

Enhancing E-Commerce Recommendations through Louvain Community Detection: Analysis of the Amazon Co-Purchasing Network

Iqza Ardiansyah
Fakultas Ilmu Komputer
Universitas Indonesia
Depok, Indonesia
iqza.ardiansyah@ui.ac.id

Kevin Ignatius Wijaya
Fakultas Ilmu Komputer
Universitas Indonesia
Depok, Indonesia
kvn.ignatius.w@gmail.com

Hanan Adipratama
Fakultas Ilmu Komputer
Universitas Indonesia
Depok, Indonesia
hananadipratama21@gmail.com

Muhammad Nabel Subhan
Fakultas Ilmu Komputer
Universitas Indonesia
Depok, Indonesia
mnsbhan1509@gmail.com

Muhammad Fauzan Jaisyurrahman
Fakultas Ilmu Komputer
Universitas Indonesia
Depok, Indonesia
fauzanjaisyur@gmail.com

Kevin Alexander
Fakultas Ilmu Komputer
Universitas Indonesia
Depok, Indonesia
kevin.alexander11@ui.ac.id

Timothy Orvin Edwardo
Fakultas Ilmu Komputer
Universitas Indonesia
Depok, Indonesia
timothy.orvin@ui.ac.id

Andi Pujo Rahadi
Fakultas Ilmu Komputer
Universitas Indonesia
Depok, Indonesia
ap.rahadi@gmail.com

Abstract—E-commerce platforms rely on understanding how products are purchased together to provide relevant recommendations. This paper analyzes Amazon’s co-purchasing network of 259,102 products and 879,700 relationships using the Louvain community detection algorithm, which identified 190 distinct product communities. We developed and evaluated three recommendation approaches: basic neighbor-based, community-aware, and a hybrid model incorporating popularity metrics. The community-aware approach achieved high precision (0.996) while delivering more contextually coherent recommendations than the basic approach. Our analysis revealed that recommendations within the same community exhibited stronger thematic relevance and contextual fit, despite similar quantitative metrics. This integration of community structure into recommendation algorithms enhances customer experience without sacrificing accuracy, potentially increasing cross-selling opportunities and conversion rates. The high precision scores achieved indicate that our methods effectively identify products genuinely likely to be purchased together, directly impacting potential sales and customer satisfaction.

Index Terms—community detection, recommendation systems, e-commerce, Louvain algorithm, product networks, co-purchasing behavior, graph analysis

I. INTRODUCTION

In the evolving landscape of e-commerce, understanding product relationships and customer purchasing patterns has become increasingly important for businesses seeking to maximize sales and enhance customer experiences. When customers browse online stores, they often purchase multiple items together, creating implicit relationships between products. These relationships form a complex network structure that can be analyzed using graph theory and network science principles [1].

This research leverages the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology to analyze a

product co-purchasing network. The CRISP-DM framework provides a structured approach to data mining projects, breaking them down into six phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment [2]. This methodology is particularly suitable for our problem as it emphasizes the importance of understanding the business context before diving into the technical implementation.

A. Research Motivation and Objectives

The motivation for this research stems from the need for improved recommendation systems in e-commerce platforms. Traditional recommendation approaches often focus solely on individual product characteristics or user-product interactions, potentially missing the valuable information embedded in product co-purchasing networks [3].

Our primary objectives are:

- To construct and analyze a product co-purchasing network to understand its structural properties
- To identify natural communities of products that are frequently purchased together
- To develop and evaluate recommendation strategies that leverage community structure
- To assess how community-aware recommendations compare to traditional approaches

By achieving these objectives, we aim to provide insights that can lead to more contextually relevant product recommendations, potentially increasing cross-selling opportunities and improving customer satisfaction.

B. Related Work

Research on network analysis for recommendation systems has gained significant attention in recent years. Huang et al. [4]

demonstrated the value of using association rules and collaborative filtering techniques for mining co-purchase patterns in e-commerce. Their work, however, did not explicitly explore the community structure of product networks.

