

UNIVERSITÀ
degli STUDI
di CATANIA

Dipartimento
di Fisica
e Astronomia
"Ettore Majorana"



PhD Programme in Complex Systems

Claudio Caprioli

NETWORK GASTRONOMY: EXPLORING THE LINKS BETWEEN RECIPES, CUISINES, AND CHEFS

PhD thesis

Supervisor: Prof. Vito Claudio Latora

2025–2026

University of Catania — Department of Physics and Astronomy "Ettore Majorana"

Acknowledgements

This thesis would not have been possible without the guidance, support, and encouragement of many people who accompanied me throughout these years.

First, I wish to thank **Federico Malizia**, whose friendship and encouragement inspired me to apply for the PhD in the first place, and whose advice and assistance were invaluable in helping me to navigate the world of doctoral studies and research. I am deeply grateful to my supervisor, **Prof. Vito Latora**, for his mentorship, guidance, and for always setting an example of scientific curiosity and rigor. I am indebted to **Prof. Iacopo Iacopini**, who kindly hosted me during my visiting period at Northeastern University London and enriched my research experience with valuable guidance and discussions. My sincere thanks also go to **Andrea Santoro**, who has supported me throughout my research journey with advice and insightful feedback. I am fortunate to have shared this path with my peers **Giovanni Palermo** and **Giambattista Albora**, whose companionship and independence of thought have taught me the value of autonomy in research. Finally, I wish to thank **Simone Poetto**, a friend and colleague with whom I shared perspectives and thoughts on academia and research, which proved both valuable and inspiring.

Beyond academia, I have been lucky to rely on the unwavering support of my friends and family. My heartfelt gratitude goes to **Nicola**, **Paolo** and **Pasquale**, who patiently listened to my endless rants about PhD life and always found time to stand by me, and to **Carlotta**, whose help and kindness made my stay in Catania much easier. Finally, I dedicate my deepest thanks to my family: my parents, **Anna** and **Luigi**, my brother **Fulvio**, and my grandfather **Giovanni**, whose unconditional love and support have provided the foundation without which this work would never have been possible.

Abstract

Food is at once a cultural artifact, a creative practice, and a complex system shaped by human ingenuity and shared traditions. This thesis approaches gastronomy from a computational and network-theoretic perspective, with the overarching aim of characterizing the patterns through which culinary systems are structured and vary across scales—from the molecular composition of ingredients, to the cultural organization of cuisines, to the social networks of chefs. By combining large-scale datasets with tools from network science, it develops a multi-level framework that connects these levels into a coherent picture of culinary organization and diversity.

At the molecular level, the thesis revisits the influential "food pairing hypothesis," which posits that ingredients sharing flavor compounds are more likely to combine harmoniously. To test this idea, we frame ingredient recommendation as a predictive task, where we remove one ingredient from a recipe and ask different models to suggest plausible candidates for replacement. In this framework, similarities—whether based on overlapping chemical compounds or on co-occurrence in existing recipes—are used to rank potential substitutes. The results show that flavor-based similarity can be useful in restricted contexts, for instance within certain ingredient categories or regional traditions, but it does not provide a consistent basis for prediction. By contrast, recipe-based similarity proves to be a stronger and more reliable signal, highlighting that culinary practice is shaped less by molecular compatibility than by culturally transmitted knowledge.

At the cultural level, cuisines are represented as networks of ingredient types, where nodes correspond to broad categories and weighted links capture their co-use in recipes. Analyzing these networks reveals not only differences in the popularity of ingredient types and typical recipe sizes, but also distinctive organizational patterns that serve as culinary fingerprints. European cuisines, for instance, tend to distribute ingredients across multiple types, while certain Asian and South American traditions emphasize the dominance of a single type such as vegetables or spices. These patterns are effectively distilled into maximum spanning trees, which provide simplified backbones that preserve the essential structure of each cuisine. Both full and simplified networks support machine learning models capable of identifying and clustering cuisines with remarkable accuracy, reflecting underlying geo-cultural groupings. In this way, ingredient-type networks expose universal patterns alongside distinctive cultural signatures, offering a new lens on culinary identity and cross-cultural comparison.

At the social level, the focus shifts to the agents of culinary innovation: chefs and their networks of influence. Drawing on biographical and relational data, we model chefs as nodes and mentorship or inspirational relations as directed edges and build a network that reconstructs the genealogical fabric of modern gastronomy, where knowledge and prestige flow across generations. Analysis of this network reveals strong asymmetries of influence, with a few central figures shaping multiple generations while most chefs leave a modest imprint. Temporal layering reinforces this pattern, as ties predominantly run from older to younger chefs across an average generational gap of two decades. Multiple centrality measures highlight the diverse ways chefs achieve prominence—whether as prolific mentors, brokers across traditions, or indispensable nodes whose removal would fragment the network.

The nationality further emerges as a primary axis of organization: the network is assortative with respect to national origin, and cross-national flows reveal clear asymmetries. French chefs dominate as exporters of influence, Italian chefs emerge primarily as receivers, and Spanish chefs project their influence across continents, consolidating Spain's role as an avant-garde hub. Community detection uncovers cohesive clusters aligned with national traditions and stylistic schools, from French and Italian lineages to the experimental Catalan scene. Finally, by extracting a maximum spanning tree weighted by information centrality, we distill the dense relational fabric into a genealogical backbone that provides a global picture of culinary evolution, in which stylistic schools, their pivotal figures, and their points of contact become clearly identifiable.

Taken together, these perspectives demonstrate the value of computational gastronomy as a field at the intersection of complex systems, cultural studies, and food science. By moving across scales—from molecules to cuisines to chefs—the thesis contributes both methodological tools and conceptual insights into how culinary knowledge is organized, transmitted, and transformed.

Contents

1 Introduction **1**
- 1.1 Why study recipes? . . . 1
- 1.2 The computational gastronomy landscape . . . 3
- 1.3 Zoom levels: from molecules to cuisines to chefs . . . 4
- 1.4 Research questions and Thesis roadmap . . . 5

2 Data, Concepts and Methods **8**
- 2.1 Culinary Data . . . 8
  - 2.1.1 Ahn et al.'s Dataset . . . 9
  - 2.1.2 CulinaryDB Dataset . . . 10
  - 2.1.3 Comparing Ingredient Categorization . . . 12
- 2.2 Graphs and Networks . . . 15
  - 2.2.1 Basic Concepts . . . 15
  - 2.2.2 Paths and Connectivity . . . 16
  - 2.2.3 Centrality . . . 18
  - 2.2.4 Assortativity . . . 21
  - 2.2.5 Communities . . . 22
  - 2.2.6 Network Motifs . . . 23
  - 2.2.7 Bipartite graphs . . . 26
- 2.3 Network representations of cuisines . . . 28
  - 2.3.1 Recipe–ingredient bipartite networks . . . 28
  - 2.3.2 Ingredient co–occurrence networks . . . 29
  - 2.3.3 The flavor network . . . 30
- 2.4 Machine Learning: General Framework and Methods . . . 32
  - 2.4.1 The general learning problem . . . 32
  - 2.4.2 Support Vector Machines . . . 33
  - 2.4.3 Hierarchical clustering . . . 35
- 2.5 Recommender Systems . . . 37
  - 2.5.1 The general framework . . . 37
  - 2.5.2 Content-based systems . . . 38
  - 2.5.3 Collaborative filtering . . . 38

3 Re-examining the Food-Pairing Hypothesis **41**
- 3.1 The Food Pairing Hypothesis: From Networks to Recipes . . . 43
  - 3.1.1 Ahn's Dataset . . . 43

3.1.2 Formulation of the Food Pairing Hypothesis . . . 43
3.2 Comparing Flavor-based and Recipe-based Similarity . . . 46
3.2.1 A unified framework for ingredient similarity . . . 47
3.2.2 Item–item similarity metrics . . . 48
3.2.3 How Recommendations Are Generated . . . 54
3.3 From Hypothesis to Prediction . . . 55
3.3.1 How We Evaluate: The Ingredient–Retrieval Task . . . 56
3.3.2 Performance metrics: Mean Average Precision and Hit Ratio . . . 57
3.4 Results of the Retrieval Task . . . 58
3.5 Results of the Retrieval Task . . . 59
3.5.1 Category-restricted retrieval . . . 64
3.5.2 Regional breakdown . . . 66
3.6 What drives compound-based recommendations . . . 67
3.6.1 Fragility check: removing the strongest anchor . . . 67
3.6.2 Fragility depth: beyond the top contributor . . . 69
3.6.3 Ingredient-type bridges . . . 71
3.7 Summary and outlook . . . 72

4 Ingredient-Type Networks as Cuisine Fingerprints 74
4.1 Data and Preprocessing . . . 75
4.1.1 Dataset Description . . . 76
4.1.2 Ingredient Typing . . . 76
4.1.3 Preprocessing Pipeline . . . 78
4.1.4 Formal Definitions . . . 79
4.2 Statistical descriptors of world cuisines . . . 80
4.2.1 Cuisines as Sets of Recipes . . . 80
4.2.2 Recipe size distributions . . . 81
4.2.3 Ingredient Type Popularity . . . 81
4.3 The Networks of Ingredient Combination . . . 85
4.3.1 Constructing the Type-Type Networks . . . 86
4.3.2 Network visualization . . . 87
4.4 Cross-cultural analysis of ingredient type usage . . . 89
4.4.1 Visual analysis of filtered type–type networks . . . 89
4.4.2 Cross-cultural distinctiveness of type usage . . . 92
4.5 Ingredient type networks as fingerprints of cuisines . . . 96
4.5.1 Maximum spanning trees: minimal structural backbones . . . 96
4.5.2 Classification: identifying cuisines from networks . . . 98
4.5.3 Clustering world cuisines . . . 100
4.6 Summary and outlook . . . 101

5 Mapping Influence: The Network of Chefs 103
5.1 The *Atlante dei Cuochi* source . . . 104
5.1.1 Introduction to the source . . . 104
5.1.2 Scope and coverage . . . 105
5.2 Data-curation pipeline . . . 108

5.2.1 Network extraction from the *Atlante* . . . 108
5.2.2 Link validation and classification . . . 108
5.2.3 Metadata enrichment . . . 109
5.3 Descriptive Network Statistics . . . 110
5.3.1 Global overview . . . 110
5.3.2 Degree distributions and inequality . . . 110
5.3.3 Temporal patterns of influence . . . 112
5.3.4 Central figures in the influence network . . . 116
5.3.5 Central ties in the influence network . . . 119
5.4 Communities of Influence . . . 122
5.4.1 Nationality and assortativity . . . 122
5.4.2 Community Structure . . . 125
5.4.3 Backbone of influence . . . 128
5.5 Motif analysis of influence patterns . . . 131
5.6 Summary and Outlook . . . 134

6 Conclusions and Future Work **136**
6.1 Summary of contributions . . . 136
6.2 Limitations and Future Work . . . 137

Appendices **140**

A Ingredient Type Network Figures **141**
A.1 Ingredient popularity and z-score profiles . . . 141
A.2 Backbone networks . . . 143
A.3 Maximum spanning trees (MSTs) . . . 143

B Extended Centrality Rankings **146**

C Community–Nationality Alignment **148**
C.1 Null Model . . . 148
C.2 Algorithmic implementation . . . 149
C.3 Test statistics . . . 151
C.4 Results . . . 154

List of Figures

2.1 Row–normalized contingency between CulinaryDB types (rows) and Ahn types (columns). For each CulinaryDB type, color gives the share of aligned ingredients that Ahn assigns to each of its types. Concentrated rows indicate close agreement; broader rows reveal systematic splits or merges across the two schemes. 13

2.2 Side-by-side examples of (a) an undirected graph and (b) a directed graph with colored nodes. Colors are consistent across panels for nodes A–F. 15

2.3 Contrasting hierarchical structures: (a) a rooted tree (a special DAG with a single root and unique parent for each non-root), and (b) a general DAG allowing multiple sources/sinks and merges without cycles. 17

2.4 Toy network with two communities: dense intra-community connectivity (solid gray) and sparse inter-community links (dashed). Nodes are colored according to their community membership (light blue and orange). 22

2.5 Canonical triads. Left: four undirected triads (0, 1, 2, and 3 edges). the full set of 16 non-isomorphic directed triads, classified according to the *triad census* [Holland and Leinhardt, 1976]. Each directed triad is labeled with its canonical census code (e.g., 021D, 030T, 111U) and its conventional name when applicable (e.g., in-star, out-star, chain). Nodes are color-coded: blue for undirected, orange for directed. 25

2.6 A toy bipartite (two–mode) network with node sets U and V. Edges only connect nodes across sets. 26

2.7 **Backbone of the flavor network.** Projection of the ingredient–compound bipartite graph onto ingredients, with edge thickness proportional to the number of shared volatile compounds. The pruned (backbone) view reveals chemically coherent clusters (e.g., meats, fruits, herbs/spices). Reproduced from Ahn et al. [2011]. 31

2.8 **Linear SVM geometry.** The separating hyperplane (solid line) maximizes the margin to the nearest points of either class. The dashed lines depict the margins. Support vectors (red outlines) are the only points that influence the decision boundary. The normal vector w is perpendicular to the separating hyperplane, and the margin width is $2/\|w\|$. 34

3.1 Bipartite ingredient–compound network (left) and its weighted projection onto the ingredient layer—the *flavor network* (right). Edge labels in the projection report the number of shared compounds between each ingredient pair. 45

3.2 **Comparison of recipe-based and compound-based recommendation methods across performance metrics and stability.** (a–b) **Slopegraphs** show paired HR@20 (a) and mAP (b) of each recipe-based method (square markers) against its compound-based counterpart (triangle markers). Each horizontal line connects a method family, revealing that recipe-based recommenders systematically achieve higher average values than compound-based ones. (c–d) **Rank stability plots** show the distribution of per-iteration ranks for HR@20 (c) and mAP (d) (lower is better), across all methods. Boxplots are colored by recommender category: recipe-based, compound-based, and direct baselines. The tighter distributions of recipe-based methods highlight their greater robustness, whereas compound-based methods exhibit more dispersed and unstable ranks across iterations. Rankings of the direct baselines (Random, Popularity) are included for context. 61

3.3 **Performance distributions across recommenders.** Ridgeline plots show the per-iteration distributions of HR@20 and mAP for recipe-based methods (a-d) and compound-based methods (b-c). Within each family, methods are ordered by their average performance, with medians marked by vertical ticks. Recipe-based similarities not only achieve higher scores on average but also exhibit narrower distributions, reflecting greater stability across train–test splits. Compound-based similarities, by contrast, show wider, noisier profiles, with only Taxonomy and Sapling emerging as relatively stronger within this weaker family. 63

3.4 **Impact of category restriction on recommendation performance.** Bars show the percentage change in (a) HR@20 and (b) mAP when candidate recommendations are restricted to the ground-truth ingredient category, compared to the unconstrained setting. Error bars correspond to propagated standard errors (Eq. 3.7). We display the ten compound-based recommenders with the largest average gains, the ten recipe-based recommenders with the largest average gains, and both direct baselines (Random and Popularity). Recommenders are ordered from highest to lowest gain, revealing which models derive genuine benefit from within-category pairing as opposed to near-duplicate substitutions. 65

3.5 **Compound-based performance by world region.** Ridgeline plots show distributions of mAP across 250 trials for the best-performing compound-based recommender in each region (orange) compared against the random baseline (green). (a) Standard recommendation (unrestricted). (b) Category-restricted recommendation, where candidate ingredients are limited to the ground-truth type. 66

3.6 **Leave-one-influence-out fragility analysis.** Scatter plots of target rank before (x-axis) and after (y-axis) removing the most influential observed ingredient. Panels (a,c,e) show recipe-based recommenders; (b,d,f) show their compound-based counterparts. From left to right: (a,b) Taxonomy index, (c,d) Sapling, (e,f) BM25-Cosine. Across all three similarities, compound-based recommenders are markedly fragile, with strong reliance on a single anchor ingredient. Recipe-based methods degrade more gently, except in the Sapling case where top-ranked targets also suffer sensible worsening. 69

3.7 **Leave-one-contributor-out fragility across recipe sizes.** Boxplots show the distribution of rank changes Δr_m after removing the m-th most similar contributor ($m = 1, \ldots, 5$), stratified by recipe size. Results are shown for the six recipe sizes with the largest number of test recipes, with coverage and total counts reported in the figure title. Recipe-based (blue, left in each pair) and compound-based (orange, right) methods highlight their different reliance on top contributors. Numbers above panels indicate the number of recipes contributing to each size. 70

3.8 **Same-type prevalence of strongest contributors.** Fraction of cases in which the most influential observed ingredient j^* shares the same type as the hidden target i^*, stratified by target rank (r_{full}). Bars show averages across iterations for recipe-based (blue) and compound-based (orange) recommenders, with error bars indicating standard deviation. High prevalence at top ranks indicates strong reliance on intra-type anchors. 72

4.1 **Scaling of recipes vs total ingredients across world cuisines.** Relationship between the number of recipes and the total number of individual ingredients across various cuisines. Each data point represents a cuisine and is color-coded according to the continent affiliation, as illustrated in the world map reported in the top-left corner. A lin-log scale has been used in the plot, and the reported best-fit line highlights the exponential growth of the number of recipes with the total number of ingredients. *Figure reproduced from Caprioli et al. [2024].* 80

4.2 **Recipe Size Distribution.**Ridge plot showing the empirical probability distribution of recipe sizes across cuisines from different continents. Each distribution, sorted by average recipe size, corresponds to a cuisine, and is colored according to the continent affiliation (as per fig. 4.1). *Figure reproduced from Caprioli et al. [2024].* 82

4.3 **Cross-cultural variation in ingredient type usage.** Comparison of absolute popularity (a) and relative over/under-use (b) of ingredient types across 23 world cuisines. *Figure reproduced from Caprioli et al. [2024].* 83

4.4 **Ingredient usage profiles in selected cuisines.** Popularity values (a) and z-score normalized values (b) for representative cuisines. *Figures reproduced from Caprioli et al. [2024].* 84

4.5 Unfiltered type–type graph of the American cuisine. Node sizes are proportional to relative abundance v_t, and edge thicknesses to co-occurrence weights w_{t_i,t_j}. he complete connectivity of the American type graph exemplifies the need for filtering methods to make structural patterns visible. 87
4.6 **Comparison of filtering strategies for the American cuisine**. The fixed-K network (a) showcases only the strongest links, producing a compact view where weaker associations are suppressed. The backbone network (b), in contrast, retains many lower-weight links that are statistically significant given local connectivity, bringing to light finer-grained structures that are hidden by rank-based filtering. Node sizes are proportional to relative abundance v_t, and edge thicknesses to co-occurrence weights w_{t_i,t_j}. 89
4.7 **Fixed-K type–type networks (top 50 edges).** Node size: relative abundance v_t; edge thickness: co-occurrence weight w_{t_i,t_j}. Using a common edge budget foregrounds the most frequent pairings per cuisine while enabling controlled visual comparison across panels. 90
4.8 **Backbone type–type networks (disparity filter).** Edges are retained when their weight is statistically significant relative to a null in which a node's strength is uniformly distributed among its incident edges. 91
4.9 **Cross-cultural distinctiveness of type usage** Distribution of standardized z-scores across cuisines. Each strip corresponds to a cuisine, with points indicating the z-score of a given type pair relative to the global distribution. Two pairs are highlighted in each cuisine: one with a high z-score and one with a low z-score, representing significant positive and negative deviations. Long-tailed distributions reveal culturally distinctive types, whereas narrow bands indicate more universal prevalence. *Figure reproduced from Caprioli et al. [2024].* 93
4.10 **Maximum spanning trees of type–type networks.** Examples for six representative cuisines. The trees highlight the most important associations while preserving connectivity. Distinctive patterns emerge across cuisines: Indian (**c**) shows a star-like structure with *Spices* at the hub, reflecting their centrality; Mexican (**e**) is similarly organized around *Vegetables*; French (**b**) emphasizes *Dairy* as a central connector to *Meat* and *Fungi*; Italian (**d**) features dual hubs, *Herbs* and *Spices*, a configuration shared only with Thai cuisine (see Fig. A.4). American cuisine (**a**) has a more distributed layout with the largest diameter, while Central European cuisine (see Fig. A.4) shows the longest diameter overall. *Figure reproduced from Caprioli et al. [2024].* . . . 97
4.11 **Identification and clustering of world cuisines**. (a–c) Confusion matrices of SVM classifiers trained on (a) ingredient frequency vectors, (b) MSTs of ingredient type networks, and (c) full ingredient type networks. Graph-based representations substantially outperform frequency-based ones. (d) Hierarchical clustering of type networks using Jensen–Shannon distance (average linkage), with six clusters identified at cut distance 0.21. (e) Geographical mapping of the clusters, highlighting both regional and cultural ties (e.g. the uniqueness of Indian cuisine, Iberian–Latin American connections, and Asian subclusters). *Figure reproduced from Caprioli et al. [2024].* 99

4.12 2d t-SNE plots of ingredient type graphs 101

5.1 **Cover of *L'Atlante dei Cuochi*** (Slow Food Editore, 2023). Image reproduced for illustrative purposes. 105
5.2 Representative biographical page for Albert Adrià and Ferran Adrià from *L'Atlante dei Cuochi*. Each entry includes a portrait sketch, birthplace and birth year, a short biography, and—crucially—the two columns of influence ties at the lower left: on the left, chefs who influenced the profiled figure; on the right, chefs influenced by them. Illustration by Martina Dirce Carcano. . 106
5.3 Geographic distribution of chefs in the *Atlante*, by country of origin or nationality (257 chefs). 107
5.4 Temporal distribution of chefs in the *Atlante* by birth year or main career era (257 chefs), showing a recency bias. 107
5.5 **Global chef influence network.** Visualization of the largest weakly connected component (225 chefs). Nodes are scaled by out-degree; names are displayed only for chefs with total degree greater than 5. Node colors represent the chef's nationality; edge colors indicate the type of influence: grey for *Working together*, green for *Inspiration without collaboration*, orange for *Uncertain*. 111
5.6 **Degree distributions and inequality.** (a) In-degree distribution (PMF, log–lin). (b) Out-degree distribution (PMF, log–lin). (c) Lorenz curves for in- and out-degree compared to the line of equality. The plots show that most chefs maintain only a handful of ties, while a minority concentrate a disproportionate share of connections, especially in terms of out-degree. . . . 112
5.7 **Temporal patterns of influence.** (a) Distribution of chefs by decade of birth, highlighting the concentration in the 1960s–1980s cohorts. (b) Histogram of age gaps between source and target of ties. The blue dashed line marks zero (younger-to-older ties), while the red solid line marks the mean gap (18.8 years). The inset reports the share of ties with positive gap ($P(\Delta > 0)$), the median gap, and the mean gap. (c) Birth year versus in-degree (blue) and out-degree (orange), with LOWESS smoothing. Out-degree peaks among the 1940s cohorts and declines in later generations, while in-degree rises steadily toward the youngest chefs. (d) Generational depth—the length of the longest directed path originating from a chef—plotted against birth year, showing a decreasing trend: earlier-born chefs anchor deeper genealogical trees, whereas younger chefs remain closer to the leaves. 113

5.8 **Temporal layout of the chef influence network.** Visualization of the largest weakly connected component, with each node placed at its birth year on the horizontal axis and assigned to a horizontal "band" corresponding to a community detected via Louvain modularity on the undirected projection. Within bands, nodes are arranged using a spring layout with small vertical jitter and collision avoidance. Directed edges are drawn as light curved lines, bending upward or downward depending on relative vertical order. Node size reflects within-community prominence (z^w-score of within-module degree), and color encodes community identity. The two most prominent hubs in each community are labeled. The layout highlights how influence flows across generations (x-axis) while remaining structured within community bands (y-axis). 115

5.9 **Cross-national flows of influence between chefs.** Alluvial diagram of mentorship and collaboration ties running between chefs of different nationalities (same-nationality ties excluded). Stream width is proportional to the number of directed edges from source nationality (left) to target nationality (right). The visualization highlights France as a dominant exporter of influence, Italy as a major receiver, and Spain as a bridge toward Latin America and Northern Europe, while other traditions appear more domestically oriented. 123

