	9	Marvell Technology Group Ltd ×8, 1
3850	8	Marvell Technology Group Ltd ×7, 1
3833	8	Marvell Technology Group Ltd ×7, 1

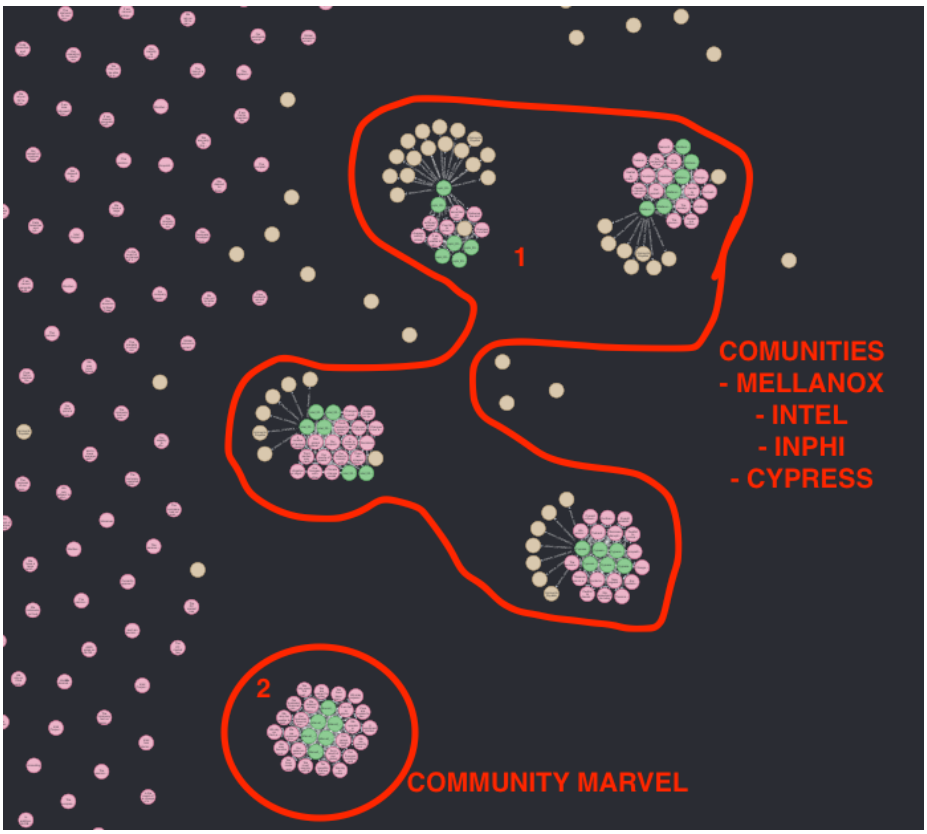
After applying the community detection algorithm, I am applying a centrality algorithm through java to find the influential nodes of the clusters that are of interest. The centrality algorithm being used across the whole document is the degree centrality algorithm. I implemented it from scratch in the Java interface and obtain text results.

Top 5 popular nodes of knowledge graph	
Node Label | Sub-label with company	*Degree power*
Node **Areas – Innovation** of **Marvel**	17
Node **Areas – Acquisition** of **Marvel**	13
Node **Areas – Legal Issues** of **Marvel**	11
Node **Areas – Cashflow** of **Marvel**	9
Node **Areas – Customers** && **Areas – Ownership** && **Areas – Personnel** of **Marvel**	8

Experiment [2]

Parameters applied with the Algorithm

- Algorithm: Connected Components
- Label: ANY
- Relationship: [DEFINE_ANNUAL_RISKS, ANNUAL_CONTAINES_DATAPOINT]
- Community Node Limit: 30 (because largest cluster is of size 30)
- Rows to show: 250 (because all the communities of size >1 were captured)

