Universiteit Leiden
The Netherlands

Bachelor Data Science and Artificial Intelligence

Network Structure of Skincare Formulations and its Association with Product Popularity

Jia Jia Zhang

Supervisors:
Dr. Akrati Saxena & Ella Has

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

01/07/2026

Abstract

This thesis investigates whether the structural organization of skincare formulations is associated with product popularity. Skincare products are represented as ingredient co-occurrence networks, where ingredients are modeled as nodes and edges indicate that two ingredients appear together in the same product. Using skincare product data from SkinSort and Sephora, the study constructs a weighted ingredient network, detects ingredient communities, and derives product-level network features including average degree centrality, average betweenness centrality, number of communities, and community proportions. The associations of these features with product ratings and review counts are studied using correlation analysis, linear regression, robustness checks, and machine learning models.

The results show that skincare formulations contain recurring ingredient combination patterns. The detected communities broadly correspond to recognizable formulation contexts, including base, moisturizing, texture-enhancing, and active ingredient groups. However, the relationship between network structure and product popularity is modest. Individual network features show weak bivariate correlations with ratings and review counts, although average degree and average betweenness are associated with product rating when modeled together. The robustness check suggests that the degree association is more stable than the betweenness association after removing water and glycerin. In the machine learning analysis, ingredient-based features generally outperform network and community features alone. Adding network features to ingredient features provides limited additional predictive value. Overall, network analysis is useful for describing skincare formulation structure, but it only partly explains product popularity.

Contents

1 Introduction **1**
- 1.1 Thesis Overview . . . 2
  - 1.1.1 Research Question . . . 2

2 Definitions **3**
- 2.1 Co-occurrence Networks . . . 3
- 2.2 Network Metrics . . . 3
- 2.3 Product Popularity . . . 4

3 Related Work **5**
- 3.1 Network Approaches to Ingredient Combinations . . . 5
  - 3.1.1 A PageRank Approach to Skincare Ingredients . . . 6
- 3.2 Machine Learning Approaches in Skincare . . . 6

4 Methodology **8**
- 4.1 Data . . . 8
- 4.2 Data Preprocessing . . . 8
- 4.3 Network Construction . . . 9
- 4.4 Network Measures and Feature Construction . . . 9
- 4.5 Statistical Analysis . . . 10
- 4.6 Robustness Checks . . . 11
- 4.7 Machine Learning Approach . . . 11

5 Results **13**
- 5.1 Description of Network Structure . . . 13
- 5.2 Ingredient Communities . . . 13
- 5.3 Network Features and Product Popularity . . . 15
- 5.4 Community Composition and Product Popularity . . . 19
- 5.5 Robustness Checks . . . 20
- 5.6 Machine Learning Prediction . . . 21

6 Discussion **24**
- 6.1 Ingredient Network Structure and Communities . . . 24
- 6.2 Interpretation of Network Feature Associations . . . 25
- 6.3 Interpretation of Community Composition . . . 26
- 6.4 Predictive Value of Network Features . . . 27
- 6.5 Limitations . . . 28

7 Conclusion **30**
- 7.1 Future Research Direction . . . 30

References **34**

1 Introduction

Cosmetic and skincare products are formulated with combinations of chemical ingredients that serve different functions, such as hydration, exfoliation, or anti-aging. From a systems perspective, each product represents a configuration of components whose interactions shape performance and consumer perception. Unlike products that only have one feature, skincare products are combinatorial systems in which value arises from a structured combination of ingredients.

Innovation in consumer product markets often follows combinatorial dynamics: new products emerge through recombining existing components with new components [1]. Similarly, in cosmetic markets, ingredients such as retinoids, peptides, or ceramides are often recombined across brands and product categories [2]. Consequently, skincare products are not independent formulations but share ingredient combination patterns. Therefore, understanding how these combinations are structured may provide insights into broader formulation patterns.

Since skincare products consist of combinations of ingredients, they can be represented as relational systems in which ingredients are connected through their joint presence in products. Network science provides a useful framework for studying such systems, where nodes represent entities and edges represent relationships between these entities [3]. In the context of skincare formulations, ingredients can be represented as nodes, and connections can be formed when two ingredients appear together in the same product. This representation allows the overall structure of ingredient combinations to be analyzed across the entire formulation system.

Similar approaches have been applied in other domains. For example, ingredient co-occurrence networks have been studied in food science, where the analysis of recipes reveals systematic patterns in how ingredients are combined [4]. This study demonstrates that network representations can uncover structural patterns that are not immediately visible when examining individual ingredient lists alone. Therefore, these findings suggest that similar structural patterns may also exist in other combinatorial systems, such as cosmetic formulations.

While structural approaches have been applied in domains such as food science, research on cosmetic products has more often focused on predicting consumer response or product success. In consumer markets, product success is commonly determined through ratings, review volume, or number of sales. Previous research in cosmetic and marketing analytics mostly relies on regression or machine learning models to predict product success [5]. However, these approaches generally treat product attributes as independent variables and do not examine the structural relationships between them. This thesis addresses this gap by combining a network analysis of ingredient relationships with machine learning models of product popularity. A similar modelling strategy is used by Stepien et al. [6], who compare attribute-based, network-based, and hybrid models in the context of fairness in opinion dynamics. Following this logic, the current thesis compares ingredient-based, network-based, and combined feature representations.

To our knowledge, structural network analysis of skincare formulations and its association with product popularity has not yet been examined in academic literature. This creates a gap in understanding whether consumer response is related only to the presence of individual ingredients, or also to the way ingredients are combined within broader formulation patterns. If structural position provides additional insight beyond the presence of individual ingredients, this suggests that product success depends not only on which ingredients are used, but also on how they are combined.

1.1 Thesis Overview

This thesis investigates the relationship between the structural organization of skincare formulations and product popularity. Specifically, the study models ingredient combinations as a co-occurrence network in which ingredients are connected if they appear together within a product. Structural properties of this network, such as centrality [7] and community structure [8, 9], are then analyzed to examine how ingredients are positioned within the formulation network. Additionally, the study considers community composition at the product level by examining the proportion of ingredients belonging to different network communities in order to assess whether specific formulation contexts are associated with product popularity. Finally, this thesis compares three predictive approaches: an ingredient-based machine learning approach, a network approach, and a combined approach that includes both ingredient and network features. Random forest and XGBoost regression models are used to predict product popularity from ingredient composition and product-level network features. This comparison makes it possible to examine whether network features provide added predictive value beyond standard ingredient-based features.

1.1.1 Research Question

The central research question of this thesis is:

To what extent is the structural position of skincare ingredients within formulation co-occurrence networks associated with product popularity?

To support this question, the study also considers:

- What structural patterns and communities emerge in skincare ingredient co-occurrence networks?
- Which ingredients occupy central or bridging positions within these networks?
- Which structural metrics are most predictive of product success?
- Are certain ingredient communities more strongly associated to product popularity than others?
- Can product popularity be predicted from ingredient-based product data using machine learning?
- Do network features improve prediction beyond ingredient-based features?

The remainder of this thesis is structured as follows. Section 2 introduces the main concepts used in the study, including co-occurrence networks, network metrics, and product popularity. Section 3 discusses related work on ingredient networks, skincare ingredient analysis, and machine learning approaches in skincare. Section 4 describes the datasets, preprocessing steps, network construction, statistical analyses, robustness checks, and machine learning approach. Section 5 presents the results of the network analysis, community detection, regression analyses, robustness checks, and predictive models. Section 6 discusses the interpretation of these findings and the limitations of the study. Finally, Section 7 summarizes the main conclusions and Section 7.1 outlines directions for further research.

2 Definitions

This section introduces the main concepts used throughout the thesis. It defines co-occurrence networks as the basis for representing skincare formulations. Furthermore, it explains the network metrics used to describe ingredient positions and community structure. Finally, it defines product popularity and explains how it is measured in the analysis.

2.1 Co-occurrence Networks

A co-occurrence network is a representation of a system in which entities are connected based on their joint appearance within a defined unit (e.g., ingredients in the same product). In such networks, nodes represent entities and edges represent the relationships between them [3]. Co-occurrence networks are commonly used to model relational structures in systems where elements appear in combinations, such as words in documents or ingredients in recipes.
In the context of this study, skincare formulations are modeled as a co-occurrence network of ingredients. Each node in the network represents a unique ingredient and an edge is formed between two ingredients if they appear together in the same product. Thus, the resulting network captures how ingredients are combined across a large set of products.
Edges in the network are weighted to reflect the frequency of co-occurrence of two ingredients. Specifically, the weight of an edge corresponds to the number of products in which the two connected ingredients appear together. This allows the network to distinguish between ingredient pairs that rarely co-occur and those that frequently appear together across formulations.
By modeling skincare formulations as a co-occurrence network, ingredient combinations can be analyzed across products. This allows for relationships between ingredients to be studied at a broader level.

2.2 Network Metrics

To analyze the structural roles of ingredients within the co-occurrence network, several standard network metrics are used. These metrics quantify how ingredients are positioned within the network and how they relate to other ingredients. The definitions used in this study follow the standard formulations in network science [3].

Degree centrality Degree centrality measures the number of connections a node has to other nodes in the network. In a weighted co-occurrence network, this corresponds to the number of distinct ingredients that a given ingredient co-occurs with across products. Ingredients with high degree centrality are connected to many other ingredients and can therefore be interpreted as widely used or highly compatible components in formulations.

Betweenness centrality Betweenness centrality measures how often a node lies on the shortest paths between other nodes in the network. It is defined as the fraction of shortest paths between all pairs of nodes that pass through a given node. Consequently, ingredients with high betweenness centrality may act as bridging components, connecting different groups of ingredients that would be weakly connected otherwise.

Community structure Community structure refers to the presence of groups of nodes that are more densely connected to each other than to the rest of the network. To identify such groups, community detection algorithms, such as the Leiden algorithm [10], can be used to identify communities of frequently co-occurring ingredients. The Leiden algorithm starts from an initial partition of the network and repeatedly improves it by moving nodes between communities when this increases the quality of the partition. It then refines the communities to ensure that they are internally well connected and aggregates them into a smaller network, after which the process is repeated. Consequently, the algorithm detects groups of ingredients that co-occur more strongly with each other than with ingredients outside the group. These communities may represent recurring formulation patterns, such as groups of ingredients that are commonly used together because they have similar functions or are part of standard skincare formulations.

Together, these metrics provide different perspectives on the structure of the ingredient network by identifying highly connected ingredients, bridging ingredients, and groups of ingredients that frequently co-occur.

2.3 Product Popularity

In this study, product popularity refers to consumer response to a product rather than direct financial outcomes such as revenue or sales. Since product popularity is itself is not directly observed, it is measured using two indicators available in the datasets: product ratings and review counts. Ratings reflect the average evaluation given by consumers, while review counts reflect the amount of consumer engagement that a product has.
These measures are commonly used as indicators of popularity on online consumer platforms [11] because they capture both perceived quality (through ratings) and attention (through the number of reviews). Together, these variables reflect both how well a product is received and how widely it is engaged with by consumers.
In the analysis, these variables are used as outcome measures to examine whether the structural properties of ingredient networks are associated with differences in product popularity.