5.10 **Communities in the working–together network.** Visualization of the largest weakly connected component (213 chefs, 344 undirected edges). Nodes are colored by Louvain community, sized by out-degree, and labeled when total degree exceeds 5. The modularity of the partition is $Q = 0.633$, indicating strong community structure. 127

5.11 **Maximum centrality spanning tree of the chef influence network.** Shown is the maximum spanning tree (rooted at the oldest chef) extracted from the largest weakly connected component using *edge information centrality* as edge weight (cf. Section 2.2.2). Node color encodes nationality (see legend at the bottom). Chef names are displayed as node labels, and both node and label sizes are proportional to each chef's out-degree in the original directed network (number of downstream influence links). Edges are oriented for visualization by preferring the original direction when unique, otherwise from older to younger based on birth year. 130

5.12 Directed 4-node motifs considered in the analysis. BiFan: two sources jointly targeting two sinks; FFBox: two feed-forward paths (A→B→D and A→C→D) plus direct A→D; P4: directed path of length 3; C4: directed 4-cycle; TT4: transitive tournament (acyclic complete orientation) on 4 nodes. 132

A.1 **Ingredient type popularity profiles for 18 additional cuisines.** Relative frequencies of ingredient types, complementing the examples shown in Section 4.2.3. 142

A.2 **z-score profiles of ingredient type popularity.** Standardized z-scores corresponding to the frequency distributions reported in Figure A.1, highlighting over- and under-represented ingredient types relative to the global baseline. 142

A.3 Networks of ingredient combinations 144

A.4 Maximum Spanning Trees of Ingredient Combination Networks 145

List of Algorithms

1 Mixing-preserving edge swap with connectivity enforcement 149

List of Tables

2.1 Regional coverage and country/area mapping for Dataset 1 (Ahn et al.). Source: Table S2 in the Supplementary Information of Ahn et al. [2011]. . . 9
2.2 Cuisine–region mapping in CulinaryDB. A cuisine may correspond to a single nation-state or to a broader geo-cultural area. 11
2.3 Comparison of the two recipe datasets used in this thesis. 12

3.1 Summary statistics of the dataset used for the analysis. 44
3.2 Similarity metrics used in item-item collaborative filtering. Each metric measures the proximity between ingredients i and j using their co-occurrence in context nodes α (recipes or flavor compounds). See main text for a description of variables. 50

4.1 **Cuisine-Countries mapping** 77
4.2 **Dominant type–type associations by cuisine.** For each of the 23 cuisines, the five heaviest type–type pairs are listed, together with the cumulative fractions explained by the top 5 and top 50 links. In most cuisines, the top 5 (about 3% of links) explain $\sim$ 15% of total weight and the top 50 (about 25%) explain $\sim$ 75%; East Asian cuisines depart from this pattern with lower concentration. 95
4.3 **Recurrent and variable type–type associations across cuisines.** (**Left**) Top ten pairs by *average* weight across cuisines (global core). (**Right**) Top ten pairs by *weight variability* (standard deviation), highlighting culturally sensitive associations (notably *Dairy*-involving pairs). 96

5.1 **Global statistics** of the chef influence network. 110
5.2 **Top chefs by in- and out-degree centrality.** In-degree measures how many mentors or predecessors point to a chef; out-degree measures how many successors or mentees a chef has influenced. 116
5.3 **Top chefs by alternative centrality measures.** The three rankings highlight complementary roles: betweenness emphasizes brokers who connect traditions, PageRank rewards chefs embedded in prestigious lineages, and efficiency loss identifies systemic anchors whose removal would disrupt global cohesion. 118
5.4 **Top 10 edges by *edge betweenness* — directed vs. undirected (LCC).** Type codes: W = working together; I = inspiration; U = uncertain. 120

5.5 **Top 10 edges by *information centrality* — directed vs. undirected (LCC).** Type codes: W = working together; I = inspiration; U = uncertain. 121

B.1 **Top 20 chefs by betweenness, PageRank, and Information Centrality.** Values are normalized centrality scores. Nationality codes: FR = France, IT = Italy, ES = Spain, UK = United Kingdom, US = United States, DK = Denmark, CH = Switzerland, AU = Australia, BR = Brazil, AR = Argentina, JP = Japan, PT = Portugal, IN = India. 147

C.1 Summary of the rewired-graph refit null model. 153

C.2 Rewired-graph null results on the work-layer LCC ($|V| = 213$, $|E| = 344$). Null preserves degree sequence, connectivity, and the full nationality mixing matrix; Louvain refit per sample. Right-tailed tests used for NMI/AMI and within-community assortativity $\bar{r}$; two-sided tests for the other measures. . . 154

Chapter 1

Introduction

1.1 Why study recipes?

Food has always occupied a central role in the human story. It is not merely a source of sustenance but a driving force behind biological evolution, cultural practices, and even geopolitical transformations. The practice of transforming raw food to make it palatable or safer for consumption predates the emergence of *Homo sapiens* itself Zohar et al. [2022], and the mastery of fire and cooking marked a decisive turning point in human evolution, enabling higher caloric intake, reducing chewing time, and profoundly shaping the physiology of our species Wrangham et al. [1999]. Later, the domestication of plants and animals during the so-called Neolithic or Agricultural Revolution established the basis for settled societies, tying food production to the rise and stability of civilizations Diamond [1997], Smith [1998].

Food has often been at the heart of conflict and cooperation alike: wars have been waged over resources such as spices, salt, or grain, while trade routes for ingredients catalyzed exchanges of knowledge, culture, and technology. The search for new trading routes was also one of the key drivers of European expansion into the Americas, profoundly reshaping global geopolitics, economies, and diets in what has been termed the "Columbian Exchange" Crosby [1972].

Within this broad human–food relationship, *recipes* occupy a privileged place. At their simplest, recipes encode a transformation: a set of ingredients combined through a sequence of operations, mediated by technologies, that include not only fire and fermentation but also utensils, cookware, and diverse cooking methods. Beyond mere instruction, however, recipes are cultural artifacts. They transmit knowledge across generations, codify shared tastes, and embody the values and taboos of a community. Contrary to common belief, recipe books are almost as old as writing itself. The earliest examples we know of are four Babylonian clay tablets from around 1700 BC, inscribed with cuneiform script and containing primordial recipes Bottéro [1995]. From then on, civilizations have repeatedly sought to record culinary practices: from the Roman *De re coquinaria* attributed to Apicius Apicius [2006], to the sophisticated medieval Arab collections such as the *Kitāb al-Ṭabīkh* al Warrāq [2007], and the Renaissance European cookbooks that circulated widely across courts and cities Capatti and Montanari [2003]. These texts not only preserve culinary techniques but also serve as historical documents, revealing how cuisines evolve over time, how ingredients appear,

disappear, or change in meaning, and how taste is continuously renegotiated.

The history of cuisine has often been shaped by slow, incremental changes. Technological innovation has consistently played a central role in this process: from the introduction of new utensils such as ovens or grinding mills, to preservation techniques like canning and refrigeration, each advancement has reshaped culinary practice and possibilities. Yet the most striking moments of transformation are mainly happening through cultural shift, reshaping the very place of cuisine within society. One such turning point was the birth of the modern restaurant in 18th-century France. Emerging in Paris in the 1760s and 1770s, these establishments offered prepared meals to paying customers in a public setting, departing from earlier traditions of inns and taverns. The restaurant crystallized a new social relation: food prepared by professional cooks, consumed in exchange for money by an urban bourgeois clientele seeking both nourishment and social distinction Spang [2000]. At its center was the figure of the *chef*, situated at the intersection of craft, engineering, art, and science. Chefs did not merely preserve culinary traditions—they actively experimented with ingredients, invented new recipes, and articulated principles to guide others. For centuries, they were the primary custodians of gastronomic knowledge.

In the 20th century, the rise of mass media redefined the chef's role. Once primarily confined to the kitchen, chefs became public figures and cultural authorities, popularizing cooking through television, magazines, and later the internet. Mass media not only celebrated individual chefs but also helped shape national culinary identities, constructing symbolic repertoires of "traditional" dishes and techniques. At the same time, radio, print, and later television programs played an important educational role, teaching generations of home cooks—most often women—how to prepare meals in accordance with evolving standards of nutrition, taste, and domestic efficiency. Cooking, once perceived as domestic labor, was thus reframed as creativity, performance, and even entertainment, paving the way for the transformations brought by the digital age.

This new visibility set the stage for another transformation: the advent of the data revolution.

With the diffusion of personal computers and Web 2.0 platforms, recipes began to be digitized at unprecedented scale. Online repositories now contain millions of recipes, collaboratively created and shared by users worldwide. Crucially, these databases are user-driven: individuals who are passionate about cooking contribute recipes, tips, and evaluations. Their active participation is itself a cultural product, nurtured by the mass media's earlier popularization of cuisine as a subject of entertainment and creative pursuit. What began as a top-down dissemination of culinary knowledge through books and television thus evolved into a bottom-up, participatory ecosystem of recipe sharing. This digitization represents a second turning point in the history of culinary knowledge: recipes, once recorded in scarce and curated books, are now available as vast, searchable databases. Beyond serving as cultural archives, these platforms have also enabled new applications such as recommender systems and ingredient substitution tools, which adapt recipes to users' nutritional, cultural, or stylistic needs Teng et al. [2012]. This abundance of data has invited scientists to approach food in new ways, marking the emergence of what is often called *scientific gastronomy*. For the first time, chefs and researchers are collaborating systematically to address age-old questions: Why do certain ingredients combine harmoniously while others clash? Which cooking techniques optimize flavor or nutrition? What principles underlie the experience of taste?

These questions, once posed anecdotally, can now be investigated quantitatively.

1.2 The computational gastronomy landscape

Although the fascination with the science of cooking has ancient roots, systematic connections between gastronomy and science are relatively recent. One of the earliest figures to explore this intersection was Benjamin Thompson (1753–1814), also known as Count Rumford, who experimented with the physics of heat and inadvertently pioneered the principles of slow cooking. Around the same period, Antoine Lavoisier (1743–1794), often considered the father of modern chemistry, carried out experiments on fermentation, respiration, and the caloric content of food, laying the groundwork for nutritional chemistry [Lavoisier, 1789]. In the 19th century, Justus von Liebig (1803–1873) advanced this line of inquiry by developing a chemical understanding of proteins, inventing meat extracts, and promoting the idea of "scientific cookery" based on the chemistry of nutrition [von Liebig, 1847]. These pioneers illustrate how the birth of modern chemistry created new tools to investigate food, taste, and cooking.

By the mid-20th century, this convergence of physics, chemistry, and gastronomy began to crystallize into systematic reference works. Among the most influential was Fenaroli's *Handbook of Flavor Ingredients* [Fenaroli, 1971], which catalogued the chemical components responsible for aroma and taste. Not long after, Harold McGee's landmark book *On Food and Cooking* McGee [1984] brought together insights from chemistry, biology, and history, providing chefs and scientists alike with a comprehensive foundation for understanding the chemical and physical basis of cooking. McGee's work, in particular, directly inspired chefs such as Heston Blumenthal to seek deeper explanations of taste and texture.

A movement was also unfolding within professional kitchens. Chefs began to experiment with science as a source of culinary innovation, giving rise in the late 20th century to what became known as *molecular gastronomy* [Cassi, 2011]. The field was pioneered by Hervé This, who formalized the term and investigated the scientific mechanisms behind traditional preparations, and soon epitomized by the work of Ferran Adrià at El Bulli. By systematically applying scientific knowledge to cooking, Adrià and contemporaries introduced techniques such as spherification and foams, expanding the culinary repertoire and demonstrating that science could serve not only as an explanatory tool but also as a creative engine. This wave of innovation encouraged scientists to turn their attention to gastronomy as a legitimate subject of study [This, 2006]. In parallel, Heston Blumenthal's pursuit of fundamental taste principles led to the formulation of the *food pairing hypothesis*, which proposed that ingredients sharing volatile compounds are more likely to taste good together.

While these contributions firmly established a link between gastronomy and science, they largely remained confined to the laboratory and the kitchen. What truly transformed gastronomy into a domain amenable to large-scale, quantitative study was the emergence of digital data and computational methods. The digitization of recipes, the cataloguing of molecular compounds, and the exponential growth of computational power collectively enabled gastronomy to enter the era of data science Ahnert [2013]. These particular conditions, combined with the significant attention attracted by Blumenthal's hypothesis, culminated in the first large-scale data-driven studies of food. In 2011, Ahn, Barabási, and collaborators

introduced the *flavor network*, a tripartite graph linking recipes, ingredients, and flavor compounds. Their work demonstrated that food pairing was indeed a strong regularity in North American and Western European cuisines, though not a universal law. More importantly, it showed that complex networks provided a natural language for describing culinary systems, opening the way for quantitative models of recipes and cuisines.

This pivotal work marked the birth of *computational gastronomy*, characterized by the data-driven perspective and a rigorous scientific approach Goel and Bagler [2022]. The field has since rapidly evolved to acquire multiple dimensions: from chemical and nutritional composition to historical, recreational, and health-related aspects of food.

This perspective treats recipes not just as cultural artifacts but also as structured data amenable to algorithmic analysis. Computational gastronomy seeks to uncover the principles that shape ingredient usage, recipe formation, and culinary evolution, situating gastronomy within the broader framework of cultural analytics. It is within this landscape that the present thesis is positioned.

1.3 Zoom levels: from molecules to cuisines to chefs

The chemical composition of food provides the most fundamental layer of information for understanding gastronomy. Harnessing this information has proven essential not only for culinary science but also for advancing fields as diverse as nutrition, epidemiology, and public health. For instance, using large-scale food purchase data and health records, Aiello et al. Aiello et al. [2019] showed that nutrient diversity and calorie amounts are the strongest predictors of the prevalence of diseases related to metabolic syndrome. Similarly, a substantial body of research has focused on analyzing the nutritional composition of foods Barabási [2020], Kim et al. [2015], Menichetti and Barabási [2022], revealing the intricate interplay between nutritional value and chemical profile. Such analyses have identified key nutrients responsible for the nutritional fitness of foods and have led to the proposal of new measures of dietary fitness that do not correlate with classical healthy diet scores Kim et al. [2015], Wang et al. [2024].

These insights underscore the centrality of food chemistry for addressing pressing global issues such as over-consumption, malnutrition, and food-related diseases. Over-consumption and malnutrition, often seen as opposite conditions, are in fact deeply interconnected, particularly in developing countries Wang et al. [2024], Barrera and Hertel [2023]. At the same time, machine learning applications have shown promise in predicting the degree of food processing Menichetti et al. [2023], opening new avenues for both scientific knowledge and the design of public health policies. Yet, as Barabási et al. Barabási et al. [2020] emphasize, the majority of food components remain systematically unquantified, making causal relations between food and health extremely difficult to establish. A complete molecular map of food is still far from achieved.

The potential of linking chemistry and cuisine is exemplified by the *food pairing hypothesis*, introduced by chef Heston Blumenthal and empirically tested by Ahn et al. Ahn et al. [2011]. The hypothesis posits that ingredients sharing volatile compounds are more likely to taste good together. Crucially, the same study revealed that the hypothesis is not universal: while Western cuisines tend to rely on compound sharing, East Asian cuisines appear to

avoid it. This cultural dependence highlighted both the power and the limits of chemistry-centered approaches. They provide mechanistic insight into taste but cannot, on their own, account for the cultural rules that guide how ingredients are actually combined.

Later work sought to refine these ideas. Simas et al. Simas et al. [2017] proposed the notion of *food bridging*, whereby two ingredients that do not share compounds directly may still be compatible if a third ingredient connects them through overlapping volatiles. This extension illustrated that cuisines can exploit both direct and indirect molecular affinities, further complicating the search for universal principles.

Moving beyond molecules, cuisines can be studied as structured cultural systems. Recipes encode not only chemical affinities but also the historical, geographical, and symbolic choices of communities. Early quantitative studies revealed striking regularities, such as the Zipf-like distribution of ingredient usage Kinouchi et al. [2008] and the bell-shaped distribution of recipe sizes Ahn et al. [2011], suggesting that despite their diversity, cuisines obey common organizational patterns. Analyzing cuisines at this scale sheds light on how cultures maintain continuity while adapting to new ingredients, and how they differentiate themselves through characteristic combinations and categories of food.

Yet cuisines are not autonomous systems: they are shaped and transformed by the actions of individuals and institutions. At a higher zoom level, the figure of the chef emerges as a key agent of culinary evolution. Chefs inherit techniques from mentors, reinterpret traditions, and introduce innovations that can ripple across entire cuisines. Their influence is amplified by restaurants, mass media, and global recognition systems, turning individual creativity into collective change. Mapping mentorship networks, stylistic lineages, and innovation trajectories thus becomes crucial for understanding how gastronomy evolves over time.

Taken together, these three zoom levels — from molecular composition, to cultural structure, to human agency — provide complementary perspectives on gastronomy. They are deeply interconnected: chemistry defines the palette of possible combinations, cuisines codify and stabilize practices within cultural frameworks, and chefs drive experimentation and transformation. It is precisely at the intersection of these levels that the present thesis situates its inquiry.

1.4 Research questions and Thesis roadmap

This dissertation is guided by the overarching aim of understanding how recipes emerge, evolve, and vary across cultures when studied through the lenses of science and computation. We formulate three research questions, each corresponding to a different zoom level of culinary analysis:

1. **At the molecular level: Do ingredients combine according to universal principles?** This question revisits the celebrated *food pairing hypothesis*, which posits that ingredients sharing flavor compounds are more likely to be compatible. While the hypothesis has been influential in both kitchens and laboratories, its generality and predictive value remain debated.

2. **At the cultural level: How are cuisines organized in terms of ingredient usage?** Cuisines encode cultural memory, resource constraints, and collective taste

preferences. By representing cuisines as networks of ingredient categories, we can capture their structural fingerprints and compare culinary systems across the globe.

3. **At the social level: How do chefs and institutions shape culinary evolution?** Beyond molecules and ingredients, cuisines are driven by the agency of chefs, whose innovations, mentorship ties, and professional trajectories influence how recipes evolve.

The dissertation is structured into four main chapters, each addressing these questions at different scales:

- **Chapter 2: Data, Concepts and Methods.** We begin by introducing the basic tools of network science that underpin the analyses developed throughout the thesis. This chapter presents fundamental definitions—including graphs, paths, trees, centralities, assortativity, and community structure—and illustrates how these concepts provide a vocabulary for reasoning about culinary systems. The goal is to establish the theoretical and methodological foundation upon which the subsequent chapters build.

- **Chapter 3: Re-examining the Food Pairing Hypothesis.** In this chapter, we revisit the food pairing hypothesis from a predictive perspective. We ask whether chemical similarity between ingredients, based on shared flavor compounds, truly offers an advantage for recommending ingredient combinations, or whether patterns observed directly in recipes provide a stronger signal. To answer this, we design a systematic comparison between chemistry-driven and data-driven approaches, evaluating their effectiveness across ingredient categories and world cuisines. This allows us to assess the circumstances under which chemistry can meaningfully contribute to culinary recommendation, and when cultural practice proves to be the stronger guide.

  The content of this chapter builds upon the work presented in *Food Pairing Unveiled: Exploring Recipe Creation Dynamics through Recommender Systems* [Palermo et al., 2024], by Palermo, Caprioli, and Albora. The chapter can be seen as an extension of the study, broadening the discussion of compound-based similarity versus recipe-based similarity as guiding principles for ingredient recommendation. The author of this thesis contributed to the development of the idea behind the paper, carried out part of the analysis, participated in the writing of the manuscript, and produced a part of the figures included therein.

- **Chapter 4: Ingredient-type Networks as Culinary Fingerprints.**

  Here we move from individual ingredients to the cultural organization of cuisines. Each cuisine is represented as a network of ingredient types, with weighted edges capturing their co-use in recipes. Through this abstraction we uncover both universal patterns and distinctive cultural signatures: cuisines reuse and recombine successful combinations rather than exhaustively exploring possibilities; ingredient type hierarchies distinguish globally common categories from marginal ones; and culturally specific associations mark the identity of culinary traditions. We show that ingredient type networks function as compact fingerprints, with maximum spanning trees exposing

minimal structural backbones and classification analyses enabling accurate identification and grouping of cuisines.

This chapter is based on the work *The networks of ingredient combinations as culinary fingerprints of world cuisines* [Caprioli et al., 2024], which systematically investigates the mesoscopic organization of global cuisines through network analysis. The present chapter illustrates and contextualizes the findings of the paper, situating them within the broader scope of this thesis. The author of this thesis conceived the idea, performed the analysis, produced the figures, and collaborated closely with the co-authors in the preparation of the manuscript.

- **Chapter 5: Mapping Influence: The Network of Chefs.** The final chapter situates cuisine within its human dimension by modeling chefs as agents embedded in networks of influence, mentorship, and institutional ties. Drawing on the *Atlante dei Cuochi* [Vignoli and Bolasco, 2023] enriched with open knowledge bases, we analyze temporal patterns of influence, identify key mentors and heirs via centrality, explore community structures shaped by national traditions and stylistic schools, and extract genealogical backbones through maximum spanning trees. This chapter highlights how haute cuisine evolves through the interplay of individuals, relationships, and regions.

  The material presented in this chapter is the basis of an ongoing collaborative project tentatively titled The Networks of Chefs: Influence, Mentorship, and Culinary Innovation (with Andrea Santoro, Iacopo Iacopini and Vito Latora as co-authors). The work is currently under preparation. The author of this thesis developed the original idea, performed the analysis contained in the chapter, and is leading the drafting of the manuscript.

Together, these four chapters form a multi-scale framework for the study of gastronomy: from molecular compounds to cultural systems to individual chefs. By moving across zoom levels, the thesis situates computational gastronomy within both the natural sciences and the humanities, bridging chemical, cultural, and social dimensions of food.

Chapter 2

Data, Concepts and Methods

This chapter lays the methodological foundations for the thesis. We begin by introducing the two recipe datasets employed throughout—Ahn et al.'s recipe corpus and CulinaryDB—emphasizing their complementary strengths (chemical annotations vs. balanced cultural coverage) and clarifying their respective roles in later chapters. Next, we present the graph–theoretic toolkit that underpins both the representation of cuisines and the study of cultural transmission among chefs. Within this framework we formalize bipartite recipe–ingredient and ingredient–compound graphs, their one-mode projections, and normalized similarity networks, and we establish the algebraic notation used consistently across the thesis. Finally, we turn to recommender systems, which provide a way to operationalize ingredient similarity and evaluate its predictive utility. We review neighborhood-based (item–item) collaborative filtering and fix the notation for recommendation scores, thereby bridging the descriptive and predictive components of our analyses.

2.1 Culinary Data

Recipes can be formally described as sets of ingredients, each ingredient in turn belonging to a broader category or type. When aggregated across culinary traditions, such collections of recipes form structured datasets that capture the cultural diversity of world cuisines as well as the biological and chemical diversity of food. In particular, ingredients can be described not only by their culinary function but also by their underlying chemical composition, which determines flavor and influences how ingredients combine. In this thesis we rely on the curated, large-scale repositories of recipes drawn from multiple online sources. These datasets differ in scope and emphasis: one provides detailed chemical annotations of ingredients, while the other prioritizes balanced cultural coverage and provides a more detailed classification of ingredient types. Together, they supply the raw material for all subsequent analyses, whether these concern the chemical composition of ingredients, the organization of culinary practices across cuisines, or the representation of food as networks. A detailed description of the two datasets used in this work, along with a comparison of their main characteristics, is provided below (see also Table 2.3).

2.1.1 Ahn et al.'s Dataset

The first dataset employed in this thesis is a recipe corpus introduced by Ahn et al. [2011], compiled from three large online platforms: AllRecipes, Epicurious, and Menupan. The dataset contains a total of 56,498 traditional recipes drawn from multiple regional cuisines and mapped to a standardized vocabulary of 381 ingredients, each annotated with the set of volatile compounds it contains (1,021 unique compounds overall).[1] On average, recipes include approximately eight ingredients. Each recipe is represented as a set of standardized ingredients, enabling the construction of both ingredient–compound and ingredient–ingredient networks.

Geographic coverage. Recipes are assigned to *eleven* higher-level regional cuisine groups by aggregating site labels. Five groups exceed 1,000 recipes and were the focus of several analyses in Ahn et al. [2011]; in Table 2.1 we report the *full* mapping, including smaller regions, to document coverage comprehensively (counts as reported in Table S2 of the Supplementary Information).

