A Work Project presented as part of the requirements for the Award of a Master's degree in Business Analytics from the Nova School of Business and Economics.

Graph Neural Networks in E-commerce: Unveiling Competitive Dynamics through Search Engine's Keywords Analysis

Gabriel Eugenio Abib (55646)

Work project carried out under the supervision of:

Qiwei Han, Alessandro Gambetti, Maximilian Kaiser

20/02/2024

Abstract:

This thesis presents a novel framework for e-commerce competitive analysis using Graph Neural Networks (GNNs), with a focus on Search Engine Optimization (SEO). Through advanced machine learning techniques, including dimensionality reduction and clustering, it addresses the challenge of identifying competitors within the dynamic SEO landscape. The study pioneers the use of a custom GraphSAGE model and proposes a Competitor Retrieval Algorithm alongside a Keyword Recommendation System. These tools not only refine e-commerce strategy but also unveil intricate market structures and relationships. While acknowledging data limitations and the absence of true labels, the research underscores the potential of GNNs in e-commerce, offering a methodological base for future exploration and continuous innovation in machine learning and marketing analytics.

Keywords:
Natural Language Processing, Network Analysis, Competitive Landscape Analysis, Knowledge Graph, FastText, E-Commerce, Bipartite Networks, Search Engine Optimization, Word2Vec

This work used infrastructure and resources funded by Fundação para a Ciência e a Tecnologia (UID/ECO/00124/2013, UID/ECO/00124/2019 and Social Sciences DataLab, Project 22209), POR Lisboa (LISBOA-01-0145-FEDER-007722 and Social Sciences DataLab, Project 22209) and POR Norte (Social Sciences DataLab, Project 22209)

1. Introduction: Competitor identification in Search Engine Optimization

In today's e-commerce landscape, for businesses to thrive it is necessary to ensure that their online presence is substantial, and their visibility is enhanced. The budgetary allocation to online advertisement is a significant investment from a company perspective (Erdmann, Arilla and Ponzoa 2022). This is reflected by the resilience of the Digital Advertising Industry in the past few years, whose revenue trends reflect the increasing investment in internet advertising by companies (Appendix I) (PwC 2023). To maximize the effectiveness of their investment, companies must ensure that their spending on digital advertising yields the desired brand visibility. Simultaneously, it is crucial to focus on minimizing these expenses, emphasizing the importance of cost-effective strategies. Search Engine Optimization (SEO) is the unpaid practice of enhancing a website's content and structure to increase its visibility in search engine results (Ologunebi and Obafemi Taiwo 2023). Chaffey and Ellis-Chadwick (2019) discovered that 63% of marketers perceive SEO generated organic website traffic effectively, emphasizing on the role of SEO in brand visibility. In the SEO Starter Guide, Google provides suggestions and requirements for E-commerce players to have a positive influence in their websites position in Search Engine Results Page (SERP) through SEO (Google for Developers 2023). Aside from the necessary requirements, these recommendations stress the creation of high-quality and people-centric content (Google for Developers (b) 2023). Google for Developers (b) (2023) encourages using relevant keywords that potential customers may use to search for content. Strategically placing these keywords in the title, main header, alt text, and link text of the web page. This strategy ensures that the content is not only user-focused but also optimized for search engines, thereby improving its visibility and accessibility.

Although this approach is crucial for optimal positioning in SERP, it does not ensure a positive desired visibility. This is caused by the complexity of the competitive landscape within the context. Given these circumstances, this study will focus on competitor identification within the landscape of SEO. Central to this investigation is the question: 'How does competition among online retailers emerge in the dynamics derived from organic search engine optimization?' This inquiry will guide the structure of the analysis, delving into the multifaceted interactions and strategies that shape the competitive e-commerce environment.

As a preliminary step in the analysis of competitor identification and network analysis this work will propose several methods of how NLP and Machine Learning can be applied towards the identification of competitors. First, a way to identifying the key clusters within the e-commerce space. Thereafter two ways of identifying direct and indirect competition is proposed. The first of these utilize Word2Vec and PCA, whereas the second utilize Minimum Spanning Trees to map the overall network of domains.

Building upon the identification of clusters and competitors, the study will redirect the focus on a specific practice in digital marketing commonly referred in the literature as competitive poaching. According to a study by Bhattacharya, Gong, and Wattal from 2022, competitive poaching refers to the marketing tactic where a brand uses a competitor's advertising keywords with the aim of attracting their customers. Throughout this research, this practice will be investigated in the context of SEO and will be referred to as advertisement poaching. The occurrence of this strategy will be observed within and across the clusters found in the preliminary part of the study and will be the basis for comparison on cluster identification.

Another path of analysis of this paper aims to develop a bipartite network graph which nodes (vertices of the graph structure) represents features of the data, such as products, retailers' domains, and keywords terms. The graph's edges not only connect properly the nodes interactions, but also contains weights that leverages numerical features of the interaction, such as cost-per-click (CPC), number of impressions and monthly search volume. This section og the work was developed based on the fact that retailers harvest on considerable advancements in computational power and can gather astounding amounts of high-quality data, as the e-commerce companies and search engines ads generate millions of user interactions on a daily basis. What can be leveraged into knowledge for better decision making to the ones that are capable to perform the necessary transformations and analysis (Gabel, S., Guhl, D., and Klapper, D. 2019; Li, Z., et al. 2020).

2. Related work

The intricate competitive landscape of the e-commerce sector necessitates the implementation of sophisticated analytical methods and strategic decision-making regarding customer acquisition, supplier selection, and product assortment. The significance of product category selection and range is emphasized by Kolk, Fisher, and Vaidyanathan (2015); it has a direct bearing on a company's capacity to attract a wide variety of consumer segments. Aligned with this notion, Gabel (2019) presents a retailer-centric methodology that evaluates competitiveness via product substitutions, thereby providing valuable perspectives on the impact of product diversity on market positioning. Gerling (2023) and Fang et al. (2012), on the other hand, argue that predefined classification systems are inadequate for capturing emergent trends in such dynamic industries and propose more flexible, adaptive alternatives.

Similarly, Gerling (2023) emphasizes the significance of quantitatively representing businesses in order to identify minute distinctions and parallels among them. The aforementioned method, which is vital for acquiring customers and selecting suppliers, is enhanced by the semantic vector representations that Mikolov et al. introduced. The aforementioned factors highlight the intricacy and ever-changing nature of the e-commerce industry's competitive environment, underscoring the importance of employing strategic judgment when it comes to product assortment, adaptive classification systems, quantitative representation, and advanced modeling.

The illustration of how clustering techniques can be effectively utilized in the extraction of social media data by Meng, Tan, and Wunsch provides additional credence for their incorporation into e-commerce (Meng, Tan og Wunsch 2019). Difficulties such as organizing web images, integrating multi-modal data, identifying user communities, conducting sentiment analysis, and detecting social network events are all tackled by clustering methodologies. In the domain of e-commerce, where they are capable of interpreting consumer behaviors, market trends, and competitive environments, their adaptability and proficiency in managing complex data sets are especially practical. This information is of immense value when it comes to making strategic decisions in the e-commerce industry, drawing parallels with the difficulties encountered in social media regarding the analysis of extensive data and comprehension of nuanced consumer interactions and market dynamics.

In addition, the report "Automated Competitor Analysis Using Big Data Analytics: Evidence from the Fitness Mobile App Business" analyzes the competitive landscape of the fitness mobile app industry using Natural Language Processing (NLP) effectively (Yin og Lu 2017). The potential of natural language processing in the wider e-commerce sector is underscored by its

application in automated competitor identification. This utilizes substantial amounts of unstructured data to efficiently extract profound insights regarding consumer behavior and market trends. The ability to comprehend the competitive e-commerce environment is of utmost importance as it facilitates strategic positioning and achieves success in the market. The findings presented in this report underscore the significance of advanced data analytics in the realm of electronic commerce and offer a strategic framework for organizations seeking to improve their approaches to competitive analysis in a constantly changing market.

Besides how clustering techniques can be used in pair with Natural Language Processing to identify competitors, other techniques like Minimum Spanning Trees can be of great help too. (C. Zhou 2009), (Battiston, et al. 2012), and (Gofman 2017) show graph theory can be enhanced with Natural Language Processing and other numerical features to determine different relationships among firms in the financial industry which can be translated into different industries to assess network structures as important players in a certain market, connectivity among companies/domains and potential snowball effects that can start with a company/domain in a certain node and then be translated to subsequent nodes or direct/indirect competitors.

3. Dataset & Methodology

This study is based on a dataset sourced from Grips, a German start-up whose mission is to ‘illuminate blind spots and create a comprehensive map of online commerce for retailers and brands’ (Grips 2023). Grips Competitive Intelligence strives to help retailers and brands understand the e-commerce market, uncover winning solutions, and further maximize their clients' potential income trajectory.

3.1 Dataset

The dataset entails observations representing products offered by diverse retailers, providing insights into organic SEO and the position in the SERP. It consists of approximately 48M observations and 16 features, with each individual row representing a distinct product for a specific domain, covering around ~50.000 unique domains. The primary information sources are Grips, Keyword Planner, and predictions made by Grips (Appendix II for data dictionary and sources). The features aim to convey various aspects of the products being showcased and promoted online, encompassing attributes of the products themselves as well as their placement in the advertising marketplace. Product features derived from metadata such as title, category, brand, and price in USD significantly influence market placement and consumer appeal (Vinutha and Prajwal 2023). Moreover, keyword-related features like keyword, estimated cost-per-click (CPC), impressions, and position in SERP play a significant role in a product's potential and visibility online. Keywords, integral to SEO, determine the positioning of digital content, influencing consumer behavior considerably (Hee Park and Agarwal 2018).

Grips' dataset provides an overview of the product offering, keywords per product, and SERP information. The extensive data allows for a myriad of analytical opportunities. For the specified research question, 'How does competition among online retailers emerge in the dynamics derived from organic search engine optimization?', variable selection, data aggregation, and feature generation have been employed to refine the dataset, detailed further in the ensuing section. In summary, this dataset offers a thorough perspective on various facets of products' online behavior, including their attributes and visibility in digital searches, aiding a comprehensive examination of their online performance.

3.2 Data Preparation

To enhance algorithmic efficiency and competitor analysis, the dataset undergoes a series of data cleaning and processing steps, tailored to the nature of the variables. Textual variables are standardized by trimming spaces, converting to lowercase, and applying lemmatization, enhancing domain comparability and NLP algorithm effectiveness. Instances of blank values erroneously recorded instead of 'None' are corrected to minimize noise. Additionally, non-ASCII characters identified during data screening are removed, ensuring data uniformity and algorithm compatibility.

After the textual data is cleansed, a comprehensive evaluation is performed on all columns to determine their sufficiency for further analysis and to address any missing values. This process, detailed in Appendix III, involves assessing the proportion of missing values in each column. Notably, columns 'low_top_of_page_bid' and 'high_top_of_page_bid' exhibit 50% missing values, leading to their removal since 'cpc_1' adequately represents the estimated CPC. Additionally, rows missing key data in 'avg_monthly_search_volume,' 'predicted_productssold,' 'impressions,' and 'producttitle' are eliminated due to the impracticality of estimation and their limited occurrence.

3.3 Feature Generation

Following initial data cleaning, the final Dataframe is built to construct algorithms on. To identify respective competitors, a feature representing the domain prefix was generated. The feature allows for a more comprehensive and integrated overview of the single retailers within each specific domain. By grouping by this feature, the count of products associated to retailers is performed, providing an overview of the brand recurrence and volume. The operations

relating to the grouping of the features allows for completeness and reduced information loss. To aggregate the quantitative variables, summary statistics such as average, mean, and standard deviation, along with the count of unique brands (NuBrands) and unique keywords (NuKeywords) are calculated. The textual characteristics are subjected to a procedure of eliminating stopwords and performing lemmatization in order to maintain uniformity and minimize noise. The final dataset is aggregated per domain containing the aforementioned numeric variables and lists of all string components (e.g., brands, categories, and producttitles).

4. Harnessing FastText and UMAP for Building Clusters

As the number of e-commerce platforms grows, identifying clusters of similar domains becomes both difficult and important. The ability to categorize these offerings based on their characteristics can provide significant insights into market niches and possible competitors, allowing to precisely identify and analyze market moves. For this part of the report a cluster is defined as a group of similar e-commerce operators that share numeric or textual characteristic, representing a differentiating segment or niche within the e-commerce world. The combination of FastText and UMAP provides an innovative and efficient way for delineating these clusters in this setting.

4.1 FastText: Beyond Simple Vectorization

FastText is an NLP algorithm from from Meta's AI Research lab, that enhance word embeddings by leveraging subword information to grasp linguistic nuances (Bojanowski, Joulin, & Mikolov, 2016). Its capacity to encode morphological nuances distinguishes it from other NLP techniques, particularly when dealing with the varied textual material in e-commerce:

1. **Embedding Generation:** FastText is used to encode product titles, categories, and keywords into the shape om embeddings, which are a combination of technical terminology, product information, brand names, and colloquial English. The subword information ensures that even out-of-vocabulary words (which are widespread in the e-commerce area due to developing jargons, brand names, and so on) be represented meaningfully. (Bojanowski, Joulin, & Mikolov, 2016)
2. **Dimensionality Reduction with SVD:** Using SVD on FastText embeddings is critical because textual data, especially when transformed into embeddings, often has significant redundancy and noise. This is efficiently reduced by SVD, which pinpoints and preserves the most latent features, resulting in representations that are both compact and intelligible. SVD, on the other hand, is not used on quantitative data in this instance, to ensure the preservation of the original context and meaning of these. In this instance the silhouette-score was used to identify the optimized number of clusters for the data, which was nine.

4.2 UMAP: Crafting Cohesive Clusters

In the domain of dimensionality reduction approaches, Uniform Manifold Approximation and Projection (UMAP) stands out (McInnes, Healy, & Melville, 2020). Its superiority over classic methods such as PCA or T-SNE is based on its unique ability to preserve data topology, making it particularly suitable for clustering in complex contexts such as e-commerce:

1. **Integration of Numeric and Textual Data:** UMAPs key benefit is its ability to retain data's topological structure, making it particularly well-suited for the sophisticated clustering scenarios typical in e-commerce (McInnes, Healy, & Melville, 2020). UMAP's ability to handle multimodal data is a tribute to its prowess: it smoothly mixes

textual and numeric formats. A detailed domain representation is required prior to using UMAP (Becht, et al., 2019). Numeric attributes scaled using Min-Max merge with SVD-reduced FastText embeddings in this case, resulting in a balanced representation that captures both quantitative and qualitative domain features.

2. **Tuning UMAP for Optimal Clustering:** Aside from its ability to handle a wide range of data, UMAP provides parameter adjustment freedom. The settings within the UMAP function are adjusted to optimize the silhouette score to achieve optimal clustering. A high silhouette score indicates clusters that are not only internally coherent but also well separated from surrounding clusters—an essential condition for accurate competitor identification (Rousseeuw, 1987).

For this methodology, the synergy between FastText and UMAP enables for efficient and meaningful clustering across domains in the e-commerce market. FastText's ability to generate subtle embeddings, along with UMAP's topological clustering, provides a solid technique for demarcating clusters, laying the groundwork for further competitor research.

4.3 Visualizing the E-commerce Terrain

The world of e-commerce is vast and varied, and as with any expansive landscape, visualization is key to truly grasping its intricacies. The provided UMAP plot, a result of the advanced methodologies, paints a vivid picture of the e-commerce domain landscape, clustered by competitive similarity.