3 Related Work

Network analysis has been used to analyze the dynamics of various complex networks, including social [12, 13], collaboration [14, 15], social media [16, 17], food [18], transaction [19], terrorist and criminal [20, 21, 22], and web [23, 24] networks. In this work, we use network analysis to study the networks of skincare ingredients. This section reviews prior research that is relevant to the present study. First, it discusses network approaches to ingredient combinations, with a focus on how formulation systems can be represented as relational structures. It then considers existing work on skincare ingredient networks and machine learning approaches in skincare. These studies provide the basis for constructing ingredient co-occurrence networks, identifying patterns in formulations, and applying machine learning models to predict product characteristics from ingredient compositions.

3.1 Network Approaches to Ingredient Combinations

A useful starting point for the present study is the paper by Ahn et al. [4], which applies network science to food composition. Their work is relevant because it demonstrates how ingredient-based products can be represented as networks and analyzed as structured systems. The paper investigates whether there are systematic principles behind why certain ingredients are combined in recipes, rather than treating culinary practice as purely cultural or intuitive. To study this, the authors construct a flavour network in which ingredients are connected if they share flavour compounds. Specifically, they construct a bipartite network that links ingredients to chemical compounds and then project this into an ingredient-ingredient network, where edges represent compound overlap between ingredients. This is relevant to the present thesis because it demonstrates how a formulation-based domain can be translated into a relational network representation, making it possible to study combinations of ingredients as a structured system instead of isolated items.
Furthermore, the authors use this network to test their hypothesis that ingredients sharing many flavour compounds are more likely to taste good together, and are therefore more likely to co-occur in recipes. Using a dataset of 56,498 recipes from several cuisines, they compare the observed ingredient combinations in real recipes with randomized baselines. This methodological step is important, because it evaluates whether the observed network patterns are statistically meaningful. Apart from revealing systematic patterns, the study also demonstrates that ingredient networks reveal meaningful variations across groups. Their results show that the relationship between ingredient similarity and co-occurrence is not the same across cuisines: Western cuisines tend to combine ingredients that share many compounds, whereas East Asian cuisines tend to avoid such combinations.
The paper is methodologically relevant because it shows how a large ingredient network can be simplified and interpreted. Ahn et al. do this by reducing the full network to the most important ingredient relationships and examining which ingredients contribute most to the combinations found in recipes. This illustrates how network methods can be used to move from a complex set of ingredient relationships to interpretable patterns. Since skincare formulations also contain many ingredient combinations, a similar network approach can be used to identify recurring combinations and influential ingredients within the skincare ingredient network.
For the current thesis, the main value of Ahn et al. lies in their methodological approach. However, the focus of this thesis is different. Rather than examining whether shared chemical properties

explain why ingredients are combined, this study investigates whether the structural position of ingredients in a co-occurrence network is associated with product popularity. Therefore, Ahn et al.'s work serves as a strong foundation for this thesis: they offer a clear proof of concept for studying formulation-based products through the lens of network science.

3.1.1 A PageRank Approach to Skincare Ingredients

A more recent study in the skincare domain is provided by Frank [25]. In contrast to work that focuses on ingredient pairing principles, this paper examines the prominence of skincare ingredients across product categories by applying the PageRank algorithm to ingredient data from 1,472 Sephora cosmetic products. The analysis covers several skincare categories, including cleansers, eye creams, face masks, moisturizers, and sun protection products.

This study shows another way in which network methods can be applied to skincare research. Methodologically, Frank constructs a bipartite graph in which one set of nodes represents products and the other represents ingredients. An edge is placed between a product and an ingredient when that ingredient is present in the product. Furthermore, the author converts this structure into a matrix representation and applies PageRank iteratively until convergence in order to estimate the prominence of ingredients in the network.

The results of the study highlight several ingredients that appear repeatedly across categories, such as water, oil, glycerin, and plant extracts. The author interprets these as multifunctional ingredients that are widely used across formulations because of their properties. The category-specific PageRank results also show that some ingredients are especially prominent within particular product categories, such as zinc oxide in sun protection products. This is particularly relevant because highly common ingredients may disproportionately shape the structure of skincare networks and dominate the ranking of important nodes. Therefore, the study shows that network analysis in skincare can be used to identify prominent ingredients and recurring formulation patterns across product categories.

3.2 Machine Learning Approaches in Skincare

A third relevant line of work concerns machine learning approaches for skincare product recommendation, in which product information and skin characteristics are used as input to predict suitable products or product effects. Within this line of work, Lee et al. [26] propose a deep learning-based skincare product recommendation system. Their study combines two sources of information: cosmetic ingredient data and AI-based facial skin condition analysis. The aim is to estimate the likely effects of skincare products based on their ingredients and then use that information in a personalized recommendation system. In contrast to studies that treat cosmetic ingredients as purely descriptive product information, Lee et al. use ingredient data as direct input for a predictive model.

A key contribution of this paper is that it demonstrates the efficacy of applying machine learning to skincare products using ingredient information. The authors report that their recommendation system performed effectively, and several examples illustrate its ability to make reasonable recommendations for different skin concerns, such as dryness and acne. Therefore, the study suggests that ingredient formulations contain meaningful predictive information, especially when combined with other user-related data. This is important because it shows that cosmetic ingredients can be incorporated into a data-driven recommendation system.

Additionally, the study is relevant because it reflects a shift in skincare analytics toward personalized and AI-supported decision-making. This shift is visible in the way the authors move beyond traditional product comparisons that rely only on general product characteristics. Rather than comparing products solely based on their composition or function, they integrate product composition with skin condition analysis to model the relationship between product characteristics and user needs. Thus, their work is an example of how machine learning can be applied in the skincare domain to generate personalized predictions and recommendations.

However, the focus of Lee et al. is different from that of the current thesis. Their paper mainly focuses on personalized skincare recommendation, whereas this thesis examines product popularity and includes machine learning as a way to evaluate whether ingredient-based features and network features are useful for prediction of popularity. Therefore, their research supports the machine learning component of this thesis by motivating the use of predictive models to evaluate whether ingredient composition, network features, or their combination can predict product popularity.

4 Methodology

This section describes the methodological steps used to examine the relationship between skincare formulation structure and product popularity. The analysis consists of three main parts. Firstly, ingredient lists are used to construct a co-occurrence network and calculate ingredient network features. Secondly, these measures are aggregated to the product level and related to ratings and review counts using correlation and regression analysis. Finally, machine learning models are used to predict product popularity using three types of input: individual ingredient presence, product-level network features, and a combination of both.

4.1 Data

The analysis in this thesis is based on two skincare product datasets: *SkinSort* [27] and *Sephora products and skincare reviews* [28]. These datasets are chosen because they complement each other in terms of the information they provide. The SkinSort dataset offers detailed information on skincare formulations, while the Sephora dataset provides measures of product popularity, such as ratings and review counts.

The SkinSort dataset is used to examine the structural organization of skincare formulations. It contains roughly 19,050 skincare products with associated ingredient lists, as well as additional values such as brand, product name, and product type. This makes it suitable for constructing an ingredient co-occurrence network. Furthermore, the size and diversity of the dataset allow for recurring formulation patterns, such as ingredients that frequently appear together in moisturizers, to be identified across a broad set of skincare products.

The Sephora dataset is used to incorporate measures of product popularity into the analysis. It includes product names, brand names, ratings, number of reviews, and ingredient lists. Although the datasets contain additional metadata, this study only uses the variables relevant to network construction, popularity measurement, and predictive modeling. These variables make it possible to examine whether network characteristics of formulations are associated with product popularity, and to evaluate predictive performance in the machine learning component.

Thus, the two datasets provide the basis for the main analyses of the study. SkinSort is used to construct the ingredient co-occurrence network, while Sephora is used to relate ingredient information to product popularity through ratings, review counts, and predictive models. Prior to analysis, the datasets are cleaned and preprocessed to standardize ingredient information and prepare the data for network construction.

4.2 Data Preprocessing

The datasets are preprocessed in order to obtain a consistent and usable representation of skincare ingredients for each product. Since the ingredient lists form the basis of the analysis, preprocessing mainly focuses on cleaning, standardizing, and restructuring these data.

For the SkinSort dataset, only the variables relevant to the analysis are retained, namely brand, product name, and ingredient list. Additionally, a unique product identifier is created by combining the brand name and product name. The raw ingredient strings are split into separate ingredients using commas as delimiters. After splitting, ingredient names are converted to lowercase and leading and trailing spaces are removed. Moreover, empty elements are excluded from the resulting lists.

A correction step is included to handle ingredient names containing numerical expressions, such as ingredients of the form '*1,2-...*'. In these cases, numerical tokens that have been incorrectly separated by a comma are merged again to ensure that the full ingredient name is preserved. After cleaning, duplicate ingredients within the same product are removed to ensure that each ingredient appears at most once per formulation.
For the Sephora dataset, the data are first restricted to products belonging to the skincare category. From this subset, only the variables required for the analysis are retained: product identifier, product name, brand name, ingredient list, rating, and review count. The ingredient lists in this dataset require additional cleaning because they are stored in a less standardized format. Brackets and quotation marks are removed, content inside parentheses is deleted, and special characters such as asterisks are excluded. Furthermore, the cleaned strings are split into individual ingredients at commas, converted to lowercase, stripped of surrounding whitespace, and stripped of trailing punctuation. As with the SkinSort data, duplicate ingredients within each product are removed, and the same correction step is applied for ingredients containing numerical expressions.
After preprocessing, both datasets contain a standardized ingredient list for each product. Subsequently, these cleaned ingredient lists are used as the basis for the analyses described in the following subsections.

4.3 Network Construction

In order to represent the structural organization of skincare formulations, an ingredient co-occurrence network is constructed. The network is based on the cleaned ingredient lists from the SkinSort dataset. Each node in the network corresponds to a unique ingredient, and an undirected edge is created between two ingredients when they co-occur in the same product. In this way, the network captures how ingredients are connected through their joint use across formulations. The network is implemented in Python using the `NetworkX` library [29].
To construct the network, all unique ingredient pairs are generated for each product based on its cleaned ingredient list. If two ingredients appear in the same product, they are counted as a co-occurring pair. Afterward, the frequency with which each pair appears across products is counted. These frequencies are used as edge weights, resulting in a weighted undirected graph in which stronger edges represent ingredient pairs that co-occur more often in the dataset.
After the initial graph is created, weak edges are removed to reduce noise and emphasize recurring formulation relationships. Specifically, only ingredient pairs with a minimum co-occurrence frequency of two are retained. By removing ingredient pairs that appear only once, the filtering step reduces the influence of rare co-occurrences and emphasizes ingredient relationships that recur across multiple products.
The filtered co-occurrence network serves as the basis for the structural analyses presented in the following subsections.

4.4 Network Measures and Feature Construction

After the co-occurrence network is constructed, network measures are computed to describe the structural role of ingredients within the formulation network. As mentioned in section 2, degree centrality, betweenness centrality, and community membership are used to describe the structural position of ingredients within the co-occurrence network and to construct product-level network

features. These measures are selected because they capture different aspects of node position in the network and are later aggregated to the product level to examine their association with product ratings and review counts.

