LLM Driven Web Profile Extraction for Identical Names & Connected Entities in Interlocking Directorships

Thesis submitted in partial fulfillment
of the requirements for the degree of

Masters of Science
in
Computer Science and Engineering
by Research

by

Prateek Sancheti
2019111041
prateek.sancheti@research.iiit.ac.in

INTERNATIONAL INSTITUTE OF
INFORMATION TECHNOLOGY
HYDERABAD

International Institute of Information Technology
Hyderabad, India

June 2024

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

This is to certify that work presented in this thesis titled ***LLM Driven Web Profile Extraction for Identical Names & Connected Entities in Interlocking Directorships*** by ***Prateek Sancheti (Roll No. 2019111041)*** has been carried out under our supervision and is not submitted elsewhere for a degree.

Date

Advisor: *Dr. Kavita Vemuri*

Advisor: *Dr. Kamalakar Karlapalem*

Acknowledgement

I would like to express my sincere gratitude to *Prof. Kavita Vemuri* and *Prof. Kamalakar Karlapalem* for their support throughout my Master's journey. Their guidance, encouragement, and approachability made me feel comfortable, allowing me to reach out to them whenever needed. I am truly grateful for their mentorship, which has been invaluable in shaping this academic pursuit.

I owe a debt of gratitude to my family for their endless love and encouragement. To my Mom, Dad, and my brother Aayush, thank you for always believing in me and pushing me to follow my dreams. I am immensely grateful for your presence in my life. A special mention goes to my grandparents, whose constant love and blessings have been a source of strength for me.

A heartfelt appreciation for my friends. In particular, I would like to thank Ms. Singhal, my *"Dulli Partner"*, for her companionship and unwavering support throughout the journey. A special thanks goes to my friend Riya Raj, who has been by my side through every phase of this journey. From moments of crankiness to late-night cribbing sessions, she has been a pillar of support and a listening ear when needed most.

I would like to express my gratitude to all those who have supported me in ways both seen and unseen. I am truly grateful for each and every one of you.

Last but not least, I wanna thank me. I wanna thank me for believing in me. I wanna thank me for doing all this hard work. I wanna thank me for having no days off. I wanna thank me for... for never quitting. I wanna thank me for just being me at all times.

"If it's flipping hamburgers at McDonald's, be the best hamburger flipper in the world" – Snoop Dogg

Abstract

The number of individuals with identical names on the internet is increasing. Thus making the task of searching for a specific individual tedious. The user must vet through many profiles with identical names to get to the actual individual of interest. The online presence of an individual forms the profile of the individual. We need a solution that helps users by consolidating the profiles of such individuals by retrieving *factual information* available on the web and providing the same as a single result. We present a novel solution that retrieves web profiles belonging to those bearing identical Full Names through an end-to-end pipeline. Our solution involves information retrieval from the web *(extraction)*, *LLM-driven Named Entity Extraction (retrieval)*, and standardization of facts using Wikipedia, which returns profiles with fourteen multi-valued attributes. After that, profiles that correspond to the same real-world individuals are determined. We accomplish this by identifying similarities among profiles based on the extracted facts using a *Prefix Tree* inspired data structure *(validation)* and utilizing *ChatGPT's* contextual comprehension *(revalidation)*. The system offers varied levels of strictness while consolidating these profiles, namely *strict, relaxed,* and *loose* matching. The novelty of our solution lies in the innovative use of *GPT* – a highly powerful yet unpredictable tool for such a nuanced task. A study involving twenty participants and other results found that one could effectively authenticate information for a specific individual.

Interlocking Directorships (IDs) have been an area of interest for researchers for several decades. Corporations sharing directors and directors sharing corporations result in connections in the corporate world. These connections carry *dual-edged influences*, from sharing resources and network expansion to collusion and *quid-pro-quo*. We present a systematic approach to identify frequently occurring groups of directors and companies and connected components within these corporate structures. We identify various *weakly (maximal cliques)* and *strongly (maximal frequent itemsets)* connected entities from these networks by extracting and analyzing a data corpus of over *55,000 Directors*, *85,000 Companies*, and over *3,00,000 Director-Company* links. We also found that *37,123* companies out of a total *87,187* – almost *30%*, have at least one pair of directors that share the same last name (possibly family-run companies). Finally, we also present a way to extract *personal* and *professional relations* between connected directors from the web with our LLM-driven profile extraction pipeline.

Contents

List of Figures

List of Tables

Abbreviations

BERT	Bidirectional Encoder Representations from Transformers
BFS	Breadth-First Search
CIN	Company Identification Number
DIN	Director Identification Number
FP	Frequent Pattern
IDs	Interlocking Directorships
LLMs	Large Language Models
LOC	Location
MCA	Ministry of Corporate Affairs
MCCs	Maximal Company Cliques
MDCs	Maximal Director Cliques
MFCIs	Maximal Frequent Company Itemsets
MFDIs	Maximal Frequent Director Itemsets
MISC	Miscellaneous
ML	Machine Learning
NE	Name Entity
NER	Name Entity Extraction
NLP	Natural Language Processing
ORG	Organisation
PER	Person
SERP	Search Engine Result Pages
SOTA	state-of-the-art

List of Related Publications

[P1] **Prateek Sancheti**, Kamalakar Karlapalem, Kavita Vemuri, **"LLM Driven Web Profile Extraction for Identical Names"**. *In Companion Proceedings of the ACM Web Conference 2024 (WWW '24 Companion). doi: 10.1145/3589335.3651946*

[P2] **Prateek Sancheti**, Kamalakar Karlapalem, Kavita Vemuri, **"Interlocking Directorship Among Corporations in India"**. *Submitted at the International Conference on Information Systems (ICIS '24)*.

Chapter 1

Introduction

This thesis delves into two critical areas: *(1) Extraction of Web Profiles for Individuals Sharing Identical Names and (2) Identifying Weakly and Strongly Connected Entities in Interlocking Directorships.*

The proliferation of identical or similar names is becoming increasingly prevalent. The *1990 U.S. Census Bureau* data revealed that out of a population of *100 million* people, there were just *90,000* unique names (Guha, 2004). This presents a significant challenge in today's interconnected world facilitated by the web. A user attempting to locate a specific individual on the internet must sift through numerous web profiles that share the same name. Each individual's web profiles form their online presence or digital footprint. Addressing this challenge requires a solution simplifying user access by consolidating these profiles, retrieving factual information across the web, and presenting it as a unified result. Notably, approximately *11 – 17%* of search engine queries involve person names, highlighting the magnitude of this issue (Artiles et al., 2010; Spink et al., 2004). The challenge stemming from this proliferation spans three critical aspects of the online world – *misrepresentation, privacy and security*, with consequences ranging from *inadvertent associations to intentional deception* of online profiles. Consider the scenario where a distinguished keynote speaker at a conference shares their name with multiple individuals, some of whom even have similar backgrounds. Without a reliable method to discern these web profiles based on factual information, users risk selecting incorrect or mismatched profiles, emphasizing the urgency of addressing this challenge.

Various studies over the years have delved into related areas, such as the occurrence of person names in newspaper articles and the disambiguation of author names in published works. These studies have introduced a range of techniques, including Name Entity Extraction (NER) and graph models, as potential solutions (Vu et al., 2007). Previous attempts at Cross Document Co-Referencing (CDC) have utilized Vector Space Models (Baeza-Yates et al., 1999; Bagga and Baldwin, 1998), while techniques leveraging biographical facts *(birthday and birthplace)* have been employed for web page clustering (Mann and Yarowsky, 2003). Commonly adopted techniques by state-of-the-art (SOTA) approaches for addressing such issues involve the extraction of Name Entity (NE) using pre-trained models and the utilization of Bag of Words (BoW) weighted by the Term Frequency - Inverse Document Frequency (TF-IDF) function (Delgado et al., 2014).

However, less studied is the domain of extraction and consolidation of web profiles of individuals sharing identical names. Current approaches often rely on Machine Learning (ML) techniques, necessitating extensive task-specific training to achieve baseline results.

In response to the need for clarity in online identity management, this thesis presents a novel solution that not only disambiguates but also consolidates distinct web profiles that share identical *Full Names (First Name and Last Name)*. Our approach is grounded in the principles of extracting valuable factual information – information that is verifiable through online sources (evidence) or direct statements within the profile, extracted using Large Language Models (LLMs) (Wang et al., 2023), Prompt Engineering (Ashok and Lipton, 2023), and the standardization of extracted facts.

Having addressed the first challenge we solve, we now introduce the second critical area this thesis targets – *Identifying Weakly and Strongly Connected Entities in Interlocking Directorships*. In the complex and modern corporate governance structure, the phenomenon of *Interlocking Directorships (IDs)* is a crucial yet nuanced aspect. A *corporation* is a legal entity separate from its owners, capable of conducting business, owning assets, and being liable for its debts. *Directors*, elected by shareholders within a corporation, hold a crucial responsibility in overseeing the strategic direction and governance of the company, ensuring its operations align with shareholder interests. A corporation can have multiple individuals serving as directors, and likewise, an individual can hold directorial positions at multiple corporations, creating a *many-to-many* network structure. Some countries, like the United States of America, even have laws stating the minimum number of directors a company must have.

An *interlocking directorship* occurs when an individual holds directorial positions across multiple corporations, creating a network of shared influence (Mizruchi, 1996). Such interlockings not only connect different individuals with diverse expertise and vision across corporate boards but also foster an intricate web of professional relationships that can have profound implications on business practices, policy formation, and strategic alliances.

The dual-edged nature of director interlocking raises significant concerns. On the one hand, engaging visionary leaders and implementing strong corporate governance can promote the dissemination of innovative policies, valuable information, resources and strategies, which can, in turn, be effective in addressing *environmental, social, and governance (ESG)* issues (James Chen, 2022). Interlocks can also help corporations obtain a competitive edge by monitoring business ties and gathering confidential information, which leads to improved management practices (Lamb and Roundy, 2016; Loderer and Peyer, 2002; Mazzola et al., 2016; Ozmel et al., 2013). On the other hand, the independence of board decisions might be compromised by sharing directors (Adams, 2017). It also harbours potential risks related to corporate malpractices or *quid-pro-quo* situations, including but not limited to corruption and collusion, the formation of shell companies, facilitation of hostile takeovers and unfair business practices (Hillman and Hitt, 1999; Holburn and Vanden Bergh, 2008).

Research on interlocking directorates and corporate networks began during the formative period of corporate capitalism *(early twentieth century)* (Dooley, 1969; Sapinski and Carroll, 2018). Earlier studies focused on the reasons behind the creation and maintenance of interlocks (Palmer et al., 1986; Pfeffer and Salancik, 1978; Stearns and Mizruchi, 1986). In contrast, in more recent studies, researchers have shifted their attention to the consequences of such interlocks on business behaviour and ideology (Caiazza et al., 2019; Hernández-Lara and Gonzales-Bustos, 2019; Mizruchi, 1996). Some researchers also tried to examine the role of Interlocking Directorates and Universities (Slaughter et al., 2014). According to Dooley (1969) *(1) Corporation's size, (2) degree of managerial control, (3) financial ties, (4) interaction with rivals and (5) existence of local economic interests* are the *five* reasons behind the formation of board interlockings. Some studies suggest resource-dependence theories as another reason for the same (Pfeffer, 2019; Pfeffer and Salancik, 1978).

Transparency and awareness among investors, stakeholders, and regulatory bodies regarding these interconnected networks are crucial for maintaining the integrity of corporate governance. Motivated by the critical need for clearer insights into the structure of director interlockings, especially within the context of Indian Corporations, we present a pipeline to navigate through the complexities of these networks. We identify *weakly and strongly connected entities* among corporate networks using data analyitcs techniques such as *Graph Cliques* and *Frequent Itemsets* of Directors and Companies within these corporate networks, highlighting how a small set of directors can share control of many companies and that a large number of directors can control a small set of companies. We further extend our study and simplify these networks by identifying relationships among these directors leveraging an adaptation of our *Large Language Model (LLM) Driven Web Profile Extraction* pipeline. These relationships are categorized as either **personal**, including family ties such as *Daughter–Father* or *Wife–Husband* etc., or **professional**, encompassing *shared work experiences, shared professional memberships,* or *shared educational backgrounds*. The objective is to provide such nuanced information as a more transparent and informed analysis of corporate structures.

1.1 Thesis Contributions

This thesis presents a comprehensive end-to-end pipeline to address the issues posed by the proliferation of identical names and person name disambiguation in today's digital world. Our innovative approach utilizes the power of Large Language Models (LLMs) to extract entities from unstructured text and provides readers with a glimpse into the transformative potential of this technology. With careful use of prompt engineering techniques, we have developed a functional prompt that facilitates extracting key information from unstructured textual sources. The method of data standardization using Wikipedia mitigates the issue of diverse representations of entities. The pipeline also addresses celebrity name conflicts by incorporating search engine optimization techniques. Lastly, we attempt to maximize the utilization of LLMs' contextual understanding to minimize human supervision of the task. The potential to utilize this technology is limitless.

Our work also presents a pipeline to identify weakly and strongly connected entities (directors and companies) present in a corporate network. A carefully designed methodology involving graph-based data traversal and data analytics techniques such as Graph Cliques and Frequent Itemsets presents a different view of the networks. We extract information for over 1,40,000 entities (85,000 companies, 55,000 directors) in the Indian Corporate network and perform various analyses. We then discuss a few observations made during our study. Lastly, we utilize an adapted version of our LLM Driven Web Profile Extraction pipeline to identify Personal and Professional Links between identified pairs of directors.

1.2 Organisation of the Thesis

This thesis is divided into **six** chapters, which are organized as follows:

- **Chapter 1** sets the stage for our thesis by explaining the motivations and reasons behind our research work. It briefly introduces the two critical areas our thesis addresses, along with our approach and the theoretical basis on which it relies. The chapter also highlights the significant contributions of this thesis to the relevant domains.

- **Chapter 2** provides an in-depth review of existing works and concepts relevant to our research focus. It offers a thorough examination of prior works concerning the two key challenges we aim to address. It also lays the ground for some of the required concepts, thus serving as a foundation for subsequent analyses in our thesis such as *NER, LLMs, Cliques and Frequent Itemsets*.

- **Chapter 3** delves into our solution to the challenge presented by the proliferation of identical names in today's interconnected world. We explore existing technologies used thus far and introduce our innovative approach to extract and consolidate web profiles of individuals sharing identical names, which harnesses LLMs for addressing this issue. The chapter outlines several critical phases, including data retrieval, entity extraction, and profile consolidation (DeDuplication), and details our comprehensive end-to-end pipeline.

- **Chapter 4** explores our efforts to enhance the transparency of interlocking directorships in the Indian corporate networks by identifying *frequent itemsets*, *cliques* and *relationships* within these networks. We delve into data extraction from various sources of information and elaborate on the data analytics concept of frequent itemsets and graph cliques and their significance in our study. Lastly, we detail our methodology for identifying relations among strongly connected directors in these networks by utilizing our work from chapter 3. We present various examples, visualizations and results to elucidate our findings.

- **Chapter 5** concludes our thesis, summarizing the entirety of our work and findings. Additionally, it offers insights into potential directions for future research and development.

Chapter 2

Related Works

2.1 LLM Driven Web Profile Extraction for Identical Names

2.1.1 Overview of Existing Literature: *Proliferation of Identical Name*

The prevalence of individuals sharing the same name is a common occurrence, increasing the challenges they pose as it increases. In today's digitally interconnected world, where one's online presence is synonymous with one's physical presence, it becomes crucial to have a method for extracting web profiles of individuals with identical names and subsequently disambiguating them based on available web information.

This field of person name disambiguation has been subject to study over time in the fields of *Natural Language Processing (NLP), Information Retrieval (IR) and Text Mining (TM)* (Delgado et al., 2018), showing its significance in effectively distinguishing between individuals who share identical names. Differentiating individuals sharing the same name is essential for various social network extraction systems (Matsuo et al., 2006; Mika, 2004).

From an NLP standpoint, researchers have tried to approach similar problems such as *Document Co – Referencing, Author name disambiguation, Web Page Clustering, etc*.

Document Co – Referencing: Bagga and Baldwin (1998) calculates document similarity using just co-occurring terms (entities) and a vector space model, whereas McCallum and Wellner (2003) presents a method that uses proper nouns to co-reference documents. Mann and Yarowsky (2003) proposed an unsupervised method that uses a collection of regular expressions to cluster the documents to their namesakes to extract biographical information relevant to each person, such as birthdate and place of birth. However, such personal information about all namesakes is not always available online. Fleischman and Hovy (2004) builds a maximum entropy classifier to determine the separations between documents, which are subsequently grouped. Their approach necessitates an extensive training set. The method proposed by Bollegala et al. (2008) extracts keywords and uses them to calculate document similarities.

Author Name Disambiguation: This is another area studied previously (Han et al., 2005; Subramanian et al., 2021). However, citations (author name problem) have a set structure, unlike other online content.

It is simple to extract fields from citations such as co-authors, title, journal name, conference name, and year of publication, which are essential for the disambiguation process.

Web Page Clustering: The Web People Search task, as defined in the *WePS* evaluation campaign (Artiles et al., 2010, 2007), consists of grouping search results for a given name according to the different people that share it. Guha (2004) presents a reordering strategy to clear individuals. The user must choose one of the pages returned as the beginning point using the algorithm. The algorithm then re-ranks all of the search results by comparing the person descriptions, giving greater ranking to pages that relate to the same person as the user-selected page. Artiles et al. (2009) discusses the role of Named Entities in the task of person name disambiguation on the web. Their findings demonstrate that even though used by most methodologies, the use of named entity identification and classification alone in solving the issue aids very little. Kalashnikov et al. (2006) presents a methodology that first extracts 'significant' entities, such as person names, organisation names, locations and hyperlinks on a webpage. They then build an entity-relationship graph, which is used to cluster webpages together. Lan et al. (2009) and Ikeda et al. (2009) bifurcates the task of web page clustering into two parts: (1) attribute/entity extraction and (2) people clustering. After successfully extracting the attributes Lan et al. (2009) implements a widely used K-means clustering algorithm to cluster web pages where as Ikeda et al. (2009) implements a two-stage clustering.

Most techniques that attempt to solve the issue of person name disambiguation try to extract some form of information from the text (entities/attributes/hyperlinks) using machine learning models and then try to cluster them by finding the relations between pages. The remaining part of this section discusses Entity Extraction, the current SOTA for the task, and their limitations.

2.1.2 Entity Extraction from Text

Entity extraction, often referred to interchangeably as *entity identification, entity chunking*, or *Named Entity Recognition (NER)*, is the process of identification and extraction of precise information from unstructured text. In a sea of words, where irrelevant data can be overwhelming, entity extraction acts as a filter, isolating important text parts like names, dates, and locations.

The entity extraction process enhances unstructured text by introducing both structural organisation and semantic meaning. In an era of exponentially expanding data volumes, the significance of the process that brings order to this overwhelming chaos cannot be overstated. Over time, a range of techniques has emerged, including utilising *Machine-Learning (ML)* algorithms to detect specific entity mentions within text or to summarise large volumes of content. This process of extracting entities has evolved into a crucial pre-processing step in *Natural Language Processing (NLP)* tasks before any subsequent analysis or processing can occur.

We build a pipeline for extracting and consolidating web profiles of individuals who share identical names. Central to this work is the need to extract critical information from various unstructured text pieces. Entity extraction, hence, emerges as an essential component of our study. Figure 2.1 is an example of the task of *Entity Extraction* within a given text. Throughout this section, we delve into the intricate workings of the process, exploring the various techniques employed in contemporary practices and their inherent limitations.

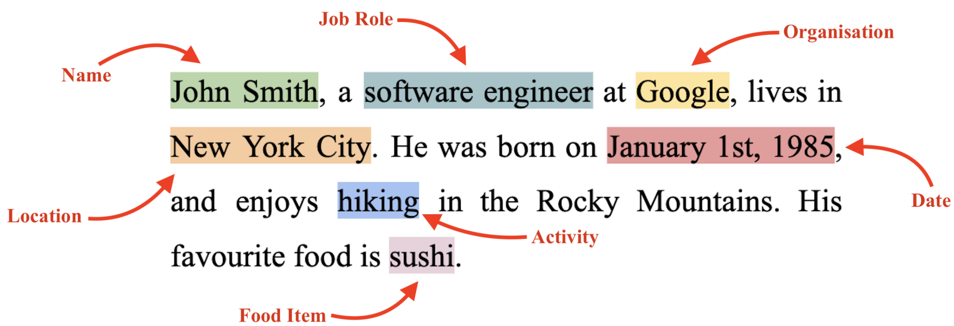


Figure 2.1: Entity Extraction from Text

2.1.2.1 What is an Entity?

An *entity*, in its broadest sense, refers to anything that possesses a distinct and self-contained existence. This encompasses tangible entities, such as animals, plants, and structures like buildings, as well as abstract or conceptual entities, including nations, organisations, and emotions. Linguistically, the concept of entities encompasses the vast majority of nouns, each representing some form of entity within a given context. Essentially, entities encapsulate the myriad elements that populate our physical or conceptual world, providing a foundation for understanding and communication within language and discourse.

2.1.2.2 What is Entity Extraction?

Entity extraction is like sorting through a stack of documents to pinpoint the essential details. It is a complex, multi-step process involving systematically identifying specific elements within unstructured raw text data, known as *entities*. These entities can range from names of individuals and places to dates and organisational affiliations. By discerning and extracting these critical pieces of information, entity extraction brings order to the chaos of raw text.

This process is integral to tasks such as information retrieval, sentiment analysis, and data mining, allowing computational systems to derive meaningful insights from textual content efficiently. In essence, entity extraction streamlines the understanding and analysis of textual data, facilitating informed decision-making and research endeavours.

2.1.2.3 What is a Named-Entity?

Named entities refer to specific entities mentioned in the text, typically identified by proper names and consistently representing the same concept throughout different contexts. For instance, *'Einstein'*, *‘Paris’* and *‘Microsoft’* are examples of named-entities where as *‘genius’*, *‘city’* or *‘company’* are not. This is because *‘Paris’* refers to particular locations, but a common noun like *‘city*’ can apply to something else concepts depending on context.

While named entities maintain a stable association with a *specific concept*, they differ slightly from rigid designators. Rigid designators exclusively denote terms with an unchanging connection to a particular concept. In Natural Language Processing (NLP), however, elements like *currency* and *time periods* are also considered named entities due to their practical relevance despite not strictly meeting the criteria of rigid designators.

2.1.2.4 *NER:* Named-Entity Extraction

Named Entity Recognition (NER) is a form of Natural Language Processing (NLP). It is a vital technique of entity extraction, going beyond mere identification to classify extracted entities into predefined categories like *“person”*, *“location”* or *“organisation”*. This classification enhances text comprehension, facilitating information retrieval and sentiment analysis tasks. NER is instrumental in extracting meaningful insights from unstructured data, benefiting various domains and applications. Put simply, it identifies *‘where’, ‘what’, ‘who’* or *‘when’* information within a sentence.

It’s crucial to recognise that although the essence of designated entities remains constant, the method of categorising them is inherently subjective. Take, for instance, the term *‘Albert Einstein’*, which consistently refers to the same individual; however, one might classify him as a *‘scientist’*, *‘physicist’* or *‘German’* among various other labels. The classifications presented here merely represent one potential approach to organising these entities. Depending on the context of the dataset, alternative labels may prove to be more apt.

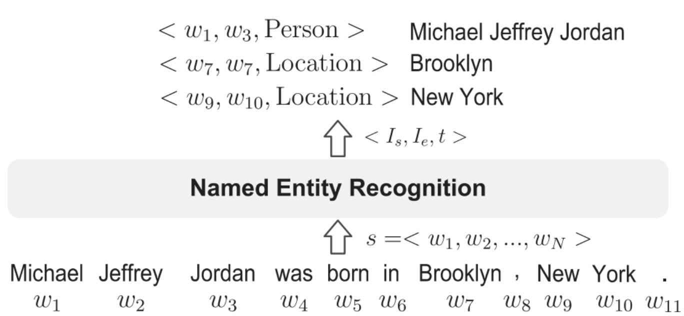


Figure 2.2: Illustration of Named-Entity Recognition Task (Li et al., 2020)

Figure 2.2 illustrates the Named-Entity Recognition task. In the upcoming sections, we’ll delve into NER for our specific context and use case of extracting web profiles of individuals with a particular name from raw text pieces. It’s important to remember that the same entities can be classified differently for different cases based on varying contexts.

2.1.2.5 How does NER Work?

The working of *NER* can be broadly explained in the following steps:

- **Tokenisation**: The first step in the process of NER involves breaking down the text into smaller units called tokens, which can encompass individual words, phrases, or even entire sentences. For example, if we take the sentence *"Albert Einstein was born in Ulm on March 14, 1879"*, it would be segmented into tokens such as *"Albert", "Einstein", "was", "born", "in", "Ulm", "on", "March", "14", ','* and *"1879"*.

- **Entity Identification:** Next, NER employs linguistic rules or statistical methods to identify potential named entities. It looks for patterns like capitalization in names *("Albert Einstein"* & *"Ulm")* or specific formats such as dates *("March 14, 1879")*.

- **Entity Classification:** Identified entities are then categorized into predefined classes like *"Person", "Organization"* or *"Location"*. Machine learning models trained on labelled datasets often handle this classification. In the given statement, *"Albert Einstein"* would be classified as a *"Person"*, *"Ulm"* as an *"Location"* and *"March 14, 1879"* as *"Date"*.

- **Contextual Analysis:** NER systems consider the broader context surrounding these identified entities to refine their accuracy. For instance, in the sentence *"MacBook Air is by far the best product offered by Apple"*, the context is used to differentiate this *"Apple"* as a company rather than the fruit.

- **Post-processing:** Following initial recognition and classification, post-processing techniques refine results. This may involve resolving ambiguities, merging multi-token entities, or utilising knowledge bases to enrich entity data.

2.1.2.6 Various Techniques & Methods for NER

Over the years, numerous Named Entity Recognition (NER) methods have been developed, each precisely adjusted to tackle the diverse challenges of extracting and categorising entities from extensive bodies of raw text. These methods encompass a range of approaches, each tailored to address particular issues encountered in the task of NER. Below, we outline various strategies that have been employed in this domain (Goyal et al., 2018).

- **Rule Based:** NER with Rule-based methods rely on manually crafted rules such as linguistic patterns, regular expressions or dictionaries to identify and classify named entities in text. These precise and interpretable methods offer developers control over the identification process. However, they suffer from scalability issues, as creating and maintaining rules for diverse datasets can be labour-intensive. Additionally, rule-based systems may struggle with generalisation to new or unseen data, and their rigidity can lead to missed entities or misclassifications, particularly in complex or evolving language landscapes. Despite these limitations, rule-based NER remains effective in specialised domains where well-defined entities and structured text are prevalent.

- **Statistical:** These methods for NER utilize statistical models such as *Hidden Markov Models (HMM)* or *Conditional Random Fields (CRF)* to identify named entities based on probability. These methods rely heavily on labelled datasets for training and can be generalised across diverse texts. However, their performance is contingent upon the amount and quality of the annotated training data they are provided. While statistical NER approaches offer the advantage of adaptability to various text types, their effectiveness hinges on the availability of extensive and accurately labelled training datasets.

- **Machine-Learning Based:** Machine learning methods for NER employ algorithms such as *decision trees* or *support vector machines*, akin to statistical approaches. These models create *feature-based representations* of observed data, overcoming many limitations of dictionary and *rule-based methods* by identifying entity names even with minor spelling variations. By learning from labelled data, they predict named entities and are widely utilised in modern NER tasks due to their capability to handle vast amounts of patterns and data. However, they also necessitate substantial labelled data for training and can be computationally demanding.

  Typically, ML-based NER involves two phases: first, training the model on annotated documents, with the time required for training varying based on model complexity; second, utilising the trained model to annotate raw documents.

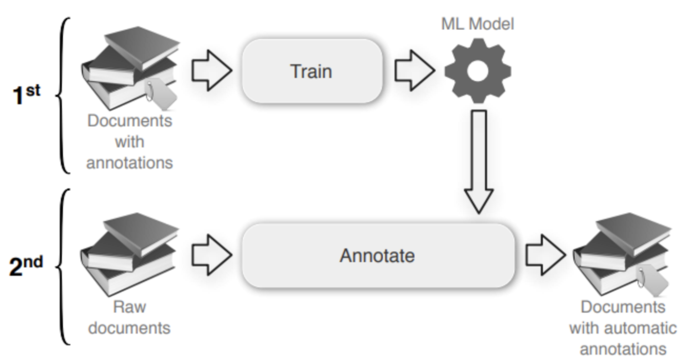


Figure 2.3: Machine Learning Based NER
source: https://www.researchgate.net/

- **Deep-Learning Based:** Deep learning methods represent the latest advancement in NER, leveraging the power of neural networks. Mainly, *Recurrent Neural Networks (RNN)* and *transformers* have emerged as prominent choices, thanks to their capacity to model long-term dependencies within text. While deep learning methods excel in large-scale tasks with abundant training data, they demand substantial computational resources.

Deep learning-based NER surpasses previous methods in accuracy by effectively assembling words, which is facilitated by techniques like word embedding, which discerns semantic and syntactic relationships among words. Moreover, it autonomously learns to analyse both topic-specific and high-level words, rendering it versatile for various tasks. With the capability to handle much of the repetitive work autonomously, deep learning NER allows researchers to allocate their time more efficiently towards other aspects of their work.

- **Hybrid:** These represent the culmination of various NER approaches, blending *rule-based, statistical, machine learning,* and *deep learning* methods to harness the strengths of each. These methods acknowledge the absence of a *one-size-fits-all* solution in NER and aim to capture the best of all worlds. Particularly beneficial for extracting entities from diverse sources, hybrid methods offer the flexibility of multiple approaches. By utilising a rule-based system for swiftly identifying straightforward entities and a machine learning system for discerning more complex ones, they can adapt to the intricacies of different datasets. This intertwined nature of hybrid methods can introduce complexity in implementation and maintenance, requiring careful management to ensure efficacy over time.

2.1.3 State-of-the-art Models for NER

We conducted experiments employing three SOTA NER systems for our pipeline. The task aims to extract entities related to the specified person from web pages. Utilising pre-trained models accessible through *Hugging Face*[1], we analysed the performance of these systems on a designated text piece. Subsequently, we evaluated the results generated by each model and conducted an in-depth analysis of their respective strengths and weaknesses within the context of our task.

This comprehensive examination allowed us to discern the nuances of each NER system's performance, providing valuable insights into their efficacy and applicability for our specific objective.

To elaborate on this analysis, we use a text piece *(1st Paragraph)* from the *Wikipedia* page of *Former U.S. President – Barack Obama* (Wikipedia, 2023)

```
Barack Hussein Obama II (Born: August 4, 1961)is an  American
politician who served as the  44th president of the United St-
ates from 2009  to  2017. A member of the Democratic Party, he
was the first African-American president in U.S. history. Obama
previously served as a U.S. senator representing  Illinois from
2005  to  2008, as an Illinois state senator from 1997 to 2004,
and as a civil rights lawyer and university lecturer.
```

Upon a cursory examination of the text, anyone can identify numerous entities in the above text. However, how humans comprehend text differs significantly from the approaches employed by *Language Models* designed for entity extraction. We aim to examine whether any of the SOTA models

[1] https://huggingface.co/

under consideration suit our specific use case. Each NER system employs a distinct approach, and we strive to evaluate their performance relative to our task. Thus, we analyse each model's handling of the provided text, examining their strengths and limitations in extracting entities to determine their suitability for our intended application.

2.1.3.1 BERT

Bidirectional Encoder Representations from Transformers (BERT), is a super-smart language model developed by folks at Google in October, 2018 (Devlin et al., 2018). Unlike its predecessors, `BERT` employs a transformer-based architecture capable of comprehending the contextual nuances of words within a sentence bidirectionally. This means that `BERT` considers both the *left* and *right* contexts of a word, a shift from previous models that processed words in a unidirectional manner *(left-to-right or right-to-left manner)*.

Another distinction in `BERT`'s innovation lies in the extensive datasets used for its pre-training. Datasets such as the entire **English Wikipedia** *(≈ 2.5B words)*[2] and the **Google's BooksCorpus** *(≈ 800M words)*[3] enable it to leverage vast amounts of textual knowledge.

`BERT` engages in **Masked Language Modeling** or *MLM*, which helps it learn from text in both directions. It works by hiding a word in a sentence and making `BERT` figure out what that missing word is by looking at the words around it from both sides. This two-way learning approach was a big breakthrough and hadn't been tried before. It's like challenging `BERT` to fill in the blanks using its understanding of the whole sentence, not just what comes before or after the missing word. This technique boosted `BERT`'s ability to grasp the meaning of language more deeply. Additionally, `BERT` also uses **Next Sentence Prediction** or *NSP*, which is like playing a game where `BERT` has to guess if one sentence logically follows the one before it. This helps `BERT` to understand how sentences relate to each other.

`BERT`'s architecture, with its inherent understanding of context by considering words in relation to their surrounding words, is perfectly suited for such tasks. NER is fundamentally a sequence tagging task where each word or token in a sequence is assigned a particular label. Fine-tuning `BERT` for NER tasks yields SOTA results, exemplified by achieving an impressive *F1 Score*[4] of **92.4**[5] on the **CoNLL-2003**[6] dataset (Devlin et al., 2018). This multifaceted approach to language comprehension underscores `BERT`'s efficacy across various natural language processing tasks. ***bert_base_NER***[7] is a pre-trained model available on *Hugging Face* to test and experiment NER using `BERT`. This model has been trained to identify **four** types of entities: **Locations (LOC), Organizations (ORG), Persons**

[2] https://huggingface.co/datasets/wikipedia
[3] https://huggingface.co/datasets/bookcorpus
[4] https://en.wikipedia.org/wiki/F-score
[5] https://paperswithcode.com/sota/named-entity-recognition-on-conll-2003-3
[6] https://huggingface.co/datasets/conll2003
[7] https://huggingface.co/dslim/bert-base-NER

(PER), and **Miscellaneous (MISC)**. Figure 2.4 is the NER output from the model upon giving the test text mentioned above *(2.1.3)* as input.

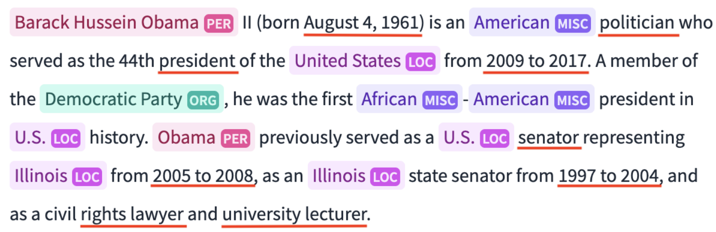


Figure 2.4: Output for NER Task on Fine Tuned BERT Model – *bert_base_NER*

Although the model identifies about *eleven* distinct entities and classifies them in the *four* categories, it misses certain entities such as *Date of Birth, Previous occupations, positions held, periods of years, etc.*, underlined with red in Figure 2.4. It is essential to understand that extracting **all** entities, emphasising on *all*, related to the name in a given text is extremely important since they can be crucial in further steps of our pipeline, which involves *disambiguating* or *consolidating* web profiles based on similarity.

2.1.3.2 Stanza

`Stanza` is a powerful Natural Language Processing toolkit developed by the **NLP Group at Stanford University**[8] that supports **66 human languages** (Qi et al., 2020). `stanza` stands out because of its language-agnostic and data-driven neural pipeline. It takes raw text as input and performs a range of text analysis tasks, including *(1) tokenisation, (2) multiword token expansion, (3) lemmatisation, (4) part-of-speech tagging, (5) morphological feature tagging, (6) dependency parsing and (7) named entity recognition.*

Trained on over *112 datasets* (including the *Universal Dependencies treebanks* and other multilingual corpora), `stanza` is well-equipped to handle diverse linguistic contexts. Moreover, it features a *Python* interface to the widely used *Java CoreNLP* package, granting users access to additional tools like co-reference resolution and relation extraction. For packages with *4* named entity types, supported categories are – `PER` (Person), `LOC` (Location), `ORG` (Organisation) and `MISC`(Miscellaneous) & for packages with *18* entity types, supported categories are – `PERSON`, `NORP` (Nationalities/Religious/Political groups), `FAC` (Facility), `ORG` (Organisation), `GPE` (Geopolitical Entity: Countries/Cities/States), `LOC` (Location), `PRODUCT`, `EVENT`, `WORK_OF_ART`, `LAW`, `LANGUAGE`, `DATE`, `TIME`, `PERCENT`, `MONEY`, `QUANTITY`, `ORDINAL` and `CARDINAL`.

The **Neural Multilingual NLP Pipeline** by `stanza` is meticulously designed to handle the complexities of processing multiple human languages efficiently. From tokenising raw text to conducting syntactic analysis, each pipeline component is tailored to accommodate the nuances found in various languages. Figure 2.5 provides an overview of `Stanza`'s Neural Multilingual NLP Pipeline.

[8] https://nlp.stanford.edu/

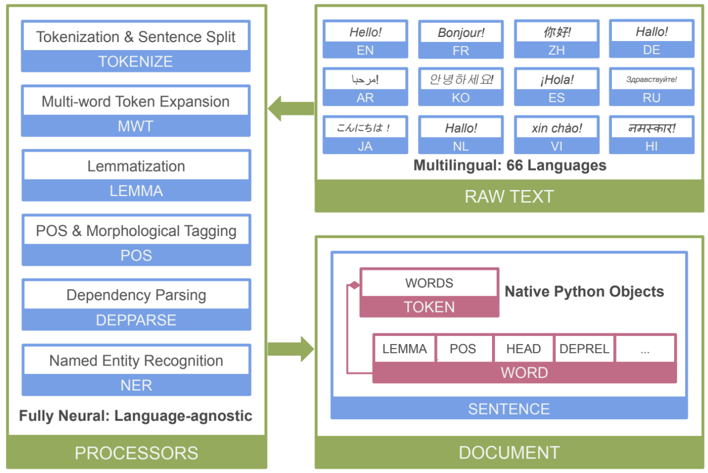


Figure 2.5: Overview of Stanza's Neural Pipeline (Qi et al., 2020)

`stanza` utilises a sophisticated approach based on contextualised string representations to handle the task of NER, as proposed by Akbik et al. (2018). This involves training both forward and backward character-level *Long Short-Term Memory (LSTM) language models*. During tagging, the representations from both language models are concatenated with word embeddings and fed into a standard one-layer *Bidirectional LSTM (Bi-LSTM)* sequence tagger, augmented with a *Conditional Random Field (CRF)* based decoder. This methodology allows `Stanza` to effectively recognise named entities such as person names and organisations in input sentences.