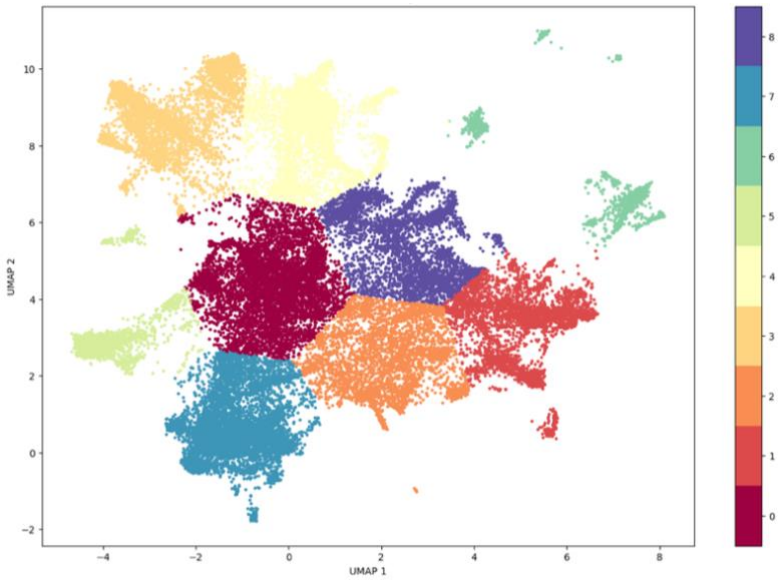


Figure 1 - Domain Clusters

From the onset, it's evident that this methodology provides a clear representation of domain proximity. Through the combination of textual (product titles, categories, and SEO keywords) and numeric data, domains have been mapped into nine differentiating clusters. This gives rise to a spectrum where domains sharing semantic similarity, whether they offer similar products, use akin marketing strategies, or have similar numeric attributes, are nestled together. Such clustering not only underscores direct competitors but also elucidates the nuanced niches within the e-commerce space.

Diving deeper into the visualization, the intricacy of the e-commerce domain landscape unveils itself. The multitude of clusters, each distinguished by unique colors, is indicative of the diversity in the market. Each of these clusters can potentially represent a distinct product niche or category. The density of these clusters, juxtaposed with their spread, becomes a commentary on market saturation and the intensity of competition. Where dense clusters might suggest fiercely contested segments, sparser regions can hint at untapped market opportunities, awaiting savvy entrepreneurs.

For businesses, the visualization is a strategic goldmine. By locating themselves within this mapped terrain, they can discern their competitive positioning. Are they ensconced within the heart of a bustling cluster, surrounded by rivals vying for the same customer base? Or do they find themselves on the fringes, pioneers in a relatively uncharted niche? Such insights can profoundly influence their market strategies, from how they differentiate their products to where they channel their marketing energies. Additionally, a closer examination of domains within specific clusters can serve as a valuable blueprint for businesses. By studying what similar domains are executing, whether in terms of product offerings, marketing campaigns, or customer engagement strategies, businesses can glean actionable insights. Identifying patterns of success within these clusters could guide a domain to tweak its strategy, adapting best practices to increase revenue and market share. Conversely, observing recurrent pitfalls can act as cautionary tales, steering businesses away from potential missteps.

In essence, this visualization serves as a compass in the vast ocean of e-commerce. By succinctly mapping domains based on their textual data, it provides businesses with a clear vantage point, from where they can chart their course with clarity and confidence, and simultaneously adapt strategies for revenue growth and market dominance.

4.4 Analysis of defined clusters

To further understand the defined clusters an analysis of the domains and overarching categories have been completed. This can be seen in Appendix IV, with the key trends and market segments of each cluster summarized below:

Table 4.4.1

Overview of clusters and key characteristics

Cluster	Definition
0	Niche markets, collectibles, and hobbies
1	Technology and electronics sector
2	Sports, active lifestyle, and outdoor equipment market
3	Cosmetics and personal care products
4	Food and beverage retailers and specialty goods
5	Luxury brands and high-end products
6	Home and lifestyle goods
7	Fashion, accessories, and products
8	Office, home appliances and furnishing

Each cluster, as revealed through the analysis of e-commerce data, symbolizes a specific segment, reflecting the varied characteristics and market dynamics in e-commerce. Now, evaluating the key numeric characteristics, found in Appendix V, of each cluster there are clear patterns that are well in line with the characteristics and segmentations. To begin with cluster 5, representing the luxury segment on average have a lower monthly search volume, which is in line with previous research on luxury consumers and their preference for in-store shopping (BOF Insights 2023). Similarly, the average price is also way above others, with only cluster 1 (technology) coming close. The opposite trend can be seen with cluster 1 where monthly search volume on average is the highest across all clusters. Analyzing key numeric attributes attest to the findings of the clusters and serve as a verification of the algorithms ability to correctly classify similar e-commerce operators.

5 Uncovering Underlying Dynamics Among Competitors in the Digital World

5.1. Background

The modern business environment has experienced a significant transformation due to the digital age. Previously, competition was primarily based on geographic proximity and similarities in products or services. However, the rise of online commerce and the importance

of search engines have created a complex, digital competitive landscape. In this era, businesses not only vie for market share but also compete for visibility on search engine result pages (SERPs). This shift necessitates a more nuanced approach to identifying competitors, as traditional methods focused on product similarity or location fall short. The research presented aims to redefine competitor identification by recognizing the importance of understanding the competitive dynamics in the pursuit of visibility on SERPs, prompting the exploration of innovative approaches in this evolving landscape.

5.2. Problem

The digital age has transformed business competition, emphasizing the importance of visibility on search engine results pages (SERPs). Traditional methods, like keyword-based SEO analyses, struggle to capture the complexity of digital competition. Identifying competitors based solely on product similarity is no longer adequate, and the intricate relationships among businesses in the digital space are often overlooked. In response, the research proposes innovative approaches, such as Minimum Spanning Trees (MSTs) and network analysis, to redefine competitor identification and address the critical challenge of contemporary digital marketing and competition analysis.

5.3. Improving the accuracy and quality of competitors identification

This research seeks to develop a more precise and comprehensive method for identifying competitors. It aims to go beyond simplistic classifications based only on standard features. Instead, the goal is to provide businesses and digital marketers with a refined understanding of their competitive landscape, considering the complex interaction of digital elements. By doing

so, its aims to equip professionals with the insights needed to improve their strategic decisions and optimize their digital marketing efforts.

5.4. Methodology

The analytical approach outlined in this section comprises two main steps. The first step involves creating numerical features at the domain level, derived from quantitative variables at the product level. Simultaneously, textual features are extracted using FastText to generate vector representations for each domain, capturing rich textual characteristics. Singular Value Decomposition (SVD) is then employed for dimensionality reduction, making the dataset manageable and informative.

The concept of Minimum Spanning Trees (MSTs) is introduced for network analysis, serving a crucial role in competitor mapping. MSTs offer a practical approach to unveil complex relationships among domains, distinguishing between direct and indirect competitors. This is especially vital in digital competition, where traditional methods struggle with intricate interdependencies. The discussion explores the significance of MSTs and their potential in interpreting competitive dynamics.

The combined use of FastText, SVD, and MSTs is deemed essential for a comprehensive understanding of competitive dynamics in the digital realm. These techniques enable the capture of both quantitative and textual characteristics, providing effective navigation through the complexities of modern digital competition.

5.5 Text embeddings and MST

In shaping our analytical approach, we place significant emphasis on leveraging text embeddings and Minimum Spanning Trees (MSTs) to gain nuanced insights into the digital competitive landscape.

5.5.1 Text Embeddings

Our approach involves utilizing FastText to extract rich textual features and generate embeddings. FastText's unique ability to consider subword information, treating words as bags of character n-grams, proves invaluable. This approach captures the intricate morphological structures and variations within short, informal texts—common characteristics in our dynamic dataset, which includes domain names and product descriptions.

5.5.2 Minimum Spanning Trees (MSTs) in Network Analysis

MSTs, drawn from graph theory, serve as a foundational tool in our network analysis. By connecting nodes (representing domains or entities) with the minimum total edge weight, MSTs simplify the intricate web of relationships into a more understandable structure. In the realm of digital competition, where traditional methods may fall short, MSTs offer a unique advantage. They enable us to distinguish between direct and indirect competitors, revealing the underlying structure of competitive relationships in a particularly relevant and insightful way.

5.5.3 Integration and Impact

The combination of text embeddings and MSTs forms a robust analytical framework. FastText allows us to represent domain-specific textual elements comprehensively, while MSTs distill complex relationships among domains, offering a clear delineation between direct and indirect competitors. This integrated approach equips us with a powerful means to navigate and

understand the competitive dynamics within the digital space, contributing to a more informed and strategic analysis. In this way, we can create a rich representation of each company and build a correlation and distance matrix that can differentiate similar companies for later used to identify potential competitors across companies.

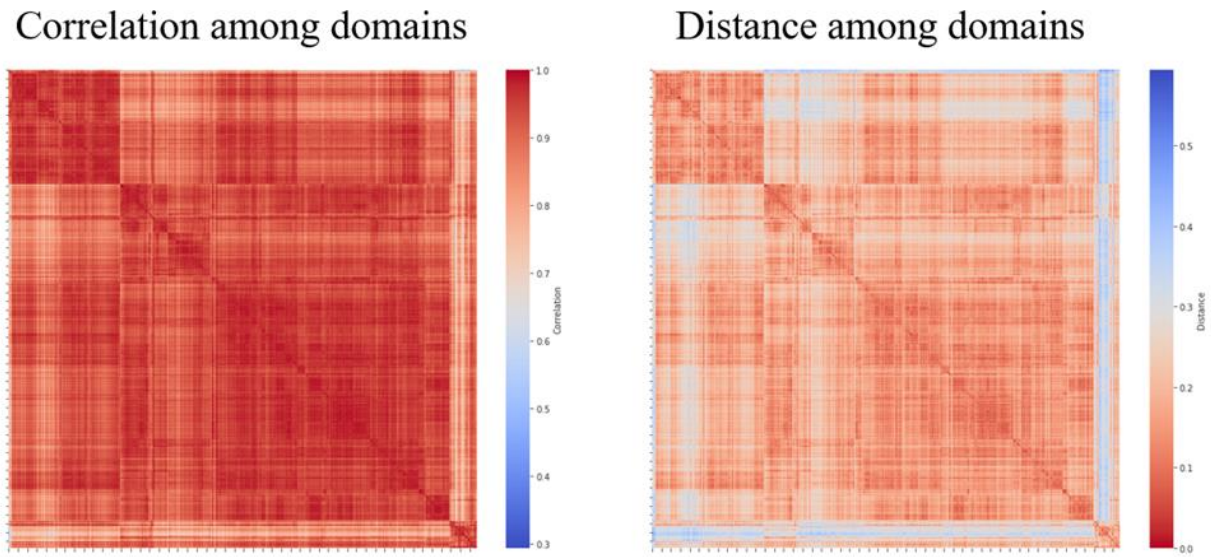


Figure 2 - Correlation vs. Distance Matrix among Domains

Note: Sample for the top 20k domains based on number of products available.

5.5.4 MST direct and indirect competitors

A deeper exploration of peer-to-peer dynamics becomes essential. To minimize the influence of irrelevant relationships between domains, we employ the shortest path algorithm (Mehlhorn and Sanders 2008) to avoid irrelevant connections among domains. This analysis helps to identify both direct and indirect competitors, significantly impacting strategic decision-making and marketing tactics. This analysis goes beyond the mere identification of competitors, shedding light on the intricate interactions that shape market dynamics. To exemplify the application of our findings, we use the prominent retail brand, Adidas. After reducing the number of connections through our previous analysis, we unveil a revealing structure of

competitors. It is important to note that this approach let us discover direct and indirect competitors for different companies right away, further analyses should be considered whenever we need to have a deeper understanding of dynamics among them to polish better strategies.

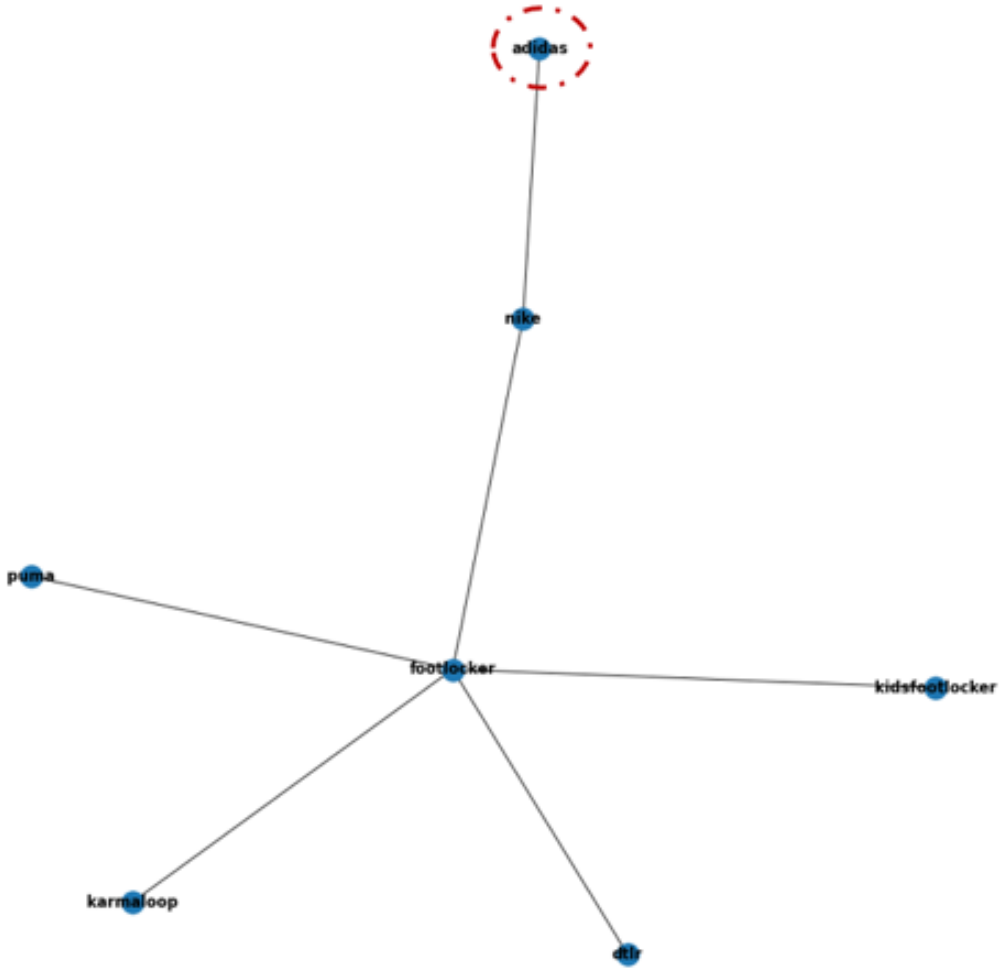


Figure 3 - Potential direct and indirect competitors for Adidas

Note: The target company is circled in red.

Some companies might look familiar right away when analyzing the competitors map for Adidas at first sight, while others might not be that familiar. To have a better understanding of why they are related in the digital world we could explore the semantic similarity behind direct and indirect competitors to understand why they are related in the digital or physical world.