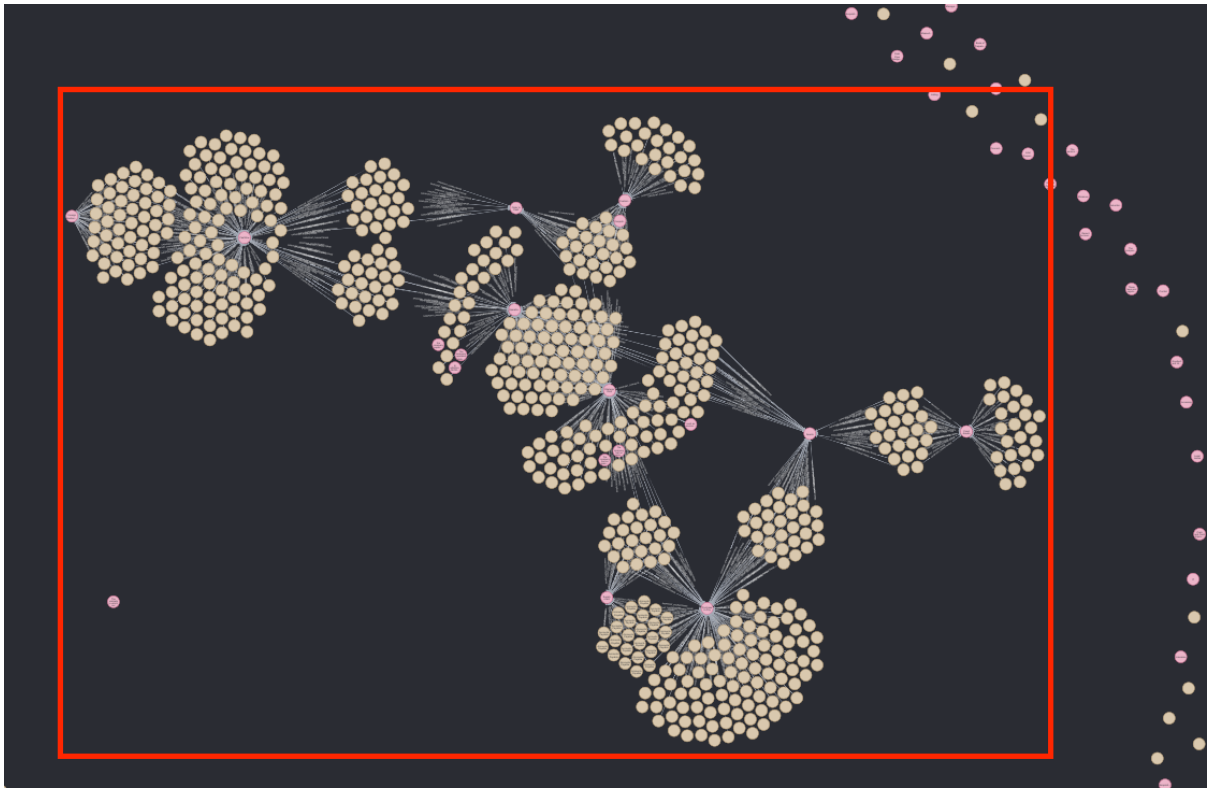


From the above results we can detect 5 communities, but in reality there are 6 companies in the dataset. This means we expected to have 6 community clusters. The company that is failing to construct a community is company ***AMD***.

From cluster [1] we can observe that companies ***Intel***, ***Inphi***, ***Mellanox***, ***Cypress*** all follow the same pattern of structure. Annual Report (green node) related to Annual Risks (pink node) and to Datapoint (brown node).

On the other hand we can observe that cluster [2] which involves company ***Marvel*** does not follow the same pattern. ***Marvel's*** Annual Report do not connect with Datapoint nodes.

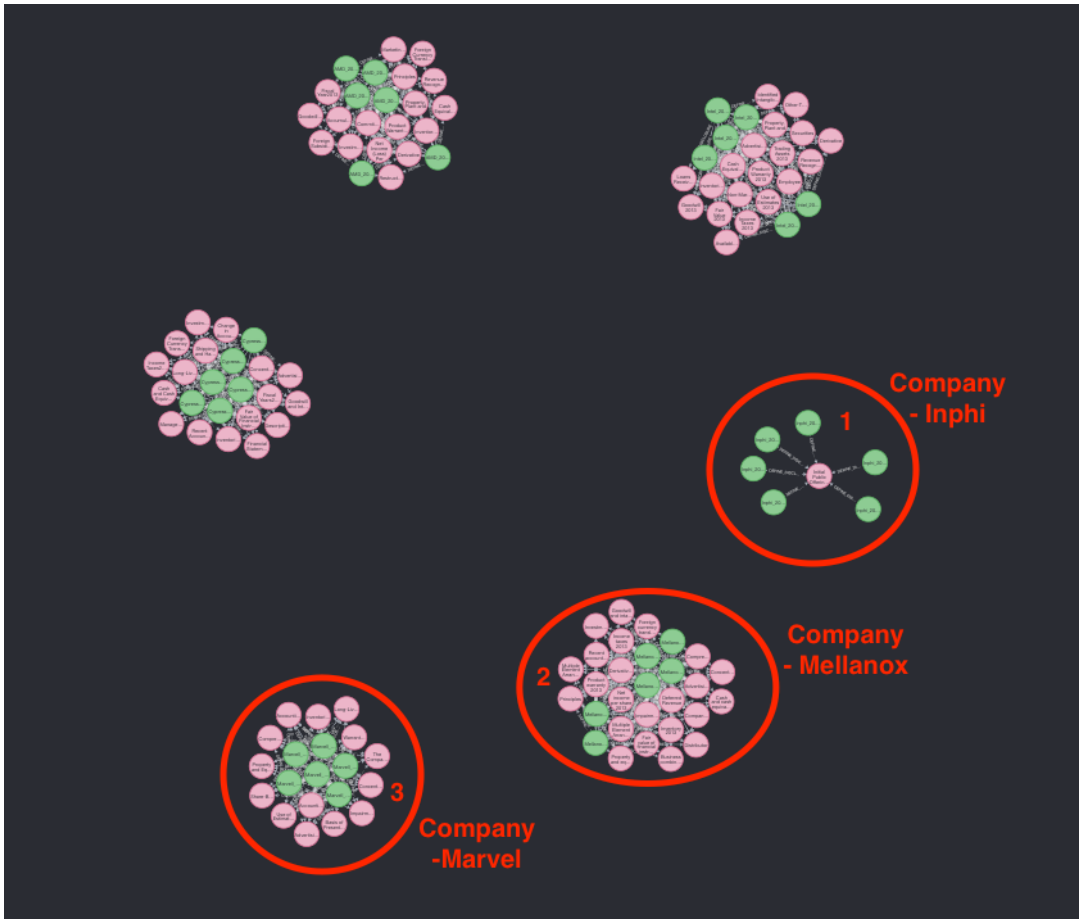
- ⇨ We again can observe abnormal behaviour associated with ***Marvel's*** Annual Risks nodes.
- ⇨ Company ***AMD*** does not form a cluster of community. While on the other hand if we increase the value of parameter "**Rows to show**", new clusters are formed from company ***AMD***, but they do not feature any Annual Reports in the community like the other companies. You can preview this in the below screenshot.