Stanza is open-source, with the entire source code, documentation, and pre-trained models for all 66 languages readily available at https://stanfordnlp.github.io/stanza/. Additionally, an online demo of `Stanza` can be accessed at http://stanza.run/.

We also compared `Stanza`'s NER performance with various other models by using the test text mentioned above (2.1.3) as input. Figure 2.6 is the NER output from `Stanza`'s pre-trained model available on the demo page. We observe that `stanza` does a slightly better job than `BERT`, particularly for our use case. It identifies more number of entities and also classifies them more precisely (Dates, NORPs, GPEs). All being said, it still *fails* to identify all the entities and misses some key terms such as *Politician, President, Senator, Civil Rights Lawyer & University Lecturer*.

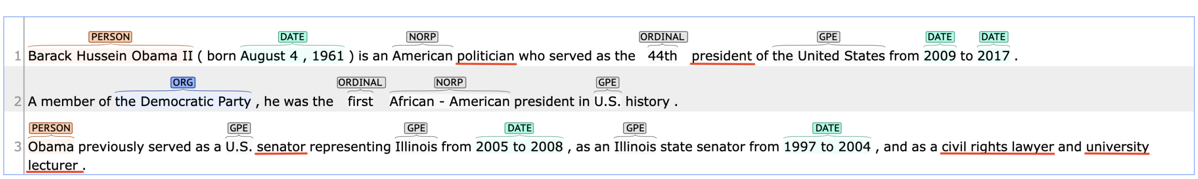


Figure 2.6: Output for NER Task on Fine Tuned `Stanza` Model – *stanza_en*

To evaluate `Stanza`'s NER performance, Qi et al. (2020) compares between `Stanza (v1.0)`, `Flair (v0.4.5)`, and `spaCy (v2.2)`. With *F1 Scores* of **92.1 (English – CoNLL03)** & **88.8**

(English – OntoNotes), `Stanza` achieved SOTA Results. It consistently achieved higher or comparable scores across all datasets compared to `Flair`. Furthermore, compared to `spaCy`, it demonstrated significantly superior NER performance. It must be noted that `stanza` achieves these impressive results with substantially smaller models compared to `Flair`, with reductions in size of up to *75%*. Figure 2.7 provides an overview of the evaluation scores of `Stanza` against `spaCy` and `Flair` (Qi et al., 2020).

Language	Corpus	# Types	Stanza	FLAIR	spaCy
Arabic	AQMAR	4	**74.3**	74.0	–
Chinese	OntoNotes	18	**79.2**	–	–
Dutch	CoNLL02	4	89.2	**90.3**	73.8
	WikiNER	4	**94.8**	**94.8**	90.9
English	CoNLL03	4	92.1	**92.7**	81.0
	OntoNotes	18	88.8	**89.0**	85.4*
French	WikiNER	4	**92.9**	92.5	88.8*
German	CoNLL03	4	81.9	**82.5**	63.9
	GermEval14	4	85.2	**85.4**	68.4
Russian	WikiNER	4	**92.9**	–	–
Spanish	CoNLL02	4	**88.1**	87.3	77.5
	AnCora	4	**88.6**	88.4	76.1

Figure 2.7: NER performances comparision for `Stanza`, `Flair` & `spaCy` across different languages and corpora.(Qi et al., 2020)
Scores reported are entity micro averaged test F_1, # Types $\Rightarrow$ No. of Entity Types

2.1.3.3 FLAIR

Developed by *Zalando Research*[9] in *2018*, `FLAIR` stands as a robust NLP framework, tailored to streamline the training and deployment of SOTA sequence labelling, text classification, and language models. It serves as a go-to resource for researchers and practitioners, offering a user-friendly interface for various word embeddings, thereby facilitating seamless integration into model architectures.

At its essence, `FLAIR` simplifies the incorporation of popular word embeddings like **GloVe** (Pennington et al., 2014), **BERT** (Devlin et al., 2018), **ELMo** (Peters et al., 2018), **Flair Embeddings** (Akbik et al., 2018) and **Character Embeddings** through its intuitive API. This process eliminates the need for additional engineering efforts, expediting model development and deployment. One can combine different word embeddings within a single model architecture. This simplifies model-building and enhances document embedding capabilities, leading to notable performance improvements across NLP tasks.

[9] https://github.com/zalandoresearch

FLAIR leverages contextual string embeddings to effectively capture nuanced linguistic nuances and context. This approach contributes to improved performance in various NLP tasks. Moreover, FLAIR offers support for numerous languages with ongoing efforts to expand its language repertoire.

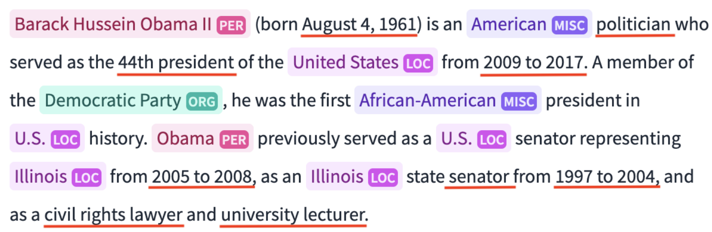


Figure 2.8: NER Task Output on FLAIR Model –*ner_english*

Qi et al. (2020) evaluates FLAIR's performance for NER task, where FLAIR stands out with the highest *F1 Scores* **(92.7 (English – CoNLL03)** and **89.0 (English – OntoNotes))** for most datasets. We also compared FLAIR's NER performance with the test text mentioned above (2.1.3) as input. To study the same, we use FLAIR's pre-trained model available on *Hugging Face* – **flairner_english**[10]. Figure 2.8 is the NER output from the same. Simiilar results were found with BERT (2.1.3.1) and Stanza's (2.1.3.2) pre-trained model.

2.1.4 Challanges with SOTA NER Techniques

Named Entity Recognition methods discussed above have demonstrated remarkable proficiency in scenarios characterised by predefined entity types and abundant, high-quality annotations. However, several challenges persist, hindering their efficacy in real-world applications, particularly for our use case.

2.1.4.1 Missing Crucial Entites

NER models often encounter difficulties in identifying certain entities, either due to a lack of appropriate training or the absence of suitable categories for classification. NER methods excel in scenarios where entity types are predefined, and ample high-quality annotations are available (Lample et al., 2016; Tjong Kim Sang and De Meulder, 2003). The same was seen in tests done on available pre-trained models of SOTA NER methods BERT, stanza and FLAIR. This inherent limitation underscores the critical influence of the training dataset on model performance. Entities crucial for our specific applications, such as disambiguation and consolidation of human web profiles, may be overlooked, highlighting the necessity for comprehensive and diverse training data to enhance model robustness.

[10]https://huggingface.co/flair/ner-english

2.1.4.2 Limited Categories of Entity Classification

Despite advancements in the field of NER, real-world NER suffers from the diversity of entity types, the emergence of new categories, and the scarcity of adequately annotated datasets (Chen et al., 2023). Even with extensive training on diverse datasets, NER models typically classify entities into a limited set of categories, such as `Location (LOC), Person (PER), Organisation (ORG)` and `Miscellaneous (MISC)` or, in some cases, a few more categories such as `NORP, GPE, ORDINAL, CARDINAL etc` (See 2.1.3.2). However, many practical applications demand the extraction of entities beyond these predefined categories. For instance, constructing a comprehensive web profile of an individual may require identifying entities like *educational institutes, tests undertaken,* or *volunteer experiences*, which may not align with existing classification schemas. The freedom to establish such custom entity categories necessitates the creation of meticulously annotated datasets on a large scale.

2.1.4.3 Entity Resolution

SOTA NER models often lack *contextual understanding* or *contextual comprehension*, posing challenges in *entity resolution*, i.e., determining whether two instances refer to the same real-world entity. In text pieces where multiple entities share common attributes, distinguishing between them becomes inherently complex. We aim to extract all entities mentioned in a given text pertaining to a specific individual and categorise them accurately. It is imperative to have a method for extracting only those relevant entities to the subject rather than capturing all entities indiscriminately. For instance, consider the phrase *"Clare, who works at JP Morgan Chase, has a friend named Jacob, employed at Google."* Here, both *"JP Morgan Chase"* and *"Google"* are identified as organisations (`ORG`), yet discerning which entity pertains to Clare, the subject remains ambiguous. Resolving such ambiguities demands a deeper contextual understanding and entity disambiguation mechanisms. The SOTA results for the task of *entity resolution* are achieved by *deep learning* based methods. Still, they typically need many labelled (matching/non-matching entity pairs) training data (Tu et al., 2022).

In summary, while SOTA NER techniques exhibit considerable efficacy, their limitations prove inadequate for our use case due to the scope of diverse and nuanced data we are dealing with. These techniques often overlook crucial information and lack the contextual comprehension necessary for distinguishing one web profile from another. Addressing these challenges necessitates concerted efforts in dataset curation, model refinement, and context-aware entity resolution techniques to enhance the applicability and robustness of NER systems in real-world settings.

2.2 Connected Entities in Interlocking Directorships

2.2.1 Cliques, Maximal Cliques & Maximum Cliques

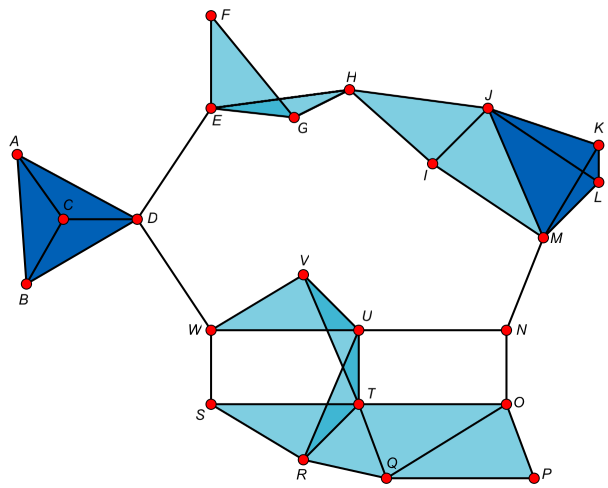


Figure 2.9: Graph Cliques

source: https://en.wikipedia.org/

In the context of graph theory, a *Clique* represents a specific subset of vertices within an undirected graph characterised by a comprehensive adjacency among its members.

As per the definition on *Wikipedia*[11], a *clique*, C, in an undirected graph $G = (V, E)$ is a subset of the vertices, $C \subseteq V$, such that every two distinct vertices are adjacent. This is equivalent to the condition that the induced subgraph of G induced by C is a complete graph. The graph depicted in Figure 2.9 contains the following cliques:

- **1 – Vertex Cliques:** All the vertices in the graph, from vertex A to vertex W. Total = *23*.
- **2 – Vertex Cliques:** All the edges in the graph, $AB, BC...VW$. Total = *42*.
- **3 – Vertex Cliques:** All the light and dark blue triangles, $ABC, BCD...UVW$. Total = *19*.
- **4 – Vertex Cliques:** Both the dark blue coloured area, $ABCD$ and $JKLM$. Total = *2*.

A *Maximal Clique* is a clique that the inclusion of any additional adjacent vertices cannot expand. This definition implies that a maximal clique is not wholly contained within the vertex set of a larger clique. It represents a subset of vertices that forms a complete graph and is as large as possible without being subsumed by a greater clique. This characteristic makes maximal cliques particularly significant in the network analysis, as they help identify the largest groups that are connected.

The graph in Figure 2.9 contains a total of *13* maximal cliques. All the *11* light blue triangles and the *2* dark blue areas are maximal cliques. Note how ABC is a clique but not a maximal clique. By adding

[11] https://en.wikipedia.org/wiki/Clique_(graph_theory)

another vertex, D, we can convert it to a larger clique i.e. $ABCD$. Similarly, ACD, BCD and ABD are not maximal cliques.

Lastly, a *Maximum Clique* is a clique which has the highest number of vertices in the entire network. The number of vertices in a maximum clique is also referred to as the *Clique Number*. The graph depicted in Figure 2.9 contains *2* maximum cliques, which are $ABCD$ and $JKLM$. The clique number for this network is *4*.

2.2.2 Frequent Itemsets & Maximal Frequent Itemsets

In Data Mining, the concept of *Frequent Itemsets*, often associated with association rule mining, is leveraged to extract relationships among items within a dataset. A frequent item set comprises a combination of items that appear together with considerable regularity in the dataset, quantified by a support count. This count reflects the number of transactions or records in which the item set occurs. For instance, in a dataset containing *100* transactions, an item set {*coffee, sugar*}, which, if found in *15* out of *100* transactions, would have a support count of *15*.

Maximal Frequent Itemsets, are similar to maximal cliques in graph theory. A maximal frequent itemset is defined as a frequent itemset that cannot be enlarged by adding another element without losing its frequent status in the dataset. None of its immediate supersets appears frequently enough to meet the minimum support criterion, making these itemsets a concise and compact representation of all frequent itemsets within a dataset. We concentrate on analysing Maximal Frequent Itemsets among the datasets of Directors and Companies, offering a compact and comprehensive view of item associations.

Various algorithms such as *Apriori*[12] or *FP-Growth*[13] are utilised to extract frequent itemsets. These methods iteratively generate item sets and eliminate those that do not meet the predefined *minimum support* threshold. These frequent itemsets are then used to generate association rules. These rules can be thought of as *"if-then"* statements that help us with the likelihood of connections between various data elements across extensive databases. The application of frequent item sets and the resultant association rules extend to multiple domains such as *market basket analysis*, *cross-selling strategies*, and *recommendation systems*, offering valuable insights into consumer behaviour and item associations.

2.2.3 Apriori Algorithm

The *Apriori Algorithm*, developed by Agrawal et al. (1994), is a fundamental method for *frequent itemset mining* and *association rule learning* within relational databases. The algorithm is aptly named *"Apriori"* because it leverages *prior* knowledge of the properties of frequent item sets to enhance its efficiency. It begins by identifying the frequent individual items in the database. It methodically expands these into increasingly larger itemsets, provided they meet a predefined frequency threshold within the database, *minimum support*.

[12] https://en.wikipedia.org/wiki/Apriori_algorithm
[13] https://www.geeksforgeeks.org/frequent-pattern-growth-algorithm/

In practical terms, the Apriori algorithm views each transaction within the database as a set of items, referred to as an itemset. Utilising a specified *minimum support* threshold, it identifies those item sets that appear as subsets in at least that many transactions. This algorithm employs a *"bottom-up"* approach for candidate generation, where initially discovered frequent subsets are incrementally expanded by adding one item at a time. This sequence of generating and testing candidate itemsets continues iteratively, ultimately concluding when forming larger frequent item sets is no longer possible. This systematic approach not only ensures comprehensive data analysis but also enhances the efficiency of discovering valid association rules. Figure 2.10 is a flowchart explaining the Apriori Algorithm.

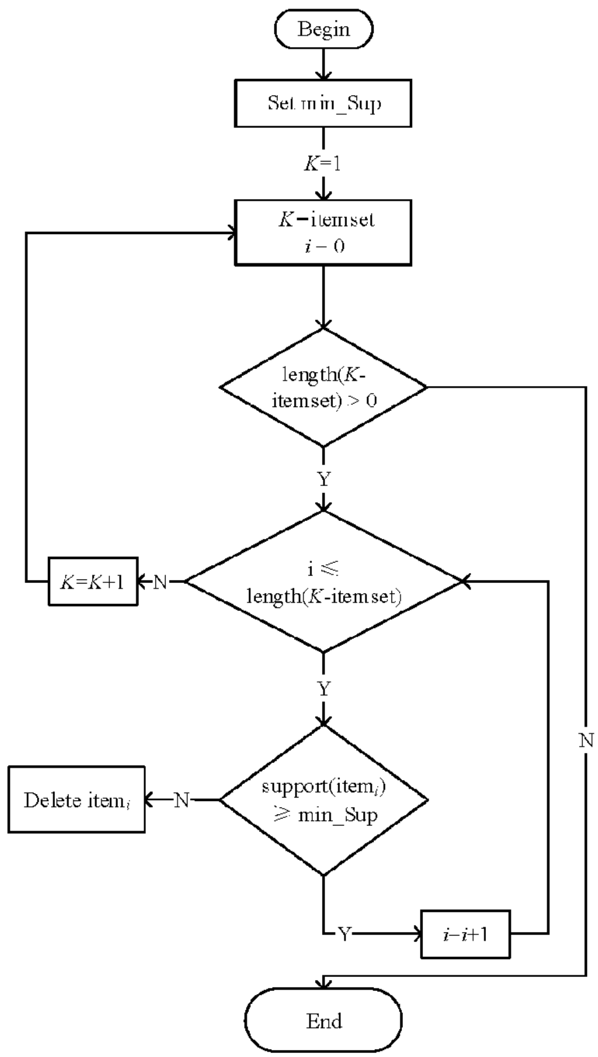


Figure 2.10: Flow Chart of *Apriori Algorithm* for Extracting Frequent Itemsets (Chen et al., 2020)

The time and space complexity of the Apriori algorithm is notably high, often characterised as exponential, specifically $O(2^{|D|})$, where D represents the horizontal width or the total number of items present in the database. This complexity arises because the algorithm potentially considers all subsets of items from the database as candidate item sets, necessitating significant computational resources as the size of D increases. This aspect of Apriori underscores the challenges in scaling the algorithm for extensive databases, where the exponential growth in possible item combinations can lead to substantial increases in both computation time and memory usage.

2.2.4 Frequent Pattern Growth (FP – Growth) Algorithm

The *Frequent Pattern (FP) Growth Algorithm*, developed by Han et al. (2000), is another popular algorithm in the field of data mining, renowned for its superior efficiency in extracting frequent itemsets from large scale datasets.

2.2.4.1 FP – Tree

Central to this algorithm is constructing a *Frequent Pattern Tree (FP – Tree)*, which uses a tree-like data structure to represent the dataset in a compact yet effective manner, capturing the essential transaction information. FP – Tree construction begins with each transaction being read and then mapped onto a specific path in the data structure. This method continues sequentially until all transactions in the dataset have been read and appropriately integrated into the tree structure.

We discuss *FP – Tree* construction with an example. Table 2.1 is a sample dataset of items bought at a store. The frequency of each item purchased in the dataset is as follows – *diaper: 4, beer: 3, nuts: 3, eggs: 3, coffee: 2, milk: 2*. Assuming the minimum support as 2, each transaction is added to the tree.

Table 2.1: Sample Transaction Data for *FP – Tree* Construction

Transaction ID	**Items Bought**
1	diaper, beer, nuts
2	eggs, diaper, beer
3	nuts, eggs, diaper, coffee, milk
4	diaper, beer, coffee
5	nuts, eggs, milk

The resulting FP - Tree for the data in Table 2.1 is depicted in Figure 2.11. Instead of using *17* distinct nodes to display the data, the FP – Tree data structure helps reduce it to just *12* nodes. The space efficiency of an FP – Tree is truly reflected in a large-scale dataset with a high number of common terms.

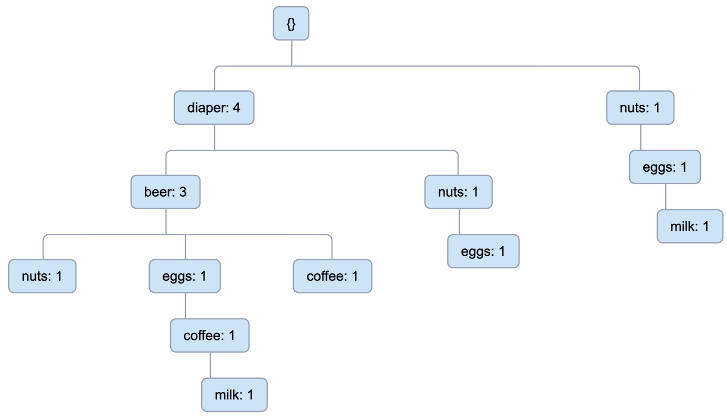


Figure 2.11: Sample *FP – Tree* Construction corresponding to transactions in Table 2.1

2.2.4.2 FP – Growth Algorithm Functioning

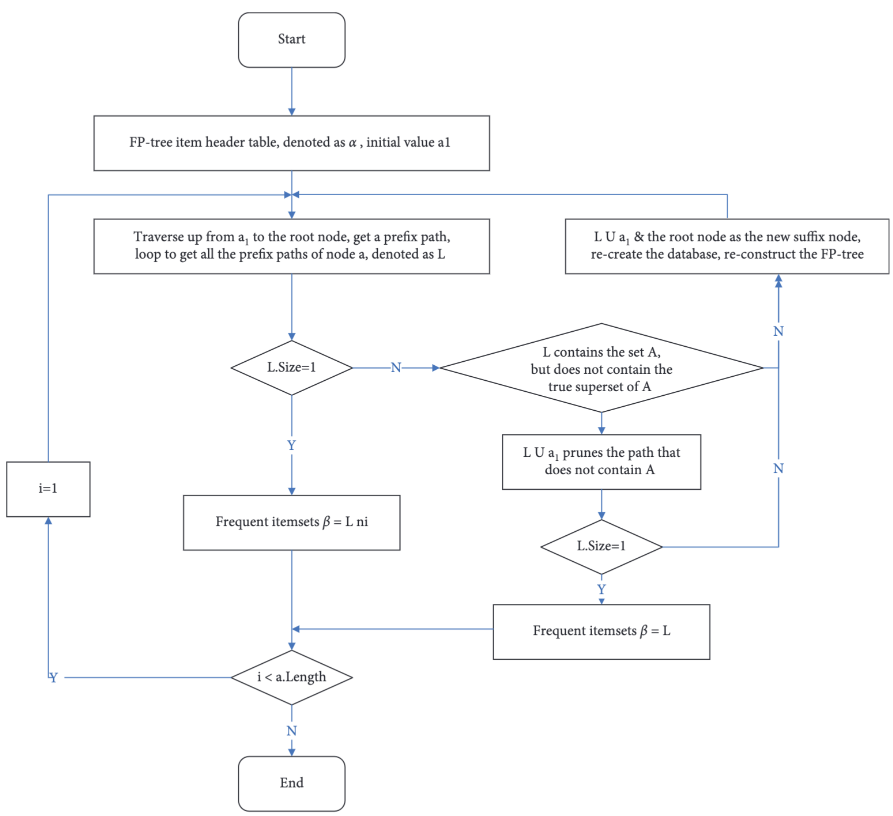


Figure 2.12: Flow Chart of *FP – Growth Algorithm* for Extracting Frequent Itemset (Hu, 2022)

Functionally, the FP Growth algorithm commences by conducting a single scan of the dataset, during which it maps each transaction into a corresponding path within the FP-tree. In building the tree, items within each transaction are sorted based on frequency, positioning the most frequent items at the forefront. This sorting ensures that the most common items are easily accessible, thereby facilitating efficient pattern extraction.

Once the FP-tree is entirely constructed, the algorithm generates frequent item sets by recursively mining the tree. This mining process is initiated from the bottom of the tree and proceeds upwards in a bottom-up manner. It systematically explores all possible combinations of itemsets, identifying those that meet or exceed the minimum support threshold. Through this systematic exploration, the FP Growth algorithm effectively uncovers all significant item sets without generating candidate sets, making it substantially faster and less resource-intensive than the Apriori algorithm.

The FP Growth algorithm's effectiveness and efficiency have led to its widespread application across various domains, including market basket analysis, bioinformatics, and web usage mining, where understanding frequent patterns is crucial. Figure 2.12 is a flowchart explaining the functioning behind the FP – Growth Algorithm for extracting frequent itemsets.

Chapter 3

LLM Driven Web Profile Extraction for Identical Names

In this chapter, we present an innovative solution to tackle the challenges stemming from the widespread occurrence of identical person names in online contexts. In today's digital landscape, an individual's online presence holds comparable significance to their physical presence, underscoring the critical need for a reliable method to locate web profiles of individuals who share the same name. Our objective is to develop a comprehensive pipeline that not only extracts web profiles of individuals with a shared name but also consolidates them based on profile similarity. This enables the user by providing them with an integrated, consolidated view of multiple web profiles associated with the same name, thereby facilitating swift and accurate identification.

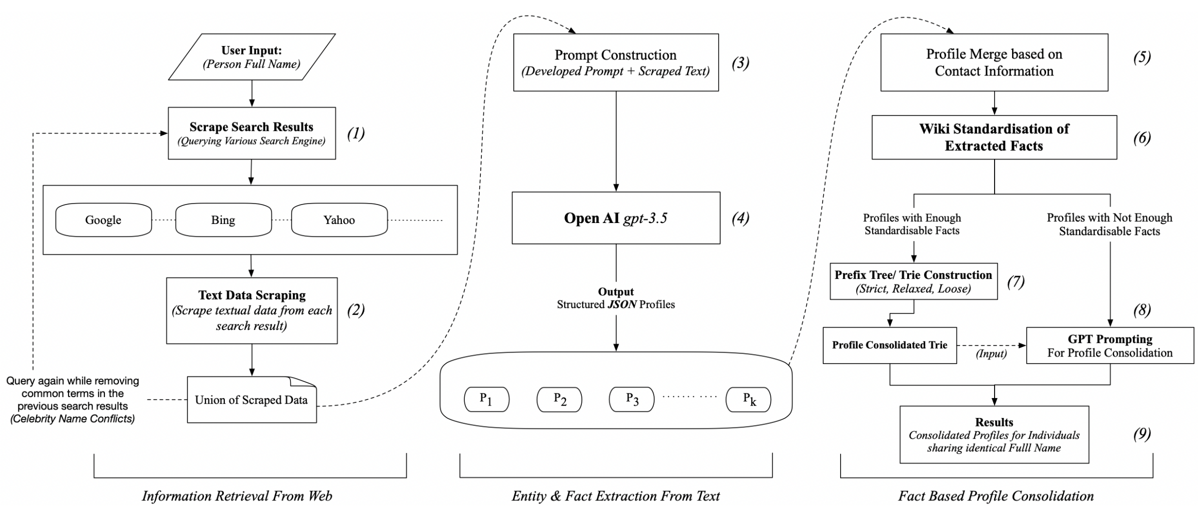


Figure 3.1: Pipeline Flowchart for LLM Driven Web Profile Extraction for Identical Names

While previous studies in this domain have primarily relied on ML techniques to achieve SOTA results, they often necessitate extensive task-specific training. In contrast, our approach endeavours to create a more adaptable solution capable of comprehending the task and its contextual nuances akin to human reasoning before consolidating or disambiguating profiles. Grounded on several foundational pillars, including entity extraction utilizing Large Language Models (LLMs), prompt engineering, standardization of extracted data using Wikipedia, a data structure-based profile consolidation (DeDuplication), and leveraging the contextual understanding of LLMs, our solution offers a holistic framework for addressing these challenges. We also discuss how our solution handles the issue of search result domination arising from celebrity name conflicts by refining search queries based on the frequency of common terms in search results.

3.1 Data Extraction from Web

When seeking information about a specific individual, a person commonly utilizes a variety of search engines over the web by inputting the person's name as the query. These search engines then provide a selection of search results, which the user must navigate to locate the intended individual.

The user provides a *Person Full Name (First Name and Last Name)* as input, which acts as the search query. The initial step involves extracting textual information from web search results relevant to the specified name to process web profiles associated with a given name. To ensure precise matching of the provided name within the search results, we encapsulate the given name within *double-quotes (" ")*. This search methodology guarantees an exact match rather than a relative one, facilitating more accurate results retrieval.

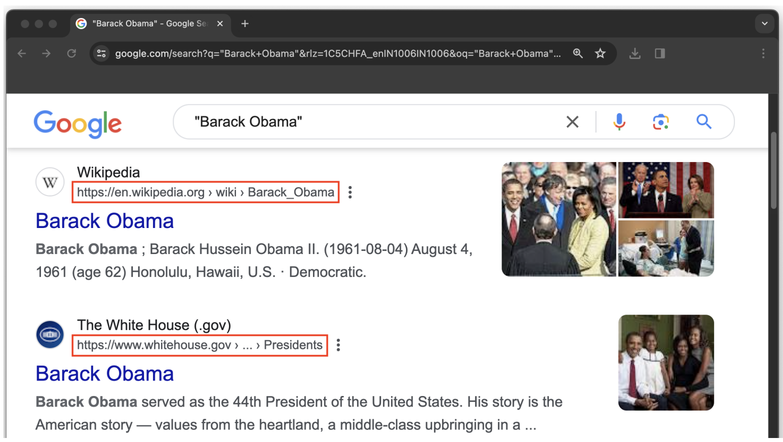


Figure 3.2: Person Name Search on Web

Figure 3.2 illustrates the appearance of a typical web search for a person. The search results are depicted with the web links highlighted in red boxes.

To extract textual information pertaining to the provided person's name over the web, we first need to retrieve these search result web links from various search engines. For our study, we focus on the three most popular search engines: *Google*[1], *Bing*[2], and *Yahoo*[3] *(Figure 3.1 (1))*.

3.1.1 Extraction of Search Results

We use different tools for different search engines to extract search results.

[1] www.google.com
[2] www.bing.com
[3] www.yahoo.com

3.1.1.1 Google Search Results

We utilize the ***search*** function provided by the ***googlesearch***[4] *Python* library, which simplifies the process of querying Google. This library leverages **requests** and **BeautifulSoup4** libraries for scraping Google Search Results.

```
googlesearch.search(query, tld='com', lang='en', tbs='0',
                    safe='off', num=10, start=0, stop=None,
                    pause=2.0, country='', extra_params=None,
                    user_agent=None, verify_ssl=True)
```

In our specific use case, the parameters are configured as follows:

- **query:** Query string. Must NOT be URL-encoded — *"Person Full Name"*
- **num:** Number of results per page — *10* (default)
- **stop:** Last result to retrieve — *100*
- **country:** We do not specify any particular country to ensure generic search results.

```
Fetching Search Results for query: Barack Obama
Number of Results requested: 100

Fetching Google Search Results
query: "Barack Obama"
Google Search Results Fetch Completed


Google Search Results:
1. https://en.wikipedia.org/wiki/Barack_Obama
2. https://barackobama.com/
3. https://www.whitehouse.gov/about-the-white-house/presidents/barack-obama/
4. https://twitter.com/BarackObama?ref_src=twsrc%5Egoogle%7Ctwcamp%5Eserp%7Ctwgr%5Eauthor
5. https://twitter.com/BarackObama/status/1771595568457957797?ref_src=twsrc%5Egoogle%7Ctwcamp%5Eserp%7Ctwgr%5Etweet
6. https://twitter.com/BarackObama/status/1771569744635707752?ref_src=twsrc%5Egoogle%7Ctwcamp%5Eserp%7Ctwgr%5Etweet
7. https://twitter.com/BarackObama/status/1771552747306791327?ref_src=twsrc%5Egoogle%7Ctwcamp%5Eserp%7Ctwgr%5Etweet
8. https://twitter.com/BarackObama/status/1771522445368787059?ref_src=twsrc%5Egoogle%7Ctwcamp%5Eserp%7Ctwgr%5Etweet
9. https://www.barackobama.com/
10. https://www.instagram.com/barackobama/?hl=en
```

Figure 3.3: Extraction of Google Search Results

3.1.1.2 Bing & Yahoo Search Results

To extract search result web links from *Bing* and *Yahoo*, we employ **Python** as our foundational language, augmented with the **Selenium**[5] and **BeautifulSoup**[6] libraries.

Selenium serves as a digital assistant for automating web browsing tasks and mimicking human-like interactions. With Selenium, users can seamlessly navigate web pages by clicking buttons, inputting text, and interacting with diverse elements on a web page. Our workflow begins with opening a search engine page (either Bing or Yahoo) using a Selenium Driver. Subsequently, we direct Selenium to locate

[4] https://pypi.org/project/googlesearch-python/
[5] https://pypi.org/project/selenium/, https://selenium.dev
[6] https://www.crummy.com/software/BeautifulSoup/

the search bar, input our query (formatted as *"Person Full Name"*), initiate the search by clicking the search button, and await the loading of results. Once the results are reflected, we extract the webpage source code, which will be processed subsequently – particularly for retrieving web links. We then instruct Selenium to navigate to the next page by clicking the designated button. This iterative process continues until the desired number of search results is attained, all seamlessly executed through automation. Figure 3.4 presents a snapshot captured during an automated web search on Bing search using the Selenium framework.

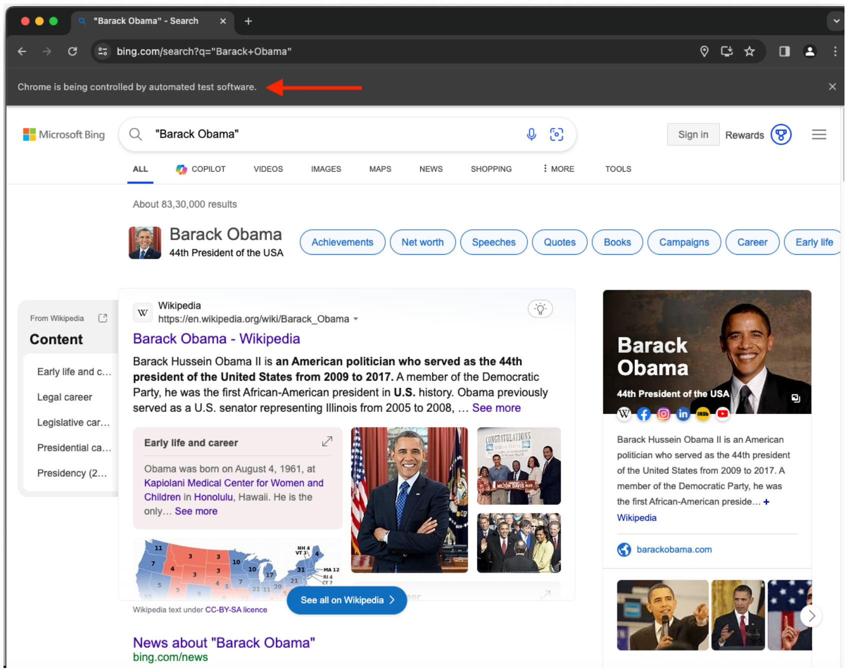
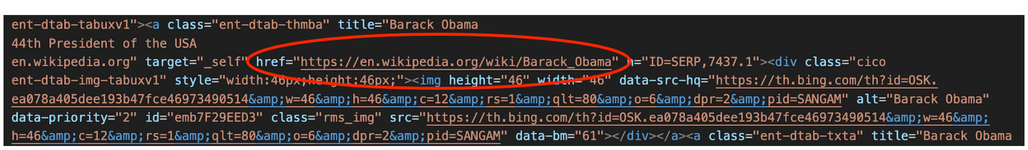


Figure 3.4: Automated web search using *Selenium WebDriver*

BeautifulSoup is another Python library complementing Selenium by parsing *HTML* and *XML* documents to simplify the extraction of pertinent data from web pages. The previously extracted source code is parsed through BeautifulSoup to extract web links effectively. In Figure 3.5, we can see how the web links for search results are hidden deep within the website's code. We meticulously extract the necessary details (text) from the extensive HTML code using functions such as **'find_all'** and **'get'** from the library.



```
ent-dtab-tabuxv1"><a class="ent-dtab-thmba" title="Barack Obama
44th President of the USA
en.wikipedia.org" target="_self" href="https://en.wikipedia.org/wiki/Barack_Obama" h="ID=SERP,7437.1"><div class="cico
ent-dtab-img-tabuxv1" style="width:46px;height:46px;"><img height="46" width="46" data-src-hq="https://th.bing.com/th?id=OSK.
ea078a405dee193b47fce46973490514&amp;w=46&amp;h=46&amp;c=12&amp;rs=1&amp;qlt=80&amp;o=6&amp;dpr=2&amp;pid=SANGAM" alt="Barack Obama"
data-priority="2" id="emb7F29EED3" class="rms_img" src="https://th.bing.com/th?id=OSK.ea078a405dee193b47fce46973490514&amp;w=46&amp;
h=46&amp;c=12&amp;rs=1&amp;qlt=80&amp;o=6&amp;dpr=2&amp;pid=SANGAM" data-bm="61"></div></a><a class="ent-dtab-txta" title="Barack Obama
```

Figure 3.5: Extraction of WebLink using *BeautifulSoup*

These libraries provide a robust framework for efficiently retrieving and processing search results from Bing and Yahoo search engines.

3.1.2 Remove Duplicate and Unnecessary Links from Search Results

When conducting searches with identical queries across different search engines, it's common to encounter overlapping search results. To streamline processing and conserve computational resources, our initial step involves eliminating duplicate search results by generating a set comprising unique search results.

Once duplicate links have been removed, the next step in our pipeline involves eliminating extraneous links that are unlikely to contribute to the creation of comprehensive, data-driven web profiles. This entails filtering out sources of information deemed unreliable or irrelevant for our purposes. By doing so, we aim to refine our dataset to include only reputable sources and websites conducive to constructing robust web profiles, as necessitated by the subsequent stages of our pipeline. This includes websites that are primarily utilized for opinion or media sharing by its users, such as:
X (formally Twitter)[7]***, Instagram***[8]***, Facebook***[9]***, Reddit***[10]***, YouTube***[11]***, Pinterest***[12] ***& Quora***[13]. Any search result that contains these domain names is removed.