Degree centrality and betweenness centrality are calculated using the `NetworkX` library [29]. Moreover, betweenness centrality is computed using node sampling in order to reduce computational complexity. Community membership is determined using the Leiden algorithm [10] on the weighted co-occurrence network, which groups together ingredients that are more strongly connected to each other than to the rest of the network.

To support the interpretation of the detected communities, a label check is performed. For each community, the 20 ingredients with the highest weighted degree are compared with cosmetic ingredient functions listed in the European Commission's CosIng database [30, 31]. These function categories describe the reported roles of cosmetic ingredients, such as humectant, emollient, preservative, or antioxidant. The resulting fit percentage is used only as an approximate check of the community labels.

The resulting network measures are stored as mappings between ingredients and their corresponding values. Subsequently, these mappings are used to construct product-level variables based on the ingredients contained in each product. More specifically, the degree centrality and betweenness centrality values of the ingredients in a product are aggregated by taking their average, resulting in the variables μ_{degree} and $\mu_{betweenness}$. In addition, the number of distinct communities represented within a product is calculated, resulting in the variable $n_{communities}$. To capture community composition at the product level, the proportion of ingredients belonging to each detected community is also calculated for every product. These proportions are used to assess whether products with a larger share of ingredients from certain formulation clusters differ in popularity.

These aggregated variables make it possible to study whether the network position and community composition of a product's ingredients are related to outcomes such as ratings and review counts. In this way, structural information from the ingredient network is translated into features that can be used in both the statistical analysis and the machine learning component of the study.

4.5 Statistical Analysis

To examine whether the structural characteristics derived from the ingredient co-occurrence network are associated with product popularity, the study applies both correlation analysis and linear regression.

Firstly, a correlation matrix is computed for the main variables used in the analysis, namely ratings, review counts, average degree centrality, average betweenness centrality, and the number of ingredient communities. Pearson correlation coefficients are used to measure the strength and direction of the linear association between each pair of variables. Furthermore, the statistical significance of these correlations is evaluated using corresponding p-values. Since the correlation matrix involves multiple simultaneous significance tests, a Bonferroni correction [32] is applied to account for multiple comparisons. This means that a stricter significance threshold is used, which reduces the likelihood of type I errors, that is, incorrectly identifying correlations as significant purely by chance.

After the correlation analysis, an ordinary least squares (OLS) linear regression model is fitted to examine whether product-level network features are associated with product rating while controlling for the other included predictors. In this model, product rating is used as the dependent variable,

while μ_{degree}, $\mu_{betweenness}$, and $n_{communities}$ are used as independent variables. Moreover, the regression model includes an intercept term.
In addition to the regression on centralities, a second regression analysis is conducted using community composition features at the product level. The proportion of ingredients belonging to each detected community is calculated for every product and used as a set of predictors for product rating. Since these proportions sum up to 1, one community is omitted from the model and serves as the reference category, meaning that it provides the baseline against which the other community coefficients are compared. This makes it possible to examine whether products with a higher share of ingredients from specific communities differ in ratings compared to the omitted community.
The regression results are interpreted using the estimated coefficients, confidence intervals, and p-values. Together, the correlation matrix and regression analysis provide an assessment of the relationship between product popularity and network features, including structural position and community composition.

4.6 Robustness Checks

To evaluate the influence of highly prevalent base ingredients on the results, the network analysis is repeated after excluding water and glycerin from the ingredient list used for network construction. Since these ingredients appear in a large number of skincare products, they may dominate co-occurrence patterns and shape the overall structure of the network disproportionately. Consequently, more meaningful relationships between less common ingredients may become harder to detect.
Thus, by using this reduced ingredient set, an alternative co-occurrence network is constructed according to the same procedure as in the main analysis. Degree centrality, betweenness centrality, and community membership are consequently recomputed and aggregated to the product level. Additionally, community detection is repeated on the reduced network to evaluate whether the detected community structure remains similar after removing water and glycerin. This robustness check makes it possible to assess whether the associations between network features and product popularity remain stable when highly prevalent base ingredients are removed.

4.7 Machine Learning Approach

To complement the network analysis, this thesis applies a tabular machine learning approach to examine whether product popularity can be predicted from standard ingredient product data. The analysis considers two outcome variables: product rating and the logarithm of review count. As shown in Figure 1, the review counts are strongly right-skewed, while the log-transformed review counts have a less skewed distribution. Therefore, the log transformation is used to reduce skewness and to obtain a more stable prediction target.
Three feature representations are constructed for the machine learning analysis. Firstly, an ingredient-based representation is created in which each product is represented by a binary vector indicating the presence or absence of individual ingredients. This yields a high-dimensional tabular dataset in which each column corresponds to a unique ingredient. To capture relationships between ingredients rather than only individual ingredient presence, a second representation is constructed using the product-level network features derived from the co-occurrence analysis, namely μ_{degree}, $\mu_{betweenness}$, and the community proportion variables ($p_{community}$). Thirdly, a combined feature representation is

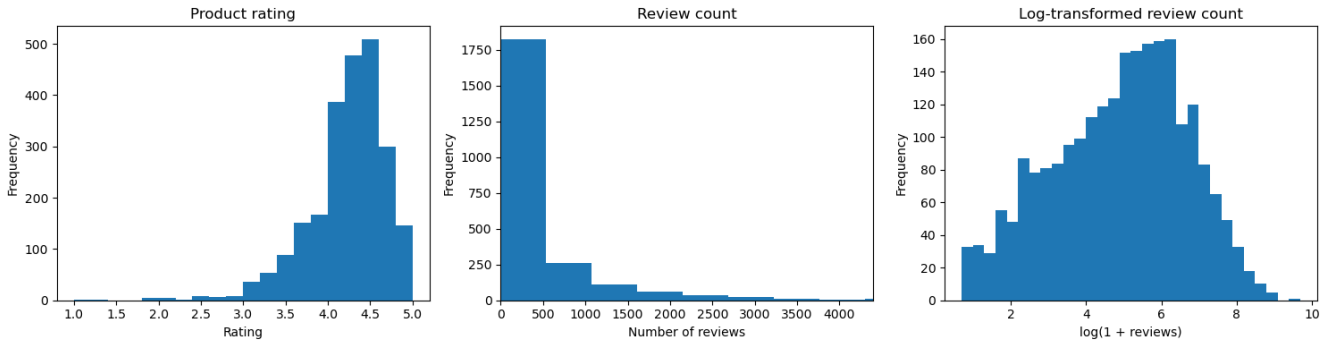

Figure 1: Distributions of product rating, review count, and log-transformed review count. Review counts are strongly right-skewed, while the log transformation reduces this skewness and produces a more balanced distribution. The review count is limited to the 99th percentile for readability.

created by concatenating the ingredient feature matrix with the network feature matrix, so that each product is represented by both its ingredient features and its network features.

For each target variable, model performance is evaluated using 5-fold cross-validation with a fixed random seed to ensure reproducibility of the data splits. Two regression models are trained on each feature representation: random forest regression and XGBoost regression. These tree-based ensemble models are chosen because they can handle high-dimensional tabular feature representations and capture non-linear relationships between ingredient features, network features, and product popularity. Random forest builds many decision trees independently and averages their predictions, while XGBoost builds an ensemble of decision trees sequentially, with each new tree correcting errors made by the previous trees. Therefore, using both models allows the analysis to compare two common tree model approaches with different ensemble strategies. Both models are trained directly on the feature values without feature scaling, since tree models do not require standardized inputs.

Model performance is evaluated using the mean cross-validated coefficient of determination (R^2). The R^2 score measures the proportion of variance in the target variable that is explained by the model, with a value of 1 indicating perfect prediction and a value of 0 corresponding to a model that predicts the mean of the target variable. Values below 0 indicate performance worse than the mean [33]. To assess whether the combined feature representation improves predictive performance compared to the ingredient-only representation, paired t-tests are conducted on the fold-level R^2 scores for each target variable and model. Since four paired comparisons are performed, a Bonferroni correction is applied to the p-values. In this way, the machine learning analysis provides an additional perspective on whether ingredient composition and network structure contain useful information for predicting product popularity.

5 Results

This section presents the results of the analyses described in Section 4. The overall structure of the ingredient co-occurrence network is described, followed by the detected ingredient communities. Furthermore, the section presents the relationships between product-level network features and product popularity. Finally, the robustness checks and machine learning prediction results are reported.

5.1 Description of Network Structure

This subsection describes the overall structure of the ingredient co-occurrence network constructed from the SkinSort dataset. The purpose of this analysis is to provide an initial overview of the size, density, connectivity, and strongest co-occurrence relationships in the network before examining community structure and product-level associations.
As shown in Table 1, the filtered network contains 6662 ingredient nodes and 670263 edges. Simultaneously, the density is low (0.0302), which indicates that only a small proportion of all possible ingredient pairs co-occur in the skincare products. The network consists of a single connected component, which implies that all retained ingredients are part of one interconnected formulation system. Additionally, five communities are detected, indicating that ingredient combinations are organized into several distinct clusters.

Table 1: Summary statistics of the ingredient co-occurrence network

Measure	**Value**
Number of products used for network construction	19050
Number of ingredient nodes	6662
Number of edges	670263
Minimum edge weight retained	2
Number of connected components	1
Average degree	201.22
Network density	0.0302
Number of communities	5

Figure 2 visualizes the top 40 strongest co-occurrence relationships in the network. The figure shows that ingredients such as water and glycerin occupy highly central positions and are connected to many other commonly used ingredients. Therefore, the figure illustrates that the strongest co-occurrences are concentrated around a set of highly prevalent ingredients. Since these strong edges correspond to ingredient pairs that appear together frequently across products, they provide evidence of recurring formulation patterns within the dataset.