6. Advertisement Poaching and Brand Dynamics in E-commerce

6.1 Introduction to Competitive Poaching in E-commerce

Building upon the robust clustering framework established in Section 4.3, which utilizes FastText and UMAP to categorize e-commerce platforms into distinct market niches, this section delves into the dynamics of competitive poaching within these clusters. As revealed through the findings, these clusters not only highlight direct competitors but also set the stage for understanding intricate competitive strategies such as advertisement poaching in the SEO landscape.

In the highly competitive landscape of e-commerce, the act of advertisement poaching, especially in the context of SEO, presents a distinct and significant challenge. E-commerce retailers use competitor names as keywords and backlinks to redirect traffic, resulting in complex patterns of influence and competitiveness within the network as identified in the clusters (referenced in Section A.4.1). This practice is both employed in organic SEO as well as Search Engine Marketing (SEM).

The subsequent sections will be structured as following. The first part of this paper will present findings on advertisement poaching within and across the previously defined clusters. This will allow the identification of players that engage in the strategy and how the occurrence of this changes across clusters. Furthermore, this research aims to observe the phenomenon’s patterns through network analysis, focusing on establishing communities based on the frequency of advertisement poaching and comparing them to clusters identified in part 4.4.

6.2 Methodological Framework for Analyzing Advertisement Poaching

6.2.1 Data Collection on Retailer Keyword Usage

To investigate advertisement poaching in online retail, the frequency at which one retailer's name appears in another's keywords is analyzed. This is reflected in a dataset with variables such as *Retailer_From*, *Retailer_To*, *Cluster_From*, *Cluster_To*, and *Count*, illustrating instances of this practice. Specifically, it shows when a retailer (Retailer_To) uses another's name (Retailer_From) in their associated SEO keywords, with 'Count' denoting the frequency of these occurrences. This analysis sheds light on the competitive strategies in the digital marketplace, where retailers may use competitors' names within their SEO strategy to draw or redirect traffic to their sites. The subsequent analysis focuses on the prevalence of this phenomenon across different clusters, each representing a specific industry (refer to 4.1. for details on cluster-specific industries).

6.2.2 Filtering Criteria for Enhanced Data Accuracy

To enhance the accuracy of this investigation on advertisement poaching, a targeted filtering process was implemented. This approach specifically excludes domains where retailer names are ambiguous, serving both as product descriptors and keywords. This measure is crucial to ensure that the analysis accurately captures instances where retailers deliberately engage in advertisement poaching as a strategic SEO tactic, rather than accidental overlaps. Examples of excluded domains are provided in Appendix VI, ensuring a focus on deliberate strategies in competitive dynamics in e-commerce.

6.3 Quantitative Insights into Advertisement Poaching.

6.3.1 Comparative Analysis of Within-Cluster and Across-Cluster Poaching

A comprehensive analysis of advertisement poaching reveals significant insights into its prevalence both within and across the e-commerce clusters defined in Section 4.3. This practice, where retailers use the names of other retailers in their SEO strategies, varies in frequency and intensity depending on the cluster context and industry.

Within clusters, the investigation reveals a substantial occurrence of poaching. The resulting findings stem from 62,480 instances of poaching, with an average of 36 occurrences per instance. The standard deviation is at 316, indicating significant variation in the frequency of poaching within clusters. The range of occurrences spans from a minimum of 1 to a maximum of 21,634, underscoring some extreme cases of poaching within specific clusters.

In contrast, when looking *across different clusters*, 93,706 instances are observed, with a lower average count of 24.53 occurrences per instance. This suggests that while poaching is still prevalent across clusters, it happens less frequently on average compared to within clusters. The maximum count in across-cluster poaching, however, peaks at an even higher value of 39,736, indicating that in some cases, the extent of poaching across clusters can be significantly high. This result can be attributed to SEO practices of larger online marketplaces.

6.3.2 Notable Retailers in Poaching Dynamics

Exploring the details, some retailers are particularly notable when it comes to advertisement poaching. For instance, 'Apple' and 'Amazon' are the most frequently appearing names used by other retailers in their SEO strategies, with counts of 2,121 and 1,872 respectively as Retailer_From. On the other side, 'Amazon' appears to be the most common Retailer_To, involved in 4,696 instances of poaching, suggesting its significant prominence as a target in

SEO poaching strategies. Other notable mentions include 'Ebay', 'Walmart', and 'Sears', indicating their roles either as common targets or users of others' names.

6.3.4 Implications of the Findings

These findings indicate a strategic pattern in the use of competitors' names in SEO, reflecting a competitive landscape where retailers frequently resort to using the names and backlinks of more prominent or directly competing businesses to enhance their own visibility. The higher frequency of poaching within clusters could suggest a more intense competitive atmosphere among closely related retailers. Conversely, the lower average yet higher peak in across-cluster instances might reflect occasional but very aggressive strategies employed by retailers when targeting businesses outside their immediate cluster.

The significant presence of leading retailers like 'Apple' and 'Amazon' in these poaching activities, both as targets and perpetrators, highlights their centrality in the competitive dynamics of the retail market. This trend raises questions about the competitive implications of SEO strategies in digital marketing and indicates a need for a more detailed understanding of online competitive practices.

6.3.5 Prevalence of Poaching within E-Commerce Clusters

The analysis of within-cluster advertisement poaching practices uncovers differences in the choice of integrating the strategy in their SEO across industries. As shown by Figure X, clusters 1 and 6 exhibit the highest relative frequency of poaching, possibly indicating more aggressive SEO competition or larger market sizes in these industries. Examining individual clusters allows to gain insight into the strategic grounds of SEO techniques in different retail sectors.

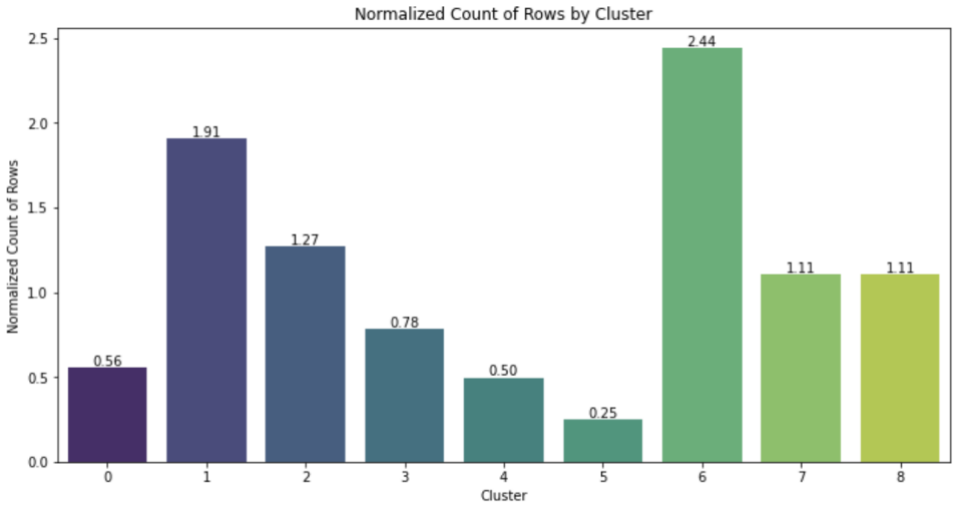


Figure 4: Counts of Advertisement Poaching by Cluster (normalized by cluster size)

Highly Competitive and Popular Sectors (Clusters 2, 3, 7, 1): In clusters related to sports and outdoor equipment, cosmetics, fashion, and technology, there's a significant trend of larger retailers like REI, Dick's Sporting Goods, Ulta, Sephora, RunRepeat, Nordstrom, B&H Photo Video, and Best Buy targeting well-known brands such as Vans, Oakley, Nike, Adidas, Clinique, and HP. This indicates a strategic focus on poaching from high-value, popular brands in highly competitive markets, reflecting the intense struggle for online dominance in these sectors.

Specialized and Niche Markets (Clusters 0, 5): Niche markets, collectibles, hobbies, and luxury brands (like Funko, Lego, Pandora, and Movado) are also prime targets for advertisement poaching. Retailers such as StockX, HotTopic, and BeCharming are leading the charge in these sectors, demonstrating a specific interest in capitalizing on the niche appeal and dedicated consumer bases of these brands.

Home, Lifestyle, and Everyday Goods (Clusters 6, 8, 4): In clusters dealing with home and lifestyle goods, home appliances, and food and beverage, the trend is towards poaching from a broader range of brands (e.g., KitchenAid, Frigidaire, Kroger, Tesco). Key players in poaching

here include Amazon, eBay, and Jewel-Osco, pointing to a strategy of capturing traffic across a wide spectrum of everyday consumer products.

Overall, these findings highlight a landscape where larger retailers and online marketplaces are actively engaging in SEO poaching across various sectors. They tend to target both high-profile brands in competitive markets and specialized or niche brands, aiming to capitalize on their established market presence and dedicated customer bases. This underscores the importance of robust and defensive SEO strategies for brands across all sectors.

6.3.6 Implications of Cluster-Specific Poaching and Overall Trends

The detailed analysis of advertisement poaching within specific e-commerce clusters sheds light on the competitive dynamics and SEO strategies prevalent in different retail sectors. By assessing the brands that are commonly targeted for poaching, the investigation reveals insights into how retailers strategically leverage the brand equity of recognized names to their advantage. In each cluster, the prevalence of poaching highlights the importance of brand visibility and recognition in SEO strategies. For example, in technology and electronics, the focus on brands like 'HP' and 'Samsung' underlines the industry's reliance on brand reputation for attracting consumer interest. Similarly, in fashion and apparel, targeting brands like 'Nike' and 'Adidas' points to the competitive nature of the industry, where brand visibility is a key driver for consumer choices.

A pervasive trend observed across all clusters is the prominent role of online marketplaces and multi-brand retailers, often referred to as 'aggregators'. Considering the multitude terminologies employed for describing these types of retailers, this study will utilize the terms 'aggregators' and 'third-party resellers' synonymously. These entities, which offer a diverse array of brands

under a single platform, consistently emerge as the most active in employing advertisement poaching strategies. This approach suggests that aggregators leverage the recognition and trust associated with established brand names not just for redirecting traffic, but also as a crucial element in enhancing their own SEO and market visibility. Their strategies indicate a keen understanding of the consumer's search behavior and the importance of brand association in the digital marketplace.

This analysis underscores the multifaceted nature of SEO strategies in the e-commerce sector, highlighting the intricate interplay between brand recognition, consumer behavior, and competitive positioning. It also raises questions about the ethical and strategic implications of such tactics in digital marketing, suggesting the need for ongoing scrutiny and adaptation of SEO practices in light of evolving market dynamics.

6.4 Network Analysis of Advertisement Poaching.

6.4.1 Methodology: Community Detection via Network Analysis

To investigate competitive dynamics of advertisement poaching among online retailers., network analysis is applied. This approach complements the FastText and UMAP clustering by providing a nuanced view of retailer interactions within these clusters. The network is visualized as a series of nodes (retailers) connected by weighted edges (indicating poaching counts), creating a bidirectional flow representing the dual role of retailers as perpetrators and victims of poaching. The networks complexity is underlined by the network's sparsity, quantified at approximately 0.000298 (for visual representation, see Appendix VII). This figure highlights the scarcity of direct connections relative to all potential connections, a state which significantly impacts algorithm choice for community detection. The chosen Community Infomap method is particularly effective for sparse networks, as it simulates the movement of

information through random walks. This approach is detailed by Rosvall and Bergstrom 2008 and further emphasized by Fortunato (2010). The algorithm's effectiveness was quantitatively assessed using the Adjusted Rand Index (ARI), a measure of the similarity between two data clustering that accounts for chance groupings. An ARI of 1 suggests perfect correspondence, whereas 0 or below indicates random or discordant clustering.

6.4.2 Results and Interpretations from Network Analysis

Referencing the original clusters identified using FastText, embeddings, and UMAP in A.4.1, the network analysis via the Community Infomap algorithm illuminates the complexity and competitiveness of advertisement poaching. The obtained ARI for Cluster_From (0.5525) suggests a moderate correlation with the originally established clusters, implying a degree of consistency in the targets of advertisement poaching. In contrast, the lower ARI for Cluster_To (0.0250) reveals a broad spectrum of poaching strategies deployed by retailers, pointing to a more dispersed and individualistic set of tactics. Such divergence in ARI scores brings to light the intricate and multifaceted nature of advertisement poaching strategies within the competitive online retail sector.

The combination of network sparsity and bidirectional connections affects the detection of uniform poaching patterns. The impact of data filtering, aimed at eliminating ambiguity from commonly used keywords, also plays a role in these findings. Together, these factors highlight the dynamic and strategic nature of SEO poaching, underlining the complex interactions between the targets and initiators of these maneuvers within the retail sector.

6.5. Conclusion: Understanding the Network of Competitive Poaching

The examination of advertisement poaching within e-commerce has revealed a dynamic interplay of competitive strategies, with a pronounced role of aggregator retailers in the use of this practice. The investigation has uncovered the variance of poaching within and across different retail clusters, indicating industry-specific strategic approaches. The findings underscore the strategic application of competitor names in SEO campaigns, suggesting a more intense competitive environment within clusters and a diverse set of tactics across clusters. Notably, the prominence of certain retailers as frequent targets points to their centrality within the digital marketplace.

7. Graph Neural Networks in E-commerce

The primary objective of this section aims to develop a comprehensive framework for analyzing the competitive landscape in the e-commerce using a Graph Neural Network. The study starts by constructing a heterogenous graph that accurately represents several e-commerce entities such as domains, keywords, products, and categories, along with their interrelationships by relying on weighted edges which leverages relevant information regarding the given relations. The advent of Graph Neural Networks (GNNs) has opened new avenues for data analysis, particularly in understanding complex relationships within large datasets. In recent years, GNNs have become an important topic that rises interest among the research community, several papers are dedicated to effectively learn node embeddings over different graph-based representations. The neural network treats the underlying graph as a computation graph and generates individual node embeddings by passing, transforming, and aggregating node's features present in the given graph. Later, the generated embedding can be widely used to downstream tasks such as classification, recommendation, or prediction. **(**Li, Z., et al. 2020)**.**

7.1 FastText Training and Most Similar Method

The next step relied on training a FastText model which was developed by Facebook AI Research (Bojanowski et al., 2017) which is designed to efficiently learn word representations and sentence classification. FastText works by treating each word as a bag of character n-grams, allowing it to capture the meaning of shorter words and understand suffixes and prefixes, making it especially useful for processing text with misspellings or mixed languages. The model is a state-of-the-art word embedding technique that extends the Word2Vec model to incorporate subword information. This approach offers richer semantic representations as FastText generates embeddings that capture not just the semantic meaning of a word but also the meaning of subword units (n-grams). The model is also capable of handling occurrences of Out-of-Vocabulary (OOV) words unlike traditional word embeddings algorithms as Word2Vec by combining the embeddings of their subword units. This feature is relevant in the context of the dataset of this project as not only the vocabulary in products is constantly evolving but also there are specific words that aren't trivial for a usual vocabulary, such as brand or domains names.

The first step to train the model was to tokenize the text variables which previously were submitted to the process of stop word removal and lemmatization. Namely this step was done for the variables: domain, keyword, product and category. Then one model was trained for each of the variables. The FastText model was configured for 256, 128 and 64 features (or embedding dimensions). The final dimensionality was chosen to balance the richness of the representation in terms of performance levels, methodology to be detailed further on this paper, against the computational efficiency. A higher number of dimensions can capture more detailed semantic relationships but at the cost of increased computational complexity.

