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.

Website Networks and Keyword Distances

Uncovering Underlying Dynamics Among Competitors in the Digital World

Diego García Rieckhof (53046)

Work project carried out under the supervision of:

Qiwei Han, Alessandro Gambetti, Maximilian Kaiser

20/12/2023

Abstract:

This study explores the intricate web of competition in e-commerce, utilizing advanced methods in natural language processing and machine learning. This analysis delves into previously undiscovered competitor networks, deciphering apparent and concealed competition using advanced techniques such as Word2Vec and Graph Neural Networks. The research redefines competitor identification in the digital era, emphasizing the importance of search engine visibility. Leveraging Minimum Spanning Trees (MSTs) and network analysis, it offers insights into digital competition dynamics. Traditional methods fall short in capturing this complexity, necessitating a multi-step analytical framework. Findings reveal a hierarchical relationship among competitors, at the same time providing insight into competition intensity and connectivity. The analysis uncovers both direct and indirect competitors, enabling businesses to refine their strategies based on semantic analysis to improve their positioning across search engine rankings.

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

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 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 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, Guhl and Klapper 2019).

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 (Kök, 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, Guhl and Klapper 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, Dutta and Datta 2013), 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 to identify minute distinctions and parallels among them. The method, which is vital for acquiring customers and selecting suppliers, is enhanced by the semantic vector representations that (Mikolov, et al. 2013) 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. (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 to maintain uniformity and minimize noise. The final dataset is aggregated per domain containing the 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 (Joulin, et al. 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. (Joulin, et al. 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 and 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 and 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.

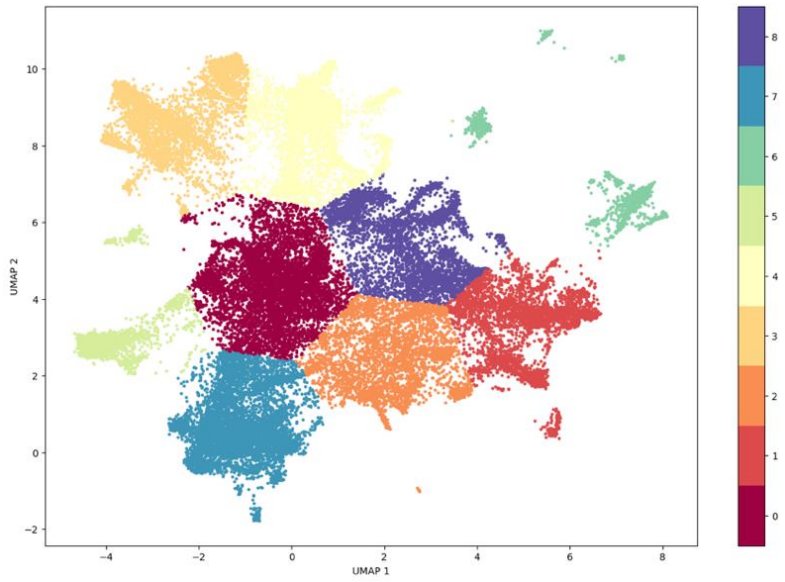


Figure 1 - Domain Clusters

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.

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.