Experiment [3]

Parameters applied with the Algorithm

- Algorithm: Connected Components
- Label: ANY
- Relationship: DEFINE_DISCLOSURE
- Community Node Limit: 29 (because largest cluster is of size 29)
- Rows to show: 6 (all communities of size >1 captured)



From the above screenshot we can observe that all clusters are of different size. However, the structure of the community is the same. An Annual Report (green node) relates to a Disclosure (pink node)

Disclosure: An accounting policy disclosure helps to prevent loss. It also helps in preventing the misuse of assets.

We can observe that cluster [1] which is owned by company ***Inphi***, has only one Disclosure across all the Annual Report's.

⇨ Is having only one disclosure appropriate? Or is it a way to hide fraudulent activities? Additionally, we can observe cluster [3] which is owned by ***Marvel*** does not have largest size of disclosures associated with them. However, since they have a lot more risks associated with them, it is be assumed they should have the most disclosures.

⇨ Help of the specialist *Auditor* and *Knowledge Engineer* is needed to verify if this is a valid assumption. If so, then ***Marvel*** has something suspicious associated with their Risks and Disclosures.

After applying the second part of the pipeline (Centrality algorithm) we gathered the most popular nodes of interest.

Top 5 popular nodes of knowledge graph	
Node Label | Sub-label with company	*Degree power*
Node **Annual Report** of **Mellanox** with name **Mellanox_2014**	23
Node **Annual Report** of **Mellanox** with name **Mellanox_2015**	23
Node **Annual Report** of **Mellanox** with name **Mellanox_2016**	23
Node **Annual Report** of **Mellanox** with name **Mellanox_2017**	23

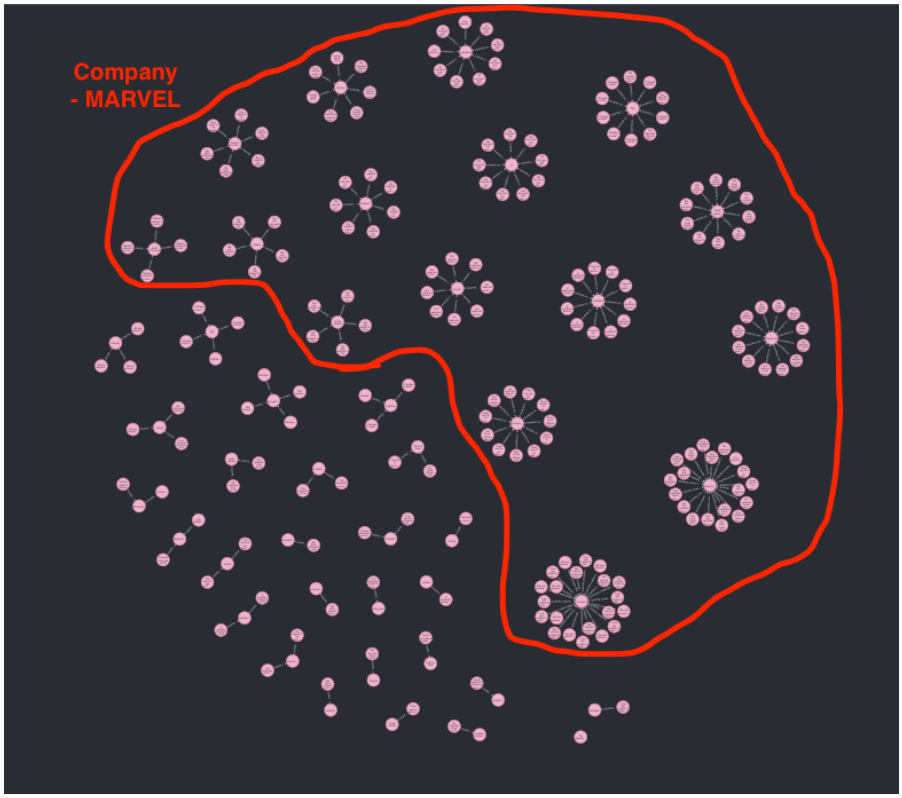
However, from the above knowledge graph we are also interested in community owned by Inphi. The community shows anomalous behaviour since it only has one disclosure, this needs further investigation.

Node Label | Sub-label with company	*Degree power*
Node **Disclosure – Acquisition** of **Inphi** with year **2013**	6

Experiment [4]