Specifically, for the variable keywords the FastText “most_similar” method was used, the core if this process was to leverage the function to identify similar words to every keyword in the dataset. A threshold of 0.8 was set to determine the degree of semantic closeness required for words to be considered in the graph structure, by setting it the focus were exclusively on words that had substantial degree of similarity ensuring relevance and accuracy on the words that will be connected on the graph. The choice of the threshold was done arbitrary based on the analysis of the output of the code to a significant number of keywords, it was observed that around this level there was a good balance of number of similar words and relevancy. The process resulted in a curated list of related words and their respective similarity score. The idea behind this feature was to enhance the graph structure which was not just structural but also conceptual, providing a more intricate and interconnected representation of keyword relationships as it is expected to allow further universal applications and enhances interpretability (Gerling, 2023).

7.3 Knowledge Graph Construction for E-Commerce Data

The initial phase of the methodology of this works involves constructing a graph that accurately represents the relationship of the variables within the dataset. This graph is not just a simple network but a heterogeneous one, meaning it includes multiple types of nodes and edges, each representing different entities and relationships. For this the Deep Graph Library (DGL) was used to build the bipartite graph whose nodes represents unique values for variables that represent components of the search engine keywords and its results as it excels in processing large-scale graph data and its optimized data structures and operations are specifically designed for graph neural networks. The total number of nodes is 4.5 million and they are divided as: 2.5 million nodes represent unique keywords, 80.1 thousand related words, 20.2 thousand domains, 8 thousand categories and 1.9 million products. Each node is enriched with a n-dimension

vector build from the correspondent previously trained FastText models. These embeddings capture the semantic essence of the text variable represented by the nodes providing a rich representation of its meaning. The nodes were then connected by weighted edges which translates various types of relationships and interactions among the nodes, examples include 'belongs to' relationships (linking products to categories), 'connects to' (associating keywords with domains or products), and other relevant e-commerce interactions. To encapsulate the strength and nature of the relationships, edges were assigned weights based on metrics like cost per click, number of impressions and monthly search volume. These weights play a crucial role in the analysis, as they quantify the intensity and significance of the connections within the network. By leveraging DGL's capabilities to handle heterogeneous graphs and enriching the node and edge representations with relevant features and weights, a detailed and nuanced graph of the e-commerce domain was created. This graph serves as the backbone of the analysis, enabling a deep dive into the complex web of relationships that define the competitive landscape in e-commerce.

Then, once the representations are available the use of GNNs is enabled. The idea is to create a embedding of the network structure that translates the competitive landscape at domain and keyword level. These embeddings will make possible downstream tasks to competitor retrieval and keyword recommendations for better SEO management. The evaluation of the tasks will be further described on section D of this paper.

7.4 Cluster Analysis

Cluster analysis was conducted to the embeddings generated out the GNN model for the domain’s nodes to distinct the competitive categories aiming to uncover underlying structures and relationships within the e-commerce landscape. For this purpose, the k-means algorithm

was chosen for the task. This algorithm partitions the nodes into k clusters in which each node belongs to the cluster with the nearest mean, the algorithm was chosen due to its efficiency and effectiveness in handling large datasets, it is notably well-suited for grouping data points, domains embeddings in this particular case. There are a few methods for determining the optimal number of clusters in a given dataset, for this step the elbow method and silhouette analysis were performed. The elbow method consists in plotting the explained variance as a function of the number of clusters and choosing the elbow of the curve as the number of clusters to use. The elbow point represents where the marginal gain in explained variance begins to diminish, indicating an optimal cluster count. To obtain additional validation to the number of clusters obtained the silhouette analysis comes handy, it measures how similar an object is to its own cluster compared to other clusters. A high silhouette score indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. After evaluation, the number of clusters were set to 3 (k = 3).

The subsequent cluster analysis, leveraging K-means clustering and validated through the elbow and silhouette methods, segmented specifically the domains embeddings generated with the GNN architecture into distinct groups. This segmentation enabled the development of a method to be described further on this report to understand of how different domains interact and compete in this space.

7.5 Top-N Recommendation Algorithm

An algorithm was created to address the goal of retrieving the most related nodes to within its type, which translates to the understanding which nodes can be considered competitors to a given vertices for the domain use case, or which are the most similar keywords in terms of graph structure and node centrality when that is the node type inserted in the analysis. The

algorithm employs the embeddings generated from the GNN to compute similarity scores and identify the most relevant embeddings related to the node of analysis.

For evaluation purposes, the algorithm applied to the domain's nodes required an initial step involving the creation of a mapping between domains and their associated keywords, what will enable the notion of the shared portion of a retailer's keyword portfolio to the competitor recommended by the algorithm. For the case of keywords, the metric suffered a slight change, the whole list of domains for a specific keyword was treated as the portfolio for that entry and the shared portion computed.

The algorithm starts by iterating to each of the node out of all of its type to retrieve its embedding. Cosine similarity is then computed between the nodes and all the nodes in the set. The cosine similarity measurement captures the degree of resemblance between the nodes in the multidimensional embedding space of the graph structure. For the additional layer of analysis, the algorithm measures the portion of shared keywords or domains, depending on the use case. The output of the algorithm is a comprehensive dataset of records about the nodes of analysis, including information such as the node, its top related nodes, similarity scores, combined scores and the overlap metrics.

The Top-N Recommendation Algorithm is a sophisticated tool in the arsenal of e-commerce competitive analysis. By integrating advanced machine learning techniques with traditional competitive analysis metrics, it provides a nuanced understanding of the competitive landscape and is able to provide recommendations for keyword bidding strategy. The algorithm's focus on cosine similarity ant is supplemented by keyword overlap analysis, offers a multi-dimensional view of the landscape.

8. Limitations

During this research, many constraints within the provided dataset and methodology were identified that are crucial to acknowledge. In the first place, the dataset utilized has limitations pertaining to the data. There is a significant lack of uniformity across many domains and brands. The inconsistency primarily arises from the precision of web scraping, as being the basis of Grips methodology (Grips 2023). These inconsistencies provide interference during the process of training the algorithm, despite that attempts have been made to reduce the impact of unusual data and limited data from specific domains.

Moreover, the assessment of the results generated by the algorithms presents substantial difficulties. An important concern revolves around the ambiguous delineation of 'competition,' which can significantly differ depending on the context. It is crucial to recognize that the notion of 'competition' in this context is subjective and can vary considerably among individuals. Moreover, the lack of true labels limits an accurate assessment of the algorithms' performance.

Additionally, when it comes to analyzing data in the field of e-commerce, it is crucial to consider various inherent restrictions. The dataset, obtained from numerous e-commerce operators, offers a fixed representation of specific domains rather than up-to-date, dynamic data appropriate for comparison analysis. The data quality in these multiple domains varies, which may introduce additional noise to the computational models. The variability in data quality is closely related to the reliability of online scraping techniques (Grips 2023).

Furthermore, the ever-changing nature of Search Engine Result Pages (SERPs), which adapt based on different domain features as of 2023, introduces an additional level of complexity to the analysis. The rapid evolution of both specific fields and combined sets of data presents a difficulty in continuously assessing the quality and uniformity of the results. The rapid and dynamic nature of this change not only makes it challenging to consistently evaluate existing models, but also makes it difficult to assess any enhancements that may arise from algorithmic adjustments.

9. General Discussion & Conclusion

The initial approach allowed the identification of competitors under different definitions and methodologies. It has been shown that the identification of competitors is possible based on similarities using clustering and network analysis techniques in **Error! Reference source not found.**, and

Graph Neural Networks in E-commerce: Unveiling Competitive Dynamics through Search Engine's Keywords Analysis

1 Introduction

1.1 Background and Motivation

In the recent years the relevance of e-commerce in sales of industry goods has grown exponentially. Specially after the world COVID-19 pandemic, online retailers saw their sales trend increase even in a higher rate than what was already happening since the beginnings digitalization of our world. Despite the fact that the pandemic is over, and brick and mortar stores are once again viable options, customers' behavior was impacted, and some tendencies won't be reversible. Therefore, understanding the competitive landscape on online markets has become more crucial than ever.

The ability to effectively analyze and understand the dynamics of the market can help retailers to better adapt strategies not only for boosting sales, but also to better allocate their investments either on portfolio management, pricing strategy or advertising expenses. Players in the e-commerce market must choose for which products to allocate resources simultaneously. These advertising related choices have a direct impact on consumer choices, final sales results and contribution margins (Gabel, S., Guhl, D., and Klapper, D. 2019).

The research of Elrod et al. (2002) argues that market structure analysis has the tendency to rely and focus on the substitution characteristics within a defined product category due to the limited availability of cheap data, computing power or scalable models (Gabel, S., Guhl, D., and Klapper, D. 2019). However, the growth of the buying behavior on online retailers conveniently provides enormous amounts of structured data. Over the last decades, there has been a

substantial advancement over collecting and managing this kind of data, but not a huge progress has been made regarding searching advertising over this data (M. Kargar, 2022).

Today, retailers harvest on considerable advancements in computational power and can gather astounding amounts of high-quality data, as the e-commerce companies and search engines ads generate millions of user interactions on a daily basis. What can be leveraged into knowledge for better decision making to the ones that are capable to perform the necessary transformations and analysis. (Gabel, S., Guhl, D., and Klapper, D. 2019; Li, Z., et al. 2020)**.** One of the many possibilities is to develop a bipartite network graph which nodes represents features as products, retailers' domains, keywords terms and so on and edges not only connects properly the nodes interactions, but also contains weights that leverages numerical features of the interaction, such as cost-per-click (CPC), number of impressions and monthly search volume.

The advent of Graph Neural Networks (GNNs) has opened new avenues for data analysis, particularly in understanding complex relationships within large datasets. In recent years, GNNs have become an important topic that rises interest among the research community, several papers are dedicated to effectively learn node embeddings over different graph-based representations. The neural network treats the underlying graph as a computation graph and generates individual node embeddings by passing, transforming and aggregating node's features present in the given graph. Later, the generated embedding can be widely used to downstream tasks such as classification, recommendation or prediction. **(**Li, Z., et al. 2020)**.**

The use of GNNs by leveraging its ability to handle graph structures present an innovative approach to decipher the intricate relationships in e-commerce data, what includes the associations among domains, products and keywords, which are important to understanding the

underlying market dynamics. This work will rely particularly on GraphSAGE (Graph Sample and Aggregation) algorithm applied in a bipartite graph structured data. The intuition behind the bipartite GraphSAGE is that at each iteration the model aggregates information from local neighbors' items. As this process repeats, vertices will incrementally gain relevant information from the nodes in its neighborhood. At each step of the iteration the node will receive more and more information from further nodes in the structure. The basic notion behind this approach is to use dimensionality reduction techniques to refine the high-dimensional information the belongs to a particular node's neighborhood into a dense vector embedding. As mentioned above, these embeddings can later be used to relevant tasks **(**Li, Z., et al. 2020; Hamilton et al. 2018).

The motivation for this research arises on the increasing complexity of the e-commerce landscape and the advertising products used on this market as the existing literature (to the best of our research) fails to provide methods for analyzing competitiveness based on features this work relies on. The Graph Neural Networks, with its ability to encode relational information, offer a promising solution to retrieve similar players (competitors). Additionally, the integration of the possibility to analyze similarity of keywords in scale enables a strategic tool for businesses to enhance their online presence and advertising strategies, leading to a more robust approach towards one's competitors.

1.2 Research Objectives

The primary objective of this research aims to develop a comprehensive framework for analyzing the competitive landscape in the e-commerce using a Graph Neural Network. The study intent to:

- Construct a heterogenous graph that accurately represents several e-commerce entities such as domains, keywords, products and categories, along with their interrelationships by relying on weighted edges which leverages relevant information regarding the given relations.
- Implement a GNN, specifically a GraphSAGE algorithm, to capture the structural and relational information embedded in the bipartite graph and its weighted edges. The resulting embeddings of each node type will gather information from the whole market structure and will enable downstream tasks for competitor's identification and keywords similarity analysis
- Develop and apply centrality measures to identify key domains and keywords aiming to better understand their influence within the whole e-commerce network.
- Propose a metric for assessing the competitive significance of entities based on GNN embeddings and node's centrality scores.
- Retrieve the most related embeddings vectors to the node's type of analysis to determine within a variable (node) type (e.g.: domains, keywords).

The study does not cover the broader aspects of e-commerce strategy that might influence the competition on the market, such as pricing, supply chain and logistics capabilities. The research is solely based on how the players position themselves to consumers throughout the advertising initiatives.

2.1 Literature Review:

2.1 E-Commerce Competitive Landscape

In other to enable the analysis, the use of Graph Neural Networks (GNN) is one of the possible approaches available in the literature. Although, not many cases documented, to the best of the dive deep, where it was applied for e-commerce players. Rolim (2022) describes the role of

GNN as a deep learning architecture able to incorporate structural information, making then suitable for tasks as node classification and link prediction in e-commerce networks. GNNs are seen as an extension of deep convolutional networks (DCN) and recurrent neural networks (RNN), enriched with the capability to process and analyze graph-structured data (Rolim, 2022). The potential use of GNNs on advanced recommender systems is attributed to Convington et al. for platforms like YouTube is also discussed (Gerling, 2023). Systems that could boost company embeddings by integrating multiple features available for the analysis.

2.2 Network Analysis for Underlying Competitiveness

Network analysis encapsulates a range of techniques for leveraging knowledge out of relationships and interactions on dataset. Recent literatures apply this technique to several different topics increasing the relevance of graph-based approaches. Cao (2023) points out that graphs are ideally suited for representing complex relationships in large datasets. In the scope of this project, this includes interactions between retailers, products and keywords. Gabel (2019) demonstrates how product maps, combined with graph overlays, can elucidate market structure drivers, showing the practical application of graph theory in market analysis.

One interesting approach is the creation of knowledge graphs (KG) which creates structures to translate the intrinsic relations of a given dataset or topic. Cao (2023) introduces CompanyKG, a knowledge graph designed to encapsulate company features and relationships. This graph, originating from a real-world investment platform, is particularly tailored for quantifying inter-company similarity. The paper proposes a representation of companies as graph nodes with attributes and their relationships as represented as weighted edges to translate distinct types of interactions among the nodes. Amazon's pioneering use of graphs for product recommendation

(Rolim, 2022) is a testament to the effectiveness of graph-based methods in enhancing e-commerce strategies, particularly in recommendation systems.

The calculation of real-numbered weights for graph edges, as described by Cao (2023), enables the knowledge graph to leverage the contrasting levels in various relationships, such quantification can be determinant for a nuanced understanding of a given network. Rolim (2022) defends that node similarity is often based on structural embeddings techniques inspired on word2vec algorithm. However, the word2vec, node2vec and struc2vec have a strict limitation towards the inclusion of node features what is crucial for capturing the full essence of node relationships in many use cases (Rolim, 2022). The superiority of GCN models over structural embedding approaches is attributed to their capacity to merge structural and node information. This combination leads to more accurate representations.

The application of GNNs, as discussed by Cao (2023), leverages network science and graph theory. This approach enables the efficient analysis of business landscapes and facilitates the use of advanced machine learning techniques. GNNs are recognized for their ability to merge node structural positions and features, offering a comprehensive view of the network, as explained by Rolim (2022). This integration is pivotal in generating accurate and meaningful node embeddings. The literature on network analysis in e-commerce converges on the effectiveness of graph-based approaches for capturing the complex dynamics of the industry. From knowledge graphs that encapsulate company features and relationships to advanced GNN models that integrate structural and attribute information, these methodologies allow for a more detailed understanding of market structures, competitive relationships, and consumer behavior.