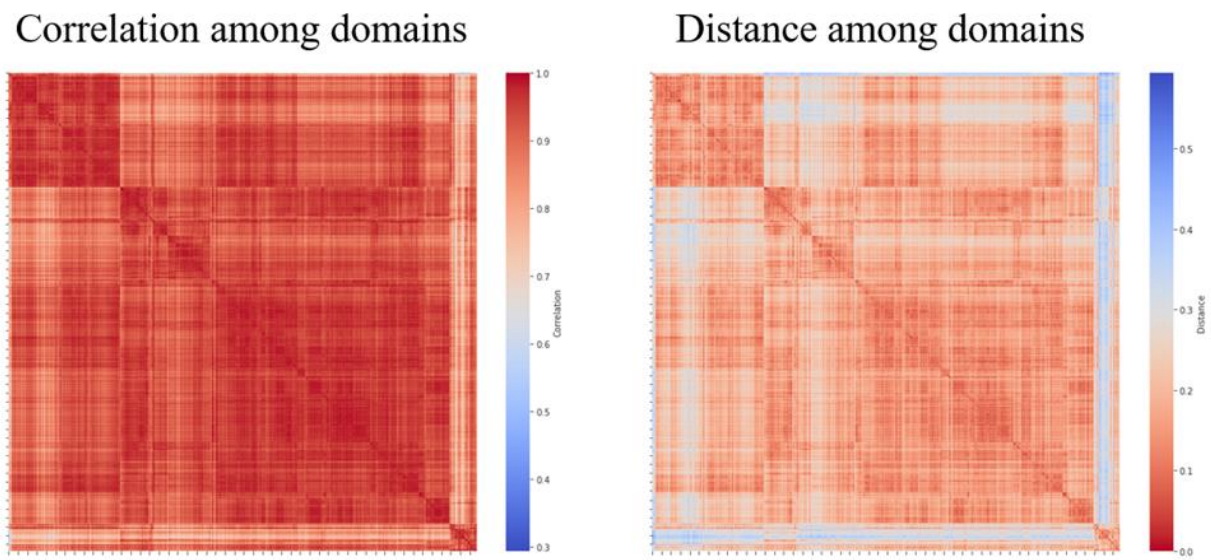


Figure 2 - Correlation vs. Distance Matrix among Domains

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.

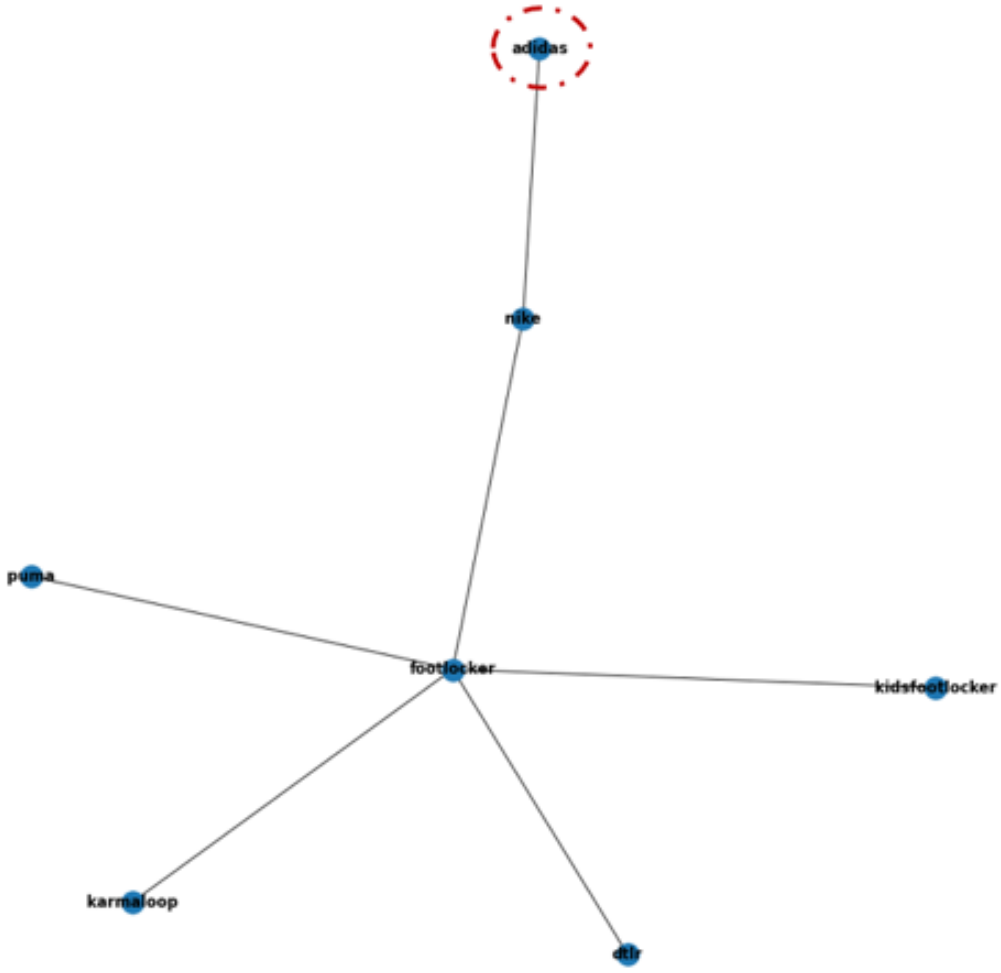


Figure 3 - Potential direct and indirect competitors for Adidas

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. 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.