5.2 Ingredient Communities

This subsection examines the communities detected in the ingredient co-occurrence network using the Leiden algorithm. The goal is to identify groups of ingredients that frequently co-occur and

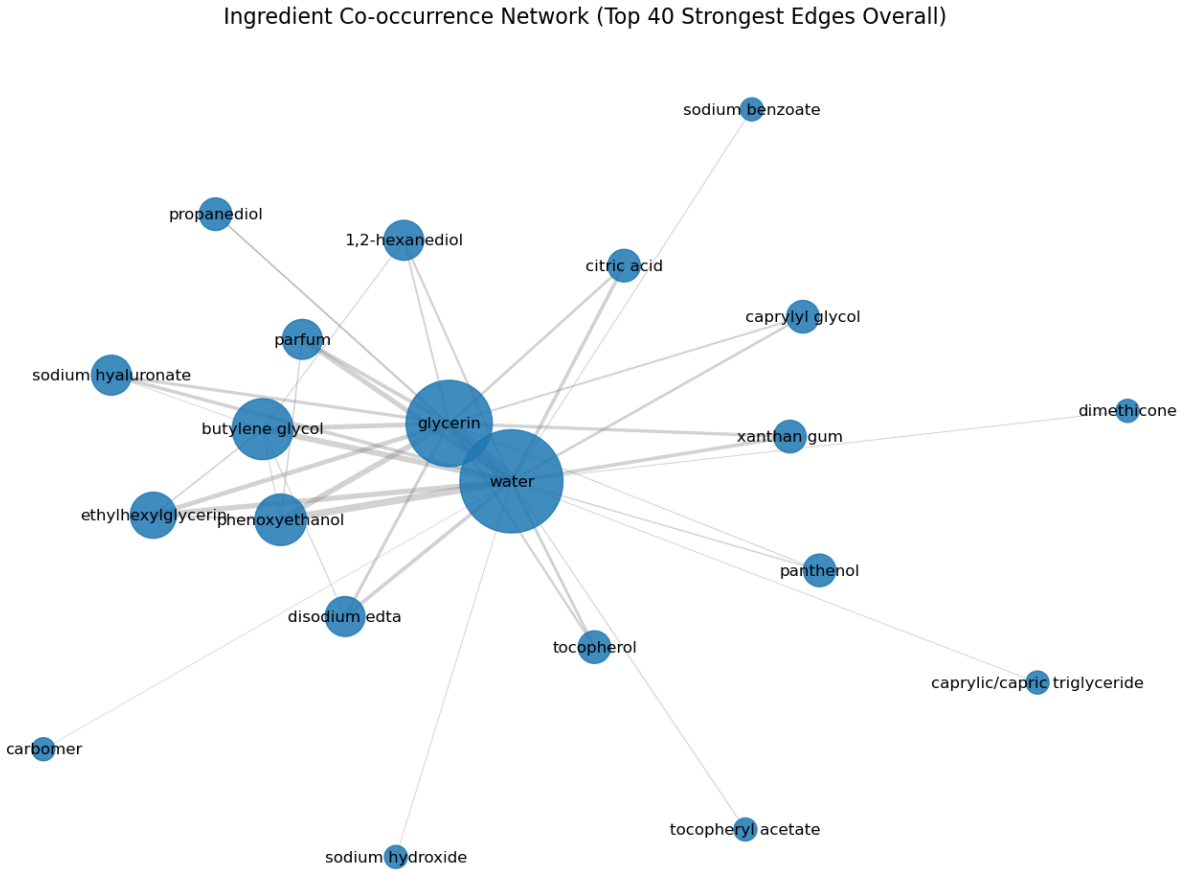

Figure 2: Top 40 strongest ingredient co-occurrences in the skincare formulation network. Node size reflects the number of connections of an ingredient within this top 40 subnetwork. Edges indicate the strongest co-occurrence relationships observed across products.

to assess whether these groups correspond to recognizable formulation contexts within skincare products.

The community detection analysis reveals that the ingredient co-occurrence network is organized into several distinct clusters of frequently co-occurring ingredients. Table 2 shows an overview of the detected communities, including their sizes, representative ingredients, approximate interpretations, and label fit percentages. These interpretations are assigned qualitatively by inspecting the most representative and frequently occurring ingredients within each community and identifying common formulation functions or product contexts. The fit percentages provide a descriptive check of these labels based on the CosIng function categories described in Section 4.4.

The community labels partly align with common functional categories described in skincare and cosmetic formulation literature, such as cleansing agents, humectants, emolients, and UV filters [2]. However, these labels should be understood as descriptive summaries rather than strict ingredient classifications. Since the Leiden algorithm groups ingredients based on how often they appear together in products, ingredients with different cosmetic functions may be assigned to the same community when they frequently occur in the same formulation context. For instance, the community labelled as formulation-support ingredients contains several supporting ingredient types, including preservatives, pH adjusters, fragrances, and base ingredients. This suggests that this community reflects a shared formulation context related to formulation support rather than a single function. The largest communities correspond broadly to formulation-support ingredients, hydrating and soothing ingredients, emollient and moisturizing lipid ingredients, and texture-enhancing ingredients. Additionally, community 4 represents a smaller community of more specialized treatment and botanical active ingredients.

Figure 3 shows the top 300 strongest ingredient associations coloured by community. The visualization suggests that the network contains identifiable communities, although these are not sharply separated from each other. Particularly, several highly prevalent base ingredients, such as water and glycerin, link multiple parts of the network, resulting in an interconnected community structure rather than fully isolated clusters.

Table 3 reports additional connectivity measures for the detected communities. The internal edge density varies across communities, with the smallest community showing a higher density than the larger communities. The internal edge share and internal weight share indicate that only a portion of each community's connections and co-occurrence strength remain within the same community. Simultaneously, the average clustering coefficients are relatively high across all communities, ranging from 0.765 to 0.863. These results show that the detected communities contain locally connected groups of ingredients, while also maintaining connections to ingredients outside their own community.

Overall, the detected communities provide evidence of recurring formulation patterns. They are also used in the later analysis, where each product is described by the proportion of its ingredients that belong to each community, and these proportions are related to product ratings.

5.3 Network Features and Product Popularity

This subsection examines whether product-level network features are associated with product popularity. To assess this relationship, the average degree centrality, average betweenness centrality, and number of ingredient communities represented in each product are compared with product rating and review count using correlation analysis and linear regression.

Table 2: Overview of the detected ingredient communities with top 20 representative ingredients ranked by weighted degree, approximate formulation interpretations, and label fit percentages.

Community	Size	Representative ingredients	Approximate interpretation	Fit (%)
0	2357	water, glycerin, phenoxyethanol, parfum, citric acid, sodium benzoate, potassium sorbate, sodium hydroxide, limonene, linalool, sodium chloride, propylene glycol, polysorbate 20, aloe barbadensis leaf juice, benzyl alcohol, sodium citrate, alcohol, lactic acid, citronellol, chlorphenesin	Formulation-support ingredients	90
1	1598	butylene glycol, ethylhexylglycerin, sodium hyaluronate, disodium edta, 1,2-hexanediol, propanediol, panthenol, carbomer, niacinamide, pentylene glycol, allantoin, camellia sinensis leaf extract, dipropylene glycol, hydrogenated lecithin, adenosine, betaine, ceramide np, acrylates/c10-30 alkyl acrylate crosspolymer, arginine, centella asiatica extract	Hydrating and soothing ingredients	80
2	1299	tocopherol, xanthan gum, caprylic/capric triglyceride, cetearyl alcohol, glyceryl stearate, butyrospermum parkii butter, squalane, helianthus annuus seed oil, cetyl alcohol, lecithin, simmondsia chinensis seed oil, polysorbate 60, rosmarinus officinalis leaf extract, ascorbic acid, behenyl alcohol, stearyl alcohol, olea europaea fruit oil, bisabolol, cocos nucifera oil, sodium phytate	Emollient and moisturizing lipid ingredients	65
3	1237	caprylyl glycol, tocopheryl acetate, dimethicone, stearic acid, silica, ci 77891, cyclopentasiloxane, peg-100 stearate, mica, ci 77492, ci 77491, palmitic acid, bha, titanium dioxide, sorbitan isostearate, c12-15 alkyl benzoate, ci 77499, octyldodecanol, ci 19140, ethylhexyl palmitate	Texture-enhancing ingredients	60
4	171	caffeine, sucrose, eucalyptus globulus leaf oil, algae extract, sodium gluconate, peg-8, zinc gluconate, laminaria digitata extract, acetyl glucosamine, yeast extract, steareth-20, calcium gluconate, magnesium ascorbyl phosphate, hydroxypropyl cyclodextrin, copper gluconate, coffea arabica seed extract, hordeum vulgare extract, plankton extract, palmitoyl hexapeptide-12, laminaria saccharina extract	Specialty treatment and active ingredients	70

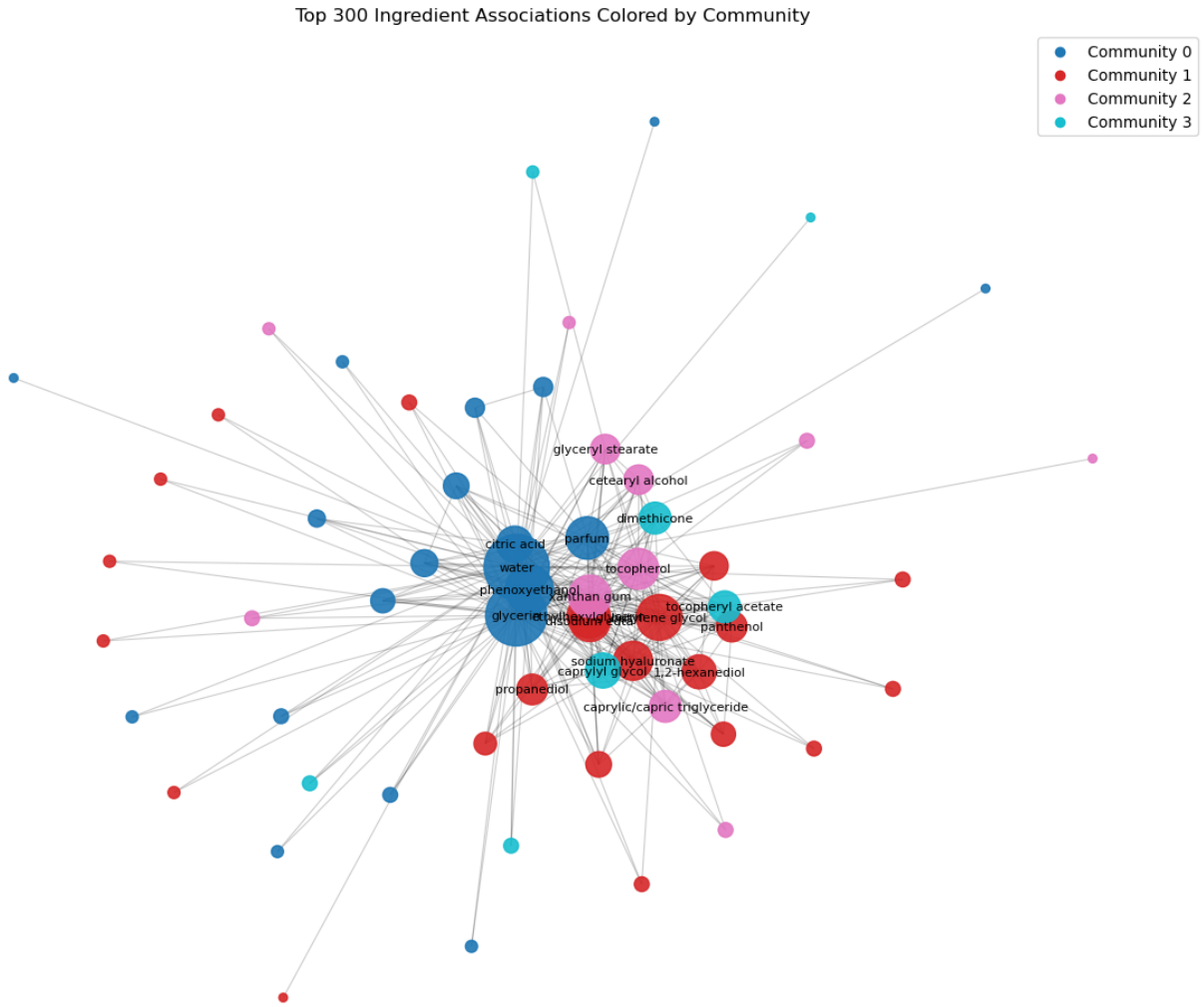

Figure 3: Top 300 ingredient associations in the co-occurrence network, coloured by community. Node size reflects the number of connections of an ingredient, while edges indicate the strongest co-occurrence relationships observed across products. Not all detected communities are represented, because the figure is based on a subnetwork of only the strongest edges.