Parameters applied with the Algorithm

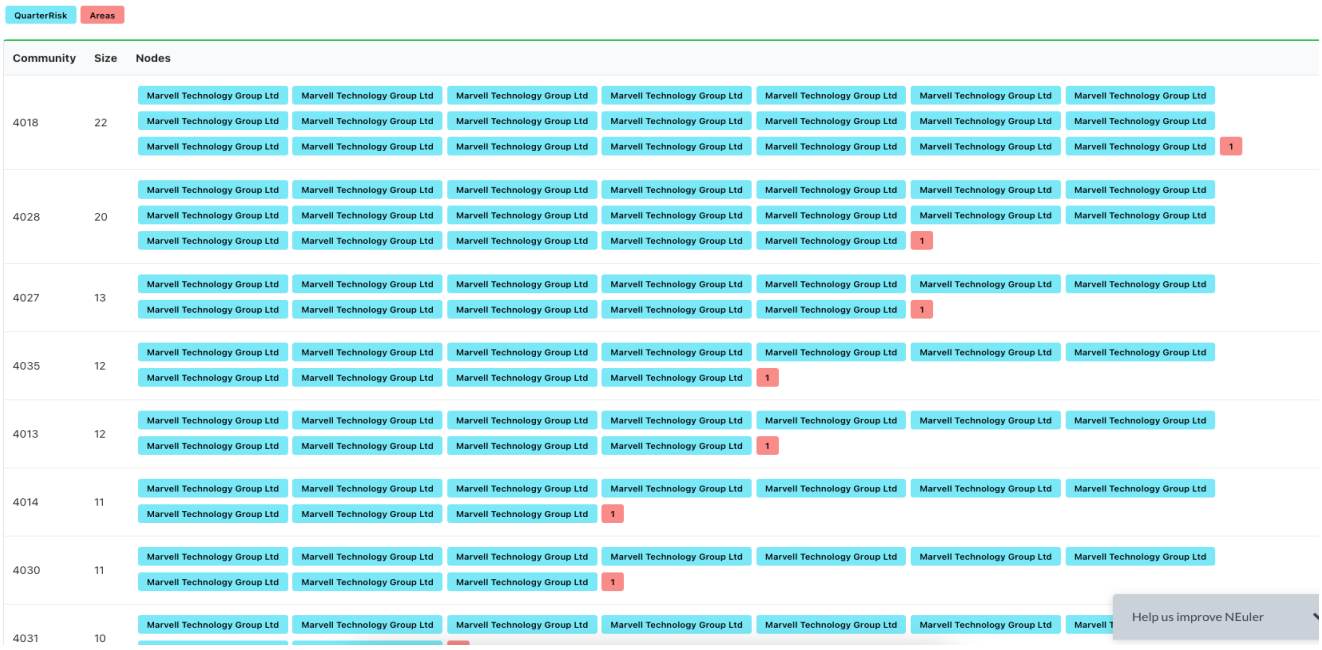
- Algorithm: Connected Components
- Label: ANY
- Relationship: QUARTER_CHARACTERISE
- Community Node Limit: 22 (because largest cluster is of size 22)
- Rows to show: 42 (all clusters of size >1 are captured)



Similar with experiment [1], we can observe that company ***Marvel*** have more risks associated with them when comparing with the other companies. However, this is something we expected. Since the annual risks are more, we expect the quarter risks to follow the same pattern.

⇨ Why does company ***Marvel*** have a lot more risks associated with them?

This can confirmed in a textual representation as well with the below screenshot.



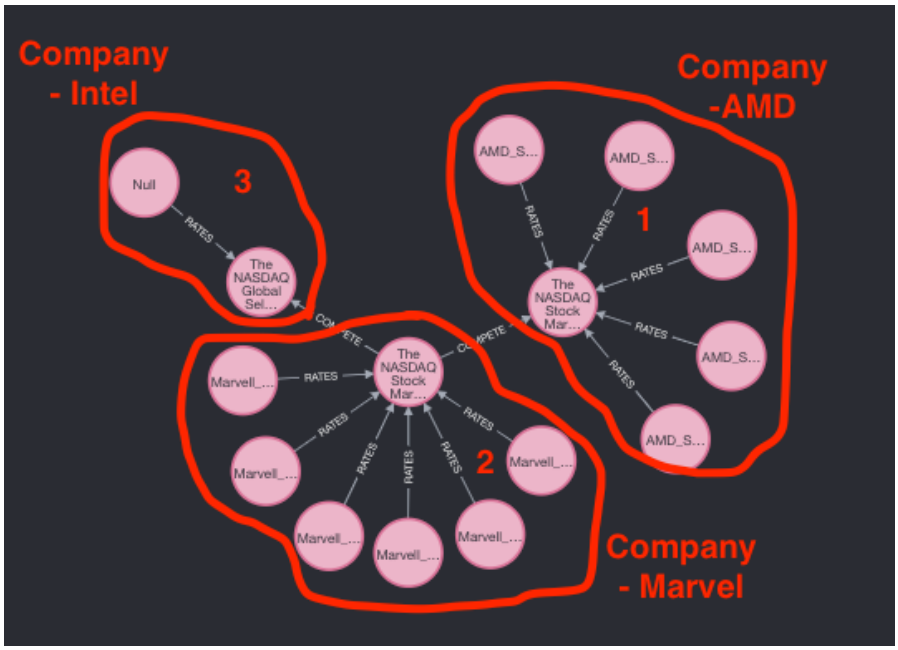
After applying the second part of the pipeline (Centrality algorithm) we gathered the most popular nodes of interest.

Top 5 popular nodes of knowledge graph	
Node Label | Sub-label with company	*Degree power*
Node **Areas – Innovation** of **Marvel**	21
Node **Areas – Acquisition** of **Marvel**	19
Node **Areas - Personnel** of **Marvel**	12
Node **Areas - Cashflow** of **Marvel**	11
Node **Areas – Ownership** of **Marvel**	11

Experiment [6]

Parameters applied with the Algorithm

- Algorithm: Connected Components
- Label: ANY
- Relationship: RATES
- Community Node Limit: 7 (because largest cluster is of size 7)
- Rows to show: 3 (only 3 clusters of size >1 exist)



All thought we have 3 communities; we can observe that only two are rated with values while the other has rated value of null.