Regional cuisine	**Recipes**	**Cuisines included**
North American	41,525	American, Canada, Cajun, Creole, Southern soul food, Southwestern U.S.
Southern European	4,180	Greek, Italian, Mediterranean, Spanish, Portuguese
Latin American	2,917	Caribbean, Central American, South American, Mexican
Western European	2,659	French, Austrian, Belgian, English, Scottish, Dutch, Swiss, German, Irish
East Asian	2,512	Korean, Chinese, Japanese
Middle Eastern	645	Iranian, Jewish, Lebanese, Turkish
South Asian	621	Bangladeshian, Indian, Pakistani
Southeast Asian	457	Indonesian, Malaysian, Filipino, Thai, Vietnamese
Eastern European	381	Eastern European, Russian
African	352	Moroccan, East African, North African, South African, West African
Northern European	250	Scandinavian

Table 2.1: Regional coverage and country/area mapping for Dataset 1 (Ahn et al.). Source: Table S2 in the Supplementary Information of Ahn et al. [2011].

[1] Counts and regional breakdown follow Table S2 in the Supplementary Information of Ahn et al. [2011]. Minor off-by-one discrepancies between subtotals and the global total come from the original report.

Ingredient inventory. The standardized ingredient list comprises 381 unique ingredients. In the original study, each ingredient is mapped to a set of volatile compounds, enabling construction of an ingredient–compound bipartite network and its ingredient–ingredient projection (the "flavor network") [Ahn et al., 2011]. While the chemistry layer is not central to this chapter, we retain the same standardized ingredient inventory for comparability with prior work.

Ingredient categorization. Additionally, each ingredient is categorized into one of 14 food categories: *Dairy*, *Vegetable*, *Fruit*, *Meat*, *Fish/Seafood*, *Herb/Spice*, *Plant derivative*, *Animal product*, *Alcoholic beverage*, *Cereal*, *Legume*, *Nut/Seed*, *Sweetener*, *Beverage*. These categories are used for coarse-grained descriptive analyses and to compare broad functional roles across cuisines.

Provenance and access. Recipes originate from three public online sources: *Allrecipes* [AllRecipes], *Epicurious* [Epicurious], and the Korean portal *Menupan* [Menupan]. Regional labels were harmonized by Ahn et al. [2011] into the eleven groups listed in Table 2.1. For methodological details on scraping, deduplication, and regional aggregation, see the Supplementary Information of Ahn et al. [2011].

2.1.2 CulinaryDB Dataset

The second dataset employed in this thesis is *CulinaryDB*, a large-scale recipe repository curated by the CoSyLab research group. It provides comprehensive coverage of world cuisines across multiple continents and is specifically designed to support computational analyses of culinary patterns. In total, the dataset comprises 45,772 recipes, collected from four major online platforms: AllRecipes (16,177 recipes), Food Network (15,917), Epicurious (11,069), and TarlaDalal (2,609). Two vocabularies are provided: 929 atomic ingredients and 103 compound ingredients (1,032 in total), of which 695 actually appear in recipes. Each recipe is represented as a set of standardized ingredients, enabling cross-cultural comparisons and the construction of abstract representations such as ingredient-type networks.

Geographic coverage. Recipes are grouped into 26 cuisines spanning five continents. A cuisine may correspond to a single country (e.g., Italian, Japanese) or to a broader geo-cultural region (e.g., Central America, Scandinavia). Table 2.2 reports the mapping between cuisines and the countries or regions they represent, following the schema provided in the original release (see Supplementary Note 1 and Table S2 of Bagler [2017]).

Ingredient categorization. Each of the 695 standardized ingredients appearing in recipes is assigned to one of 20 food categories, or *ingredient types*, designed to reflect biological origin, culinary function, or degree of processing. The categories are: *Additive, Bakery, Beverages, Alcoholic Beverages, Cereals, Dairy, Essential Oils, Fish, Flowers, Fruit, Fungi, Herbs, Legumes, Maize, Meat, Nuts & Seeds, Plants, Seafood, Spices, Vegetables*. This classification enables a coarse-grained representation of recipes that abstracts from specific ingredient identities while retaining functional and cultural interpretability.

Cuisine	Countries/regions covered
Indian	India, Pakistan, Bangladesh
African	Nigeria, Ethiopia, DR Congo, Tanzania, South Africa, Kenya, Uganda, Sudan, Algeria, Morocco, Angola, Ghana, Mozambique, Madagascar, Côte d'Ivoire, Cameroon, Niger, Mali, Burkina Faso, Malawi, Zambia, Chad, Somalia, Senegal, Zimbabwe, Guinea, Rwanda, Benin, Burundi, Tunisia, South Sudan, Togo, Sierra Leone, Libya, Congo, Central African Republic, Liberia, Mauritania, Eritrea, Gambia, Botswana, Namibia, Gabon, Lesotho, Guinea-Bissau, Equatorial Guinea, Mauritius, Eswatini, Djibouti, Comoros, Cabo Verde, São Tomé & Príncipe, Seychelles
Chinese	China, Taiwan
Thai	Thailand
South East Asian	Cambodia, Laos, Vietnam, Philippines, Brunei, Malaysia, Singapore, Indonesia, Timor-Leste
Middle Eastern	Yemen, Oman, United Arab Emirates, Kuwait, Saudi Arabia, Jordan, Israel, Lebanon, Syria, Iraq, Iran, Türkiye, Armenia, Egypt, Qatar, Bahrain
Korean	South Korea, North Korea
Japanese	Japan
Australasian	Australia, New Zealand
Mexican	Mexico
Caribbean	Cuba, Jamaica, Haiti, Dominican Republic, Puerto Rico, Trinidad and Tobago
South American	Argentina, Bolivia, Brazil, Chile, Colombia, Ecuador, Guyana, Paraguay, Peru, Suriname, Uruguay, Venezuela
Central European	Germany, Austria, Switzerland
Greek	Greece
French	France
Misc. Portuguese	Portugal
Italian	Italy
British Isles	England, Scotland, Wales, Ireland, Isle of Man, Inner and Outer Hebrides, Northern Isles
Scandinavian	Norway, Sweden, Finland, Åland Islands, Svalbard
East European	Poland, Latvia, Lithuania, Belarus, Ukraine, Moldova, Romania, Hungary, Slovakia, Czechia, Bulgaria, Russia
Spanish	Spain
American	United States of America
Canadian	Canada

Table 2.2: Cuisine–region mapping in CulinaryDB. A cuisine may correspond to a single nation-state or to a broader geo-cultural area.

Provenance and access. The dataset was assembled by CoSyLab through systematic collection from the four online sources listed above. Further details on preprocessing, standardization, and classification can be found in the official dataset release Bagler [2017].

Complementarity of the two datasets. The two datasets play complementary roles in this thesis. The Flavor Network (Dataset 1) is uniquely valuable because it augments recipes with *chemical information*, linking ingredients to the volatile compounds that determine flavor. This layer enables us to revisit the food pairing hypothesis in a principled way by comparing chemical similarity with empirical co-occurrence. CulinaryDB (Dataset 2), by contrast, offers broader and more *balanced cultural coverage*, with recipes grouped into 26 cuisines spanning five continents. Its standardized taxonomy of ingredients and 20 functional categories makes it especially suited for cross-cultural analyses of recipe structure and for constructing compact, comparable network representations of cuisines. A side-by-side comparison of their main properties is reported in Table 2.3.

	Dataset 1: Flavor Network (Ahn et al.)	**Dataset 2: CulinaryDB**
Recipes	56,498	45,772
Ingredients	381 standardized ingredients	929 atomic + 103 compound (1,032 total); 695 appear in recipes
Compounds	1,021 volatile flavor compounds linked to ingredients	Not included
Cuisines / Regions	11 regional cuisine groups (Table 2.1)	26 cuisines spanning 5 continents (Table 2.2)
Sources	Allrecipes, Epicurious, Menupan	Allrecipes, Food Network, Epicurious, TarlaDalal
Ingredient categories	**14** food groups (e.g., Dairy, Vegetable, Fruit, Meat, Fish/Seafood, Herb/Spice, ...)	**20** culinary categories (e.g., Vegetables, Spices, Dairy, Meat, ...)
Focus	Ingredient–compound associations; flavor-sharing hypothesis	Cross-cultural recipe structure; ingredient-type networks

Table 2.3: Comparison of the two recipe datasets used in this thesis.

2.1.3 Comparing Ingredient Categorization

How do the two datasets "think" about ingredients? To answer this, we first lined up their vocabularies and then compared how each side labels the same items. The alignment com-

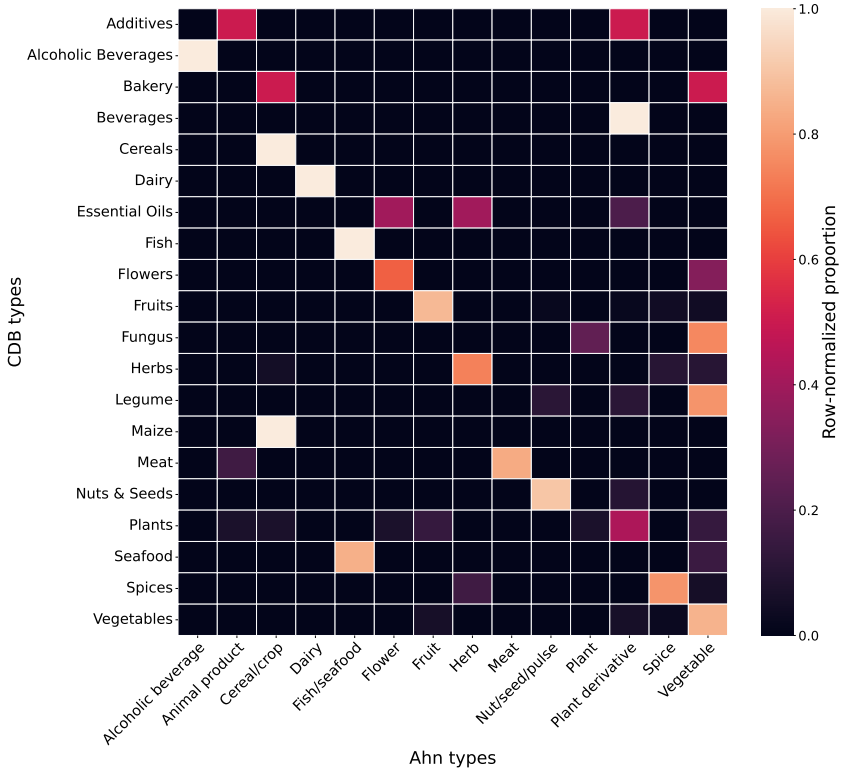


Figure 2.1: Row–normalized contingency between CulinaryDB types (rows) and Ahn types (columns). For each CulinaryDB type, color gives the share of aligned ingredients that Ahn assigns to each of its types. Concentrated rows indicate close agreement; broader rows reveal systematic splits or merges across the two schemes.

bines exact matches, light normalization of names, and a final pass to resolve tricky cases. In total, *233* ingredients are shared by both datasets; beyond this overlap, CulinaryDB contributes *361* additional ingredients and the Ahn's Dataset adds *140* more. These aligned pairs let us quantify where the two categorizations agree and where they differ.

From the 233 shared items we built a table whose rows are CulinaryDB's 20 types and whose columns are Ahn's 14 types. Each cell counts how many shared ingredients fall into that row–column combination. We then row–normalized the table so each CulinaryDB row sums to one, yielding the share of a CulinaryDB type that Ahn assigns to each of its types (Figure 2.1). When most of a row's mass is concentrated in a single column, the two schemes largely agree for that CulinaryDB type; when the mass is spread across columns, the schemes diverge. If the same Ahn column dominates several different rows, this signals that Ahn is using a broader category that absorbs multiple, more specific CulinaryDB types.

Several families show near–complete agreement: *Alcoholic Beverages*, *Dairy*, *Meat*, *Spices*, and *Fruits*. In these groups, the vast majority of shared ingredients assigned to a CulinaryDB type receive the corresponding label in Ahn's taxonomy.

The figure also makes clear where CulinaryDB is more specific. CulinaryDB distinguishes *Fish* and *Seafood*. Items in *Fish* correspond cleanly to Ahn's *Fish/Seafood*. Items in *Seafood* mostly map to *Fish/Seafood* as well, with a small share landing in *Vegetable*

on the Ahn side—largely algae, which CulinaryDB treats as *Seafood* but Ahn places with plant/vegetable–like categories. CulinaryDB also separates *Cereals* from *Maize*; in Ahn, both are absorbed by the broader *Cereal/Crop* label, producing concentrated mass in a single column.

Some categories exist on one side only, and the heatmap shows where they go. CulinaryDB has *Legumes* (largely mapping to *Vegetable* in Ahn) and *Fungi* (mostly *Vegetable* in Ahn, with a small fraction *Plant*). Non–alcoholic *Beverages* in CulinaryDB correspond entirely to Ahn's *Plant derivative*. *Additives* split roughly between Ahn's *Animal product* and *Plant derivative*. *Bakery* items fall primarily under Ahn's *Cereal/Crop* and, to a lesser extent, *Vegetable*. *Essential Oils* distribute across Ahn's *Flowers*, *Herb/Spice*, and *Plant derivative*. Conversely, Ahn includes two broad "container" classes—*Animal product* and *Plant derivative*—that have no single counterpart in CulinaryDB; items that Ahn assigns to these tend to be spread across *Dairy*, *Meat/Seafood*, *Additives*, *Plants*, *Flowers*, *Spices*, and *Essential Oils* in CulinaryDB. Even where both taxonomies define the same family, correspondence is not always perfect. For instance, some *Flowers* in CulinaryDB appear as *Vegetables* in Ahn. The *Plants* category is present in both taxonomies but aligns only weakly: ingredients labeled *Plants* in CulinaryDB are mostly assigned to *Plant derivative* in Ahn.

Overall, Ahn's scheme leans toward biological origin and compact "catch–all" classes, whereas CulinaryDB combines origin with culinary function and product form, introducing finer distinctions (e.g., *Herbs* vs. *Spices*; *Fish* vs. *Seafood*; standalone *Fungi*, *Maize*, *Essential Oils*, *Bakery*). The 233 aligned ingredients provide a robust backbone for cross–dataset comparisons, while the additional 361 (CulinaryDB) and 140 (Ahn) unique items illustrate complementary coverage rather than contradiction.

2.2 Graphs and Networks

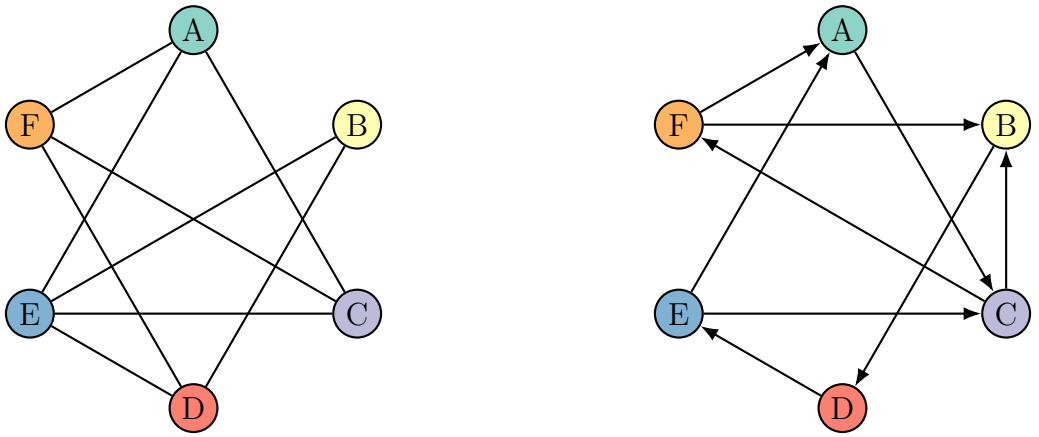
Graphs are among the most widely used mathematical structures for representing complex systems. Intuitively, a graph is a collection of entities together with the relations among them: individuals linked by social ties, proteins connected by interactions, or web pages joined by hyperlinks. This simple abstraction makes it possible to model heterogeneous systems within a common framework [Newman, 2018, Barabási, 2016].

2.2.1 Basic Concepts

Formally, a graph is defined as

$$G = (V, E),$$

where V is a finite set of *nodes* (or *vertices*) and $E \subseteq V \times V$ is a set of *edges* (or *links*) that connect pairs of nodes. Edges can be *undirected*, meaning that $(u, v) \in E$ implies $(v, u) \in E$, or *directed*, when orientation matters. Figure 2.2a shows simple toy examples of an undirected graph (a) and a directed graph (b). Edges may also carry weights, in which case each edge is associated with a real value $w(u, v)$ representing, for instance, the strength or frequency of the interaction. Graphs that contain no multiple edges between the same pair of nodes and no self-loops (edges of the form (v, v)) are called *simple graphs*.



(a) Undirected graph (b) Directed graph

Figure 2.2: Side-by-side examples of (a) an undirected graph and (b) a directed graph with colored nodes. Colors are consistent across panels for nodes A–F.

A convenient algebraic representation of a graph is given by its *adjacency matrix*. For a graph with $n = |V|$ nodes, the adjacency matrix is an $n \times n$ matrix A whose entry A_{ij} encodes the edge between nodes i and j. In an unweighted, undirected graph, $A_{ij} = 1$ if $(i, j) \in E$ and $A_{ij} = 0$ otherwise, with A symmetric by construction. In directed graphs, A is generally asymmetric, while in weighted graphs the entries A_{ij} store the corresponding edge weights. This matrix-based representation is fundamental for algorithmic and theoretical treatments of graphs, as it enables the application of linear algebraic tools to network analysis.

Degree. One of the most basic and informative properties of a node is its *degree*, defined as the number of edges incident upon it. In terms of the adjacency matrix, the degree of node i is

$$k_i = \sum_{j=1}^{n} A_{ij}.$$

In directed graphs, one distinguishes between the *in-degree*, given by the number of incoming edges, and the *out-degree*, given by the number of outgoing edges. In weighted graphs, the analogous quantity is the *strength* of a node, defined as the sum of the weights of its incident edges. Degree provides a simple measure of a node's connectivity within the network and serves as the foundation for more sophisticated measures of importance.

Subgraphs. A graph $H = (V_H, E_H)$ is a *subgraph* of $G = (V, E)$ if $V_H \subseteq V$ and $E_H \subseteq E \cap (V_H \times V_H)$. Intuitively, a subgraph is obtained by selecting a subset of nodes together with any subset of the edges between them. Special cases include *induced subgraphs*, where E_H contains all edges of E that connect pairs of nodes in V_H, and *spanning subgraphs*, where $V_H = V$ but only part of the edge set is retained.

2.2.2 Paths and Connectivity

A *path* in a graph is a sequence of nodes in which each consecutive pair is connected by an edge. The *length* of a path is the number of edges it contains, and the *shortest path* (or *geodesic*) between two nodes is the one with minimal length. The distribution of shortest path lengths influences how quickly information, resources, or influence can spread through a network [Latora et al., 2017].

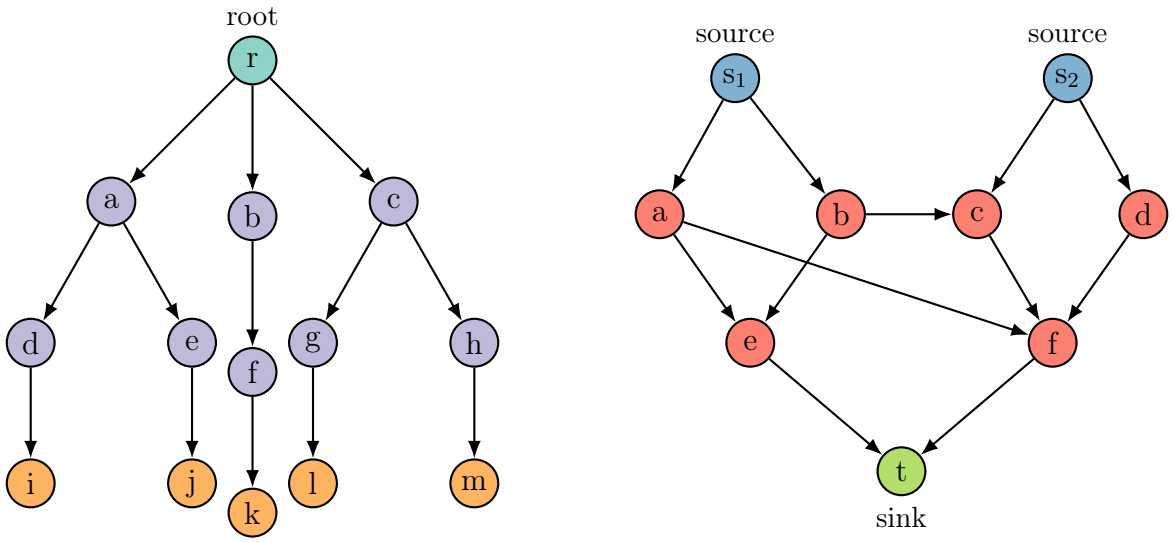
A special type of path is a *cycle*, defined as a closed path that starts and ends at the same node and in which every other node appears only once. Cycles are central in distinguishing different classes of graphs: acyclic graphs contain none, while graphs with many cycles typically exhibit higher redundancy and robustness.

In an undirected graph, two nodes u and v are said to be *connected* if there exists at least one path between them. In this case, the *distance* between u and v is defined as the length of the shortest connecting path. The notion of connected nodes allows us to define *connected components*, which are maximal subgraphs in which every pair of nodes is mutually reachable [Newman, 2018, Barabási, 2016]. A graph consisting of a single connected component is called *connected*; otherwise, it is *disconnected*. Identifying connected components is often a first step in analyzing empirical networks, since many structural properties are only meaningful within connected regions.

In directed graphs, the notion of connectivity must respect edge orientations. A *strongly connected component* is a maximal subgraph in which every pair of nodes u, v is reachable from one another by following the direction of edges. By contrast, a *weakly connected component* is defined by ignoring edge directions: two nodes belong to the same component if they are connected when all edges are treated as undirected. The distinction is crucial when directionality encodes asymmetric relations such as flows, influence, or causal links.

In many large real-world networks, one component is much larger than all others, containing the majority of nodes and edges. This *giant component* is the natural focus of

most analyses, since it captures the main connected backbone of the system, while smaller components often correspond to isolated or peripheral elements.



(a) Rooted tree (edges oriented away from the root).

(b) Directed acyclic graph (DAG) with multiple sources and a sink.

Figure 2.3: Contrasting hierarchical structures: (a) a rooted tree (a special DAG with a single root and unique parent for each non-root), and (b) a general DAG allowing multiple sources/sinks and merges without cycles.

Acyclic graphs and trees. Graphs without cycles are called *acyclic*. In the undirected case, an acyclic connected graph is a *tree*, while a collection of disjoint trees forms a *forest*. Trees are important because they provide the minimal connected structure linking all nodes, with exactly $n - 1$ edges for n nodes. Equivalently, a tree can be defined as a connected graph in which any two nodes are joined by a unique simple path [Diestel, 2017]. This alternative characterization emphasizes that trees exclude redundancy: the removal of any edge disconnects the graph, while the addition of any edge necessarily creates a cycle.

A common refinement is the notion of a *rooted tree*, where one node is distinguished as the *root* (Fig. 2.3 (a)). This designation introduces a natural hierarchy: edges can be interpreted as oriented away from the root toward its *children*, creating notions of parent–child relationships, levels (or depth), and subtrees. Rooted trees are particularly useful in contexts where hierarchical or genealogical structures are modeled, such as organizational charts, taxonomies, and evolutionary trees.

These properties make trees natural models for hierarchical or branching structures, such as family trees, organizational charts, or phylogenies.

In directed graphs, the absence of cycles defines a *directed acyclic graph* (DAG) (Fig. 2.3b). DAGs naturally encode systems with directionality and no feedback loops, such as task dependencies, genealogies, or citation networks. Unlike rooted trees, which enforce a single designated root and a strict parent–child structure, DAGs allow multiple sources and sinks and capture general partial orderings among nodes. Thus, every rooted tree is a DAG, but

not every DAG is a tree: trees represent the most constrained form of hierarchy, while DAGs describe broader causal or dependency structures.