2.3 Graph Neural Networks and GraphSAGE:

Graph Neural Networks (GNNs) and GraphSAGE represent a significant evolution in the field of graph representation, extending the capabilities of traditional graph analysis methods. Rolim (2022) describes GNNs as a framework that integrates entity features with the graph structure. Unlike traditional methods that focus solely on graph topology, GNNs learn representations that encapsulate both the attributes of entities and their relational context within the graph. The author makes the point that some tech companies applied these models to their recommendation's algorithms, for example Pinterest created the PinSage operating on bipartite graph to boost user experience by providing tailored recommendations (Rolim, 2022).

The model's framework relies on the idea of message passing and features aggregation. Hamilton et al. (2018) introduced GraphSAGE to extend the reach of GCN (Graph Convolutional Networks) to inductive unsupervised learning. This was achieved through trainable neighborhood aggregation functions, allowing GraphSAGE to generate node embeddings even for nodes not seen during training. Cao (2023) notes that GraphSAGE is particularly effective for edge prediction tasks. Its training process ensures adjacent nodes have similar representations while maintaining distinct representations for non-adjacent nodes, addressing scalability challenges in graph representation. Cao (2023) also observes that GNN-based methods, including GraphSAGE, tend to group companies within the same sector more closely. This attribute is beneficial for tasks like industry sector prediction. The author points out that on his research for purposes of similarity prediction GNN-based approaches, which consider node embeddings, edges, and their weights, have shown superior performance (Cao, 2023).

GraphSAGE has been applied in different kind of business for different purposes, its ability to handle bipartite weighted graphs allowed Uber Eats to recognize the importance user-item interactions (Rolim, 2022). The author also discusses how convolutional GNN architectures, including GraphSAGE, aggregates neighborhood information based on edge weights allowing for representations which consider the strength of the relationship between nodes. Although most works rely on simple aggregation and order permutation strategies, it's possible to achieve good results applying a canonical order to neighbors and using functions that consider order like LSTM (Rolim, 2023).

GNNs, and specifically GraphSAGE, represent a paradigm shift in graph analysis, offering a more holistic approach to understanding complex networks. By incorporating both node features and structural information, these models provide deeper insights into the relationships and dynamics within networks. The adaptability of these models to different contexts, including weighted networks and various business sectors, further highlights their versatility and potential effectiveness to the understanding of competitiveness as this project aspires to do.

2.4 Nodes Similarity and Recommendation:

Recent advancements in the field focus on quantifying nodes similarities, enhancing recommendation systems, and employing sophisticated natural language processing (NLP) and machine learning techniques to achieve the purpose. Cao (2023) discusses the development of evaluation tasks like similarity prediction and ranking, crucial for quantifying company similarities. This approach involves creating annotated test sets for assessing the effectiveness of similarity measures. Fang et al. (cited by Gerling, 2023) introduce advanced NLP models for calculating company similarities, marking a significant advancement in the field. Gerling (2023) notes the importance of cosine distance in measuring the similarity between vector

representations of firms, a method widely used due to its effectiveness in capturing the geometric or correlational relationship between vectors. Gabel (2019) emphasizes product proximity as an essential feature in product maps, where Euclidean distances represent product similarities. The calculation of cosine similarity scores between companies is a standard method for determining the most similar firms, as detailed by Gerling (2023).

Dimensionality transformation is vastly suggested in the literature to the purposes of improving interpretability and accuracy, facilitating visualizations or reduce computational complexity. Gerling (2023) highlights the role of dimensionality transformation in normalizing company similarities, enhancing the interpretability of relationships. This transformation can significantly alter similarity scores, providing more accurate representations of company relations. Cao (2023) utilizes UMAP (Uniform Manifold Approximation and Projection) to reduce embeddings to two dimensions, facilitating the visualization of company clusters in the embedding space. Gerling (2023) mentions the use of t-SNE (t-distributed stochastic neighbor embedding) to refine embedding spaces and reduce computational complexity. This technique is particularly useful in handling high-dimensional financial company data on the author's work.

The literature on node similarity and recommendation systems highlights the growing importance of advanced analytical techniques, such as NLP models, dimensionality reduction algorithms, and visualization methods. Although, visualization techniques, while powerful, can sometimes produce maps that are difficult to interpret due to the clustering of products (Gabel, S., Guhl, D., and Klapper, D. 2019). To quantify the similarity or competitiveness across players, algorithmic approaches were developed, as in Gerling (2023) describes an algorithmic approach to identify similar firms, utilizing cosine similarities and sorting mechanisms to

determine the most relevant peers. The integration of these advanced techniques represents a significant leap forward in the field, offering more sophisticated and nuanced approaches to understanding and leveraging the intricacies networks.

3: Implementation and Results

3.1 Data Collection and Preprocessing

The dataset preprocessing was vastly described earlier in the report thus no need for scrutinize longer. However, its relevant to point out that for the development of the network to this section we disregarded the aggregation of data on domain level described on aiming to preserve the distribution of numerical features within domains. As the size of the dataset was not reduced by the grouping process the dataset was strategically segmented to focus specifically on keywords leading to products within the fashion (Apparel and Accessories) category due to computational constraints. Limiting the dataset to the specific category reduced the computational load, making the processing and analysis more manageable and efficient, although the size was significantly reduced, the number of observations is still relevant to the project goal (2.5 million keywords, 10.5% of total) and the reproducibility of the methodology for other categories is an easy implementation.

This segmentation allowed for a more in-depth analysis and the category was chosen not only due to its dynamic nature and substantial presence on the e-commerce, but also due to the better results for the classification of products resulted from this search terms. Which contributed to minimize the noise of misclassified products that were more frequently happening in categories such as media. Therefore, the resulting dataset is a representative and insightful subset of the larger dataset.

3.2 GNN Model Implementation

This section focuses on the training process and the loss function employed to optimize the model. The model is a customized version of GraphSAGE, designed to handle a heterogeneous graph with different types of nodes and edges. It consists of different layers as described above, each tailored to handle specific node types in the e-commerce dataset. Different kind of architectures were tested before settling to the final model, MSE and cross-entropy were tested as loss functions, a range of different epochs were tested (5, 10, 20, 30, 40 and 50) and the convolution functions LSTM, mean and pool were applied. Regarding the input network, tests were performed with 3 distinct graphs each containing node's features in varying dimensions (64, 128 and 256), testing results are available in appendix D1. The evaluation criteria will be further described in the following sections of this paper, but it's important to outline the configuration that achieved best results within the trial process: the model comprises two layers of SAGEConv , for each node type, with a 'LSTM' aggregator. This configuration enables the model to aggregate features from neighboring nodes effectively.

Regarding the training steps, the model is initialized with input, hidden, and output feature dimensions for each node type. As the goal is to learn representations that capture the inherent structure and features of the graph without explicit labels, a cross entropy loss function was used for unsupervised learning. The Adam optimizer is employed with a learning rate of 0.001. Adam is known for its efficiency in handling large datasets and its adaptability in adjusting learning rates during training. Later, it undergoes training for a defined number of epochs, 10 in total. In each epoch, the model performs a forward pass where it computes embeddings for each node type. A backward pass is conducted where gradients are computed, and the model parameters are updated to minimize the loss. This iterative process continues for each epoch, with the loss being monitored to track the model's learning progress. Once training is complete,

the model is set to evaluation mode. A forward pass without gradient tracking is run to obtain the final embeddings for each node type. The output from the model includes embeddings for each node type, which are vital in capturing the structural and feature-based information of the nodes.

These embeddings are crucial for downstream tasks like clustering, visualization, and as input features for other predictive models. The learned embeddings provide deep insights into the relationships and significance of various entities in the e-commerce domain. They can be used to identify competitive clusters, measure centrality, and perform various analyses that contribute to understanding the competitive dynamics in e-commerce, as detailed in the following sections.

3.3 Cluster Analysis and Visualization

In this phase, the project focused on clustering and visualizing the domain embeddings derived from the GraphSAGE. To do so, the code relies on using Principal Component Analysis (PCA) to reduce the dimensions of domain embeddings while retaining a significant portion of the variance (target explained variance set at 95%), this step was chosen aiming to simplify the data without losing essential information and making it manageable for subsequent clustering and visualization. Uniform Manifold Approximation and Projection (UMAP), another

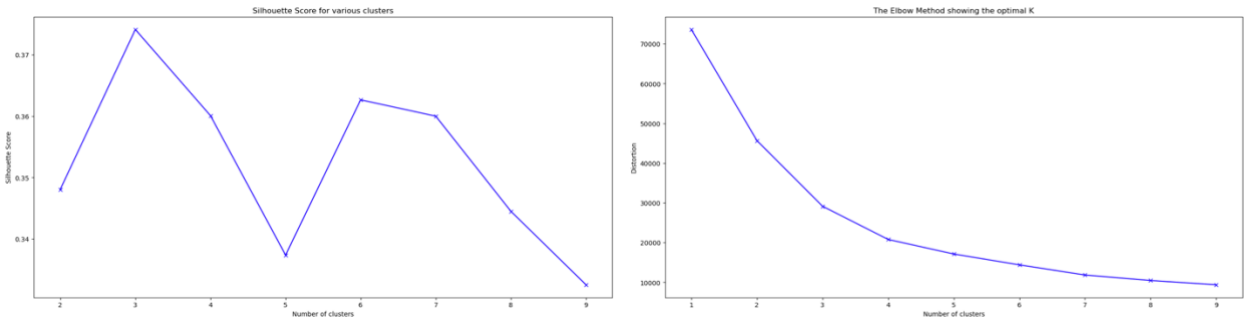


Figure 3.1 – Evaluation for optimal number of clusters. Silhouette score on the left and elbow method on the right indicates that the k-means algorithm should be applied for 3 clusters.

dimensionality reduction technique, was applied to the PCA-transformed embeddings to further reduce them to two dimensions. This two-step process helps in balancing the preservation of global data structure (achieved by PCA) and the exploration of local groupings or clusters (achieved by UMAP), providing a more comprehensive view of the data.

The number of clusters for K-means was set to three, but this was not an arbitrary choice. Specific evaluation methods as described earlier on this report, namely the elbow method and silhouette scores were applied, results on figure 3.1. K-means clustering was then applied to the 2D embeddings from UMAP, grouping similar domains into distinct clusters. A scatter plot was created to visualize the 2D embeddings post-UMAP reduction. The plot used different colors to represent various clusters, providing a clear visual distinction between them, as demonstrated on figure 3.2.

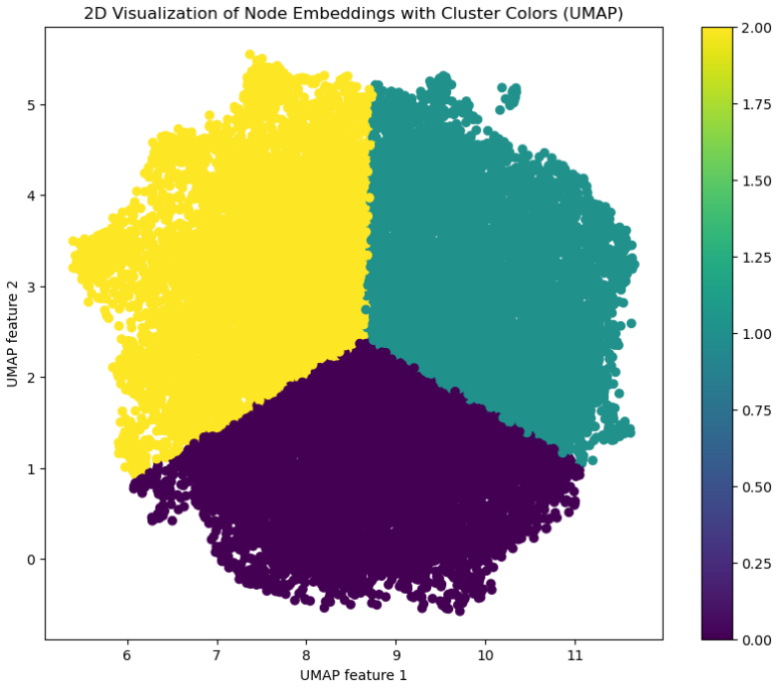


Figure 3.2 – Visualization of the domain's embeddings after dimensionality reduction with UMAP. Colors represent each cluster labels retrieved from k-means.

3.4 Node similarity analysis: Competitor Retrieval Algorithm

The Competitor Retrieval Algorithm enables the identification of competitive. This section will outline the algorithm's methodology and its application to node similarity analysis. The algorithm utilizes the embeddings which encapsulate the graph structure providing a rich representation of each node's position within the network, figure 3.3 translates the structure of the code. Cosine similarity is calculated between nodes to determine their resemblance. This measurement is based on the angle between the embedding vectors in the multidimensional space, offering a measurement of similarity.

Algorithm 1: Top-N Competitors Recommendation
Input:
Node index i of a given node (domain)
Expected number of recommendations n (where n < total number of nodes K)
Embeddings matrix E containing the vector representations of nodes
Output:
A sorted list of top n recommended nodes R with highest combined scores based on cosine similarity and competitive significance
Algorithm Steps:
1: $\vec{v}$ ← Retrieve the embedding vector E[i] for node i
2: If $\vec{v}$ is empty, then
3: Return an empty list
4: End if
5: Initialize an empty list G to store node names and similarity scores
6: For all embedding vectors $\vec{v}$_k in E do
7: If k is not i then
8: Retrieve the node name n_k corresponding to embedding $\vec{v}$_k
9: Compute cosine similarity cos_sim between $\vec{v}$ and $\vec{v}$_k
10: Calculate keyword overlap keyword_overlap as the intersection over the union of keywords between i and k, normalized by the number of keywords for k
11: Append (n_k, cos_sim, keyword_overlap) to list G
11: End if
12: End for
13: Sort list G in descending order by descending cosine similarity
14: Return the top n entries from list G

Figure 3,3 – Top-N algorithm for competitor retrieval

It's observed that applying PCA to the embeddings results in a broader range of similarity scores, enhancing the algorithm's capability to differentiate between nodes. Post-PCA embeddings tend to have a wider distribution of similarity scores, which is essential for a more granular analysis of node similarity. This transformation allows for a more differentiated and nuanced understanding of the nodes' relationships. According to Gerling, this happens as result

of the normalization due to the dimensionality reduction on the domains similarities thus enhancing the interpretability of the relations (Gerling, 2023).

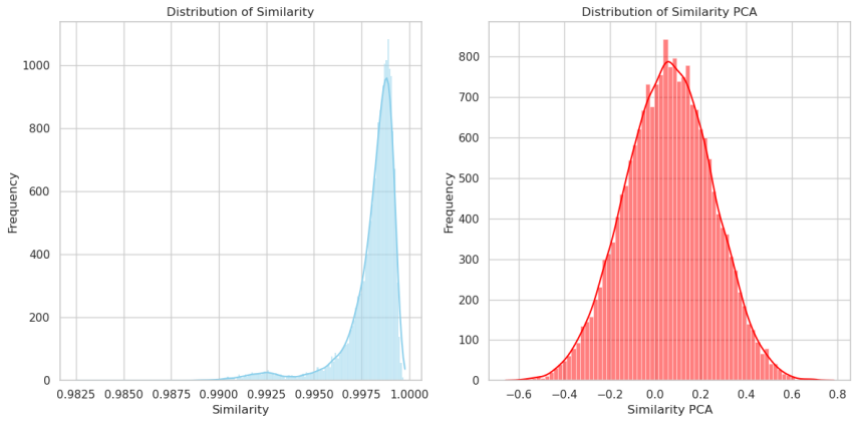


Figure 3.4 – Distribution of cosine similarities for a given domain. On the left, the similarities for 64 dimensions vector. On the right, the cosine similarities on the same embeddings after PCA set to preserve 0.95% of significance.