3.1.3 Extraction of Textual Information from Search Results

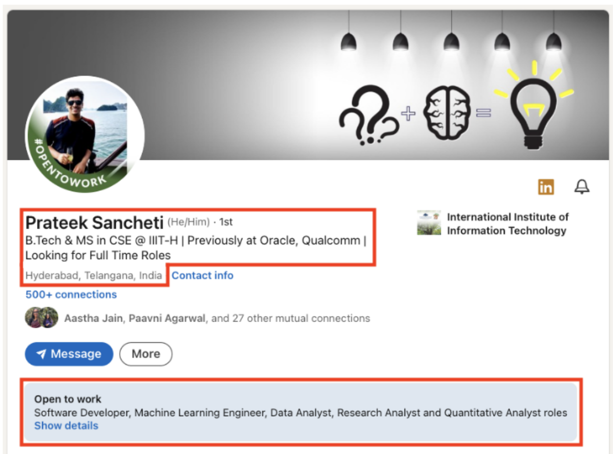
After filtering out duplicate and unnecessary links, we are left with distinct search results corresponding to the provided input name. To proceed with our pipeline to extract and consolidate web profiles for identical names, it is imperative to extract plain text from the webpages associated with these search results *(Figure 3.1 (2))*. This task is accomplished through a combination of the fundamental web page automation techniques offered by *Selenium* and the **get_text** function offered by *BeautifulSoup*.

[7] https://twitter.com/home
[8] https://www.instagram.com/
[9] https://www.facebook.com/
[10] https://www.reddit.com/
[11] https://www.youtube.com/
[12] https://www.pinterest.com/
[13] https://www.quora.com/

3.1.3.1 LinkedIn Text Scraping

A LinkedIn profile is an important source of information for a given name. Several fields such as *Headlines, Education, Work Experience, License, etc* provide professional information about a person.

Figure 3.6 and Figure 3.7 and attached text below depict textual scrapping from a LinkedIn Profile[14].

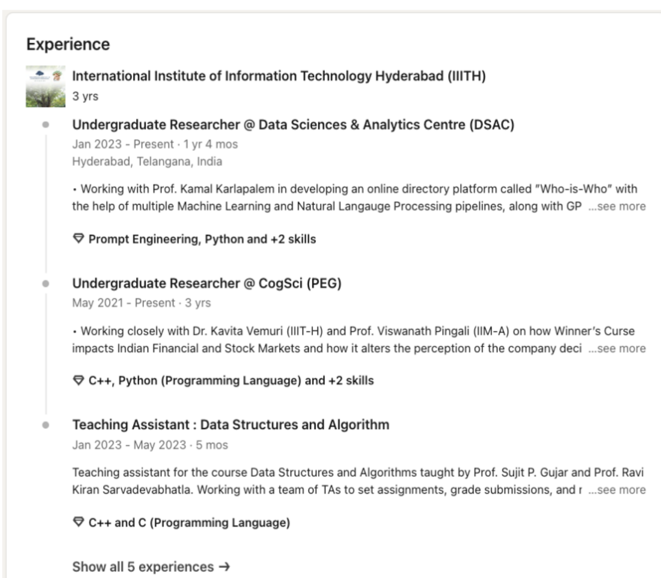


Extracted Text:

```
Prateek Sancheti (He/Him) B.Tech & MS in CSE @ IIIT-H | Previously at Oracle, Qualcomm | Looking for Full Time Roles Hyderabad, Telangana, India Contact info

Open to work Software Developer, Machine Learning Engineer, Data Analyst, Research Analyst and Quantitative Analyst roles
```

Figure 3.6: Sample Text Scraping from LinkedIn (Headline)



Extracted Text:

```
International Institute of Information Technology Hyderabad (IIITH) 3 yrs Undergraduate Researcher @ Data Sciences & Analytics Centre (DSAC) Jan 2023 - Present·1 yr 4mos Hyderabad, Telangana, India. Working with Prof. Kamal Karlapalem in developing an online directory platform called "Who-is-Who" with the help of multiple Machine Learning and Natural language processing pipelines, along with GPT integration to increase the accuracy and efficiency of the platform.
.
.
```

Figure 3.7: Sample Text Scraping from LinkedIn (Experience)

[14] www.linkedin.com/in/sanchetiprateek/

3.2 Large Language Models Driven Name Entity Recognition

3.2.1 The Emergence of Large Language Models for NLP Tasks

LLMs and GPTs with their *Deep Neural Network architecture* and *Pre-Trained Large Language Models (LLMs)* have shown significant performance enhancements in recent years across various NLP tasks (Brown et al., 2020; Chowdhery et al., 2023; Hoffmann et al., 2022; Rae et al., 2021; Smith et al., 2022). They exhibit unprecedented complexity pertaining to their architecture and extensive training, enabling them to comprehend diverse human queries without explicit training, effectively replacing many conventional NLP models.

Their capabilities span across various tasks such as *Machine Translation* (Freitag et al., 2022), *Question Answering* Robinson et al. (2022) and *Named-Entity Recognition* Wang et al. (2023) and beyond, thus revolutionizing the field, showcasing remarkable prowess in understanding and generating natural language text by leveraging extensive corpora and capturing intricate linguistic patterns and semantic relationships. The ability to comprehend contextual information across different domains makes it suitable for various linguistic tasks (Baktash and Dawodi, 2023). Continuous efforts are being made to enhance the capabilities of LLMs towards improved efficiency and accuracy in tasks of NER (Ashok and Lipton, 2023; Wang et al., 2023).

Two prominent strategies that emerge while leveraging the Large Language Models (LLMs) for downstream tasks are: *fine-tuning* and *in-context learning (ICL)* (Wang et al., 2023). The Fine-tuning strategy involves initializing a pre-trained model and conducting additional training epochs on downstream supervised data, (Gururangan et al., 2018; Guu et al., 2020; Raffel et al., 2020; Roberts et al., 2020). Conversely, in-context learning or ICL prompts LLMs to generate text based on few-shot demonstrations. This approach was pioneered by Radford et al. (2019), who introduced prompts containing demonstrations to reframe downstream tasks. Brown et al. (2020) expanded on this concept through systematic analyses and experiments with GPT-3 across various tasks. Moreover, various research underscores the significance of refining prompts and demonstrations, showcasing that enhancements in these areas can significantly augment the performance of in-context learning strategies (Lu et al., 2021; Perez et al., 2021; Rubin et al., 2021).

Over the past few years, there has been a significant emergence of generative deep learning models based on Transformer architectures, marking a notable advancement in the field. Models like **GPT-3**[15] *(May 2020)*, **GPT-4**[16] *(March 2023)*, **LLaMA**[17] *(Feb 2023)*, and **Mixtral 8x7B**[18] *(Dec 2023)* stand at the forefront of this revolution, profoundly impacting numerous traditional Natural Language Processing use cases. Notably, named-entity recognition is among the areas experiencing a transformative shift.

To explore the applicability of Large Language Models for our specific use case, we turn to ***gpt 3.5 - turbo****, a fine-tuned version of gpt-3*, developed by *OpenAI*. The following section covers our journey in

[15]https://en.wikipedia.org/wiki/GPT-3
[16]https://openai.com/gpt-4
[17]https://llama.meta.com/
[18]https://mistral.ai/news/mixtral-of-experts/

leveraging LLMs for Named Entity Recognition. We delve into the various recent studies to cover the domain and the concept of prompts and explain our engineering approach to tailored prompts per our requirements. Additionally, we examine the advantages of employing LLMs for NER tasks, highlighting their potential to revolutionize traditional methodologies. Amidst this exploration, we also acknowledge various challenges inherent to this approach and detail our strategic address, ultimately paving the way for a nuanced understanding of Driven NER. This section delves into how this groundbreaking technology becomes a solution for our application, effectively addressing the challenges encountered with SOTA NER models discussed earlier.

3.2.2 Experiments with NER Task facilitated by Large Language Models

We perform experiments with several prompts instructing the Language Model to execute Named Entity Recognition (NER) tasks. To accomplish this, we'll utilize the **ChatGPT**[19] interface, which is also developed by OpenAI and employs the *GPT-3.5* model. To provide a comparison, we use the same text used for testing SOTA NER Models (2.1.3).

Experiment 1

Table 3.1 shows how, with a simple one-line prompt, ChatGPT extracts all the entities from the given text and assigns them relevant entity categories by itself. It can also identify the entities, such as DOB, Office Years and Previous Jobs, that previously tested SOTA NER missed (See 2.1.3).

Table 3.1: *Named-Entity Recognition (NER)* Task output using *ChatGPT* – Experiment *1*

→ ***Text:*** "Barack Hussein Obama II (born August 4, 1961) is an American politician who served as the 44th president of the United States from 2009 to 2017. A member of the Democratic Party, he was the first African-American president. Obama previously served as a U.S. senator representing Illinois from 2005 to 2008 and as an Illinois state senator from 1997 to 2004, and worked as a civil rights lawyer and university lecturer." (Wikipedia, 2023)

→ ***Prompt:*** Read the text attached carefully and extract all entities

→ ***Output:*** Here are the extracted entities from the given text:

- Full Name: Barack Hussein Obama II
- Birthdate: August 4, 1961
- Title/Position: 44th president of the United States (2009 to 2017)
- Political Affiliation: Member of the Democratic Party
- Historical Significance: First African-American President
- Region: Illinois
- Previous Political Roles: [U.S. Senator – Illinois (2005-2008), Illinois Senator (1997-2004)]
- Previous Professions: [Civil rights lawyer, University Lecturer]

Although the NER task performed by ChatGPT exhibits commendable accuracy, certain aspects merit attention. Firstly, while the extracted entities are generally correct, the model's categorization method appears self-assigned. This means that when presented with a similar prompt and different text, the model might assign dissimilar entity categories to comparable entities. For instance, it could categorise *"Barack Hussein Obama II"* under *"Full Name"* or *"Person Name"* instead of simply *"Person"*. Secondly, the format in which the entities are extracted appears suitable for human comprehension but may not be optimized for unsupervised computer systems. This text-based format may necessitate further processing for seamless integration into such systems.

[19]https://chat.openai.com/

Experiment 2

Impressed by the efficiency demonstrated by GPT in extracting entities from a given text through a one-line prompt, we were curious whether the same approach could effectively address the *entity resolution* challenge posed by SOTA NER methods (See 2.1.4.3). To further investigate this, we conducted another experiment with ChatGPT. Table 3.2 illustrates how adding a single phrase to the prompt enabled the model to precisely extract information related to the subject (in this instance, Clare). This capability holds significant relevance for our specific use case, as it allows for extracting pertinent information about a particular individual from vast datasets without supervision.

Table 3.2: *Named-Entity Recognition (NER)* Task output using *ChatGPT* – Experiment *2*

<table>
<tr><td colspan="2">→ Text: “In today’s dynamic corporate world, Clare stands as a pillar of JP Morgan Chase, leveraging her financial understanding and dedication to propel the institution forward. Meanwhile, Jacob thrives within the innovative ecosystem of Google, and his expertise in technology shapes the company’s pioneering ventures. Clare’s role encompasses financial analysis, risk management, and strategic planning, while Jacob delves into software engineering, data analytics, and artificial intelligence. Their professional synergy underscores the diverse skill sets and collaborative spirit driving success in modern organizations.”</td></tr>
<tr><td>→ Prompt 1: Read the text attached carefully and extract all entities</td><td>→ Prompt 2: Read the text attached carefully and extract all entities only related to <u>Clare</u></td></tr>
<tr><td>→ Output 1: Here are the extracted entities from the given text:
• Person: Clare
• Organization: JP Morgan Chase
• Person: Jacob
• Organization: Google
• Job Title/Skill: Financial Analysis
• Job Title/Skill: Risk Management
• Job Title/Skill: Strategic Planning
• Job Title/Skill: Software Engineering
• Job Title/Skill: Data Analytics
• Job Title/Skill: Artificial Intelligence</td><td>→ Output 2: Here are the extracted entities only related to “Clare” in the given text:
• Person: Clare
• Organization: JP Morgan Chase
• Job Title/Skill: Financial Analysis
• Job Title/Skill: Risk Management
• Job Title/Skill: Strategic Planning</td></tr>
</table>

After observing the results from experiments involving diverse texts and subjects, we recognized the potential of this powerful technology for our specific use case. Understanding the necessity of providing precise instructions to the GPT model is paramount for achieving accurate and efficient data retrieval. This journey toward meticulously crafting a prompt to accomplish a specific task is known as **prompt engineering**. In the subsequent section, we delve into the various studies that discuss the techniques for the same.

3.2.3 Evolution of GPT Prompt for *NER* Task

Prompt Engineering encapsulates the systematic process of structuring text to convey instructions that can be comprehended and executed by generative AI models. At its core, a prompt represents natural language text delineating the task the model is expected to undertake. This discipline leverages the concept of *in-context learning*, characterized by a model's capacity to assimilate knowledge from provided prompts temporarily. The process of *prompt engineering* constitutes a sophisticated technique aimed at refining the input instructions directed to generate desired outputs. Successful prompt engineering necessitates a profound comprehension of model capabilities and inherent biases, empowering greater control over responses. It constitutes a strategic endeavour to harness the potential of AI while upholding principles of responsibility and accuracy.

Wang et al. (2023) and Ashok and Lipton (2023), suggest multiple techniques to fine-tune *GPT* results and *engineer prompts* for our NER task without *extensive task-specific engineering*. We use various techniques such as ***(A) Task Description, (B) Predefined Requirements, (C) Edge Case Handling, (D) Output Formatting, (E) Self Verification & Reinforced Learning*** *(Figure 3.1 (3))*, to converge on a prompt that extracts information from a given text about the subject in a structured format.

(A) Task Description: An introductory insight into the task

→ *You are an excellent linguistic assistant.*

→ *Your task is to help users extract information about a person's profile from a given text in a structured format.*

(A) Task Description

With the above text as the initial context, the LLM functions as a *Linguistic Assistant*. It leverages its expertise while performing the task of extracting information from a given text.

(B) Predefining Requirements: As discussed before GPT *self-assigns* the *categories* to the extracted entities. While GPT identifies and categorizes all entities within the text, the resulting output is unsuitable for direct processing. Relying solely on the automatically assigned keys can introduce ambiguity as GPT could assign different keys to similar entity types, potentially complicating subsequent stages that involve text comparison. Therefore, to refine the model's outputs, we explicitly define the requirements that GPT should consider before returning results. This includes specifying how entities should be labelled and how to handle exceptional cases. To tackle this issue, we employ the following instructions for guiding the model:

Retrieve the information requested below about the subject "Name" and return the same in a structured format. Perform the entity extraction on the text delimited by triple back ticks:

→ 'Education' *(Keys → 'Institute Name', 'Course Name', 'Branch', 'Grade', 'Duration')*

→ 'Work Experience' *(Keys → 'Organization Name', 'Role', 'Job 'Description', 'Duration')*

(B) Predefining Entity Attributes / Sub-Attributes

We pass the following text as a test for the above prompt,

Text: ```*Eric Wu is a dedicated and ambitious B.Tech student currently pursuing his degree in the field of CSE at NVS since 2019. Hailing from the vibrant city of Shanghai (China), Eric gained valuable work experience within renowned companies such as Oracle and Qualcomm. Eric has a friend, Jim Lee, who works at Google.*```

Output:

```
{
    "Education"
    {
        "Institute Name" : NVS,
        "Course Name" : B.Tech,
        "Branch" : CSE,
        "Duration" : 2019 -- Currently Pursuing
    }
    "Work Experience"
    [
        {
            "Organization Name" : Oracle
        },
        {
            "Organization Name" : Qualcomm
        }
    ]
}
```

Passing the *Keys* explicitly ensures that the model does not self assign keys to entities. The model also takes care of *entity resolution* and returns entities *only* related to the subject *"Eric Wu"* and not to their friend, as we strictly instructed in the above prompt *(B)*.

(C) Edge Cases: Notice that some keys, such as *'Grade', 'Job Role', 'Job Description'*, are missing in the output. This is because the requested information is not present in the given text. We instruct the model to handle such cases as:

> $\rightarrow$ *If you are unable to find any information requested by the user, return it as 'Unknown'*

(C) Edge Case: Missing Information

We passed the same text as input with *Edge Case* instruction *(C)*. The output for the same was:

Output:

```
{
    "Education"
    {
        "Institute Name" : NVS,
        "Course Name" : B.Tech,
        "Branch" : CSE,
        "Duration" : 2019 -- Currently Pursuing,
        "Grade" : Unknown
    }
    "Work Experience"
    [
        {
            "Organization Name" : Oracle,
            "Job Role" : Unknown,
            "Job Description" : Unknown
        },
        {
            "Organization Name" : Qualcomm,
            "Job Role" : Unknown,
            "Job Description" : Unknown
        }
    ]
}
```

(D) Output Formatting: GPT has the capacity to provide the generated output in various formats Wang et al. (2023), which include *XML, HTML, JSON,* or even *Plain Text*. We must explicitly instruct the model to return the results in a useful format for our use case. We decided to retrieve results from GPT in *JSON* format.

> → *Return the results from the above Entity Extraction task in a structured JSON Format.*

(D) Output Formatting Instruction

(E) Self Verification & Reinforced Learning: Another concern with LLMs, as highlighted in prior studies, is the *hallucination* and *bias* issue. LLMs tend to confidently identify empty or null inputs as important entities Braverman et al. (2020); Jiang et al. (2021); Zhao et al. (2021). To handle this, Wang et al. (2023) and Pang et al. (2023) suggest a *Self-Verification* approach that is positioned right after the entity extraction phase. It prompts the LLMs to check whether the identified entities correspond to their labelled entity tag and that the requested task is properly completed. In our case, we use the following set of instructions to add self-verification:

> → *Check again that all the extracted entities are present in the input text and are assigned to the correct attributes as requested by the user.* – (Self Verification)
>
> → *Double check if all the requirements stated by the user are completely fulfilled. Take your time before returning results to the user.* – (Self Verification)
>
> → *If any of the instruction is not followed, you will be heavily penalized.* – (Reinforced Learning)

(E) Self Verification Instructions

We observed that the *Self Verification (SV) and Reinforced Learning (RL)* Pang et al. (2023) based approach effectively alleviates the issue of hallucination and results in increased accuracy while extracting entities.

We added a few more attributes in the *prompt*, similarly as described previously *(B)*, that need to be extracted to cover almost all relevant information that builds an online profile. We now have our *Final Prompt: [(A)+(B)+(C)+(D)+(E)]* that extracts all entities from the provided text, assigns keys to these entities based on instructions, manages any potential edge cases, validates its output and returns the results in the desired output format. *Table 3.3* is an exhaustive list of all the distinct attributes *(Total: 14)* obtained from the final prompt, along with their description.

Table 3.3: Description of Categories for Extracted Entities Using GPT

Note: If no entity is detected for an attribute/sub-attribute, it is mentioned as *'Unknown'* in the output

Category Name	**Category Description**
Name	*Name* of the subject in provided text
Age	*Age* of the subject
Birth Date	*Birth Date* of the subject in *DD-MM-YYYY* Format
Contact Information	List of *Email IDs*, *Mobile Numbers*, *Weblinks (Website, Blogs etc)*
Locations	List of *Locations* mentioned in relation to the subject
Education	List of *Institute Name, Course Name, Branch, Grade and Duration*
Work Experience	List of *Organization Name, Role, Job Description and Duration*
Projects	List of *Project Name, Organization Name, Project Description and Duration*
Interests	List of *Interests* mentioned in relation to the subject
Licenses & Certifications	List of *Name, Issued By, Credential ID, Issue Date*
Volunteer Experience	List of *Organization Name, Duration, Description*
Causes	List of *Causes* mentioned in relation to the subject
Test Scores	List of *Test Name, Test Score, Test Date*
Miscellaneous	List of all other information that is relevant in distinguishing a person from another but does *NOT* fit in any of the above attributes.

Andrew Ng[20], a well-known computer scientist in the industry, in collaboration with *OpenAI* introduced an online course *"ChatGPT Prompt Engineering for Developers"*[21]. This course covers a few more fundamental principles that go handy in crafting a descriptive prompt for a given task.

Basic Principles of Prompt Engineering:

1. **Write clear and specific instructions**

   - Short Prompt $\neq$ Good Prompt: The Prompt does not need to be short. It can be long and descriptive to give more context to the model.
   - Use delimiters such as *Quotes(" "), Dash (- -), Angular Brackets(<>) & XML Tags (<tag>)* to clearly indicate distinct parts of the input
   - Ask for structured output in *HTML / XML / JSON* formats.
   - Ask the Model whether the Conditions are Satisfied and Check the Assumptions required for the task. This technique is used to handle potential edge cases and works as an instruction for the model on how to revert back the response to avoid unexpected results and errors.
   - Few-Shot Learning: Give some successful examples of the model in the prompt itself to make it understand what is expected of the task.

[20] https://www.andrewng.org/

[21] https://www.deeplearning.ai/short-courses/chatgpt-prompt-engineering-for-developers/

2. **Give the model time to "think"**

    - Specify the steps required to complete a task
    - Instruct the model to work out its solution before rushing to a conclusion

3. **Iteratively Develop the Prompt**

    - Begin with a *Prompt* by adding a basic task description.
    - Test the prompt with a sample input-output.
    - Iteratively keep on developing this prompt with the techniques suggested above.
    - Repeat this process until a final prompt that returns the desired output is achieved.

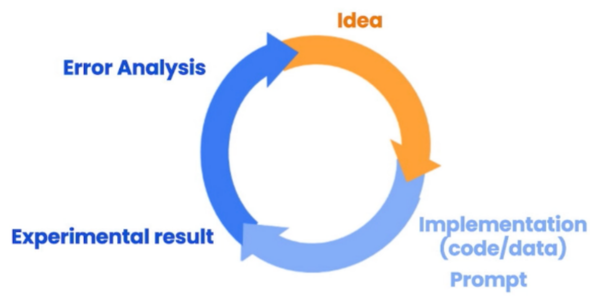


Figure 3.8: Iterative Cycle of Prompt Engineering
source: https://www.deeplearning.ai/

Advantages of LLM Driven NER Over SOTA

- **Contextual Understanding:** GPT's contextual comprehension leads to precise extraction of basic entities like names, dates, locations, and more intricate and hard-to-identify entities.

- **Enhanced Accuracy:** Its broad knowledge and linguistic nuance result in more accurate entity recognition, particularly in diverse or specialized domains.

- **Reduced Data Dependency:** The pre-trained proficiency of GPT reduces the necessity for extensive domain-specific training data, saving time and resources.

- **Adaptability:** GPT's continuous learning and updates enable it to stay current with evolving language trends, maintaining high recognition performance.

- **User-Friendly Integration:** User-friendly APIs facilitate seamless integration into various applications, offering efficient and dependable entity recognition solutions.

- **Superior Performance:** GPT outperforms traditional NLP models with more effective and accurate entity recognition, benefiting tasks such as information retrieval and sentiment analysis.

3.2.4 Final Prompt

You are an AI Assistant that helps users extract information about a person's profile from a given text in a structured JSON format.
Note:

→ The person you have to extract entities about is: *"PERSON NAME"*. Make sure you only extract entities related to the subject and no one else.

→ If you are unable to find any particular information requested by the user, mark it strictly as *'Unknown'*

→ All the values to the keys should not be more than *3-4* words

→ Double-check if all the requirements stated by the user are completely fulfilled. Take your time before returning results to the user.

→ If any instructions are not followed, you will be heavily penalised.

Retrieve the information requested below and return the same in structured format as mentioned:

→ Name: Name of the person in context

→ Age: Age of the person in context

→ Birth Date: In *DD/MM/YYYY* Format as much as possible

→ Contact Information: Keys are *'Email IDs', 'Mobile Numbers', 'Weblinks'*

→ Locations: Exhaustive List of all the locations mentioned in the given text in context to the person

→ Education: List *(keys: 'Institute Name', 'Course Name', 'Branch', 'Grade', 'Duration')*

→ Work Experience: List *(keys: 'Organisation Name', 'Role', 'Job Description', 'Duration')*

→ Projects: List *(keys: 'Project Name', 'Organisation Name', 'Project Description', 'Duration')*

→ Interests: List of interests the person in context keeps

→ Licenses & Certifications: List *(keys: 'Name', 'Issued By', 'Credential ID', 'Issue Date')*

→ Volunteer Experience: List *(keys: 'Organisation Name', 'Duration', 'Description')*

→ Causes: List

→ Test Scores: List *(keys: 'Test Name', 'Score', 'Date')*

→ Miscellaneous: List of all the entities that are relevant to distinguish the profile from any other profile and still not mentioned in any of the above categories. Give keys to entities in the Miscellaneous category by yourself

Perform the entity extraction on the text delimited by triple quotes "'{text}"'

3.2.5 OpenAI API Integration

Having carefully engineered a prompt for our specific use case, the next step involves integrating the same GPT model into our Python script. We utilised the OpenAI API for this step, which is crafted for developers seeking to harness GPT models within programs, web applications, scripts, and beyond. To facilitate this integration, we follow a set of straightforward steps outlined below:

→ **Install the *OpenAI* Package**

Ensure that the OpenAI package is installed and up-to-date within the Python environment. Execute the following command on the terminal/command line:

```
pip install --upgrade openai
```

→ **Setting up an API Key**

Setting up an API Key involves two main steps. First, one must create the API Key[22] itself through the OpenAI platform, which serves as the means of authentication for accessing the API's functionalities. Once the API Key is generated, it must be integrated into the user's system by editing the Bash Profile (for Unix-like systems) or Environment Variables (for Windows). This entails adding the API Key as an environment variable within the system configuration. Once added, the API Key enables secure authentication to access the OpenAI API's capabilities.

→ **Sending the API request**

```python
from openai import OpenAI
client = OpenAI()

completion = client.chat.completions.create(
  model="gpt-3.5-turbo",
  messages=[
    {"role": "system", "content": "You are an excellent linguistic assistant
                                   that helps users extract
                                   information about a person's
                                   profile from a given text in a
                                   structured format."},
    {"role": "user", "content": PROMPT}
  ],
  temperature = 0
)

print(completion.choices[0].message)
```

Python code of OpenAI API Request

[22] https://platform.openai.com/api-keys

Overview of OpenAI API Service

The code snippet showcases the utilization of the OpenAI API for text completion tasks within a conversational framework. It allows customization through keywords such as *"model"*, *"role"* and *"content"*, *"temperature"*. We provide an overview of all the terms here:

Tokens: Requests made to the API incur charges based on the total number of **tokens** involved in both the input provided and the output generated for a given query. In the context of OpenAI GPT models, tokens represent clusters of characters that serve as the foundational units of text. These tokens are produced by a tokenizer algorithm, which partitions the text into smaller segments based on predefined rules, including spaces, punctuation marks, and special characters.

Model: Based on the number of tokens and the task at hand, one can choose from an array of GPT Models[23]. We choose the *"gpt-3.5-turbo-0613 (Snapshot of gpt-3.5-turbo from June 13th 2023.)"* model, which is particularly well-suited for tasks like text generation, such as Named Entity Recognition (NER). The limit for the number of tokens in this model is *4,096 tokens*. In some cases where the number of tokens exceeds this limit, we opt for the *gpt-3.5-turbo-16k-0613* model, which has a limit of *16,385 tokens*.

Role: The API offers us three distinct roles to use: *"system"*, *"user"*, and *"assistant"*, Each role serves a specific purpose within the conversation. The *system role* provides high-level instructions to the model, guiding its behaviour. The *user role* represents input from the user, typically in the form of prompts or queries. Lastly, the *assistant role* encompasses the model's responses to the user's input. In the context of our use case, which involves extracting information about a person's profile from a given text, we primarily leverage the *system* and *user* roles.

Temperature: It is a number between *0* and *2* that defines the degree of randomness of the model's output. By randomness, we mean the disparity between two outputs generated by the model in response to identical prompts. The higher the temperature, the more random the outputs from the GPT model are; the lower the temperature, the more deterministic the outputs are. We set this value to *0* since we want the same answer for a given prompt every single time.

Completion *OR* **Output**: The output from the OpenAI API request provides essential details regarding the completion task in *JSON* format. Figure 3.9 is a sample output depicting the results from the OpenAI API Chat Completion function. It includes information such as the "finish_reason", which indicates the reason for completion termination, where *"stop"* implies completion of the response or that the API returned a complete message. The *"message"* object contains the generated content, with the *"content"*

[23] https://platform.openai.com/docs/models

field presenting the result. The *"role"* field denotes the role of the output, with "*assistant*" signifying that the response is from the GPT model. Additionally, the *"model"* field specifies the model used for the completion task. Details on token usage, including *"prompt tokens"* and *"completion tokens"*, offer insights into the processing involved in generating the response.

```
[1]: import openai
     import os
     import time

     def get_completion(prompt, model="gpt-3.5-turbo"):
         messages = [{"role": "user", "content": prompt}]
         response = openai.ChatCompletion.create(
             model=model,
             messages=messages,
             temperature=0, # this is the degree of randomness of the model's output
         )
         return response

     res = get_completion("Write a short note about OpenAI")
     res
```



```
[1]: <OpenAIObject chat.completion id=chatcmpl-9C1BHKFwYj19F7Yujps2fE2228PsO at 0x113348b90> JSON: {
       "choices": [
         {
           "finish_reason": "stop",
           "index": 0,
           "logprobs": null,
           "message": {
             "content": "OpenAI is an artificial intelligence research lab that aims to ensure that artificial g
     eneral intelligence (AGI) benefits all of humanity. Founded in December 2015, OpenAI conducts research in v
     arious areas of AI, including reinforcement learning, robotics, and natural language processing. The organi
     zation is known for developing cutting-edge AI models, such as GPT-3, which has demonstrated impressive cap
     abilities in generating human-like text. OpenAI also promotes transparency and collaboration in the field o
     f AI research, with a focus on ethical considerations and responsible deployment of AI technologies.",
             "role": "assistant"
           }
         }
       ],
       "created": 1712651503,
       "id": "chatcmpl-9C1BHKFwYj19F7Yujps2fE2228PsO",
       "model": "gpt-3.5-turbo-0125",
       "object": "chat.completion",
       "system_fingerprint": "fp_b28b39ffa8",
       "usage": {
         "completion_tokens": 109,
         "prompt_tokens": 14,
         "total_tokens": 123
       }
     }
```


Figure 3.9: Chat Completion Function using OpenAI API

3.2.6 Example of Extracted Web Profiles

Source: *https://www.linkedin.com/in/sanchetiprateek/*

Start

Extracted Web Profile:


```
{
    "Name": Prateek Sancheti,
    "Age": Unknown,
    "Birthdate": July 25,
    "Contact Information":
        {
            "Email IDs": prateek.sancheti@research.iiit.ac.in,
            "Mobile Numbers": Unknown,
            "Weblinks":[https://psancheti110.github.io/, linkedin.com/in/
               sanchetiprateek]
        },
    "Locations": [Jodhpur, Rajasthan, Hyderabad, Telangana, India],
    "Education":
    [
        {
            "Institute Name": International Institute of Information
               Technology Hyderabad (IIIT - H),
            "Course Name": Bachelor of Technology - B.Tech,
            "Branch": Computer Science,
            "Grade": Unknown,
            "Duration": 2019 - 2024
        }
    ],
    "Work Experience":
    [
        {
            "Organisation Name": IIIT Hyderabad,
            "Role": Undergraduate Researcher @ DSAC,
            "Job Description": Working with Prof. Kamal Karlapalem ... ,
            "Duration": Jan 2023 - Present
        },
        {
            "Organisation Name": IIIT Hyderabad,
            "Role": Undergraduate Researcher @ CogSci (PEG),
```


Continue . . .

```
        "Job Description": Working closely with Dr. Kavita Vemuri ... ,
        "Duration": May 2021 - Present
    },
    {
        "Organisation Name": Product Labs,
        "Role": Software Development Intern,
        "Job Description": ... ,
        "Duration": Nov 2022 - Feb 2023
    },
    {
        "Organisation Name": IIIT Hyderabad,
        "Role": Teaching Assistant (Data Systems),
        "Job Description": Teaching Assistant for the course Data
            Systems offered by Prof. Kamalakar Karlapalem. ... ,
        "Duration": Jul 2022 - Dec 2022
    },
    {
        "Organisation Name": Indian School of Business,
        "Role": Research Assistant,
        "Job Description": ... ,
        "Duration": Oct 2021 - Present
    },
    {
        "Organisation Name": Oracle India,
        "Role": Application Development (SWE) Intern,
        "Job Description": ... ,
        "Duration": May 2023 - Jul 2023
    },
    {
        "Organisation Name": Qualcomm,
        "Role": Interim Engineering Intern,
        "Job Description": ... ,
        "Duration": Feb 2023 - May 2023
    },
    {
        "Organisation Name": Qualcomm,
        "Role": Interim Engineering Intern,
        "Job Description": ... ,
        "Duration": May 2022 - Jul 2022
    },
```

Continue . . .

```
        {
            "Organisation Name": Chaincode Labs (Summer of Bitcoin),
            "Role": Bitcoin Core Developer,
            "Job Description": Open Source work in Bitcoin Core: ...,
            "Duration": Jul 2021 - Sep 2021
        }
    ],
    "Projects": [],
    "Interests": [],
    "Licenses & Certifications":
    [
        {
            "Name": Investment Foundations Program,
            "Issued By": CFA Institute,
            "Credential ID": 520***,
            "Issue Date": Feb 2022
        },
        {
            "Name": 30 Days of Google Cloud,
            "Issued By": Google for Developers,
            "Credential ID": Unknown,
            "Issue Date": Dec 2020
        }
    ],
    "Volunteer Experience": [],
    "Causes": [],
    "Test Scores": [],
    "Miscellaneous": [ Prompt Engineering, Named Entity Recognition (NER),
       Data Analysis, Python, Financial Analysis, Quantitative Research,
       React.js, Machine Learning, Computer Vision, Databases, Selenium,
       Google Cloud Platform (GCP), Blockchain, Qt Framework, HTML5, Java,
        C, JavaScript, CSS, MySQL, Teamwork, Public Speaking, Problem
       Solving, Leadership, Algorithms, Data Structures]
}
```

. . . End

3.3 DeDuplication – Consolidation & Disambiguation of Web Profiles

After extracting web profiles associated with a given name from search results across diverse search engines and processing them through GPT using our meticulously crafted NER prompt, we've obtained a collection of individual web profiles stored in structured JSON format. The next phase in our pipeline involves transitioning to the crucial task of ***DeDuplication***, or we can say ***Profile Consolidation***. This pivotal step involves identifying and merging duplicate or overlapping data instances. Our objective here is to scrutinize each pair of profiles within the extracted dataset systematically, evaluate the information contained in each profile (based on extracted entities), and group together those that pertain to the same real-world individual. This section comprehensively examines the sequence of steps undertaken to consolidate these profiles. It encompasses *Profile Consolidation Based on Contact Information (3.3.1), Standardization of Extracted Entities (3.3.2), the implementation of a Prefix Tree inspired Data Structure for Profile Consolidation (3.3.3)*, and finally, *Profile Consolidation with Prompting (3.3.4)*.

3.3.1 Profile Consolidation Based on Contact Information

Contact Information, such as *Email IDs or Mobile Numbers or Weblinks*, serves as a unique identifier for each individual. When two profiles share contact information. It indicates that they represent the same person in the real world, so the two profiles can be merged. By merging these profiles across all attributes, we create an enhanced profile carrying data of both profiles, thus eliminating redundant data *(Figure 3.1 (5))*. Figure 3.10 depicts an example of two profiles sharing contact information. *Profile 1* and *Profile 2* share the same Mobile Number *(01234-56789)*. We merge the two profiles upon finding such an instance across any pair of the extracted web profiles. To identify all groups of profiles that share contact information, we utilize the idea of *Disjoint Set Union-Find Algorithm*[24].

	Profile 1	**Profile 2**	**Merged Profile**
Email IDs	*"Unknown"*	*["abc@xyz.com", "def@uvw.com"]*	*["abc@xyz.com", "def@uvw.com"]*
Mobile Number	*["(+91)0123456789", "1234512345"]*	*["01234-56789", "0987654321"]*	*["01234-56789", "0987654321", "1234512345"]*
Web Links	*[Link 1, Link 2]*	*[Link 3, Link 4]*	*[Link 1, Link 2, Link 3, Link 4]*

Figure 3.10: Example of Profiles Sharing Contact Information

After merging profiles that share contact information, we move on to finding similarities among profiles based on factual information in the profiles. Not all the entities extracted in these web profiles are facts. A *fact* is information that can be verified through evidence, e.g. *Education Details, Date of Birth,*

[24]https://en.wikipedia.org/wiki/Disjoint-set_data_structure

Work Experience. In contrast, a *non-facts* are subjective, ambiguous, or unverifiable statements. Determining Profile Match is checking whether two profiles correspond to a single individual or two distinct individuals in the real world. We need to compare all the extracted factual information across each attribute value to decide on the profiles' similarity and hence *Fact Focused Web Profile Consolidation*

3.3.2 Standardization of Extracted Entities

The process of standardization of extracted facts (entities) addresses the need to homogenize diverse representations of entities within the extracted profiles. Individuals may mention their details in multiple formats, influenced by the intended audience — *abbreviations, full names, common names (known to those familiar with the entity), and more*. The challenge arises when comparing these textual variations. To overcome this challenge, we *Standardize these Facts using Wikipedia (Figure 3.1 (6))*.

Wikipedia is a vast repository of structured information. *Wiki Standardization of Facts* refers to the practice of replacing textual information *(extracted facts)* within a document or dataset with its corresponding *Wikipedia URL*. The process enhances data clarity by creating a single reference point and replacing diverse textual representations with a URL. Once standardized, we can directly compare the URL corresponding to the entities. The process of standardizing extracted entities in these web profiles using Wikipedia involves sending different search queries to search engines for each factual entity or attribute present in the given profile.