Table 3: Community connectivity measures for the detected ingredient communities. Internal edge density refers to the proportion of possible within-community edges that are present. Internal edge share and internal weight share indicate the percentage of a community's connections and co-occurrence strength that remain within the same community. The average clustering coefficient measures the local connectedness within each community.

Community	Size	Int. edge density	Int. edge share (%)	Int. weight share (%)	Avg. clustering coef.
0	2357	0.0218	22.6	24.1	0.863
1	1598	0.0796	30.4	31.4	0.830
2	1299	0.0483	17.6	16.2	0.810
3	1237	0.0584	21.1	22.0	0.817
4	171	0.2894	9.3	8.4	0.765

Figure 4 presents the Pearson correlations between the product popularity variables and the network features at the product level. Product rating shows very weak correlations with the three network features: average degree ($r = 0.01$), average betweenness ($r = -0.03$), and the number of communities ($r = 0.01$). The number of reviews also shows very weak correlations with average degree ($r = 0.02$), average betweenness ($r = 0.00$), and the number of communities ($r = 0.01$). Furthermore, product rating and number of reviews show a weak but statistically significant positive correlation ($r = 0.08$). In contrast, stronger correlations are observed between the network features themselves. Average degree and average betweenness are strongly positively correlated ($r = 0.78$). The number of communities is also positively correlated with average degree ($r = 0.52$) and average betweenness ($r = 0.20$).

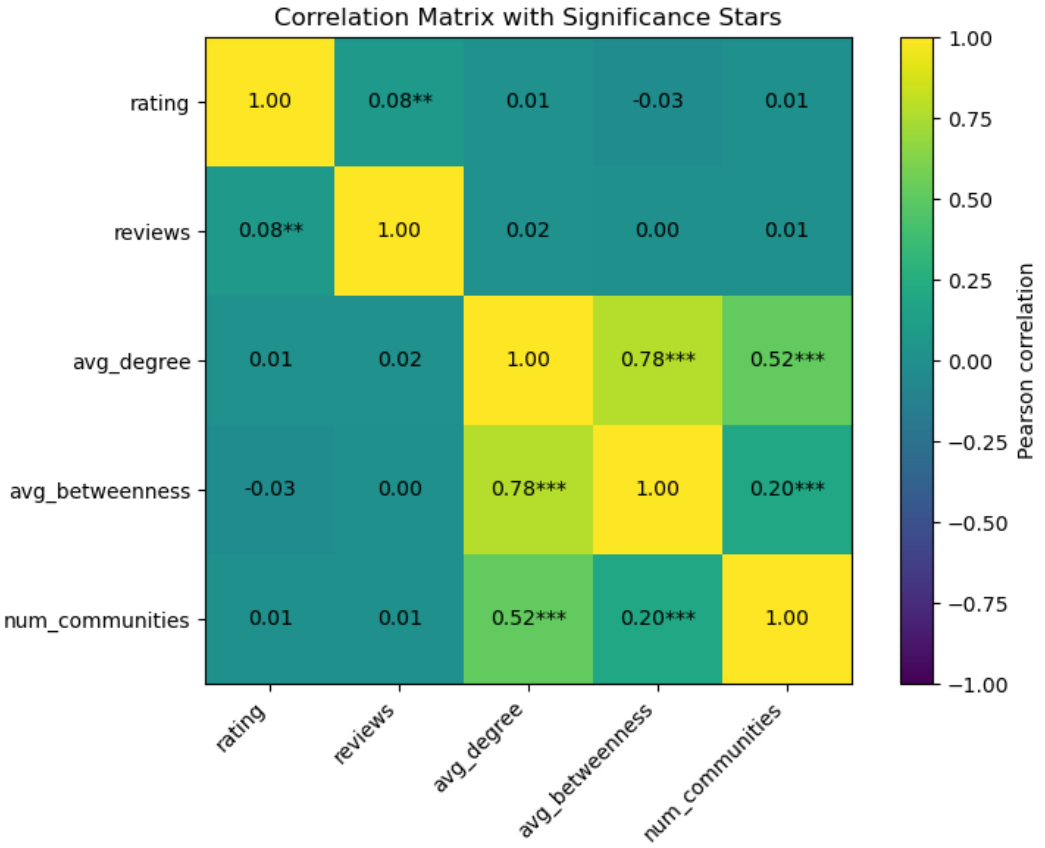

Figure 4: Pearson correlation matrix between product popularity measures and product-level network features. Each cell shows the Pearson correlation coefficient between two variables. Significance stars indicate the statistical significance of the correlation coefficient: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Figure 5 shows the estimated coefficients from the regression model predicting product rating using the product-level network features. The confidence interval for average degree and average betweenness do not cross zero, with average degree having a positive coefficient and average betweenness having a negative coefficient. In contrast, the confidence interval for the number of communities does cross zero.
Overall, the bivariate correlations between product popularity and the network features are close

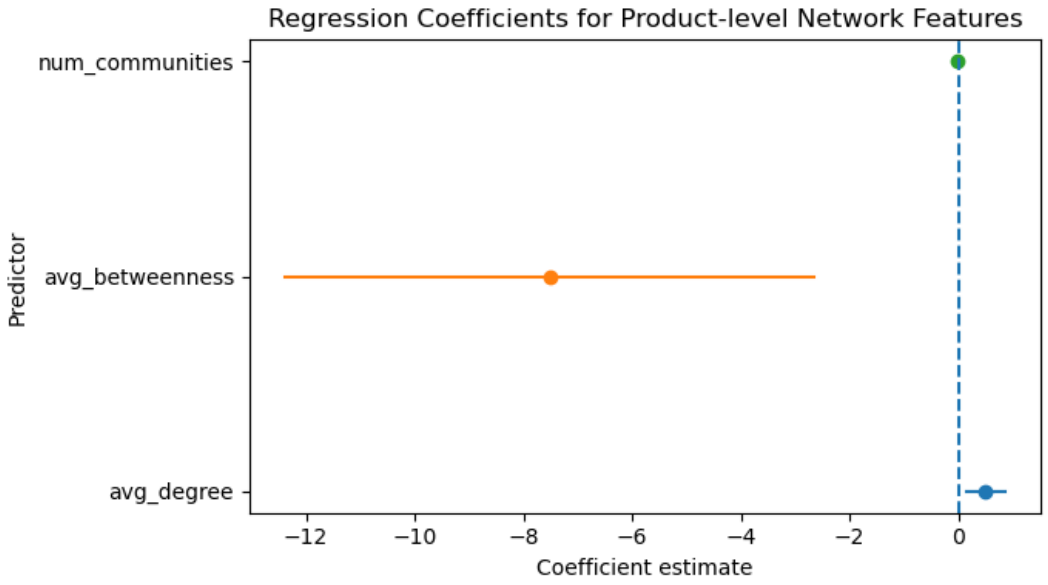

Figure 5: Regression coefficient plot for product-level network features predicting product rating. Points represent the estimated coefficients for each predictor, while horizontal lines show the corresponding 95% confidence intervals. The dashed vertical line at zero represents no estimated association: confidence intervals crossing this line indicate coefficients that are not statistically distinguishable from zero at the 5% level.

to zero. However, when the network features are included together in the regression model, both average degree and average betweenness have non-zero coefficients.

5.4 Community Composition and Product Popularity

This subsection examines whether the community composition of a product's ingredients is associated with product rating. Community composition is measured by calculating the proportion of each product's ingredients assigned to each community. These proportions are used as predictors in a regression model.

Figure 6 presents the estimated regression coefficients for the ingredient community proportion variables predicting product rating. Community 0 is used as the reference category and is therefore omitted from the figure.

The estimated coefficients vary across communities. Community 2 has the largest positive coefficient, and its confidence interval lies entirely above zero. Community 1 also has a positive coefficient, suggesting a positive relationship between the proportion of ingredients from this community and product rating. However, its confidence interval crosses the zero line, indicating that the coefficient is not statistically significant. Moreover, community 3 and community 4 both have negative coefficients, suggesting a negative relationship between the proportion of ingredients from these communities and product rating. The confidence interval for community 3 lies below zero, while the confidence interval for community 4 crosses zero.

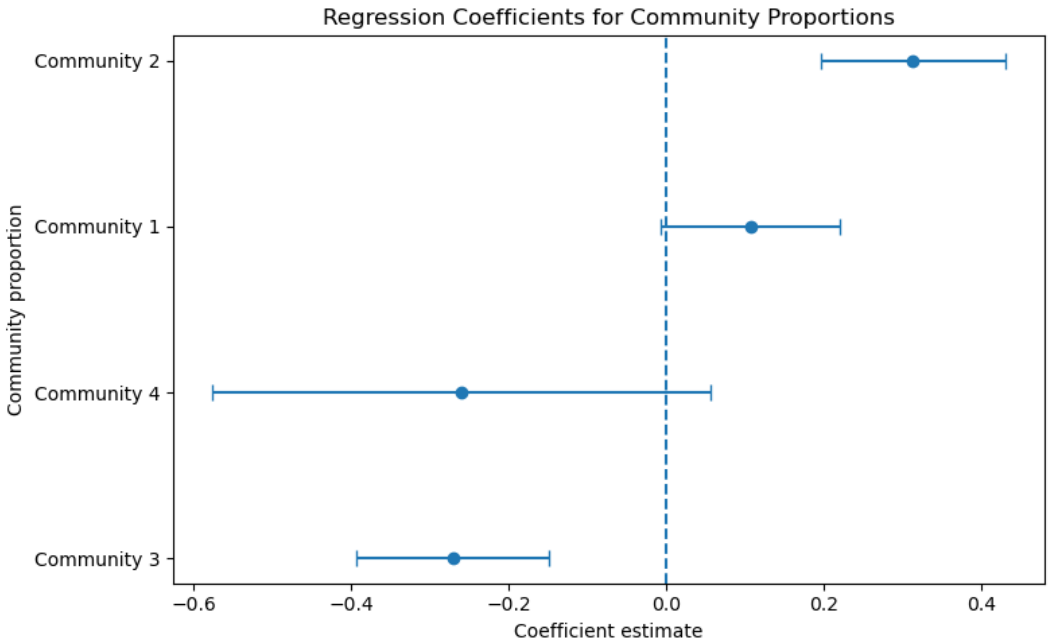

Figure 6: Regression coefficient plot for ingredient community proportions predicting product rating. Coefficients indicate the association between the proportion of a product's ingredients assigned to each community and its rating. Points show coefficient estimates, horizontal lines show 95% confidence intervals, and the dashed vertical line indicates zero effect.

5.5 Robustness Checks

This subsection evaluates whether the regression results for the network features are robust to the removal of highly prevalent base ingredients. Water and glycerin are excluded because they occur most frequently across skincare formulations and may therefore have a disproportionate influence on the structure of the co-occurrence network. To assess this, the co-occurrence network is reconstructed after removing water and glycerin, the network features are recalculated, and the regression model is fitted again.