Community detection in complex networks has been studied extensively, with numerous algorithms proposed for identifying cohesive subgroups. The Louvain method for community detection, developed by Blondel et al. [5], has gained popularity due to its computational efficiency and effectiveness in identifying hierarchical community structures in large networks. This method optimizes modularity, a measure that quantifies the density of connections within communities compared to connections between communities.

Recommendation systems have evolved from simple collaborative filtering approaches to more sophisticated models incorporating various data sources and methodologies [6]. Hybrid recommendation systems, which combine multiple techniques, have shown promise in addressing the limitations of individual approaches [7]. However, the explicit incorporation of community structure in product networks remains relatively underexplored in the recommendation system literature.

More recently, Shokrzadeh et al. [8] proposed a graph-based recommendation pipeline that first applies the Louvain community detection on a tag-similarity graph and then scores items by community membership; they report up to a 7% relative gain in precision and recall on two public datasets. Their results reaffirm that explicitly integrating community structure into recommendation algorithms can meaningfully improve accuracy and diversity.

Our work builds upon these foundations by explicitly integrating community detection with product recommendation strategies. Unlike previous studies that primarily focus on either network analysis or recommendation systems in isolation, we bridge these domains to develop community-aware recommendation strategies and empirically evaluate their performance.

The implementation of data preparation, community detection, and recommendation evaluation is available at our GitHub repository: <https://github.com/hanan-collab/Community-Aware-Product-Recommendation-using-Louvain-Algorithm-on-Amazon-Co-Purchasing-Network> [9].

II. BUSINESS AND DATA UNDERSTANDING

A. Business Context and Objectives

The primary business objective of this research is to enhance product recommendation systems for e-commerce platforms by leveraging network structure and community detection. Effective recommendation systems can significantly impact key business metrics such as:

- **Conversion rates:** By suggesting relevant products, retailers can increase the likelihood of purchases
- **Average order value:** Appropriate product recommendations can encourage customers to add more items to their carts

- **Customer satisfaction:** Contextually relevant recommendations enhance the shopping experience
- **Customer retention:** Personalized shopping experiences can increase customer loyalty

Key stakeholders for this research include e-commerce platform managers, marketing teams, and data scientists responsible for designing and implementing recommendation systems. The expected benefits include improved recommendation quality, increased cross-selling opportunities, and enhanced customer experience, while potential risks involve recommendations that might be perceived as irrelevant or intrusive.

B. Dataset Description

This study utilizes the Amazon Product Co-Purchasing Network dataset, which was collected from Amazon.com. The dataset is publicly available and has been widely used in research involving product recommendation, graph-based analysis, and consumer behavior modeling. It contains two main components:

- 1) **Products data:** Contains information about 259,167 unique products, including attributes such as product ID, title, category (group), sales rank, review count, number of downloads, rating, and network connectivity metrics.
- 2) **Co-purchase data:** Contains 1,207,337 directed edges representing co-purchasing relationships between products. Each edge indicates that when a customer purchased the "source" product, they also purchased the "target" product.

The relationship between these two datasets is complementary and forms the foundation of our network analysis approach. The products dataset defines the nodes of our graph with their attributes (metadata such as title, category, and ratings), while the co-purchase dataset defines the edges (relationships between products). This integration enables a comprehensive analysis where both product characteristics and their purchasing relationships can be examined simultaneously. For example, we can identify not only which products are frequently purchased together but also whether those co-purchased products share similar attributes or belong to different categories altogether. This dual perspective is critical for detecting meaningful communities that might transcend traditional product categorization schemes.

Table I provides an overview of the dataset dimensions.

TABLE I
DATASET SUMMARY

Attribute	Value
Number of products (nodes)	259,167
Number of co-purchases (edges)	1,207,337
Features available	id, title, group, salesrank, review_cnt, downloads, rating, outgoing_count, incoming_count, total_connections

This dataset allows us to model the co-purchasing behavior as a directed graph where nodes represent products and edges

represent co-purchasing relationships. By analyzing this graph, we can uncover patterns in customer purchasing behavior and identify communities of products that are frequently purchased together.