For each of the domains, the algorithm considers the shared keywords as a portfolio overlap. The algorithm produces a dataset containing detailed records about the nodes, including the node of interest, the most related nodes (competitors), their similarity scores and overlap metrics. Serving as a nuanced analytical tool, the algorithm enhances e-commerce competitive analysis by integrating advanced machine learning techniques with traditional metrics.

D.3.5 Node similarity analysis: Keyword Recommendation System

The Keyword Recommendation System is a sophisticated component that leverages node similarity analysis to suggest strategic keywords for e-commerce platforms, figure 3.5 demonstrates the structure of it. This system uses the advanced embeddings and similarity measures to identify keywords that could potentially improve search visibility and user engagement or better allocation of advertising resources.

Algorithm 3: Keyword Recommendation for a Domain

Input:

domain: The specific domain for which the keyword recommendations are being made.
keywords: A list of m keywords associated with the domain.
top_n: The number of similar keywords to recommend.

Output:

A list of top n similar keywords for each keyword in the domain's portfolio.

Algorithm Steps:

1: Initialize an empty aggregate vector $\vec{p}$
2: For each keyword fk in the domain's keywords portfolio do
3: $\vec{f}\_k \leftarrow$ Retrieve the embedding vector for keyword fk
4: $\vec{p} \leftarrow \vec{p} + \vec{f}\_k$
5: End for
6: Normalize $\vec{p}$ by dividing it by m (number of keywords in the portfolio)
7: If $\vec{p}$ is empty, then return an empty list
8: Initialize an empty list G to store recommended keywords and their shared portions
9: For each keyword fk in the general keyword pool do
10: Compute cosine similarity cos_sim between $\vec{p}$ and the embedding of fk
11: Retrieve the shared portion shared_portion between fk and the domain's keywords
12: Append (fk, cos_sim, shared_portion) to list G
13: End for
14: Sort list G in descending order by cosine similarity
15: Return the top n entries from list G

Figure 3.5 – Top-N algorithm for keywords recommendation

Cosine similarity scores are computed between the embeddings of the keywords to find the most similar keywords based on the graph's multidimensional space. These scores are crucial for determining the closeness of keywords in the context of the graph, which is expected to translate not only the semantic relation among terms but also the structural information of the graph. The system ranks the keywords based on the calculated similarity scores thus keywords with higher similarity scores are considered more relevant and are prioritized in the recommendation list.

Table 3.1 brings the output of the tool for a set of keywords. To validate the recommendations, the system performs an overlap analysis that assesses the shared portion of the recommended keywords within the set of terms of the given domain. The goal of the analysis is intended to enhance the existing keyword portfolio of a domain, as it informs if there is potential for increasing investments for some terms or even to decrease investments for terms that might be saturated. They are selected to enlighten both breadth and depth in keyword strategy. By utilizing the recommended keywords, e-commerce platforms can make informed decisions

about which keywords to target for SEO and advertising campaigns, ultimately aiming to improve their visibility and performance in search engines.

Table 3.1- Keyword recommender results for 3 distinct keywords for Adidas, H&M, Zalando and Nike

Domain	Original Keyword	Recommended Keyword	Shared Portion
adidas	adidas freerider pro	'adidas freerider pro'; 'adidas 510 freerider pro'; 'adidas five ten freerider pro' 'adidas 510 freerider'; 'adidas 5 10 freerider'	0.8
adidas	adidas tubular doom sock	'adidas tubular doom sock'; 'adidas tubular doom sock p'; 'adidas tubular doom'; 'adidas tubular high top'; 'adidas tubular doom sock primeknit'	0.4
adidas	mickey mouse addidas	'mickey mouse addidas'; 'mickey mouse adidas'; 'minnie mouse adidas'; 'mickey adidas'; 'disney minnie mouse adidas'	1
hm	damen pullover mit hundemotiv	'damen pullover mit hundemotiv'; 'pullover mit hundemotiv'; 'pullover hundemotiv'; 'wikinger pullover herren'; 'pullover mit pferdemotiv'	0.2
hm	v neck cotton blouse	'v neck cotton blouse'; 'sleeveless cotton blouse'; 'cotton sleeveless blouse'; 'silk cotton blouse'; 'cotton gauze blouse'	0.6
hm	v blouse	'v blouse'; 'o blouse'; 'no blouse'; '90s blouse'; 'h m blouse'	0.2
zalando	megadeth shirt	'megadeth shirt'; 'megadeth t shirt'; 'megadeth t-shirt'; 'megadeth lamb of god shirt'; 'megadeth hoodie'	0.2
zalando	bear air sweatshirt	'bear air sweatshirt'; 'bear sweatshirt'; 'bel air sweatshirt'; 'penn law sweatshirt'; 'iowa sweatshirt'	0.2
zalando	tfnc green dress	'tfnc green dress'; 'pea green dress'; 'hm green dress'; 'plt green dress'; 'h&m green dress'	0.2
nike	jordan 1 hyperroyal	'jordan 1 hyperroyal'; 'jordan hyperroyal'; 'jordan1 hyper royal'; 'jordan 1 hyper royal'; 'jordan one hyper royal'	1
nike	jordan 11 ie black cement	'jordan 11 ie black cement' 'jordan 1 black cement'; 'jordan 8 black cement' 'jordan 11 low ie black cement'; 'jordan 4 black cement'	0.4
nike	nike vapormax chukka slip	'nike vapormax chukka slip'; 'nike air vapormax chukka slip'; 'vapormax chukka slip'; 'nike vapormax pink oxford'; 'nike vapormax plus pink oxford'	1

As demonstrated on Table 3.2, the strategy behind each competitor set of keywords is different, for instance Nike relies not only on individual terms but also on the ones similar to them whereas Zalando does not bid as much on the recommended terms.

Table 3.2 – Descriptive statistics for Shared Portion metric on competitors Keywords

Domain	Mean	Std Dev	Min	0.25	0.50	0.75	Max
adidas	57%	27%	20%	40%	60%	80%	100%
hm	49%	24%	20%	20%	40%	60%	100%
nike	64%	26%	20%	40%	60%	80%	100%
zalando	38%	22%	20%	20%	40%	40%	100%

3.6 Evaluation Metrics and Results

The evaluation of the competitor identification algorithm is multi-faceted, combining quantitative measures with qualitative insights. It not only assesses the algorithm's ability to identify competitors accurately but also evaluates the meaningfulness and business relevance of these identifications. To address the task, we assume that competitors within the context of this research would most likely be the ones that bids on a same search term. Thus, the keyword overlap metric is used to measure the extent to which two domains are competing directly with each other based on their search engine optimization (SEO) efforts. This metric is calculated by determining the proportion of shared keywords between any two given domains.

For each domain, the set of associated keywords is identified. The keyword overlap between a domain and its potential competitors is then calculated by finding the intersection of their keyword sets. The degree of overlap is expressed as a percentage, representing the proportion of shared keywords relative to the total keyword set of a competitor. This percentage reflects how much of their SEO strategy overlaps, indicating direct competition in search engine results. To focus the analysis on truly competitive relationships and reduce noise, a subset of domains is selected based on their shared keyword profiles. A minimum threshold of 15% shared keywords is established. This means that only those domains that have at least an average of 15% of their keywords in common with the top 100 domains ranked on the overlap metric are considered for retrieving its competitors.

The final stage of the evaluation is to compare the set of competitors identified by the Top-N algorithm and the set of competitors based on the keywords overlap rule. By computing the intersection of both sets, it's possible to measure how effective the tool was to identify the competitors based on the assumptions developed for the evaluation.

3.7 Discussion of Results

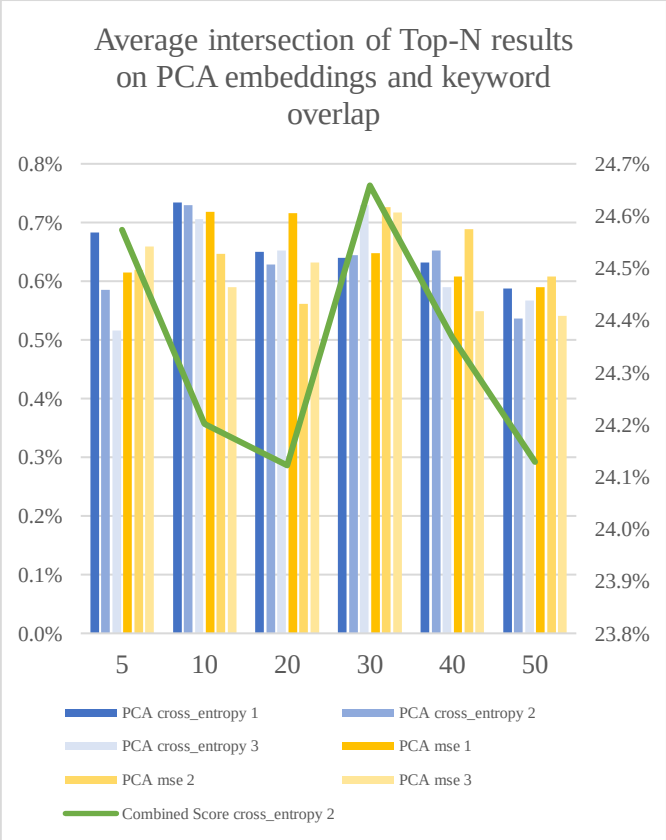


Figure 3.6 – Average intersection of the Top-N algorithm output and the Top-N competitors by keyword overlap. The green line represents the best model with the degree centrality combined with cosine similarity on the secondary axis.

The initial implementation of the model, which utilized cosine similarity measures derived from GNN embeddings, showed limited success in accurately identifying competitors. This was evident in the comparative analysis with the keyword overlap metric. The application of PCA to the embeddings brought about a notable improvement in differentiating the cosine similarities between nodes. This step enhanced the distinctiveness of the embeddings, leading to better separation of entities in the similarity calculation. However, despite this improvement, the overall performance in competitor identification remained unsatisfactory, results on figure 3.6.

A significant enhancement to the algorithm was the introduction of degree centrality as a weight in the similarity calculation. This approach is relatively novel in the literature and seeks to combine traditional network centrality measures with advanced machine learning outputs. The integration of degree centrality addresses the need to factor in the network influence of a domain, alongside its contextual similarity. Degree centrality reflects the number of direct connections a node has in the network, indicating its influence or prominence. The inclusion of

degree centrality as a weight in the similarity measure resulted in a more nuanced and accurate identification of competitors, aligning more closely with the competitive dynamics observed in the e-commerce sector.

Table 3.2 – Results for Top-3 competitors. On the left, the competitor's algorithm output for the combined score methodology on the algorithm. On the right side, the top-3 competitors based on the keyword overlap assumption.

Top-N Algorithm output (top 3)			Highest Keyword Overlap (top 3)		
domain	competitor_domain	keyword_overlap	domain	competitor_domain	keyword_overlap
adidas	stockx	8.4%	adidas	dickssportinggoods	16.3%
adidas	dickssportinggoods	16.3%	adidas	stockx	8.4%
adidas	kohls	4.8%	adidas	finishline	4.8%
hm	amazon	5.1%	hm	amazon	5.1%
hm	ebay	3.0%	hm	lulus	4.9%
hm	zalando	4.2%	hm	zalando	4.2%
zalando	amazon	9.1%	zalando	ebay	9.1%
zalando	ebay	9.1%	zalando	amazon	9.1%
zalando	nordstrom	3.3%	zalando	office	4.0%
nike	otto	0.6%	nike	shoecarnival	2.3%
nike	rei	1.7%	nike	rei	1.7%
nike	redbubble	0.0%	nike	runningwarehouse	0.8%

4: Conclusion and Future Work

4.1 Summary of Findings and Contributions to the Field

The study applied Graph Neural Networks (GNNs) and advanced machine learning techniques, to the e-commerce domain, providing significant insights into the competitive landscape. The findings include handling of high-dimensional data through PCA and UMAP for visualization, cluster analysis, the effective use of GNNs for node representation, the development of a Competitor Retrieval Algorithm that leverages node similarity for identifying competitors, and a Keyword Recommendation System that suggests strategic keywords for search engine optimization and advertising. Providing a methodological approach for implementing these algorithms in a real-world e-commerce setting

4.2 Limitations of the Study and Suggestions for Future Research

While the study achieved its primary objectives, it acknowledges certain limitations:

The model's performance was constrained by the availability and quality of the data. The absence of a unique identifier to a same product across the market (e.g.: EAN) limits the capability of the network structure to fully translate the relationship of products. The absence of true labels bounds the evaluation of the model performance, what limited approach to domain expertise. There is a significant opportunity for future research to explore the effects and implications of combining centrality measures with cosine similarity. Understanding how this integration influences the identification of competitors and their strategic implications is crucial. Future studies could focus on analyzing the performance or returns of keywords with high or low shared portions. Such research could provide deeper insights into the effectiveness of keywords in driving traffic and conversions, leading to more informed SEO and marketing strategies.

4.3 Conclusion

The research presented in this thesis demonstrates the potential of GNNs in transforming e-commerce analytics. The Competitor Retrieval Algorithm and the Keyword Recommendation System are poised to become valuable tools for e-commerce platforms. The study also opens up several avenues for future research, promising further advancements in the field. Through continuous innovation and exploration, the intersection of machine learning and e-commerce analytics will likely yield even more powerful insights and tools for businesses.

Bibliography

Battiston, Stefano, Michelangelo Puliga, Rahul Kaushik, Paolo Tasca, and Guido Caldarelli. 2012. "Debtrank: Too central to fail? financial networks, the fed and systemic risk." *Scientific reports* (Nature Publishing Group UK London) 2: 541.

Becht, Etienne, Leland McInnes, John Healy, Charles-Antoine Dutertre, Immanuel W H Kwok, Lai Ng Ng, Florent Ginhoux, and Evan W Newell. 2019. *Dimensionality reduction for visualizing single-cell data using UMAP.* Nature Biotechnology.

Bhattacharya, Siddharth , Jing Gong, and Sunil Wattal. 2022. “Competitive Poaching in Search Advertising: Two Randomized Field Experiments.” *Information Systems Research* 599–619.

BOF Insights. 2023. *The Evolving Art of Luxury Experiential Retail.* Business of Fashion.

Bojanowski, Piotr, Armand Joulin, and Tomas Mikolov. 2016. *Meta.* August 18. https://research.facebook.com/blog/2016/8/fasttext/.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. "Enriching word vectors with subword information." *Transactions of the association for computational linguistics* (MIT Press One Rogers Street) 5: 135-146.

Cao, Lele, et al. 2023. *"CompanyKG: A Large-Scale Heterogeneous Graph for Company Similarity Quantification."*. arXiv preprint arXiv:2306.10649,. https://arxiv.org/abs/2306.10649.