Spanning trees. Given a connected undirected graph $G = (V, E)$, a *spanning tree* is a subgraph $T = (V, E_T)$ such that $E_T \subseteq E$, T is a tree, and T contains all nodes of G. In other words, a spanning tree preserves the full vertex set but retains only a subset of edges sufficient to ensure connectivity while avoiding cycles. Every connected graph admits at least one spanning tree, and in general multiple distinct spanning trees may exist. Note that every spanning tree is a spanning subgraph (cf. Section 2.2.1) that is acyclic and connected.

In a weighted graph, where edges are associated with weights $w : E \to \mathbb{R}$, one can introduce optimization criteria. The two most studied cases are:

- The *maximum spanning tree* (MaxST), defined as the spanning tree $T^* = (V, E^*)$ such that
$$\sum_{(u,v)\in E^*} w(u, v) \;\geq\; \sum_{(u,v)\in E'} w(u, v)$$
for any other spanning tree (V, E'). The MaxST is the spanning tree with the largest possible total weight.

- The *minimum spanning tree* (MST), defined analogously as the spanning tree that minimizes the total weight of its edges:
$$\sum_{(u,v)\in E^*} w(u, v) \;\leq\; \sum_{(u,v)\in E'} w(u, v)$$
for any other spanning tree (V, E').

Neither the MST nor the MaxST is necessarily unique, but all optimal solutions share the same total weight. In practice, uniqueness is guaranteed under a simple condition: if all edge weights are distinct, then the MST (or MaxST) is unique. This follows because greedy algorithms such as Kruskal's or Prim's [Kruskal, 1956, Prim, 1957] never encounter ties in choosing edges, so the resulting spanning tree is determined unambiguously. When edge weights are not distinct, multiple optimal trees may exist, although they all have the same total weight.

For a complete treatment of the existence, uniqueness, and necessary and sufficient conditions (including cut and cycle properties), see sources such as Diestel [2017] and the Princeton Algorithms course material on MSTs [Sedgewick and Wayne, 2025].

Notation. In much of the literature, the acronym MST refers to the *minimum spanning tree*. In this thesis, however, we are exclusively concerned with *maximum spanning trees*, and we adopt the acronym MST to denote them throughout.

2.2.3 Centrality

In social networks we do not only ask *who is connected to whom*, but also *who plays a pivotal role* in shaping how the network functions. Some individuals act as *hubs*, maintaining

ties with many others and ensuring that information or resources spread quickly. Others serve as *bridges*, linking otherwise separate communities and enabling communication across structural boundaries [Newman, 2018]. If a hub is removed, many actors lose immediate access to a large portion of the network; if a bridge disappears, entire groups may become isolated and interaction across communities severely limited.

Centrality measures formalize these intuitions by providing quantitative ways of identifying and ranking key actors. Different measures capture complementary notions of importance—ranging from local prominence to global accessibility, brokerage, recursive prestige, and structural vulnerability. For the purposes of this thesis, we group centrality measures into four broad categories: (i) local centralities, (ii) path-based centralities, (iii) recursive/spectral and random-walk centralities, and (iv) efficiency-based centralities.

1. Local Centralities These measures depend only on a node's immediate neighborhood, ignoring global structure. The prototypical example is *degree centrality*, which counts how many direct connections a node has. This is basically the degree of a node, defined in a different guise in 2.2.1. In weighted networks this generalizes to node strength, the sum of edge weights, defined as $s_i = \sum_{j=1}^{n} w_{ij}$. Local centrality is computationally cheap and interpretable, but does not reflect a node's reach or influence beyond its neighbors.

2. Path-Based Centralities Path-based measures emphasize the strategic positions nodes occupy in facilitating or obstructing flows across the network. They are especially relevant in contexts such as communication or transportation networks, where efficiency depends on shortest paths, though their reliance on path enumeration makes them computationally intensive in large graphs [Freeman, 1979, Borgatti, 2005]:

- **Closeness centrality.** A measure of accessibility, defined as the inverse of the average shortest path length from node i to all others:

$$C_C(i) = \frac{n-1}{\sum_{j \neq i} d(i,j)}, \tag{2.1}$$

  where $d(i,j)$ is the shortest path distance between i and j [Freeman, 1979]. Nodes with high closeness are, on average, closer to all others in the network. For disconnected graphs, variants such as harmonic closeness are often preferred.

- **Betweenness centrality.** A measure of brokerage, based on how often a node lies on shortest paths between other pairs [Freeman, 1979]:

$$C_B(i) = \sum_{s \neq i \neq t} \frac{\sigma_{st}(i)}{\sigma_{st}}, \tag{2.2}$$

  where σ_{st} is the number of shortest paths from s to t, and $\sigma_{st}(i)$ is the number of those paths that pass through i. Nodes with high betweenness often control flows of information or resources.

3. Recursive / Spectral and Random-Walk Centralities Recursive and random-walk measures are especially powerful in contexts where influence is not merely local but propagates through chains of connections, as in web search, citation analysis, or reputation systems. Their reliance on linear algebra and stochastic processes makes them more sophisticated mathematically, but also well-suited to capture global patterns of prestige or authority.

- **Eigenvector centrality.** A recursive measure of prestige, assigning high scores to nodes connected to other highly central nodes [Newman, 2018]. It is defined as the leading eigenvector of the adjacency matrix:

$$C_E(i) = \frac{1}{\lambda} \sum_{j=1}^{n} A_{ij} C_E(j),$$

  where λ is the largest eigenvalue of A.

- **PageRank.** A widely used variant of eigenvector centrality, originally introduced for ranking web pages [Brin and Page, 1998]. It is based on the stationary distribution of a random walk with restart (or damping factor), which ensures convergence and robustness to dangling nodes. PageRank captures both popularity and the quality of incoming links, making it well suited to directed networks.

4. Information Centrality. Another perspective on centrality builds on the concept of *network efficiency*, which quantifies how effectively a graph supports communication among its nodes. The efficiency between two nodes i and j is defined as the inverse of their shortest path distance,

$$\epsilon_{ij} = \frac{1}{d(i,j)},$$

with the convention $\epsilon_{ij} = 0$ if i and j are disconnected. The *global efficiency* of a graph G is then the average efficiency over all unordered pairs of distinct nodes:

$$E[G] \;=\; \frac{1}{n(n-1)} \sum_{i \neq j} \epsilon_{ij} \tag{2.3}$$

This formulation captures how efficiently information or resources can be exchanged across the entire network, balancing both path lengths and overall connectivity [Latora and Marchiori, 2001].

Building on this notion, Latora and Marchiori [2007] introduced *information centrality*, which evaluates the importance of a node i by the relative drop in global efficiency when i (and its incident edges) are removed:

$$C_I(i) \;=\; \frac{E[G] - E[G_{\setminus i}]}{E[G]} \tag{2.4}$$

Information centrality thus lies in $[0, 1]$, with higher values indicating that the removal of i severely degrades the network's efficiency. Conceptually, it combines features of closeness (dependence on distances) and betweenness (sensitivity to intermediate positions), but frames them in terms of the node's contribution to the network's global communicability.

This approach belongs to the broader family of *delta centralities*, where the importance of a node (or set of nodes) is defined by the change induced in a global graph functional—such as efficiency, connectivity, or robustness—when the node is deleted. Delta-based measures are particularly informative in empirical applications, as they reveal structural vulnerabilities and the extent to which individual nodes sustain the network's ability to function.

Taken together, these measures capture complementary dimensions of node importance: hubs (degree/strength), accessibility (closeness), brokerage (betweenness), recursive prestige (eigenvector/PageRank), and structural vulnerability (information centrality).

2.2.4 Assortativity

Assortativity, or mixing by similarity, quantifies whether connections in a network occur more (or less) frequently between nodes that share similar properties than would be expected at random [Newman, 2003, 2018]. Two common cases are (i) *degree assortativity*, which measures correlation between the degrees at the two ends of an edge, and (ii) *attribute assortativity* for categorical or real-valued node features (e.g., types, labels).

Degree assortativity. The standard measure of degree assortativity is based on the Pearson correlation between the degrees at the two ends of an edge [Newman, 2002]. For an undirected graph with $m = |E|$ edges, let j_e and k_e denote the degrees of the two endpoints of edge e. The assortativity coefficient is defined as

$$r \;=\; \frac{\dfrac{1}{m}\displaystyle\sum_{e\in E} j_e k_e \;-\; \left[\dfrac{1}{m}\displaystyle\sum_{e\in E} \tfrac{1}{2}(j_e + k_e)\right]^2}{\dfrac{1}{m}\displaystyle\sum_{e\in E} \tfrac{1}{2}(j_e^2 + k_e^2) \;-\; \left[\dfrac{1}{m}\displaystyle\sum_{e\in E} \tfrac{1}{2}(j_e + k_e)\right]^2} \tag{2.5}$$

The numerator measures how much the average product $j_e k_e$ exceeds the value expected if the degrees at the two ends of an edge were statistically independent. The denominator normalizes by the variance of endpoint degrees, ensuring that r is confined to $[-1, 1]$. Thus, $r > 0$ indicates that high-degree nodes tend to attach to other high-degree nodes (assortative mixing), $r < 0$ indicates the opposite (disassortative mixing, as in hub–periphery structures like star graphs), and $r = 0$ corresponds to no correlation. Empirically, social networks are often assortative, whereas biological and technological networks tend to be disassortative [Newman, 2002, 2018].

Attribute assortativity (categorical). Degree assortativity is in fact a special case of attribute assortativity, where the attribute is the degree itself. More generally, when nodes carry categorical attributes (e.g., types or labels), assortativity is measured using the *mixing matrix* e_{ab}, where e_{ab} is the fraction of edges that connect a node of type a to a node of type b. The row and column marginals are given by $a_a = \sum_b e_{ab}$ and $b_b = \sum_a e_{ab}$ (for undirected graphs e is symmetric, so $a = b$). The assortativity coefficient is then

$$r \;=\; \frac{\mathrm{Tr}(e) - \sum_a a_a b_a}{1 - \sum_a a_a b_a}\,. \tag{2.6}$$

Here, $\mathrm{Tr}(e)$ represents the fraction of same-type edges, while $\sum_a a_a b_a$ is the expected fraction of such edges if endpoints were matched at random given the marginals. The coefficient therefore quantifies how much the observed within-type connectivity exceeds random expectation, normalized to lie in $[-1, 1]$. A value of $r = 1$ means all edges fall within types, $r = 0$ corresponds to random mixing, and $r < 0$ indicates that edges preferentially connect nodes of different types [Newman, 2003].

Directed and weighted variants. In directed graphs, one can correlate in/out degrees at source and target to obtain four coefficients (in–in, in–out, out–in, out–out) [Newman, 2018]. For weighted networks, a natural extension replaces edge counts by edge weights in the formulas, thereby correlating degrees/attributes with respect to connection strength.

Assortativity summarizes mesoscopic organization in a single number: positive values indicate a tendency to form homogeneous clusters (by degree or attribute), while negative values indicate "cross-type" or hub–periphery wiring. It complements community structure: modularity detects dense groups, whereas assortativity tests whether links preferentially join similar nodes irrespective of density patterns.

2.2.5 Communities

While assortativity characterizes edge-level mixing (who tends to connect to whom across single links), many networks also exhibit mesoscopic organization into *communities* (also called *modules* or *clusters*)—cohesive groups of nodes with denser internal connections than external ones [Fortunato, 2010, Newman, 2018]. In a social network, communities may correspond to circles of friends or professional groups; in culinary graphs, they can represent sets of ingredient types that frequently co-occur in recipes. Figure 2.4 illustrates this idea with a toy network displaying a simple two–community structure. Detecting community structure is important because it reveals mesoscopic patterns of organization that are not apparent from local measures such as degree or centrality.

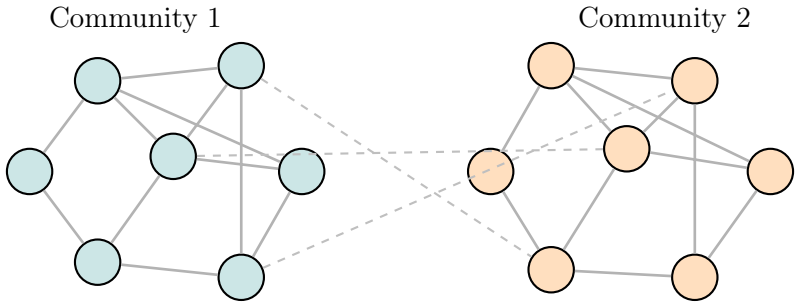


Figure 2.4: Toy network with two communities: dense intra-community connectivity (solid gray) and sparse inter-community links (dashed). Nodes are colored according to their community membership (light blue and orange).

Formally, community detection seeks a partition of the node set into subsets that optimize some criterion of internal density versus external sparsity. One widely used approach is the

maximization of *modularity* [Newman and Girvan, 2004, Newman, 2006], a quality function that compares the observed density of edges within groups to the expected density under a suitable null model. Given a partition of the network into communities, modularity is defined as

$$Q = \frac{1}{2m} \sum_{i,j} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j), \tag{2.7}$$

where A_{ij} is the adjacency matrix, k_i and k_j are the degrees of nodes i and j, m is the total number of edges, c_i is the community assignment of node i, and $\delta(c_i, c_j)$ equals 1 if i and j belong to the same community and 0 otherwise. A high value of Q indicates that many edges fall within communities compared to a random graph with the same degree sequence [Newman, 2006]. While modularity has become a standard objective for community detection, it is not without limitations, such as the *resolution limit*, whereby small communities may remain undetected [Fortunato and Barthélemy, 2007].

Community detection. A wide variety of algorithms have been developed to uncover community structure in empirical networks. Greedy optimization methods, most notably the Louvain and Leiden algorithms [Blondel et al., 2008, Traag et al., 2019], are widely used because they efficiently maximize modularity. Alternative approaches include divisive methods such as the Girvan–Newman algorithm, which iteratively removes edges with the highest betweenness centrality [Newman and Girvan, 2004], and spectral methods, which exploit the eigenvectors of graph Laplacians to guide partitions [Newman, 2006]. Each approach comes with trade-offs between accuracy, resolution, and computational cost, but collectively they provide a rich toolkit for extracting mesoscopic patterns of organization from network data.

The presence and composition of communities provide insights into the functional organization of networks [Fortunato, 2010]. For instance, in a culinary network, communities may reveal sets of ingredient types that act as cohesive building blocks of a cuisine, while in social networks they reflect the formation of tightly knit groups.

2.2.6 Network Motifs

Beyond global descriptors such as degree distributions and mesoscopic structures such as community partitions, networks often exhibit characteristic *motifs*: small, recurring subgraphs that appear significantly more often than expected under a suitable null model [Milo et al., 2002]. Motifs capture local patterns of interaction that may underlie functional or cultural processes in a system. For instance, in biological networks they can represent regulatory circuits, while in social networks they may encode triadic closure or hierarchical influence.

Network motifs are formally defined as small connected subgraphs that occur in a network at a frequency significantly different from what would be expected in a suitably randomized null model of the same network [Milo et al., 2002].

Let $G = (V, E)$ be a graph. For a fixed size k, let $G[W]$ denote the subgraph *induced* by a k-node vertex set $W \subset V$. Two k-node graphs H_1, H_2 are *isomorphic* if there exists a bijection $\phi : V(H_1) \to V(H_2)$ that preserves adjacency. We label $[H]$ the *isomorphism class*

of a k-node pattern H. We count occurrences up to isomorphism:

$$N_G([H]) \;=\; \big|\{\, W \subset V :\; |W| = k,\; G[W] \cong H \text{ and } G[W] \text{ is connected} \,\}\big|.$$

Thus, a *motif of size k* is an isomorphism class $[H]$ of connected induced subgraphs that is statistically overrepresented compared to an appropriate null model.

(Disconnected induced subgraphs are typically excluded from motif enrichment; they are useful in triad censuses but are not treated as motifs.)

Null models. To assess whether $[H]$ is a motif, we compare $N_G([H])$ against an ensemble $\mathcal{G}_0$ of randomized graphs that preserve salient low-level constraints of G. Common choices include:

- **Degree-sequence preserving (configuration model, undirected):** randomize via degree-preserving edge swaps so that the degree of every node matches that in G.
- **Directed in/out-degree preserving:** randomize orientations via double-edge swaps while preserving each node's in- and out-degree (optionally also the number of mutual dyads).
- **Bipartite degree-preserving:** for two-mode data, randomize the biadjacency M by swaps that keep row/column sums (recipe sizes and ingredient frequencies) fixed.
- **Sparser baselines (for sensitivity checks):** Erdős–Rényi $G(n,p)$ (matching edge density) or models preserving degree *distribution* but not the exact sequence.

For a chosen $\mathcal{G}_0$, the statistical significance is quantified by a z-score

$$z([H]) \;=\; \frac{N_G([H]) - \big\langle N_{\mathcal{G}_0}([H]) \big\rangle}{\sigma_{\mathcal{G}_0}([H])},$$

where $\langle\cdot\rangle$ and $\sigma(\cdot)$ denote the mean and standard deviation over the null ensemble.

A pattern $[H]$ is deemed a *motif* if $z([H])$ is sufficiently large, indicating overrepresentation relative to random expectation. Positive z indicates enrichment (suggesting a non-random structural role), while negative z indicates suppression (also informative, e.g., avoidance of certain cycles). The vector of z-scores across all k-node isomorphism classes constitutes a *motif profile*, a local structural "fingerprint" complementary to global or community-level statistics.

Triads. The simplest and most widely studied class of motifs are *triads*, i.e., subgraphs on $k = 3$ vertices. In undirected graphs, there are only two non-isomorphic *connected* triads: the open triad (a path of length two) and the closed triad (a triangle). Disconnected configurations (an isolated edge plus a singleton, or three isolated nodes) exist but are generally excluded from motif analysis, as motifs are typically restricted to connected subgraphs.

In directed graphs, by contrast, the situation is richer. On three labeled vertices, each unordered pair can be in one of four dyadic states: no tie, a single arc in one direction, a single edge in the opposite direction, or a mutual (bidirectional) tie. Since there are three unordered pairs, this yields $4^3 = 64$ possible labeled directed graphs (with no loops

and at most one arc per ordered pair). Up to isomorphism—disregarding vertex labels and retaining only structure—these collapse into *16* distinct classes, known collectively as the *triad census* [Holland and Leinhardt, 1976]. Each class is labeled by a canonical code that summarizes its dyad census (M, A, N): the counts of mutual ($\leftrightarrow$), asymmetric ($\rightarrow$), and null ($\varnothing$) dyads. Additional suffixes (D/U/C for "down/up/cyclic" orientations) distinguish orientations relative to a mutual dyad.

Figure 2.5 illustrates the two connected undirected triads (left) and a representative subset of the directed census types (right), ranging from simple chains and stars to cycles and reciprocated structures. Together, these provide the building blocks for higher-order patterns of interaction.

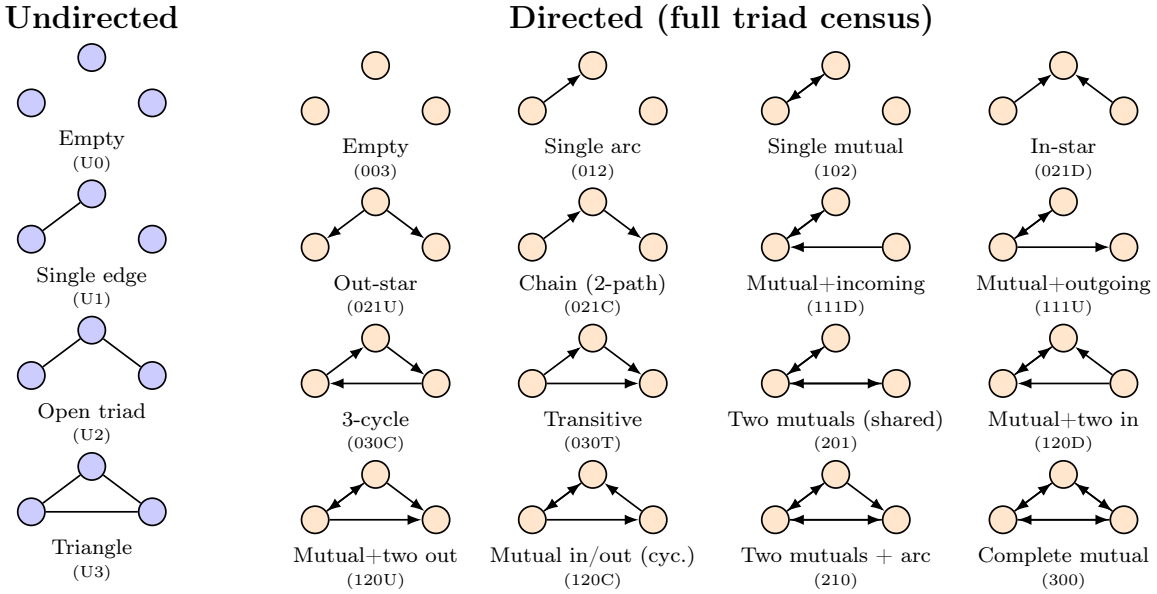


Figure 2.5: Canonical triads. Left: four undirected triads (0, 1, 2, and 3 edges). the full set of 16 non-isomorphic directed triads, classified according to the *triad census* [Holland and Leinhardt, 1976]. Each directed triad is labeled with its canonical census code (e.g., 021D, 030T, 111U) and its conventional name when applicable (e.g., in-star, out-star, chain). Nodes are color-coded: blue for undirected, orange for directed.

Triads play a fundamental role in network analysis because they capture the most elementary patterns of interaction beyond dyads. In social networks, for instance, the prevalence of triangles reflects the principle of triadic closure—"a friend of my friend is also my friend." In information networks, directed triads such as the *feed-forward loop* serve as building blocks of regulatory systems.

Let n_Δ and $n_\wedge$ denote the numbers of triangles and open triads in G, respectively. The relative abundance of closed triads is often quantified by the *clustering coefficient*,

$$C = \frac{3n_\Delta}{n_\wedge + 3n_\Delta},$$

where the factor of 3 accounts for the number of closed triplets per triangle. Thus, the clustering coefficient can be interpreted as the probability that an open triad is "completed" into a closed triad.

In this thesis, motif analysis is applied specifically to the chef influence network, where it helps uncover recurrent micro-structures of cultural transmission. Triads highlight hierarchical and cyclical relations, distinguishing, for example, linear chains of mentorship from patterns of mutual inspiration. Thus, motifs complement the global and mesoscopic measures introduced earlier by focusing on the elementary building blocks of influence among chefs.

2.2.7 Bipartite graphs

Many systems of interest do not consist of a single homogeneous class of entities, but rather of two distinct types that interact only across categories. Examples include people attending events, authors writing papers, users purchasing items, or—in the culinary domain—recipes linked to the ingredients they contain, and ingredients linked to the volatile compounds they share. Such data are naturally represented as *bipartite* (or two–mode) graphs [Newman, 2018].

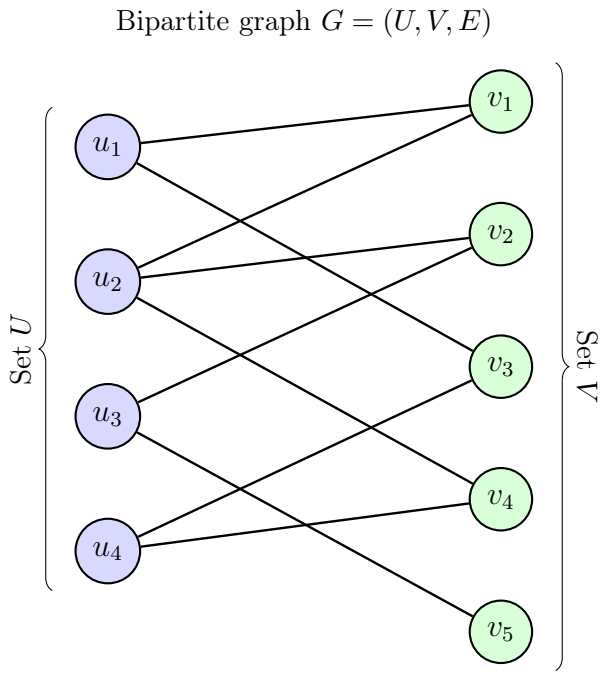


Figure 2.6: A toy bipartite (two–mode) network with node sets U and V. Edges only connect nodes across sets.