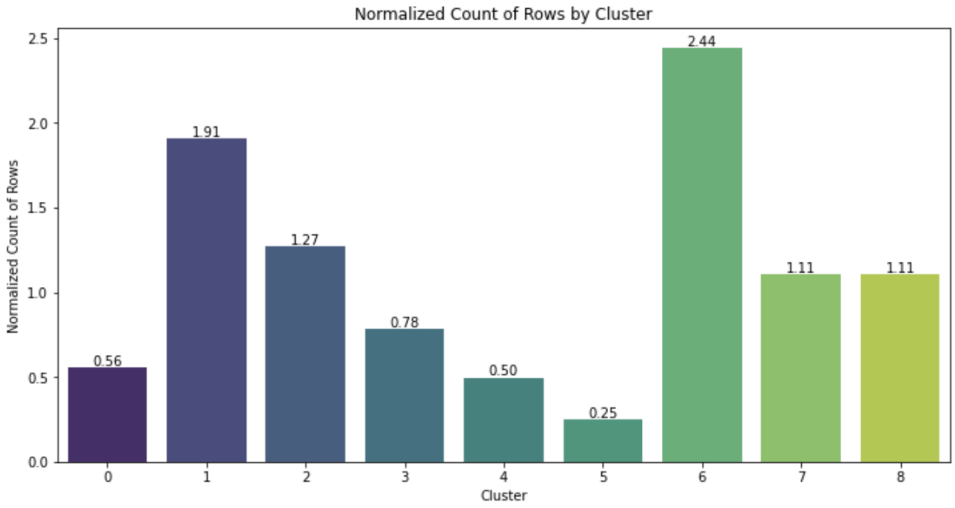


Figure 4: Counts of Advertisement Poaching by Cluster (normalized by cluster size)

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. 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, 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 case. There are a few methods for determining the optimal number of clusters in each 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 each 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 strategies. 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. Both methodologies will dig deeper into different nuances that can be used to assess and craft better strategies for different companies and their positioning in the Search Engine Ranking.

The research found a dynamic interplay in advertisement poaching among e-commerce competitors, heavily influenced by aggregator retailers. It identified industry-specific poaching tactics within and across different retail clusters. The use of competitor names in SEO campaigns suggested intense competition within clusters and varied tactics across them. Notably, frequent targeting of certain retailers indicated their market centrality. Network analysis using Community Infomap algorithm, referencing original clusters from FastText embeddings and UMAP, showed moderate to diverse advertisement poaching strategies, with varying ARI scores indicating complexity and individualistic tactics in the online retail sector. This complexity is further highlighted by the impact of network sparsity, bidirectional

connections, and data filtering on the detection of poaching patterns, underscoring the strategic nature of SEO poaching.

As a final step, the study explores an innovative approach to leverage the creation of knowledge graphs and Graph Neural networks to address the proposal of this research. The methodology proposes to combine the structural features of the graph into single embeddings which can be used not only to retrieve competitors, but also to recommend keywords for better management of advertising strategies.

In addressing the research question 'How does competition among online retailers emerge in the dynamics derived from organic search engine optimization?', this study comprehensively demonstrates that the intricacies of SEO metrics, such as keyword usage, search volume, and domain interaction patterns, are pivotal in unveiling the competitive landscape of online retail. It extends beyond the confines of advertisement poaching to explore how these SEO dynamics facilitate the identification of both direct and indirect competitors. By delving into the complex interplay of these metrics, the research sheds light on the strategic maneuvers employed by online retailers to gain an edge in a highly competitive digital marketplace. The findings underscore the multifaceted nature of e-commerce competition, where SEO proficiency not only influences visibility but also defines the contours of rivalry in the online domain.

A. Uncovering Underlying Dynamics Among Competitors in the Digital World

A.1. Introduction

The traditional approach to defining business competitors, based solely on physical products and geographical proximity, is rapidly becoming obsolete. In today's fast-paced digital landscape, the competitive arena has expanded to encompass the virtual world, with search engines at its epicenter. Companies no longer compete exclusively for market share or customer attention based on their offerings; they are now engaged in a fierce race to secure prominent positions within search engine results. This paper dives into the complexities of identifying competitors in this digital age, where understanding the competitive landscape goes beyond recognizing companies with similar products and services and extends into the land of getting superior visibility within search engine rankings.

A.1.1 Background & Problem Statement

The current business landscape has undergone a deep transformation, driving companies into an era where the definition of competition can not only be bounded to their physical products. Before the internet era arrived, businesses predominantly gauged their competition assessing competitors within their geographic proximity and looking for similar products or services to the ones they are selling. However, the digital age has altered this model. The propagation of online commerce, coupled with the importance of search engines, has given rise to a multilayered, digital competitive landscape.

In this digital age, competition is no longer narrowed by the attributes of products or services; it now includes the digital footprints of businesses. This change is accentuated by search engines, which have become the arena where businesses compete against each other. Companies are no longer only interested in securing a market share; they are competing too for a prime position within search engine result pages (SERPs).