Chaffey, Dave , and Fiona Ellis-Chadwick. 2019. *Digital Marketing: Strategy, Implementation & Practice.* Pearson UK.

Fail (Management of Innovation and Change). Boston: Harvard Business Review.

Elrod, Terry, Gary J. Russell, Allan D. Shocker, Rick L. Andrews, Lynd Bacon, Barry L. Bayus, J. Douglas Carroll, et al. 2002. *"Inferring Market Structure from Customer Response to Competing and Complementary Products."*. Marketing Letters 13, no. 3: 221–3.

Erdmann, Anett, Ramón Arilla, and José M. Ponzoa. 2022. "Search Engine Optimization: The Long-Term Strategy of Keyword Choice." *Journal of Business research.*

Fang, Fang, Kaushik Dutta, and Anindya Datta. 2013. *"Lda-based industry classification."* .

Fortunato, Santo. 2010. "Community Detection in Graphs." *Physics Reports* 486: 75–174.

Gabel, S., Guhl, D., and Klapper, D. 2019. *"P2V-MAP: Mapping Market Structures for Large Retail Assortments."* . Journal of Marketing Research, vol. 56, no. 4, pp. 557-580. https://doi.org/10.1177/0022243719833631.

Gerling, Christopher. 2023. *"Company2Vec — German Company Embeddings Based on Corporate Websites."* . International Journal of Information Technology & Decision Making, pp. 1–35. World Scientific Pub Co Pte Ltd. http://dx.doi.org/10.1142/s0219622023500694.

Gofman, Michael. 2017. "Efficiency and stability of a financial architecture with too-interconnected-to-fail institutions." *Journal of Financial Economics* (Elsevier) 124: 113--146.

Google for Developers (b). 2023. *Google Search Essentials (Formerly Webmaster Guidelines).* December. Accessed December 2023. https://developers.google.com/search/docs/essentials.

Google for Developers. 2023. *SEO Starter Guide: The Basics.* December. Accessed December 2023. https://developers.google.com/search/docs/fundamentals/seo-starter-guide.

Grips. 2023. Accessed 2023. https://gripsintelligence.com/data-methodology.

—. 2023. https://gripsintelligence.com/about.

Hamilton, W. L., Ying, R., and Leskovec, J. 2018. *"Inductive Representation Learning on Large Graphs."* . arXiv preprint arXiv:1706.02216, . https://arxiv.org/abs/1706.02216.

Han, Qiwei, Pedro Ferreira, and João Paulo Costeira. 2016. "Asymmetric Peer Influence in Smartphone Adoption in a Large Mobile Network." *arXiv.*

Hee Park, Chang, and Manoj K. Agarwal. 2018. “The Order Effect of Advertisers on Consumer Search Behavior in Sponsored Search Markets.” *Journal of Business Research.*

Kargar, M. 2022. *"Search System over e-Commerce Data for Business Users,"* . Istanbul, Turkey: International Conference on Engineering & MIS (ICEMIS), pp. 1-5, doi: 10.1109/ICEMIS56295.2022.9914107.

Kumar, V, and Werner Reinartz. 2012. *Customer Relationship Management: Concept, Strategy, and Tools.* Springer.

Li, Z., et al. 2020. *"Hierarchical Bipartite Graph Neural Networks: Towards Large-Scale E-commerce Applications."* . IEEE 36th International Conference on Data Engineering (ICDE), pp. 1677-1688. https://doi.org/10.1109/ICDE48307.2020.00149.

Lopes Rolim, Lucas. 2022. *"Embedding of Bipartite Graphs via Graph Neural Networks with Application to User-Item Recommendations."* . UFRJ/COPPE. https://www.cos.ufrj.br/uploadfile/publicacao/3046.pdf.

McInnes, Leland, John Healy, and James Melville. 2020. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction."

Mehlhorn, Kurt, and Peter Sanders. 2008. "Shortest paths." In *Algorithms and Data Structures: The Basic Toolbox*, by Kurt and Sanders, Peter Mehlhorn, 191-215. Springer.

Meng, Lei, Ah-hwee Tan, and Donald C Wunsch. 2019. *Clustering and its extensions in the social media domain.* Singapore Management University.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. "Distributed Representations of Words and Phrases and their Compositionality."

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. "Efficient estimation of word representations in vector space." *arXiv preprint arXiv:1301.3781.*

Ologunebi, John, and Ebenezer Obafemi Taiwo. 2023. “The Importance of SEO and SEM in improving brand visibility in E-commerce industry; A study of Decathlon, Amazon and ASOS.” *MPRA Paper* (University Library of Munich, Germany).

PwC. 2023. *IAB.* 12 April. Accessed November 2023. https://www.iab.com/insights/internet-advertising-revenue-report-full-year-2022/.

Rosvall, Martin, and Carl T. Bergstrom. 2008. “Maps of Random Walks on Complex Networks Reveal Community Structure.” *Proceedings of the National Academy of Sciences of the United States of America* 1118–23.

Rousseeuw, Peter J. 1987. "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis."

Vinutha, M. S., and M. R. Prajwal. 2023. “A Survey on Search Engine Optimization-Types, Techniques and Factors.” *International Journal of All Research Education and Scientific Methods (IJARESM)* 1933-1938.

Yin, Lei, and Ruodan Lu. 2017. *Automated Competitor Analysis Using Big Data Analytics: Evidence from the Fitness Mobile App Business.* Business Process Management Journal.

Zhou, Chen. 2009. "Are banks too big to fail? Measuring systemic importance of financial institutions." *Measuring Systemic Importance of Financial Institutions.*

Zhou, Jie, et al. 2020. *"Graph Neural Networks: A Review of Methods and Applications."* . AI Open, vol. 1, pp. 57-81. ScienceDirect, https://doi.org/10.1016/j.aiopen.2021.01.001.

Appendix

Appendix I: Internet Advertising Revenue trend (2020-2022)

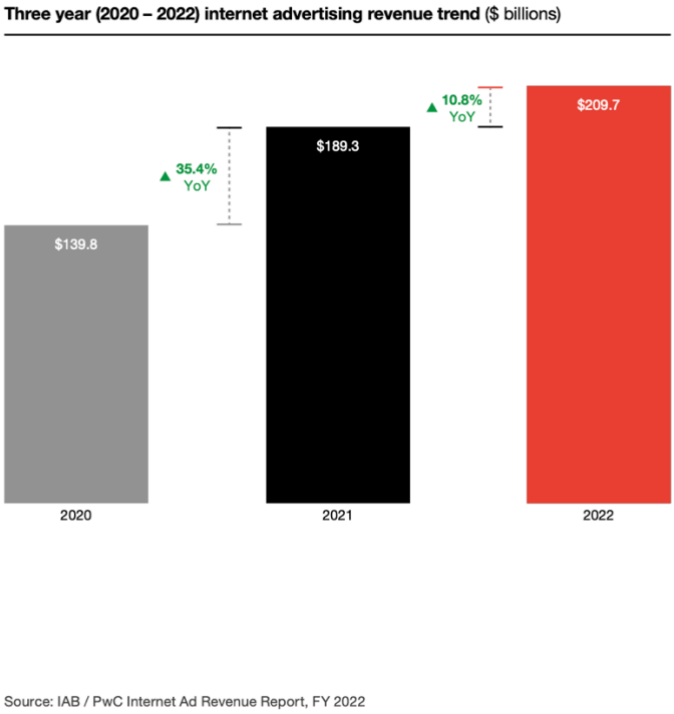


Figure 5 - Internet Advertising Revenue. Source: IAB/PwC Internet Ad Revenue report, FY 2022

Appendix II: Data Dictionary

Variable	Data Type	Description	Source
Keyword	Categorical variable	Stands for the term inserted in the search engine related to products that are shown. The column contains one or more keywords.	Grips
Domain	Categorical variable	Indicates the “top level domain of the website”.	Grips
Line_ID	Numerical Value	Used for identification of the product, however not standardized across retailers.	Grips
URL	String variable	Represents the web address of the specific product.	Grips
Position	Numerical Variable	When a specific term is searched, this feature shows the product URL's rank or position on the Search Engine Results Page (SERP). The position is represented by a positive number, where "1" is the SERP's top result.	Grips
Branded	Numerical Variable	Represented as an integer indicating the quantity of keywords that are connected to a particular brand.	Grips
CPC	Numerical continuous variable	Quantifies an estimated amount to be spent by brands/suppliers/retailers whenever a person clicks on the link showcased by the search engine.	Grips
Low_top_of_page_bid	Numerical variables	Represents the lower boundary for the amount bidders are willing to pay for clicks based on specific variables.	Keyword Planner (Google)
High_top_of_page_bid	Numerical variables	Numerical variables that represent the upper boundary for the amount bidders are willing to pay for clicks based on specific variables.	Keyword Planner (Google)
Avg_monthly_search_volume	Continuous variable	Average number of times the keyword is searched per month.	Keyword Planner (Google)
Product Title	String variable	The title given to a product.	Scraped from metadata
Brand	String variable	Name of the brand	Scraped from metadata
Price (USD)	Continuous Numerical variable	Price of the product sold	Scraped from metadata
Category	Categorical variable, string	Product hierarchy of the respective product	Prediction by Grips
Predicted Products Sold	Continuous numerical variable	Serves for the estimation of sales of the product under the given keyword.	Grips
Impressions	Continuous variable	Represents a monthly estimation of products sold.	Grips

Appendix III: Data Cleaning

Variable	Percentage of missing values
low_top_of_page_bid	47.2%
high_top_of_page_bid	47.2%
brand	30.8%
category	11.1%
avg_monthly_search_volume	2.6%
predicted_productssold	0.1%
impressions	0.05%
producttitle	0.05%

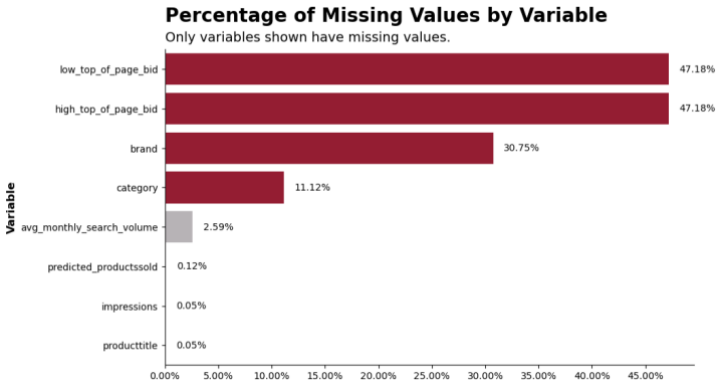
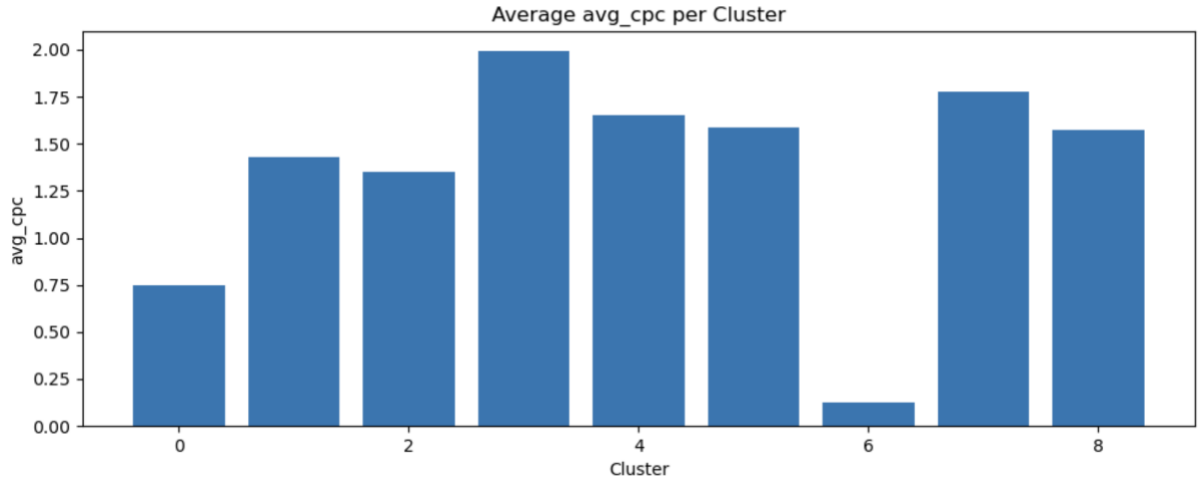


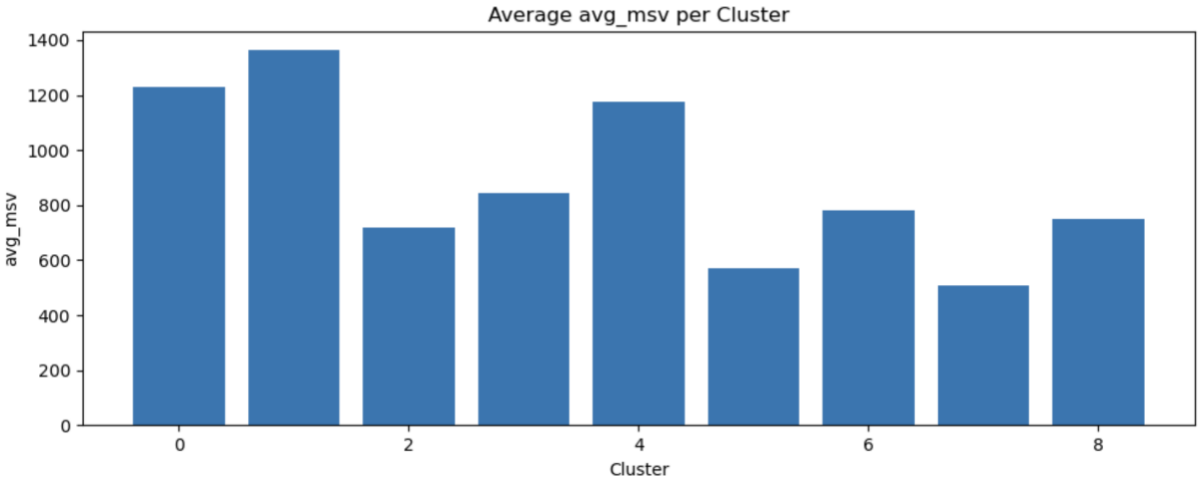
Figure 6 - % of Missing Values within dataset

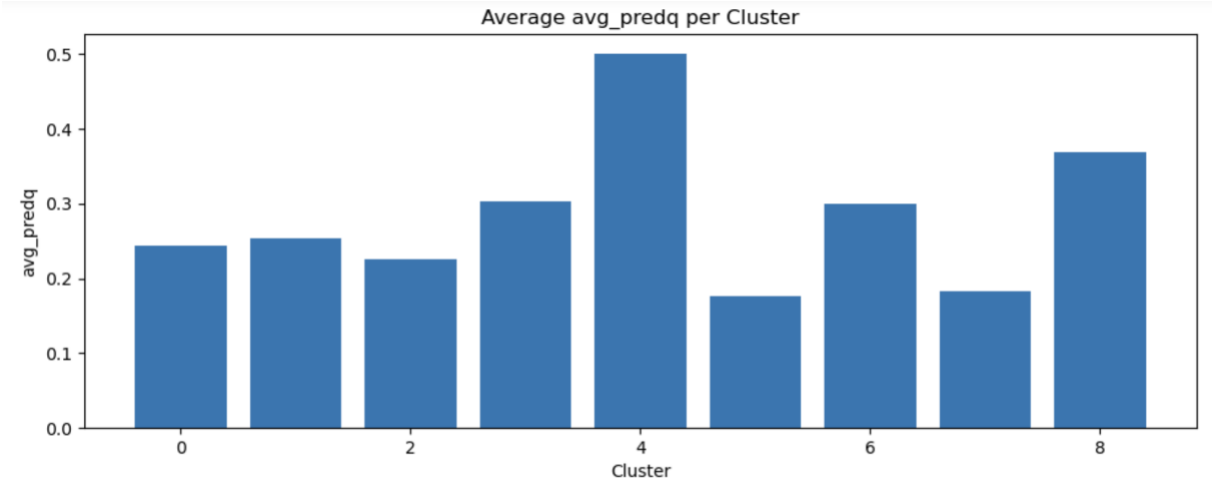
Appendix IV: Overview of clusters, selected domains, and top categories

