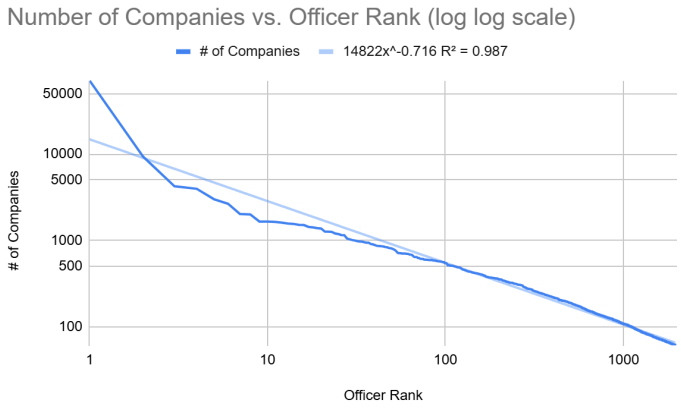
HW #1 by Tal Schul - Exploring ICIJ Offshore Leaks Dataset

This assignment aims to analyze the structure and hidden patterns within the ICIJ Offshore Leaks dataset, which maps relationships between companies, individuals, and addresses in offshore financial networks. The exploration will focus on identifying an anomaly in the data and explain these behaviors.

A phenomenon that may interest us will be an officer overseeing a lot of entities, this will give us insight on how such officers behave and help us discover patterns in their behaviors.
In order to achieve this we can query the graph DB with the following query and see such officers:

```
MATCH (o:officer)-[:officer_of]->(e:entity)
RETURN o.name, COUNT(e) AS num_companies
ORDER BY num_companies DESC
LIMIT 1000;
```

Firstly, an interesting pattern emerges when sorting the officers by the number of entities they oversee, and then plotting the amount of companies vs their rank we get a graph with power law properties, meaning that a few officers participate in most of the connections to entities.



A second pattern becomes apparent when investigating the top officer query in the DB. The top officer this query returns is “THE BEARER” with 70950 different entities which seems significantly higher than expected (large residual in the plot) and might be an outlier. In order to check this discrepancy I ran the following query that counts the number of officers with this name:

```
MATCH (o:officer)
WHERE o.name = "THE BEARER"
RETURN COUNT(o) AS num_officers;
```

I discovered there are multiple nodes with this name, specifically 70873, and in order to look into it I modified the original query to split the count by entity and not by name and filtered it to have only nodes with the name “THE BEARER”, then counted the number of officers with each number of companies:

```
MATCH (o:officer)-[:officer_of]->(e:entity)
WHERE o.name = "THE BEARER"
WITH o, COUNT(e) AS num_companies
RETURN num_companies, COUNT(o) AS num_officers
ORDER BY num_companies DESC;
```

And it will produce the following output:

Number of Companies	Number of Officer that are Linked to This Number of Companies
1	70774
2	58
3	20

Looking at the output we can see that most of these nodes are only connected to one company and a few are connected to more. Looking up “THE BEARER” online I found out that the definition of bearer is “a person or thing that carries or holds something” leading me to believe that this is just a filler name for officers that don't have an actual name in the system and that each such node represent one such officer that is only connected to a few entities like we would expect from the tail of the power law relation.
In order to check if there are more such names, like “THE BEARER”, I wanted to see for each name, how does the number of officers with that name compare to the number of entities related to the name, I explored this connection by looking at the ratio: *num_officers / num_companies*.
In order to explore this ratio, I ran the following query:

```
MATCH (o:officer)-[:officer_of]->(e:entity)
WITH o.name AS officer_name, COUNT(e) AS num_companies, COUNT(DISTINCT o) AS num_officers
RETURN officer_name, num_officers, num_companies,
    CASE WHEN num_companies > 0 THEN toFloat(num_officers) / num_companies ELSE 0 END AS
officer_to_company_ratio
ORDER BY num_companies DESC
LIMIT 100;
```

And by looking at such names with a ratio greater than 0.9 I got the following:

officer_name	num_officers	num_companies	officer_to_company_ratio
THE BEARER	70852	70950	0.9986187456
EL PORTADOR	9325	9326	0.9998927729
Bearer 1	2655	2655	1
Bearer	818	837	0.9772998805
The Bearer	813	813	1
BEARER	537	587	0.9148211244

Interestingly, all the names have the same meaning. With five of them containing the word “bearer” and the last one is the Spanish translation of the word, “EL PORTADOR”.

In conclusion, my analysis of the ICIJ Offshore Leaks dataset revealed a power law distribution in officer-entity relationships, with a few officers overseeing most entities. However, the most common officer name, "THE BEARER," appears to be a placeholder rather than a real individual. Further investigation found similar generic names which highlights the need for careful interpretation of offshore financial data to distinguish between real officers and fake ones.