Definition and matrices. Formally, a bipartite graph is a triple

$$G = (U, V, E), \quad U \cap V = \varnothing, \quad E \subseteq U \times V,$$

where U and V are two disjoint node sets (often called "modes"), and edges E connect nodes only across the sets (no U–U or V–V edges are allowed). An illustrative toy example is shown in Figure 2.6. Let $|U| = m$ and $|V| = n$. The structure of G is conveniently encoded

in the *biadjacency matrix* $M \in \{0,1\}^{m \times n}$ with entries

$$M_{\alpha i} = \begin{cases} 1 & \text{if } (\alpha, i) \in E, \\ 0 & \text{otherwise.} \end{cases}$$

Thus rows correspond to nodes in U and columns to nodes in V. An alternative representation is the *block adjacency matrix* of dimension $(m+n) \times (m+n)$,

$$A = \begin{pmatrix} 0 & M \\ M^\top & 0 \end{pmatrix},$$

which places the biadjacency matrix in the off–diagonal blocks and zeros elsewhere. This makes bipartite graphs a natural generalization of the familiar adjacency matrix representation of one–mode graphs. Weighted bipartite graphs are easily accommodated by letting $M_{\alpha i} \in \mathbb{R}_{\geq 0}$ store the strength of interaction (e.g., counts, ratings, frequencies) rather than a binary indicator.

Degrees and paths. As in one–mode graphs, degrees are a fundamental local descriptor. For nodes in U and V, degrees are simply the row and column sums of M:

$$d_\alpha = \sum_{i=1}^{n} M_{\alpha i} \quad (\alpha \in U), \qquad u_i = \sum_{\alpha=1}^{m} M_{\alpha i} \quad (i \in V).$$

Intuitively, d_α counts how many neighbors node α in U has in V, while u_i counts how many neighbors node i in V has in U. In weighted bipartite graphs, the corresponding quantity is called the *strength*, i.e. the sum of edge weights incident on a node. Paths in bipartite graphs necessarily alternate between the two node sets, and notions of connectivity and components extend directly to this setting. In practice, these concepts are most often implemented by applying standard algorithms for connected components to the block adjacency matrix A. Analyses that involve distances or shortest paths are usually restricted to the largest (weakly) connected component.

One–mode projections. While bipartite graphs faithfully represent the dual structure of the data, it is often useful to summarize them by focusing on only one of the two modes. This leads to the construction of *one–mode projections*, where nodes correspond to entities of a single type, and edges encode their co–participation in the other mode. Formally, these projections can be written as

$$W^{(V)} \;=\; M^\top M \;-\; \operatorname{diag}(M^\top M), \qquad W^{(U)} \;=\; M M^\top \;-\; \operatorname{diag}(M M^\top),$$

so that $W^{(V)}_{ij} = \sum_\alpha M_{\alpha i} M_{\alpha j}$ counts how many U–nodes connect i and j, and similarly for $W^{(U)}_{\alpha\beta}$. In words, two V–nodes (e.g., ingredients) are connected if they co–occur in at least one U–node (e.g., recipe), with the edge weight reflecting the number of such co–occurrences.

A practical concern when working with one–mode projections is *density inflation*: because popular nodes connect to many others, the projection of a dense bipartite graph can be nearly

complete. To counter this, backbone extraction techniques (such as the disparity filter of Serrano et al. 2009) are used to prune links and retain only those that are statistically significant relative to a null model.

Bipartite graphs and their projections thus provide a flexible algebraic framework for moving between different levels of description: from recipes to ingredients, from ingredients to compounds, and from detailed entities to broad categories. In the next section we specialize these constructions to the culinary domain, outlining the main network representations of cuisines that will be employed throughout this thesis.

2.3 Network representations of cuisines

Recipes can be mapped into graphs in multiple, complementary ways depending on whether the focus is on individual ingredients, their chemical composition, or their functional roles in the recipe. In this thesis we make systematic use of three principal representations, each derived from an underlying bipartite graph structure.

2.3.1 Recipe–ingredient bipartite networks

The most direct way to represent a recipe dataset as a graph is through a *bipartite network* connecting recipes to the ingredients they contain. Let $\mathcal{R}$ denote the set of recipes and $\mathcal{I}$ the set of ingredients, with $\mathcal{R} \cap \mathcal{I} = \varnothing$. The bipartite recipe–ingredient graph is then defined as

$$G = (\mathcal{R}, \mathcal{I}, E), \qquad E \subseteq \mathcal{R} \times \mathcal{I},$$

where $(r, i) \in E$ indicates that ingredient $i \in \mathcal{I}$ appears in recipe $r \in \mathcal{R}$. By construction, edges run only across sets, never between two recipes or two ingredients.

Incidence matrix. The structure of G is conveniently encoded by the *biadjacency matrix* $M \in \{0,1\}^{|R|\times|I|}$, defined by

$$M_{ri} = \begin{cases} 1 & \text{if ingredient } i \text{ is used in recipe } r, \\ 0 & \text{otherwise.} \end{cases}$$

Each row corresponds to a recipe, each column to an ingredient, and each entry records presence or absence. In weighted versions, entries may store the frequency of occurrence of an ingredient (e.g., in corpora with multiple recipe variants) or quantities such as normalized amounts.

Degrees. Simple row and column sums of M provide informative statistics:

$$d_r = \sum_{i \in I} M_{ri}, \qquad u_i = \sum_{r \in R} M_{ri}.$$

Here d_r is the *degree* of recipe r, i.e. the number of ingredients it contains (its *recipe size*); u_i is the degree of ingredient i, i.e. the number of recipes in which it appears (its *ingredient*

frequency). These two distributions are among the most basic descriptors of a cuisine: the recipe size distribution reflects conventions about dish complexity, while ingredient frequencies capture cultural staples and peripheral ingredients.

The bipartite representation preserves the full structure of the dataset without aggregation. It forms the basis for almost all subsequent analyses: ingredient co–occurrence networks arise as one–mode projections of G (Section 2.3.2), while recommendation and similarity measures rely on weighting and reweighting of the incidence matrix. This representation is also particularly suitable for null-model generation, since shuffling edges in a bipartite structure allows one to preserve recipe sizes and ingredient frequencies independently.

2.3.2 Ingredient co–occurrence networks

A widely used representation in computational gastronomy is the *ingredient co–occurrence network*, where nodes are ingredients and edges capture their tendency to appear together in recipes [Teng et al., 2012, Jain and Bagler, 2015]. This representation is obtained as the one–mode projection of the recipe–ingredient bipartite graph described in Section 2.3.1.

Projection. Starting from the incidence matrix $M \in \{0,1\}^{|\mathcal{R}|\times|\mathcal{I}|}$, the co–occurrence matrix is defined as

$$W = M^\top M - \mathrm{diag}(M^\top M).$$

Here W_{ij} counts the number of recipes in which ingredients i and j co–occur, while the subtraction removes diagonal entries (which would otherwise record trivial self–counts). The result is a weighted, undirected graph on the set of ingredients I, with adjacency matrix W.

Interpretation. The degree of an ingredient in this projection,

$$k_i = \sum_{j\in I} W_{ij},$$

reflects how many different ingredients it co–occurs with, weighted by frequency. Highly connected nodes correspond to versatile, widely used ingredients (e.g., salt or onion), while low–degree nodes are specialized or rare. The strength of a link W_{ij} represents the empirical association between two ingredients in the cuisine: stronger edges indicate frequent co–usage.

From co–occurrence to similarity networks. While raw co–occurrence counts reflect empirical frequency, they are inherently biased toward popular ingredients that appear in many recipes. To mitigate this effect, one often normalizes edge weights to obtain a *similarity network*. Here, the goal is to capture the strength of association between two ingredients relative to their overall usage. For example, the cosine similarity between co–occurrence profiles is

$$B^{\cos}_{ij} = \frac{W_{ij}}{\sqrt{u_i\, u_j}},$$

where

$$u_i = \sum_{j\in\mathcal{I}} W_{ij}$$

denotes the total frequency of ingredient i (its degree or strength in the co–occurrence network). Other measures, such as the Jaccard index,

$$B_{ij}^{\text{jac}} = \frac{W_{ij}}{u_i + u_j - W_{ij}},$$

or pointwise mutual information,

$$\text{PMI}_{ij} = \log \frac{W_{ij} \cdot |\mathcal{R}|}{u_i \, u_j},$$

emphasize the *specificity* of co–usage rather than its absolute frequency.

Similarity networks are thus derived from co–occurrence data but highlight different aspects of culinary structure: they privilege distinctive associations over sheer popularity and are especially useful when comparing across cuisines, where baseline frequencies of ingredients may differ substantially. In this thesis, both raw co–occurrence and normalized similarity networks are employed, depending on whether the focus is on common structural backbones or on context–sensitive affinities between ingredients.

Applications. Ingredient co–occurrence networks form the backbone of many studies of culinary data. They provide a compact representation of a cuisine's combinatorial space and have been used to investigate questions such as the *food pairing hypothesis* [Ahn et al., 2011], cross–cultural similarities and differences between cuisines, and the identification of core versus peripheral ingredients. In this thesis, they will serve as a central tool for quantifying structural patterns of recipe composition across world cuisines.

2.3.3 The flavor network

A complementary representation to co–occurrence graphs models the chemistry underlying flavor. Each ingredient is annotated with the set of volatile molecules ("flavor compounds") detected in it, giving rise to a bipartite (two–mode) network between ingredients and compounds [Ahn et al., 2011]. Formally, let $\mathcal{I}$ be the set of ingredients, $\mathcal{C}$ the set of compounds, and

$$M^{\text{flavor}} \in \{0,1\}^{|\mathcal{C}| \times |\mathcal{I}|}$$

the *biadjacency* (incidence) matrix with entries

$$M_{\alpha i}^{\text{flavor}} = \begin{cases} 1, & \text{if compound } \alpha \in \mathcal{C} \text{ is present in ingredient } i \in \mathcal{I}, \\ 0, & \text{otherwise.} \end{cases}$$

The (compound) degree of an ingredient i is $u_i = \sum_{\alpha \in \mathcal{C}} M_{\alpha i}^{\text{flavor}}$, i.e., the number of compounds detected in i, while the (ingredient) degree of a compound α is $d_\alpha = \sum_{i \in \mathcal{I}} M_{\alpha i}^{\text{flavor}}$, i.e., how ubiquitous that molecule is across ingredients.

Projection to the flavor network. Projecting the bipartite graph onto the ingredient layer yields the *flavor network*, a weighted, undirected graph whose nodes are ingredients and whose edge weights count shared compounds [Ahn et al., 2011]. Let $W^f \in \mathbb{N}^{|\mathcal{I}|\times|\mathcal{I}|}$ with

$$W_{ij}^f = \sum_{\alpha\in\mathcal{C}} M_{\alpha i}^{\text{flavor}}\, M_{\alpha j}^{\text{flavor}} = \big|\, C_i \cap C_j \,\big|, \qquad i \neq j,$$

where $C_i = \{\alpha \in \mathcal{C} : M_{\alpha i}^{\text{flavor}} = 1\}$ is the compound set of ingredient i. The node strength $s_i = \sum_{j\neq i} W_{ij}^f$ quantifies the total compound overlap that ingredient i shares with the rest of the pantry.

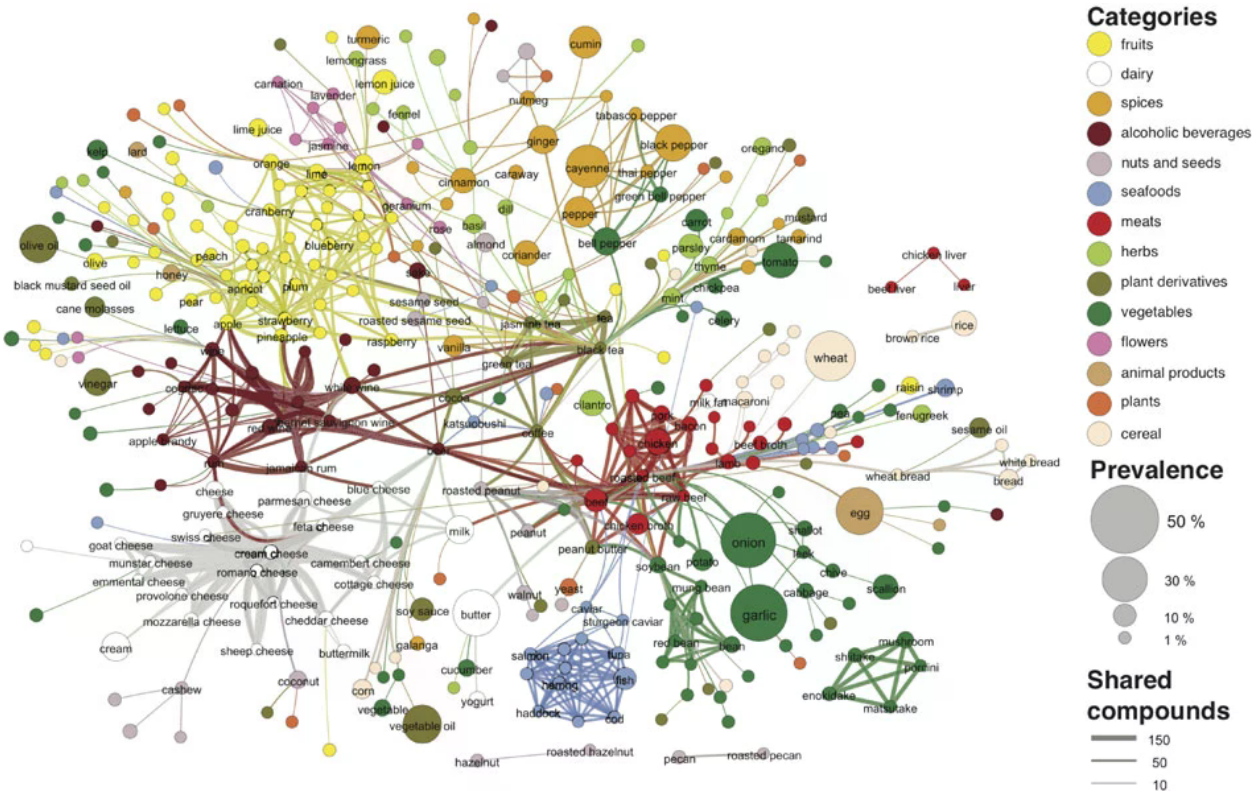


Figure 2.7: **Backbone of the flavor network.** Projection of the ingredient–compound bipartite graph onto ingredients, with edge thickness proportional to the number of shared volatile compounds. The pruned (backbone) view reveals chemically coherent clusters (e.g., meats, fruits, herbs/spices). Reproduced from Ahn et al. [2011].

Normalizations and backbones. As in the case of co–occurrence networks, raw overlaps tend to be dominated by a few very common elements — in this case, highly ubiquitous compounds that are present across a large fraction of ingredients. Such compounds contribute large baseline similarities between many ingredient pairs, even if those pairs are not particularly distinctive in practice. Without normalization, the flavor network is therefore biased toward generic chemical signals rather than specific affinities that might better explain culinary usage. To correct for this effect, one can apply normalizations such as cosine similarity

or resource allocation weights, which discount the influence of compounds that appear too widely and yield more interpretable measures of chemical proximity (see Section 2.3.2).

Even after normalization, the projected network remains very dense; this *density inflation* is especially evident in visual representations, where the abundance of links obscures meaningful patterns. For this reason, Ahn et al. [2011] applied backbone extraction using the disparity filter [Serrano et al., 2009] to retain only statistically significant links. The resulting pruned network, reproduced in Figure 2.7, reveal broad food groups (e.g., meats, fruits, herbs/spices), offering a chemically grounded map of the culinary space.

Interpretation and caveats. Flavor–overlap edges encode a notion of *chemical proximity*: ingredients that share many volatile compounds are considered similar. This is the quantitative basis of the *food pairing* hypothesis explored in Chapter 3. However, high compound ubiquity and cooking transformations (which alter volatile profiles) can blur signal, and overlap alone does not account for texture, preparation, or cultural usage—motivating the multi–representation approach used in this thesis.

2.4 Machine Learning: General Framework and Methods

In addition to network analysis, this thesis also employs methods from machine learning. Several of the tasks addressed, such as recognizing the cuisine of a recipe or recommending ingredients, can be formulated within the general framework of the *learning problem*. We begin with a brief overview of this framework and then introduce Support Vector Machines (SVMs) and Hierarchical Clustering, the two classifiers used in this thesis.

2.4.1 The general learning problem

At its core, machine learning addresses the problem of inferring patterns or rules from data. Formally, let $\mathcal{X}$ denote the space of inputs (e.g. feature representations of recipes or cuisines). Depending on the task, we may also have an associated output space $\mathcal{Y}$, representing labels or targets. The learner is provided with a dataset $\mathcal{D}$ of n examples drawn from an unknown distribution $P(X, Y)$, and the objective is to construct a function f that captures the relationship between inputs and outputs, or the structure of the data itself [Abu-Mostafa et al., 2012].

Within this predictive framework, different types of tasks can be defined depending on the nature of $\mathcal{Y}$:

- **Classification.** When $\mathcal{Y}$ is a finite set of discrete categories, the goal is to assign each input to one of them. An example in this thesis is the identification of a recipe's cuisine label from its ingredient-type network (Chapter 4).
- **Recommendation.** When $\mathcal{Y}$ encodes a relevance or preference signal, the goal is to predict which items are suitable in a given context. In the culinary domain, this corresponds to suggesting plausible missing ingredients for a partial recipe (Chapter 3).

- **Prediction of continuous outcomes.** More generally, if $\mathcal{Y}$ takes values in $\mathbb{R}$, the task is regression. While regression is not a main focus of this thesis, the same framework could be applied to forecasting nutritional values or consumer ratings.

Supervised learning. In supervised settings, each example comes with an input–output pair (x_i, y_i) with $x_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$. The goal is to learn a function $f : \mathcal{X} \to \mathcal{Y}$ that generalizes well to unseen data, rather than merely memorizing the training examples. To this end, the available dataset $\mathcal{D}$ is typically split into two disjoint parts: a *training set*, used to fit the model parameters, and a *test set*, used exclusively to evaluate predictive performance on unseen examples. In practice, cross-validation schemes are often employed, where the data are repeatedly partitioned into training and test subsets to obtain robust estimates of generalization accuracy (see Section 2.5). Typical supervised tasks include classification, recommendation, and regression, as discussed above [Abu-Mostafa et al., 2012, Bishop, 2006].

Unsupervised learning. In unsupervised settings, only inputs $\{x_i\}$ are available without corresponding labels. The goal is then to discover structure in the data, such as clusters, low-dimensional representations, or latent factors. In this thesis, unsupervised methods are used in the clustering of cuisines based on similarity of their ingredient-type networks.

2.4.2 Support Vector Machines

Among the classifiers considered, Support Vector Machines (SVMs) are of particular importance to this work. SVMs are a family of supervised learning algorithms introduced by Cortes and Vapnik [1995], designed to find a decision boundary that maximizes the separation margin between classes.

Linear case

In the simplest binary case, let $\{(x_i, y_i)\}_{i=1}^n$ be training examples with $y_i \in \{-1, +1\}$. The primal optimization problem of a linear SVM is

$$\min_{w,b,\xi} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i$$

subject to

$$y_i(w^\top x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, \ldots, n.$$

Here w is the normal vector of the separating hyperplane, b the bias, ξ_i slack variables allowing margin violations, and $C > 0$ a regularization parameter controlling the trade-off between margin maximization and classification errors.

Solving this constrained optimization problem via the method of Lagrange multipliers leads to its dual formulation. In the dual, the solution is expressed in terms of coefficients $\alpha_i \geq 0$, which are the Lagrange multipliers associated with the margin constraints $y_i(w^\top x_i + b) \geq 1 - \xi_i$. These coefficients measure how strongly each training point influences the decision boundary.

The weight vector w can then be written as a linear combination of the training inputs:

$$w = \sum_{i=1}^{n} \alpha_i y_i x_i.$$

Only a subset of the training points have nonzero multipliers $\alpha_i > 0$. These are called the *support vectors*, since they lie closest to the separating hyperplane and fully determine its orientation. Points farther from the margin have $\alpha_i = 0$ and do not contribute to the decision boundary. The resulting decision function takes the form

$$f(x) = \operatorname{sign}\left(\sum_{i=1}^{n} \alpha_i y_i (x_i^\top x) + b \right),$$

which depends only on the support vectors through their inner products with the new input x.

Geometrically, an SVM fits a separating hyperplane that maximizes the margin, i.e., the distance from the hyperplane to the nearest training points of any class. These nearest points are the *support vectors*. They are precisely the points that "press" against the margin: moving any one of them would tilt or shift the hyperplane, while points farther away (with $\alpha_i = 0$) have no effect on the decision boundary. The margin width equals $2/\|w\|$, so maximizing the margin corresponds to minimizing $\|w\|$ under the classification constraints.

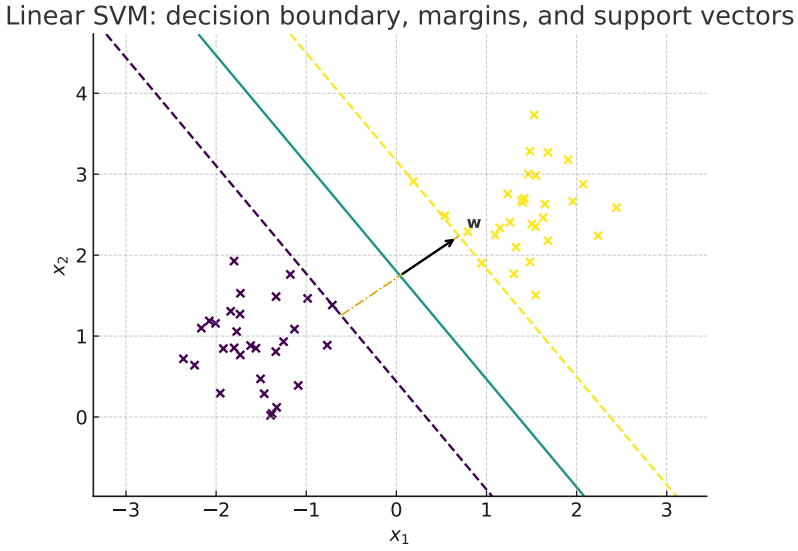


Figure 2.8: **Linear SVM geometry.** The separating hyperplane (solid line) maximizes the margin to the nearest points of either class. The dashed lines depict the margins. Support vectors (red outlines) are the only points that influence the decision boundary. The normal vector w is perpendicular to the separating hyperplane, and the margin width is $2/\|w\|$.

Kernel extension

Many real-world problems are not linearly separable in the original feature space. The kernel trick generalizes the above formulation by replacing the inner product $x_i^\top x$ with a

kernel function

$$K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle,$$

which implicitly maps inputs to a higher-dimensional feature space $\phi : \mathcal{X} \to \mathcal{H}$. The decision function then becomes

$$f(x) = \operatorname{sign}\left(\sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + b \right).$$

This shows that the linear case is a special case of the kernelized formulation with $K(x_i, x_j) = x_i^\top x_j$. Different choices of kernel correspond to different notions of similarity, allowing SVMs to capture nonlinear decision boundaries in the original input space.

In this thesis we use the radial basis function (RBF) kernel

$$K(x_i, x_j) = \exp\left(-\gamma \|x_i - x_j\|^2\right),$$

where $\gamma > 0$ controls the width of the Gaussian. The RBF kernel is well-suited to capturing nonlinear relationships in structured, high-dimensional data.

Multiclass extension

While the formulation above is binary, real tasks such as cuisine recognition involve multiple classes. We employ the multiclass extension implemented in `scikit-learn`'s `SVC` class [scikit-learn developers, 2025], which decomposes the problem into a set of one-vs-one binary classifiers. Given K classes, this results in $K(K-1)/2$ binary classifiers whose outputs are combined by majority voting.

