GRAPH-BASED MARKET SEGMENTATION USING A COMMUNITY DETECTION ALGORITHM: A CASE STUDY ON THE YELP DATASET

Marko Marković*,
[0000-0001-6036-4146]

Katarina Marković,
[0000-0002-0153-4069]

Biljana Tešić
[0000-0003-3226-8477]

Singidunum University,
Belgrade, Serbia

Abstract:

Market segmentation based on user-generated data becomes challenging when business entities cannot be reliably grouped using only attributive features, and it is necessary to take patterns of a shared user base into account. This problem is particularly relevant in digital marketing. Relationships between businesses revealed through user behavior may indicate hidden market segments and connect IT-based network analysis methods with practical marketing decisions. In this paper, a graph-based approach is used to construct a business-business network from the Yelp dataset and apply a community detection algorithm. The results indicate that this approach can identify structurally coherent groups of businesses with similar audiences, which may serve as a basis for more precise segmentation, targeting, and competitive environment analysis.

Keywords:

Market Segmentation, Community Detection Algorithms, Greedy Modularity Optimization.

INTRODUCTION

Market segmentation is one of the key activities of modern marketing, as it enables the identification of groups of users or market entities with similar needs, interests, and behavioral patterns. In the digital environment, where organizations have access to large volumes of data on user interactions, reviews, and preferences, traditional segmentation approaches based solely on demographic and descriptive characteristics are often insufficient for uncovering more complex behavioral patterns. For this reason, machine learning has gained an increasingly important role in marketing in recent years, especially in tasks such as customer segmentation, offer personalization, and decision support [1] [2].

A particularly important challenge arises in the analysis of platforms based on user-generated content, where the data are not only attributive but also relational. On platforms such as Yelp, users interact with different businesses through ratings and reviews, thereby forming a complex network of relationships between users and business entities.

Correspondence:
Marko Marković

e-mail:
mmarkovic@singidunum.ac.rs

DOI: 10.15308/Sinteza-2026-233-239

In such systems, segmentation cannot be viewed only as grouping based on individual attributes, but also as the problem of identifying structurally connected groups within a network. *Wang* shows that a network-based approach to the analysis of online reviews can provide significant insights into segmentation and user behavior [3].

Graph models provide a suitable formal framework for analyzing such data, as they enable the modeling of entities as nodes and their relationships as edges. In marketing, network analysis is used to examine the structure of relationships among market actors, including users, products, brands, and business entities [4]. Within this approach, *community detection* is particularly important, that is, the identification of groups of nodes that are more densely connected to each other than to the rest of the network. This problem can be viewed as a form of unsupervised machine learning, since the algorithm discovers natural groupings based on the structure of the network itself, without predefined classes [5] [6].

This paper examines the application of a community detection algorithm to the Yelp dataset, with the aim of identifying market segments based on the network of relationships between users and businesses. The initial model is based on a bipartite user-business graph, which is then projected into a business network connected through a shared user base. An algorithm based on modularity maximization is applied to the resulting graph, identifying communities that can be interpreted as groups of businesses with similar audiences or market positions. In this way, the paper connects an IT perspective, through graph algorithms and data processing, with a marketing perspective, through market segmentation and the practical applicability of the results in targeting, local positioning, and the planning of promotional activities.

2. THEORETICAL BACKGROUND AND MOTIVATION

A graph is a mathematical structure composed of nodes and edges, where nodes model entities and edges represent the relationships between them. This type of model is particularly suitable in cases where the subject of analysis is not merely a set of isolated attributes, but a system in which patterns of connectivity between elements are important. In marketing, network analysis enables the study of relationships among users, products, brands, and business entities, which is especially significant in a digital environment where a large portion of behavior takes place through interactions on platforms, portals, and social networks [4].

Modern marketing systems generate data that is simultaneously attributive and relational. Users leave ratings, reviews, comments, and recommendations, thereby forming complex networks of relationships with different businesses and services. Under such conditions, classical tabular analysis is often insufficient to reveal deeper behavioral patterns, because it does not take into account the network structure of the data. Precisely for this reason, the graph-based approach is becoming increasingly important in contemporary analytical systems and marketing intelligence [2] [4].