This dynamic shift stresses the need for a more precise and comprehensive approach to identifying competitors. Conventional methods, focused solely on product similarity or geographic proximity, fall short in this new landscape. The ongoing business setting calls for a deeper understanding of the complexities involved in the pursuit of prominence on the SERPs—a place where visibility often equates to success.

As such, this research is motivated by the need to redefine the considerations of competitor identification. In the digital age, it is no longer sufficient to be aware of businesses with analogous product offerings; the imperative is to grasp the competitive dynamics within the race for heightened visibility on search engines. This shift motivates a search for innovative approaches to discern and categorize competitors in this new landscape, thus forming the foundation of the analysis.

Because of this, implications of inadequate competitor mapping are extensive. Businesses relying on outdated or incomplete competitive intelligence risk misallocating resources, deploying wrong strategies, and missing opportunities for growth and optimization within the internet. Focusing on this problem is crucial for businesses struggling to grow their digital presence. Therefore, our research seeks to investigate innovative approaches, such as Minimum Spanning Trees (MSTs) and network analysis, to redefine competitor identification in the digital age. By doing so, this research aims to provide a new perspective on competitor networks within the digital world, offering a solution to this critical challenge in contemporary digital marketing and competition analysis.

A.1.2 Objectives

The primary objectives of the research are driven by the need to address the challenges posed by evolving digital competition:

Improve 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.

Differentiate competitors in the race for search engine rankings

Recognizing that competition within the digital world focuses around securing prime positions in search engine rankings, it aims to redefine competitor mapping to include this critical dimension. The research seeks to distinguish not only traditional competitors but also those competing for a position on SERPs. By doing so, it aims to provide a simpler and more practical understanding of competition in this new age.

Uncover interconnections and interdependencies among companies

Businesses are deeply interconnected, forming a complex web of relationships that surpass product features. To address this, we employ network analysis techniques and Minimum Spanning Trees (MSTs) to reveal the underlying structures. The research seeks to uncover direct competitors, as well as potential indirect competitors, whose interactions may not be immediately evident. This in-depth exploration of interdependencies among companies within the digital domain offers valuable insights for businesses seeking to position themselves strategically within the digital landscape. We aim to provide a holistic approach that surpasses traditional competitor mapping, ultimately enhancing the field of digital marketing and competition analysis.

A.1.3 Research Questions

To guide the research, the following questions are made:

- How can Minimum Spanning Trees be utilized to distinguish between direct and indirect SEO competitors?
- What valuable insights can be gained from the use of Minimum Spanning Trees to analyze SEO competition networks?
- In what ways does the MST-based approach enables a richer semantic analysis?

A.2. Methodology

In this section, it's discussed the analytical approach that serves as the basis for this research. This approach is divided into two steps. The first step involves the creation of numerical features at the domain level, which are derived from quantitative variables that originally exist at the product level. At the same time, its carried an extraction of different features from the textual variables to generate vector representations for each domain. To achieve this, FastText works as the engine for generating embeddings from textual features. FastText's capabilities are vital in enhancing our analysis by capturing the rich textual characteristics associated with each domain. In addition to FastText, Singular Value Decomposition (SVD) is used as a tool for dimensionality reduction, allowing the process to convert the extensive dataset into a manageable and informative form.

Also, the concept of Minimum Spanning Trees (MSTs) is introduced and its application within network analysis. MSTs play an important role in our research as they offer a unique and practical approach to competitor mapping. By utilizing MSTs, the research can reveal the complex relationships among domains, distinguishing between direct and indirect competitors. This aspect is particularly critical in the context of digital competition, where traditional methods often fall short in accounting for complex interdependencies.

The combined use of FastText, SVD, and MSTs as an analytical approach is essential for comprehensively understanding the competitive dynamics in the digital world. These techniques provide the study with the ability to capture both quantitative and textual characteristics, allowing it to navigate the complexities of modern digital competition effectively.

A.2.1 Feature Generation and Text Embeddings using FastText

In this research, feature engineering is a key element of the analytical approach. The study has dedicated significant efforts to creating a robust and comprehensive set of features. This process involves two distinct yet interconnected aspects: the extraction of numerical elements from quantitative variables and the extraction of textual elements using FastText.

The creation of numerical features at the domain level is a multi-step process. It begins by aggregating and summarizing quantitative variables originally associated with individual products. These variables include metrics such as cost per click, monthly search volume, number of products being sold, and product positions in the search engine ranking, among other variables. Aggregating these at the domain level allows to establish domain-specific quantitative characteristics. It also calculates statistical properties such as mean, median, standard deviation, and confidence intervals to capture various facets of each domain.