Only companies ***Marvel*** and ***AMD*** have external resources rating them and reporting them.

- ⇨ Why do we have external resources only on these two companies?
- ⇨ Are they already suspects of fraudulent activities?

After applying the second part of the pipeline (Centrality algorithm) we gathered the most popular nodes of interest.

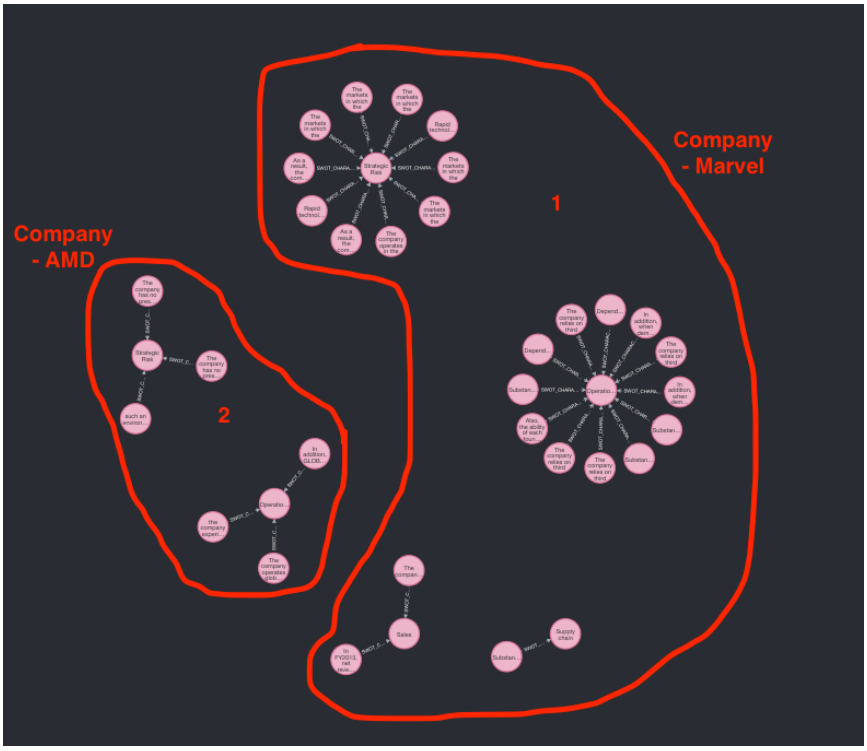
Top 2 popular nodes of knowledge graph	
Node Label | Sub-label with company	*Degree power*
Node **Company** of **Marvel** with UId 1	6
Node **Company** of **AMD** with UId 2	5

Only two nodes have degree power larger > 1 however in this case it does not offer any extra information.

Experiment [7]

Parameters applied with the Algorithm

- Algorithm: Connected Components
- Label: ANY
- Relationship: SWOT_CHARACTERISE
- Community Node Limit: 13 (because largest cluster is of size 13)
- Rows to show: 6 (only 6 clusters with size >1)



We have 6 communities formed using the above parameters however, 4 of them are owned by company ***Marvel*** and 2 are owned by company ***AMD***. This was expected since there exist Swot Reports only for these two companies.

We can observe that even the external resources are identifying a lot more risks to be associated with company ***Marvel***.

⇨ Is ***Marvel*** applying fraudulent activities through the large number of risks?

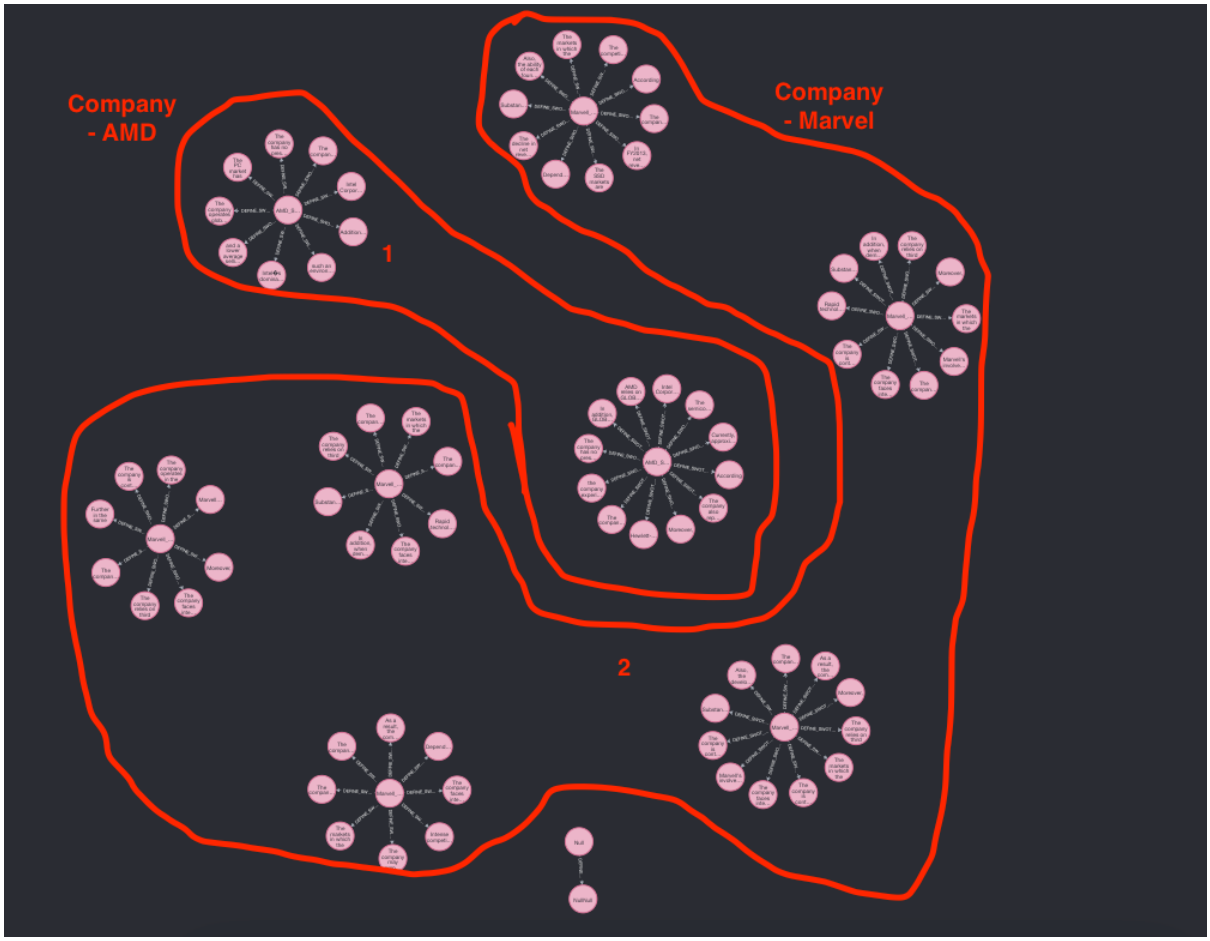
After applying the second part of the pipeline (Centrality algorithm) we gathered the most popular nodes of interest.