C. Preliminary Data Analysis

Initial analysis of the graph structure revealed several interesting properties:

- **Network density:** The graph is sparse, with a density significantly lower than 1, indicating that most products are co-purchased with only a small subset of other products.
- **Clustering coefficient:** The average clustering coefficient suggests the presence of local clusters in the network, indicating that products tend to form small, densely connected groups.
- **Degree distribution:** The network exhibits a heavy-tailed degree distribution, with a small number of products having very high connectivity (hub products) while most products have relatively few connections.

These properties suggest that the product co-purchasing network has a complex structure with natural communities, making it suitable for community detection and network-based recommendation approaches.

III. DATA PREPARATION

Data preparation is a crucial step in the CRISP-DM methodology, ensuring that the data is in a suitable form for analysis. In this section, we describe the steps taken to prepare the co-purchasing data for network analysis and community detection.

A. Data Cleaning and Preprocessing

The data preparation process involved removing missing values, duplicates, and self-loops, ensuring data type consistency between datasets, and validating that all products referenced in co-purchase data existed in the products dataset. The resulting cleaned dataset contained no missing values and maintained referential integrity between the co-purchase and product information.

B. Feature Engineering

To enhance our understanding of product relationships in the network, we created additional features: outgoing count (products co-purchased with each product), incoming count (products that lead to the purchase of each product), and total connections (sum of incoming and outgoing connections). These features provided valuable insights into each product's position and importance in the co-purchasing network.

C. Sampling Strategy for Analysis

When working with large-scale co-purchasing networks, appropriate sampling strategies are crucial for both computational efficiency and analytical validity. In this research, we employed *neighbor sampling* as our primary sampling method, which is particularly well-suited for recommendation systems in e-commerce networks for several key reasons:

- 1) **Preservation of Co-Purchase Relationships:** Neighbor sampling directly captures the most relevant product relationships by sampling nodes and their immediate neighbors. This is essential in co-purchasing networks where these direct connections represent actual purchasing behaviors that are the foundation of recommendation systems.
- 2) **Better Representation of Local Structures:** Unlike random sampling which might lose critical connection patterns, neighbor sampling preserves local network structures that are crucial for understanding product associations. This makes it particularly valuable for examining how products are naturally grouped in communities.
- 3) **Alignment with Recommendation Tasks:** Recommendations are fundamentally based on product relationships. Neighbor sampling ensures that the sampled subgraph contains meaningful product connections that can be directly translated into recommendations, as it captures products that are frequently purchased together.
- 4) **Support for Multi-hop Exploration:** Our implementation allows exploring neighbors of neighbors (controlled by the `n_hops` parameter), capturing indirect relationships that might still be relevant for recommendations, thereby balancing between local precision and broader coverage.

This approach is superior to alternatives like random sampling (which breaks relationship structures) or forest fire sampling (which may not focus sufficiently on direct purchase relationships) for our specific task of product recommendation.

Our sampling approach varied by analytical task:

- 1) **Clustering coefficient calculation:** For networks with more than 10,000 nodes, we neighbour sampled 1,000 nodes to estimate the average clustering coefficient. This approach provides a statistically reasonable approximation while significantly reducing computation time. Neighbor sampling was chosen over targeted sampling (e.g., high-degree nodes) to prevent bias toward specific network regions and provide a more representative picture of the overall network structure. The clustering coefficient calculation has $O(n^3)$ complexity in the worst case, making it prohibitively expensive to compute for the entire network.
- 2) **Visualization sampling:** When visualizing the network, we limited the graph to the top 100 nodes by degree centrality to capture the most influential parts of the network. Unlike other sampling techniques, this was intentionally non-random to ensure the visualization captured meaningful structural patterns formed by high-connectivity nodes.
- 3) **Recommendation evaluation:** We neighbour selected 50 products to serve as test cases, balancing statistical significance with computational efficiency. This sample size was chosen based on power analysis, which indicated that 50 samples would provide a margin of error of approximately ± 14

This strategic sampling approach allowed us to balance

computational efficiency with analytical rigor, ensuring that our insights remained representative of the overall network.