At the same time, it extracts features from the textual variables associated with each domain. To get the most out of the textual features it applies different steps to process them such as removal of stop words, strings only showing numbers, special characters. After the initial cleaning, it lemmatizes the words to get the root form of each one and extract the nouns from each string to spot either the brands, products, or relevant keywords. These features are essential in understanding the nuances and context of digital competition.

To get a richer set of features it uses FastText (Bojanowski, et al. 2017) which is a technology developed by Facebook's AI Research (FAIR) lab that extends the concepts of Word2Vec by utilizing subword information. While Word2Vec operates at the word level, FastText goes a step further by considering sub-word information, which is particularly advantageous when dealing with complex morphological structures or when analyzing short and informal texts.

FastText, like Word2Vec, operates on the distributional hypothesis, which posits that words that appear in similar contexts tend to have similar meanings. It learns vector representations for words by training on large corpora of text. These vector representations, known as word embeddings, capture the semantic and syntactic information associated with words. This makes them useful for a wide range of NLP tasks, including text classification, named entity recognition, sentiment analysis, and information retrieval.

One of the main advantages of FastText against Word2Vec is that it treats each word as a bag of character n-grams. N-grams are contiguous sequences of characters within a word. For instance, in the word "orange," the 3-grams would be "ora," "ran," "ang," and "nge." FastText creates vector representations for each of these n-grams and aggregates them to obtain the vector representation of the word. This approach allows FastText to capture not only the meaning of whole words but also the meanings of smaller morphemes and subwords within them.

This method is preferred due to its ability to capture rich textual characteristics, handle out-of-vocabulary words, and efficiently process large datasets (Mikolov, et al. 2013). This choice is particularly advantageous when dealing with our type of dataset where domain names, product names, and other textual elements can be highly variable and subject to change.

A.10.2.2 Singular Value Decomposition (SVD) for Dimensionality Reduction

The feature space obtained from the previous step is typically high-dimensional, posing challenges for subsequent analysis and interpretation. To address this, the study employs

Singular Value Decomposition (SVD) as a tool for dimensionality reduction (Stewart 1993). SVD has proven to be invaluable in various fields, including data compression, image processing, and recommendation systems. In the research, SVD plays a critical role in transforming the extensive feature space into a more compact and informative representation. This technique transforms the original data as a combination of orthogonal singular vectors. By selecting a subset of these singular vectors, it retains the most significant information while discarding noise and less important variations. This reduction in dimensionality simplifies the dataset and focuses on the essential aspects of competition. For this study, the number of dimensions to be chosen is defined using the amount of variance explained when it's equal to 90%.

A.2.3 Minimum Spanning Trees (MST) in network analyses

Minimum Spanning Trees (MSTs) are a fundamental concept in graph theory (Graham and Hell 1985). They are a subgraph of a larger network that connects all the nodes with the minimum possible total edge weight. In our context, these nodes represent domains or entities, and the edge weights signify the strength of relationships between them. MSTs allow the study to distill the complex web of connections into a simpler, more understandable structure.

MSTs are a fundamental concept in graph theory. They are a subgraph of a larger network that connects all the nodes with the minimum possible total edge weight. In this context, these nodes represent domains or entities, and the edge weights signify the strength of relationships between them.

With the help of MSTs, it’s possible to uncover the complex relationships among domains, distinguishing between direct and indirect competitors. This distinction is particularly critical in the context of digital competition.

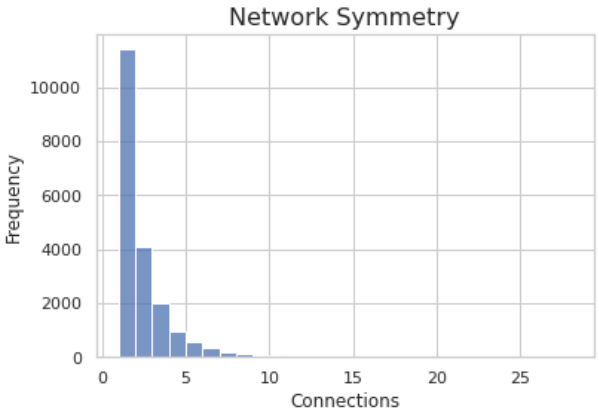
A.3. Results

Within this section, the results derived from the analytical approach introduced in the previous section are explored. One of the notable findings is the absence of symmetry in the network structure, a trend evident in Figure 3.1.

This observation hints at the unequal distribution of competitors within the competitive landscape. Interestingly, approximately a quarter of these domains have not just one, but two or more competitors. This imbalance suggests the potential existence of hierarchical relationships or disparities in competition, where a select few companies dominate the competitive landscape. Such imbalances in competition can influence industry dynamics, including factors such as pricing strategies, consumer choices, and overall competitive market structures like search engine rankings.