→ **Preparing Search Query**: For a given entity E with entity category E_c, the search query is:

$$\texttt{E}_{\texttt{c}}\ \texttt{E site:wikipedia.org}$$

For example, if *Institute Name* in a given case is *MIT*, the search query will be

```
Institute Name MIT site:wikipedia.org
```

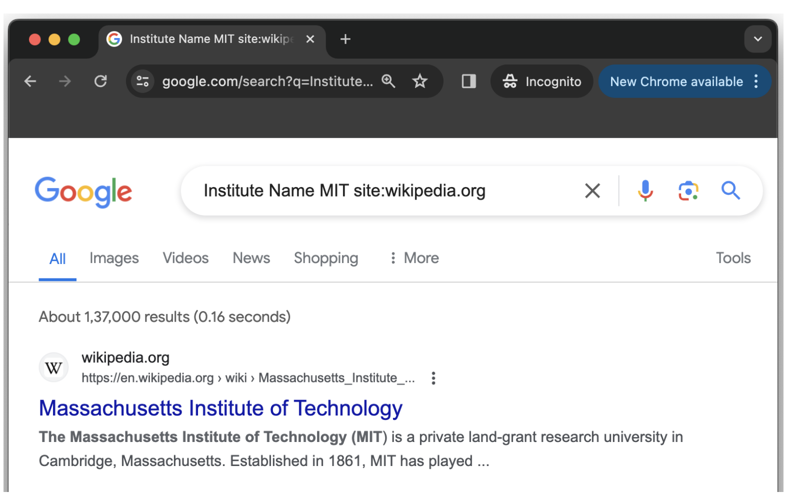


Figure 3.11: Example of Search Query for Wiki Standardisation of Extracted Entities

This search query pattern ensures the correlation between E and E_c is kept in check and also that the results are only from *Wikipedia*.

→ **Query Using Selenium & Scrape the First Search Result**: We use our search result extraction pipeline (See 3.1.1) to extract the *first* search result for a given search query.

→ **Excess Search Queries – reCAPTCHA Issue**: The standardization process may involve over *100* entities within a single web profile, necessitating a substantial number of search queries for their standardization on Google through Selenium automation. All search engines employ set systems to detect robotic or automated actions, primarily to safeguard against malicious activities like *web crawlers* or *DDoS* attacks. When such an activity is detected, it requires human verification through reCAPTCHA[25]. To overcome this challenge, we employ *Search Engine Result Pages (SERP) Scrappers* that have designated technologies to overcome this challenge with a simple API integration. We send our query to these scrappers, which then searches the search engine and returns the desired results without any delay or lag. A few popular SERP scrappers are *OxyLabs*[26], *SerpApi*[27] & *ScraperAPI*[28].

Example of a Wiki - Standardised Web Profile:

```
{
    "Education"
    {
        "Institute Name" : MIT,
        "Course Name" : B.Eng,
        "Branch" : CSE,
        "Duration" : Unknown,
        "Grade" : Unknown,
        "Institute_Name_Wikipedia_Link" : https://en.wikipedia.org/wiki/
            Massachusetts_Institute_of_Technology,
        "Course_Name_Wikipedia_Link" : https://en.wikipedia.org/wiki/
            Bachelor_of_Engineering,
        "Branch_Wikipedia_Link" : https://en.wikipedia.org/wiki/
            Computer_science_and_engineering
    }
}
```

To determine whether two profiles correspond to the same individual, one needs to compare each attribute category of one profile with the other profile. We use a *tree-based* approach.

[25]https://www.google.com/recaptcha/about/
[26]https://oxylabs.io/
[27]https://serpapi.com/
[28]https://www.scraperapi.com/blog/top-google-serp-api-search-engine-proxies-and-scraping-tools/

3.3.3 Prefix Tree | *Trie*

A *Prefix Tree*, also known as a *Trie*, is a tree-like data structure that efficiently stores data. We use the fundamental property of a trie of grouping together elements sharing similar attributes to construct a data structure for storing the extracted profiles.

3.3.3.1 Trie Construction

The profiles are added to the *Trie* like data structure one by one *(Figure 3.1 (7))*, with each *Attribute Category* at a designated *depth* from the root node. The order in which the attributes of a given profile are inserted into the Trie is determined by the frequency of data occurrence in each attribute category among all the profiles, starting with the most frequently occurring attribute to the least frequent one. This order reduces the number of leaf nodes, hence reducing the number of distinct profiles. The factual information of each profile is stored in *Nodes*, which are connected internally using *Edges*. Before inserting information from an extracted profile to the *Trie*, we examine whether there is an existing sibling node with identical values. If such a node exists, we append the current profile to that node. Otherwise, if no matching Node is found, we create a new node at the same depth to accommodate this profile. This process repeats until all the profiles are inserted into the trie.

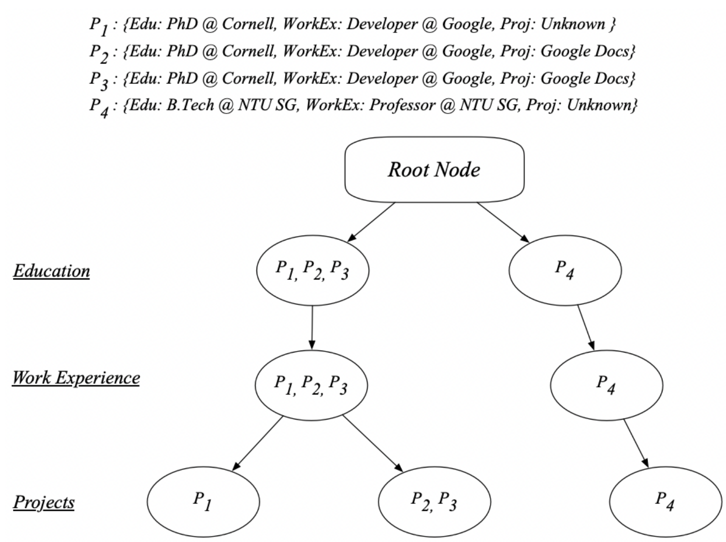


Figure 3.12: Depiction of Profile Consolidation with a Prefix Tree *(Trie)* Inspired Data Structure

Consider four Profiles P_1, P_2, P_3 and P_4. Figure 3.12 is a sample depiction of the *Profile Consolidation* of these profiles, having only *3* entity categories – *Education, Work Experience & Projects*, using the *Prefix Tree* inspired Data Structure. Since P_1, P_2 and P_3 share a common education *(PhD at Cornell University)*, they were put in the same node where P_4 was placed separately in another node at the Education Level *(Depth 1)*. As we proceed further, P_1, P_2 and P_3 are yet again put in the same node at Work Experience Level *(Depth 2)* since a common link was found *(Google)* while P_4 proceeds in its own separate branch. Lastly, at the Project Level *(Depth 3)*, Profile P_1 was bifurcated from P_2 and P_3

due to lack of project information in P_1. As the depth increases, our ability to determine the similarity between two given profiles also improves. One can with certainty say that P_2 and P_3 correspond to the same individual in the real world since they belong to the same leaf node *(Depth 3)*. Similarly, we can say P_4 corresponds to a distinct individual in the real world than the other three Profiles.

3.3.3.2 Varied Level of Profile Consolidation

To cater for a diverse set of user preferences, our solution offers *three* levels of precision while checking similarity with other profiles – *Strict, Relaxed, and Loose*.

- ***Strict Profile Match:*** Consider a Profile P_1 consisting of *five* entries in the *Education* category, each having all the predefined sub-attributes. For a Profile to be called *strictly matching* with P_1, either P_1 must contain *all* the entries present in the other profile *(all sub-attributes in each element)* or vice-versa.

- ***Relaxed Profile Match:*** A match of *at least one* entire entry *(all sub-attributes in that element)* for that Attribute Category. Say Profile P_1 has *five* elements in the *Education* category. For a Profile to be called *relaxedly matching* with P_1, it must contain at least one element that entirely matches *(all the sub-attributes)* with an element present in P_1.

- ***Loose Profile Match:*** A match of *at least one* sub-attribute from any of the elements in that Attribute Category. For e.g., Profile P_1 has *five* elements in the *Education* category. For a Profile to be called *loosely matching* with P_1, it must contain at least one sub-attribute, either Institute Name or Course Name or Branch Name, that matches with the corresponding sub-attribute from an element in P_1.

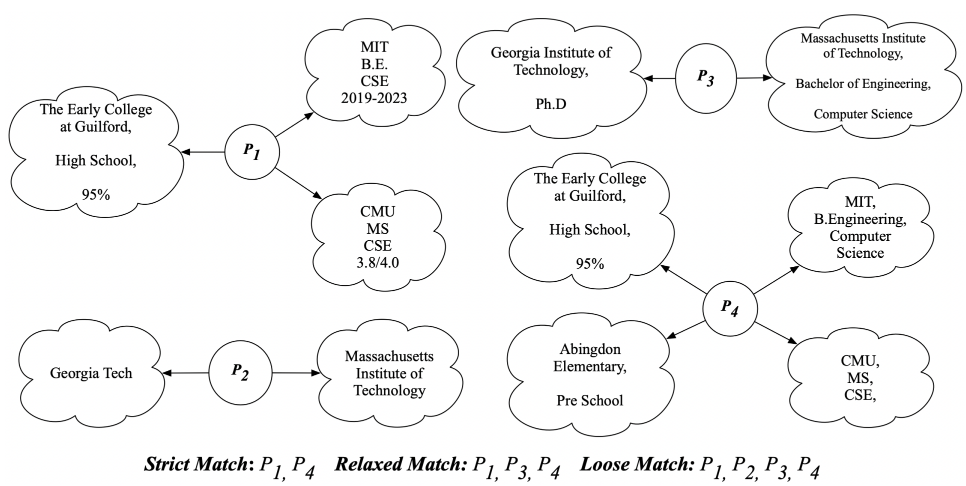


Figure 3.13: Profiles Matching Scenarios *(Strict – Relaxed – Loose)*

Figure 3.13 depicts four profiles with different educational descriptions. These profiles show varying levels of similarity, each with a distinct precision. For instance, P_1 and P_4 show a *Strict* match since all three education entries in P_1 are fully present in P_4, including all subcategories (Institute, Course, Branch). P_3 shows a *Relaxed* match with both P_1 and P_4 because only one complete education entry *(MIT-B.Eng-CSE)* matches in all subcategories. Lastly, P_2 matches *Loosely* with all other profiles since it only matches the other profiles at a sub-category level *(Institute Name: MIT)*.

3.3.3.3 Example Discussing Profile Consolidation with Prefix Tree

Let us consider *5* distinct web profiles for the name *Prateek Sancheti*. We aim to identify if any of the web profiles correspond to the same real-world individual or not. To get a clearer understanding of the profile consolidation process, we only focus on *3* entity categories: *Location, Education, Work Experience*. Attached below is the reduced version of the extracted web profiles.

Start

Profile 1: www.linkedin.com/in/sanchetiprateek/
Extracted Profile 1: See subsection *3.2.6*

Profile 2: www.linkedin.com/in/prateek-sancheti-7302192b/
Extracted Profile 2:


```
{
    "Name": Prateek Sancheti,
    "Contact Information": {
        "Weblinks": [www.linkedin.com/in/prateek-sancheti-7302192b,
           instaprintz.in]},
    "Locations": [Indore, Madhya Pradesh, India, Hyderabad, Bhopal],
    "Education":
    [
        {
            "Institute Name": Indian Institute of Management, Indore,
            "Course Name": PGCPM,
            "Branch": Business Administration,
            "Grade": Unknown,
            "Duration": Jan 2023 - Jan 2024
        },
        {
            "Institute Name": Samrat Ashok Technological Institute,
            "Course Name": B.E,
            "Branch": Computer Engineering,
            "Grade": 8.3,
            "Duration": 2010 - 2014
        },
```


Continue . . .

```
        {
            "Institute Name": Joy Senior Secondary School,
            "Course Name": HSC,
            "Branch": Maths-Science,
            "Grade": 81.6,
            "Duration": 2006 - 2010
        }
    ],
    "Work Experience": [
        {
            "Organisation Name": Department Of Sports & Youth Welfare,
            "Role": Member State Youth Advisory Council,
            "Job Description": Unknown,
            "Duration": May 2023 - Present
        },
        {
            "Organisation Name": Atal Innovation Mission,
            "Role": Mentor of Change,
            "Job Description": Unknown,
            "Duration": Mar 2020 - Present
        },
        {
            "Organisation Name": Insta Printz,
            "Role": Chief Executive Officer (CEO),
            "Job Description": Started rapid prototyping and customization
               for clients from various sectors.
            "Duration": May 2015 - Present
        },
        {
            "Organisation Name": Mangosteen Eco-Life Style,
            "Role": Owner/Partner,
            "Job Description": Rooftop cafe serving fresh food and fitness
               programs.
            "Duration": Sep 2015 - Sep 2016
        },
        {
            "Organisation Name": Cognizant Technology Solutions,
            "Role": Software Engineer,
            "Job Description": Worked as a software engineer.
            "Duration": Aug 2014 - May 2015
        },
```


Continue . . .

```
        {
            "Organisation Name": Netlink Software,
            "Role": Trainee,
            "Job Description": I worked as a trainee in JavaSE, HTML, and
               Oracle.
            "Duration": Sep 2011 - Jul 2012
        }
    ]
}
```

Profile 3: www.linkedin.com/in/prateeksancheti/

Extracted Profile 3:

```
{
    "Name": Prateek Sancheti,
    "Contact Information": "Weblinks": www.linkedin.com/in/prateeksancheti,
    "Locations": [Pune, Maharashtra, India ],
    "Education":
    [
        {
            "Institute Name": Pune University,
            "Course Name": PGDFT, International Economics,
            "Branch": Unknown,
            "Grade": First class with distinction,
            "Duration": 2017 - 2018
        },
        {
            "Institute Name": Pune University,
            "Course Name": Bachelor's degree, Business/Commerce, General,
            "Branch": Unknown,
            "Grade": First class with distinction,
            "Duration": 2013 - 2017
        },
        {
            "Institute Name": The Institute of Chartered Accountants of
               India,
            "Course Name": Charted Accountant (CA),
            "Branch": Unknown,
            "Grade": Unknown,
            "Duration": Unknown
        }
    ],
```

Continue . . .

```
    "Work Experience":
    [
        {
            "Organisation Name": Anand Rathi & Associates,
            "Role": Sr. Wealth Manager - Preferred Wealth Management,
            "Job Description": Managing a complete Investment balance sheet
                 for an individual. Focusing on building a strategy with
                detailed asset allocation.
            "Duration": Sep 2022 - Present
        },
        {
            "Organisation Name": Purnartha SEBI Registered Equity
                Investment Advisor,
            "Role": Sr. Wealth Manager (Equity Asset Class),
            "Job Description": At Purnatha, I have learned real equities,
                like how a company's fundamentals create wealth for
                investors.
            "Duration": Feb 2019 - Sep 2022
        },
        {
            "Organisation Name": Motilal Oswal Financial Services Ltd,
            "Role": Investment Advisor - Portfolio Management Services,
            "Job Description": At MO, I learned the basics of wealth
                products and how they are helpful for different sets of
                clients like PMS, AIF, Structured Products, etc.
            "Duration": Aug 2018 - Jan 2019
        },
        {
            "Organisation Name": MIDAS WEALTH ADVISORY PRIVATE LIMITED,
            "Role": Portfolio Analyst,
            "Job Description": Unknown,
            "Duration": Jun 2018 - Aug 2018
        },
        {
            "Organisation Name": MylesCars,
            "Role": Digital Marketing Executive,
            "Job Description": Unknown,
            "Duration": Sep 2017 - Nov 2017
        }
    ],
}
```

Continue . . .

Profile 4: www.f6s.com/member/prateeksancheti

Extracted Profile 4:


```
{
    "Name": Prateek Sancheti,
    "Contact Information": {
        "Weblinks": [www.f6s.com/member/prateeksancheti, instaprintz.in]},
    "Locations": [Indore, India],
    "Education":
    [
        {
            "Institute Name": Samrat Ashok Technological Institute (SATI),
            "Course Name": B.Tech,
            "Branch": Unknown,
            "Grade": Unknown,
            "Duration": Aug 2010 - May 2014
        },
        {
            "Institute Name": Joy Senior Secondary School,
            "Course Name": HSC,
            "Branch": Unknown,
            "Grade": Unknown,
            "Duration": April 2005 - March 2010
        }
    ],
    "Work Experience": [
        {
            "Organisation Name": Insta Printz,
            "Role": CEO,
            "Job Description": Unknown,
            "Duration": May 2015 - Present
        },
        {
            "Organisation Name": Cognizant Technology Solutions,
            "Role": Engineer,
            "Job Description": Unknown,
            "Duration": Aug 2014 - May 2015
        },
        {
            "Organisation Name": Trivima,
            "Role": CTO,
```


Continue . . .

```
            "Job Description": Unknown,
            "Duration": Unknown
        }
    ]
}
```

Profile 5: https://www.urbanpro.com/jodhpur/prateek-sancheti/24456287
Extracted Profile 5:

```
{
    "Name": Prateek Sancheti,
    "Contact Information": {
        "Weblinks": www.urbanpro.com/jodhpur/prateek-sancheti/24456287 },
    "Locations": [Jodhpur, India],
    "Education":
    {
        "Institute Name": IIIT Hyderabad,
        "Course Name": B. Tech,
        "Branch": Computer Science Engineering,
        "Grade": Unknown,
        "Duration": Unknown
    },
    "Work Experience":
    {
        "Organisation Name": Unknown,
        "Role": Online Tutor,
        "Job Description": Teaching students pursuing competitive exams
           like JEE, BITSAT, NTSE, KVPY,
        "Duration": 2 years
    }
}
```

. . . End

Figures 3.14, 3.15 and 3.16 are depiction of *Strict, Relaxed* and *Loose* Prefix Tree profile consolidation for the *five* profile P_1, P_2, P_3, P_4, P_5 extracted above.

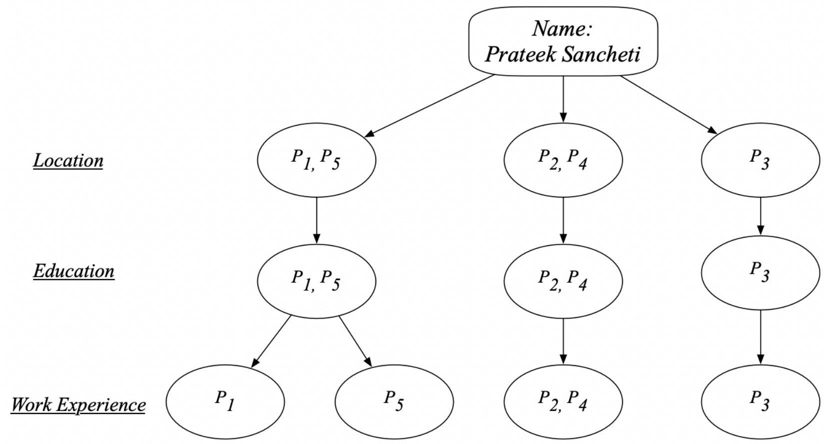


Figure 3.14: Example Profile Consolidation *(Strict)*

In *Strict Profile Consolidation* (Figure 3.14), Profiles P_1 and P_5 share the same location *(Jodhpur, India)* as well as same education *(B.Tech @ IIIT Hyderabad)* but due to lack of work experience information in P_5 they are separated at the last level. P_2 and P_4 share the same Location *(Indore, India)*, Education *(Engineering @ SATI, Highschool @ Joy Senior Secondary School)* and Work Experience *(CEO @ Insta Printz, Engineer @ Cognizant)*. P_3 doesn't have any entity strictly matching with any of the other *four* profiles.

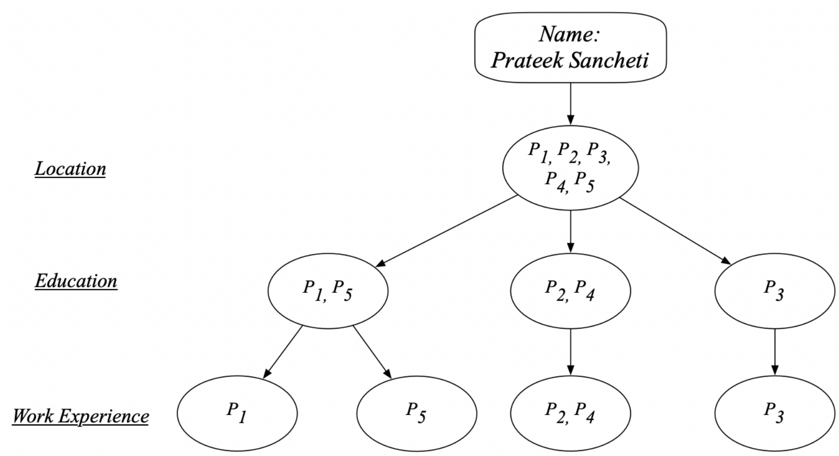


Figure 3.15: Example Profile Consolidation *(Relaxed)*

In *Relaxed Profile Consolidation* (Figure 3.15), All the profiles P_1, P_2, P_3, P_4 and P_5 share the same location *(India)*. Profiles P_1 and P_5 share same education *(B.Tech @ IIIT Hyderabad)* but due to lack of work experience information in P_5 they are separated at the last level. P_2 and P_4 share the Education *(Engineering @ SATI, Highschool @ Joy Senior Secondary School)* and Work Experience *(CEO @ Insta Printz, Engineer @ Cognizant)*. Profile P_3 doesn't have any Education or Work Experience entity relaxedly matching with any of the other *four* profiles; hence, it stands separate in the prefix tree.

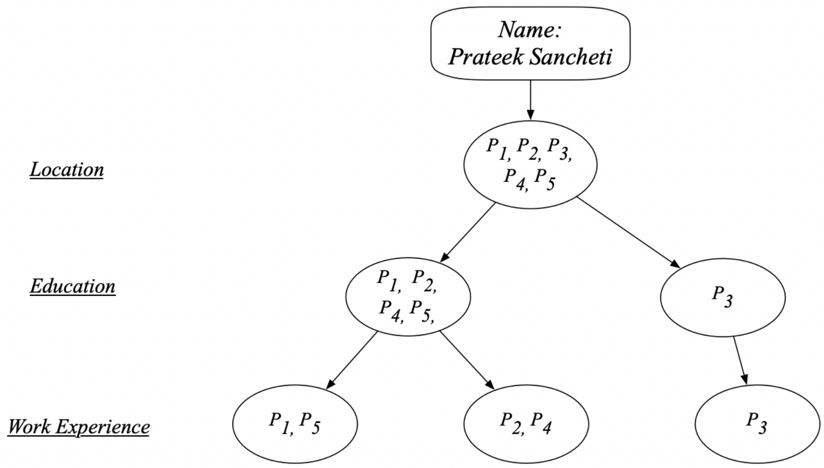


Figure 3.16: Example Profile Consolidation *(Loose)*

In *Loose Profile Consolidation* (Figure 3.16), All the profiles P_1, P_2, P_3, P_4 and P_5 share the same location *(India)*. Profiles P_1, P_2, P_4 and P_5, all share one entity in education *(Engineering)* hence the fall in the same node. Profiles P_2 and P_4 share the same work profile *(CEO, InstaPrintz etc.)*, whereas Profile P_3 doesn't share any entity in either education or work experience hence it is in a separate.

By analysing several results, we find *Relaxed Matching* scenario to give the most realistic matching among Strict, Relaxed and Loose.

3.3.4 Profile Consolidation with Prompting

While Wikipedia provides an extensive collection of structured information that is used to standardize most profiles, there still remains some factual information related to a profile that cannot be standardized using Wikipedia. This includes information that requires contextual understanding, such as job descriptions, career paths, and background. Because of this issue, we are left with a small percentage of profiles that do not have enough standardized data to be inserted in the Trie. GPT, with its ability to comprehend context, serves as a solution to consolidate such profiles with the previously consolidated profiles (Baktash and Dawodi, 2023).

We prepare the prompt using the prompt construction techniques discussed earlier *(3.2.3)* to determine if two profiles belong to the same real-world individual. We designed the prompt to be cautious to reduce the possibility of incorrect matches. *(Figure 3.1 (8))*

→ *You will be presented with a set of profiles for individuals who share a common name.*

→ *Each profile has multiple attributes about the person.*

→ *Your objective is to thoroughly analyze every factual information in these profiles to determine if any profiles represent the same real-world individual.*

→ *To consider two profiles belonging to the same individual, they must share entities in at least two entity categories.*

→ *Return the answer in the following format:*

  - *Do the profiles correspond to the same individual? "Yes/No".*
  - *Score the profile similarity on a scale of 1 to 10*
  - *List of reasons behind the above answer, comparing each entity category*

→ *Verify your answer before returning it to the user. If any false profile match is provided, you will be penalized.*

GPT Prompt to Determine Profile Similarity

These instructions prompt the GPT to assess the similarity between input profiles while considering the context in which individual attributes appear.

3.3.4.1 Example Discussing Profile Consolidation with GPT Prompting

Consider an example where *three* individuals share *"Huang Wei"* as their full name

Profile 1: Huang Wei is a dynamic software developer at QC San Diego (CA), one of the leading tech firms in the world. Hailing from the vibrant city of HK (China), he earned his Bachelor's degree in Beijing (Tsing Hua Imperial College). Huang's professional journey has led him to specialize in the world of DSP simulators along with telecommunications and wireless technology. His commitment to excellence and knack for solving complex technical challenges make him an invaluable asset to his team and the company at large.

Extracted Profile 1:

```
{
    "Name" : Huang Wei,
    "Locations" : [San Diego (CA), HK, Beijing, China],
    "Education" :
    {
        "Institute Name" : Tsing Hua Imperial College,
        "Course Name" : Bachelor's Degree,
        "Branch Name" : Unknown,
        "Duration" : Unknown
    },
    "Work Experience" :
    {
        "Organization Name" : QC San Diego,
        "Role" : Software Developer,
        "Job Description" : [DSP Simulators, Telecommunication and
                             Wireless Technology]
    }
}
```

Profile 2: In the bustling world of technology, Huang Wei found his calling as an SW Engineer at Qualcomm. Originally from Hong Kong, Huang embarked on his educational journey at Beijing's Tsinghua University (B. Tech CSE), where he nurtured his innate technical aptitude. Wei is entrusted with a pivotal role that revolves around developing Digital Signal Processing simulators, making him an invaluable asset to his team and a driving force behind Qualcomm's technological advancements.

Extracted Profile 2:


```
{
    "Name" : Huang Wei,
    "Locations" : [Hong Kong, Beijing],
    "Education" :
    {
        "Institute Name" : Tsinghua University,
        "Course Name" :  B. Tech,
        "Branch Name" : CSE,
        "Duration" : Unknown
    },
    "Work Experience" :
    {
        "Organization Name" : Qualcomm,
        "Role" : SW Engineer,
        "Job Description" : [Digital Signal Processing Simulators]
    }
}
```


Profile 3: Huang Wei is an experienced journalist known for his expertise in crime reporting. Huang holds the position of Senior Journalist at Shanghai Daily's main branch, where he delves deep into compelling narratives. He earned his Bachelor's degree in Journalism from Tsinghua School of Journalism and Communication (China). Huang's dedication to reporting and investigative abilities have solidified his reputation as a prominent figure in journalism.

Extracted Profile 3:


```
{
    "Name" : Huang Wei,
    "Locations" : [Shanghai, China],
    "Education" :
    {
        "Institute Name" : Tsinghua School of Journalism and Communication,
        "Course Name" :  Bachelor's degree,
        "Branch Name" : Journalism,
        "Duration" : Unknown
    },
    "Work Experience":
    {
        "Organization Name" : Shanghai Daily,
        "Role" : Senior Journalist,
        "Job Description" : [Crime Reporting]
    }
}
```

We provide each pair of the extracted profiles mentioned above along with GPT Prompt *(3.3.4)* to determine if they correspond to the same real-world individual. The output from GPT was as follows:

→ **Results from Profile Consolidation using GPT for Profile 1 & Profile 2**

- ***Do the profiles correspond to the same individual?*** *Yes.*
- ***Profile Similarity Score:*** *8 / 10*
- ***Reasons:***
  1. Locations partially match with common locations (Hong Kong ∼ HK, Beijing)
  2. Education (Bachelor Degree at Tsing Hua Imperial College ∼ B.Tech at Tsinghua University)
  3. Work Experience (QC San Diego ∼ Qualcomm, DSP ∼ Digital Signal Processing)

→ **Results from Profile Consolidation using GPT for Profile 1 & Profile 3**

- ***Do the profiles correspond to the same individual?*** *No.*
- ***Profile Similarity Score:*** *2 / 10*
- ***Reasons:***
  1. Location partial match (China)
  2. Education & Work Experience (Although there is a partial match in University name & Course name, but completely distinct work profiles)

→ **Results from Profile Consolidation using GPT for Profile 2 & Profile 3**

- ***Do the profiles correspond to the same individual?*** *No.*
- ***Profile Similarity Score:*** *1 / 10*
- ***Reasons:*** Although partial match in Institute name, the context describes distinct work profiles

Notice in the above example how GPT, with its contextual understanding, can precisely detect the false positives and true negatives for identifying profile matches.

We use both the above-mentioned techniques *(Trie Based & GPT Prompting Based)* to consolidate profiles of individuals sharing the same name conclusively. We employ the technique of Wiki Standardization followed by Trie Construction to manage profiles with sufficient content that can be standardized. In contrast, we use the GPT Prompting technique to identify similarities among the remaining profiles (that require contextual understanding) and profiles that have already been consolidated with Trie.

GPT, with its NLP skills, makes it possible to perform tasks like disambiguating the extracted profiles. However, its effectiveness can diminish as the size of the input prompt increases. As a result, the probability of GPT hallucinating and returning inaccurate increases. Our approach is strategic and judiciously uses GPT only for the profiles lacking substantial standardizable content, which represents a smaller portion of the entire dataset. This selective application helps us mitigate the issues posed by large prompts.

3.4 Testing & Results

To ensure the robustness of our system, we conducted comprehensive testing. This testing encompassed a wide range of scenarios:

3.4.1 Common Names

We subjected the system to the challenge of handling over *500 different names (common & uncommon)*. The list of names also included the most common names in the US, as determined by *fivethirtyeight*[29] – *"James Smith"* being the most common. Table *3.4* summarises the results of this testing. As one would expect, we observe more distinct profiles in the case of common names than in the case of less or non-common names.

Table 3.4: Result Summary : Testing done on a total of *500* distinct *Full Names*. A summary of *average* number of distinct profiles observed for common/non-common names for different matching precision. f: female, m: male, common m/f names used 150, non-common m/f names used : 100

	Common		**Non-Common**	
	f	*m*	*f*	*m*
Total Search Results	*300*	*300*	*300*	*300*
Avg. Unique Search Results	*175*	*180*	*145*	*165*
Avg. Strict Matching	*135*	*150*	*110*	*125*
Avg. Relaxed Matching	*110*	*120*	*85*	*100*
Avg. Loose Matching	*65*	*75*	*45*	*60*

Observation: As the commonness of a name increases, more distinct profiles are observed compared to less common names when scraping the same number of search results.

E.g., while testing for the name *"James Smith" (the most common name in the US)*, we extracted a total of 300 results from *Google, Yahoo and Bing*. This resulted in around *160, 135 and 105* unique profiles in a strict, relaxed and loose match, respectively. These unique profiles included a boxer, an actor, a fitness influencer, an award-winning author, a software developer, a businessman, and many others, all united by their shared name. Whereas while testing for the name *"Jian Nie"*, a relatively less common name than *"James Smith"*, the search resulted in around *110, 90 and 50* unique profiles in a strict, relaxed and loose match respectively.

3.4.2 Celebrity Name Conflicts

Following the successful extraction of textual data from search results based on the individual's full name, we address the issue of celebrity name conflicts. Employing a hash table approach for words present in the extracted textual data, we identify recurring common words across all web pages. This

[29]https://fivethirtyeight.com/features/whats-the-most-common-name-in-america/

process yields keywords that potentially represent a single individual, in most cases, celebrities. Subsequently, if such words are prevalent across a majority of links, we refine the search query. This optimization involves appending a ' - ' followed by the keyword to the search query. It ensures that web results containing that specific keyword - and consequently, the individual associated with it - are excluded from the search results.

This refinement of the search query allows our system to retrieve information about non-celebrity individuals who share the same names. Consider input name as *"Michael Jordan"*, a famous basketball player. A web search for this name will return results, with almost all referring to the renowned sportsman. To search for individuals who share this name, the system refines the search queries:

- *Query 1:* "Michael Jordan"
- *Query 2:* "Michael Jordan" -basketball
- *Query 3:* "Michael Jordan" -basketball -chicago
- *Query 4:* "Michael Jordan" -basketball -chicago -nike

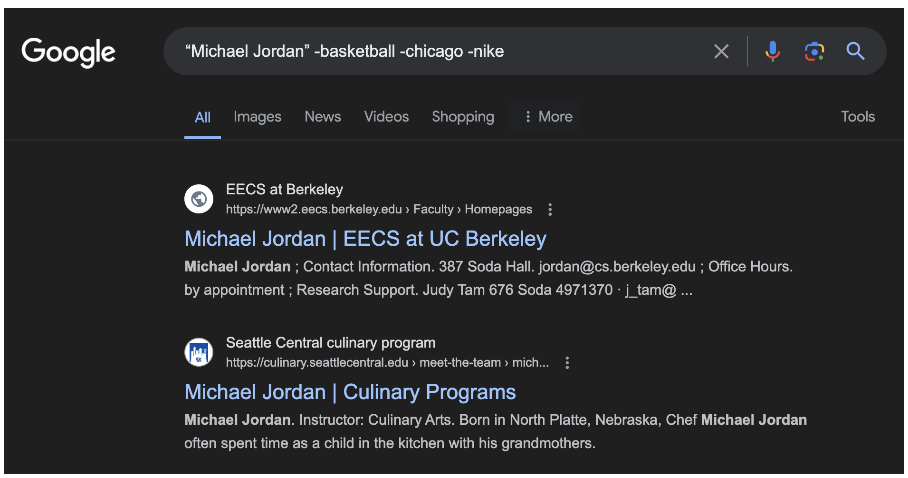


Figure 3.17: Refined Search Query for Celebrity Name Conflict

The refined search query now returns results for individuals who only share the name with the basketball player. Our system returned various profiles for *Michael Jordan*, a few of them are listed below:

Michael Jordan:

1. Chef at a Culinary Institute
2. An Engineer in New York
3. Attending Physician at Tufts
4. Pehong Chen Distinguished Professor
5. Book Writer from Michigan
6. Law Associate in New Jersey
7. Managing Director (Tax Partner)
8. Investment Banker for Private Equity
9. American Businessman
10. VP (Customer Success) in Utah

3.4.3 Other Practical Use Cases

Online Presence Management for individual users Our system carries practical utility for individuals seeking to manage their online presence and verify their online identity. Users can utilize the system to compile a *list of links* where their profile information is being used and review the websites and sources where their profile information is being used. Consider the case of *Michael Jordan – The Professor (iv)* as discovered above in *3.4.2*. Using our system, we successfully obtained a list of around *80+* links associated with this specific individual. A few links from the list are mentioned below:

- **Academic Links**
  - *EECS Berkeley* https://www2.eecs.berkeley.edu/Faculty/Homepages/jordan.html
  - *MIT IDSS:* https://idss.mit.edu/staff/michael-i-jordan/
  - *Yale University*: https://yale2020.yale.edu/honorary-degrees/michael-i-jordan
- **Scholarly Links**
  - *IEEE Xplore:* https://ieeexplore.ieee.org/author/37269675800
  - *WLA Prize:* https://www.thewlaprize.org/Laureates/2022/Michael_I._Jordan/
  - *Google Scholar:* https://scholar.google.com/citations?user=yxUduqMAAAAJ&hl=en
- **Non-Academic Links**
  - *Wikipedia:* https://en.wikipedia.org/wiki/Michael_I._Jordan
  - *LinkedIn:* https://www.linkedin.com/in/michael-jordan-767032125/
  - *Crunchbase:* https://www.crunchbase.com/person/michael-i-jordan

By regularly monitoring the usage of their online profile, users can quickly identify any instances of *unauthorized representation* or *impersonation*. This use case empowers individuals to ensure the legitimacy of their online presence and protect themselves against potential impersonation or misuse of their identity.

Improve Web People Search Algorithms Our pipeline can help enhance web searches for person names. With our pipeline as a layer on top of the Search Engine, users can quickly and effectively identify the individual they are searching for while gaining access to more factual information about that person.

3.4.4 User Study

To evaluate the practical usability of our system, we conducted a user study involving *fifty* participants *(college students & working professionals)*. Each participant was tasked with searching for individuals they knew who shared names with others to assess whether our system could accurately curate information about these individuals. The outcomes of the user evaluation showed the pipeline successfully extracted various web profiles of individuals carrying the input name. Participants not only located the profiles of their intended individuals but also observed that these profiles comprehensively aggregated information from all relevant links. In multiple cases where people looked up their own names, they were surprised to discover several other individuals who shared the exact same name as them. Our method displayed various individuals who possibly shared entities such as location, education, or work experience with the user. For example, when an individual searched their name, they found another individual that happens to be from the same home town and also works at the same organisation.

3.4.5 User Supervision

Human supervision cannot be avoided while differentiating and consolidating profiles.

"Venkatesh Reddy" is a fairly common name in *India*. We found certain individuals shared the same educational background *(B.Tech / B.E. @ IIT)* as well as work background *(Software Engineer)*. We found a total of *six* such individuals who are actually different in the real world. The distinct *locations*, along with other information about these individuals, rendered the profiles separate. Both the *Trie* and *GPT* techniques of profile consolidation produced ambiguous results in such cases, highlighting the irreplaceable value of human oversight in profile disambiguation.

Chapter 4

Identifying Weakly & Strongly Connected Entities in Interlocking Directorships

This study systematically investigates the network of interlocking directorships within Indian corporations, leveraging a systematic data collection and analysis approach. The study aims to *Identify Weakly and Strongly Connected Entities in Interlocking Directorships* and analyze the intricate web of relationships between companies and their directors using data analytics and visualization techniques.

4.1 Data Extraction using BFS Traversal

The department which maintains a comprehensive database of all corporations and their respective directors in India is the **Ministry of Corporate Affairs (MCA)**[1]. Due to certain limitations pertaining to data scraping from a government website, the primary data source of information for our pipeline is the **ZaubaCorp**[2]. Zauba is an organization that functions as an online directory for information about Indian corporations and their corresponding directors. It provides two main categories of web pages: *Company Pages* and *Director Pages*. A Company Page lists basic information about the company, including the Company Name, Company Identification Number (CIN), company address, email IDs of point of contact, and hyperlinks to current and past directors. Similarly, a Director Page contains information about the directors, such as Director Name, Director Identification Number (DIN), and links to all companies where they hold or have held directorial positions. This structure allows us to perceive a potentially endless network of corporate connections as a graph. For ease of understanding, all individuals on the board of directors for a company, including *Individual Partner, Nominee-Body Corp Partner, Director, Managing Director, Partner*, are considered as Directors Nodes.