2.4.3 Hierarchical clustering

In addition to supervised classification, this thesis also employs unsupervised methods to discover structure in culinary data. A central tool is *hierarchical clustering*, a family of algorithms that build a hierarchy of nested clusters without requiring the number of clusters to be specified in advance [Bishop, 2006].

The input to hierarchical clustering is a set of objects together with a pairwise dissimilarity measure $d(x_i, x_j)$. In this work, objects are ingredient-type networks representing cuisines, and dissimilarity is measured by the Jensen–Shannon (JS) distance between their edge-weight distributions. Given two probability distributions P and Q defined over the same support, the Jensen–Shannon distance is defined as:

$$JS(P \,\|\, Q) = \tfrac{1}{2} KL(P \,\|\, M) + \tfrac{1}{2} KL(Q \,\|\, M), \qquad M = \tfrac{1}{2}(P + Q), \tag{2.8}$$

where $KL(P \,\|\, Q) = \sum_i P(i) \, \log \frac{P(i)}{Q(i)}$ is the Kullback–Leibler divergence [Kullback and Leibler, 1951]. The square root of this divergence, $d_{JS}(P, Q) = \sqrt{JS(P \,\|\, Q)}$, is a true metric and is used in clustering analyses.

The agglomerative (bottom-up) procedure starts with each object in its own cluster. At each iteration, the two clusters with the smallest distance are merged, where inter-cluster distances are computed according to a *linkage criterion*. Common choices include:

- **Single linkage**: distance between closest elements of two clusters.
- **Complete linkage**: distance between farthest elements.
- **Average linkage**: average distance between all pairs across clusters.

This iterative merging continues until all objects are grouped into a single cluster, producing a binary tree known as a *dendrogram*. By cutting the dendrogram at a chosen distance threshold, a partition of the data into clusters is obtained. The choice of threshold reflects the desired level of granularity: higher cuts yield coarser groupings, while lower cuts produce finer distinctions.

2.5 Recommender Systems

Modern societies generate and consume information, products, and cultural artifacts at unprecedented scale. Recommender systems are algorithms designed to guide individuals through this abundance by suggesting items that are likely to be of interest. Nowadays, they represent a ubiquitous component of online platforms, powering personalized suggestions across domains as diverse as e-commerce, streaming services, and social media [Adomavicius and Tuzhilin, 2005, Ricci et al., 2022].

At their core, recommender systems exploit patterns of preference and similarity to identify relevant items. They can rely on explicit signals (ratings, likes, purchases) or implicit signals (browsing history, co–occurrence, contextual information). A classical distinction is between *content-based methods*, which recommend items similar to those a user has previously liked, and *collaborative filtering*, which leverages behavioral patterns across many users. This taxonomy, formalized in early surveys such as Adomavicius and Tuzhilin [2005], has shaped much of the field. Among these paradigms, collaborative filtering has been particularly influential, with early methods demonstrating that user–item interactions can be effectively predicted by exploiting similarity patterns among users or items.

In recent years, recommender systems have been reshaped by advances in machine learning. Deep learning approaches integrate sequence modeling, attention mechanisms, and representation learning, enabling models to capture complex, nonlinear patterns of preference that were out of reach for earlier techniques [Zhang et al., 2019].

In the culinary domain, recommender systems provide a natural framework for suggesting ingredients to extend or complete a recipe, for proposing substitutions when certain items are unavailable, and for recommending whole recipes to users based on their tastes or dietary constraints. This versatility has resulted in a rich application area, where personalization intersects with health, culture, and taste [Trattner and Elsweiler, 2017]. In this thesis, recommendation is used both as a methodological tool to probe culinary structure and as a way to test hypotheses about the principles that govern ingredient combinations.

2.5.1 The general framework

Despite methodological differences, most recommender systems operate within a common framework. Given a set of users $\mathcal{U}$ and a set of items $\mathcal{I}$, the goal is to estimate, for a target user $\alpha \in \mathcal{U}$, the relevance of each candidate item $i \in \mathcal{I}$. This is formalized through a *recommendation score* $\hat{r}_{\alpha i}$, which quantifies how likely user α is to engage with item i.

The computation of $\hat{r}_{\alpha i}$ depends on the paradigm: in content-based systems it is derived from the similarity between item i and those previously liked by α; in collaborative filtering it is inferred from patterns of co–usage across the entire user–item matrix M; in more advanced approaches it may result from latent factor models or deep learning architectures. Once scores are computed, items are ranked in decreasing order of $\hat{r}_{\alpha i}$, and the top–k unseen items are returned as recommendations.

This scoring–ranking process provides the unifying backbone of recommender systems, while leaving flexibility in how the scores themselves are estimated.

2.5.2 Content-based systems

Content-based methods recommend items that are similar to those a user has previously interacted with, where similarity is defined in terms of item attributes. For movies, these may include genre, actors, or keywords; for recipes, attributes can be nutritional profiles, ingredient lists, or preparation styles. The main advantage of content-based approaches is that they do not require data from other users, making them effective even in sparse settings or when preferences are highly individual. However, they tend to produce overly narrow suggestions, since recommendations remain confined to items resembling past choices. Content-based approaches are not employed in this thesis, but they are mentioned here for completeness as part of the standard taxonomy of recommender methods [Adomavicius and Tuzhilin, 2005, Ricci et al., 2022].

2.5.3 Collaborative filtering

Collaborative filtering (CF) is one of the most influential paradigms in recommendation, built on the idea that users with similar past behavior will have similar future preferences [Adomavicius and Tuzhilin, 2005]. Rather than relying on explicit item attributes, CF exploits implicit structures in the user–item interaction matrix. This makes it especially suited to contexts, such as culinary data, where items (ingredients, recipes) are richly structured but where implicit patterns of co–usage already carry strong signals.

The user–item matrix. At the core of CF lies the *interaction matrix*

$$M \in \mathbb{R}^{|\mathcal{U}|\times|\mathcal{I}|},$$

where $\mathcal{U}$ is the set of users and $\mathcal{I}$ the set of items. The entry $M_{\alpha i}$ records whether user $\alpha \in \mathcal{U}$ has interacted with item $i \in \mathcal{I}$ (e.g., watched a movie, purchased a product, or used an ingredient). In explicit-feedback settings, $M_{\alpha i}$ may be a rating; in implicit-feedback settings, it is often binary (0/1) or a frequency count. In this thesis, the recipe–ingredient biadjacency matrix introduced earlier (Section 2.2.7) is an instance of such a structure.

Neighborhood methods. The earliest collaborative filtering algorithms are known as *neighborhood-based methods*, because they rely on identifying a set of similar entities—a "neighborhood"—and using their behavior to make predictions. Neighborhood methods are based on a notion of *similarity*, which quantifies how alike two users or two items are, based on the interaction matrix M. There are two main variants, depending on whether similarity is computed across users or across items. In *user–user collaborative filtering*, recommendations for a target user α are generated by looking at other users with similar interaction profiles. In the simplest cases, similarity between two users is computed by considering their respective interaction profile, and therefore is a function of the two corresponding rows in the interaction matrix.

$$\mathrm{sim}(\alpha, \beta) = f(M_\alpha, M_\beta).$$

A basic example is co–occurrence,

$$\mathrm{sim}_{\mathrm{co}}(\alpha, \beta) = \sum_i M_{\alpha i} M_{\beta i},$$

which simply counts how many items both users have interacted with. Normalizations such as cosine similarity or correlation adjust this measure for differences in overall activity. Once a neighborhood of similar users has been identified, recommendations for α are made by aggregating the items most common among them.

The complementary perspective is offered by *item–item collaborative filtering*, where similarity is computed between columns of M:

$$\mathrm{sim}(i, j) = f(M^i, M^j).$$

Two items i and j are considered similar if they are frequently chosen by the same users; co–occurrence, cosine similarity, and Jaccard index are all standard options. A recommendation is then obtained by suggesting items that are most similar to those the target user has already engaged with.

While row–row and column–column similarities cover the classical case, many methods go further by defining similarity as a more general function of the entire interaction matrix:

$$\mathrm{sim}(x, y) = f(M),$$

where x, y may be users or items. This more general perspective accommodates methods based on diffusion processes on the user–item graph, probabilistic models of co–occurrence, or latent factor representations learned by matrix factorization.

Recommendation scores. Once similarities are defined, collaborative filtering methods compute recommendation scores by aggregating the preferences of the neighborhood.

In *user–user collaborative filtering*, the predicted score for user α and item i is a weighted average of how similar users β interacted with i:

$$\hat{r}_{\alpha i} = \frac{\sum_{\beta \neq \alpha} \mathrm{sim}(\alpha, \beta)\, M_{\beta i}}{\sum_{\beta \neq \alpha} |\mathrm{sim}(\alpha, \beta)|}.$$

Here $M_{\beta i}$ indicates whether user β interacted with item i, and the similarity acts as a weight. Intuitively, user α is predicted to like i if similar users did.

In *item–item collaborative filtering*, the predicted score for user α and candidate item i is obtained by looking at the items j that α has already consumed:

$$\hat{r}_{\alpha i} = \frac{\sum_{j \in \mathcal{I}} \mathrm{sim}(i, j)\, M_{\alpha j}}{\sum_{j \in \mathcal{I}} |\mathrm{sim}(i, j)|}.$$

In this case, the recommendation is stronger if i is similar to many items already present in α's profile.

Finally, once scores $\hat{r}_{\alpha i}$ are computed for all candidate items i, they are ranked in decreasing order and the top–k are returned as recommendations.

Relevance to this thesis. In practice, neighborhood methods are widely adopted as baseline algorithms across domains. Their effectiveness depends on the density of the interaction matrix: sparse datasets may limit predictive power, but in domains such as recipes—where co–occurrence data are abundant—they provide a natural and powerful framework.

In the context of culinary data, the item–item approach is particularly natural. If a recipe already contains a set of ingredients, additional ingredients can be recommended by looking at those that most frequently co-occur with the existing ones across the corpus of recipes. Accordingly, throughout this thesis we adopt the *item–item collaborative filtering* approach as the primary recommendation strategy.

Chapter 3

Re-examining the Food-Pairing Hypothesis

Humans have historically faced the fundamental challenge of selecting foods that satisfy nutritional needs while avoiding those that are toxic or unsafe. In early societies, food palatability likely evolved as a proxy for nutritional value. For instance, the pleasurable sensation associated with high-fat foods may reflect the energy density of fatty acids, while the sweetness of simple sugars—rapidly absorbed and metabolically useful—has long been associated with reward. In an ancestral environment where access to energy was limited, these preferences likely conferred an evolutionary advantage by promoting the consumption of calorie-dense foods needed for high-energy activities such as hunting or escaping predators.

However, the composition of modern human diets is not governed solely by biology. Climate and geography strongly influence the availability of ingredients: agricultural output is shaped by soil composition, weather conditions, and local ecosystems, while animal farming depends on the regional presence of domesticated species. Cultural practices and religious beliefs further contribute to the structure of local cuisines, giving rise to a globally diverse and historically layered gastronomic landscape.

In recent decades, food has transcended its original role as a means of survival and has become increasingly tied to pleasure, identity, and social status. In parallel, the industrialization of food systems and the growing availability of calorie-dense, highly palatable products have led to an epidemiological transition. With many infectious causes of death under control, modern societies now face a surge in lifestyle-related diseases, such as obesity, diabetes, and cardiovascular disorders. Ironically, the same preferences that once helped our ancestors survive may now undermine our health.

This tension—between the hedonic appeal of food and the need for sustainable, health-conscious diets—presents a pressing challenge for contemporary societies. Designing recipes and diets that are both nutritionally balanced and pleasurable is increasingly recognized as a critical problem in fields ranging from public health to gastronomy to food technology.

Addressing this challenge requires a deeper understanding of the principles that govern how ingredients are combined into recipes. Why do some combinations work better than others? Why do many cuisines revolve around a limited set of recurring ingredient pairings that define their core flavor identity? Is the relatively small number of recipes used worldwide—a few million at most—merely a snapshot of an ongoing exploration of culinary space,

or does it reflect intrinsic constraints in what combinations are perceived as acceptable?

While factors such as texture, temperature, and even sound contribute to our perception of food, flavor remains the dominant dimension of palatability. Flavor encompasses olfactory signals (mediated by volatile molecules binding to olfactory receptors), taste (mediated by molecules stimulating taste buds), and trigeminal sensations such as spiciness or freshness. As such, the chemical composition of ingredients—specifically their flavor compounds—provides a natural starting point for a systematic investigation into the principles of recipe construction.

While some of the richest and most complex national cuisines (e.g. french, indian, italian, chinese) have developed their own rules guiding the use and pairing of ingredients in recipes, these rules are confined to the field of single cultural traditions, and tipically fail when applied outside their realm of application. They are therefore more of an ensamble of heuristics arising from consolidated practices rather than universal principles.

The first attempt to identify a universal principle behind food palatability emerged not in a laboratory, but in the kitchen of a Michelin-starred restaurant. Around the year 2000, chef Heston Blumenthal was experimenting with savory ingredients and chocolate when he discovered that caviar and white chocolate formed an unexpectedly harmonious pairing. Intrigued, he consulted a flavor scientist at Firmenich, a major flavor manufacturer. Their comparative chemical analysis revealed that both caviar and white chocolate shared several key flavor compounds. This observation sparked the idea that novel, tasty food combinations could be created by identifying ingredients with overlapping chemical profiles—a concept that would become known as the *food pairing* hypothesis.

Blumenthal's early combinations—such as chocolate and blue cheese, or pork liver and jasmine—gained both culinary acclaim and media attention. What began as a chef's intuition soon attracted the interest of scientists, and the idea evolved into a formal hypothesis: that ingredients sharing flavor compounds are more likely to taste good together than those that do not. Ahn et al. [2011] were the first to provide a quantitative formulation of this idea, introducing a flavor network and metrics to capture co-occurrence patterns across cuisines.

Since then, the food pairing hypothesis has become a central theme in computational gastronomy, particularly due to its descriptive success in modeling Western culinary patterns. However, its utility as a *generative* principle—particularly in tasks such as ingredient recommendation—remains largely unexplored. If food pairing were a core mechanism of human taste preferences, then it should allow us to predict and construct appealing ingredient combinations based on molecular profiles alone. This possibility opens the door to algorithmic creativity in cooking and recipe generation.

In this chapter, we revisit the food pairing hypothesis from a predictive perspective. We investigate whether flavor-based similarity, derived from shared chemical compounds, provides a useful signal for ingredient recommendation, and compare it to recipe-based similarity, which reflects actual usage patterns in empirical data. To this end, we cast the problem as a controlled retrieval task, repeatedly removing one ingredient from a recipe and testing whether different recommenders can recover it. Alongside recipe- and compound-based models, we also include direct baselines to benchmark performance. Our framework allows us to examine not only overall accuracy but also the effect of restricting candidate recommendations to the same ingredient category, the extent to which performance varies across world regions, and the mechanisms through which successful compound-based predic-

tions arise. In doing so, we provide a systematic assessment of the conditions under which chemistry-based similarity can—or cannot—support culinary recommendation.

3.1 The Food Pairing Hypothesis: From Networks to Recipes

The empirical foundation of this chapter is provided by the dataset introduced by Ahn et al. [2011], which played a central role in testing the original formulation of the food pairing hypothesis. This dataset links recipes, ingredients, and flavor compounds, thus enabling a systematic investigation of whether chemical similarity can explain culinary practice. While its main feature and its formal graph-theoretic representation has already been introduced in Chapter 2, here we briefly recall its essential features and emphasize those most relevant to the present analysis.

3.1.1 Ahn's Dataset

The dataset comprises 56,498 traditional recipes collected from three major online repositories (`allrecipes.com`, `epicurious.com`, and `menupan.com`), covering five broad culinary regions: North American, Western European, Southern European, Latin American, and East Asian. Each recipe is represented as a set of ingredients, consistent with the formal definition adopted in Chapter 2, and each ingredient is annotated with its associated set of volatile *flavor compounds*, which are known to contribute to olfactory and gustatory perception. Overall, the dataset includes 381 unique ingredients and 1,021 distinct compounds, with an average of about 51 compounds per ingredient.

Ingredients are further grouped into 14 high-level food categories (e.g. *Dairy*, *Vegetable*, *Meat*, *Herb/Spice*), a type of classification we use in later sections to analyze structural and functional patterns in culinary networks (see Section 2.1.1 for the complete list of categories and their role in network construction).

Table 3.1 summarizes the key characteristics of the dataset. This representation naturally lends itself to a bipartite graph between ingredients and compounds, from which one can derive a projected *flavor-sharing network* G_I by linking pairs of ingredients that share one or more compounds. The bipartite and projected representations introduced in Chapter 2 form the basis for the food-pairing analysis carried out in this chapter.

3.1.2 Formulation of the Food Pairing Hypothesis

Building on this dataset, Ahn et al. [2011] proposed the *food pairing hypothesis*, which posits that ingredients sharing a greater number of flavor compounds are more likely to co-occur in recipes. While initially rooted in anecdotal observations and chef intuition, Ahn and colleagues provided the first large-scale quantitative test of the hypothesis using network theory and molecular data. The ingredient–compound bipartite network and its projection onto the ingredient layer (the *flavor network*) provide the natural mathematical setting for testing this idea. Within this framework, flavor compound overlap between ingredients can be formalized and compared against empirical co-occurrence patterns extracted from recipes.

Dataset characteristic	**Value**
Total number of recipes	56,498
Total number of ingredients	381
Total number of flavor compounds	1,021
Average compounds per ingredient	~ 51
Number of ingredient categories	14
Number of broad culinary regions	5
Original data sources	Epicurious, AllRecipes, Menupan

Table 3.1: Summary statistics of the dataset used for the analysis.

This subsection introduces the network representations and quantitative measures used to evaluate the hypothesis.

Formally, let $M_{\alpha i}^{\text{flavor}}$ be the biadjacency matrix of the *ingredient–compound bipartite graph*, where α indexes flavor compounds and i indexes ingredients (see Chapter 2, Section 2.3.3). An entry $M_{\alpha i}^{\text{flavor}} = 1$ indicates that compound α is present in ingredient i, and 0 otherwise. Projecting this bipartite network onto the ingredient layer yields the *flavor network* G_I, where the weight of an edge between ingredients i and j is given by the number of shared compounds:

$$w_{ij}^{f} = |C_i \cap C_j|,$$

with C_i the set of compounds associated with ingredient i. The node strength

$$s_i = \sum_{j \neq i} w_{ij}^{f}$$

then quantifies the total number of compounds ingredient i shares with the rest of the ingredient set. These definitions provide the foundation for quantifying the extent to which recipes favor compound-sharing ingredient pairs.

In parallel, the *recipe–ingredient bipartite graph* (Chapter 2, Section 2.3.1) captures how ingredients are actually combined in culinary practice, with biadjacency matrix $M_{\rho i}^{\text{recipes}}$ indicating whether ingredient i appears in recipe ρ. This representation naturally links network statistics such as degree distributions to recipe size distributions.

To test the food pairing hypothesis, Ahn et al. [2011] introduced a recipe-level measure of shared compounds. For a recipe R with n_R ingredients, the average number of shared compounds among ingredient pairs is

$$N_s(R) = \frac{2}{n_R(n_R - 1)} \sum_{\substack{i,j \in R \\ i \neq j}} |C_i \cap C_j|.$$

The cuisine-level statistic $\langle N_s \rangle$ is then defined as the mean of $N_s(R)$ across all recipes in that cuisine. By comparing empirical values of $\langle N_s \rangle$ with those obtained from randomized recipe

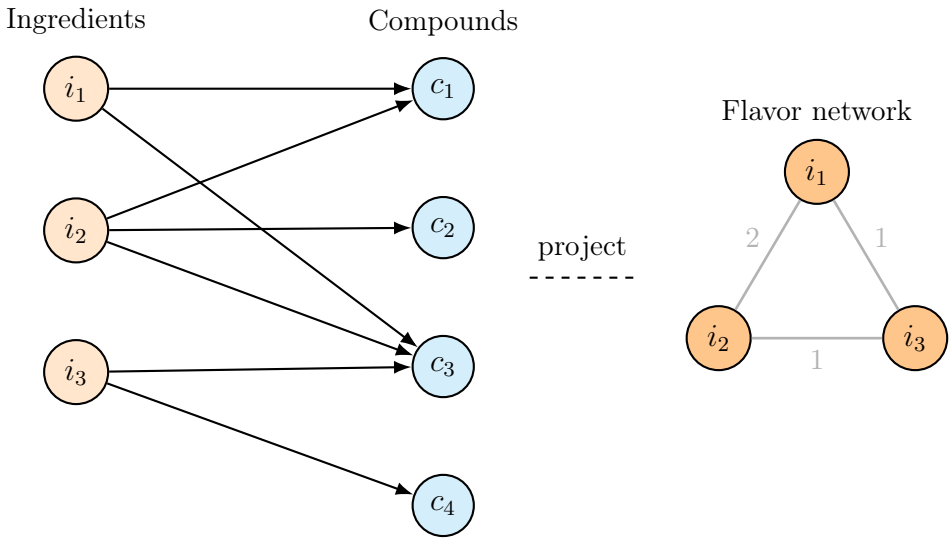


Figure 3.1: Bipartite ingredient–compound network (left) and its weighted projection onto the ingredient layer—the *flavor network* (right). Edge labels in the projection report the number of shared compounds between each ingredient pair.

sets (constructed by shuffling ingredients while preserving frequencies), Ahn et al. assessed whether real cuisines exhibit systematic deviations from random expectations.

Their results revealed striking cross-cultural differences: North American and Western European cuisines display strong positive food pairing (empirical $\langle N_s \rangle$ greater than randomized baseline), whereas East Asian cuisines exhibit negative food pairing, favoring ingredient combinations with little or no compound overlap. Community detection on the backbone of the flavor network further showed modular structures aligned with broad ingredient categories (e.g. meats, fruits, herbs).

Thus, the food pairing hypothesis can be reformulated as a *topological property*: recipes preferentially combine ingredient pairs that are strongly connected in the flavor network. While this view provides a compelling descriptive framework, it leaves open a crucial question: does compound-level overlap also carry *predictive* value? In other words, if the food pairing hypothesis truly reflects a generative principle of human taste, then flavor-based similarity should enable us not only to describe existing cuisines, but also to recommend new, plausible ingredient combinations. Conversely, if prediction based on chemical profiles performs poorly compared to empirical co-occurrence data, this would suggest that food pairing is more a retrospective pattern than a forward-looking mechanism. In the following sections, we therefore evaluate compound-level similarity as a signal for ingredient recommendation and benchmark it against recipe-based similarity derived directly from observed culinary practice, using the recommendation framework introduced in Chapter 2,(Section 2.5).

3.2 Comparing Flavor-based and Recipe-based Similarity

Food recommendation systems have emerged as an active area of research, spanning applications from personalized diet planning to recipe generation and ingredient substitution [Trattner and Elsweiler, 2017]. As discussed in Chapter 2, Subsection 2.5.1, recommender systems adapt classical paradigms in recommender research—content-based filtering, matrix factorization, and collaborative filtering—across domains. In the culinary context, recipes are commonly treated as "users", while ingredients are treated as "items" to be recommended or withheld depending on compatibility, preference, or nutritional constraints. This abstraction provides a general and flexible framework for modeling culinary data.

A particularly intuitive paradigm is *item–item collaborative filtering* [Adomavicius and Tuzhilin, 2005], introduced in Chapter 2, Subsection 2.5.3. In this approach, recommendations are generated by comparing items (here, ingredients) according to the similarity of their usage contexts. If a recipe already contains a set of ingredients, additional ingredients can be suggested by identifying those most similar to the existing ones. This logic connects directly to the food pairing hypothesis: if similarity is defined in terms of shared flavor compounds, and if co-occurrence in recipes is indeed driven by such molecular similarity, then an item–item recommender using compound overlap should successfully reproduce real ingredient combinations. Conversely, if empirical usage patterns dominate, then similarity measured from recipe co-occurrence should provide a stronger predictive signal.

This chapter builds on this observation. The food pairing hypothesis can be naturally embedded within the item–item collaborative filtering framework, where similarity between ingredients is derived either from the ingredient–compound bipartite graph (flavor-based) or from the recipe–ingredient bipartite graph (recipe-based), both introduced in Chapter 2. Flavor-based similarity reflects what we call *chemical proximity*: ingredients with overlapping molecular profiles are considered compatible. This signal is static and culture-agnostic, since it arises directly from chemistry. By contrast, recipe-based similarity is empirical, encoding how ingredients are actually used together across cuisines, thereby incorporating not only flavor but also preparation methods, traditions, and human preferences.