In the context of market segmentation, graphs enable a shift from the analysis of individual characteristics to the analysis of similarity and connectivity. For example, two businesses may be connected not because they belong to the same formal category, but because they are rated or visited by the same users. This type of structural connectivity may have greater marketing value than administrative classification, because it more accurately reflects actual audience behavior [3] [7].

2.1. COMMUNITY DETECTION AS UNSUPERVISED MACHINE LEARNING

Community detection represents one of the central problems in the analysis of complex networks. The basic idea is to identify groups of nodes within a network that are more densely connected than to the rest of the network. Such groups are referred to as communities, modules, or clusters. *Clauset, Newman*, and *Moore* emphasize that community detection is particularly important in large networks, as it enables the identification of their internal organization and simplifies the interpretation of complex systems [5].

From a machine learning perspective, community detection can be viewed as a form of unsupervised learning. Unlike classification, where predefined classes exist, here the algorithm independently discovers the structure of the data based on the pattern of connections within the network. This is especially useful in situations where the researcher does not know in advance how many segments exist or how they are organized. In a marketing context, the identified communities can be interpreted as groups of businesses or users with similar behavioral patterns, interests, or interaction structures [1] [5] [8].

In platforms based on reviews and evaluations, such as Yelp, communities may indicate the presence of market niches, groups of similar businesses, or shared user bases. [9] Wang shows that a network-based approach to the segmentation of online reviews can provide significant insights into market structure and the roles of different

types of nodes, including central, peripheral, and bridging segments [3]. Therefore, community detection is suitable not only as a technical method for graph analysis, but also as a tool for understanding market structure from a marketing perspective.

2.2. MODULARITY

One of the most influential criteria for evaluating the partition of a network into communities is modularity. *Newman* defines modularity as a measure that compares the actual number of edges within proposed communities to the expected number of such edges in a corresponding random network model. High modularity indicates that the partition is good, in the sense that there are more connections within communities than would be expected by chance. The standard form of modularity can be expressed as shown in Equation 1:

$$Q = \frac{1}{2m}\sum_{ij}\left(A_{ij} - \frac{k_i k_j}{2m}\right)\delta\left(c_i, c_j\right)$$

Equation 1. Standard form of modularity.

where A_{ij} is an element of the adjacency matrix, k_i is the degree of node i, m is the total number of edges, and $\delta(c_i, c_j)$ is an indicator function denoting whether two nodes belong to the same community. In practical terms, modularity enables the quality of segmentation to be evaluated not only visually but also numerically. [6]

In this paper, modularity plays a dual role. On the one hand, it provides the theoretical foundation for understanding what constitutes a "good" segmentation in a network. On the other hand, it is directly related to the algorithm used in the experimental part of the study, since the greedy community detection approach operates by iteratively increasing the modularity value at each step of merging communities [5] [6] [10].

2.3. MOTIVATION FOR PROBLEM SELECTION

A particular challenge lies in the fact that consumer behavior in modern environments cannot be sufficiently described using only individual attributes of users or products. A substantial amount of information is embedded in the relationships between actors: which businesses users visit, rate, or recommend, which businesses share the same audience, and how recognizable groupings emerge within the network. For this reason, network analysis and community detection represent an effective combination of IT approaches and marketing applications.

The motivation of this study stems from the need to demonstrate how a relatively clear and interpretable algorithmic approach can provide valuable insights for market segmentation. Instead of relying solely on classical statistical or attribute-based methods, this paper assumes that relationships between users and businesses constitute an important source of information about market structure. In this way, the topic remains primarily within the domain of information technology, as it is based on graphs, algorithms, and data processing, while simultaneously maintaining a clear marketing focus through market segmentation, with detected communities serving as representations of market segments.