Top 5 popular nodes of knowledge graph	
Node Label | Sub-label with company	*Degree power*
Node **Areas – Operations** of **Marvel**	12
Node **Areas – Strategic Risk** of **Marvel**	10
Node **Areas – Strategic Risk** of **AMD**	3
Node **Areas - Operations** of **AMD**	3
Node **Areas – Sales** of **Marvel**	2

Experiment [8]

Parameters applied with the Algorithm

- Algorithm: Connected Components
- Label: ANY
- Relationship: DEFINE_SWOT_RISKS
- Community Node Limit: 13 (because largest cluster is of size 13)
- Rows to show: 9 (because only 9 communities with size >1 exist)



We can observe again that risks through Swot Reports are only associated with company ***AMD*** [1] and company ***Marvel*** [2].

Again we can see that ***Marvel*** has more risks associated with them however the largest community in terms of size is owned by ***AMD***.

After applying the second part of the pipeline (Centrality algorithm) we gathered the most popular nodes of interest.

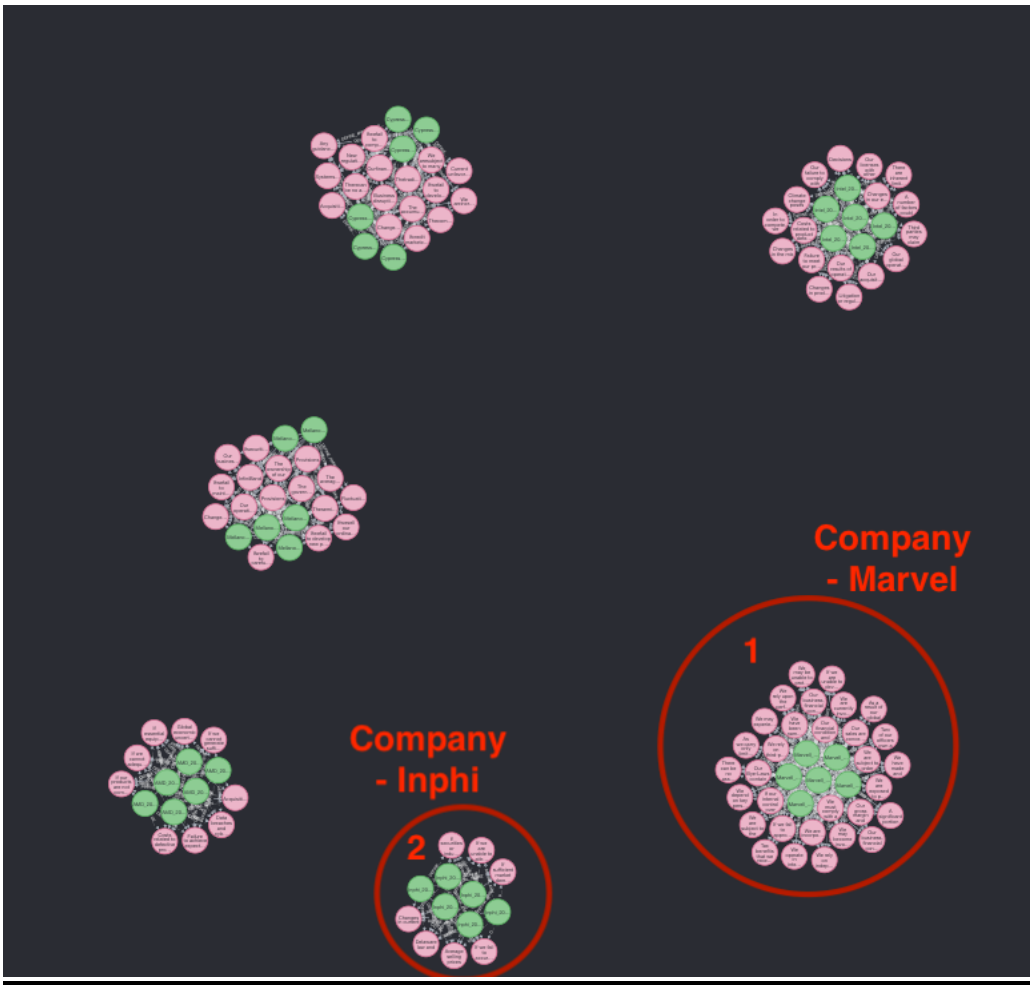
Top 5 popular nodes of knowledge graph	
Node Label | Sub-label with company	*Degree power*
Node **Swot Report** of **AMD** with UId "AMD_SWOT_2013"	12
Node **Swot Report** of **Marvel** with UId "Marvell_SWOT_2015"	11
Node **Swot Report** of **Marvel** with UId "Marvell_SWOT_2016	10