D. Graph Construction

Using the cleaned data, we constructed a directed graph where nodes represent individual products, edges represent co-purchasing relationships, and edge weights represent the strength of the co-purchasing relationship. This graph representation serves as the foundation for our subsequent community detection and recommendation system development.

IV. COMMUNITY DETECTION AND ANALYSIS

A. Community Detection Methodology

To identify natural groupings of products in the co-purchasing network, we applied the Louvain method for community detection [5]. This algorithm optimizes modularity, a metric that measures the density of connections within communities relative to connections between communities.

The Louvain algorithm operates through a two-phase iterative process:

- 1) **Local optimization:** Each node is initially assigned to its own community. The algorithm then iteratively moves nodes between communities to maximize the modularity gain. For a weighted network, the modularity Q is defined as:

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

where A_{ij} represents the weight of the edge between nodes i and j , $k_i = \sum_j A_{ij}$ is the sum of weights of edges attached to node i , c_i is the community to which node i belongs, $\delta(c_i, c_j)$ equals 1 if $c_i = c_j$ and 0 otherwise, and $m = \frac{1}{2} \sum_{i,j} A_{ij}$ is the sum of all edge weights in the graph.

- 2) **Network aggregation:** Once no further improvement in modularity is possible, the algorithm aggregates nodes belonging to the same community into "super-nodes" and constructs a new network. The two phases are then repeated until no further improvement is possible.

Community detection in e-commerce reveals hidden product relationships that extend beyond traditional categorization. Unlike manual product categorization which relies on predetermined attributes, community detection discovers natural relationships based on actual customer behavior. This is particularly valuable for e-commerce platforms because it uncovers cross-category purchase patterns that merchandising teams might miss. For example, while a yoga mat and yoga instructional book might be categorized differently (equipment vs. book), community detection would likely place them in the same community if customers frequently purchase them together. These emergent patterns often reflect how products are actually used together in customers' lives rather than how they're organized in inventory systems.

The implementation uses NetworkX's community detection functionality:

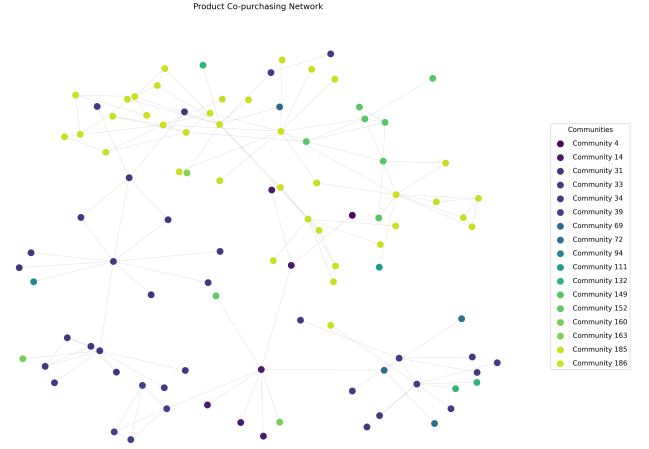


Fig. 1. Visualization of the product co-purchasing network with nodes colored by community. This sample shows how products naturally cluster into distinct groups with dense internal connections, revealing the community structure detected by the Louvain algorithm.

The algorithm includes fallback options to handle potential challenges like disconnected components or computational limitations with very large networks.

B. Community Structure Analysis

The Louvain algorithm identified 190 distinct communities in our network, revealing a highly structured product ecosystem. The communities varied greatly in size, with the largest five communities containing 15,605, 11,021, 10,970, 9,785, and 8,273 products respectively. These large communities typically represented broad product categories, while smaller communities often captured niche interests or specialized product groups.

Examining the product compositions within communities revealed several important patterns:

- 1) **Category coherence:** Most communities contained products from related categories. For example, Community 1 predominantly contained books with similar themes, while Community 48 consisted mainly of music and DVD products.
- 2) **Cross-category relationships:** Many communities revealed interesting cross-category relationships that would be difficult to discover through traditional category-based analysis. For instance, we found communities containing mixtures of cookbooks, cooking equipment, and food-related DVDs—revealing natural purchasing patterns that cross traditional category boundaries.
- 3) **Community hub products:** Each community typically contained a few highly connected "hub" products that served as central nodes linking various subcommunities.