3. DATASET

The experimental part of the study uses the *Yelp Open Dataset*, which is intended for educational purposes and contains real-world data on businesses and user activities, including reviews and related business information. Public descriptions of the dataset indicate that it consists of real-world data and that its key components are organized in JSON files covering businesses, reviews, and users. [11]

In this study, three files are used: *review.json, business.json*, and *user.json*. The review file contains at least the user identifier (*user_id*) and the business identifier (*business_id*), enabling the linking of users and businesses into a unified network model. Descriptions of the Yelp data structure further confirm that *user_id* and *business_id* represent the key relationships between tables, while *business.json* includes information such as business name, city, categories, and other attributes. [12] [13]

Due to computational limitations and the scope of a conference paper, the analysis is not conducted on the full dataset, but rather on a sample of approximately 50,000 reviews constructed from multiple segments of the data file in order to reduce sampling bias. This approach is methodologically justified for a demonstrative and experimental study, as the goal is to illustrate the applicability of the method and the interpretability of the detected communities, rather than to exhaustively analyze all statistical properties of the complete dataset.

4. METHODOLOGY

The methodological framework of the study is based on several sequential steps. As a first step, a bipartite graph $G = (U \cup B, E)$ is constructed from the review data, where the set U represents users, the set B represents businesses, and an edge E exists if a given user has reviewed a given business. This representation is natural, as it directly models the interaction between users and services. Although the bipartite graph itself contains significant information, for market segmentation its projection onto the set of businesses is more suitable. This results in a new graph in which two businesses are connected if they share common users, while the edge weight can be defined as the number of users who have reviewed both businesses. In this way, the network no longer represents individual reviews, but rather the similarity between businesses in terms of their shared user base.

Such a projection has a clear marketing interpretation. If two businesses share a larger number of common users, it can be assumed that they belong to the same or adjacent market niche, compete for a similar audience, or participate in shared consumption patterns. Therefore, the projected business graph becomes a suitable basis for further segmentation. In the experimental procedure, it is also possible to filter weak connections, for example by retaining only those edges whose weight exceeds a predefined threshold, to reduce noise and emphasize more structurally relevant relationships. Although this filtering is not theoretically mandatory, in practical work with real data it often improves the interpretability of the network.

A community detection algorithm, *greedy_modularity_communities*, from the *NetworkX* library is then applied to the constructed graph. This is an implementation of the **Clauset–Newman–Moore** approach, i.e., *greedy modularity maximization*. The algorithm starts from a trivial partition in which each node represents its own community, and then iteratively merges pairs of communities that yield the greatest increase in modularity, until no further improvement is possible. This approach is particularly suitable for medium-scale analyses, as it provides a good balance between partition quality and computational efficiency. [5] [14]

After community detection, the quality of the resulting partition is evaluated using the modularity value. In addition, each community is interpreted based on attributes from the *business.json* file, primarily business name, city, and categories. In this way, *the analysis moves from an abstract network partition to meaningful marketing segments*. The visual component of the experiment may include: (1) a partial visualization of the business network, (2) a color-coded representation of communities, and (3) selected examples of dominant categories or representative businesses within each of the larger communities. Modularity as a quality criterion and communities as the algorithm's output directly connect the theoretical and experimental parts of the study. [6]

5. RESULTS AND ANALYSIS

After constructing the projected business-business graph, the resulting network contained 3,402 nodes and 9,579 edges, while the largest connected component comprised 84 nodes and 108 edges. The application of the greedy modularity maximization algorithm identified 9 communities, with a modularity value of 0.6252. This value indicates a *high level of structural coherence*, suggesting that connections within communities are substantially denser than connections between communities. Such a result confirms that the graph derived from shared user interactions contains a meaningful latent structure rather than randomly connected businesses.

The obtained result is particularly important from the perspective of market segmentation. Since the graph was constructed by linking businesses reviewed by the same users, the detected communities may be interpreted as *groups of businesses sharing a similar audience base*. In other words, the algorithm did not rely on predefined business categories but instead discovered clusters directly from user behavior patterns. This makes the resulting segmentation especially relevant for marketing analytics, where the goal is often to reveal hidden structures in customer activity.