A corporate network spanning companies and their directors is a *two-mode* or *bipartite graph*, where the nodes can be classified into two categories with an edge only from one category to another. In our case, the categories are companies and directors, with an edge from one to another implying directorship. We employ a *Breadth-First Search (BFS)* graph traversal technique to navigate through the vast amount of data efficiently. BFS traversal is an algorithm that systematically explores a graph's nodes layer

[1]https://www.mca.gov.in/
[2]https://www.zaubacorp.com/

by layer, starting from a selected node and moving outward to all its immediate neighbours before proceeding to the next level of nodes. The process of data extraction begins with the user giving a base node, which acts as the starting vertex for the graph traversal.

Figure 4.1 is an example of a two-mode network. Think of it as the blue nodes *(A, B, C, D, E, F)* signifying companies or company nodes and the orange nodes *(1, 2, 3, 4, 5, 6)* signifying directors or director nodes. The edge between the company node 'A' and director node '1' implies that '1' is a director at the company 'A' and similarly for all other nodes in the network. In order to extract the information about the entire network, we need to start with one node, let 'A' be the base node. Since 'A' is connected (has an edge) to the director nodes '1' and '2', when visiting company node 'A', we have access to the hyperlinks to '1' and '2'. Note that we only have access to hyperlinks of '1' and '2' as of now and not the information inside them. To extract information from these director nodes, we first need to visit them, or we can say we need to visit the web pages linked to them.

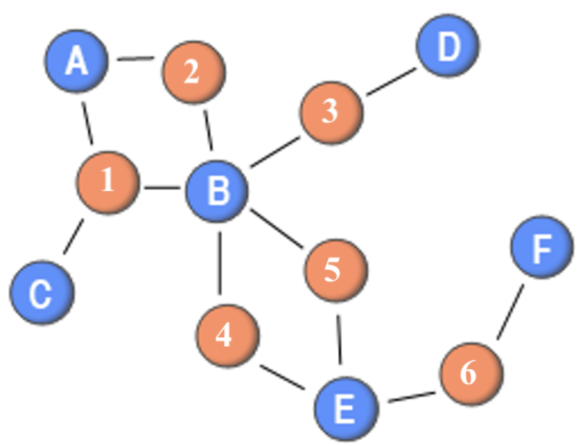


Figure 4.1: *BFS* Traversal in a two-mode graph.
Blue Nodes *(A – F)* depict Companies, Orange Nodes (1 – 6) depict Directors

The BFS traversal goes as follows

→ The BFS traversal initiates with company node 'A'. It is added to the BFS sequence or queue as the starting point.

→ Upon visiting the company node 'A', we extract the information about this company (CIN, Company Name, etc.), mark it as visited and remove it from the BFS queue. We now explore all its directly connected neighbours, which are director nodes '1' and '2' in this case and add them to the back of the BFS queue.

→ Moving on, we visit the next element in the queue, which is director node '1' and extract the information about this director (DIN, Director Name, etc.). While '1' is being marked as visited and being removed from the queue, we add the neighbouring nodes of '1', company nodes 'B' and ‘C’, to the back of the BFS queue. Since ‘A’ (another neighbour node of ‘1’) is already a visited node, it is not added to the BFS queue.

→ The element next in the queue is director node ‘2’. Upon visiting this node, we remove it from the BFS queue and add the same to the list of visited nodes. We now try to add the neighbouring company nodes of ‘2’, nodes ‘A’ and ‘B’, to the back of the BFS queue, but since ‘A’ is already visited, and ‘B’ is already present in the queue we do not add them again.

→ This entire process repeats until the queue is empty, and hence, all the nodes of the network are visited. The traversal order for the above graph with ‘A’ as the starting node is A, 1, 2, B, C, 3, 4, 5, D, E, 6, F. This order illustrates the breadth-first strategy of visiting all neighbours at the current depth before moving on to the next level.

By maintaining a *BFS Queue* and a list of *Visited* nodes during the graph traversal, this process carefully avoids revisiting the same nodes, hence ensuring an efficient exploration of the entire network. This extracted data lays the groundwork for a detailed examination of the interlocking directorates network and sets the stage for subsequent steps in our methodology.

4.2 Structure of Extracted Data

We extract the information about all the nodes in a corporate network and systematically organize them into three distinct CSV files:

1. **Company Information:** List of all the companies in the entire network with all the relevant information – *CIN, Company Name, Zauba URL*. Table 4.1 is an example of Company Info CSV.

Table 4.1: Company Information CSV

CIN	Company Name	Zauba URL
U85191MH2014NPL253500	Tata Foundation	www.zaubacorp.com/company/ TATA-FOUNDATION/ U85191MH2014NPL253500
U85110MH2010NPL207270	Reliance Foundation	www.zaubacorp.com/company/ RELIANCE-FOUNDATION/ U85110MH2010NPL207270
L21016DL2013PLC386045	Greenlam Industries Limited	www.zaubacorp.com/company/ GREENLAM-INDUSTRIES-LIMITED/L21016DL2013PL C386045
....	....	....

2. **Director Information:** List of all the directors in the entire network with all the relevant information – *DIN, Director Name, Zauba URL*. Table 4.2 is an example of Director Info CSV.

Table 4.2: Director Information CSV

DIN	Director Name	Zauba URL
00000001	Ratan Naval Tata	www.zaubacorp.com/director/RATAN-NAVAL-TATA/00000001
00019080	Jalaj Ashwin Dani	www.zaubacorp.com/director/JALAJ-ASHWIN-DANI/00019080
03115198	Nita Mukesh Ambani	www.zaubacorp.com/director/NITA-MUKESH-AMBANI/ 03115198
00006273	Gautam Shantilal Adani	https://www.zaubacorp.com/director/GAUTAM-SHANTILAL-ADANI/00006273
00010883	Anil Kumar Agarwal	https://www.zaubacorp.com/director/ANIL-KUMAR-AGARWAL/00010883
. . . .	. . . .	. . . .

3. **Company – Director Information:** A list of all companies and their directors. Each row has a company name and a director name, implying the directorship of an individual at that particular company. Table 4.3 is an example of Company – Director Info CSV.

Table 4.3: Company – Director Information CSV

Company Name	Director Name
Reliance Foundation	Jalaj Ashwin Dani
Reliance Foundation	Nita Mukesh Ambani
Reliance Foundation	Isha Mukesh Ambani
Reliance Foundation	Akash Mukesh Ambani
Reliance Foundation	Anant Mukesh Ambani
. . . .	. . . .

4.3 Visualising the Extracted Data

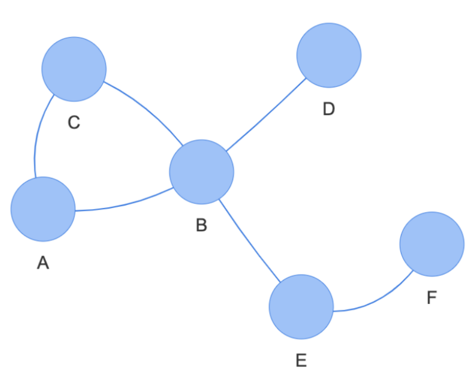
The data extracted from the source is pretty straightforward. From this clear dataset, there are *five* distinct types of visualizations that can be effectively utilized. Each visualization approach is designed to highlight different aspects and patterns within the data, providing a comprehensive understanding through graphical representation. In this section, we discuss these three distinct visualisations and understand the inferences each one of them provides.

4.3.1 Company – Director Graph

The Company – Director Graph is a foundational visualization derived from the dataset. It represents a network structure where both companies and directors are nodes. Each connection, or edge, between a director and a company signifies a directorship. Figure 4.1 is a company – director graph and this type of visualization employs a two-mode graph framework. The blue nodes depict companies, and the orange nodes depict directors. Company Nodes such as 'B' connect the majority of the nodes in the network. We call such nodes a company node that has a lot of directors, as *Star Company Node*. If it ceases to exist for some reason, for example, insolvency, it will lead to this entire network disconnecting. Similarly, a Director node, which connects a lot of other directors together through shared directorship or is on board of a lot of companies, is called a *Star Director Node*. In case of Death or Retirement, such directors disconnect the entire network.

4.3.2 Company – Company Graph

The Company – Company Graph offers a distinct perspective, focusing solely on the direct relationships between companies within the dataset. It is a one-mode graph network, where companies are the sole nodes connected by edges that signify shared directorial ties. When a single individual assumes directorial roles in two or more companies, these companies become interconnected within the graph. By visualizing these connections, we gain insights into the web of inter-dependencies among companies, highlighting potential areas of collaboration, competition, or shared strategic interests.



Edge	Connection Node(s) (Directors)	Edge Weight (No. of Connection Nodes)
A – B	[1 , 2]	2
A – C	[1]	1
B – C	[1]	1
B – D	[3]	1
B – E	[4 , 5]	2
E – F	[6]	1

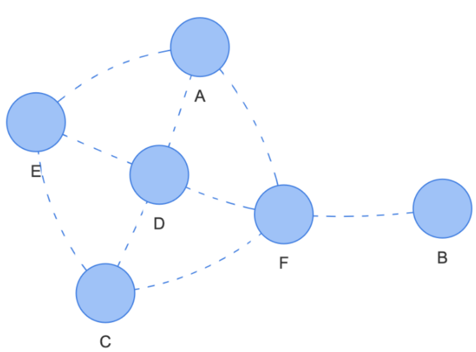
Figure 4.2: *Company – Company* Graph

Figure 4.2 serves as an illustrative depiction of the Company-Company Graph, derived from the corporate network outlined in Figure 4.1. The accompanying table in Figure 4.2 provides a detailed account of each edge, elucidating the rationale behind the connections between pairs of companies that are connected with an edge as well the strength of each connection with Edge Weight. For instance, a direct edge between Company Node ‘A’ and Company Node ‘B’ signifies the shared directorship of individuals designated as Director ‘1’ and Director ‘2’ across both companies, thereby establishing a tangible link between the two companies. Similarly, for all the other edges in Figure 4.2. A higher edge weight implies more number of shared directors and hence a stronger connection between two companies. Companies with a shared director can be referred to as 1st Degree connections.

4.3.3 Company – Company Indirect Graph

In corporate networks, addressing indirect connections between companies is also important. These connections occur when two companies lack a shared director, yet a possible path exists linking them together. While not immediately apparent, such connections can wield considerable influence within the corporate landscape.

Figure 4.3 offers a graphical representation of the Company – Company Indirect Graph for the corporate network depicted in Figure 4.1. The accompanying table in Figure 4.3 furnishes a comprehensive breakdown of each indirect connection observed in the network by describing all possible connection paths for each one of them. For example, Company Nodes ‘A’ and ‘D’ do not share a director but there exist two different paths between the two: *(i)* $A \to 1 \to B \to 3 \to D$ and *(ii)* $A \to 2 \to B \to 3 \to D$. The connection degree of an edge here is the number of distinct directors present along the shortest path between the two companies, giving us an idea of how far the two company nodes are from each other.



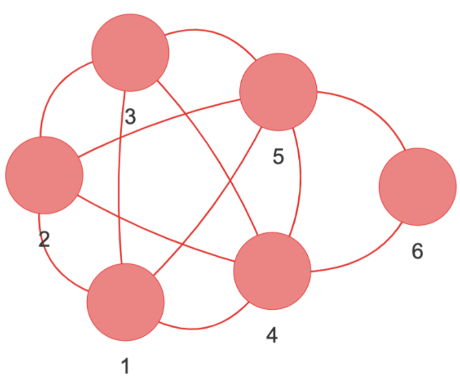
Edge	Connection Path(s) (From one Company to another)	Edge Weight (No. of Paths)	Connection Degree
A – D	A → [1, 2] → B → 3 → D	2	2
A – E	A → [1, 2] → B → [4, 5] → E	4	2
A – F	A → [1,2] → B → [4,5] → E → 6 → F	4	3
B – F	B → [4,5] → E → 6 → F	2	2
C – D	C → 1 → B → 3 → D, C → 1 → A → 2 → B → 3 → D	2	2
C – E	C → 1 → B → [4,5] → E, C → 1 → A → 2 → B → [4,5] → E	4	2
C – F	C → 1 → B → [4,5] → E → 6 → F, C → 1 → A → 2 → B → [4,5] → E → 6 → F	4	3
D – E	D → 3 → B → [4, 5] → E	2	2
D – F	D → 3 → B → [4, 5] → E → 6 → F	2	3

Figure 4.3: *Company – Company* Indirect Graph

Notice how Company Node 'B' is common among all the paths for indirectly connected companies in Figure 4.3. This shows how Company 'B' plays a pivotal role in connecting several companies directly and as well as indirectly in the network. Another observation to make is the *Edge Weight* or the *Number of Paths* from one company to another. Given two pairs of indirectly connected companies with an equal degree of connection, the pair with more paths between them will be more connected than the pair with fewer paths. For instance, the company pairs *'A' – 'D'* and *'A' – 'E'* are both 2nd degree connections, and they have a total of *2* and *4* paths, respectively. Hence, the company pair *'A' – 'E'* can be said to be more connected than the company pair *'A' – 'D'*.

4.3.4 Director – Director Graph

A Director – Director Graph mirrors the structure of the Company-Company Graph, but it centres on human individuals, specifically directors, as its nodes. In this one-mode graph network, directors are interconnected by edges denoting shared directorships within companies. When two individuals assume directorial roles in the same company, they become linked within the graph because of that company. This visualization offers valuable insights into the interconnections among directors.



Edge	Connection Node(s) (Companies)	Edge Weight	Edge	Connection Node(s) (Companies)	Edge Weight
1 – 2	[A , B]	2	2 – 5	[B]	1
1 – 3	[B]	1	3 – 4	[B]	1
1 – 4	[B]	1	3 – 5	[B]	1
1 – 5	[B]	1	4 – 5	[B , E]	2
2 – 3	[B]	1	4 – 6	[E]	1
2 – 4	[B]	1	5 – 6	[E]	1

Figure 4.4: *Director – Director* Graph

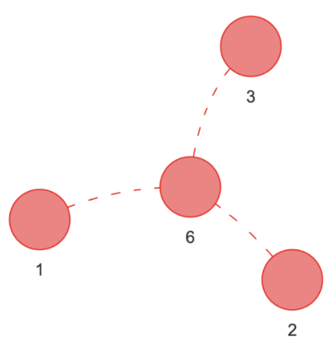
Figure 4.4 serves as an illustrative depiction of the Director-Director Graph, derived from the corporate network outlined in Figure 4.1. The accompanying table in Figure 4.4 provides a detailed rationale behind each edge, elucidating the directorial ties between pairs of individuals as well the strength of each connection with Edge Weight. For instance, a direct edge between Director node '1' and Director node '2' signifies their shared directorships at Company 'A' and Company 'B'. A higher edge weight implies more number of shared companies and hence a stronger connection between two directors. Directors who share a company can be referred to as 1st Degree connections.

What distinguishes the Director – Director Graph as particularly significant is its focus on the *human factor* within corporate networks. Directors serve as pivotal decision-makers and influencers within companies, shaping organizational strategies, guiding business operations and inspiring others with their vision.

4.3.5 Indirect Director – Director Graph

Similar to indirect connections between companies, corporate networks also have indirect connections between directors. These connections occur when two directors are not on the same board of any company, yet a possible path exists linking them together, which can very well create influence in certain ways.

Figure 4.5 offers a graphical representation of the Director – Director Indirect Graph for the corporate network depicted in Figure 4.1 as well as the accompanying table in Figure 4.5 describes each indirect connection observed between directors in the network with the possible connection paths for each one of them. For example, Director Nodes '1' and '6' do not share a company but there exist four different paths between the two: *(i)* $6 \rightarrow E \rightarrow 4 \rightarrow B \rightarrow 1$, *(ii)* $6 \rightarrow E \rightarrow 5 \rightarrow B \rightarrow 1$, *(iii)* $6 \rightarrow E \rightarrow 4 \rightarrow B \rightarrow 2 \rightarrow A \rightarrow 1$ and *(iv)* $6 \rightarrow E \rightarrow 5 \rightarrow B \rightarrow 2 \rightarrow A \rightarrow 1$. The connection degree of an edge here is the number of distinct companies present along the shortest path between the two directors, giving us an idea of how far the two director nodes are from each other.



Edge	Connection Path(s) (From one Director to another)	Edge Weight (No. of Paths)	Connection Degree
6 – 1	6 → E → [4, 5] → B → 1, 6 → E → [4, 5] → B → 2 → A → 1	4	2
6 – 2	6 → E → [4, 5] → B → 2, 6 → E → [4, 5] → B → 1 → A → 2	4	2
6 – 3	6 → E → [4, 5] → B → 3	2	2

Figure 4.5: *Company – Company* Indirect Graph

Similar to the Company – Company Indirect Graph, given that two pairs of directors are indirectly connected and have an equal degree of connection, the pair of directors with more paths between them will be more connected than the pair with fewer paths. For instance, director pairs *'1' – '6'*, *'2' – '6'* and *'3' – '6'* are all 2nd degree connections and they have a total of *4*, *4* and *2* paths, respectively. Hence, the director pairs *'1' – '6'* and *'2' – '6'* can be said to be more connected than the director pair *'3' – '6'*.

4.4 Weakly Connected Entities: *Graph Cliques*

Mathematically and computationally, identifying all cliques of all sizes (maximal/maximum cliques) is a *NP-Complete* Problem (Hua et al., 2020). In a graph G with vertices V, in order to extract all the cliques, we need to identify all subsets and check if all the vertices in the subgraph are connected to each other. For a subgraph with n nodes, to be a clique needs to have ${}^{n}C_2$ edges and hence, the time complexity for extracting all maximal cliques in the graph G can reach an exponential time of the vertex number V (Moon and Moser, 1965). Various approaches such as *Pivot Bron-Kerbosch Algorithm and Tomita Algorithm* have been studied to identify the Number of Cliques in a given graph.

To address this computational complexity while still extracting meaningful insights, we limit the network analysis to the *third-degree* connections of a specific node (either a director or a company), effectively creating a smaller, more manageable sub-network. This allows for the identification of potentially influential communities within the immediate vicinity of the chosen node.

We use the `find_cliques`[3] function from `NetworkX`[4] Library in *Python*. The implementation of this function is based on the discussion by Cazals and Karande (2008) on the *Bron and Kerbosch Algorithm* (Bron and Kerbosch, 1973) and its extension by Tomita et al. (2004). The function takes a graph G, an undirected graph, as input and returns an iterator over all the maximal cliques present in the graph.

4.4.1 Interpretation of Maximal Cliques in Corporate Networks

The concept of *Cliques* and *Maximal Cliques* is crucial in analyzing the interconnectivity and network robustness in various fields, including corporate networks where understanding tightly-knit groups can reveal insights into organizational structure and communication patterns. Our focus will be on identifying maximal cliques for a given network in order to extract all large groups of connected companies and directors.

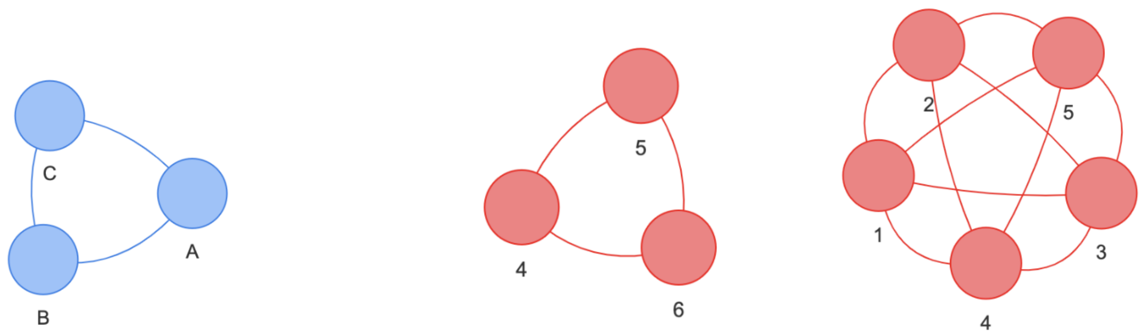
Figure 4.6 shows maximal cliques for the sample network of directors and companies discussed earlier in Figure 4.1. Various cliques (size > 2) can be identified for the corresponding Company – Company Graph and Director – Director Graph.

Maximal Cliques in Company – Company Graph (Fig 4.2): This network contain only one maximal clique. Companies *'A'*, *'B'* and *'C'* form a maximal clique of size *3*. It can be interpreted that all the companies of this maximal clique are connected and hence have *at least* one director in common for each pair. Companies *'A'* and *'B'* share *2* directors – *'1'* and *'2'*, where as companies *'A'* and *'C'* as well as companies *'B'* and *'C'* share only one director – *'1'*.

Maximal Cliques in Director – Director Graph (Fig 4.4): This network contain two maximal cliques.

[3] https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms.clique.find_cliques.html
[4] https://networkx.org/

Directors *'1'*, *'2'*, *'3'*, *'4'*, *'5'* form a maximal clique of size *5* and Directors *'4'*, *'5'*, *'6'* for a maximal clique of size *3*. It can be interpreted that all the directors of this maximal clique are connected and hence have *at least* one company in common for each pair. Companies *'1'* and *'2'*, and companies *'4'* and *'5'* share *2* directors – *'A', 'B'* and *'B','E'* respectively. All the other pairs of companies share *1* director – *'B'*.



3 – Vertex Clique *(ABC)*

3 – Vertex Clique *(456)*

5 – Vertex Clique *(12345)*

(a)

(b)

Figure 4.6: Maximal Cliques for Sample Corporate Network in Fig 4.1 (a) Maximal Cliques in Company – Company Graph in Fig *4.2*, (b) Maximal Cliques in Director – Director Graph in Fig *4.4*

A larger clique or a clique with a larger number of nodes (directors or companies) has more power to create influence in a corporate network as compared to a smaller clique. For instance, In Figure 4.6(b) the *5 – vertex* Director clique *(12345)* is a larger in size than the *3 – vertex* director clique *(456)*. With more Directors connected with each other, it has more power to create influence.

Although every pair of directors or companies is connected in a clique, the level or strength of each connection is different. We cannot say that all the pairs are equally connected. A pair of directors sharing more companies is strongly connected compared to a pair with fewer companies. Similarly, this can be extended to company cliques. In order to identify strongly connected components, we turn to mining *Frequent Itemsets* of directors and companies in a corporate network.

4.5 Strongly Connected Entities: *Frequent Itemsets*

4.5.1 Interpretation of Frequent Itemsets in Corporate Networks

The extracted data is reorganised into specific formats required for this analysis to identify frequent itemsets within a corporate network, such as groups of directors or companies that appear together frequently.

The resulting data after the processing looks like this:

Updated Company Information: We add a column *"List of Directors"* to the Company Info (Table 4.1) with the help of Company – Director Info (Table 4.3). The column contains all the directors corresponding to a particular company. Table 4.4 is an example of Updated Company Info.

Table 4.4: Updated Company Information CSV

CIN	Company Name	List of Directors
CIN_1	C_1	$[D_1, D_2, D_3, D_4]$
CIN_2	C_2	$[D_3, D_4, D_5, D_6, D_7]$
CIN_3	C_3	$[D_1, D_4, D_7, D_8]$
....	....	....

Updated Director Information: We add a column *"List of Companies"* to the Director Info (Table 4.2) with the help of Company – Director Info (Table 4.3). The column contains all the companies for which an individual is a director. Table 4.5 is an example of Updated Director Info.

Table 4.5: Updated Director Information CSV

CIN	Director Name	List of Companies
DIN_1	D_1	$[C_1, C_2, C_3]$
DIN_2	D_2	$[C_3, C_4, C_5, C_6, C_7]$
DIN_3	D_3	$[C_1, C_4, C_7, C_8]$
....	....	....

4.5.1.1 Frequent Director Itemsets

Aim: Our aim is to identify sets of Directors that appear together, sharing directorships, frequently in a corporate network. This helps us identify *strongly connected directors*. We identify the companies corresponding to these directors for each frequent director itemset. To extract the same, we utilize the *List of Directors* corresponding to each company in the *Updated Company Info* file.

Attached below is the Python code that uses `fpmax`[5] function developed by `mlxtend` to extract Maximal Frequent Itemsets from a given database. The implementation of this function is based on the algorithm developed by Grahne and Zhu (2003). `min_support`, *0.0001* here, is the threshold for an

[5] https://rasbt.github.io/mlxtend/user_guide/frequent_patterns/fpmax/

itemset to be considered as frequent. For ex., if the length of the entire dataset is *10,000* rows, an itemset needs to occur at least *10,000 x 0.0001 = 10* times to be considered as *frequent*.

```
from mlxtend.preprocessing import TransactionEncoder
from mlxtend.frequent_patterns import fpmax
import pandas as pd

# Reading Updated Company Info for List of Directors corresponding to each Company
data = pd.read_csv('./updated_company_info.csv')
director_lists = data['List of Directors'].str.split(',')

# Convert the List of Directors into a format suitable using TransactionEncoder
te = TransactionEncoder()
te_ary = te.fit(director_lists).transform(director_lists)
df_te = pd.DataFrame(te_ary, columns=te.columns_)

maximal_frequent_itemsets = fpmax(df_te, min_support=0.0001, use_colnames=True)
```

Python code to extract Maximal Frequent Itemsets using `fpmax` *function*

Table 4.6: Example Results Maximal Frequent Director Itemset and Corresponding Companies

Support	**Frequent Director Itemset**	**Intersecting Companies**
s_1	$[D_4...]$	$[C_x, ...]$
s_2	$[D_4, D_7, D_5...]$	$[C_a, ...]$
s_3	$[D_2, D_3...]$	$[C_x, C_a, C_z...]$
. . . .	. . . .	. . . .

4.5.1.2 Frequent Company Itemsets

Aim: To identify sets of Companies that appear together, sharing directors, frequently in a corporate network. We also want to identify the directors shared between these frequently occurring companies for each frequent company itemset. This helps us identify *strongly connected companies*. To extract the same, we utilize the same with code as described above with *List of Companies* corresponding to each company in the *Updated Director Info* file.

Table 4.7: Example Results Maximal Frequent Company Itemset and Corresponding Directors

Support	**Frequent Company Itemset**	**Intersecting Directors**
s_1	$[C_1...]$	$[D_x, ...]$
s_2	$[C_2, C_4, C_5...]$	$[C_a, ...]$
s_3	$[C_2, C_7...]$	$[D_x, D_a, D_z...]$
. . . .	. . . .	. . . .

4.6 Extracted Data Corpus

The dataset we extracted for this study encompasses a comprehensive network of corporate directors and the companies they are affiliated with. This section details the characteristics and scale of the data used to conduct the analysis.

Since we need to provide a base node to begin graph traversal as explained before, in order to have access to a large dataset, we decided our base node to be one of India's largest conglomerate – ***Reliance Foundation***. The dataset comprises of ***54,216 Directors***, ***87,187 Companies*** and ***2,99,970 Company – Director Edges*** in total. This network of directors and companies includes some of India's largest conglomerates such as *Adani Group, Tata Group, Vedanta Limited, Aditya Birla Group, Airtel, HDFC Bank, State Bank of India, Air India, Infosys Limited* etc.

We performed a *Last Name* analysis on the *List of Directors* similar to Chandrashekar and Muralidharan (2012). Mentioned below are the *20* Last Names sorted in descending order of occurrence. It shows a dominance by certain groups *(Marwaris, Baniyas, Gujratis, Jains, etc)* over others in Indian corporations.

S no.	Last Name	Frequency	S no.	Last Name	Frequency
1	Shah	*1250*	11	Agrawal	*234*
2	Patel	*809*	12	Desai	*230*
3	Singh	*704*	13	Joshi	*204*
4	Jain	*640*	14	Mittal	*198*
5	Gupta	*634*	15	Parikh	*189*
6	Mehta	*580*	16	Goyal	*180*
7	Agarwal	*533*	17	Gandhi	*180*
8	Kumar	*493*	18	Reddy	*177*
9	Sharma	*468*	19	Aggarwal	*174*
10	Rao	*252*	20	Arora	*159*

Table 4.8: Top *20* Last Names Based on Frequency in Dataset

We also found that *37,123* companies of total *87,187* – almost *42.5%*, have at least one pair of directors that share the same last name. One can say that in India, corporations are primarily family-concentrated compared to other nations.

Figure 4.7 represents the frequency distribution of the number of directors for a given company. We observe that most companies have either *2* (*43%* of the total companies) or *3* (*26%* of the total companies) distinct directors. On the other hand, figure 4.8 represents the frequency distribution of a number of companies corresponding to a given individual (director). We observe that most individuals hold directorial positions at either *1* (*41%* of the total directors) or *2* (*17%* of the total directors) distinct companies and around *58.6%* are directors for *2* or more companies.

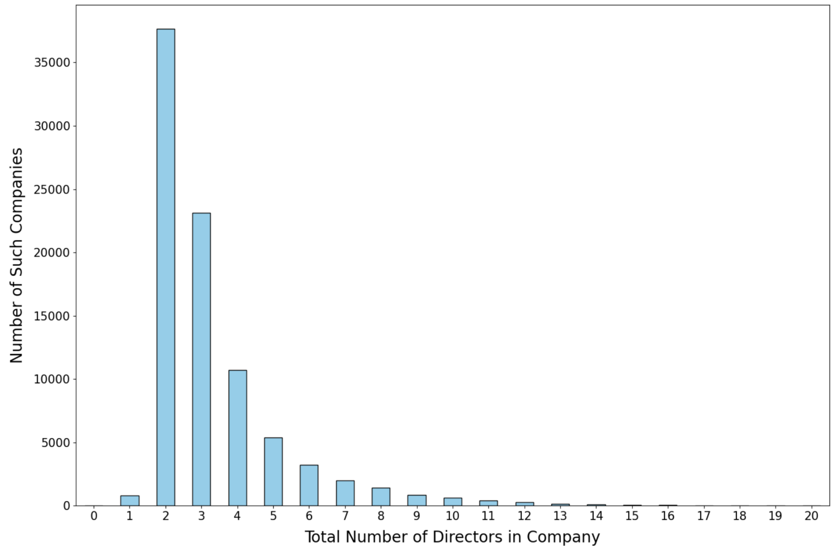


Figure 4.7: Frequency Distribution of Number of Directors Corresponding to a Company

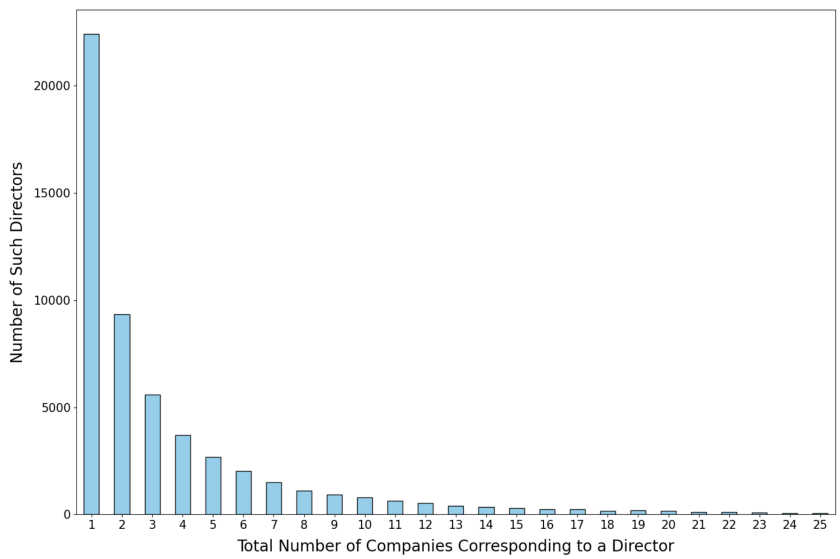


Figure 4.8: Frequency Distribution of Number of Companies Corresponding to a Director

The Extracted Data can be found at:

- **Director Information:** https://shorturl.at/zJLP0
- **Company Information:** https://shorturl.at/svRSZ
- **Company – Director Information:** https://shorturl.at/isuIW

4.7 LLM Driven Director – Director Relation Identification

The examination of corporate networks extends beyond the mere identification of connected individuals. Our study delves deeper into uncovering latent connections and underlying factors contributing to these director-director associations. Employing an adapted iteration of the *LLM Driven Web Profile Extraction* pipeline delineated in Chapter 3, we extract pertinent insights regarding the interrelations between two directors. These connections traverse a spectrum from personal affiliations, such as familial bonds, to professional ties encompassing shared work histories, concurrent memberships in professional organizations, and other affiliations within the professional sphere.

4.7.1 Personal Relation Identification

In order to check for *Personal Relation* between two directors, we employ a diverse array of web search techniques to procure information concerning personal affiliations available across search engines. Utilizing varied web queries such as *"Director 1, Director 2"* or *"Director 1, Director 2 relation"* or *"Director 1, Director 2 Family tree"* directed towards search engines, we retrieve the top *five* search results and proceed to extract textual information through web scraping.

We develop a specialized prompt for the task of identifying personal relations among directors from a text using the same Prompt Engineering techniques as in Section 3.2.3.

→ ***Task Description:*** You are a linguistic analyst. The task is to analyze a given text and determine if there is any familial relationship implied between {*Director 1*} and {*Director 2*}.

→ ***Predefining Requirements:*** Use this list to return the identified relation: *"husband - wife", "daughter - father", "nephew - uncle", "mother - son", "sister - brother", "grandfather - granddaughter", "grandmother - grandson", "cousin - cousin", "aunt - nephew", "stepmother - stepson", "stepfather - stepdaughter", "godmother - godson", "adoptive mother - adopted son", "sister-in-law - brother-in-law", "friend - friend".*

→ ***Edge Case Handling:***

- If the text does not mention either of the directors, return the answer as *'Not Available'*.
- If the text mentions both directors but does not imply any familial relationship between them, return the answer as *'Not Available'*.

→ ***Output Formatting:*** Return the identified familial relationship in *JSON* format with key as *"Relation"*: *"husband - wife"*

→ ***Self-Verification and Reinforced Learning:*** Make sure the identified relationships are based solely on explicit textual evidence. Do not make assumptions from director names.

Prompt for Personal Relation Identification

The carefully devised prompt mentioned above ensures consistency and standardization, with the model primed to discern from a predefined list of *15* relation pairs such as *Wife – Husband, Daughter – Father, Nephew – Uncle, etc*. By attaching the textual data that we have scraped along with this prompt, we get our desired output of any inference regarding any familial ties implicit within the provided text. If no such relation is identified, the model returns the answer as *'Not Available'* and in such a case, we move on to find *Professional Link* between the directors.

Example on Personal Relation Identification:

→ Director 1: *Mukesh Dhirubhai Ambani (DIN: 00001695)*

→ Director 2: *Hital Rasiklal Meswani (DIN: 00001623)*

→ Search Query: *Mukesh Dhirubhai Ambani, Hital Rasiklal Meswani*

→ Search Engine Results: ***https://www.india.com/business/meet-nikhil-and-hital-meswani-the-cousins-who-are-reliances-highest-paid-employees-6214986/***

→ Extracted Text:

```
Meet Nikhil and Hital Meswani, The Cousins Who Are Reliance's
Highest-Paid Employees Meet Nikhil and Hital Meswani,
                        . . .
                        . . .
    About Hital MeswaniHital Meswani is the nephew
    of Mukesh Ambani and the son of Rasiklal Meswani.
                    . . . . . . . .
```

→ GPT Output:

```
{
    "Relation": uncle - nephew
}
```

4.7.2 Professional Relation Identification

In order to extract Professional links among directors, we leverage the Web Profile Extraction pipeline discussed in Chapter 3. This involves a systematic process encompassing web searches, data scraping, LLM-driven Named Entity Recognition (NER), and standardization of information sourced from Wikipedia. This culminates in the creation of structured web profiles for both directors, which are presented in JSON format. Subsequently, through the identification of commonalities among organizations, institutions, and other pertinent details extracted from these profiles, we identify any matches indicative of professional affiliations.