Figure 7 presents the regression coefficients for the product-level network features after excluding water and glycerin from the network construction. The coefficient for average degree remains positive and statistically significant, indicating that the positive association between average degree and product rating is robust to the removal of these highly prevalent ingredients. The coefficient for average betweenness remains negative, but is no longer statistically significant, suggesting that the betweenness result is more sensitive to the presence of water and glycerin in the network. Moreover, the coefficient for the number of communities remains close to zero and is also not statistically significant, consistent with the main regression results.

The community detection analysis is also repeated after excluding water and glycerin from the network construction. The reduced network produces eight communities instead of five. Furthermore, the main community interpretations remain broadly similar, including formulation-support, hydrating and soothing, emollient, texture-enhancing, and active ingredients. However, the reduced network also separates additional smaller communities, including a community of hydrating amino acid ingredients. This indicates that the broad community structure is partly stable after removing

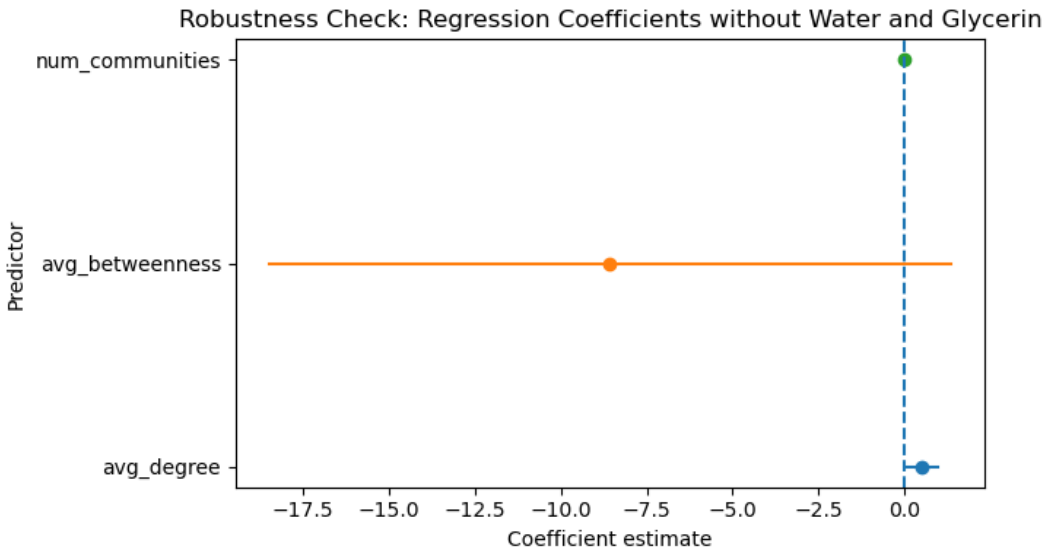

Figure 7: Robustness check regression coefficient plot for product-level network features predicting product rating after excluding water and glycerin from the network construction. Points show coefficient estimates, horizontal lines show 95% confidence intervals, and the dashed vertical line indicates zero effect.

water and glycerin, while more specific ingredient groupings are affected by the removal of these prevalent base ingredients.

5.6 Machine Learning Prediction

This subsection evaluates whether product popularity can be predicted from ingredient-based features, network features, or a combination of both. Random forest and XGBoost regression models are compared using 5-fold cross-validation for two target variables: product rating and log-transformed review count.

Figure 8 presents the mean cross-validated R^2 scores for the random forest and XGBoost models across the three feature sets. For predicting the log-transformed review count, the highest mean R^2 is obtained by the XGBoost model using the combined feature set. For predicting product rating, the random forest model with the combined feature set shows the highest mean R^2. However, the network/community-only feature set performs worse than the ingredient-only and combined feature sets for both target variables. In several cases, the network/community-only feature set produces negative mean R^2 scores, which indicates that these models perform worse than a baseline model that always predicts the average value of the target variable.

Table 4 compares the combined feature set with the ingredient-only feature set using paired t-tests on the cross-validated R^2 scores. After applying the Bonferroni correction, most differences are not statistically significant at the 0.05 level. The only significant improvement is observed for XGBoost when predicting the log-transformed review count ($p = 0.0086$). However, the mean R^2 difference is small (0.0215), indicating that the combined feature set provides only limited additional predictive value compared to the ingredient-only feature set.

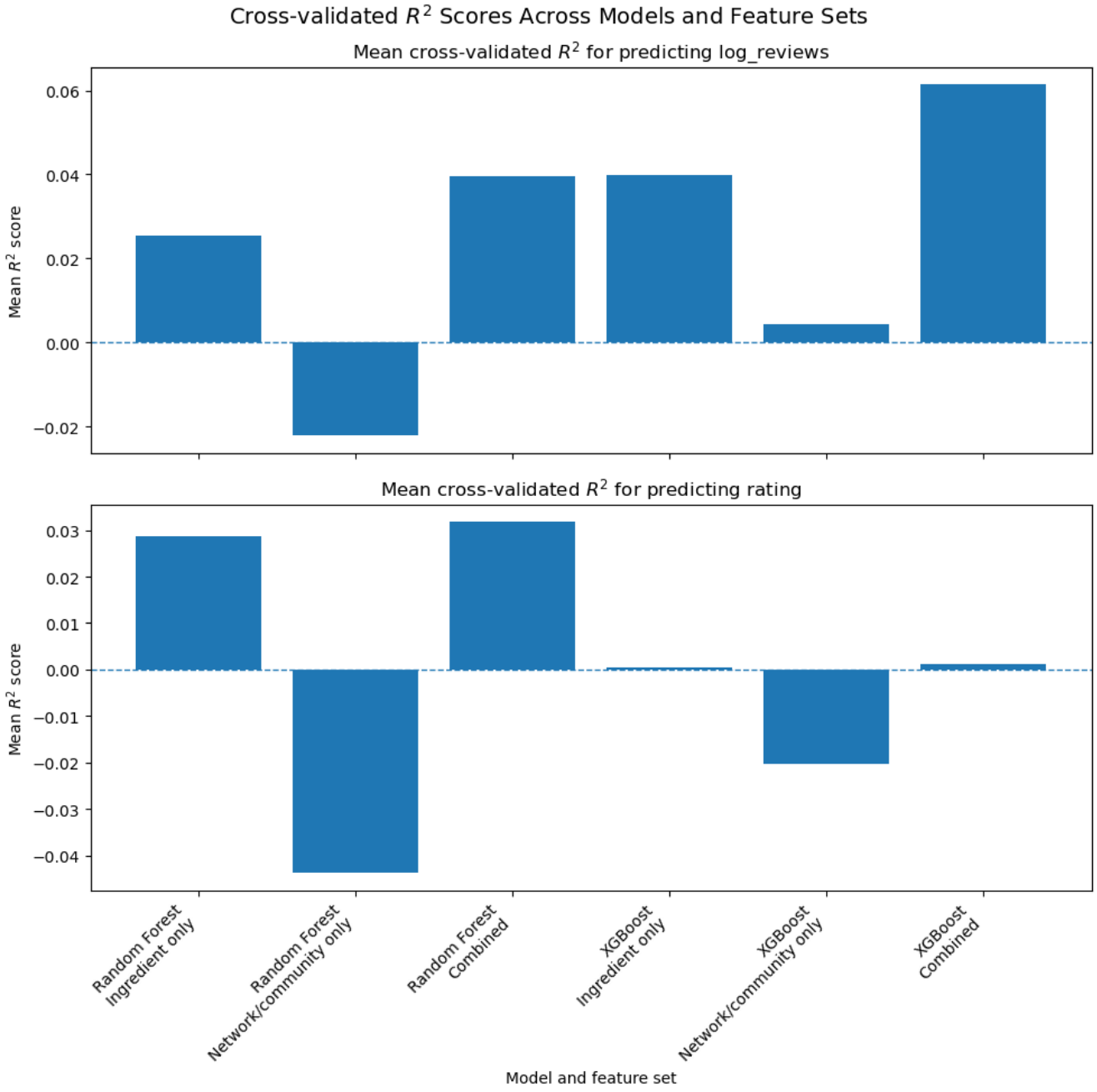

Figure 8: Mean cross-validated R^2 scores across random forest and XGBoost models using different feature sets. The upper panel shows prediction results for log-transformed review count, and the lower panel shows prediction results for product rating. Feature sets include ingredient-only features, network/community features, and a combined feature set. The dashed horizontal line indicates an R^2 score of zero.

Table 4: Paired comparison of cross-validated R^2 scores between the combined and ingredient-only feature sets. Bonferroni corrected p-values are reported for the four paired comparisons. Positive differences indicate a higher mean R^2 for the combined feature set.

Target	**Model**	**Mean R^2 difference**	**p-value**
Rating	Random forest	0.0032	1.0000
Rating	XGBoost	0.0007	1.0000
Log reviews	Random forest	0.0141	0.1452
Log reviews	XGBoost	0.0215	0.0086

6 Discussion

Overall, the results suggest that skincare formulations contain meaningful network structure. However, the relationship between network features and popularity is modest and depends on how the features are modeled. The following subsections discuss the interpretation of the ingredient network, the relationship between network features and product popularity, the role of community composition, the machine learning results, and the main limitations of the study.

6.1 Ingredient Network Structure and Communities

The ingredient co-occurrence network (Figure 2) shows that skincare formulations are organized around recurring combinations of ingredients rather than isolated ingredient choices. The filtered network forms one connected component, which implies that even ingredients with different functions are indirectly connected through products. Simultaneously, the low network density indicates that not all ingredients are combined with all others. This suggests that although skincare products can be formulated in many different ways, ingredients are combined according to recurring patterns rather than arbitrarily.

The detected communities further support this interpretation. The communities broadly correspond to recognizable formulation contexts, namely formulation-support, hydration and soothing, emollient and lipid-based moisturization, texture enhancement, and a smaller group of more specialized active ingredients. The label fit percentages in table 2 show that these interpretations match many of the representative ingredients, but not all of them. This suggests that the community detection does not merely produce arbitrary clusters, but captures meaningful groups of ingredients that often appear together in formulations. Nevertheless, the label percentages confirm that the labels are approximate summaries rather than strict functional classifications. In this sense, the network approach provides a useful way to summarize formulation structure at a broader level than individual ingredient lists. However, these communities should not be interpreted as perfectly separated functional categories. The visualization of the strongest ingredient associations (Figure 3) shows that the communities are not fully separated, but remain connected through highly prevalent ingredients such as water and glycerin. This is expected in skincare formulations, because many ingredients serve as base components or formulation aids across product types. Therefore, the network does not consist of isolated groups, but rather of partially overlapping ingredient groups. This means that the detected communities can be used to identify recurring formulation contexts, but they should not be interpreted as strictly separated ingredient categories.

The community connectivity measures in table 3 further show that the detected communities have clear internal structure, but are not fully separated from each other. The high average clustering coefficients indicate that ingredients within the same community form locally connected groups, which supports the idea that the communities capture recurring formulation patterns. However, the internal edge weight and shares show that many co-occurrence relationships also connect ingredients across different communities. Thus, the communities are locally meaningful, but still part of one broader interconnected network.