The size distribution of the detected communities also suggests that the business-business network is *not* homogeneous, as summarized in *Table 1*. The three largest communities contained 16, 14, and 11 nodes, respectively, while the remaining communities were smaller and more specialized. Such a distribution is consistent with the idea that the local market contains several broader behavioral clusters together with narrower niche segments.

Figure 1 presents a partial visualization of the projected business-business graph. Each node represents a business, while edges indicate that two businesses share common users in the analyzed sample. Even in this partial representation, the graph does not appear random; rather, it shows visible local clusters and denser substructures.

This visually supports the assumption that user review behavior creates meaningful relations between businesses.

Figure 2 shows the same network with detected communities highlighted in different colors. The visual separation of colored clusters is consistent with the modularity score and provides additional confirmation that the algorithm was able to identify distinct structural groups. The figure also demonstrates that some communities are relatively compact, while others are more diffuse, which may reflect differences in market specialization or overlap in customer behavior.

A qualitative examination of the detected communities *reveals meaningful business patterns*. The largest community, containing 16 nodes, was dominated by the categories *Restaurants, Bars*, and *Nightlife*. Representative businesses in this cluster included *Southgate, Tria Cafe Wash West, Dirty Franks* and *Pub & Kitchen*. This suggests a segment composed primarily of hospitality and social venues with overlapping customer bases. The second-largest community, with 14 nodes, also included a significant number of restaurants, but showed a more heterogeneous composition, including businesses such as *Penang, PetSmart, Sephora*, and *Rescue Spa*. This may indicate a broader urban consumer routine in which food, retail, and personal care businesses are behaviorally linked through the same population of users. The third-largest community, with 11 nodes, was characterized by categories such as *Restaurants, Japanese,* and *American (New)*, with representative businesses including *Fish & Co, Samurai Sushi, RuSan's Sushi* and *Seafood, Prime 108,* and *Tavern*. This cluster appears to represent a food-oriented segment with a more specific culinary profile.

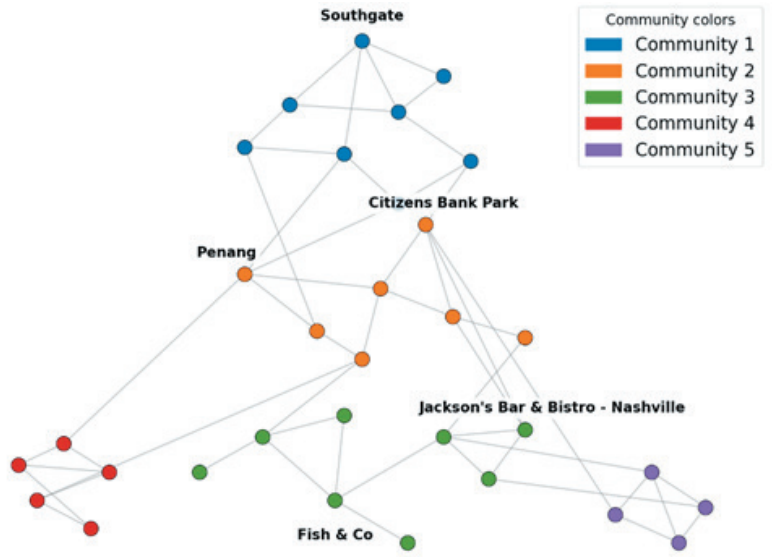

Figure 1. A partial visualization of the projected business-business network using *NetworkX*

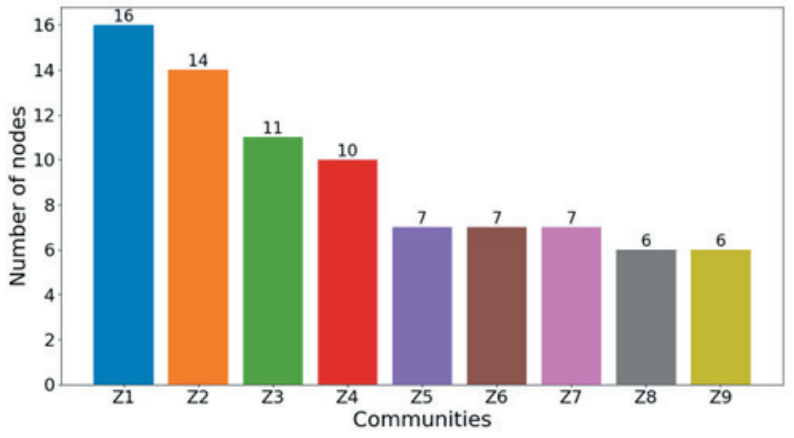

Figure 2. Detected communities in the business-business network; node colors indicate community membership