Figure 3.1

Competitors Network Symmetry



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

Since our network lacks symmetry, its imperative to turn to metrics like degree, closeness, and betweenness centrality to gain deeper insights into the inner workings of each company. These metrics enable us to understand various aspects, such as the number of direct competitors within the network, the level of connectivity with other companies, and the potential role of a company as a bridge between different companies. This multifaceted analysis not only helps uncover the

competitive network but also underscores the strategic significance of each company within the network.

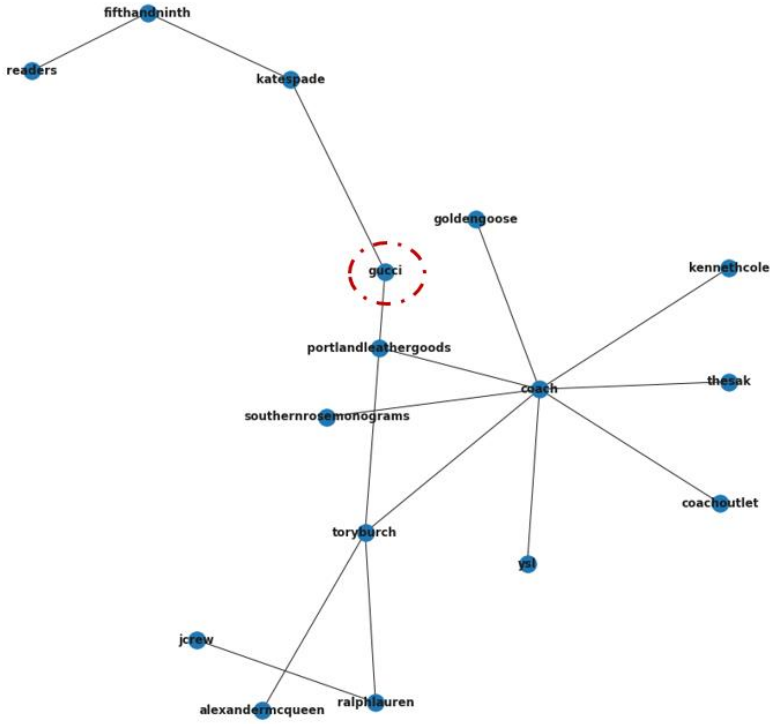
It is essential to understand degree centrality, as it allows us to gauge the intensity of competition and evaluate the network's overall structure. Companies exhibiting a high degree of centrality, like StockX (An online marketplace and clothing reseller, primarily of sneakers), typically contend with a broader array of direct competitors. This requires the formulation of distinct competitive strategies. Closeness centrality, on the other hand, draws attention to a company's proximity to others within the network, indicating its ability to rapidly influence or be influenced by competitors. Meanwhile, betweenness centrality serves as a vital metric for identifying nodes that serve as intermediaries or bridges in the network, shaping the flow of information or competition. These insights are important in strategic decision-making, ranging from establishing partnerships with highly central companies to even how companies invest in their presence in the digital world.

After comprehending the structural features of our network, 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 the findings, the prominent fashion brand Gucci is used. After reducing the number of connections through our previous analysis, we unveil a revealing structure.

Figure 3.2

Potential direct and indirect competitors for Gucci



Note: The target company circled in red.

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

Within Table 3.1, it's possible to gain some valuable insights into the semantics behind these strategies based on Top-N keywords, products, and categories for each competitor. This granular level of analysis aids in understanding the specific tactics employed by each company, offering a comprehensive view of the keywords and strategies that influence their digital presence and competitiveness.

Table 3.1

Semantics behind direct competitors of Gucci, Top-N words

Keywords	Products	Categories
Bag	Bag	Accessories
Mini	Mini	Apparel

Wallet	Leather	Wallets
Black	Shoulder	Cases
Tote	Small	Handbags
Small	Sandal	Shoes
Belt	Tote	Clothing
Leather	Belt	Jewelry
Crossbody	Wallet	Bags

Identifying indirect competitors also holds significant strategic implications. It opens opportunities for companies to explore new markets and customer bases. Understanding indirect competitors expands horizons, offering the potential for market expansion. This understanding paves the way for diversification, differentiation, or the exploration of untapped market segments indirectly influenced by these competitors. Simultaneously, it equips companies to anticipate potential threats and adjust strategies accordingly. This awareness may involve developing contingency plans, refining marketing messages, or fortifying market positions to navigate changing competitive dynamics.