The central question of this chapter is therefore: *does flavor-based similarity provide a meaningful and predictive signal for ingredient recommendation, or is empirical co-occurrence the stronger guide to culinary compatibility?* Addressing this question allows us to test the food pairing hypothesis as a *generative* model of recipe formation, rather than a purely descriptive pattern.

The central question of this chapter is therefore: *does flavor-based similarity provide a meaningful and predictive signal for ingredient recommendation, or is empirical co-occurrence the stronger guide to culinary compatibility?* Addressing this question allows us to test the food pairing hypothesis as a *generative* model of recipe formation, rather than a purely descriptive pattern. To this end, we evaluate both similarity measures within the collaborative filtering and graph-based frameworks introduced in Chapter 2, using standard similarity metrics (cosine, Jaccard, PMI) and more advanced models such as LightGCN.

3.2.1 A unified framework for ingredient similarity

To answer this question, we require a common mathematical framework that allows flavor-based and recipe-based similarity to be expressed on equal footing. Following the notation introduced in Chapter 2, we represent both cases using a biadjacency matrix $M_{\alpha i}$ of a bipartite graph, where i indexes ingredients and α indexes either:

- recipes ($\alpha = \rho$), yielding *recipe-based similarity*, or
- compounds ($\alpha = c$), yielding *flavor-based similarity*.

Similarity between two ingredients i and j is then expressed as a function of the biadjacency matrix M:

$$S_{ij} = f(M)$$

As introduced in Chapter 2, similarity functions play a central role in both neighborhood-based recommender systems (Section 2.5.3) and in the normalization of ingredient co–occurrence networks (Section 2.3.2). Here, we specialize the same general formalism to the two bipartite graphs at the core of the food pairing framework: the ingredient–compound network and the recipe–ingredient network. This allows us to compute ingredient–ingredient similarity either in terms of molecular overlap (flavor-based) or in terms of empirical co–usage (recipe-based). Standard similarity measures include (see also Chapter 2, Section 2.3.2):

- **Cosine similarity**:

$$S_{ij} = \frac{\sum_\alpha M_{\alpha i} M_{\alpha j}}{\sqrt{\sum_\alpha M_{\alpha i}^2}\sqrt{\sum_\alpha M_{\alpha j}^2}},$$

- **Jaccard similarity**:

$$S_{ij} = \frac{|\{\alpha : M_{\alpha i} = 1 \wedge M_{\alpha j} = 1\}|}{|\{\alpha : M_{\alpha i} = 1 \vee M_{\alpha j} = 1\}|},$$

- **Pointwise Mutual Information (PMI)**:

$$S_{ij} = \log \frac{P(i,j)}{P(i)P(j)},$$

  where probabilities are estimated over the rows of M.

Each definition produces a fully connected, weighted ingredient–ingredient graph: one reflecting molecular overlap, the other empirical co-occurrence. This unified representation allows us to directly compare the predictive value of flavor-based and recipe-based similarity within collaborative filtering and graph-based recommendation models, thereby assessing whether food pairing functions as a generative principle or merely a descriptive regularity. An alternative perspective, consistent with the discussion in Chapter 2, Subsection 2.3.2, is to interpret the similarity matrices derived here as weighted ingredient networks. In that earlier section, co–occurrence counts were normalized to construct similarity networks that emphasize distinctive associations between ingredients rather than absolute frequency. The same principle applies in the present setting: whether similarity is computed from recipes or

from compounds, the resulting S_{ij} can be understood as defining an ingredient–ingredient similarity network. This viewpoint highlights the structural parallel between flavor-based and recipe-based similarity, and clarifies their role as the backbone over which item–item collaborative filtering operates.

3.2.2 Item–item similarity metrics

As introduced in Chapter 2, similarity functions provide the backbone of item–item collaborative filtering and of normalized co–occurrence networks. In this chapter, we extend that general framework to the recommendation task and systematically test a broad family of similarity metrics. The goal is not to advocate for a particular measure, but to ensure that our evaluation of flavor–based versus recipe–based similarity is robust to methodological choices. To this end, we consider eighteen alternative metrics drawn from network science, information retrieval, and recommender systems. Each metric instantiates the same general formalism but captures a different intuition about what it means for two ingredients to be "similar"—from raw co–occurrence frequency to degree–corrected association, from statistical correlation to higher–order path–based proximity.

All metrics are defined on the same general formalism, using the biadjacency matrix $M_{\alpha i}$ of a bipartite graph, where i indexes ingredients and α indexes context nodes (recipes or compounds). For convenience, we recall the key quantities used throughout:

- $CO_{ij} = \sum_\alpha M_{\alpha i} M_{\alpha j}$ is the number of co-occurrences between ingredients i and j, i.e., the number of recipes (or compounds) in which both appear together.
- $u_i = \sum_\alpha M_{\alpha i}$ is the **ubiquity** of ingredient i, i.e., the number of context nodes α it is linked to.
- $d_\alpha = \sum_i M_{\alpha i}$ is the **diversification** of context node α, i.e., the number of ingredients linked to it.
- N is the total number of context nodes in the bipartite network.
- $\bar{d} = \frac{1}{N} \sum_\alpha d_\alpha$ is the **average diversification**, i.e., the average number of ingredients per recipe (or average context size).
- $D_i = \mathrm{diag}(u_i)$ is the diagonal matrix of ingredient degrees (ubiquities).
- $D_u = \mathrm{diag}(d_\alpha)$ is the diagonal matrix of context degrees (diversifications).

The similarity between two ingredients i and j is denoted B_{ij} and reflects the strength of association as computed by a given metric. Taken together, the metrics span several methodological families, each capturing a different intuition about how co-occurrence encodes similarity:

Raw co-occurrence metrics quantify similarity simply by the number of times two ingredients appear together, without any correction. They provide a baseline measure of joint frequency, against which more sophisticated normalizations can be compared (e.g., Co-occurrences).

Normalized overlap metrics rescale co-occurrence counts to control for ingredient ubiquity. Additive forms (e.g., Jaccard, Sørensen) divide by the union or average degree, while multiplicative forms (e.g., Cosine, Asymmetric Cosine) normalize by products or square roots of popularities. Variants such as Hub Depressed/Promoted further adjust how hub ingredients influence similarity.

Degree-aware link prediction metrics penalize heterogeneous context nodes, reducing the contribution of ubiquitous contexts. They emphasize rare but specific co-occurrences and suppress hub effects, as in Adamic/Adar, Resource Allocation and Taxonomy Network similarity.

Weighting-based metrics (IR-inspired) borrow from information retrieval, reweighting ingredients to suppress hubs and boost informative ones. This family emphasizes discriminative associations and mitigates the bias of ubiquitous ingredients (e.g., TF–IDF Cosine, BM25 Cosine).

Path- and diffusion-based metrics exploit higher-order connections in the bipartite graph. Probabilistic Spreading simulates directed resource flow, while P3α and RP3β model multi-step random walks with degree dampening. These metrics capture indirect associations and are particularly suited for recommendation tasks.

Statistical and null-model metrics assess similarity relative to expected usage patterns. The Pearson correlation coefficient measures centered co-usage, controlling for global popularity, while Sapling Similarity incorporates null-model expectations to detect signal enrichment beyond structural bias.

We summarize these metrics and their mathematical definitions in Table 3.2, and we provide detailed rationales in the following dedicated paragraphs.

Co-occurrence (CN) Also known as *Common Neighbors*, the co-occurrence metric simply counts how many times two ingredients i and j appear together in the same context (e.g., recipe or compound profile):

$$B_{ij}^{CN} = CO_{ij} = \sum_{\alpha} M_{\alpha i} M_{\alpha j}.$$

Because it uses raw counts, this measure is highly sensitive to ingredient popularity: pairs involving very frequent ingredients (hubs) receive large scores even if the association is not specific or distinctive. It thus serves as a baseline benchmark without any correction for popularity or context effects.

Jaccard Index (JA) The Jaccard index normalizes co-occurrence by the size of the union of neighbors of i and j:

$$B_{ij}^{JA} = \frac{CO_{ij}}{u_i + u_j - CO_{ij}}$$

Recall that u_i and u_j denote the ubiquities of ingredients i and j, i.e., the total number of contexts in which they appear. In this sense, ubiquity also captures ingredient popularity:

Method	Formula	Reference
Co-occurrences (B^{CN})	CO_{ij}	Liben-Nowell and Kleinberg [2007]
Jaccard (B^{JA})	$\frac{CO_{ij}}{u_i+u_j-CO_{ij}}$	Jaccard [1901]
Sørensen index (B^{SO})	$\frac{2CO_{ij}}{u_i+u_j}$	Sørensen [1948]
Hub Depressed index (B^{HD})	$\frac{CO_{ij}}{\max(u_i,u_j)}$	Ravasz et al. [2002]
Hub Promoted index (B^{HP})	$\frac{CO_{ij}}{\min(u_i,u_j)}$	Ravasz et al. [2002]
Cosine Similarity (B^{CS})	$\frac{CO_{ij}}{\sqrt{u_i u_j}}$	Salton and McGill [1983]
Asymmetric Cosine (B^{AC})	$B^{AC}_{ij} = \frac{\sum_\alpha M_{\alpha i} M_{\alpha j}}{\left(\sum_\alpha M^2_{\alpha i}\right)^{\alpha/2}\left(\sum_\alpha M^2_{\alpha j}\right)^{(1-\alpha)/2}}$	Desrosiers and Karypis [2011]
Leicht-Holme-Newman (B^{LHN})	$\frac{CO_{ij}}{u_i u_j}$	Leicht et al. [2006]
Adamic/Adar (B^{AD})	$\sum_\alpha \frac{M_{\alpha i} M_{\alpha j}}{\log(d_\alpha)}$	Adamic and Adar [2003]
Resource Allocation (B^{RA})	$\sum_\alpha \frac{M_{\alpha i} M_{\alpha j}}{d_\alpha}$	Zhou et al. [2009]
Taxonomy Network (B^{TN})	$\frac{1}{\max(u_i,u_j)} \sum_\alpha \frac{M_{\alpha i} M_{\alpha j}}{d_\alpha}$	Zaccaria et al. [2014]
TF–IDF Cosine ($B^{\text{TF-IDF}}$)	$x_{\alpha i} = M_{\alpha i}\, idf_i,\ \ idf_i = \log \frac{N+1}{u_i+1} + 1;\ \ B^{\text{TF-IDF}}_{ij} = \frac{\sum_\alpha x_{\alpha i} x_{\alpha j}}{\sqrt{\sum_\alpha x^2_{\alpha i}}\sqrt{\sum_\alpha x^2_{\alpha j}}}$	Salton and McGill [1983]
BM25 Cosine (B^{BM25})	$w^{\text{BM25}}_{\alpha i} = \frac{M_{\alpha i}(k_1+1)}{M_{\alpha i}+k_1(1-b+b\, d_\alpha/d)} \cdot \log\left(\frac{N-u_i+0.5}{u_i+0.5}+1\right);\ \ B^{\text{BM25}}_{ij} = \frac{\sum_\alpha w_{\alpha i} w_{\alpha j}}{\sqrt{\sum_\alpha w^2_{\alpha i}}\sqrt{\sum_\alpha w^2_{\alpha j}}}$	Robertson and Zaragoza [1009]
Probabilistic Spreading (B^{PS})	$\frac{1}{u_j} \sum_\alpha \frac{M_{\alpha i} M_{\alpha j}}{d_\alpha}$	Zhou et al. [2007]
P3α ($B^{\text{P3}\alpha}$)	$B = D_i^{-\alpha} M^\top D_u^{-\alpha} M$	Zhou et al. [2007]
RP3β ($B^{\text{RP3}\beta}$)	$B = \left(D_i^{-\alpha} M^\top D_u^{-\alpha} M\right) D_i^{-\beta}$	Zhou et al. [2007], Volpe et al. [2022]
Pearson Correlation (B^{PCC})	$\frac{\sum_\alpha (M_{\alpha i}-\frac{u_i}{N})(M_{\alpha j}-\frac{u_j}{N})}{\sqrt{\sum_\alpha (M_{\alpha i}-\frac{u_i}{N})^2}\sqrt{\sum_\alpha (M_{\alpha j}-\frac{u_j}{N})^2}}$	Shardanand and Maes [1995]
Sapling Similarity (B^{SAP})	$\left(\frac{CO_{ij}(1-\frac{CO_{ij}}{u_j})+(u_i-CO_{ij})(1-\frac{u_i-CO_{ij}}{N-u_j})}{u_i(1-\frac{u_i}{N})} - 1\right) \text{sign}\left(1-\frac{CO_{ij}N}{u_i u_j}\right)$	Albora et al. [2023]

Table 3.2: Similarity metrics used in item-item collaborative filtering. Each metric measures the proximity between ingredients i and j using their co-occurrence in context nodes α (recipes or flavor compounds). See main text for a description of variables.

highly ubiquitous ingredients are also the most popular. By dividing by the union size, the Jaccard index discounts the contribution of popular ingredients: even if two hubs co-occur often, their similarity score is reduced because the denominator grows with ubiquity. This makes the measure less sensitive to hubs than raw co-occurrence, highlighting more distinctive associations.

Sørensen Index (SO) This metric normalizes co-occurrence by the arithmetic mean of the ubiquities:

$$B^{SO}_{ij} = \frac{2\, CO_{ij}}{u_i + u_j}.$$

The normalization softens the effect of highly popular ingredients, though less strongly than cosine. At the same time, the arithmetic mean in the denominator makes the metric particularly sensitive to low-degree (rare) ingredients, amplifying their contribution.

Hub Depressed Index (HD) The Hub Depressed Index explicitly penalizes hubs by normalizing with the maximum ubiquity:

$$B_{ij}^{HD} = \frac{CO_{ij}}{\max(u_i, u_j)}$$

If one ingredient is very popular, the denominator is large and the score is reduced. The underlying assumption is that hubs provide little discriminative information for similarity.

Hub Promoted Index (HP) In contrast, the Hub Promoted Index rewards co-occurrence with hubs by normalizing with the minimum ubiquity:

$$B_{ij}^{HP} = \frac{CO_{ij}}{\min(u_i, u_j)}$$

Here, if at least one of the two ingredients is rare, the denominator is small, producing a higher score. This assumes that co-usage with highly popular ingredients strengthens similarity rather than diluting it.

Cosine Similarity (CS) Cosine similarity normalizes raw co-occurrence by the geometric mean of the ubiquities of i and j:

$$B_{ij}^{CS} = \frac{CO_{ij}}{\sqrt{u_i u_j}}.$$

Unlike additive normalizations such as Jaccard or Sørensen, Cosine uses a multiplicative form: the denominator thus scales sub-linearly as the square root of the product of popularities. This multiplicative normalization reduces the influence of hubs more gently: two popular ingredients are not overly penalized as long as their overlap grows proportionally, while similarity decreases if one ingredient is popular without sufficient co-occurrence with the other. As a result, cosine is less sensitive to ingredient hubs yet remains responsive to disproportional differences in degree.

Leicht-Holme-Newman Index (LHN) This measure applies a stronger penalization of popularity by dividing by the product of the ubiquities:

$$B_{ij}^{LHN} = \frac{CO_{ij}}{u_i u_j}$$

Compared to cosine, which uses the geometric mean, LHN employs a stricter multiplicative normalization. As a result, pairs of popular ingredients are heavily down-weighted, while rare pairs are amplified. This makes LHN particularly effective at highlighting unusual combinations, but it can also suppress similarities involving hubs more aggressively than cosine.

Asymmetric Cosine (B^{AC}). A degree-controlled generalization of cosine that introduces directionality via a power parameter $\alpha \in (0, 1)$:

$$B_{ij}^{AC} = \frac{\sum_\alpha M_{\alpha i} M_{\alpha j}}{\|M_{\cdot i}\|_2^{\alpha} \; \|M_{\cdot j}\|_2^{1-\alpha}}.$$

For binary M, $\|M_{\cdot i}\|_2 = \sqrt{u_i}$. When $\alpha \neq \frac{1}{2}$ the similarity is *asymmetric* ($B_{ij}^{AC} \neq B_{ji}^{AC}$), allowing one to emphasize the "query" item i (larger α) or the "neighbor" item j (smaller α). By tuning α, one can control how strongly ingredient popularity affects the normalization. We use $\alpha = 0.6$ by default, slightly privileging the query side.

Adamic/Adar (AD) This metric rewards overlaps that occur in rare contexts. Each shared context α contributes inversely to the logarithm of its degree d_α, so that ubiquitous contexts have little effect:

$$B_{ij}^{AD} = \sum_\alpha \frac{M_{\alpha i} M_{\alpha j}}{\log d_\alpha}.$$

Originally introduced in social networks, this measure highlights niche-based affinities: two ingredients are considered more similar if they meet in few, specific contexts rather than in many common ones.

Resource Allocation Index (RA) A close relative of Adamic/Adar, but using a linear penalization of high-degree contexts:

$$B_{ij}^{RA} = \sum_\alpha \frac{M_{\alpha i} M_{\alpha j}}{d_\alpha}$$

It emphasizes shared connections through rare contexts even more than Adamic/Adar.

Taxonomy Network Similarity (TN) This measure combines two forms of correction: overlaps are penalized both at the context level (by dividing by d_α) and at the ingredient level (by dividing by $\max(u_i, u_j)$):

$$B_{ij}^{TN} = \frac{1}{\max(u_i, u_j)} \sum_\alpha \frac{M_{\alpha i} M_{\alpha j}}{d_\alpha}.$$

The first factor suppresses the contribution of ubiquitous contexts, while the second penalizes pairs involving hub ingredients. TN therefore emphasizes specific overlaps between relatively non-hub ingredients through rare contexts.

TF–IDF weighted Cosine ($B^{\text{TF-IDF}}$). A popular information retrieval weighting scheme that down-weights ubiquitous ingredients and boosts informative ones. Each ingredient i is reweighted by its inverse document frequency:

$$idf_i = \log\!\left(\tfrac{N+1}{df_i+1}\right) + 1, \qquad df_i = u_i.$$

With $x_{\alpha i} = M_{\alpha i}\, idf_i$, cosine similarity is applied to the reweighted columns. The effect is to suppress high-popularity (hub) ingredients while amplifying rare but distinctive ones. This balances frequency with informativeness, curbing popularity bias while preserving the multiplicative scaling of cosine.

BM25-weighted Cosine ($B^{\mathbf{BM25}}$). A refinement of TF–IDF weighting that incorporates both context length normalization and term burstiness. The weight of ingredient i in context α is

$$w_{\alpha i}^{\mathrm{BM25}} = \frac{tf_{\alpha i}(k_1+1)}{tf_{\alpha i} + k_1\big(1-b+b\,d_\alpha/\bar{d}\big)} \cdot \log\!\left(\frac{N-df_i+0.5}{df_i+0.5}+1\right),$$

where $tf_{\alpha i} = M_{\alpha i}$, $df_i = u_i$, $d_\alpha = \sum_i M_{\alpha i}$, and $\bar{d} = \frac{1}{N}\sum_\alpha d_\alpha$. Cosine is then applied to the BM25-weighted vectors. Compared to TF–IDF, BM25 more flexibly adjusts for varying recipe sizes and ingredient burstiness across contexts, often yielding stronger retrieval performance in domains with heterogeneous context lengths (e.g., recipes of different sizes). We adopt the standard parameters $k_1 = 1.2$, $b = 0.75$.

Probabilistic Spreading (PS) Inspired by diffusion processes on bipartite graphs, this measure simulates the probability of "resource flow" from item i to item j through shared contexts:

$$B_{ij}^{PS} = \frac{1}{u_j}\sum_\alpha \frac{M_{\alpha i}M_{\alpha j}}{d_\alpha}.$$

The d_α in the denominator penalizes popular contexts, while the factor $1/u_j$ normalizes by the target ingredient's ubiquity. As a result, $B_{ij} \neq B_{ji}$ in general: the similarity flow is *asymmetric*, reflecting the intuition that ingredient i may be a stronger predictor of j than vice versa.

P3α (three-step walk with degree dampening). This metric is based on a three-step random walk on the bipartite graph: item $\to$ context $\to$ item. Let M be the biadjacency matrix, and D_u and D_i the diagonal degree matrices on the context and item sides, respectively. The similarity is computed as

$$B = D_i^{-\alpha}\, M^\top D_u^{-\alpha} M.$$

The exponents $\alpha \in [0,1]$ control how much high-degree nodes are penalized. When $\alpha = 0$, the walk is undamped and hubs dominate; increasing α down-weights popular contexts and items, giving more weight to rare associations. This allows the metric to balance exploration through the network with robustness against popularity bias. In the culinary domain, this means that frequent co-uses of ubiquitous ingredients (such as salt or water) are deemphasized, while rarer, more distinctive ingredient pairings emerge more strongly.

RP3β (asymmetric popularity-penalized P3α). RP3β extends P3α by introducing an additional popularity penalty on the *target* side:

$$B = \big(D_i^{-\alpha}\, M^\top D_u^{-\alpha} M\big) D_i^{-\beta}.$$

The right multiplication by $D_i^{-\beta}$ makes the similarity asymmetric, since the penalty applies only to candidate items. The parameter $\beta \geq 0$ suppresses recommendations of very popular targets, encouraging the retrieval of less ubiquitous ingredients. In practice, moderate values of α and β mitigate both source- and target-side popularity effects, improving top-k retrieval stability in recommender settings.

Pearson Correlation Coefficient (PCC) This classic statistical measure computes the correlation between the centered co-usage profiles of i and j:

$$B_{ij}^{PCC} = \frac{\sum_\alpha (M_{\alpha i} - \frac{u_i}{N})(M_{\alpha j} - \frac{u_j}{N})}{\sqrt{\sum_\alpha (M_{\alpha i} - \frac{u_i}{N})^2} \sqrt{\sum_\alpha (M_{\alpha j} - \frac{u_j}{N})^2}}$$

By subtracting the expected frequency u_i/N for each ingredient, PCC controls for global popularity bias and highlights whether two items co-occur more (positive correlation) or less (negative correlation) than expected under independent usage. It thus detects both attraction and repulsion between ingredients.

Sapling Similarity (SAP) A more recent measure designed for recommendation systems, SAP incorporates a probabilistic null model and explicit popularity balancing. Its formulation is:

$$B_{ij}^{SAP} = \left(\frac{CO_{ij}(1 - \frac{CO_{ij}}{u_j}) + (u_i - CO_{ij})(1 - \frac{u_i - CO_{ij}}{N - u_j})}{u_i(1 - \frac{u_i}{N})} - 1 \right) \cdot \text{sign}\left(1 - \frac{CO_{ij}N}{u_i u_j}\right)$$

SAP evaluates whether the observed co-occurrence CO_{ij} exceeds what is expected under randomized association, correcting for the degrees u_i and u_j. By combining null-model expectations with degree balancing, it aims to capture genuine signal enrichment while mitigating structural popularity bias.

3.2.3 How Recommendations Are Generated

Having defined how similarity matrices B are constructed from either recipes or compounds (Sec. 3.2.2), we now turn to how these similarities are operationalized to produce ingredient recommendations. That is, given a partially observed recipe, how do we decide which ingredient to suggest as a likely candidate for addition?

As introduced in Chapter 2, Subsection 2.5.1, the goal of a recommender system is to compute a *recommendation score* $\hat{r}_{\alpha i}$ for each candidate item i, relative to a target user α.

In our culinary setting, recipes play the role of users and ingredients the role of items. Within the item–item collaborative filtering framework (Chapter 2, Subsection 2.5.3), a candidate ingredient is evaluated according to its similarity to those already present in the recipe. Once a similarity matrix B has been fixed, we compute for each candidate ingredient i a recommendation score relative to recipe α:

$$\hat{r}_{\alpha i} = \frac{\sum_j B_{ij} M_{\alpha j}}{\sum_j |B_{ij}|}, \tag{3.1}$$

where $M_{\alpha j}$ is the recipe–ingredient incidence matrix introduced in Chapter 2.