If ingredient communities are partly shaped by common base ingredients, then community membership reflects both recurring product contexts and the widespread use of certain ingredients across formulations. For example, a community may contain ingredients that are often used in similar types of products, but it may also be connected to other communities through ingredients

that appear in many different types of products. Therefore, the detected communities should be understood as empirical co-occurrence groups rather than strict skincare categories.
The robustness check further shows that the exact community structure depends partly on the network construction. After removing water and glycerin, the broad community interpretations remain similar, but the reduced network produces additional smaller communities. This indicates that the communities capture recurring co-occurrence patterns, while their exact composition can change when highly prevalent ingredients are removed.
Overall, the network structure indicates that skincare formulations contain recurring ingredient combination patterns. The detected communities provide a useful summary of these patterns, but their interconnections and partial sensitivity to network construction mean that they should be interpreted as empirical co-occurrence groups rather than fixed skincare categories.

6.2 Interpretation of Network Feature Associations

The relationship between product-level network features and product popularity is not straightforward. The correlation analysis shows that average degree, average betweenness, and the number of communities have only very weak bivariate correlations with product rating and review count. This suggests that no single network feature, considered in isolation, strongly explains differences in product popularity. Thus, products with more central or more diverse ingredient compositions are not necessarily rated higher or reviewed more often.
These findings should also be interpreted as evidence about linear relationships only. The correlation analysis and OLS regression can show whether ratings tend to increase or decrease as a network feature increases, but they do not capture non-linear patterns, such as cases where products with medium values of a feature receive higher ratings than products with either low or high values.
However, the regression results provide a more nuanced picture. When the product-level network features are included in the same model, average degree has a positive association with product rating, while average betweenness has a negative association. This difference between the correlation and regression results is important, because the network features are strongly correlated themselves. Average degree and average betweenness especially capture overlapping aspects of ingredient centrality. Consequently, their separate relationships with product rating may only become visible when they are modeled together.
The positive coefficient for average degree may indicate that products containing ingredients with many direct co-occurrence connections tend to be rated slightly higher. A possible explanation is that high-degree ingredients are often common, versatile, or compatible ingredients that appear across many products. Their positive association with rating may therefore reflect aspects of ingredient familiarity, perceived quality, or consumer trust. This interpretation is consistent with cosmetic consumer research showing that trust in ingredient information is associated with perceived value and purchase intention [34]. However, this does not mean that high-degree ingredients directly cause higher ratings.
The negative coefficient for average betweenness is more difficult to interpret. Betweenness centrality captures ingredients that connect different parts of the network, which means that ingredients with high betweenness may appear in formulations that bridge otherwise separate ingredient groups. In the context of skincare, this could reflect more complex or unusual formulations, but it could also reflect ingredients that appear across different product types for technical reasons. Therefore, the negative association should not be interpreted as evidence that bridging ingredients are less

desirable. Rather, it suggests that degree centrality and betweenness centrality capture different aspects of ingredient position, which are related to product rating in different ways.
The robustness check adds an important qualification to this interpretation. After removing water and glycerin, the coefficient for average degree remains positive and statistically significant, whereas the coefficient for average betweenness remains negative but is no longer statistically significant. This suggests that the association between average degree and rating does not depend entirely on the inclusion of water and glycerin in the network, whereas the association between betweenness and rating is more sensitive to the presence of highly prevalent base ingredients. Since water and glycerin connect many ingredient combinations, they can act as shortcuts between different parts of the network. Therefore, their inclusion can strongly affect betweenness centrality, which is based on shortest paths.
Overall, these results suggest that network position contains more information about product ratings, but the associations are modest and depend on the specific structural measure used. Degree centrality appears to provide the most stable signal, while betweenness centrality is more sensitive to network construction choices. Thus, the findings support the idea that ingredient structure is relevant to product popularity, but they also show that this relationship is indirect and should not be reduced to an assumption that more central ingredients always lead to more popular products.

6.3 Interpretation of Community Composition

The community composition results add another layer to the relationship between formulation structure and product popularity. As discussed in Section 2.3, ratings reflect consumer evaluations rather than direct measures of ingredient effectiveness. Therefore, the community coefficients should be interpreted as associations with perceived product response rather than as evidence that certain ingredient groups perform better or worse. This is consistent with research on online product ratings, which treats ratings as indicators of consumer perceived product quality [35].
The results suggest that community composition is not uniformly related to product rating. Compared with the reference community, the proportion of ingredients from community 2 shows a positive association with rating, while the proportions of ingredients from communities 3 and 4 show negative associations. Community 1 has a positive coefficient, but this is not statistically significant. This indicates that the relationship between ingredient communities and ratings depends on which parts of the network are represented in a product, rather than on whether a product contains ingredients from many communities.
The positive coefficient for community 2 may reflect the role of emollient and lipid-rich formulation patterns. Ingredients in this community include oils, fatty alcohols, esters, and moisturizing ingredients, which are common in products designed to improve skin feel, softness, and barrier support [36, 37]. These properties may be more directly reflected in consumer ratings, since texture and moisturization are aspects of a product that users can often evaluate quickly. This interpretation is consistent with cosmetic sensory research, which emphasizes that sensory attributes such as texture play an important role in consumer satisfaction and product evaluation [38]. However, the link between this community and consumer ratings should be treated cautiously, because the community represents ingredients that frequently co-occur in the data rather than a strictly defined functional category.
The negative coefficients for communities 3 and 4 are more difficult to interpret. Community 3 broadly relates to sun protection and silicone-based ingredients, while community 4 contains more

specialized treatment and active ingredients. Lower ratings for products with higher proportions of these communities do not necessarily imply that these ingredient groups are less effective. Rather, they may reflect differences in consumer expectations, product category, or user experience. For example, sunscreen products are often evaluated not only on effectiveness, but also on properties such as greasiness, white cast, or compatibility with makeup. Similarly, products with specialty treatment or active ingredients may be used by consumers with more specific skin concerns, which could lead to stricter evaluations.

Thus, this means that the community proportion results should be interpreted as associations between formulation patterns and ratings, rather than as evidence that certain ingredient communities directly cause higher or lower popularity. The communities capture recurring combinations in the ingredient network, but ratings are shaped by many additional factors, such as marketing, brand reputation, price, skin type, and consumer expectations. Therefore, community composition appears to provide useful information about formulation patterns at the product level, but the analysis does not investigate the influence of other factors that may affect product ratings.

Overall, the community composition analysis suggests that a product's position within the ingredient community structure is relevant to product ratings. However, these results should not be interpreted as a ranking of ingredient communities from better to worse. They show that different formulation contexts are associated with ratings in different ways.

6.4 Predictive Value of Network Features

The machine learning results provide a different perspective on the role of network structures. The regression analyses examine specific linear associations between network features and product ratings, while the machine learning models evaluate whether ingredient and network information can be used to predict product popularity. Unlike the OLS models, the tree-based machine learning models can capture more flexible and non-linear relationships.

The weaker performance of the network/community-only feature set indicates that the aggregated network features cannot fully represent the information in the original ingredient lists. This may be explained by the fact that the network features summarize each product into a relatively small number of variables. Consequently, they capture broad structural patterns, but they also remove many of the details specific to ingredients. For example, two products may have similar average degree or similar community proportions while still containing very different ingredients. Therefore, the network features are useful as a compact structural representation, but they are less detailed than the ingredient-only representation.

The ingredient-only feature set generally performs better than the network/community-only feature set, which suggests that the presence or absence of individual ingredients remains highly informative for prediction. This does not mean that network structure is irrelevant, but it suggests that much of the predictive information available in the co-occurrence network is contained in the ingredient composition itself. Thus, network features may provide additional context about how ingredients are positioned in the formulation system, but they do not replace the direct information contained in the ingredient list.

The comparison between the ingredient-only and combined feature sets further supports this interpretation. The combined feature set shows the largest improvement over the ingredient-only feature set for predicting log-transformed review count with XGBoost. This improvement is statistically significant under the Bonferroni corrected significance threshold. Therefore, the

machine learning results provide only partial evidence that adding network and community features improves prediction beyond ingredient-only features. More specifically, this pattern may suggest that network structure is more informative for product visibility or engagement than for average ratings. However, this interpretation should not be overstated, since review count is influenced by many factors outside the formulation itself.
The limited predictive performance for product rating suggests that average ratings are difficult to explain from ingredient and network features alone. Ratings are likely shaped by subjective and external factors. These factors are not fully captured by either ingredient lists or network features. Since product ratings are shaped by multiple product-related and contextual factors [35], even a detailed representation of formulation structure may only explain a small part of the variation in ratings.
Overall, the machine learning results suggest that network features provide a useful representation of skincare formulations, but their predictive value is limited when used independently. The strongest predictive information comes from ingredient features, while adding network features provides only a small additional benefit. This supports the broader conclusion that network structure can help capture formulation patterns, but product popularity is influenced by factors other than ingredient structure alone.

6.5 Limitations

Several limitations should be considered when interpreting the findings of this thesis. Firstly, product popularity is measured using ratings and review counts, which are only proxies for consumer response. Ratings may reflect perceived product quality, and review counts capture consumer engagement, but neither of them directly represent product effectiveness or number of sales. Therefore, the popularity measures used in this study should be interpreted as observable indicators of consumer response rather than direct measures of product success.
Secondly, product ratings and review counts can be influenced by factors outside the ingredient network, such as marketing, consumer expectations, skin type, product category, price, and brand reputation. Although price is included in the Sephora dataset, it is not included as a control variable because the focus of this thesis is on ingredient composition and network structure rather than commercial product characteristics. Since these external factors are not fully included in the statistical models, the observed associations between network features and popularity may partly capture differences between products that are unrelated to ingredient composition.
Additionally, the analysis is based on ingredient lists, which provide information about whether an ingredient is present in a product but not about its exact concentration. This is important because the same ingredient may have different effects depending on ingredient concentration or the strength of chemical interactions between ingredients [2]. Although ingredient lists contain ordering information, this study treats ingredients as present or absent and does not model concentration. Consequently, the analysis captures structural patterns in ingredient occurrence, but not the full chemical properties of the formulations.
Furthermore, the co-occurrence network makes it possible to study recurring formulation patterns, but it is limited by representing ingredients as connected whenever they appear in the same product. Co-occurrence should therefore not be interpreted as evidence that ingredients chemically interact or work better together. Two ingredients may co-occur because they serve similar roles, because they are common base ingredients, or because they are frequently used in the same product category.

Therefore, the network captures observed ingredient combinations rather than the underlying mechanisms that explain why those combinations occur.

Moreover, the network structure is influenced by highly prevalent ingredients. Ingredients such as water and glycerin appear in many products and can connect ingredient groups that would otherwise be less directly connected. The robustness check addresses this issue by removing water and glycerin. However, other common ingredients may still affect centrality measures and community detection. This means that the detected communities and centrality values should be interpreted as dependent on the construction of the network and the preprocessing choices used in the analysis.

Finally, the results are observational and not causal. The regression and machine learning analyses can identify associations and predictive patterns, but they cannot determine whether specific network positions or ingredient communities directly cause higher or lower product popularity. Therefore, the findings suggest a relationship between ingredient network structure and product popularity, but they do not demonstrate a direct causal effect.