Example on Professional Relation Identification:

→ Director 1: *Abhiraj Singh Bhal* [6] *(DIN: 07005253)*

→ Director 2: *Varun Khaitan*[7] *(DIN: 07005033)*

→ Extracted Web Profile for Director 1:


```
{
    "Name": Abhiraj Singh Bhal,
    "Contact Information":
    {
        "Email IDs": Unknown,
        "Weblinks": [linkedin.com/in/abhirajbhal,urbancompany.com]
    },
    "Locations": [Gurgaon, Haryana, India, Mumbai],
    "Education":
    [
        {
            "Institute Name": Indian Institute of Management
               Ahmedabad,
            "Institute_Name_Wikipedia_Link": https://en.wikipedia.
               org/wiki/Indian_Institute_of_Management_Ahmedabad,
            "Course Name": MBA,
            "Course_Name_Wikipedia_Link": https://en.wikipedia.org
               /wiki/Master_of_Business_Administration,
            "Branch": Business Administration,
            "Grade": Unknown,
            "Duration": 2009 - 2011
        },
        {
            "Institute Name": Indian Institute of Technology,
               Kanpur,
            "Institute_Name_Wikipedia_Link": https://en.wikipedia.
               org/wiki/IIT_Kanpur,
            "Course Name": Bachelor of Technology,
            "Course_Name_Wikipedia_Link": https://en.wikipedia.org
               /wiki/Bachelor_of_Technology,
            "Branch": Electrical Engineering,
            "Grade": Unknown,
            "Duration": 2005 - 2009
        }
    ],
```


[6] https://www.linkedin.com/in/abhirajbhal/
[7] https://www.linkedin.com/in/varunkhaitan/

```
        "Work Experience":
        [
            {
                "Organisation Name": Urban Company,
                "Role": Co-Founder,
                "Job Description": Largest home services in Asia,
                "Duration": Sep 2014 - Present
            },
            {
                "Organisation Name": The Boston Consulting Group,
                "Organisation_Name_Wikipedia_Link": https://en.
                   wikipedia.org/wiki/Boston_Consulting_Group,
                "Role": Consultant,
                "Job Description": Consulting,
                "Duration": May 2011 - Apr 2014
            }
        ]
    }
```

→ Extracted Web Profile for Director 2:

```
    {
        "Name": Varun Khaitan,
        "Contact Information":
        {
            "Email IDs": Unknown,
            "Weblinks": linkedin.com/in/varunkhaitan
        },
        "Locations": [New Delhi Area, India, New York, Istanbul],
        "Education":
        [
            {
               "Institute Name": Indian Institute of Technology,
                  Kanpur,
                "Institute_Name_Wikipedia_Link": https://en.wikipedia.
                   org/wiki/IIT_Kanpur,
                "Course Name": Bachelor of Technology,
                "Course_Name_Wikipedia_Link": https://en.wikipedia.org
                   /wiki/Bachelor_of_Technology,
                "Branch": Electrical Engineering,
                "Grade": Unknown,
                "Duration": 2005 - 2009
            }
        ],
```

```
    "Work Experience":
    [
        {
            "Organisation Name": Urban Company,
            "Role": Co-founder,
            "Job Description": Home services platform,
            "Duration": Sep 2014 - Present
        },
        {
            "Organisation Name": The Boston Consulting Group,
            "Organisation_Name_Wikipedia_Link": https://en.
               wikipedia.org/wiki/Boston_Consulting_Group,
            "Role": Consultant,
            "Job Description": Consulting,
            "Duration": Oct 2013 - Apr 2014
        },
        {
            "Organisation Name": Qualcomm,
            "Organisation_Name_Wikipedia_Link": https://en.
               wikipedia.org/wiki/Qualcomm,
            "Role": Engineer,
            "Job Description": Inventor on 6 US patents,
            "Duration": Dec 2009 - Jul 2011
        }
    ],
}
```

→ Professional Links Identified:

- *Electrical Engineering @ Indian Institute of Technology, Kanpur. Same Batch 2005 – 2009.*
- *Consultant @ The Boston Consulting Group.*
    - Abhiraj Singh Bhal as Consultant from *May 2011* to *April 2014*.
    - Varun Khaitan as Consultant from *October 2013* to *April 2014*.
- Both Co – Founded *Urban Company* in *September 2014*

4.8 Results

4.8.1 Maximal Director Cliques: *MDCs*

To illustrate the utility of identifying Maximal Director Cliques (MDCs) within a corporate network, we conducted an analysis of several example cases. This analysis involved the following steps:

- Identifying the *first, second* and *third – degree* connections (directors) of a chosen (base) director.
- Cliques observed among *first* and *second degree* connections will be a subset of Cliques observed among *third degree* connection, so we extract cliques for *third degree* connection.
- Extracting MDCs from this network where the base director is present.
- Sorting these MDCs based on two criteria: (1) the number of directors in the clique and (2) the number of intersecting companies that these directors share board seats.
- Highlighting any significant observations.

Example 1 *Ratan Naval Tata (DIN: 00000001)* as the *Base Node*

- **1st Degree Connections:**
  - *No. of Directors: 16*
  - *No. of Companies: 9*
- **2nd Degree Connections:**
  - *No. of Directors: 256*
  - *No. of Companies: 88*
- **3rd Degree Connections:**
  - *No. of Directors: 2996*
  - *No. of Companies: 1100*

- **Analysis of 3rd Degree Network MDCs containing *Ratan Naval Tata***
  - Total number of MDCs: *4*
  - Average number of directors in an MDC: *4 – 5*
  - Largest MDC in size (Maximum number of Directors):
    * Clique: *6* Directors – *Ratan Naval Tata, Bindu Ananth, Nandan Mohan Nilekani, Annette Austin, Sukanya Kripalu, Subramaniam Somasundaram, Vijay Laxman Kelkar*
    * Companies Shared: *3* Companies – *Avanti Microfinance Private Limited, Indian Institute for Human Settlements, Avanti Finance Private Limited*
  - Smallest MDC in size (Minimum number of Directors):
    * Clique: *2* Directors – *Ratan Naval Tata, Venkataramanan Ramachandran*
    * Companies Shared: *3* Companies – *Ratan Tata Foundation, RNT Capital Advisers LLP, Avanti Capital Advisors LLP*
  - Clique with directors sharing maximum number of companies:

    * Clique: *7* Directors – *Ratan Naval Tata, Shantanu Naidu, Raman Kalyanakrishnan, Raghavan Ramachandra Shastri, Burzis Shapur Taraporevala, Mehli Kersasp Mistry, Nishita Manoj Murarka*
    * Companies Shared: *4* Companies – *Foundation for Rural Entrepreneurship Development, RNT Associates Private Limited, Ratan Tata Endowment Foundation, Advanced Veterinary Care Foundation*
  - Clique with directors sharing minimum number of companies:
    * Clique: *3* Directors – *Ratan Naval Tata, Sridhar Narayan, Samir Yajnik*
    * Companies Shared: *1* Companies – *Electrodrive Powertrain Solutions Private Limited*

Example 2: *Mukesh Dhirubhai Ambani (DIN: 00001695)* as the *Base Node*

- **1st Degree Connections:**
  - *No. of Directors: 40*
  - *No. of Companies: 9*
- **2nd Degree Connections:**
  - *No. of Directors: 617*
  - *No. of Companies: 190*
- **3rd Degree Connections:**
  - *No. of Directors: 6102*
  - *No. of Companies: 2960*

- **Analysis of 3rd Degree Network MDCs containing *Mukesh Dhirubhai Ambani***
  - Total number of MDCs: *17*
  - Average number of directors in an MDC: *8 – 9*
  - Largest MDC in size (Maximum number of Directors):
    * Clique: *12* Directors – *Akash Mukesh Ambani, Isha Mukesh Ambani, Anant Mukesh Ambani, Pankaj Mohan Pawar, Mukesh Dhirubhai Ambani, Manoj Harjivandas Modi, Shumeet Banerji, Raminder Singh Gujral, Haigreve Khaitan, Dileep Chinubhai Choksi, Donald Stewart Harrison, Dinesh Hasmukhrai Kanabar*
    * Companies Shared: *10* Companies – *Shripal Enterprises LLP, Reliance Foundation, Jio Platforms Limited, Reliance Foundation Institution of Education and Research, Reliance Jio Infocomm Limited, Samarjit Enterprises LLP, Reliance Retail Ventures Limited, Reliance Industries Limited, Adani Green Energy Limited, Reliance Retail Limited*
  - Smallest MDC in size (Minimum number of Directors):
    * Clique: *4* Directors – *Hital Rasiklal Meswani, Shiv Kumar Bhardwaj, Pawan Kumar Kapil, Mukesh Dhirubhai Ambani*
    * Companies Shared: *5* Companies – *The Indian Film Combine Pvt Ltd, Reliance Petroleum Limited, Reliance Industries Limited, Reliance Sibur Elastomers Private Limited, Indian Petrochemicals Corporation Limited*
  - Clique with directors sharing maximum number of companies:

    - Clique: *9* Directors – *Akash Mukesh Ambani, Isha Mukesh Ambani, Anant Mukesh Ambani, Nita Mukesh Ambani, Shumeet Banerji, Adil Zainulbhai, Raminder Singh Gujral, Manoj Harjivandas Modi, Mukesh Dhirubhai Ambani*
    - Companies Shared: *16* Companies – *Devarshi Commercials LLP, Karuna Commercials LLP, Shripal Enterprises LLP, EIH Limited, Tattvam Enterprises LLP, Reliance Foundation, Jio Platforms Limited, Reliance Foundation Institution of Education and Research, Reliance Jio Infocomm Limited, Samarjit Enterprises LLP, Reliance Retail Ventures Limited, Janardan Commercials LLP, Reliance Industries Limited, Chakradev Enterprises LLP, Reliance Retail Limited, Shivangi Commercials LLP*
  - Clique with directors sharing minimum number of companies:
    - Clique: *8* Directors – *Anand Jaikumar Jain, Rajendra Singh Lodha, Sandeep Junnarkar, Mukesh Dhirubhai Ambani, Nikhil Rasiklal Meswani, Kamal Nanavaty Pantilal, Shiv Kumar Bhardwaj, Sandeshkumar Jagannath Anand*
    - Companies Shared: *2* Companies – *Indian Petrochemicals Corporation Limited, Reliance Industries Limited*

Example 3: *Gautam Shantilal Adani (DIN: 00006273)* as the *Base Node*

- **1st Degree Connections:**
  - *No. of Directors: 44*
  - *No. of Companies: 12*
- **2nd Degree Connections:**
  - *No. of Directors: 1002*
  - *No. of Companies: 284*
- **3rd Degree Connections:**
  - *No. of Directors: 9178*
  - *No. of Companies: 4002*

- **Analysis of 3rd Degree Network MDCs containing *Gautam Shantilal Adani***
  - Total number of MDCs: *22*
  - Average number of directors in an MDC: *5 – 6*
  - Largest MDC in size (Maximum number of Directors):
    - Clique: *9* Directors – *Gautam Shantilal Adani, Shashi Shanker, Friga Noy Ahlem, Pranav Vinod Adani, Olivier Marc Sabrie, Suresh Manglani, Shailesh Vishnubhai Haribhakti, Gauri Trivedi, Chandra Iyengar*
    - Companies Shared: *8* Companies – *IndianOil - Adani Gas Private Limited, Adani TotalEnergies Biomass Limited, Adani TotalEnergies E-Mobility Limited, Adani Total Gas Limited, Adani Power Limited, Total Adani Fuels Marketing Private Limited, Adani Green Energy Limited, Adani Enterprises Limited*
  - Smallest MDC in size (Minimum number of Directors):
    - Clique: *3* Directors – *Gautam Shantilal Adani, Shailesh Vishnubhai Haribhakti, Palamadai Sundararajan Jayakumar*

        * Companies Shared: *3* Companies – *Adani Ports and Special Economic Zone Limited, Adani Total Gas Limited, Future Generali India Life Insurance Company Limited*

  – Clique with directors sharing maximum number of companies:

        * Clique: *5* Directors – *Gautam Shantilal Adani, Rajesh Shantilal Adani, Pranav Vinod Adani, Bhavik Bharatkumar Shah, Ameetkumar Hiranyakumar Desai*
        * Companies Shared: *21* Companies – *Adani Institute for Education and Research, Adani Welspun Exploration Limited, B2B India Private Limited, Adani Ports and Special Economic Zone Limited, Adani Transmission Limited, Adani Enterprises Limited, Mundra Special Economic Zone Limited, Adani Logistics Limited, Adani Tradeline LLP, Adani Tradeline Private Limited, Ambuja Cements Limited, Adani Total Gas Limited, Adani Advisory LLP, Karnavati Museum of Leadership Foundation, Adani Green Energy Limited, Adani Infrastructure Services Private Limited, Adani Infrastructure Private Limited, Adani Tradings Services LLP, Adani Power Limited, Adani Cements Limited, Baramati Power Private Limited*

  – Clique with directors sharing minimum number of companies:

        * Clique: *3* Directors – *Gautam Shantilal Adani, Shailesh Vishnubhai Haribhakti, Palamadai Sundararajan Jayakumar*
        * Companies Shared: *3* Companies – *Adani Ports and Special Economic Zone Limited, Adani Total Gas Limited, Future Generali India Life Insurance Company Limited*

Example 4: *Kiran Mazumdar Shaw (DIN: 00347229)* as the *Base Node*

- **1st Degree Connections:**
  - *No. of Directors: 68*
  - *No. of Companies: 18*
- **2nd Degree Connections:**
  - *No. of Directors: 927*
  - *No. of Companies: 312*
- **3rd Degree Connections:**
  - *No. of Directors: 8272*
  - *No. of Companies: 4037*

- **Analysis of 3rd Degree Network MDCs containing *Kiran Mazumdar Shaw***
  - Total number of MDCs: *16*
  - Average number of directors in an MDC: *6 – 7*
  - Largest MDC in size (Maximum number of Directors):
    * Clique: *12* Directors – *Jahnavi Keshao Phalkey, Vijaya Chandru, Senapathy Gopalakrishnan, Inutri Srinivas Nagesh Prasad, Rohini Nilekani, Govindan Rangarajan, Geetha Narayanan, Shashidhara Subrahmanya Lingadahalli, Anil Parameswaransarala Kumar, Ajjahalli Basavaiah Basavaraju, Kiran Mazumdar Shaw, Ekkati Venkataramana Reddy*

- * Companies Shared: *4* Companies – *Foundation for Science Innovation and Development, Science Gallery Bengaluru, IISC Medical School Foundation, i-Hub for Robotics and Autonomous Systems Innovation Foundation*
- Smallest MDC in size (Minimum number of Directors):
  - * Clique: *3* Directors – *Kiran Mazumdar Shaw, Senapathy Gopalakrishnan, Naushad Darius Forbes*
  - * Companies Shared: *3* Companies – *CSEP Research Foundation, Science Gallery Bengaluru, Nayanta Education Foundation*
- Clique with directors sharing maximum number of companies:
  - * Clique: *8* Directors – *Emmanuel Rupert, Viren Prasad Shetty, Shankar Arunachalam, Naveen Tewari, Nachiket Mor, Terri Smith Bresenham, Devi Prasad Shetty, Kiran Mazumdar Shaw*
  - * Companies Shared: *9* Companies – *Mazumdar Shaw Medical Foundation, Narayana Vaishno Devi Specialty Hospitals Private Limited, Narayana Hrudayalaya Surgical Hospital Private Limited, Narayana Institute for Advanced Research Private Limited, Narayana Hospitals Private Limited, Narayana Hrudayalaya Limited, Meridian Medical Research & Hospital Ltd., Narayana Health Institutions Private Limited, NH Integrated Care Private Limited*
- Clique with directors sharing minimum number of companies:
  - * Clique: *5* Directors – *Ranjan Ramdas Pai, Kiran Mazumdar Shaw, Siddhartha Mukherjee, Prem Pavoor, Arun Anand*
  - * Companies Shared: *1* Companies – *Immuneel Therapeutics Private Limited*

Example 5: *Falguni Sanjay Nayar (DIN: 00003633)* as the *Base Node*

- **1st Degree Connections:**
  - *No. of Directors: 31*
  - *No. of Companies: 10*
- **2nd Degree Connections:**
  - *No. of Directors: 730*
  - *No. of Companies: 176*
- **3rd Degree Connections:**
  - *No. of Directors: 7151*
  - *No. of Companies: 3452*

- **Analysis of 3rd Degree Network MDCs containing *Falguni Sanjay Nayar***
  - Total number of MDCs: *6*
  - Average number of directors in an MDC: *7 – 8*
  - Largest MDC in size (Maximum number of Directors):
    - * Clique: *14* Directors – *Ajay Kumar Dua, Pattamadai Natarajasarma Vijay, Satyavati Berera, Saket Burman, Mukesh Hari Butani, Pritam Das Narang, Mohit Burman, Amit Burman, Ravindra Chandra Bhargava, Subbaraman Narayan, Falguni Sanjay Nayar, Aditya Chand Burman, Rajiv Mehrishi, Ajit Mohan Sharan*

- ∗ Companies Shared: *7* Companies – *Lite Bite Foods Private Limited, Dabur India Limited, Fem Care Pharma Limited, Aviva Life Insurance Company India Ltd, Dabur Foods Limited, H & B Stores Limited, Interx Laboratories Private Limited*

– Smallest MDC in size (Minimum number of Directors):

- ∗ Clique: *2* Directors – *Falguni Sanjay Nayar, Rashmi Vinodchandra Mehta*
- ∗ Companies Shared: *1* Companies – *Golf Land Developers Private Limited*

– Clique with directors sharing maximum number of companies:

- ∗ Clique: *10* Directors – *Falguni Sanjay Nayar, Anchit Nayar, Alpana Parida, Pradeep Parameswaran, Sanjay Omprakash Nayar, Adwaita Sanjay Nayar, Seshashayee Sampathiyengar Sridhara, Milan Bhagwandas Khakhar, Milind Shripad Sarwate, Anita Ramachandran*
- ∗ Companies Shared: *13* Companies – *FSN International Private Limited, FSN Brands Marketing Private Limited, Metropolis Healthcare Limited, Nykaa Fashion Private Limited, Geometric Limited, Sealink View Probuild Private Limited, Heritage View Developers Private Limited, Nykaa E-Retail Private Limited, Sanjay & Falguni Nayar Foundation, Sea View Probuild Private Limited, Sorin Advisors LLP, 72 Ventures LLP, FSN E-Commerce Ventures Limited*

– Clique with directors sharing minimum number of companies:

- ∗ Clique: *2* Directors – *Falguni Sanjay Nayar, Rashmi Vinodchandra Mehta*
- ∗ Companies Shared: *1* Companies – *Golf Land Developers Private Limited*

Table 4.9 is a consolidated representation of the results for the above examples discussing MDCs.

Director Name	**Col 1**	**Col 2**	**Col 3**	**Col 4**	**Col 5**	**Col 6**
Ratan Naval Tata	2996, 1100	4, 4.75	6, 3	2, 3	7, 4	3, 1
Mukesh Dhirubhai Ambani	6102, 2960	17, 8.05	12, 10	4, 5	9, 16	8, 2
Gautam Shantilal Adani	9178, 4002	22, 5.45	9, 8	3, 3	5, 21	3, 3
Kiran Mazumdar Shaw	8272, 4037	16, 6.31	12, 4	3, 3	8, 9	5, 1
Falguni Sanjay Nayar	7151, 3452	6, 7.5	14, 7	2, 1	10, 13	2, 1

Table 4.9: Consolidated Results for *MDCs* Examples. **Col 1:** *No. of Directors in 3rd degree connections, No. of Companies in 3rd degree connections*, **Col 2:** *No. of MDCs that include the corresponding director, Average No. of directors in such MDCs*, **Col 3:** *Size of MDC with maximum No. of directors, No. of companies shared*, **Col 4:** *Size of MDC with minimum No. of directors, No. of companies shared*, **Col 5:** *Size of MDC which shares maximum no. of companies, No. of companies shared*, **Col 6:** *Size of MDC which shares minimum no. of companies, No. of companies shared*

4.8.2 Maximal Company Cliques: *MCCs*

Similar to MDCs, to illustrate the utility of identifying Maximal Company Cliques (MCCs), we conducted an analysis on several examples. This analysis involved the following steps:

- Identifying the *first, second* and *third – degree* connections (companies) of a chosen (base) company.
- Cliques observed among *first* and *second degree* connections will be a subset of Cliques observed among *third degree* connection, so we extract cliques for *third degree* connection.
- Extracting MCCs from this network where the base company is present.
- Sorting these MCCs based on two criteria: (1) the number of companies in the clique and (2) the number of intersecting directors that these companies share.
- Highlighting any significant observations.

Example 1: *Hindustan Petroleum Corporation Limited (CIN: L23201MH1952GOI008858)* as the *Base Node*

- **1st Degree Connections:**
  - *No. of Companies: 15*
  - *No. of Directors: 6*
- **2nd Degree Connections:**
  - *No. of Companies: 134*
  - *No. of Directors: 32*
- **3rd Degree Connections:**
  - *No. of Companies: 720*
  - *No. of Directors: 175*

- **Analysis of 3rd Degree Network MCCs containing *Hindustan Petroleum Corporation Limited***
  - Total number of MCCs: *6*
  - Average number of companies in an MCC: *5*
  - Largest MCC in size (Maximum number of Companies):
    * Clique: *7* Companies – *Ratnagiri Refinery and Petrochemicals Limited, Prize Petroleum Company Limited, HPCL Biofuels Limited, Mangalore Refinery and Petrochemicals Limited, Hindustan Petroleum Corporation Limited, HPCL Rajasthan Refinery Limited, HPCL-Mittal Energy Limited*
    * Directors Shared: *4* Directors – *Bharathan Shunmugavel, Pushp Kumar Joshi, Rajneesh Narang, Suresh Kasargod Shetty*
  - Smallest MCC in size (Minimum number of Companies):
    * Clique: *3* Companies – *CREDA - HPCL Bio Fuel Limited, Hindustan Petroleum Corporation Limited, Petronet India Limited*
    * Directors Shared: *4* Directors – *Shrikant Madhukar Bhosekar, Ramaswamy Jagannathan, Pushp Kumar Joshi, Satya Prakash Gupta*

- Clique with companies sharing maximum number of directors:
  - Clique: *7* Companies – *HPCL-Mittal Energy Limited, HPCL Rajasthan Refinery Limited, Hindustan Petroleum Corporation Limited, Prize Petroleum Company Limited, Hindustan Colas Private Limited, South Asia LPG Company Private Limited, HPCL-Mittal Pipelines Limited*
  - Directors Shared: *5* Directors – *Pushp Kumar Joshi, Rajan Tandon, Bharathan Shunmugavel, Suresh Kasargod Shetty, Rajneesh Narang*
- Clique with companies sharing minimum number of directors:
  - Clique: *5* Companies – *Fource Energy Consultants LLP, CREDA - HPCL Bio Fuel Limited, NCDEX e Markets Limited, ONGC Petro Additions Limited, Hindustan Petroleum Corporation Limited*
  - Directors Shared: *2* Directors – *Ramaswamy Jagannathan, Pushp Kumar Joshi*

Example 2: *Infosys Limited (CIN: L85110KA1981PLC013115)* as the *Base Node*

- **1st Degree Connections:**
  - *No. of Companies: 37*
  - *No. of Directors: 7*
- **2nd Degree Connections:**
  - *No. of Companies: 621*
  - *No. of Directors: 150*
- **3rd Degree Connections:**
  - *No. of Companies: 6863*
  - *No. of Directors: 1955*

- **Analysis of 3rd Degree Network MCCs containing *Infosys Limited***
  - Total number of MCCs: *5*
  - Average number of companies in an MCC: *8 – 9*
  - Largest MCC in size (Maximum number of Companies):
    - Clique: *10* Companies – *Stockpole Investments Pvt Ltd, Karmayogi Bharat, Infosys Limited, SVP Philanthropy Foundation, Lagoon Investments Private Limited, Ashland Plastics Private Limited, Comtrade Leasing and Investment (India) Pvt Ltd, Give Foundation, Withya HR Fund LLP, GVI Associates LLP*
    - Directors Shared: *3* Directors – *Govind Iyer, Priti Warrier, Santrupt Misra*
  - Smallest MCC in size (Minimum number of Companies):
    - Clique: *5* Companies – *Pune Software Park Private Limited, Thesys Technologies Private Limited, Infosys Limited, Capgemini India Private Limited, Capgemini Consulting India Private Limited*
    - Directors Shared: *5* Directors – *Pierre Yves Cros, Hubert Paul Henri Giraud, Aruna Jayanthi, Sreenivas Rao Baru, Salil Satish Parekh*
  - Clique with companies sharing maximum number of directors:

    - Clique: *5* Companies – *Pune Software Park Private Limited, Thesys Technologies Private Limited, Infosys Limited, Capgemini India Private Limited, Capgemini Consulting India Private Limited*
    - Directors Shared: *5* Directors – *Pierre Yves Cros, Hubert Paul Henri Giraud, Aruna Jayanthi, Sreenivas Rao Baru, Salil Satish Parekh*
  - Clique with companies sharing minimum number of directors:
    - Clique: *9* Companies – *BMR Global Services Private Limited, BMR Business Solutions Private Limited, Taxand Advisors Private Limited, Infosys Limited, Biocon Limited, Indostar Capital Finance Limited, Biocon Biologics Limited, Reliance Petroleum Limited, Avaana Capital Advisors LLP*
    - Directors Shared: *3* Directors – *Mukesh Hari Butani, Bobby Kanubhai Parikh, Kiran Mazumdar Shaw*

Example 3: *Apollo Pharmacies Limited (CIN: U52500TN2016PLC111328)* as the *Base Node*

- **1st Degree Connections:**
  - *No. of Companies: 22*
  - *No. of Directors: 4*
- **2nd Degree Connections:**
  - *No. of Companies: 161*
  - *No. of Directors: 33*
- **3rd Degree Connections:**
  - *No. of Companies: 912*
  - *No. of Directors: 210*

- **Analysis of 3rd Degree Network MCCs containing *Apollo Pharmacies Limited***
  - Total number of MCCs: *3*
  - Average number of companies in an MCC: *8 – 9*
  - Largest MCC in size (Maximum number of Companies):
    - Clique: *16* Companies – *KEI-RSOS Petroleum and Energy Private Limited, Apollo Energy Company Limited, Apollo Pharmalogistics Private Limited, Spectra Hospital Services Private Limited, Apollo Mumbai Hospital Limited, International Software India Limited, PCR Investments Limited, Kalpatharu Infrastructure Development Company Private Limited, Prime Time Logistics Technologies Private Limited, Sindya Infrastructure Development Company Private Limited, Apollo Nellore Hospital Limited, Apollo Health Resources Limited, Active Spine Care India Private Limited, Indian Hospitals Corporation Limited, Apollo Pharmacies Limited, Apollo Asha Bioelectro Private Limited*
    - Directors Shared: *7* Directors – *Mrinalini Reddy Chigullarevu, Subramanian Vridhakasi, Lodugureddygari Lakshminarayanareddy, Venkata Narasa Reddy Adapala, Puthen Veetil George Eapen, Muthukrishna Ganesan, Singana Obul Reddy*
  - Smallest MCC in size (Minimum number of Companies):

    - Clique: *3* Companies – *PCR Investments Limited, Apollo Pharmacies Limited, Apollo Medicals Private Limited*
    - Directors Shared: *4* Directors – *Mrinalini Reddy Chigullarevu, Venkata Chalam Durvasula, Shobana Kamineni, Singana Obul Reddy*
  - Clique with companies sharing maximum number of directors:
    - Clique: *16* Companies – *KEI-RSOS Petroleum and Energy Private Limited, Apollo Energy Company Limited, Apollo Pharmalogistics Private Limited, Spectra Hospital Services Private Limited, Apollo Mumbai Hospital Limited, International Software India Limited, PCR Investments Limited, Kalpatharu Infrastructure Development Company Private Limited, Prime Time Logistics Technologies Private Limited, Sindya Infrastructure Development Company Private Limited, Apollo Nellore Hospital Limited, Apollo Health Resources Limited, Active Spine Care India Private Limited, Indian Hospitals Corporation Limited, Apollo Pharmacies Limited, Apollo Asha Bioelectro Private Limited*
    - Directors Shared: *7* Directors – *Mrinalini Reddy Chigullarevu, Subramanian Vridhakasi, Lodugureddygari Lakshminarayanareddy, Venkata Narasa Reddy Adapala, Puthen Veetil George Eapen, Muthukrishna Ganesan, Singana Obul Reddy*
  - Clique with companies sharing minimum number of directors:
    - Clique: *7* Companies – *NSL Nagapatnam Power and Infratech Limited, Amaravati Thermal Power Private Limited, Capital Fortunes Private Limited, Apollo Medicals Private Limited, Constellation Sports & Entertainment Partners LLP, Apollo Pharmacies Limited, Capital Fortunes Ventures Private Limited*
    - Directors Shared: *1* Directors – *Venkata Chalam Durvasula*

Example 4: *SBI Foundation (CIN: U85100MH2015NPL266051)* as the *Base Node*

- **1st Degree Connections:**
  - *No. of Companies: 19*
  - *No. of Directors: 7*
- **2nd Degree Connections:**
  - *No. of Companies: 392*
  - *No. of Directors: 125*
- **3rd Degree Connections:**
  - *No. of Companies: 5044*
  - *No. of Directors: 1527*
- **Analysis of 3rd Degree Network MCCs containing *SBI Foundation***
  - Total number of MCCs: *5*
  - Average number of companies in an MCC: *5 – 6*
  - Largest MCC in size (Maximum number of Companies):
    - Clique: *9* Companies – *State Bank Operations Support Services Private Limited, SBI Foundation, Indian Institute of Banking and Finance, SBICAP Ventures Limited, SBI*

Capital Markets Limited, SBI Funds Management Limited, SBI Cards and Payment Services Limited, SBI General Insurance Company Limited, SBI Life Insurance Company Limited

    * Directors Shared: *6* Directors – *Shamsher Singh, Dinesh Khara, Om Prakash Mishra, Tejendra Mohan Bhasin, Rajay Kumar Sinha, Alok Kumar Choudhary*

  - Smallest MCC in size (Minimum number of Companies):
    * Clique: *3* Companies – *Association of Mutual Funds in India, SBI Foundation, SBI Funds Management Limited*
    * Directors Shared: *2* Directors – *Shamsher Singh, Dinesh Khara*

  - Clique with companies sharing maximum number of directors:
    * Clique: *9* Companies – *State Bank Operations Support Services Private Limited, SBI Foundation, Indian Institute of Banking and Finance, SBICAP Ventures Limited, SBI Capital Markets Limited, SBI Funds Management Limited, SBI Cards and Payment Services Limited, SBI General Insurance Company Limited, SBI Life Insurance Company Limited*
    * Directors Shared: *6* Directors – *Shamsher Singh, Dinesh Khara, Om Prakash Mishra, Tejendra Mohan Bhasin, Rajay Kumar Sinha, Alok Kumar Choudhary*

  - Clique with companies sharing minimum number of directors:
    * Clique: *5* Companies – *Aga Khan Rural Support Programme (India), Indian Council on Global Relations, Indian Public Schools Society, National Investment and Infrastructure Fund Limited, SBI Foundation*
    * Directors Shared: *1* Directors – *Ishaat Hussain*

Example 5: *Larsen & Turbo Limited (CIN: L99999MH1946PLC004768)* as the *Base Node*

- **1st Degree Connections:**
  - *No. of Companies: 106*
  - *No. of Directors: 18*
- **2nd Degree Connections:**
  - *No. of Companies:z2625*
  - *No. of Directors: 543*
- **3rd Degree Connections:**
  - *No. of Companies:21465*
  - *No. of Directors:6331*

- **Analysis of 3rd Degree Network MCCs containing *Larsen & Turbo Limited***
  - Total number of MCCs: *36*
  - Average number of companies in an MCC: *6 – 7*
  - Largest MCC in size (Maximum number of Companies):

* Clique: *19* Companies – *Mahindra And Mahindra Limited, Shell Technology India Private Limited, Shell India Private Limited, Reliance Foundation Institution Of Education And Research, Interglobe Aviation Limited, Global Health Limited, CSEP Research Foundation, Shell Bitumen India Private Limited, Apollo Tyres Limited, Mahindra Electric Automobile Limited, Hazira Gas Private Limited, Larsen And Toubro Limited, Organogami Consultants Private Limited, TMA Estates LLP, Bharat Shell Limited, Vodafone India Limited, Jubilant FoodWorks Limited, Colgate-Palmolive (India) Limited, L&T Hydrocarbon Engineering Limited*
* Directors Shared: *12* Directors – *Ramamurthi Shankar Raman, Donald Samuel Anderson, Subramanian Sarma, Nitin Chandrashanker Shukla, Vikram Singh Mehta, Shikha Sanjaya Sharma, Shyamala Gopinath, Meleveetil Damodaran, Hari Shanker Bhartia, Pallavi Shardul Shroff, Rajesh Ganesh Jejurikar, Ashwani Windlass*

– Smallest MCC in size (Minimum number of Companies):
  * Clique: *3* Companies – *ITC Limited, ING Vysya Bank Limited, Larsen And Toubro Limited*
  * Directors Shared: *3* Directors – *Meleveetil Damodaran, Hemant Bhargava, Alka Marezban Bharucha*

– Clique with companies sharing maximum number of directors:
  * Clique: *10* Companies – *Mindtree Limited, LTI Mindtree Limited, L And T EMSYS Private Limited, L&T Realty Developers Limited, L&T Infrastructure Development Projects Limited, L&T Realty Limited, L&T Metro Rail (Hyderabad) Limited, Larsen And Toubro Limited, LandT Welfare Company Limited, L&T Employees Welfare Foundation Private Limited*
  * Directors Shared: *18* Directors – *Dip Kishore, Deepa Gopalan Wadhwa, Ramamurthi Shankar Raman, Vijayalakshmi Iyer, Hariharan Narayanswamy, Anand Narotamdas Desai, Krishnamurti Venkataramanan, Jayant Damodar Patil, Vaishali Prasad Koparkar, Bijou Kurien, Sanjeev Aga, Chandrasekaran Ramakrishnan, Ramnath Ramdittamal Mukhija, Shailendra Narain Roy, Sekharipuram Narayanan Subrahmanyam, Anilkumar Manibhai Naik, Shrikant Prabhakar Joshi, Apurva Purohit*

– Clique with companies sharing minimum number of directors:
  * Clique: *4* Companies – *ITC Limited, SMC Global Securities Limited, UGRO Capital Limited, Larsen And Toubro Limited*
  * Directors Shared: *1* Directors – *Hemant Bhargava*

Table 4.10 is a consolidated representation of the results for the above examples discussing MCCs.

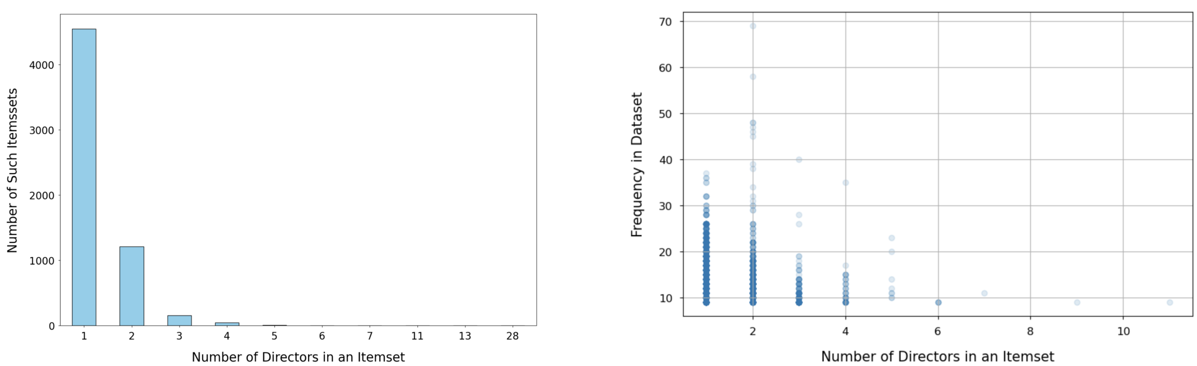
Company Name (Industry)	Col 1	Col 2	Col 3	Col 4	Col 5	Col 6
HPCL (Petroleum & Natural Gas)	720, 175	6, 5.0	7, 4	3, 4	7, 5	5, 2
Infosys Limited (Consulting & IT)	6863, 1955	5, 8.20	10, 3	5, 5	5, 5	9, 3
Apollo Pharmacies (Pharmaceutical)	912, 210	3, 8.67	16, 7	3, 4	16, 7	7, 1
SBI Foundation (Banking & Finance)	5044, 1527	5, 5.2	9, 6	3, 2	9, 6	5, 1
Larsen & Turbo Ltd. (Construction)	21465, 6331	36, 6.69	19, 12	3, 3	10, 18	4, 1

Table 4.10: Consolidated Results for *MCCs* Examples. **Col 1:** *No. of Companies in 3rd degree connections, No. of Directors in 3rd degree connections,*, **Col 2:** *No. of MCCs that include the corresponding company, Average No. of companies in such MCCs*, **Col 3:** *Size of MCC with maximum No. of companies, No. of directors shared*, **Col 4:** *Size of MCC with minimum No. of companies, No. of directors shared*, **Col 5:** *Size of MCC which share maximum no. of directors, No. of directors shared*, **Col 6:**, *Size of MCC which shares minimum no. of director, No. of directors shared*

4.8.3 Maximal Frequent Director Itemsets: *MFDIs*

We extract Maximal Frequent Director Itemsets (MFDIs) using the above pipeline and the *Company Information* file. The total number of Companies is ***87,187***. Minimum Support ***0.0001***. Hence, the minimum number of times a director item needs to occur to be considered as frequent is ***9 (0.0001 * 87,187)***. A total of ***5,972 MFDIs*** were extracted.

Figure 4.9 describes the general distribution of the extracted MFDIs. Figure 4.9(a) is a frequency graph of the Number of Directors in an itemset, and Figure 4.9(b) is a distribution graph of the Number of Items in an itemset and its Frequency in the dataset. The *frequency* of an itemset here is the number of times the itemset occurs in the dataset, and hence, it is the number of companies shared between its corresponding directors.



(a) (b)

Figure 4.9: General Data Distribution Plots for MFDIs (a) Frequency of Number of Directors in an Itemset (b) Number of Directors in Itemset vs Frequency of Itemset in Data

Table 4.11 and Table 4.12 summarise the top *5* results of MFDIs sorted based on Support and Number of Directors in Itemset respectively. In both the tables, the *"Support (Freq)"* column denotes the support value the itemset carries. The value inside parenthesis is the number of times this itemset occurs in the Company Information file and hence it is the number of distinct companies shared between the directors listed in the *"Frequent Director Itemset"* column. Attached next to the Director name in parenthesis in both the tables is the Director Identification Number or DIN to avoid duplicate name confusion. Sections 4.8.3.1 and Section 4.8.3.2 cover the in-depth analysis and observations on results from Table 4.11 and Table 4.12 respectively.