Like in Table 3.1, in Table 3.2, it presents some insights into the Top-N keywords, products, and categories for each competitor. This detailed information empowers companies to fine-tune their strategies and remain competitive in the ever-evolving digital landscape.

Table 3.2

Semantics behind indirect competitors of Gucci, Top-N words

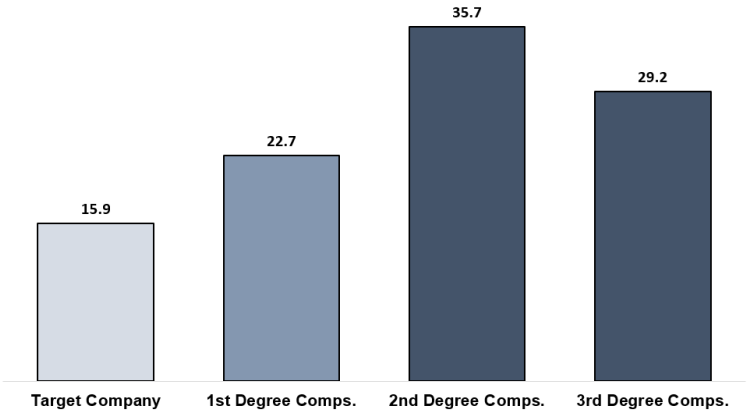
Keywords	**Products**	**Categories**
Bag	Bag	Accessories
Ring	Leather	Apparel
Black	Small	Handbags
Wallet	Mini	Wallets
Belt	Shoulder	Cases

Leather	Tote	Clothing
Hat	Belt	Shoes
Case	Medium	Jewelry
Chain	Wallet	Bags

Assessing information about competitors also enables companies to benchmark their internet visibility expenditures, facilitating a comparison of the efficacy of their practices with those of competitors in driving volume more efficiently.

Figure 3.3

Efficiency of Expenditure on Visibility Among Competitors



Note: Efficiency measure as number of impressions divided by the average cost per click.

In summary, the preceding measurements and shortest path analysis extend beyond the mere identification of direct competitors. These analyses play a pivotal role in showing both indirect and direct competitors. While direct competitors are more straightforward to detect due to their similar features, indirect competitors may not exhibit an obvious overlap in product or other features. However, they can still significantly influence one another's market positions. These indirect competitors may include companies offering complementary products, firms targeting overlapping customer segments, or competing for a prominent position in search engine rankings. Inspecting the strength of these relationships allows us to discover potential indirect competitors whose influence might not be immediately evident but holds considerable influence

over the competitive landscape. In essence, this analytical approach provides a comprehensive view of the complex of relationships within the digital world, enabling us to distinguish both direct and indirect competitors. This comprehensive approach empowers companies to make informed decisions and adapt to the evolving landscape of digital competition, ultimately driving success and growth in the digital age.

A.4. Conclusions

This research has undertaken the challenge of redefining competitor identification for the digital age. Through this multi-step approach, we have dived into the competitive landscape, discovering insights that can drive strategic decision-making and success in the digital world.

- The application of Minimum Spanning Trees (MSTs) reveals direct and indirect SEO competitors within the digital landscape. MSTs allowed us to create a simplified yet highly informative structure that revealed complex relationships among domains. By identifying the shortest paths and connections, we can highlight both direct competitors who exhibit similarities in product features and indirect competitors whose influence might not be immediately evident.
- The use of MSTs provides a deeper understanding of the structural dynamics within the competitive network, as evidenced by the asymmetry and the uneven distribution of competitors. The centrality metrics, including degree, closeness, and betweenness centrality, have provided multifaceted insights into the intensity of competition, the level of connectivity, and the role of companies as intermediaries.
- The MST-based approach offers several advantages over established methods for competitor mapping in SEO. It captures the complex interdependencies and hierarchical relationships among competitors, something traditional methods struggle to accomplish. The use of textual features, dimensionality reduction through SVD, and the application of MSTs combine to provide a comprehensive view of the competitive landscape.

The insights gained from this approach go beyond the identification of competitors, offering a nuanced understanding of both direct and indirect competitors and their potential influence.

A.5. Further Steps

In addition to the positive results of the research, it is possible to enhance them by implementing the following recommendations:

- Exploring the direction of improving Minimum Spanning Trees (MSTs) using Graph Neural Networks (GNNs) is highly advantageous due to the potential to enhance graph representations, optimize MST structures, and adapt them to evolving graphs. GNNs provide an opportunity to predict missing edges, make intelligent decisions, and uncover complex relationships within intricate graph structures.
- Investigate alternatives for text embedding using transformer-based models like BERT or GPT for contextual embeddings, which might capture semantic relationships among word

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 WH Kwok, Lai Guan Ng, Florent Ginhoux, and Evan W Newell. 2019. "Dimensionality reduction for visualizing single-cell data using UMAP." *Nature biotechnology* 38-44.