Node **Swot Report** of **Marvel** with UId “Marvel_SWOT_2013”	10
Node **Swot Report** of **AMD** with UId “AMD_SWOT_2014”	9

Experiment [9]

Parameters applied with the Algorithm

- Algorithm: Connected Components
- Label: ANY
- Relationship: DEFINE_ANNUAL_RISKS
- Community Node Limit: 37 (because largest cluster is of size 37)
- Rows to show: 6 (because only 6 communities with size >1 exist)



We can again observe that company ***Marvel*** has a larger number of risks associated with them. In addition to this, we can see company ***Inphi*** has a smaller number of risks associated with them when compared with the other companies.

⇨ Could the difference in the number of risks be associated to a fraudulent activity?

After applying the second part of the pipeline (Centrality algorithm) we gathered the most popular nodes of interest.

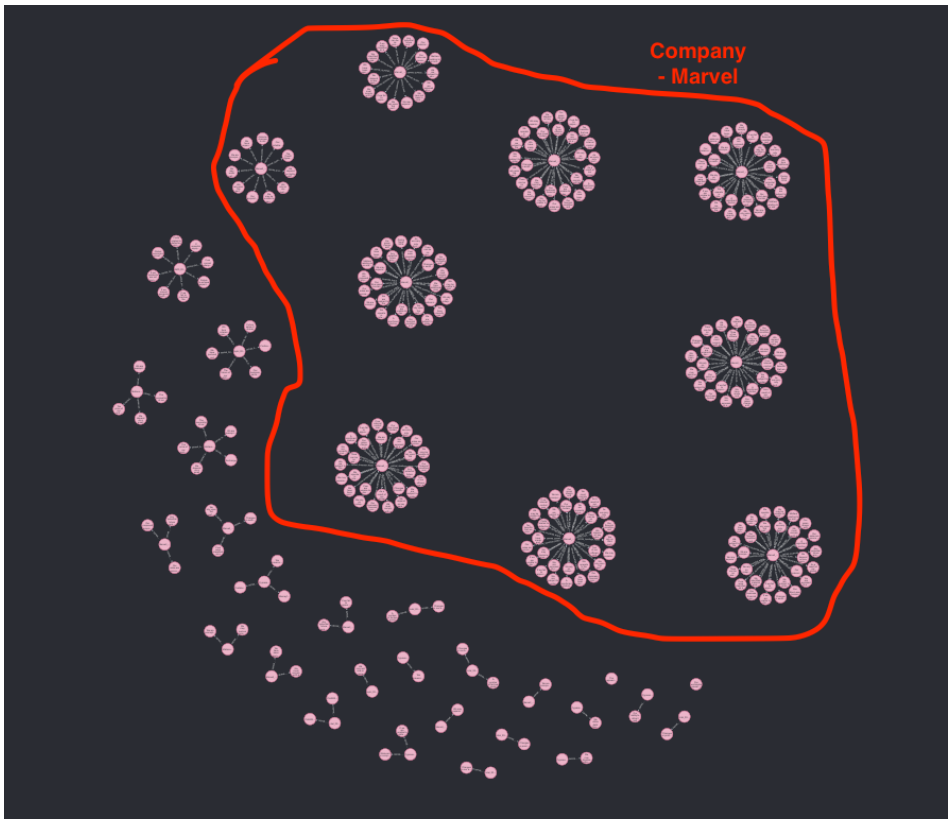
Top 5 popular nodes of knowledge graph	
Node Label | Sub-label with company	*Degree power*
Node **Annual Report** of **Marvel** with UId "Marvell_20182013"	31

Node **Annual Report** of **Marvel** with UId “Marvell_20132013”	31
Node **Annual Report** of **Marvel** with UId “Marvell_20142013”	31
Node **Annual Report** of **Marvel** with UId “Marvel_20152013”	31
Node **Annual Report** of **Marvel** with UId “Marvell_20162013”	31

Experiment [11]

Parameters applied with the Algorithm

- Algorithm: Connected Components
- Label: ANY
- Relationship: DEFINE_QUARTER_RISKS
- Community Node Limit: 32 (because largest cluster is of size 32)
- Rows to show: 35 (because only 35 communities with size >1 exist)



We get the same insights as before; company ***Marvel*** has a larger number of risks associated with them. Since all the companies come from the same industry, we do expect a similar number here. For this reason, further examination is needed.