4.8.3.1 Observation and Analysis of Top 5 MFDIs Results Based on Frequency or Support

Table 4.11: Top *5 MFDIs* Based on Frequency or Support

Support (Freq)	Frequent Director Itemset
0.000791402 (69)	*2 Directors* – Shyam Sundar Patodia *(00203989)*, Manju Patodia *(00209095)*
0.000665237 (58)	*2 Directors* – Atul Chordia *(01737471)*, Ashok Dhanraj Chordia *(00569054)*
0.000550541 (48)	*2 Directors* – Kunjbihari Shah *(00622460)*, Jay Kunjbihari Shah *(08954281)*
0.000550541 (48)	*2 Directors* – Bhushan Vilaskumar Palresha *(01258918)*, Nilesh Vilaskumar Palresha *(00414963)*
0.000539071 (47)	*2 Directors* – Manoj Nawalrai Hingorani *(01238210)*, Sanjiv Chamanlal Aurora *(01238188)*

Table 4.11 – *Row 1*: Director ***Shyam Sundar Patodia*** and ***Manju Patodia*** share the same Last Name *(Patodia)* indicating to a family-relation. They share directorship across a total of *69* companies, which are listed below:

List of Companies: Mandiv Real Estate Private Limited, Dhanrashi Builders LLP, Glowing Developers LLP, Topaz Infra Developers LLP, Legend Nirman LLP, Manvijay Builders Private Limited, Hilife Towers LLP, Manvijay Vyapaar Private Limited, Manvijay Projects Private Limited, Manvijay Real Estate LLP, Manvijay Heights LLP, Manvijay Realty LLP, Manvijay Niketan LLP, Marshal Promoters LLP, Glowing Enclave Private Limited, Manvijay Vanijya Private Limited, Linton Infrastructure LLP, Manvijay Properties Private Limited, Softlink Properties Private Limited, Venus Controls & Switchgear Pvt Ltd, Deserve Projects LLP, Glowing Residency Private Limited, Glowing Real Estate Private Limited, Marshal Promoters Private Limited, Logic Infra Projects LLP, Glowing Developers Private Limited, Softlink Properties LLP, Manvijay Realty Private Limited, Glowing Residency LLP, Review Developers LLP, Mandiv Real Estate LLP, Manvijay Constructions LLP, Glowing Enclave LLP, Manvijay Enclave Private Limited, Glowing Projects Private Limited, Manvijay Buildcon LLP, Manvijay Enclave LLP, Manvijay Niketan Private Limited, Hytone Estates LLP, Mandiv Enclave LLP, Glowing Constructions Private Limited, Glowing Niketan LLP, Manvijay Constructions Private Limited, Vicarage Developers LLP, Tenement Developers LLP, Kalamunj Nirman Private Limited, Glowing Heights Private Limited, Glowing Projects LLP, Glowing Real Estate LLP, Manvijay Vyapaar LLP, Dhankuber Real Estate Private Limited, Manvijay Vanijya LLP, Kalamunj Nirman LLP, Glowing Heights LLP, Glowing Niketan Private Limited, Venus Infra Realtors LLP, Manvijay Builders LLP, Manvijay Properties LLP, Legend Nirman Private Limited, Deserve Projects Private Limited, Manvijay Residency Private Limited, Mandiv Enclave Private Limited, Samir Commerce & Finance Pvt Ltd, Manvijay Projects LLP, Manvijay Buildcon Private Limited, Manvijay Heights Private Limited, Manvijay Real Estate Private Limited, Dhankuber Real Estate LLP, Glowing Constructions LLP

Observations for the frequent director itemset and the corresponding list of shared companies:

- **Types of Businesses:**
  - *Real Estate Development:* Mandiv Real Estate Private Limited, Glowing Developers LLP, and Topaz Infra Developers LLP
  - *Construction:* Manvijay Builders Private Limited and Manvijay Constructions LLP
  - *Property Management:* Manvijay Realty LLP and Manvijay Properties Private Limited
- **Common Names & Specific Terms:**

- *Manvijay: 25* times
- *Glowing: 16* times
- *Mandiv: 4* times
- *Nirman: 4* times

- **Legal Structure:** *LLPs (Limited Liability Partnerships)* and *Private Limited Companies*: The list includes both LLPs and Private Limited companies, indicating different legal structures and ownership arrangements.
- **Same Name Different Legal Structure:** Several companies have the exact same name but different legal structures (LLP / Private Limited). This could be because of firm restructuring in order to raise more capital or add more partners. These include:
  - Manvijay Builders Private Limited – Manvijay Builders LLP
  - Legend Nirman LLP – Legend Nirman Private Limited
  - Glowing Enclave LLP – Glowing Enclave Private Limited
  - Venus Infra Realtors LLP – Venus Infra Realtors Private Limited
  - Deserve Projects LLP – Deserve Projects Private Limited

Table 4.11 – *Row 2*: Director ***Atul Chordia*** and ***Ashok Dhanraj Chordia*** also share the same Last Name *(Chordia)* indicating to a family-relation. They share directorship across a total of *58* different companies, which are listed below:

List of Companies: Incline Spaces LLP, Ashdan Township Holdings Private Limited, Classic Royal Realty Developer LLP, Pacer Real Estate Developers LLP, Aquaris Projects LLP, Ashdan Land Developers Private Limited, Ashdan Projects Private Limited, Jairaj Realty LLP, Florett Real Estate Developers Private Limited, Western City Townships LLP, Mahalunge Land Developers Private Limited, Ela Construwell LLP, Chordia Developers LLP, Ashdan Realty Private Limited, Fluxura Land Developers LLP, Tathawade Land Developer Private Limited, Jairaj Realty Unit 9 LLP, Ashdan Properties Private Limited, Classic Promoters and Builders Private Limited, Baner Land Developers LLP, Integrated Business Ecosystem Private Limited, D.C. Retail Private Limited, AC Realty Market LLP, Mahalunge Township Developers LLP, Mexus Real Estate Developers LLP, Arose Projects LLP, ADC Real Estate Developers Private Limited, Ashdan Buildcon Private Limited, Experior Land Developers LLP, ANP Construwell LLP, Chordia Spaces LLP, Arose Properties Private Limited, Highspot Landmarks LLP, Built to Live Realty Development No. 1 L, Built to Live Realty LLP, AC Realty LLP, Maan-Hinje Township Developers LLP, Manjari Housing Projects LLP, Fraction Projects LLP, Athrvi Project Management LLP, Avencore Properties LLP, Real Estate Centre Private Limited, Mahalunge Land Developers LLP, Eliture Land Developers LLP, Ashdan Advisors LLP, Lettuce Works LLP, Renozo Developers LLP, North Pune Realty LLP, ADC Realty LLP, East Pune Realty LLP, Steller Spaces LLP, Future Sector Land Developers LLP, Magnite Properties Private Limited, Aspure Land Developers LLP, AC Realty Spaces LLP, Astaria Land Developers LLP, Chordia Holdings and Construction Private Limited, Hinjewadi Land Developers LLP

Observations for the frequent director itemset and the corresponding list of shared companies:

- **Types of Businesses:**
  - *Real Estate Development, Construction & Property Management:* The company names in the list include various terms such as *Land Developers*, *Real Estate*, *Township* and *Realty* indicating the domain of business.

- **Common Names & Specific Terms:**
  - *Ashdan: 9* times
  - *Chordia: 2* times
  - *Built to Live: 2* times
  - *Developers: 14* times

- **Same Name Different Legal Structure:**
  - Ashdan Township Holdings Private Limited – Ashdan Township Holdings LLP
  - Ashdan Land Developers Private Limited – Ashdan Land Developers LLP
  - Ashdan Projects Private Limited – Ashdan Projects LLP
  - Ashdan Realty Private Limited – Ashdan Realty LLP
  - Ashdan Properties Private Limited – Ashdan Properties LLP
  - AC Realty Market LLP – AC Realty Market Private Limited

- **Other Observations:**
  - *Geographic Focus*: Several company names include specific locations, such as *"North Pune Realty LLP"* and *"Hinjewadi Land Developers LLP"*, indicating a focus on developments in those areas.
  - *Partnership and Collaboration:* Several companies include terms like *"Projects LLP"*, suggesting collaborative efforts or joint ventures in real estate projects.

Table 4.11 – *Row 3*: Directors ***Kunjbihari Shah*** and ***Jay Kunjbihari Shah*** also share the same Last Name *(Shah)* indicating to a family-relation. They share directorship across a total of *48* different companies, which are listed below:

List of Companies: Harivansh Solar Projects LLP, Lambodar Solar Projects LLP, Gadadhar Solar Projects LLP, Girdhari Solar Projects LLP, Parvati Solar Projects LLP, Maatangi Solar Projects LLP, Dhwarkesh Solar Projects LLP, Umaputra Solar Projects LLP, Krishnapriya Solar Projects LLP, Zenith Power Projects Private Limited, Parthivi Solar Projects LLP, Vaarahi Solar Projects LLP, Govind Solar Projects LLP, Sahishnu Solar Projects LLP, Kshipra Solar Projects LLP, Devvrata Solar Projects LLP, Narayani Solar Projects LLP, Subhangi Solar Projects LLP, Srimati Solar Projects LLP, Radhamadhav Solar Projects LLP, Vilasini Solar Projects LLP, Vinali Solar Projects LLP, Bhalchandra Solar Projects LLP, Uttama Solar Projects LLP, Sumitra Solar Projects LLP, Svayambhoo Solar Projects LLP, Nahush Solar Projects LLP, Jahu Solar Projects LLP, Dadhichi Solar Projects LLP, Kripalu Solar Projects LLP, Ansuya Solar Projects LLP, Vrushabh Solar Projects LLP, Sadgati Solar Projects LLP, Dharmik Solar Projects LLP, Shamli Solar Projects LLP, Shrikant Solar Projects LLP, Vrushbhanu Solar Projects LLP, Ridhika Solar Projects LLP, Rasmandal Solar Projects LLP, Madanmohan Solar Projects LLP, Indranuj Solar Projects LLP, Vakratund Solar Projects LLP, Shivdoot Solar Projects LLP, Amala Solar Projects LLP, Harikrishna Solar Projects LLP, Janeshwar Solar Projects LLP, Dhrupad Solar Projects LLP, Satyabhama Solar Projects LLP

Observations for the frequent director itemset and the corresponding list of shared companies:

- **Types of Businesses:** All the companies in the list are involved in *Solar Projects*.

- **Common Names & Specific Terms:** The word *Solar Projects* appears *47* times
- **Legal Structre:** Except for *Zenith Power Projects Private Limited*, all other companies are LLPs.
- **Other Observations:** The first name of all these companies, such as *Harivansh, Lambodar, Girdhari, Parvati . . .*, show a common origin in Hindu mythology and ancient Indian history.

Table 4.11 – *Row 4*: Directors ***Bhushan Vilaskumar Palresha*** and ***Nilesh Vilaskumar Palresha*** also share the same Last Name *(Palresha)* indicating to a family-relation. They share directorship across a total of *48* different companies, which are listed below:

List of Companies: Aadit Infra Construwell LLP, Vtp Ventures Private Limited, Home Rising Construction LLP, Magnite Realty LLP, Nnp Buildcon LLP, Ujwal Home Rising Construction LLP, Business Ventures (Nasik) Private Limited, Ujwal Construwell LLP, Nvp Rising LLP, Arose Real Estate Developers LLP, Bvp Real Estate Developers Private Limited, Home Construwell LLP, Aap Construction LLP, Vilro Constructions LLP, Radical Spaces (India) Private Limited, Vtp Urban Nirvana LLP, Ira Erectors LLP, Maxtra Constructions Private Limited, Arhum Erectors LLP, Rising Welworth Enterprises LLP, Nivaan And Brothers Constructions LLP, Integrated Business Ecosystem Private Limited, Vtp Urban Projects (Pune) LLP, Ela Real Estate Developers LLP, Home Welworth Construction LLP, Home Enviiro Buildmate LLP, Manjari Township Private Limited, Anp Construwell LLP, Tribute Food Rising LLP, Ashdan Developers Private Limited, Vtp Urban Life Spaces LLP, Vtp Projects LLP, Vtp Corporation LLP, West Pune Properties Private Limited, Vtp Rairah Foods Private Limited, Earthwise Realty LLP, Bvp Construwell LLP, Vtp Urban Realty LLP, Vighnesh Landmark LLP, Rairah Vtp Realty LLP, Arhum Erectors Private Limited, Vtp Construction Private Limited, Magnite Developers Private Limited, Nnp Buildcon Private Limited, Mahalunge Construction LLP, Tribute Construction Rising LLP, Rairah Vtp Ventures LLP, West Pune Realty LLP

Observations for the frequent director itemset and the corresponding list of shared companies:

- **Types of Businesses:** Real Estate Development, Construction, Infrastructure, Property.
- **Common Names & Specific Terms:**
  - *VTP: 18* times
  - *Construction: 10* times
  - *Rising: 6* times
  - *Realty: 4* times
- **Other Observations:** Some companies include specific location names such as *"Pune"* and *"Nasik"* suggesting their geographic focus or area of operation.

Table 4.11 – *Row 5*: Directors ***Manoj Nawalrai Hingorani*** and ***Sanjiv Chamanlal Aurora*** share directorship across a total of *47* different companies which are listed below:

List of Companies: Global Mint Realty LLP,Indo Global Fulgaon Industries Private Limited,Fulgaon Khule Promoters & Developers LLP,Kripalu Industries Private Limited,Ranjangao Bio Projects & Infra LLP,Ajax Estates Private Limited,Hinjewadi IT Park Private Limited,Indo Global Promoters & Developers LLP,Global Sharyu Infrastructures LLP,Global Square Realty LLP,Indo Global Business Park LLP,Indo Global Infotech City Private Limited,Indo Global Erectors Private Limited,Indo Global Buildcon LLP,Unoterra Global Infrastructures Private Limited,Titania Software Park Private Limited,Yadnesh Infrastructures Private Limited,Indo Global Erectors LLP,Synergy Infotech Private Limited,Indo Global Innovative Solutions Private Limited,Indo Global Industrial Park LLP,Gajakarna Infrastructures Services

Private Limited,Indo Global Infrastructures (Ranjangao) Services Private Limited,Mind Space Shelters LLP,Indo Global Real Estates & Infra LLP,Indo Global Hinjewadi Software Park Private Limited,Global Ashoka Fulgao Industrial Park LLP,Bhairavnath Infrastructures Private Limited,Indo Global Precast Systems LLP,Bhairavnath Infrastructures LLP,Indo Global Infotech City LLP,Global Vithai Infra LLP,Benyel Engineering Private Limited,Indo Global Township Private Limited,Cayman Properties LLP,Indo Global Fulgaon Industries LLP,Uranus Softech Park Private Limited,Ajax Estates LLP,Indo Global Infotech Park Private Limited,Indo Global Infra Projects Private Limited,Lambodara Infrastructure & Development Services Private Limited,Hinjewadi Business Park Private Limited,IGM Engineers And Contractors Private Limited,Vithai Developers LLP,White Star Buildcon LLP,Asthavinayak Logistics LLP,Unoterra Global Business Parks Private Limited

Observations for the frequent director itemset and the corresponding list of shared companies:

- **Types of Businesses:** Real Estate and Infrastructure Development
- **Common Names & Specific Terms:** *Indo Global (18)*, *Hinjewadi (3)*

4.8.3.2 Observation and Analysis of Top 5 MFDIs Results Based on Number of Directors in an Itemset

Table 4.12: Top *5 MFDIs* Based on Number of Directors in Frequent Itemsets

Support (Freq)	Frequent Director Itemset
0.000103226 (9)	*16 Directors* – Suresh Baid *(00030585)*, Shibsankar Paul *(05248838)*, Sujoy Ganguli *(02164332)*, Sudhanshu Sethia *(01472043)*, Sumit Bhansali *(00361918)*, Pawan Kumar Bhansali *(01500993)*, Manik Pal *(00770574)*, Bikramjit Mallik *(05244547)*, Arvind Kumar Jha *(05244537)*, Keya Saraswati *(00344750)*, Rekha *(05249627)*, Sushil Kumar Bhansali *(00344931)*, Raj Kumar Nahata *(00028205)*, Gautam Bhansali *(05195426)*, Binod Shaw *(05244481)*, Dan Chand Surana *(01500827)*
0.000103226 (9)	*11 Directors* – Avanti Ashwini Malhotra *(05285616)*, Smriti Malhotra *(07067050)*, Ritu Mukesh Malhotra *(06897799)*, Priti Ashwini Malhotra *(02611837)*, Mukesh Satpal Malhotra *(00129504)*, Ashwini Baldevraj Malhotra *(00129609)*, Pooja Sood *(02426080)*, Saachi Ashwini Malhotra *(07131247)*, Urvashi Sahni *(01287787)*, Bhushan Chintamani Wadikar *(09752588)*, Akshay Mukesh Malhotra *(00129626)*
0.000126166 (11)	*7 Directors* – Sanjeev Srivastva *(00040401)*, Ashok Kumar Gupta *(00139581)*, Mukesh Agarwal *(00052032)*, Manoj Gaur *(00582603)*, Ashok Chaudhry *(00286070)*, Ram Kishor Arora *(00021491)*, Amit Jain *(00916016)*
0.000103226 (9)	*6 Directors* – Ravi Kumar Dugar *(01549253)*, Surendra Kumar Dugar *(00424900)*, Pradip Kumar Chopra *(00425171)*, Madhu Dugar *(07045784)*, Arun Kumar Sancheti *(00025453)*, Pratiti Chopra *(01584605)*
0.000103226 (9)	*6 Directors* – Vandana Ajay Shriram *(00195184)*, Ajay Shridhar Shriram *(00027137)*, Kavita Vikram Shriram *(00193436)*, Ajit Shridhar Shriram *(00027918)*, Richa Ajit Shriram *(00193419)*, Vikram Shridhar Shriram *(00027187)*

Table 4.12 – *Row 1*: The *16* Directors share *9* companies which are listed below:

List of Companies: Linkplan Nirman LLP, Shivmahima Residency LLP, Oversure Complex LLP, Shivphal Realtors LLP, Silverfine Commotrade LLP, Topflow Niketan LLP, Mangaldham Residency LLP, Pushapdham Properties LLP, Pawansathi Vincom LLP

Observations for the frequent director itemset and the corresponding list of shared companies:

- Last name *"Bhansali"* appears four times (Sumit Bhansali, Pawan Kumar Bhansali, Sushil Kumar Bhansali, Gautam Bhansali)
- The company names suggest involvement in real estate, with terms like "Residency", "Realtors," and "Properties".

Table 4.12 – *Row 2*: The *11* Directors share *9* companies which are listed below:

List of Companies: Antherium Farms & Properties LLP, Deodar Farms & Properties LLP, Erythrina Farms & Properties LLP, Sesame Farms & Resorts LLP, Gladiolus Farms & Properties LLP, Casuarina Farms & Properties LLP, Camomile Farms And Properties LLP, Fir Farms & Properties LLP, Alamander Farms & Resorts LLP

Observations for the frequent director itemset and the corresponding list of shared companies:

- Last name *"Malhotra"* appears eight times (Avanti Ashwini Malhotra, Smriti Malhotra, Ritu Mukesh Malhotra, Priti Ashwini Malhotra, Mukesh Satpal Malhotra, Ashwini Baldevraj Malhotra, Saachi Ashwini Malhotra, Akshay Mukesh Malhotra)
- Father – Child relation: The name *"Ashwini"* is present in Ashwini Baldevraj Malhotra and is also common among multiple directors as a middle name in this list (Avanti Ashwini Malhotra, Priti Ashwini Malhotra, Saachi Ashwini Malhotra).
- Father – Child relation: The name *"Mukesh"* is present in Mukesh Satpal Malhotra and also appears as both middle names for two directors (Ritu Mukesh Malhotra, Akshay Mukesh Malhotra).
- The phrase *"Farms & Properties"* is consistent across all company names, suggesting a common theme of agricultural and property-related business activities.
- Plant Names: Each company name begins with the name of a plant or flower, such as *"Antherium", "Deodar", "Erythrina", "Sesame", "Gladiolus", "Casuarina", "Camomile", "Fir",* and *"Alamander"*
- Some company names include *"Resorts"*, suggesting potential involvement in the hospitality industry, particularly resort development or management.

Table 4.12 – *Row 3*: The *7* Directors share *11* companies which are listed below:

List of Companies: Crossings Buildwell Private Limited, Crossings City Private Limited, Crossings Towers Private Limited, Dynamic Towers Private Limited, Crossings Home Solutions Private Limited, Crossings Realty Private Limited., Crossings Infrastructure Private Limited, Crossings Township Private Limited., Crossings Residency Private Limited., Crossings Corporate Park Private Limited, Crossings Greens Property Private Limited

Observations for the frequent director itemset and the corresponding list of shared companies:

- Each director's name appears to have been unique, suggesting a diverse range of backgrounds or possibly a lack of familial connections among them.

- Type of Business: The consistent use of terms like *"Buildwell", "Towers", "Realty", "Infrastructure", "Township", "Residency", "Corporate Park"*, and *"Greens Property"* indicates involvement in Real Estate Development/Management.
- Use of the term *"Crossings"* at the beginning of each company name suggests a unified brand.

Table 4.12 – *Row 4*: The *6* Directors share *9* companies which are listed below:

List of Companies: Minolta Agencies LLP, Zoom Vincom LLP, Neelamber Hi-Rise LLP, Platinum Vyapaar LLP, Goldmine Commercial LLP, Presidency Traders LLP, Planet Vanijya LLP, P K C & Associates LLP, Jupiter Dealers LLP

Observations for the frequent director itemset and the corresponding list of shared companies:

- The last name *"Dugar"* (Ravi Kumar Dugar, Surendra Kumar Dugar, Madhu Dugar) and *"Chopra"* (Pradip Kumar Chopra, Pratiti Chopra) appears multiple times, suggesting possible familial ties among the directors.
- *LLM Driven Director – Director relation identification:*
  - *Pratiti Chopra* and *Pradip Kumar Chopra* as *Wife – Husband*.
  - *Madhu Dugar* and *Surendra Kumar Dugar* as *Wife – Husband*.
  - *Ravi Kumar Dugar* and *Surendra Kumar Dugar* as *Nephew – Uncle (Ravi Kumar Dugar is the son of Surendra Kumar Dugar's brother Santosh Kumar Dugar)*.
- These companies span a variety of Business Types, including construction, buying, trading, brokering of goods, as well as mining.

Table 4.12 – *Row 5*: The *6* Directors share *9* companies which are listed below:

List of Companies: Orchard Plantations LLP, Belli And Lilly Farms LLP, Alfresco Estates LLP, Red Rose Infrastructure LLP, P S D Estates LLP, Grapes Farms LLP, Univenta Agricultural LLP, Realitus Trading & Farming LLP, Camp Infrastructure LLP

Observations for the frequent director itemset and the corresponding list of shared companies:

- All the directors share the same last name *"Shriram"*, and many share the same middle name *"Shridhar"* indicating a familial ties.
- The director names can be seen as pairs:
  - Vandana Ajay Shriram & Ajay Shridhar Shriram
  - Kavita Vikram Shriram & Vikram Shridhar Shriram
  - Richa Ajit Shriram & Ajit Shridhar Shriram
- *LLM Driven Director – Director relation identification:*
  - Ajay Shridhar Shriram, Vikram Shridhar Shriram and Ajit Shridhar Shriram are siblings (share Brother – Brother relation). Sons on *Shridhar* and *Prabha Shriram*.

- *Vandana Ajay Shriram* and *Ajay Shridhar Shriram* as *Wife – Husband*
- *Kavita Vikram Shriram* and *Vikram Shridhar Shriram* as *Wife – Husband*
- *Richa Ajit Shriram* and *Ajit Shridhar Shriram* as *Wife – Husband*

- The combination of agricultural (“Plantations”, “Farms”, “Agricultural”) and property-related (“Infrastructure”, “Estates”) terms in the company names suggests diversified business interests or integrated business models.
- The use of plant names (“Orchard”, “Belli”, “Lilly”, “Grapes” and “Red Rose”) in some company names may indicate a specific focus on agriculture or horticulture, possibly involving the cultivation of those plants for commercial purposes.

Table 4.13 represents consolidated statistics for Maximal Frequent Director Itemsets or MFDIs.

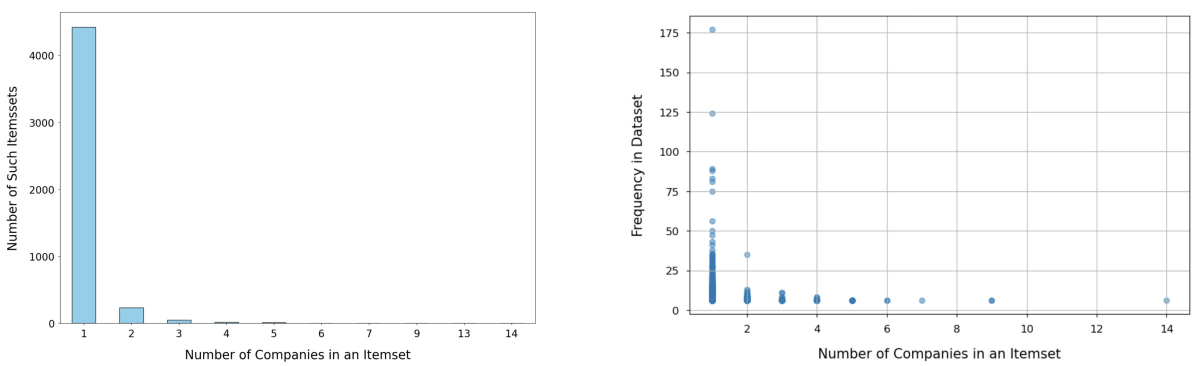
No. of Rows in Company Information File	*87,147*
Minimum Support (Frequency)	*0.0001 (9)*
Average No. of Directors in an MFDI	*1.1295*
Average Frequency OR No. of Companies Shared	*12.738* $\sim$ *13*
Size of MFDI with Maximum No. of Directors	*16*
Size of MFDI with Minimum No. of Directors	*1*
Maximum No. of Companies Shared by an MFDI	*69*
Minimum No. of Companies Shared by an MFDI	9

Table 4.13: Consolidated Statistics for *MFDIs*

4.8.4 Maximal Frequent Company Itemsets: *MFCIs*

We extract Maximal Frequent Company Itemsets (MFCIs) using the above pipeline and the *Director Information* file. The total number of Directors is ***54,216***. Minimum Support ***0.0001***. Hence, the minimum number of times a director item needs to occur to be considered as frequent is ***6 (0.0001 * 54,216)***. A total of ***4,742 MFCIs*** were extracted.

Figure 4.10 describes the general distribution of the extracted MFCIs. Figure 4.10(a) is a frequency graph of the Number of Companies in an itemset, and Figure 4.10(b) is a distribution graph of the Number of Companies in an itemset and its Frequency in the dataset. The *frequency* of an itemset here is the number of times the itemset occurs in the dataset, and hence, it is the number of directors shared between its corresponding companies.



(a) (b)

Figure 4.10: General Data Distribution Plots for MFCIs (a) Frequency of Number of Companies in an Itemset (b) Number of Companies in Itemset vs Frequency of the Itemset in Data

Table 4.14 and Table 4.15 summarise the top *5* results of MFCIs sorted based on Support and Number of Companies in an Itemset respectively. In both the tables, the *"Support (Freq)"* column denotes the support value the itemset carries. The value inside parenthesis is the number of times this company itemset occurs in the Director Information file and hence it is the number of distinct directors shared between the companies listed in the *"Frequent Companies Itemset"* column. Attached to the Company name in parenthesis in both the tables are the Company Identification Number or CIN to avoid duplicate name confusion. Sections 4.8.4.1 and Section 4.8.4.2 cover the in-depth analysis, and observations on results from Table 4.14 and Table 4.15 respectively.

4.8.4.1 Observation and Analysis of Top 5 MFCIs Results Based on Frequency or Support

Table 4.14: Top *5 MFCIs* Based on Frequency or Support

Support (Freq)	Frequent Company Itemset
0.003264719 (177)	*1* Company – Palmyrah Micro Finance *(U93000TN2012NPL086868)*
0.002711377 (147)	*1* Company – Alpha Alternatives MSAR LLP *(AAW-2814)*
0.002084255 (113)	*1* Company – Altacura AI Absolute Return Fund LLP *(AAX-3837)*
0.001807584 (98)	*1* Company – Goan Villas LLP *(AAK-8635)*
0.001752250 (95)	*1* Company – Nava Venture Advisory Services LLP *(AAD-7160)*

Table 4.14 – *Row 1*: The Company *Palmyrah Micro Finance* has a total of *(177)* Directors which are listed below:

List of Directors: Murugesapandian, Thangaiah Manickaraj, Alaguraj Aravindan, Ramasamy Parthiban, Parthiban Nandhan, Karuppaiah Soundrapandian, Kalasamy Suthanthira Pandian, Karthik Karikkolraj, Chandrasekaran Shanthi, Thilagar Issac Muthuraj, Karthick Pandi, Gopinath, Suthanthira Pandian Vasuki, Dhanakodi Maharani, Muthumanickam Kamaraj, Ganeshkumar Kowsikan, Manohar Rajesh, Subbiah Velusamy Rajan Babu, Marimuthu Mareeswaran, Chockalinga Pandian Laxmi Devi, Karthikkeyan Niranjana, Joseph Giftson Enoch, Amalraj Nagalakshmi, James Jebaraj, Rajaram Sharmila, Sankareswaran Sundarapandi, Kathiresan Suganthi, Karuppasamy Santhosh Pandiyan, Pandian Srithar, Maheswaran, Balagurusamy Ambika, Karuppasamy Arunachalapandian, Kalaiappa Nadar Meenakshi, Soundrapandian Madhumila, Ranjit Soyzza Abrue Cross Dsoyza, Mathalai Kumar Srimathi Hemalatha, Ananth Mariswari, Natrayan Ramesh, Druaisamy Nadar Arunjunai Rajan, Thamodharan Issac David, Ganesan Ashokan, Mookaiah Nadar Subramanian, Kuppusamy, Paulpandi, Karmegam Sankar, Palanichamy Selvakumar, Kodeswaran Rajesh, Rengarajan, Karthikkeyan Nehruthilagam, Karuppiahnadar Suthandradas Sugajeevitha, Balamurugan Kavitha, Soundararaj Jabez Sam Durai, Antony Raja Joseph Deva Heliton, Rajasekaran Nadar Sundar, Murugesapandian Suresh, Kamaraj Karthikeyan, Karikkolraj Vigneshwaran, Surendrakumar Amarsigapanidiyan, Marikani Chellathurai, Raju Anbalagan, Mathan, Brahaspathy, Thangapandi Balamurugan, Seenivasan Ananthi, Kodeeswaran Dilip, Chelladurai Beulah Elizabeth, Nadarajanassari Rajakumar, Pasupathy Vijaybabu, Thangaraj Chandrasekar, Rajapandian, Sankar Vani, Parthiban Malathy, Rohini Rajasekarar Nadar, Nagendran Tharani Tharan, Patchaimal Periyathambi Nadar, Jebaraj Ebenezer Samuel, Velnadar Prabhakar, Ayyappan Anandappan, Kanagasabhapathy Karthikkeyan, Thomas Issaacdaniel, Annamalaikani Jeyachandran, Ayyappan Sugindran, Muthukrishnan Maanvizhi, Ayyappannadar Dhamodharan, Mohan Buvanaathan, Janarthanan Ashok, Manoj, Pandian, Thangaswamy Kumar, Premkumar Gracy Angela, Thirumalaikani Preethi, Ashok Jaya Gowri, Palanichamy Mahendran, Mohan Kiruthiga Meenakshi, Natarajan Mukkanthan, Ramasamy Kamaraj, Subramanian Mathubala, Chinnarajanadar Palanishamy, Kodeswaran Gowri, Chellakani Rajamanickam Abinesh, Rajasekaran Chinnakalai Nadar, Bose Selvakumar, Ponniah Thangaraj, Karthikkeyan Diwakar Vairavel, Hemanth Kumar Selvakumar, Suthanthira Pandian Ram Vignesh, Raja Nancyangeline, Ganeshpandian Muthukumar, James Arul Soundaraj, Kalarani, Manickaraj Vijaya Shanmuga Prakash, Ramaswami Poongkuzhali, Gurusamy Marishkumar, Raja Manikandan, Jegadeesan Thulasiram, Murugesapandian Tamilselvi, Suthanthira Pandian Ilakkia, Veluchamy Senthikumar, Ramayanadar Nagarathinam Rajaram, Arasakumar Anbarasan, Marikani Sakthivel, Parthiban Vennila, Vivekananthan, Paulnadar Thangapandi Janakiraman, Ramasamy Chandrasekar, Yogeswari, Diraviam Rameshbabu, Mukkandan Manjuladevi, Ganesan Marikani, Velraj Karthic, Dilip Susila Devi, Mahalingam Soundarapandian, Rajiah Darathy Nallammal, Thangappan Karthika Devi, Sivalingam Vivekanandhan, Subash Chandra Bose Vijayakumar, Gurusamy Nadar Paramasivam, Mukkandan Jegannath, Kadarkaraichamy Nadar Chinnamani Nadar S, Senthikumar Shyam Sundar, Thatchinamoorthy Ananthy, Gurusamy Pandiarajan, Raja Pandian Raja Sekaran, Mohanasundaram Jeyapraba, Pandi Sankareswaran, Manohar Ravichandran, Raju Muthukumar, Tharmaraj Malaiyarasan, Rajesh Pandiselvam, Vaithilingam Gurusamy, Koilpitchai Muthiahkoilraj Joshua, Issace Jesudass, Deepalakshmi Murugesan, Asokan Sakkarai ., Abirami, Velayutham Ravichandran, Thavasiappan, Anandhappan Issac Aravind, Duraipandi, Kasi Nadar Vanarajan, Mayan Janardhanam, Anandhappan Athilakshmi, Natarajan Karunakaran Thillairajan, Mariappan Ayyadurai, Regunathan Manicka Raja, Puthuraja Kamalakannan, Guruswamy Vetrivel, Chenguttuvan Nithan Chandrasekaran, Poundass Kanda Samy, Mohan ., Jabezsam Durai Petricia, Karmegam, Thavamani, Jeyabal Arasananth, Veeranagu Ashokraj, Anandhappan Umaiswarya, Ramia Chinnasamy Ramkumar

Observations for the frequent company itemset and the corresponding list of directors controlling it:

- The company provides financial services to the *rural, semi-urban* and *urban* areas of *Tamil Nadu*.
- Managing Director: *Ayyappan Anandappan*
- Not a lot of common names were identified, but some recurring patterns can be observed, such as the repetition of *"Karthikkeyan" "Suthanthira Pandian"* and *"Mukkandan"*.
- Many of the names seem to reflect traditional Tamil naming conventions, with surnames often indicating the person's ancestral village, father's name, or caste affiliation.
- The diversity of names suggests a broad spectrum of individuals likely hailing from various regions of Tamil Nadu or having Tamil heritage.