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, 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.

Chaffey, Dave, and Fiona Ellis-Chadwick. 2019. *Digital marketing.* Pearson uk.

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* 75-174.

Gabel, Sebastian, Daniel Guhl, and Daniel Klapper. 2019. "P2V-MAP: Mapping market structures for large retail assortments." *Journal of Marketing Research* 557-580.

Gerling, Christopher. 2023. "Company2Vec--German Company Embeddings based on Corporate Websites." *arXiv preprint.*

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.

Graham, Ronald L, and Pavol Hell. 1985. "On the history of the minimum spanning tree problem." *Annals of the History of Computing* (IEEE) 1: 43-57.

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

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

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.*

Joulin, Armand, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. "Bag of tricks for efficient text classification." *arXiv.*

Kök, A Gürhan, Marshall L Fisher, and Ramnath Vaidyanathan. 2015. "Assortment planning: Review of literature and industry practice." *Retail supply chain management: Quantitative models and empirical studies* 175-236.

Li, Zhao, Xin Shen, Yuhang Jiao, Xuming Pan, Pengcheng Zou, Xianling Meng, Chengwei Yao, and Jiajun Bu. 2020. "Hierarchical bipartite graph neural networks: Towards large-scale e-commerce applications." In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, 1677-1688. IEEE.

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, 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* 1118-1123.

Rousseeuw, Peter J. 1987. "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis." *Journal of computational and applied mathematics* 53-65.

Stewart, G. W. 1993. "On the Early History of the Singular Value Decomposition." *SIAM Review* (Society for Industrial and Applied Mathematics) 35: 551-566.

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.*

Appendix

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

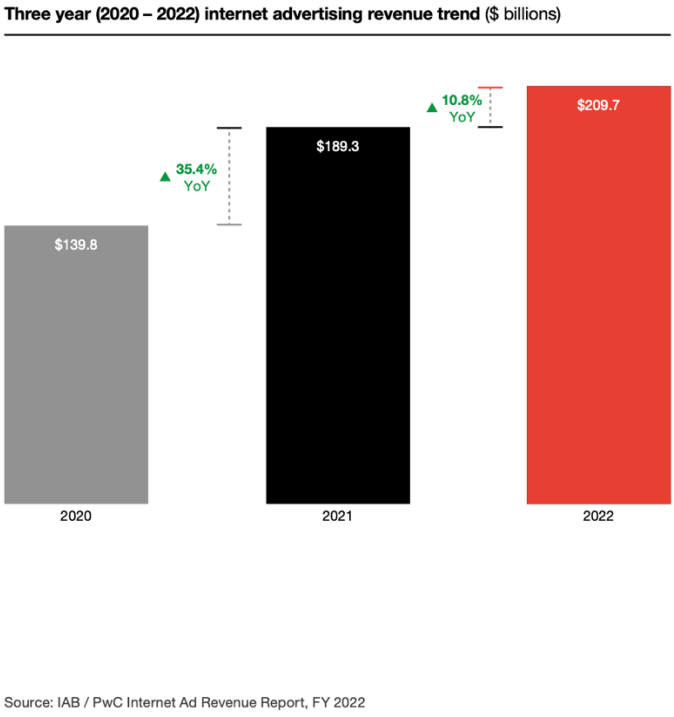


Figure 5 - Internet Advertising Revenue. Source: IAB/PwC Internet Ad Revenue report, FY 2022

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%

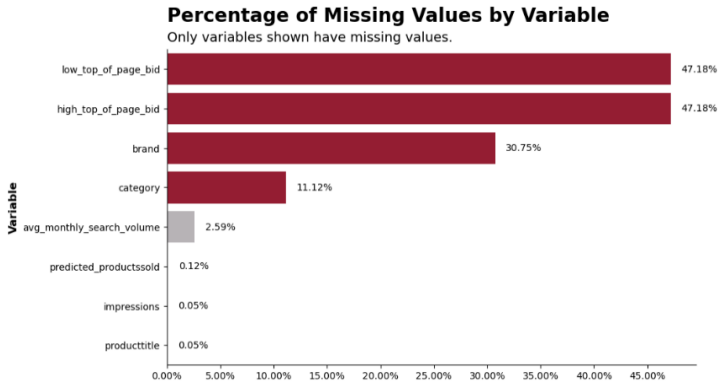


Figure 6 - % of Missing Values within dataset