An especially important finding is that the detected communities do not always correspond directly to the formal business categories provided in the dataset. In several cases, businesses from different administrative categories appeared in the same community because *they shared a similar customer base*. From a marketing perspective, this is highly valuable, since it reflects **actual consumer behavior** rather than only descriptive classification. Therefore, the graph-based approach reveals relations that may remain hidden in traditional *attribute-based* segmentation.

The results indicate that community detection can provide an effective basis for identifying market segments in review-based digital platforms. By combining network projection with modularity-based clustering, the analysis reveals groups of businesses that are linked not merely by formal labels, but by real user interaction patterns. This makes the method particularly suitable for applications in local marketing, customer targeting, cross-promotion analysis, and competitive positioning.

6. DISCUSSION: MARKETING IMPLICATIONS

The obtained results have several potential implications for marketing practice. First, communities of businesses can be interpreted as market segments based on a shared user base. This means that a marketing manager can better understand which competitors or complementary actors their business actually intersects with in terms of consumer behavior, rather than relying solely on declarative categories or intuitive market assessments. Second, the identified communities can be used for planning local partnerships and cross-promotion activities, as businesses within the same community are likely to attract similar users. Third, bridging nodes between communities may indicate businesses that connect different niches and therefore have particular strategic importance for spreading messages or promotions across multiple segments. This interpretation is consistent with findings in the literature that emphasize the importance of a network-based approach for understanding marketing relationships and segmenting online reviews. [3]

From an information technology perspective, the study demonstrates that a relatively simple combined procedure—parsing *semi-structured JSON* data, constructing a bipartite graph, projecting the network, and applying *modularity-based* community detection—can yield results that are both technically grounded and business-interpretable. This confirms that the application of machine learning in marketing does not necessarily require deep neural networks or complex predictive models; network analysis algorithms can also represent a highly relevant form of intelligent analytics when the nature of the problem is relational. This is consistent with contemporary literature reviews on the role of machine learning and AI in marketing, which indicate that the value of these methods lies not only in prediction, but also in supporting consumer understanding, segmentation, and decision-making. [1]

Table 1. Summary of network structure and detected community characteristics

Metric / Community	Value / Description
Number of nodes in business-business graph	3,402
Number of edges in business-business graph	9,579
Number of nodes in the largest connected component	84
Number of edges in the largest connected component	108
Number of detected communities	9
Modularity	0.6252
Largest community size	16
Second-largest community size	14
Third-largest community size	11
Community 1 dominant categories	Restaurants, Bars, Nightlife
Community 1 representative businesses	Southgate, Tria Cafe Wash West; Dirty Franks; Pub & Kitchen
Community 2 dominant categories	Restaurants, American (Traditional), Pets
Community 2 representative businesses	Vetri Cucina; Penang; PetSmart; Sephora; Rescue Spa
Community 3 dominant categories	Restaurants, Japanese, American (New)
Community 3 representative businesses	Fish & Co; RuSan's Sushi and Seafood; Prime 108; Tavern

However, the approach also has limitations. The projection of the bipartite graph onto a business network inevitably reduces part of the information, as it no longer captures the nuances of individual user behavior but instead focuses on aggregated relationships between businesses. In addition, the results may depend on the sample size, the threshold used for filtering edges, and the choice of algorithm. The greedy modularity approach is efficient and practical, but it does not always guarantee a unique or globally optimal partition. Therefore, future work could include comparisons with alternative methods such as Louvain or Leiden algorithms, as well as the incorporation of textual features from reviews to enable a more refined semantic interpretation of the detected communities. [14]

7. CONCLUSION

This paper presents a framework for graph-based market segmentation using the Yelp dataset, where segmentation is modeled as a community detection problem in a business network. The underlying idea is that real patterns of user behavior cannot always be adequately described using only tabular attributes, but that relationships between users and businesses contain additional structural information. For this reason, a bipartite user–business graph was constructed, projected into a business-business network, and analyzed using a greedy modularity maximization algorithm.