Table 4.14 – *Row 2*: The Company *Alpha Alternatives MSAR LLP* has a total of *(135)* Directors which are listed below:

List of Directors: Madhukar Vinayak Kotwal, Sanjay Omprakash Nayar, Mala Ranjit Bhavnani, Pradeep Natverlal Kotak, Abhishek Dalmia, Chengalath Jayaram, Anand Shreevallabh Kabra, Sujata Parekh Kumar, Banwari Lal Jatia, Ushadevi Jatia, Sunil Satyapal Gulati, Nelum Pahlaj Gidwani, Suresh Kumar Poddar, Deepak Chopra, Balakrishnan Anantharaman, Prashant Gopal Kandoi, Suketu Viren Shah, Sachin Bhatia, Anil Dass, Sahir Sami Khatib, Mukesh Rajnikant Kapadia, Ramesh Ramnani, Ajay Bector, Binoy Pradip Somaia, Abhishek Amitabh Bachchan, Sadananda Ramanna Shetty, Anand Yashavant Mahajan, Rajesh Gupta, Deepakbhai Ramanbhai Patel, Ashokkumar Purushottamdas Mahansaria, Nayantara Thapar, Mukesh Maganlal Shah, Jayen Ramesh Shah, Prakash Damodar Kamat, Tushar Ramesh Shah, Atul Kumar Gupta, Anita Ramachandran, Manu Mahmud Parpia, Harshad Maganlal Patel, Piyushbhai Champaklal Shah, Sunil Srinivasan Chari, Vikram Singh, Jasbir Singh Gujral, Samir Mahendra Mehta, Chaitanya Ramesh Kejriwal, Mahabir Prasad Agarwal, Deepak Jamnadas Thakkar, Somel Deepak Thakkar, Saurabh Bahadurlal Jain, Anang Kunjviharibhai Shah, Jagtiani Harish Jethmal, Samar Sharad Chauhan, Adil Hassan, Taab Agnes Siddiqi, Satish Kumar Garg, Aditya Prakash Gupta, Ashok Vishwanath Hiremath, Rameshchandra Ramnath Dhoot, Mita Kanoria, Damini Kumarpal Desai, Perumal Srinivasan, Ketan Kothari Chandulal, Deena Dukle Waman, Manish Mohan Motwani, Vinod Shekhar, Kameswara Sarma Bulusu, Lekhaben Girishbhai Maheshwari, Vandana Lal, Yogesh Sachdeva, Pankaj Raghbeer, Kirit Sunderlal Patel, Mukesh Kumar Agarwala, Arvind Lal, Mayank Jalan, Venkatesan Narayanan Thenpathi, Hiten Pravin Shah, Gopal Krishnan, Poonam Chandra Tibrewal, Anilkumar Mithalal Jain, Neeta Bharat Khemka, Anshul Anil Goel, Rajesh Kumar Saroj, Eshwar Ramana Karra, Aditya Bhartia, Prachi Pedat Bardeshkar, Sugata Sircar, Rashmi Kant, Ravi Nanda, Abdulkadar Adamali Lokhandwala, Paulastya Sachdev, Radheshyam Purushottam Malu, Raj Kumar Jain, Rupen Mukesh Jhaveri, Chilamilika Lalini Hariani, Wilfred Desouza, Vipul Mahendra Parekh, Abhishek Gupta, Sachin Ramesh Tendulkar, Ishita Bowry, Aman Gupta, Devinjit Singh, Sanjiv Shyam Kela, Praveen Ranjan Sinha, Neena Gupta, Sameer . Jain, Harneet Singh Chandhoke, Ritesh Arora, Nakul Aggarwal, Monisha Pankaj Sharma, Sameer Ashok Mehta, Kaushal Agarwalla, Anshuman Singh, Himanshu Varshney, Karan Kumar Modi, Bimal Mukesh Shah, Sandeep Kumar Singh, Kartekeya Myadam, Abhijit Manohar Gupta, Shyam Jee Bhan, Rohit Shenai, Kiran Poddar, Kanchan Samtani, Rajat Gupta, Dinesh Kumar Godara, Shobha Mukesh

Kapadia, Yodhan Sachdev, Mudit Singhania, Shreyans Hitendrakumar Mehta, Luv Natwarlal Ukani, Krishnan Srinivasan, Anal Amit Patel, Srinivas Sadu, Mohd Farid Ahsan, Bala Vamsi Tatavarthy, Milan Sharma, Priyank Mehul Shah, Anant Kumar Daga, Adwaita Sanjay Nayar, Nishant Goyal, Shantanu Sahai, Pramod Bhandari, Nishitkumar Umeshchandra Jain, Laveena Amol Bardeshkar, Chander Prakash Gurnani, Pearlkanchan Singh Chandhoke, Shradha Anish Kariwala, Pooja Vivek Navandar

Observations for the frequent company itemset and the corresponding list of directors controlling it:

- A multi-asset class asset management firm creating investment solutions.
- *MSAR:* Multi Strategy Absolute Return
- Prominent Directors Identified:
  - Sachin Ramesh Tendulkar (former Indian Cricketer)
  - Abhishek Amitabh Bachchan (Bollywood Actor)
  - Rohit Shenai (Indian Musician)
- Common Last Names: *"Shah", "Patel", "Jain", "Gupta", "Desai"*

Table 4.14 – *Row 3*: The Company *Altacura AI Absolute Return Fund LLP* has a total of *(113)* Directors which are listed below:

List of Directors: Roopa Kudva, Sunandan Kapur, Brijesh Saxena, Balvinder Singh Kalsi, Sameer Manchanda, Anil Chawla, Atul Raheja, Sanjay Gupta, Shiv Agrawal, Vikram Lal, Shashi Bhushan Budhiraja, Raj Hajela, Supriya Kapur, Anupam Prakash, Radhika Sharma, Chandra Kumar Jain, Soundararajan Bangarusamy, Sriram Khattar, Gaurav Satyapal Jain, Navin Juneja, Hemant Narang, Satyamurti Ramasundar, Vidur Talwar, Rajeev Dass Kumar, Ashwani Talwar, Anurag Dalmia, Promila Shroff, Pradeep Jaipuria, Neelabh Dalmia, Shruti Jaipuria, Chinnaswamy Anandprasad, Madhuvanti Kanoria, Daksh Ahluwalia, Deepak Kapoor, Madhukar Bajaj, Laxmi Narayan Sain, Love Khosla Kumar, Bindu Kapur, Gautam Kapur, Vinay Dharamchand Jain, Mohit Thukral, Varun Talwar, Rajshena Kamlesh Talera, Bhaskar Pramanik, Aneeta Chopra, Mohit Bhatnagar Anand, Vikram Dhirani, Nagesh Maganlal Patel, Syed Ahmed Raja Zaidi, Vineet Khuller, Tarun Aggarwal, Sanjiv Shanmugam, Promila Kukreja, Pawan Khanna, Sharad Prem Jagtiani, Jayanti Dalmia, Benoy Roychowdhury, Deepak Talwar, Radha Ahluwalia, Arvind Walia, Kamlesh Kumar Misra, Sundararajan Srikumar, Sujatha Srikumar, Arun Nath Maira, Preety Kumar, Sanjay Bhandarkar, Vivek Madanlal Hinduja, Simran Lal, Neelam Aggarwal Singh, Priya Gupta, Rajit Mehta, Paulastya Sachdev, Varun Khosla, Nihar Joshi, Dipika Nagpal, Rajesh Razdan, Kimsuka Narasimhan, Sunil Kumar Chaturvedi, Rajiv Kuchhal, Priya Chawla, Mansha Vijay Bulchandani, Rohit Bhasin, Pooja Malhotra, Madhvi Malhotra, Jyotinder Singh Randhawa, Pallavi Mohan, Huma Hussain, Jaikumar Vinodkumar Sheth, Shekhar Dasgupta, Devashish Ohri, Navjit Ahluwalia, Prabhpreet Singh Shah, Raman Nagpal, Ambika Narang, Karanjit Bajwa Singh, Ravin Minocha, Harish Chawla, Anil Bathija, Amrita Kapur, Suresh Raina, Parul Alok Mittal, Juhi Hajela, Ravindra Bandhakavi, Aaradhana Dalmia, Rajkamal Sharma, Ashish Agrawal, Vignesh Soundararajan, Charanjit Singh Shah, Usha Rama Rao, Neeraj Aggarwal, Rajiv Mehta, Ravi Bhola, Sumita Juneja

Observations for the frequent company itemset and the corresponding list of directors controlling it:

- The company is an Artificial Intelligence-backed multi-strategy hedge fund.
- Designated Directors: *Raman Nagpal*

- Several last names appear in pairs such as Kapur *(Sunandan Kapur, Supriya Kapur)*, Talwar *(Vidur Talwar, Deepak Talwar)*, Jain *(Chandra Kumar Jain, Vinay Dharamchand Jain)*, Dalmia *(Anurag Dalmia, Neelabh Dalmia)*, Khosla *(Love Khosla Kumar, Varun Khosla)*, Malhotra *(Pooja Malhotra, Madhvi Malhotra)*, Aggarwal *(Neelam Aggarwal Singh, Neeraj Aggarwal)*, Chawla *(Priya Chawla, Harish Chawla)*, Ahluwalia *(Daksh Ahluwalia, Navjit Ahluwalia)* and Narang *(Hemant Narang, Ambika Narang)*.
- *Celebrities Identified:* Suresh Raina (former Indian Cricketer)

Table 4.14 – *Row 4*: The Company *Goan Villas LLP* has a total of *(98)* Directors which are listed below:

List of Directors: Madakshira Rao Vasudeva Shanthala, Prakash Yallapa Patil, Birdavolu Surjeet Rao, Ashwani Kumar Mehra, Srinivas Mani, Prakash Vaman Nayak, Vishwas Mehendale Vasudev, Jadav Manirao Sampath, Shriniwas Vasantrao Thakur, Ketan Kamalakar Gokhale, Girish Dattatraya Gadgil, Prabhakara Sreekantaiah, Sanjay Kothari, Parag Ramniklal Chhatbar, Ashok Narayan Apte, Dhruv Jayantilal Shah, Shekhar Vasant Malwade, Bharatkumar Keshubhai Desai, Prakash Hangarkatta Rao, Anil Kumar Gupta, Amarpal Sehmbey, Bibhash Chakravorty, Summerwell Socorro Dagama, Balamurugan Parthasarthy, Balachandran Prakash, Aparana Makhijani, Mary Thompson, Mangudi Subramanian Nagarajan, Vatsal Jyotindra Mehta, Kailash Srichand Golani, Armaity Elavia Raymond, Kultar Singh Sambi, Parmod Kumar Rishi, Ashok Vishnulal Goswami, Venkataswami Manohar Paul, Devashis Sudhanshu Roy, Nayant Maneklal Savani, Simon John Patrao, Ravi Rangachari, Venkagouda Krishnagouda Patil, Uday Kumar Gurkar, Rayomund Keki Patel, Sudhir Prasad, Atul Vipinchandra Shah, Asha Raghavan, Naga Mandava Venkateswararao, Mohan Shantigrama Alasingrachar, Alok Ramdev Kanagat, Dilip Chandulal Bania, Anand Bangalore Muthukrishna, Suresh Pattathil, Padmini, Roopashri Nayak, Sudesh Laxminarayan Honnavalli, Sriram Varadarajan, Rohit Sureshchandra Sharma, Kalmaaligaigowder Balalingiah, Rustom Bomanji Ginwalla, Siddharth Mehra, Pushkar Raj, Sangeeta Veerendra Jamdade, Settihalli Manjunath Narayanaswamy Gowda, Deepak Raghunath Parab, Amar Singh, Subrata Chatterjee, Ramaswamy Ganapathy, Ramesh Krishnamohan Pallekonda, Ravindra Kumar, Narayanaswamy Nagarajaiah, Vincent Patrick Rodrigues, Girish Deshpande, Namburi Nageswara Rao, Firoze Homi Vazifdar, Shunmugam Senthil Kumaran, Sunita Rakesh Nanda, Gopalakrishnanj Jararamank, Ajay Ugramohan Nandkeolyar, Praveen Reddy Cheruku, Vinita Joseph Dsouza, Anjali Muttoo, Lakshmi Narayan Taneja, Sriram Prasad Papani, Saurabh Saraswat, Nilesh Ramchandra Dhopade, Camysetti Devarajulu Balamukundaraj, Shiva Ganesh Jyothinagara Venkatachalama, Ashwinikumar Ashok Khandekar, Candida Melwyn Dsouza, Nilesh Shamji Gala, Richie Gupta, Sudhir Narayan Gayawal, Harshita Devadas Das, Ajay Ramkrishna Bhandarkar, Anil Sharma, Sameer Narendra Bhoir, Jaiprakash Ramanna Namdar, Amit Guha Neogi, Jinesha Labhachandra Shah

Observations for the frequent company itemset and the corresponding list of directors controlling it:

- There is not much information available about the company on the Internet.
- Designated Directors: *Balachandran Prakash* & *Girish Deshpande*
- Several other companies such as Holiday Recreations Villas, Vacation Residentials LLP, Vacation Villas LLP, Goan Apartments LLP, share the address with this company suggesting business type as hospitality specifically in Goa (India).

Table 4.14 – *Row 5*: The Company *Nava Venture Advisory Services LLP* has a total of *(95)* Directors which are listed below:

List of Directors: Kartik Ganapathy, Vandana Devi, Gurmeet Singh Chawla, Vasant Kathal, Ajay Jaysingh Anjaria, Neha Chowdhary, Pradeep Vasant Dhobale, Uma Ghurka, Jagannath Mudumbi Selvanarayan, Manish Trehan, Venkata Suresh Babu Pasupuleti, Namala Srinivasan, Narendra Harilal Dadia, Kartik Sharma, Samit Khanna, Puneet Singhania, Akshay Jain, Ruchir Dixit, Deepak Kini, Satya Kumar Dontamsetti, Jignesh Vasant Kenia, Subhendu Mandal, Adrian Luis Pinto, Pushpa Kochar, Ankur Mittal, Shipij Trivedi, Raj Rani, Chakradhr Gadde, Subramanian Narayanan, Vinit Nair, Sandeep Bansal, Sundeep Chebrolu, Sanjeev Kumar Gupta, Abhijit Balwant Selukar, Sanjiv Ray, Nishant Verman, Sunil Prabhakar Kashelkar, Dinesh Singh, Hem Chandra Gupta, Bijal Hasmukh Shah, Medha Kabra Ghurka, Mohit Mohan, Kollath Puthanveethil Rameshmenon, Payal Aggarwal, Nitin Suresh Khanna, Navanee Saxena, Sunil Bhaskar Desai, Kaustubh Ramesh Roplekar, Mattipalle Saisivaram, Chandrashekar Srinivasamurthy, Aaditya Paras Shah, Gaurav Kedia, Pollayil Sagi Sabastian, Mitesh Jitendra Shah, Amitesh Sinha, Venkatram Jandhyala, Syed Fasiullah Hussaini, Vipul Garg, Rahul Verma, Prince Shekhar Valluri, Abhishek Menon Ramachandran, Abhishek Paras Shah, Viswanath Nanjunda Rao, Popuri Venkata Nrisimha Ravi Kumar, Vinay Kumar Bansal, Mohammadabbas Naushadali Patel, Pushpa Narasimhan, Parul Goel, Varsha Agrawal, Meghna Kumat, Apoori Suman, Sathish Subraya Borker, Vijaya Sarathi Nerellapalli, Girish Sudarshan Aggarwal, Amreshwar Sati, Prasanna Aithal, Priya Gupta, Sravan Prasad Vankina, Sucharita Singh, Vipin Arora, Amit Bhardwaj, Umesh Hora, Aniket Bhalchandra Nikumb, Gurpreet Prashant Sisodia, Ashim Jolly, Ramya Srinath, Deepak Keshavlal Maniyar, Mahapatra Susmita, Mohanraj Jagannivasan, Chaitanya Atul Anjaria, Kinjal Jimish Kapadia, Pawan Kumar Dasaraju, Aditya Shrikant Prabhu, Hrushti Dipak Shah, Vivek Amarnani

Observations for the frequent company itemset and the corresponding list of directors controlling it:

- The company provides advisory services in Management, Scientific, and Technical Consulting.
- Designated Directors: *Ankur Mittal* & *Vinay Kumar Bansal*
- Last names like *"Gupta", "Shah"* and *"Singh"* occur multiple times.

4.8.4.2 Observation and Analysis of Top 5 MFCIs Results Based on Number of Companies in an Itemset

Table 4.15: Top *5 MFCIs* Based on Number of Companies in Frequent Itemsets

Support (Freq)	Frequent Company Itemset
0.000110668 (6)	*14* Companies – Cypress Farms & Resorts LLP *(AAE-0806)*, Vinca Rosea Farms And Properties LLP *(AAE-1039)*, Erythrina Farms & Properties LLP *(AAE-1014)*, Deodar Farms & Properties LLP *(AAE-1037)*, Antherium Farms & Properties LLP *(AAE-1021)*, Sesame Farms & Resorts LLP *(AAE-1020)*, Hollyhock Farms & Properties LLP *(AAE-1183)*, Shoeflower Farms And Properties LLP *(AAE-1031)*, Yucca Farms And Properties LLP *(AAE-1420)*, Gladiolus Farms & Properties LLP *(AAE-0914)*, Casuarina Farms & Properties LLP *(AAE-1417)*, Camomile Farms And Properties LLP *(AAG-0556)*, Fir Farms & Properties LLP *(AAE-0871)*, Alamander Farms & Resorts LLP *(AAE-1020)*
0.000110668 (6)	*9* Companies – Red Rose Infrastructure LLP *(AAX-2326)*, Alfresco Estates LLP *(AAW-2537)*, Orchard Plantations LLP *(AAW-2495)*, Belli And Lilly Farms LLP *(AAN-2729)*, P S D Estates LLP *(AAG-2501)*, Univenta Agricultural LLP *(AAG-8367)*, Grapes Farms LLP *(AAW-2547)*, Realitus Trading & Farming LLP *(AAQ-5410)*, Camp Infrastructure LLP *(AAO-2937)*
0.000110668 (6)	*9* Companies – Inhaled Technologies LLP *(AAF-9288)*, One World Pharma LLP *(AAN-6813)*, Nebumed Pharma LLP *(AAF-6608)*, Lancet Pharma LLP *(AAF-8769)*, Medule Pharma LLP *(AAF-7016)*, Mediorals Laboratories LLP *(AAI-5407)*, Oncocare India LLP *(AAF-6439)*, Advanced Remedies LLP *(AAF-8772)*, Pharmasolve Specialities India LLP *(AAF-9152)*
0.000110668 (6)	*7* Companies – Growx Startups LLP *(AAF-3028)*, Growx Deals LLP *(AAF-1716)*, Growx Tech Partners LLP *(AAG-0145)*, Growx Advisory LLP *(AAE-9032)*, Growx Techconcepts LLP *(AAI-0740)*, Growx Projects LLP *(AAF-9447)*, Growx Benchmark LLP *(AAJ-7600)*
0.000110668 (6)	*6* Companies – Garware Industries Limited *(U74999MH1989PLC053573)*, Envision Properties Private Limited *(U45400MH2010PTC202621)*, Via Investment Consultants Private Limited *(U45400MH2007PTC172262)*, Naigaon Chemicals Private Limited *(U24119MH2003PTC139109)*, Great Design Properties Private Limited *(U70102MH2009PTC197095)*, Best Design Properties Private Limited *(U45202MH2010PTC205261)*

Table 4.15 – *Row 1*: The *14* Companies listed in this row, share a total of *6* Directors which are listed below:

List of Directors: Ashwini Baldevraj Malhotra, Mukesh Satpal Malhotra, Smriti Malhotra, Ritu Mukesh Malhotra, Akshay Mukesh Malhotra, Urvashi Sahni

Observations for the frequent company itemset and the corresponding list of directors controlling it:

- All the companies are currently *Active*. Previously, Pvt. Ltd. converted to LLPs.
- Common Last Name: *Malhotra* (Shwini Baldevraj Malhotra, Mukesh Satpal Malhotra, Smriti Malhotra, Ritu Mukesh Malhotra, Akshay Mukesh Malhotra)

- Father – Child relation: The name *"Mukesh"* is present in Mukesh Satpal Malhotra and also appears as both middle names for two directors (Ritu Mukesh Malhotra, Akshay Mukesh Malhotra).
- The phrase *"Farms & Properties"* is consistent across all company names, suggesting a common theme of agricultural and property-related business activities.
- Plant Names: Each company name begins with the name of a plant or flower, such as *"Cypress", "Vinca Rosea" "Antherium", "Deodar", "Shoeflower"*
- Some company names include *"Resorts"*, suggesting potential involvement in the hospitality industry, particularly resort development or management.

Table 4.15 – *Row 2*: The *9* Companies listed in this row, share a total of *6* Directors which are listed below. This row is the same as *Table 4.12 – Row 5*.

Table 4.15 – *Row 3*: The *9* Companies listed in this row, share a total of *6* Directors which are listed below:

List of Directors: Shonaal Amar Lulla, Geeta Amar Lulla, Shankar Srinivasan, Srinivas Hariharan, Shirin Hamied, Rumana Hamied

Observations for the frequent company itemset and the corresponding list of directors controlling it:

- Company Status:
  - Active: *Advanced Remedies LLP*
  - Defunct: *Mediorals Laboratories LLP*
  - Striken Off: *Inhaled Technologies LLP, One World Pharma LLP, Nebumed Pharma LLP, Lancet Pharma LLP, Medule Pharma LLP, Oncocare India LLP, Pharmasolve Specialities India LLP*.
- Common Last Name: *Lulla* (Shonaal Amar Lulla, Geeta Amar Lulla), *Hamied* (Shirin Hamied, Rumana Hamied)
- These companies appear to be involved in the pharmaceutical industry, as indicated by terms like *"Pharma", "Pharmaceuticals", "Laboratories" and 'Specialties"* in their names.

Table 4.15 – *Row 4*: The *7* Companies listed in this row, share a total of *6* Directors which are listed below:

List of Directors: Sheetal Bahl, Ashish Taneja, Dharak Navin Dedhia, Mehul Praful Jobanputra, Manu Rikhye, Jimish Jobanputra Naresh

Observations for the frequent company itemset and the corresponding list of directors controlling it:

- Common Last Name: *Jobanputra* (Mehul Praful Jobanputra, Jimish Jobanputra Naresh)

- *"Growx"* term appears at the beginning of every company name, suggesting a common branding.
- Terms like *"Startups", "Deals", "Tech Partners", "Advisory", "Techconcepts", "Projects"*, and *"Benchmark"* suggest involvement in business-related services such as startup support, deals brokerage, technology partnerships, advisory services, and project management.

Table 4.15 – *Row 5*: The *6* Companies listed in this row, share a total of *6* Directors.
List of Directors: Tushar Madhuvandas Parikh, Monika Garware, Sheela Garware Shashikant, Sonia Garware, Sarita Garware Ramsay, Shashikant Bhalchandra Garware

Observations for the frequent company itemset and the corresponding list of directors controlling it:

- Unique name for each company suggests diversified business types.
  - Companies like *"Garware Industries Limited"* and *"Naigaon Chemicals Private Limited"* suggest involvement in manufacturing or industrial activities.
  - Companies like *"Envision Properties Private Limited"*, *"Great Design Properties Private Limited"*, and *"Best Design Properties Private Limited"* indicate involvement in real estate and property-related activities.
  - Companies like *"Via Investment Consultants Private Limited"* suggest involvement in investment and consulting services.
- Common Last Name: *Garware* (Monika Garware, Sheela Garware Shashikant, Sonia Garware, Sarita Garware Ramsay, Shashikant Bhalchandra Garware)

Table 4.16 represents consolidated statistics for Maximal Frequent Company Itemsets or MFCIs.

No. of Rows in Directors Information File	*54,216*
Minimum Support (Frequency)	*5.4216* $\sim$ *(6)*
Average No. of Companies in an MFCI	*1.1077*
Average Frequency OR No. of Companies Shared	*8.383* $\sim$ *9*
Size of MFCI with Maximum No. of Companies	*14*
Size of MFCI with Minimum No. of Companies	*1*
Maximum No. of Directors Shared by an MFCI	*177*
Minimum No. of Directors Shared by an MFCI	6

Table 4.16: Consolidated Statistics for *MFCIs*

Chapter 5

Conclusion, Limitations & Future Work

This thesis presents solution to two critical issues: *(1) Extraction of Web Profiles for Individuals Sharing Identical Names and (2) Identifying Weakly and Strongly Connected Entities in Interlocking Directorships.*

In the digital landscape, the challenge of distinguishing between individuals who share the same name is an important issue. Users seeking specific persons often confront a large number of web links that further require extensive processing and comparison to locate the desired profile. Our approach addresses this problem by providing users with an *integrated, consolidated view* of multiple web profiles associated with the same name, facilitating swift and accurate identification. Our solution leverages *LLMs* for both *extraction of factual information* and *consolidation of profiles* along with other tools such as *Wikipedia* and streamlining the selection of the correct profile.

To ensure the robustness of the pipeline, we test it with over *500* distinct common and uncommon names and present corresponding results. We also perform a *fifty* participant user study to evaluate the practical usability of our system. What makes our work novel is the *strategic use of LLMs* only for tasks where LLMs outperform current SOTA techniques (Wang et al., 2023) for our use case. Various prompts were carefully engineered to ensure *output consistency & hallucinations* – a significant challenge with LLMs (Wolf et al., 2023).

Corporate Networks are made of two major entities: *Directors* and *Companies*. An interlock can occur with one director sitting on board of *2* or more companies, but it does not make all such individuals a point of interest. It is not enough to just identify an interlock. Like any other social network, among corporate networks, certain groups of entities occur *together* and *frequently*, hence implying connections between them. Similarly, it is vital in corporate networks to identify the director names or company names that appear together often and determine the reason behind these links to understand the hidden dynamics behind major corporations and markets.

We present a method that uses data analytics techniques such as *maximal graph cliques* and *frequent itemsets* to identify groups of directors and companies connected weakly and strongly. A data corpus of over *85,000* companies and *55,000* directors in the *Indian Corporate Network* was prepared using a graph traversal-based data scrapping approach. Our findings reveal that around *58.6%* are directors for *2* or more companies, implying a large number of interlocking directorships. We perform various analyses on this data, presenting several results. The proposed methodology not only extracts groups of companies and directors that appear together frequently, implying strong connections between them, but it also attempts to identify the reason behind a particular connection. Using an *adapted version of our LLM Driven Web Profile Extraction* pipeline, we try to extract *personal (familial ties)* and *professional (shared work experience etc)* links between connected directors for a more informed analysis. We present several examples of LLM Driven Director–Director Relation Identification in our results.

A *Last Name analysis* on the list of approximately *55,000* distinct Directors to identify the most recurring last names showed dominance by certain groups in *Indian corporate networks*. Top *10* results for the same are: *Shah (1250), Patel (809), Singh (809), Jain (640), Gupta (634), Mehta (580), Agarwal (533), Kumar (493), Sharma (468), Rao (252)*. We also found that nearly *37,123* companies of total *87,187* – almost *42.5%*, have at least one pair of directors that share the same last name. One can say that Indian corporations are primarily *family concentered* as compared to other nations.

5.1 Limitations

While our LLM Driven Web Profile Extraction pipeline is intricately designed, *human oversight* remains indispensable. Despite entity extraction facilitated by LLMs and data standardisation through Wikipedia, certain instances emerge where both Prefix Tree disambiguation and Disambiguation with GPT Prompting yielded ambiguous outcomes. Although rare, such cases underscore the inherent limitations wherein sophisticated systems may not surpass human intellect and comprehension. Furthermore, the pipeline to extract web profiles of individuals corresponding to a given name is hindered by *time* and *monetary constraints*. Extracting text from each web link necessitates data scraping, consuming valuable time. Subsequently, querying LLMs and receiving JSON-structured profiles via the OpenAI APIs is both time and financially expensive. Hence, as the volume of links escalates, so do these expenses.

Source of information plays an important role in corporate structures. Data scraping is the sole viable solution unless one gets the government's data. No source of information is flawless. We need to trust the source to handle basic elements such as spelling of company names and director names or ensuring accurate hyperlink connections. One wrong link can alter the entire network structure since we are using a BFS graph traversal technique. The data scraping process is also very *time-consuming*. Corporate interlocking networks can be extremely dense, and visualising such networks can be difficult. Extensive *data cleaning* and pre-processing are required, often resulting in some *data loss* during the process. Also, the director – director relation identification pipeline depends on information available on the web, hence it is not always possible to extract such information even with such a sophisticated pipeline.

5.2 Future Work

Our proposed solution to extract web profiles of individuals sharing the same names offers the potential to evolve into a paid web service for various use cases, enabling individuals and authorities to extract information about a specific person or monitor the possible misuse of their own profiles across the web. This research not only advances our understanding of person name disambiguation but also presents a practical solution with broader implications for information retrieval and online identity management.

As of May 2024, a total of *27,07,587* Companies are registered with the *Ministry of Corporate Affairs, India* whereas the total number of directors registered is not known. As time passes by, this number will only increase. We scraped about *85,000* companies, which is only *3.1%* of the total. One can perform the proposed analysis on a more extensive database to extract larger and stronger groups of connected entities for even more interesting observations on connected companies and directors. Also, we only used the connections between present directors and companies. The study can be extended to connections, including past directors of companies. All the analyses can be pre-processed, and this data on connected components can be made public so people with more knowledge in the domain can make valuable inferences.

Bibliography

Adams, R. B. (2017). Boards, and the directors who sit on them. In *The handbook of the economics of corporate governance*, volume 1, pages 291–382. Elsevier.

Agrawal, R., Srikant, R., et al. (1994). Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499. Santiago.

Akbik, A., Blythe, D., and Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. In *Proceedings of the 27th international conference on computational linguistics*, pages 1638–1649.

Artiles, J., Amigó, E., and Gonzalo, J. (2009). The role of named entities in web people search. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pages 534–542.

Artiles, J., Borthwick, A., Gonzalo, J., Sekine, S., and Amigó, E. (2010). Weps-3 evaluation campaign: Overview of the web people search clustering and attribute extraction tasks. In *CLEF (Notebook Papers/LABs/Workshops)*.

Artiles, J., Gonzalo, J., and Sekine, S. (2007). The semeval-2007 weps evaluation: Establishing a benchmark for the web people search task. In *Proceedings of the fourth international workshop on semantic evaluations (semeval-2007)*, pages 64–69.

Ashok, D. and Lipton, Z. C. (2023). Promptner: Prompting for named entity recognition. *arXiv preprint arXiv:2305.15444*.

Baeza-Yates, R., Ribeiro-Neto, B., et al. (1999). *Modern information retrieval*, volume 463. ACM press New York.

Bagga, A. and Baldwin, B. (1998). Entity-based cross-document coreferencing using the vector space model. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Baktash, J. A. and Dawodi, M. (2023). Gpt-4: A review on advancements and opportunities in natural language processing. *arXiv preprint arXiv:2305.03195*.

Bollegala, D., Honma, T., Matsuo, Y., and Ishizuka, M. (2008). Automatically extracting personal name aliases from the web. In *International Conference on Natural Language Processing*, pages 77–88. Springer.

Braverman, M., Chen, X., Kakade, S., Narasimhan, K., Zhang, C., and Zhang, Y. (2020). Calibration, entropy rates, and memory in language models. In *International Conference on Machine Learning*, pages 1089–1099. PMLR.

Bron, C. and Kerbosch, J. (1973). Algorithm 457: finding all cliques of an undirected graph. *Communications of the ACM*, 16(9):575–577.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Caiazza, R., Cannella Jr, A. A., Phan, P. H., and Simoni, M. (2019). An institutional contingency perspective of interlocking directorates. *International Journal of Management Reviews*, 21(3):277–293.

Cazals, F. and Karande, C. (2008). A note on the problem of reporting maximal cliques. *Theoretical computer science*, 407(1-3):564–568.

Chandrashekar, S. and Muralidharan, K. (2012). Networks of power and influence: Board interlocks in india 1995-2007–an empirical investigation. *IIM Bangalore Research Paper*, (374).

Chen, J., Lu, Y., Lin, H., Lou, J., Jia, W., Dai, D., Wu, H., Cao, B., Han, X., and Sun, L. (2023). Learning in-context learning for named entity recognition. *arXiv preprint arXiv:2305.11038*.

Chen, M., Luo, X., Zhu, Y., Li, Y., Zhao, W., and Wu, J. (2020). An apriori-based learning scheme towards intelligent mining of association rules for geological big data.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. (2023). Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Delgado, A. D., Martínez, R., Fresno, V., and Montalvo, S. (2014). A data driven approach for person name disambiguation in web search results. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 301–310.

Delgado, A. D., Montalvo, S., Unanue, R. M., and Fresno, V. (2018). A survey of person name disambiguation on the web. *IEEE access*, 6:59496–59514.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dooley, P. C. (1969). The interlocking directorate. *The American Economic Review*, 59(3):314–323.

Fleischman, M. and Hovy, E. (2004). Multi-document person name resolution. In *Proceedings of the Conference on Reference Resolution and Its Applications*, pages 1–8.

Freitag, M., Vilar, D., Grangier, D., Cherry, C., and Foster, G. (2022). A natural diet: Towards improving naturalness of machine translation output. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3340–3353, Dublin, Ireland. Association for Computational Linguistics.

Goyal, A., Gupta, V., and Kumar, M. (2018). Recent named entity recognition and classification techniques: a systematic review. *Computer Science Review*, 29:21–43.

Grahne, G. and Zhu, J. (2003). High performance mining of maximal frequent itemsets. In *6th International workshop on high performance data mining*, volume 16, page 34.

Guha, R. (2004). Disambiguating people in search. In *The Thirteenth International World Wide Web Conference, WWW2004*.

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S. R., and Smith, N. A. (2018). Annotation artifacts in natural language inference data. *arXiv preprint arXiv:1803.02324*.

Guu, K., Lee, K., Tung, Z., Pasupat, P., and Chang, M. (2020). Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Han, H., Zha, H., and Giles, C. L. (2005). Name disambiguation in author citations using a k-way spectral clustering method. In *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries*, pages 334–343.

Han, J., Pei, J., and Yin, Y. (2000). Mining frequent patterns without candidate generation. *ACM sigmod record*, 29(2):1–12.

Hernández-Lara, A. B. and Gonzales-Bustos, J. P. (2019). The impact of interlocking directorates on innovation: the effects of business and social ties. *Management Decision*, 57(10):2799–2815.

Hillman, A. J. and Hitt, M. A. (1999). Corporate political strategy formulation: A model of approach, participation, and strategy decisions. *Academy of management review*, 24(4):825–842.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A., Welbl, J., Clark, A., et al. (2022). Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Holburn, G. L. and Vanden Bergh, R. G. (2008). Making friends in hostile environments: Political strategy in regulated industries. *Academy of Management Review*, 33(2):521–540.

Hu, L. (2022). Research on english achievement analysis based on improved carma algorithm. *Computational Intelligence and Neuroscience*, 2022:1–11.

Hua, X., Zhong, M., Liu, Q., and Wang, M. (2020). List all maximal cliques of an undirected graph: A parallable algorithm. In *IOP Conference Series: Materials Science and Engineering*, volume 790, page 012076. IOP Publishing.

Ikeda, M., Ono, S., Sato, I., Yoshida, M., and Nakagawa, H. (2009). Person name disambiguation on the web by two-stage clustering. In *2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference*.

James Chen (2022). Interlocking directorates: What it means, how it works.

Jiang, Z., Araki, J., Ding, H., and Neubig, G. (2021). How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

Kalashnikov, D. V., Mehrotra, S., Chen, Z., Nuray-Turan, R., and Ashish, N. (2006). Disambiguation algorithm for people search on the web. In *2007 IEEE 23rd International Conference on Data Engineering*, pages 1258–1260. IEEE.

Lamb, N. H. and Roundy, P. (2016). The "ties that bind" board interlocks research: A systematic review. *Management Research Review*, 39(11):1516–1542.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. (2016). Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

Lan, M., Zhang, Y. Z., Lu, Y., Su, J., and Tan, C. L. (2009). Which who are they? people attribute extraction and disambiguation in web search results. In *2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference*.

Li, J., Sun, A., Han, J., and Li, C. (2020). A survey on deep learning for named entity recognition. *IEEE transactions on knowledge and data engineering*, 34(1):50–70.

Loderer, C. and Peyer, U. (2002). Board overlap, seat accumulation and share prices. *European Financial Management*, 8(2):165–192.

Lu, Y., Bartolo, M., Moore, A., Riedel, S., and Stenetorp, P. (2021). Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *arXiv preprint arXiv:2104.08786*.

Mann, G. and Yarowsky, D. (2003). Unsupervised personal name disambiguation. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*, pages 33–40.

Matsuo, Y., Mori, J., Hamasaki, M., Ishida, K., Nishimura, T., Takeda, H., Hasida, K., and Ishizuka, M. (2006). Polyphonet: an advanced social network extraction system from the web. In *Proceedings of the 15th international conference on World Wide Web*, pages 397–406.

Mazzola, E., Perrone, G., and Kamuriwo, D. S. (2016). The interaction between inter-firm and interlocking directorate networks on firm's new product development outcomes. *Journal of Business Research*, 69(2):672–682.

McCallum, A. and Wellner, B. (2003). Toward conditional models of identity uncertainty with application to proper noun coreference.

Mika, P. (2004). Bootstrapping the foaf-web: an experiment in social network mining. In *Proc. 1st Workshop Friend of a Friend, Social Networking and the Semantic Web*.

Mizruchi, M. S. (1996). What do interlocks do? an analysis, critique, and assessment of research on interlocking directorates. *Annual review of sociology*, 22(1):271–298.

Moon, J. W. and Moser, L. (1965). On cliques in graphs. *Israel journal of Mathematics*, 3:23–28.

Ozmel, U., Reuer, J. J., and Gulati, R. (2013). Signals across multiple networks: How venture capital and alliance networks affect interorganizational collaboration. *Academy of management Journal*, 56(3):852–866.

Palmer, D., Friedland, R., and Singh, J. V. (1986). The ties that bind: Organizational and class bases of stability in a corporate interlock network. *American Sociological Review*, pages 781–796.

Pang, J.-C., Wang, P., Li, K., Chen, X.-H., Xu, J., Zhang, Z., and Yu, Y. (2023). Language model self-improvement by reinforcement learning contemplation. *arXiv preprint arXiv:2305.14483*.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Perez, E., Kiela, D., and Cho, K. (2021). True few-shot learning with language models. *Advances in neural information processing systems*, 34:11054–11070.

Peters, M. E., Neumann, M., Zettlemoyer, L., and Yih, W.-t. (2018). Dissecting contextual word embeddings: Architecture and representation. *arXiv preprint arXiv:1808.08949*.

Pfeffer, J. (2019). Size and composition of corporate boards of directors: The organization and its environment. In *Corporate Governance*, pages 53–64. Gower.

Pfeffer, J. and Salancik, G. R. (1978). *The external control*. New York.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. (2020). Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Rae, J. W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., Aslanides, J., Henderson, S., Ring, R., Young, S., et al. (2021). Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Roberts, A., Raffel, C., and Shazeer, N. (2020). How much knowledge can you pack into the parameters of a language model? *arXiv preprint arXiv:2002.08910*.

Robinson, J., Rytting, C. M., and Wingate, D. (2022). Leveraging large language models for multiple choice question answering. *arXiv preprint arXiv:2210.12353*.

Rubin, O., Herzig, J., and Berant, J. (2021). Learning to retrieve prompts for in-context learning. *arXiv preprint arXiv:2112.08633*.

Sapinski, J. P. and Carroll, W. K. (2018). Interlocking directorates and corporate networks. In *Handbook of the international political economy of the corporation*, pages 45–60. Edward Elgar Publishing.

Slaughter, S., Thomas, S. L., Johnson, D. R., and Barringer, S. N. (2014). Institutional conflict of interest: The role of interlocking directorates in the scientific relationships between universities and the corporate sector. *The Journal of Higher Education*, 85(1):1–35.

Smith, S., Patwary, M., Norick, B., LeGresley, P., Rajbhandari, S., Casper, J., Liu, Z., Prabhumoye, S., Zerveas, G., Korthikanti, V., et al. (2022). Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. *arXiv preprint arXiv:2201.11990*.

Spink, A., Jansen, B. J., and Pedersen, J. (2004). Searching for people on web search engines. *Journal of Documentation*, 60(3):266–278.

Stearns, L. B. and Mizruchi, M. S. (1986). Broken-tie reconstitution and the functions of interorganizational interlocks: A reexamination. *Administrative science quarterly*, pages 522–538.

Subramanian, S., King, D., Downey, D., and Feldman, S. (2021). S2and: A benchmark and evaluation system for author name disambiguation. In *2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 170–179. IEEE.

Tjong Kim Sang, E. F. and De Meulder, F. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Tomita, E., Tanaka, A., and Takahashi, H. (2004). The worst-case time complexity for generating all maximal cliques. In *International Computing and Combinatorics Conference*, pages 161–170. Springer.

Tu, J., Fan, J., Tang, N., Wang, P., Chai, C., Li, G., Fan, R., and Du, X. (2022). Domain adaptation for deep entity resolution. In *Proceedings of the 2022 International Conference on Management of Data*, pages 443–457.

Vu, Q. M., Masada, T., Takasu, A., and Adachi, J. (2007). Disambiguation of people in web search using a knowledge base. In *2007 IEEE International Conference on Research, Innovation and Vision for the Future*, pages 185–191.

Wang, S., Sun, X., Li, X., Ouyang, R., Wu, F., Zhang, T., Li, J., and Wang, G. (2023). Gpt-ner: Named entity recognition via large language models. *arXiv preprint arXiv:2304.10428*.

Wikipedia (2023). Barack obama — Wikipedia, the free encyclopedia. [Online; accessed 13-October-2023].

Wolf, Y., Wies, N., Levine, Y., and Shashua, A. (2023). Fundamental limitations of alignment in large language models. *arXiv preprint arXiv:2304.11082*.

Zhao, Z., Wallace, E., Feng, S., Klein, D., and Singh, S. (2021). Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pages 12697–12706. PMLR.