Cluster	Selected Domains	Top Categories
0	StockX, Lego, Scrapbook, Bigbadtoystore, …	Art, Entertainment, Medium, Hobby, …
1	HP, Samsung, Canon, Asus, Dell, Apple, …	Accessory, Electronics, Part, …
2	Vans, Oakley, Wilson, Ping, Burton, Titleist, Evo, …	Sporting, Good, Outdoor, …
3	Chanel, Olay, Dermalogica, Skinceuticals, …	Health, Care, Beauty, Personal, …
4	Tesco, Nespresso, Sodastream, Totalwine, …	Food, Beverage, Tobacco, Item, …
5	Pandora, Cartier, Tiffany, Swarowski, …	Apparel, Jewelry, Accessory, Necklace, …
6	Ikea, Petsafe, Bauhaus, Ebay, Mediamarkt, …	Accessory, Medium, Supply, Garden, …
7	Nike, Adidas, Converse, Ugg, Birkenstock, …	Accessory, Apparel, Clothing, …
8	Kitchenaid, Weber, Dyson, Walmart, …	Garden, Home, Accessory, Kitchen, …

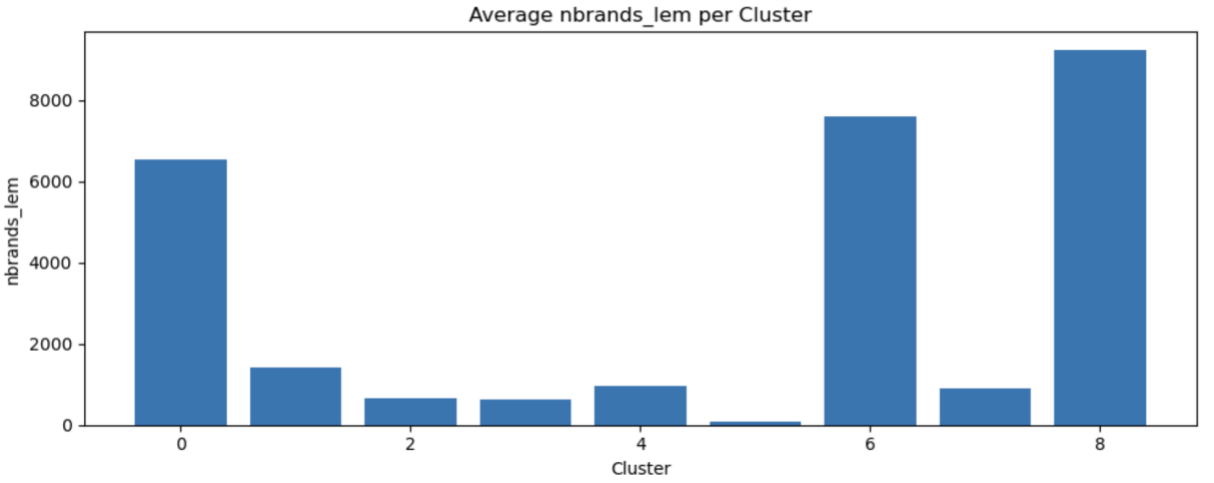
Appendix V: Average numerical statistics per cluster

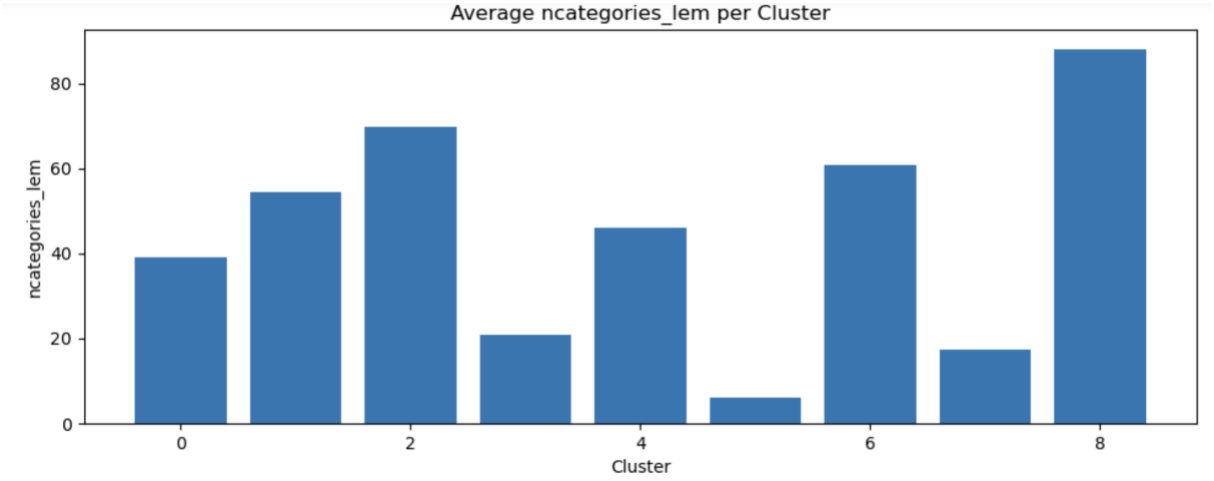


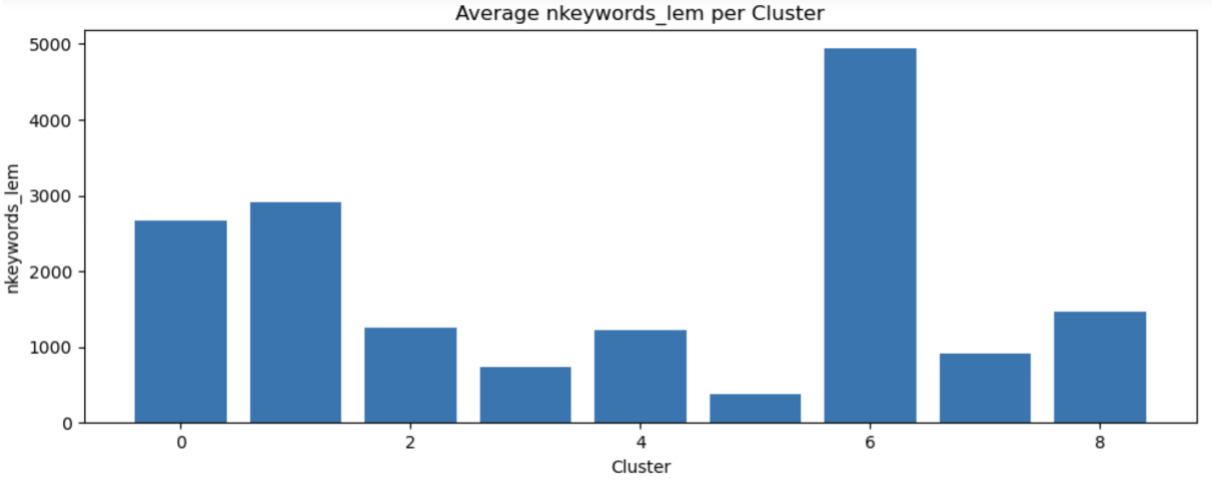


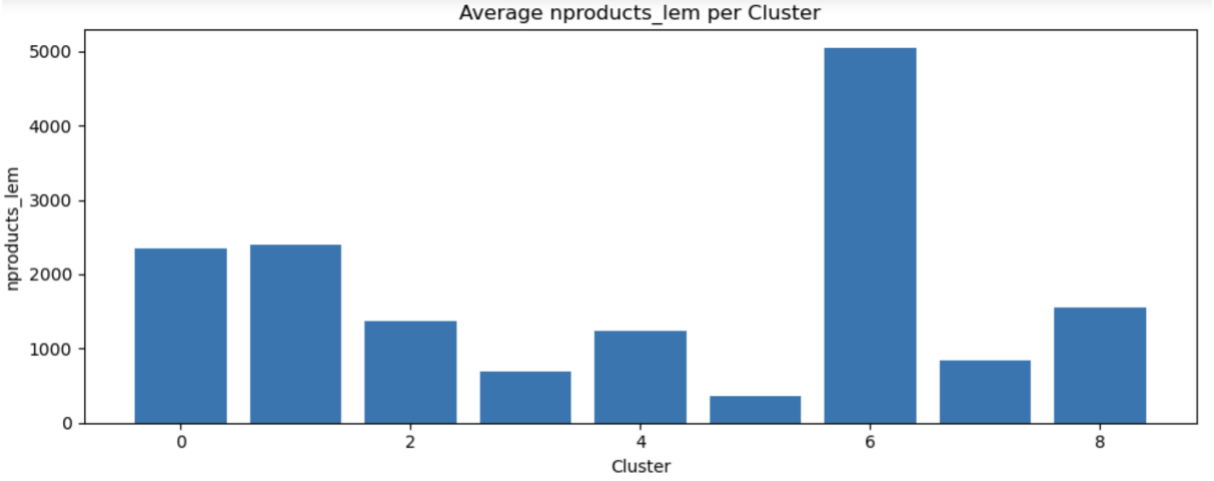

Average avg_predq per Cluster
avg_predq
0.0
0.1
0.2
0.3
0.4
0.5
0
2
4
6
8
Cluster


Average avg_imps per Cluster
avg_imps
0
25
50
75
100
125
150
175
0
2
4
6
8
Cluster


Average nbrands_lem per Cluster
nbrands_lem
0
2000
4000
6000
8000
0
2
4
6
8
Cluster


Average ncategories_lem per Cluster
ncategories_lem
0
20
40
60
80
0
2
4
6
8
Cluster


Average nkeywords_lem per Cluster
nkeywords_lem
0
1000
2000
3000
4000
5000
0
2
4
6
8
Cluster


Average nproducts_lem per Cluster
nproducts_lem
0
1000
2000
3000
4000
5000
0
2
4
6
8
Cluster

Appendix VI: Filtered Derived Retailer Names

Category	Example Retailer Name
Numeric Keywords	220, 1928, 686, 8020, 925
Common Adjectives/Adverbs	did, in, one, only, no, easy, my, at, very
Common Product Names	keyboard, spoon, stethoscope, oven, ring, canoe, headphones, thermometer, fridge, glasses, wigs
Activity/Sport Related	soccer, browning, pilates, bowling
Geographical Locations	uk, eu
Currency Names	euro, dollar
Religious/Spiritual Terms	christ, dame, angel, catholic
Animals and Personas	teenage, animale, corgi, frog, rhino, singer, seals, bird

Appendix VII: Advertisement Poaching Network Sparsity

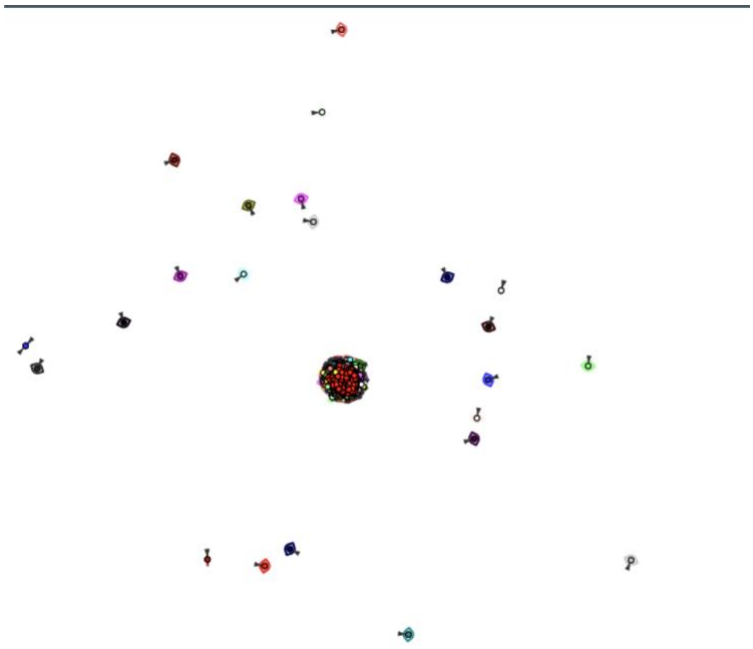


Figure 7 - Visual Representation Of Network Sparsity

Appendix D1: GNN architecture parameters evaluation

Function	loss	Layer	epoch	intersection_pca_overlap	intersection_combined_overlap	intersection_similarities_overlap
		1	5	0.60%	23.21%	0.50%
		2	5	0.63%	24.50%	0.48%
		3	5	0.59%	24.98%	0.41%
		1	10	0.77%	24.93%	0.65%
	cross_entropy	**2**	**10**	**0.83%**	**24.45%**	**0.57%**
		3	10	0.72%	24.40%	0.56%
		1	20	0.67%	25.70%	0.45%
		2	20	0.79%	24.37%	0.55%
lstm		3	20	0.63%	23.37%	0.51%
		1	5	0.41%	25.70%	0.37%
		2	5	0.76%	24.24%	0.50%
		3	5	0.71%	24.55%	0.49%
		1	10	0.84%	24.52%	0.76%
	mse	2	10	0.72%	25.05%	0.50%
		3	10	0.67%	23.09%	0.57%
		1	20	0.62%	24.46%	0.52%
		2	20	0.56%	24.54%	0.43%
		3	20	0.72%	24.88%	0.54%
		1	5	0.84%	25.04%	0.60%
		2	5	0.63%	23.87%	0.55%
		3	5	0.60%	23.78%	0.55%
mean	mse	1	10	0.78%	24.40%	0.59%
		2	10	0.74%	24.88%	0.49%
		1	20	0.76%	24.46%	0.55%
		2	20	0.51%	24.72%	0.40%
		1	5	0.84%	24.48%	0.66%
		2	5	0.68%	24.65%	0.54%
		3	5	0.52%	23.49%	0.43%
		1	10	0.77%	24.11%	0.66%
	cross_entropy	2	10	0.72%	23.95%	0.61%
		3	10	0.71%	23.98%	0.46%
		1	20	0.65%	24.01%	0.50%
		2	20	0.54%	23.88%	0.49%
pool		3	20	0.72%	24.67%	0.62%
		1	5	0.67%	24.22%	0.61%
		2	5	0.48%	23.05%	0.41%
		3	5	0.72%	24.12%	0.45%
		1	10	0.74%	23.91%	0.50%
	mse	2	10	0.66%	24.94%	0.52%
		3	10	0.59%	25.06%	0.49%
		1	20	0.94%	22.99%	0.62%
		2	20	0.71%	22.29%	0.43%
		3	20	0.65%	24.18%	0.50%