7 Conclusion

This thesis investigated whether the structural position of skincare ingredients in formulation co-occurrence networks is associated with product popularity. By representing ingredients as nodes and their co-occurrences in products as edges, the study examined skincare formulations as a relational system rather than as isolated ingredient lists. The results show that skincare formulations contain clear recurring patterns of combination. Furthermore, the ingredient co-occurrence network forms an interconnected structure, and the detected communities broadly reflect recurring formulation contexts. This indicates that network analysis can provide a useful descriptive perspective on how skincare ingredients are commonly combined across products.

The association between network structure and product popularity is more modest. The bivariate correlations between product popularity and the product-level network features are close to zero, suggesting that individual network features do not strongly explain ratings or review count on their own. However, the regression results show that average degree and average betweenness have statistically significant associations with product rating when included together. Additionally, the robustness check further indicates that the positive association of average degree remains stable after excluding water and glycerin, while the association for average betweenness is more sensitive to this change.

The community composition analysis also suggests that different ingredient communities are associated with product ratings in different ways. However, these communities should not be interpreted as strict skincare categories or as a ranking from better to worse. Rather, they represent co-occurrence groups that capture recurring formulation contexts within the data. The community coefficients should therefore be interpreted as the estimated association between a higher proportion of ingredients from a given community and product rating, compared with the omitted reference community.

Finally, the machine learning analysis shows that ingredient features generally have stronger predictive value than network/community features alone. Adding network and community features to ingredient-based features produced the largest improvement for XGBoost when predicting log-transformed review count. This improvement is statistically significant after Bonferroni correction, but the size of the improvement is small. Therefore, the added predictive value of network features beyond ingredient features is limited.

In conclusion, the findings indicate that ingredient network structure is relevant to understanding skincare formulations, but its relationship with product popularity is limited. Network analysis is therefore most useful as a complementary approach. It helps reveal how ingredients are structurally organized and how formulation patterns relate to popularity, but it does not fully explain consumer ratings or engagement on its own.

7.1 Future Research Direction

While this thesis provides an initial analysis of skincare formulations as ingredient co-occurrence networks, several extensions could further improve the understanding of how formulation structure relates to product popularity.

Firstly, future research could use richer formulation data. The current analysis is based on ingredient presence and co-occurrence, but it does not include ingredient concentration, exact ingredient order, product claims, branding, price, or sensory properties. Including these variables would make it

possible to examine whether the observed associations between network structure and product popularity remain when more product information is taken into account. Particularly, ingredient concentration and formulation order could provide a more detailed representation of how ingredients function within products.

Furthermore, future research could explore alternative ways of constructing and analyzing skincare ingredient networks. The current study constructs one overall co-occurrence network using a minimum edge weight threshold. Future work could compare different threshold values, weighting strategies, or community detection algorithms to assess how sensitive the detected structures are to methodological choices. Moreover, it could be useful to construct separate networks for different product categories, such as moisturizers, cleansers, serums, and sunscreens. This would create the possibility to examine whether ingredient communities and central ingredients differ across product categories.

Additionally, future research could use more direct or detailed measures of product popularity. Ratings and review counts are useful indicators of consumer response, but they do not directly measure sales or product effectiveness. Thus, future work could use sales data or repeated product evaluations to better capture how consumer response develops over time.

Finally, future research could further validate the detected ingredient communities. In this thesis, community labels are assigned through qualitative inspection and a comparison with CosIng ingredient functions. Future studies could improve this step by using input from cosmetic formulation experts to label the representative ingredients. This would make it possible to evaluate whether the community interpretations are reproducible across different classification methods.

References

[1] L. Fleming, "Recombinant uncertainty in technological search," *Management Science*, vol. 47, no. 1, pp. 117–132, 2001.

[2] S. Sharma, U. Ahmad, J. Akhtar, A. Islam, M. M. Khan, and N. Rizvi, "The art and science of cosmetics: Understanding the ingredients," in *Cosmetic Products and Industry - New Advances and Applications*. IntechOpen, 2023.

[3] M. E. J. Newman, *Networks: An Introduction*. Oxford: Oxford University Press, 2010.

[4] Y.-Y. Ahn, S. E. Ahnert, J. P. Bagrow, and A.-L. Barabási, "Flavor network and the principles of food pairing," *Scientific Reports*, vol. 1, p. 196, 2011.

[5] E. W. T. Ngai and Y. Wu, "Machine learning in marketing: A literature review, conceptual framework, and research agenda," *Journal of Business Research*, vol. 145, pp. 35–48, 2022.

[6] S. Stepien, M. Janik, M. Nurek, A. Saxena, and R. Michalski, "Fairness in opinion dynamics," 2026.

[7] A. Saxena and S. Iyengar, "Centrality measures in complex networks: A survey," *CoRR*, vol. abs/2011.07190, 2020. [Online]. Available: https://arxiv.org/abs/2011.07190

[8] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, no. 3-5, p. 75–174, Feb. 2010. [Online]. Available: http://dx.doi.org/10.1016/j.physrep.2009.11.002

[9] A. Arya, P. K. Pandey, and A. Saxena, "Node classification using deep learning in social networks," in *Deep learning for social media data analytics*. Springer, 2022, pp. 3–26.

[10] V. A. Traag, L. Waltman, and N. J. van Eck, "From louvain to leiden: guaranteeing well-connected communities," *Scientific Reports*, vol. 9, no. 1, Mar. 2019. [Online]. Available: http://dx.doi.org/10.1038/s41598-019-41695-z

[11] P. P. Analytis, A. Delfino, J. Kämmer, M. Moussaïd, and T. Joachims, "Ranking with social cues: Integrating online review scores and popularity information," 2017. [Online]. Available: https://arxiv.org/abs/1704.01213

[12] C. Bird, A. Gourley, P. Devanbu, M. Gertz, and A. Swaminathan, "Mining email social networks," in *Proceedings of the 2006 international workshop on Mining software repositories*, 2006, pp. 137–143.

[13] A. Saxena, P. Saxena, H. Reddy, and R. Gera, "A survey on studying the social networks of students," *arXiv preprint arXiv:1909.05079*, 2019.

[14] G. De Prato and D. Nepelski, "Global technological collaboration network: Network analysis of international co-inventions," *The Journal of Technology Transfer*, vol. 39, no. 3, pp. 358–375, 2014.

[15] A. Saxena and S. Iyengar, "Evolving models for meso-scale structures," in *2016 8th international conference on communication systems and networks (COMSNETS)*. IEEE, 2016, pp. 1–8.

[16] D. Ediger, K. Jiang, J. Riedy, D. A. Bader, C. Corley, R. Farber, and W. N. Reynolds, "Massive social network analysis: Mining twitter for social good," in *2010 39th international conference on parallel processing*. IEEE, 2010, pp. 583–593.

[17] I. Bermudez, D. Cleven, R. Gera, E. T. Kiser, T. Newlin, and A. Saxena, "Twitter response to munich july 2016 attack: Network analysis of influence," *Frontiers in big Data*, vol. 2, p. 17, 2019.

[18] C. Brinkley, "The small world of the alternative food network," *Sustainability*, vol. 10, no. 8, p. 2921, 2018.

[19] A. Saxena, Y. Pei, J. Veldsink, W. van Ipenburg, G. Fletcher, and M. Pechenizkiy, "The banking transactions dataset and its comparative analysis with scale-free networks," in *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2021, pp. 283–296.

[20] R. Gera, R. Miller, A. Saxena, M. MirandaLopez, and S. Warnke, "Three is the answer: Combining relationships to analyze multilayered terrorist networks," in *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, 2017, pp. 868–875.

[21] D. M. Schwartz and T. Rouselle, "Using social network analysis to target criminal networks," *Trends in Organized Crime*, vol. 12, no. 2, pp. 188–207, 2009.

[22] R. Miller, R. Gera, A. Saxena, and T. Chakraborty, "Discovering and leveraging communities in dark multi-layered networks for network disruption," in *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 2018, pp. 1152–1159.

[23] H. D. Boekhout, A. A. Blokland, and F. W. Takes, "Early warning signals for predicting cryptomarket vendor success using dark net forum networks," *Scientific Reports*, vol. 14, no. 1, p. 16336, 2024.

[24] A. Saxena, A. Singh, H. Saxena, and R. Helles, "Tracking the trackers: Structural evolution of web-tracking on educational websites," in *Proceedings of the 18th ACM Web Science Conference 2026*, 2026, pp. 392–397.

[25] D. Frank, "Unveiling skincare ingredient networks," *Acta Pharmaceutica Hungarica*, vol. 95, no. 1, pp. 37–40, 2025.

[26] J. Lee, H. Yoon, S. Kim, C. Lee, J. Lee, and S. Yoo, "Deep learning-based skin care product recommendation: A focus on cosmetic ingredient analysis and facial skin conditions," *Journal of Cosmetic Dermatology*, vol. 23, no. 6, pp. 2066–2077, 2024.

[27] K. Hasan. (2024) 19,000 skincare products database of skinsort. Kaggle. [Online]. Available: https://www.kaggle.com/datasets/kazireyazulhasan/19000-skincare-products-database-of-skinsort

[28] N. Inky. (2024) Sephora products and skincare reviews. Kaggle. [Online]. Available: https://www.kaggle.com/datasets/nadyinky/sephora-products-and-skincare-reviews

[29] A. A. Hagberg, D. A. Schult, and P. J. Swart, "Exploring network structure, dynamics, and function using networkx," in *Proceedings of the 7th Python in Science Conference*, 2008, pp. 11–15.

[30] European Commission, "Cosing: Cosmetic ingredient database," 2026. [Online]. Available: https://single-market-economy.ec.europa.eu/sectors/cosmetics/cosmetic-ingredient-database_en

[31] ——, "Cosing: Functions," 2026. [Online]. Available: https://ec.europa.eu/growth/tools-databases/cosing/reference/functions

[32] H. Abdi *et al.*, "Bonferroni and šidák corrections for multiple comparisons," *Encyclopedia of measurement and statistics*, vol. 3, no. 01, p. 2007, 2007.

[33] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: With Applications in R*, 2nd ed. Springer, 2021.

[34] E. Choi and K. C. Lee, "Effect of trust in domain-specific information of safety, brand loyalty, and perceived value for cosmetics on purchase intentions in mobile e-commerce context," *Sustainability*, vol. 11, p. 6257, 11 2019.

[35] M. Sánchez-Pérez, M. D. Illescas-Manzano, and S. Martínez-Puertas, "Online review ratings: An analysis of product attributes and competitive environment," *Journal of Marketing Communications*, vol. 28, no. 5, pp. 487–505, 2022.

[36] M. Ogorzałek, E. Klimaszewska, M. Mirowski, A. Kulawik-Pióro, and R. Tomasiuk, "Natural or synthetic emollients? physicochemical properties of body oils in relation to selected parameters of epidermal barrier function," *Applied Sciences*, vol. 14, no. 7, 2024. [Online]. Available: https://www.mdpi.com/2076-3417/14/7/2783

[37] S.-Y. Kang, J.-Y. Um, B.-Y. Chung, S.-Y. Lee, J.-S. Park, J.-C. Kim, C.-W. Park, and H. O. Kim, "Moisturizer in patients with inflammatory skin diseases," *Medicina*, vol. 58, p. 888, 07 2022.

[38] F. Andrei, "Sensory science in cosmetics," in *Cosmetic Industry: Trends, Products and Quality Control*. IntechOpen, 2025.