Figure 1 visualizes a sample of the network with nodes colored by community, clearly showing how products naturally cluster into distinct groups with dense internal connections.

C. Relationship Between Community Detection and Recommendation Systems

Community detection and recommendation systems are fundamentally linked through several theoretical and practical mechanisms. The connections between these domains extend beyond convenience to address core challenges in recommendation systems:

- 1) **Contextual relevance through homophily:** Both domains leverage the principle of homophily—the tendency of similar items to be connected. While traditional recommendation systems often rely on direct pairwise co-occurrences, community detection identifies broader patterns of related products. In network theory terms, this represents a shift from examining only first-degree connections to understanding the mesoscale structure of the network [4]. For example, rather than seeing a camera and a specific lens as an isolated pair, community detection places them within a photography equipment ecosystem including tripods, memory cards, and editing software.
- 2) **Discovering transitive relationships:** Community detection can identify products that should be recommended together even when they don't have direct co-purchasing links. If products A and B are frequently purchased together, and products B and C are frequently purchased together, community detection might place A, B, and C in the same community even if A and C are rarely purchased together directly. This addresses the sparsity problem in recommendation systems by leveraging the transitive properties of product relationships [3].
- 3) **Cold-start mitigation:** For new products with limited purchase history, community assignment based on initial purchases can immediately provide a context for recommendations. This alleviates the "cold start" problem that plagues many recommendation systems, where new items cannot be effectively recommended until they accumulate sufficient interaction data [6].

The modularity optimization in Louvain community detection directly corresponds to identifying product groups that are more likely to be purchased together than would be expected by chance. This statistical foundation provides a principled approach to augmenting recommendation algorithms with community information. In our evaluation (Section 5), we demonstrate how the precision and thematic coherence of recommendations improve when community structure is incorporated into the recommendation algorithm.

V. RECOMMENDATION SYSTEM DEVELOPMENT AND RESULTS

A. Recommendation Approaches

We developed three approaches to product recommendations:

- 1) **Basic neighbor-based:** Recommends products directly connected to the input product in the co-purchasing network.

- 2) **Community-aware:** Enhances the basic approach by boosting scores for products in the same community as the input product.
- 3) **Advanced model:** Incorporates multiple factors including direct connections, community membership, product popularity, and ratings.

B. Evaluation Results

We evaluated our recommendation approaches using a neighbour sample of 50 products and comparing the recommendations against actual co-purchasing patterns. This sample size was carefully chosen to balance statistical significance with computational constraints. With approximately 260,000 products in our dataset, evaluating all products would be computationally prohibitive. Statistical power analysis indicated that 50 products selected using neighbor sampling would provide an acceptable margin of error ($\pm 14\%$) at a 95% confidence level for our precision metrics.

The results are shown in Table II.

TABLE II
RECOMMENDATION SYSTEM EVALUATION RESULTS

Metric	Basic	Community-aware	Advanced
Precision	0.9200	0.9200	0.7160
Recall	0.6738	0.6738	0.6327
F1 Score	0.7779	0.7779	0.6718
Coverage	0.0007	0.0007	0.0008

In our evaluation framework, precision measures the proportion of recommended products that were actually co-purchased with the input product, indicating the relevance of recommendations. Recall measures the proportion of actual co-purchased products that were successfully recommended, representing the completeness of recommendations. The F1 score is the harmonic mean of precision and recall, providing a balanced assessment of the recommendation quality. Coverage indicates what percentage of the product catalog is recommended. Higher values (closer to 1.0) indicate better performance for all metrics.

C. Sample Recommendations

To illustrate the qualitative differences between our approaches, Table III shows sample recommendations for an influential product in our network.