From a theoretical perspective, the study demonstrates how the concepts of modularity and community detection can be naturally linked to marketing segmentation. From a practical perspective, it shows that Yelp data provide a suitable basis for such analysis, as they enable modeling of networks based on shared user interactions among businesses. The expected contribution of the study lies in combining an IT perspective—data processing, graph modeling, and algorithms—with a marketing focus on segmentation, targeting, and understanding market structure.

REFERENCES

[1] M. Alves Gomes and T. Meisen, "A review on customer segmentation methods for personalized customer targeting in e-commerce use cases," *Information Systems and e-Business Management*, vol. 21, p. 527–570, 2023, doi.org/10.1007/s10257-023-00640-4.

[2] A. De Mauro, A. Sestino and A. Bacconi, "Machine learning and artificial intelligence use in marketing: a general taxonomy," *Italian Journal of Marketing*, p. 439–457, 2022, doi.org/10.1007/s43039-022-00057-w.

[3] H.-J. Wang, "Market segmentation of online reviews: A network analysis approach," *International Journal of Market Research*, vol. 64, no. 5, 2022, doi.org/10.1177/14707853211059076.

[4] C. M. Webster and P. D. Morrison, "Network analysis in marketing," Australasian Marketing Journal, vol. 12, no. 2, 2004, doi.org/10.1016/S1441-3582(04)70094-4.

[5] A. Clauset, M. E. J. Newman and C. Moore, "Finding community structure in very large networks," *Physical Review E,* vol. 70, no. 6, 2004, doi.org/10.1103/PhysRevE.70.066111.

[6] M. E. J. Newman, "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences*, vol. 103, p. 8577–8582, 2006, doi.org/10.1073/pnas.0601602103.

[7] M. Bouguessa and K. Nouri, "BiNeTClus: Bipartite Network Community Detection Based on Transactional Clustering," *ACM Transactions on Intelligent Systems and Technology*, vol. 12, no. 1, pp. 1 - 26, 2020, doi.org/10.1145/3423067.

[8] F. R. Khawaja, Z. Zhang, Y. Memon and A. Ullah , "Exploring community detection methods and their diverse applications in complex networks: a comprehensive review," *Social Network Analysis and Mining*, vol. 14, p. 115, 2024, doi.org/10.1007/s13278-024-01274-1.

[9] T. Jendal, M. Corfixen, M. Olesen, P. Dolog, K. Hose, D. Dell'Aglio and M. Lissandrini, "The Yelp Collaborative Knowledge Graph," *CIKM '25: Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pp. 6414 - 6419, 2025, doi.org/10.1145/3746252.3761615.

[10] V. D. Blondel, J.-L. Guillaume, R. Lambiotte and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, p. 10008, 2008, doi.org/10.1088/1742-5468/2008/10/P10008.

[11] "Yelp, Open Dataset," [Online]. Available: business.yelp.com/data/resources/open-dataset. [Accessed 16 4 2026].

[12] D. Lucey , "Tapping Yelp data with Apache Drill from Mac using {sergeant}," 2020. [Online]. Available: redwallanalytics.com/2020/10/27/tapping-yelp-data-with-apache-drill-from-mac-using-sergeant/. [Accessed 17 4 2026].

[13] "Yelp Source and Description of the Data," [Online]. Available: https://the-examples-book.com/projects/data-sets/Yelp. [Accessed 16 4 2026].

[14] "Documentation, NetworkX, greedy_modularity_communities," [Online]. Available: networkx.org/documentation. [Accessed 16 4 2026].