⇨ Why does ***Marvel*** have so many risks?

After applying the second part of the pipeline (Centrality algorithm) we gathered the most popular nodes of interest.

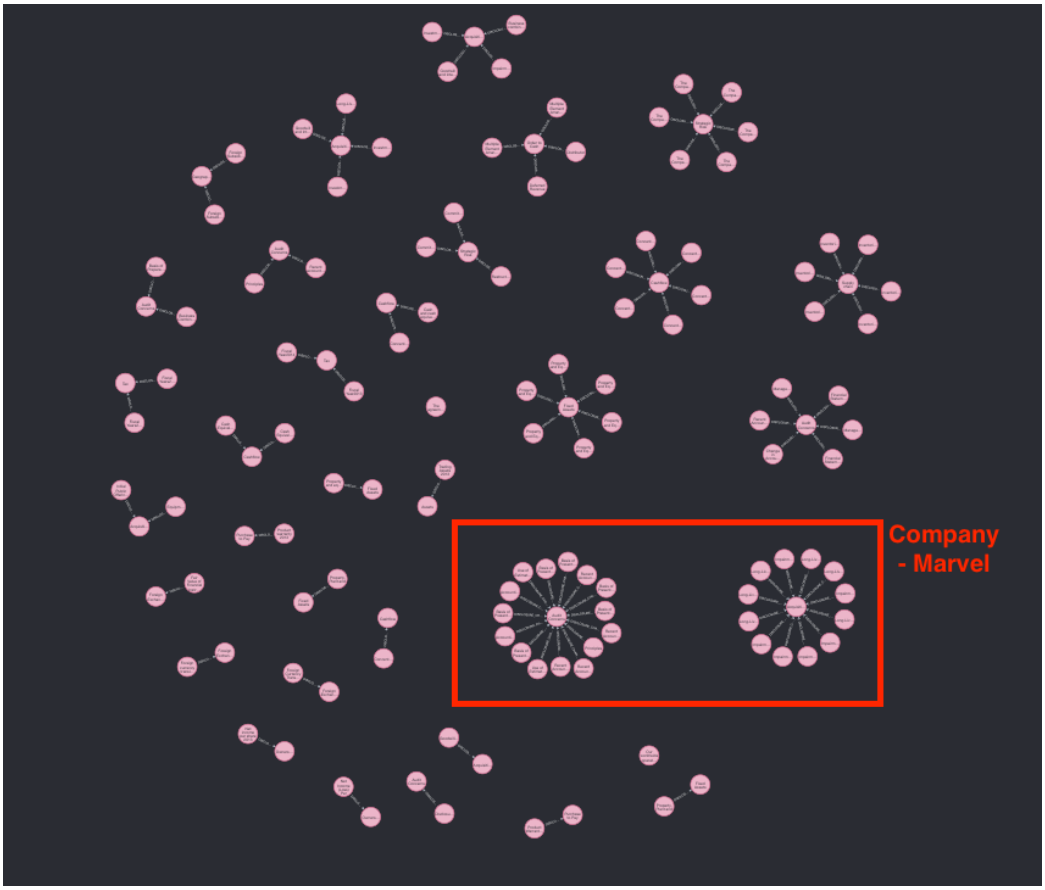
Top 5 popular nodes of knowledge graph	
Node Label | Sub-label with company	*Degree power*
Node **Quarter Report** of **Marvel** with UId "Marvell_2013_Q103/02/201304/05/2013"	31
Node **Quarter Report** of **Marvel** with UId "Marvell_2016_Q131/01/201630/04/2016"	30

Node **Quarter Report** of **Marvel** with UId “Marvell_2013_Q205/05/201303/08/2013”	29
Node **Quarter Report** of **Marvel** with UId “Marvel_2014_Q102/01/201403/05/2014”	29
Node **Quarter Report** of **Marvel** with UId “Marvell_2015_Q101/02/201502/05/2015”	29

Experiment [11]

Parameters applied with the Algorithm

- Algorithm: Connected Components
- Label: ANY
- Relationship: DISCLOSURE_CHARACTERISE
- Community Node Limit: 16 (because largest cluster is of size 16)
- Rows to show: 31 (because only 31 communities with size >1 exist)



Even though Marvel does not have the most disclosures, they have the largest clusters of them. This could be an indication of fraudulent activity in a specific area.

⇨ Large communities of disclosures mean something?

After applying the second part of the pipeline (Centrality algorithm) we gathered the most popular nodes of interest.

Top 2 popular nodes of knowledge graph	
Node Label | Sub-label with company	*Degree power*
Node **Areas – Audit Concern** of **Marvel**	15
Node **Areas - Acquisition** of **Marvel**	12

I only kept the top two nodes, because the rest are of equal size across all companies and show normal behaviour.

Conclusion

Going through multiple and various experiments which we modified parameters of the ***Connected Component*** algorithm we observed the following:

1. Suspicious and abnormal behaviour is associated with company ***Marvel*** risks.
2. Suspicious and abnormal behaviour is associated with company ***Inphi*** disclosures.
3. Only company ***Marvel*** and ***AMD*** have swot reports/risks.

Most of the abnormal behaviour of company ***Marvel*** has to do with the large number risks that are associated with their ***Areas*** nodes highlighted in the resultant tables from applying the centrality algorithm.