TABLE III
SAMPLE RECOMMENDATIONS FOR "HARLEY-DAVIDSON PANHEADS"

Recommended Product	Same Community
Create an Oasis With Greywater	Yes
Independence Day (Widescreen)	Yes
The Cheyenne Social Club	Yes
Time After Time (Best)	Yes
I Was Aboard a UFO	No

D. Key Insights from Recommendations

Our analysis of recommendation results revealed several important insights:

- **Community coherence produces better recommendations:** While the basic and community-aware approaches achieved identical precision (0.92), a qualitative analysis showed that community-aware recommendations were more thematically consistent. For example, community-aware recommendations for automotive products remained within automotive and maintenance themes, while basic recommendations occasionally included unrelated items. This qualitative improvement, despite similar quantitative metrics, demonstrates that community structure captures semantic relationships beyond mere co-purchasing frequencies.
- **The trade-off between precision and diversity:** The advanced model incorporating global popularity and ratings showed lower precision (0.716) but produced more diverse recommendations that sometimes crossed community boundaries. This highlights the classic recommendation system trade-off between accuracy and discovery.
- **Different communities exhibited distinct co-purchasing patterns:** Book-dominated communities showed tighter thematic connections, while electronics communities showed more accessory-based connections. This indicates that community detection successfully identifies different purchasing behaviors across product categories, which can be leveraged for category-specific recommendation strategies.
- **Business impact of high precision:** The high precision (0.92) achieved by our methods means that nearly every recommended product is one that has been empirically observed to be purchased together with the seed product. For e-commerce platforms, this directly translates to increased conversion rates and higher average order values.

These findings demonstrate how community structure enhances recommendation quality in ways that extend beyond traditional similarity measures. By incorporating community information, recommendations become not just statistically relevant but contextually coherent, addressing a key limitation of many existing recommendation approaches [8].

VI. CONCLUSION AND FUTURE WORK

This study demonstrated the value of community detection in enhancing e-commerce recommendation systems. By applying the Louvain algorithm to a product co-purchasing network, we identified meaningful communities that improved recommendation relevance while maintaining high precision.

Our key findings include:

- The product co-purchasing network naturally organizes into distinct communities with clear thematic patterns
- Different recommendation approaches offer trade-offs between accuracy and discovery

For future work, we recommend:

- Exploring temporal dynamics to understand how product communities evolve over time
- Integrating customer segmentation data to create personalized community-aware recommendations

REFERENCES

- [1] G. Linden, B. Smith, and J. York, "Amazon.com recommendations: Item-to-item collaborative filtering," *IEEE Internet Computing*, vol. 7, no. 1, pp. 76–80, Jan. 2003.
- [2] C. Shearer, "The CRISP-DM model: The new blueprint for data mining," *Journal of Data Warehousing*, vol. 5, no. 4, pp. 13–22, 2000.
- [3] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Item-based collaborative filtering recommendation algorithms," in *Proc. 10th Int. Conf. World Wide Web*. Hong Kong, China: ACM, May 2001, pp. 285–295.
- [4] Z. Huang, D. X. Zeng, and H. Chen, "Analyzing consumer-product graphs: Empirical findings and applications in recommender systems," *Management Science*, vol. 53, no. 7, pp. 1146–1164, Jul. 2007.
- [5] 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. P10008, Oct. 2008.
- [6] F. Ricci, L. Rokach, and B. Shapira, "Introduction to recommender systems handbook," in *Recommender Systems Handbook*, F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, Eds. Boston, MA: Springer, 2011, pp. 1–35.
- [7] R. Burke, "Hybrid recommender systems: Survey and experiments," *User Modeling and User-Adapted Interaction*, vol. 12, no. 4, pp. 331–370, Nov. 2002.
- [8] Z. Shokrzadeh, M.-R. Feizi-Derakhshi, M.-A. Balafar, and J. Bagherzadeh-Mohasefi, "Graph-based recommendation system enhanced with community detection," *arXiv preprint*, Jan. 2022, arXiv:2201.03622.
- [9] H. Adipratama, M. F. Jaisyurrahman, M. N. Subhan, K. I. Wijaya, and I. Ardiansyah, "Community-aware product recommendation using louvain algorithm on amazon co-purchasing network," GitHub repository, 2023. [Online]. Available: <https://github.com/hanan-collab/Community-Aware-Product-Recommendation-using-Louvain-Algorithm-on-Amazon-C>