# Unsupervised Entity Resolution for Financial Datasets

Data Science Institute
COLUMBIA UNIVERSITY

**Basil Vetas, Deepak Maran, Ritu Tak, Sha Yu Han, Yang Qiao**
**Amir Rahmani, Jihan Wei**
**Eleni Drinea**

**Data Science Capstone Project with Capital One**

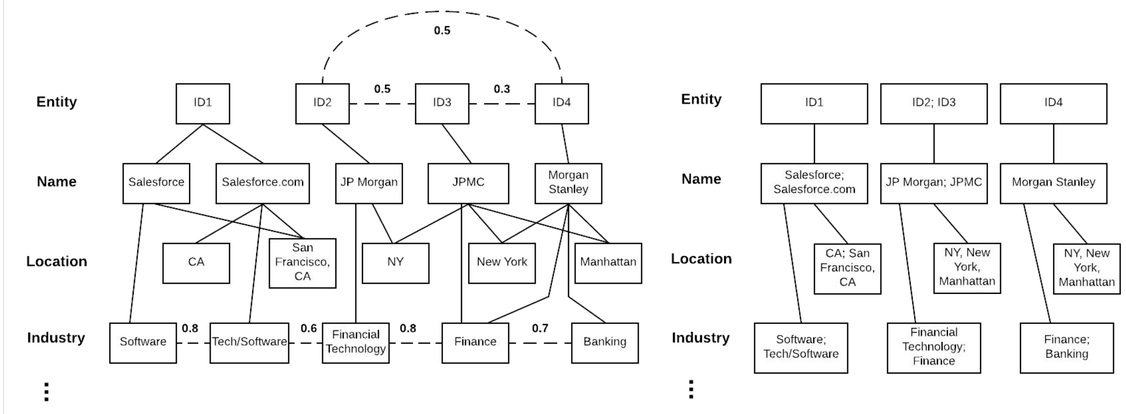
## Introduction

**Many databases contain uncertain and imprecise references to real-world entities. This happens often when data is aggregated from different sources. The absence of unique identifiers for the underlying entities often results in ambiguity or multiple references.**

**Entity Resolution is the task of identifying and deduplicating unique entities from various data records based on the semantics hidden in the data. It discovers the underlying entities and maps duplicate records to these unique entities. There are many approaches to entity resolution, but in the unsupervised learning setting we do this by linking and clustering the data.**

## Methods

**We approach the problem with an unsupervised graphical model that produces a summarized K-partite graph. We transform the original datasets into multiple K-type graphs, with nodes representing fields of the row entries of the dataset. Nodes are connected by weighted edges representing relationships between entities themselves as well as between entities and their attributes. The weights on edges within layer represent the similarity scores while edges between two layers are calculated using TFIDF score. Our method clusters nodes from the original graphs into super nodes in the new summarized graphs (example below).**




(a) An Input Graph

(b) A Summary Graph

## Results

**To create our experimental data we sample 50 unresolved entities from the provided datasets. We use above mentioned methodology to calculate a summarized bipartite graph. The first layer of the graph corresponds to entity names forming supernodes of resolved entities. The second layer corresponds to unique words from available attributes such as address and industry.**

**In Figure 3 we look at clusters corresponding to supernodes in layer 1 of our summary graph. As can be seen, the algorithm clusters different mentions of entity names together. We will investigate the performance of this methodology further using quantitative metrics for cluster similarity such as silhouette score and rand index.**

## Experiment: Entity Resolution for 50 Entities

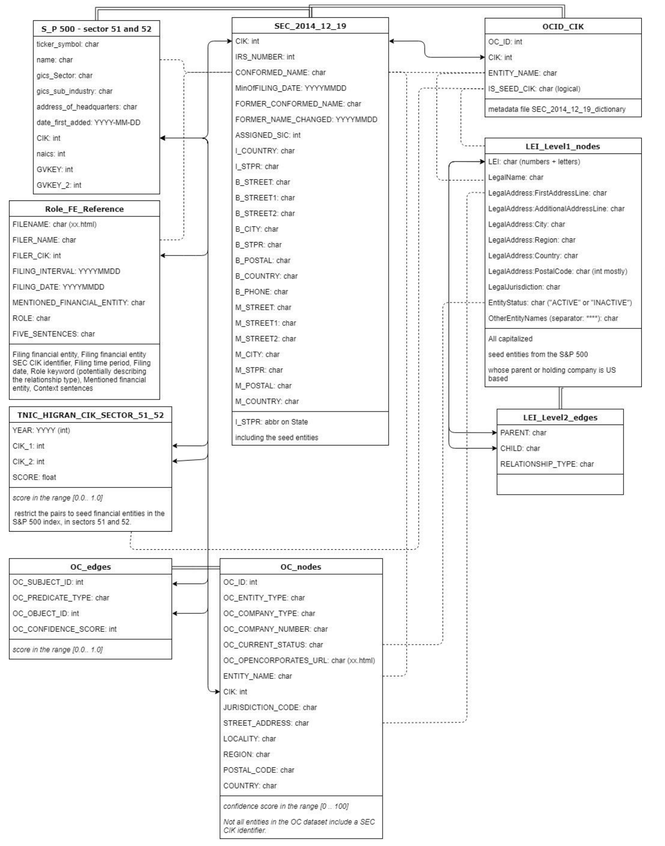



**Figure 2. Raw Data Structure**

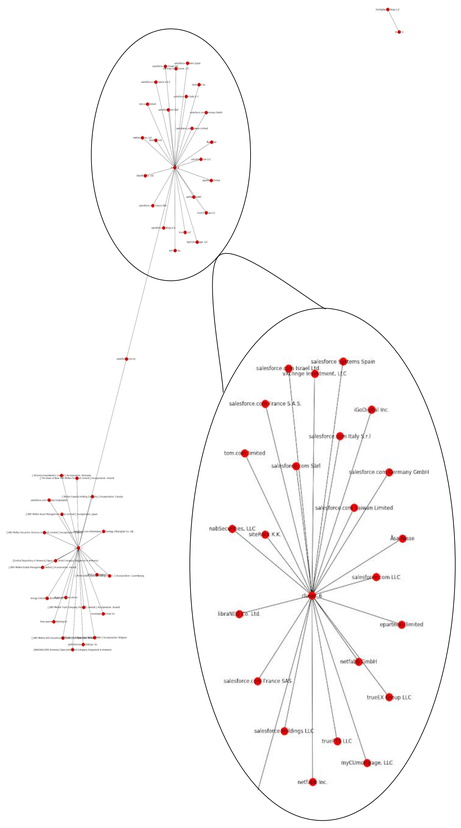


**Figure 3. Clustered Entities**

## Conclusions

**In this work we use multi type graph summarization method that identifies entities in an unsupervised setting. We applied this approach on a small subset of provided dataset and visually inspected the results. We plan to make an quantitative assessment of this approach and scale it to complete dataset.**

## Acknowledgments

**We are very grateful to Amir Rahmani, Jihan Wei from Capital One for many helpful discussions and comments on our approaches and reports. We also gratefully acknowledge Eleni Drinea for her support and guidance throughout the process.**

## References

**Unsupervised Entity Resolution on Multi-type Graphs. ISWC 2016 - 15th International Semantic Web Conference. 2016. Linhong Zhu and Majid Ghasemi-Gol and Pedro Szekely and Aram Galstyan and Knoblock, Craig A. http://usc-isi-i2.github.io/papers/zhu16-iswc.pdf**