Equation (3.1) takes the form of a weighted average: it aggregates the similarities between candidate i and the observed ingredients $j \in \alpha$, normalized by the total weight of i's similarity profile. This normalization prevents highly connected ingredients from dominating simply due to their ubiquity. Intuitively, a candidate ingredient receives a high score if it is strongly similar—under the chosen metric—to many of the ingredients already present.

This computation is repeated for all ingredients $i \notin \alpha$, i.e., those not already present in recipe α. The result is a vector of recommendation scores, one for each candidate ingredient, which we denote as the *recommendation vector*:

$$\vec{\hat{r}}_\alpha = (\hat{r}_{\alpha 1}, \hat{r}_{\alpha 2}, \ldots, \hat{r}_{\alpha N}), \tag{3.2}$$

where N is the total number of ingredients in the dataset. To avoid recommending ingredients that are already present in the recipe, we set $\hat{r}_{\alpha i} = -\infty$ for all $i \in \alpha$.

In practice, recommender systems typically operate on the *ranking* induced by these scores rather than their raw magnitudes. The absolute values are not directly comparable across different similarity metrics, since each metric defines its own scale and normalization. Moreover, downstream tasks such as top-k retrieval, precision, and recall depend on orderings rather than absolute values [Aggarwal, 2016]. For this reason, we sort the entries of the recommendation vector in descending order to obtain the *ordered recommendation vector*:

$$\vec{R}_\alpha = \mathrm{argsort}_{\mathrm{desc}}(\vec{\hat{r}}_\alpha), \tag{3.3}$$

where $\vec{R}_\alpha = (i_1, i_2, \ldots, i_N)$ is a permutation of ingredient indices such that $\hat{r}_{\alpha i_1} \geq \hat{r}_{\alpha i_2} \geq \ldots \geq \hat{r}_{\alpha i_N}$. The top-ranked elements of this vector—those with the highest scores—are interpreted as the most plausible ingredients to be added to recipe α, according to the system.

This framework is general: once a similarity matrix B has been obtained—whether flavor-based, recipe-based, or embedding-based—Eq. (3.1) provides a common mechanism for producing actionable recommendations and ensures comparability across methods.

Direct baselines. In addition to similarity-based recommenders, we employ two simple baselines that establish reference points for performance evaluation. The first is a **random recommender**, which assigns equal probability to all candidate ingredients not already present in the recipe. This produces a uniformly random ranking and serves as a true lower bound: any meaningful recommender must outperform chance. The second is a **popularity-based recommender**, which ranks candidates according to their overall ubiquity u_i in the dataset, independent of the specific recipe context. Its recommendation score is defined as

$$\hat{r}^{\mathrm{pop}}_{\alpha i} = \begin{cases} u_i, & i \notin \alpha, \\ -\infty, & i \in \alpha, \end{cases} \tag{3.4}$$

where $u_i = \sum_\alpha M_{\alpha i}$ denotes the frequency of ingredient i across all recipes. Because ingredient frequencies follow a heavy-tailed distribution, this naive baseline already achieves non-trivial performance, and thus sets a strong reference point. Demonstrating improvements beyond both baselines is therefore a necessary condition for establishing the added value of collaborative filtering or learned similarity.

3.3 From Hypothesis to Prediction

What we set out to learn. We ask four practical questions: (i) can flavor-compound overlap *predict* which ingredient belongs in a recipe, (ii) does this depend on region, (iii) how does it compare with signals learned from actual usage in recipes, and (iv) what mechanisms drive correct predictions in practice?

3.3.1 How We Evaluate: The Ingredient–Retrieval Task

Having defined in Sec. 3.2.3 how a similarity matrix B is turned into recommendation scores $\hat{r}_{\alpha i}$ and ranked candidate lists, we now introduce the evaluation protocol used to test predictive performance. Our aim is to mimic a practical culinary challenge: given a partially observed recipe (with one ingredient removed), can the system correctly guess which ingredient is missing?

The setup follows a standard train–test split. We randomly designate 90% of the recipes as the *training set*, denoted $\mathcal{R}_{\text{train}}$, and the remaining 10% as the *test set*, $\mathcal{R}_{\text{test}}$. For each recipe $\alpha \in \mathcal{R}_{\text{test}}$, we remove exactly one ingredient $i^\star$. The task of the recommender system is then to recover $i^\star$, given only the partially observed recipe and similarity information learned from $\mathcal{R}_{\text{train}}$.

Formally:

- M_{full}: the complete recipe–ingredient biadjacency matrix, built from all recipes in the dataset.
- M_{train}: the restriction of M_{full} to recipes in $\mathcal{R}_{\text{train}}$ (90% of the dataset).
- M_{test}: the restriction of M_{full} to recipes in $\mathcal{R}_{\text{test}}$ (10% of the dataset).
- $\tilde{M}_{\text{train}}$: a modified version of M_{full} in which exactly one ingredient has been removed from every recipe in $\mathcal{R}_{\text{test}}$, while recipes in $\mathcal{R}_{\text{train}}$ remain unaltered. This matrix is used during recommendation so that the system sees incomplete test recipes alongside complete training recipes.
- $\tilde{M}_{\text{test}}$: the test-only counterpart, obtained from M_{test} by removing one ingredient from each recipe in $\mathcal{R}_{\text{test}}$. This is the direct input representation for test recipes at evaluation time.

Because our study compares two different families of recommenders—one based on ingredient co-occurrence in recipes and the other on flavor compound overlap—the precise construction of the similarity matrix B differs, but the retrieval protocol remains identical.

Specifically, for recipe–based recommenders, the similarity matrix B is estimated from M_{train} using one of the metrics of Sec. 3.2.2. For compound–based recommenders, B is computed once from the fixed ingredient–compound network, independent of train/test splits. In both cases, test recipes $\alpha \in \mathcal{R}_{\text{test}}$ are evaluated by scoring candidates with Eq. (3.1) and ranking the resulting vector $\vec{\hat{r}}_\alpha$.

Cross-validation methodology. The ingredient–retrieval task is evaluated under a repeated cross-validation scheme. In each experimental run, we generate a new random train–test split of recipes, with 90% used for training and 10% held out for evaluation. Every recipe-based recommender is trained on the identical training set, and all recommenders–recipe-based and compound-based—-are evaluated on the same test set. In this way, any observed differences in performance can be attributed solely to the recommendation strategy itself rather than to discrepancies in data exposure.Repeating the procedure across many independent splits allows us to capture both the central tendency and the variability of each method's performance.

We stress that cross-validation is performed at the *recipe level*. Consequently, ingredient coverage across folds follows their empirical frequency distribution: ubiquitous ingredients (e.g., salt, onion) are likely to appear in both training and test sets, while rarer ingredients may occasionally be absent from training in some splits. This setup mirrors realistic recommendation scenarios, where not every item is guaranteed to be represented in every partition, and aligns with standard evaluation practices in recommender systems [Aggarwal, 2016].

3.3.2 Performance metrics: Mean Average Precision and Hit Ratio

Evaluating the quality of a recommender system requires metrics that reflect not only whether the correct ingredient is recovered, but also *how well* it is ranked among the possible alternatives. In the ingredient–retrieval task, every test recipe α has exactly one relevant target—the removed ingredient $i^\star$—and the system's output is a ranked list of candidate additions. Two natural and complementary questions then arise: (i) does the system place $i^\star$ near the top of its ranked list, and (ii) how often does $i^\star$ appear within the top portion of recommendations a cook might reasonably inspect?

To capture these aspects we adopt two widely used information–retrieval metrics: Hit Ratio at k (HR@k) and Mean Average Precision (mAP). Both metrics are standard in recommender systems and offer complementary perspectives. HR@k measures the *coverage* of correct predictions within a fixed cutoff, reflecting the probability that a user scanning only the top-k items would encounter the true ingredient. In contrast, mAP evaluates the *ranking quality* across the full list, rewarding systems that consistently place the correct ingredient earlier rather than later. Taken together, these measures allow us to assess not just whether recommenders eventually retrieve the correct answer, but also whether they do so in a practically useful way.

Formally, let $\mathcal{C}_\alpha = \{i : i \notin \alpha\}$ be the candidate set for recipe α, and let $\vec{r}_\alpha$ denote the recommendation vector of scores (Sec. 3.2.3). Sorting the candidates in descending order of score produces the ordered recommendation vector

$$\vec{R}_\alpha = \big(i_1, i_2, \ldots, i_{|\mathcal{C}_\alpha|}\big) = \mathrm{argsort}_{\mathrm{desc}}(\vec{r}_\alpha)\,,$$

and the rank of the true ingredient $i^\star$ is

$$\mathrm{rank}(i^\star;\, \vec{R}_\alpha) \;=\; \min\{k :\; i_k = i^\star\} \;\in\; \{1, \ldots, |\mathcal{C}_\alpha|\} \cup \{\infty\}.$$

Hit Ratio at k (HR@k). HR@k evaluates whether the correct ingredient appears within the top-k recommendations:

$$\mathrm{Hit@}k(\alpha) \;=\; \mathbb{I}\Big[\mathrm{rank}(i^\star;\, \vec{R}_\alpha) \le k\Big].$$

Averaging over all test recipes $\mathcal{R}_{\mathrm{test}}$ gives

$$\mathrm{HR@}k \;=\; \frac{1}{|\mathcal{R}_{\mathrm{test}}|} \sum_{\alpha \in \mathcal{R}_{\mathrm{test}}} \mathrm{Hit@}k(\alpha). \tag{3.5}$$

In our single-relevant-item setting, HR@k coincides with Recall@k. We typically report HR@20, which reflects a scenario in which a cook might reasonably examine the top twenty suggestions.

Mean Average Precision (mAP). While HR@k treats all ranks within the top-k equally, mAP differentiates between earlier and later retrieval. Precision at cut k is defined as

$$P_\alpha(k) \;=\; \frac{1}{k}\sum_{t=1}^{k}\mathbb{I}\big[i_t = i^\star\big],$$

Average Precision (AP) for recipe α is

$$\mathrm{AP}_\alpha \;=\; \sum_{k=1}^{|\mathcal{C}_\alpha|} P_\alpha(k)\,\mathbb{I}\big[i_k = i^\star\big].$$

Since there is exactly one relevant item in our setting, AP simplifies to

$$\mathrm{AP}_\alpha \;=\; \frac{1}{\mathrm{rank}(i^\star;\,\vec{R}_\alpha)}, \qquad \mathrm{AP}_\alpha = 0 \;\text{ if }\; \mathrm{rank}(i^\star;\,\vec{R}_\alpha) = \infty.$$

Intuitively, this means that if the removed ingredient is ranked first ($r_\alpha = 1$), then $\mathrm{AP}_\alpha = 1$; if it is ranked second ($r_\alpha = 2$), then $\mathrm{AP}_\alpha = 0.5$; if it is ranked third ($r_\alpha = 3$), then $\mathrm{AP}_\alpha \simeq 0.33$, and so on, decaying as the inverse of the rank position. At the opposite extreme, if the ingredient is ranked for example last among $|\mathcal{C}_\alpha| = 200$ candidates, $\mathrm{AP}_\alpha = 1/200 = 0.005$, while if it never appears in the ranking ($r_\alpha = \infty$), then $\mathrm{AP}_\alpha = 0$. Thus AP directly reflects how early in the ranking the correct ingredient appears, heavily rewarding top-ranked hits and strongly penalizing late or missing ones. Taking the mean over all test recipes gives

$$\mathrm{mAP} \;=\; \frac{1}{|\mathcal{R}_{\mathrm{test}}|}\sum_{\alpha\in\mathcal{R}_{\mathrm{test}}}\mathrm{AP}_\alpha.$$

Thus mAP rewards systems that not only find the correct ingredient, but rank it as close to the top as possible. Unless otherwise stated, we compute mAP over the entire ranking (untruncated).

In summary, HR@k provides a coarse but practical measure of success within a fixed shortlist, while mAP offers a finer-grained assessment of ranking quality across all candidates. Together, they provide a balanced view of retrieval performance in our culinary setting.

3.4 Results of the Retrieval Task

The ingredient–retrieval task is designed so that all recommenders face exactly the same challenge: in each experimental run, we generate a new random train–test split of recipes, with 90% used for training and 10% held out for evaluation, as described in 3.3. Every recipe-based recommender is trained on this identical training set, and all recommenders –recipe based and compound based– and the direct baselines are evaluated on the corresponding test set. In this way, any observed differences in performance can be attributed solely to the recommendation strategy itself, rather than to discrepancies in the data they see.

Because a single split might produce results that are idiosyncratic to that partition of the data, we repeat the entire procedure 250 times, each with an independent random split.

This repetition allows us to assess not just the central tendency of each method, but also its variability across random resamplings.

During every run, we record two key performance metrics introduced in 3.3.2 —Hit Ratio at 20 (HR@20) and mean Average Precision (mAP)—for every recommender. These values are then aggregated across iterations to produce both average scores and full performance distributions. The resulting dataset of 250 repeated evaluations per method forms the basis for the comparative analyses reported in this section, ensuring that our conclusions rest on stable statistical evidence rather than on the outcome of a single experiment.

3.5 Results of the Retrieval Task

The ingredient–retrieval task was introduced in Sec. 3.3.1, where we detailed the cross-validation procedure used to ensure fair and robust evaluation across recommenders. Here, we focus on the outcomes of this protocol.

In each experimental run, every recipe-based recommender is trained on the identical training set, while compound-based recommenders and direct baselines are evaluated on the same held-out test set. Thus, any observed differences in performance can be attributed solely to the recommendation strategy itself. To reduce the influence of any particular split, the procedure is repeated 250 times with independent random partitions. This repetition yields stable estimates of both the central tendency and the variability of each method's performance. During every run, we record the two key performance metrics introduced in Sec. 3.3.2—Hit Ratio at 20 (HR@20) and mean Average Precision (mAP)—for every recommender. Aggregating results across repetitions produces both average scores and full performance distributions. The resulting dataset of 250 evaluations per method forms the basis for the comparative analyses reported in this section, ensuring that our conclusions rest on stable statistical evidence rather than on the outcome of a single experiment.

Overall comparison: recipe vs. compound recommenders. A first, striking observation is that *recipe-based recommenders consistently and substantially outperform compound-based recommenders.* Across all similarity measures and in all 250 runs, recipe-based variants achieve higher mean Average Precision and higher Hit Ratio at 20. Figure 3.2(a–b) illustrates this point. On the x-axis we report the average performance over all iterations, while the y-axis groups methods by the underlying similarity metric. For each similarity, the performance of its compound-based version (triangle marker) and recipe-based version (square marker) are connected by a dashed line. In every case, the triangle marker is on the left of the square, highlighting a uniform advantage for recipe-derived similarities.

Beyond accuracy, recipe-based recommenders are also *more stable.* To quantify this, we compute—for each iteration—the relative rank of every method according to its mAP and HR@20. We then aggregate these ranks into distributions and visualize them as boxplots in Fig. 3.2(c–d). Each boxplot is colored by method family, and for a given similarity the boxplots of its recipe- and compound-based versions are aligned at the same x-position. On the far right we add the two direct baselines (Random and Popularity). To enrich the visualization, we layer the boxplots over the underlying per-iteration ranks, shown as a

stripplot. Finally, similarities are ordered on the x-axis by the median rank of their recipe-based version.

Two clear patterns emerge. First, the rankings of recipe-based recommenders are consistently higher than those of compound-based ones, as already suggested by the raw performance plots. Second, recipe-based recommenders also exhibit much tighter rank distributions, whereas compound-based methods show on average more dispersed ranks across iterations. This indicates that usage-based similarities provide not only higher accuracy but also greater robustness to random resampling of the data.

Direct baselines offer further perspective. Random recommendation performs worst, as expected, while the Popularity baseline achieves surprisingly strong results: it is perfectly stable across iterations, outperforms all compound-based recommenders, and even surpasses four recipe-based methods. This confirms that ingredient frequency alone is a powerful—but ultimately limited—predictor, and that more sophisticated recipe-based similarities are needed to extract finer-grained signals of compatibility.

This systematic gap reflects a fundamental difference in the information available to each family of models. Whereas compound-based recommenders rely solely on molecular similarity, recipe-based approaches learn compatibility directly from patterns of ingredient co-occurrence in observed recipes. The way ingredients are actually combined in culinary practice encodes latent compatibilities that may not be apparent from chemistry alone. By grounding similarity in usage context, recipe-based models implicitly capture not only molecular affinities but also cultural conventions, preparation constraints, and flavor–texture synergies that are difficult to formalize explicitly. These contextual signals explain why, in practice, usage-based similarity provides much stronger predictive power than compound overlap, even when the latter is refined with structural corrections such as Taxonomy or Sapling.

Detailed performance distributions. To probe performance beyond averages, we examine the full distribution of results across iterations. Figure 3.3 summarizes these outcomes using ridgeline plots, restricted to the top 14 recommenders in each family. Each ridge represents the distribution of one method's per-iteration score, normalized for comparability, with a vertical tick marking its mean. Methods are ordered top to bottom by their average performance, so that stronger recommenders appear higher on the panel. Fig. 3.3 (a) and Fig. 3.3 (b) report HR@20 and mAP for the recipe-based recommenders, while Fig. 3.3(c) and Fig. 3.3 (d) report the corresponding distributions for compound-based recommenders.

For **recipe-based recommenders**, the highest-performing similarities are RP3β and BM25 Cosine. Both stand out across mAP and HR@20, confirming their robustness. Their strong performance can be explained by their sophistication relative to raw co-occurrence counts. RP3β is a *path-based similarity* that accounts for second-order connections between ingredients (i.e., co-occurrence through shared neighbors) while penalizing overly popular items, thereby highlighting meaningful but non-trivial associations. BM25 Cosine, in contrast, is a *weighting-based similarity* that normalizes ingredient frequencies by recipe length and average recipe size, attenuating the dominance of ubiquitous ingredients and emphasizing more discriminative co-usage patterns. Taken together, these results suggest that methods which incorporate either second-order relational structure or contextual weighting

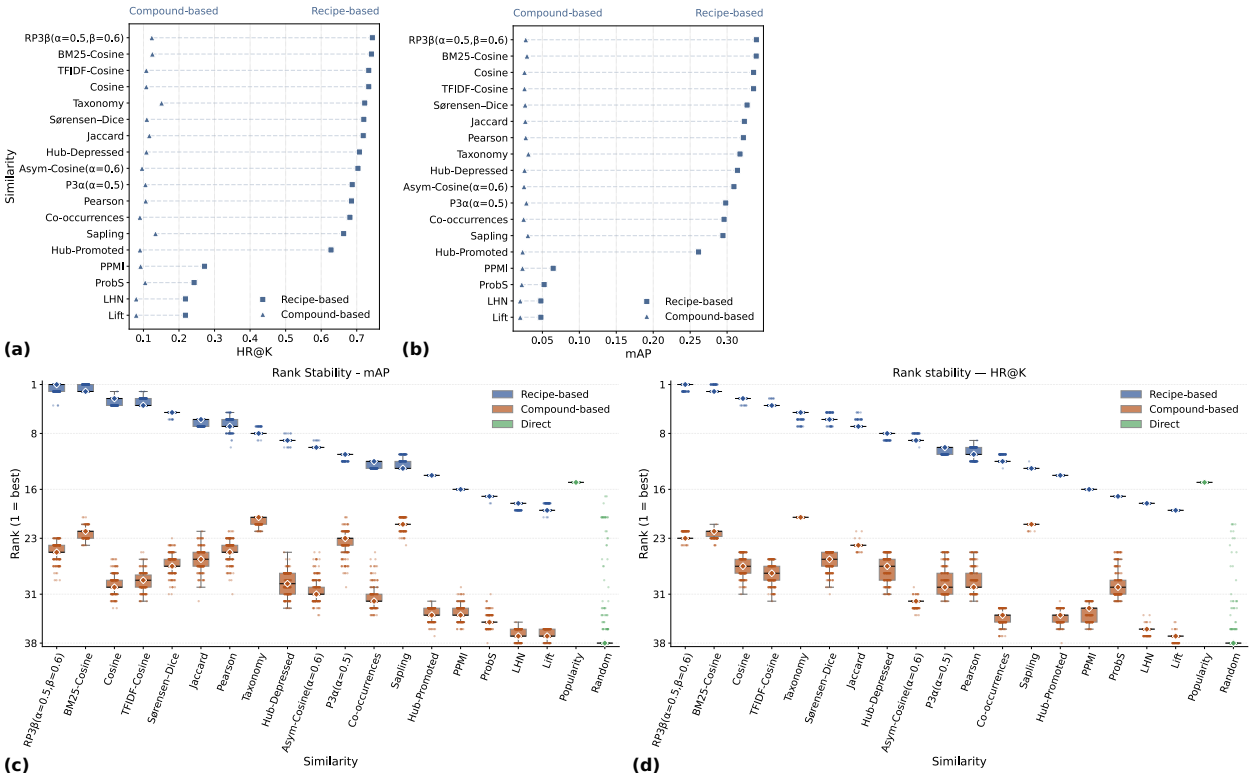


Figure 3.2: **Comparison of recipe-based and compound-based recommendation methods across performance metrics and stability.** (a–b) **Slopegraphs** show paired HR@20 (a) and mAP (b) of each recipe-based method (square markers) against its compound-based counterpart (triangle markers). Each horizontal line connects a method family, revealing that recipe-based recommenders systematically achieve higher average values than compound-based ones. (c–d) **Rank stability plots** show the distribution of per-iteration ranks for HR@20 (c) and mAP (d) (lower is better), across all methods. Boxplots are colored by recommender category: recipe-based, compound-based, and direct baselines. The tighter distributions of recipe-based methods highlight their greater robustness, whereas compound-based methods exhibit more dispersed and unstable ranks across iterations. Rankings of the direct baselines (Random, Popularity) are included for context.

capture the practical logics of ingredient usage more effectively than simpler measures.

For **compound-based recommenders**, performance is generally limited by the noisy nature of the chemical space: many flavor molecules are highly ubiquitous, appearing across a wide range of ingredients, and thus contribute little discriminative power. In this context, the top performers are those using the Taxonomy and Sapling similarities. Unlike raw overlap measures, these metrics introduce structural penalties and corrections that reduce the impact of trivial or redundant co-presence and highlight more informative patterns of compound sharing. Consistent with this refinement, Taxonomy and Sapling also achieve relatively more stable rankings—especially for HR@20, as shown in the boxplots of Fig. 3.2(c)—further distinguishing them from the noisier compound-based baselines.

Sapling Similarity incorporates a probabilistic perspective: it measures how much knowing that two ingredients share one compound changes the expectation that they will also share others. Crucially, this allows the score to be negative if the association is misleading—penalizing pairings that overlap mainly through generic compounds. By distinguishing between informative and uninformative overlaps, Sapling sharpens the weak signal embedded in the chemical network. The *Taxonomy* index provides a complementary adjustment by down-weighting the contribution of ubiquitous compounds. In practice, this favors overlaps mediated by rarer, more distinctive molecules, filtering out widespread signals that are unlikely to carry culinary significance.

This behavior is particularly important in the flavor-compound network, where a large fraction of molecules are extremely common across many ingredients. Simple similarity scores (e.g., raw cosine or Jaccard overlap) are dominated by these ubiquitous compounds and thus fail to provide predictive power. By reweighting the contribution of each compound according to its distinctiveness, Taxonomy and Sapling recover some predictive signal. Still, even with these refinements, compound-based performance remains markedly below that of recipe-based recommenders (cf. Fig. 3.2), underscoring the limited predictive power of chemical overlap alone.

Recipe-based recommenders: strengths and caveats. Recipe-based approaches learn compatibility directly from observed patterns of ingredient co-occurrence in recipes. This has the advantage of grounding similarity in real culinary practice, capturing not only molecular affinities but also cultural conventions, preparation constraints, and flavor–texture synergies that are hard to formalize explicitly.

These advantages come with potential limitations. Co-occurrence signals are inevitably influenced by ingredient popularity: ubiquitous staples such as salt, butter, or onion will co-occur with many other ingredients, risking spurious similarity inflation. As a result, naive similarity measures may over-recommend highly frequent items, diluting the specificity of predictions. This popularity bias explains why simple raw co-occurrence counts or unadjusted cosine similarity, while serviceable, leave room for improvement.

The top-performing measures in our experiments—RP3β and BM25 Cosine—succeed precisely because they address these limitations. RP3β, a path-based measure, accounts for second-order connections in the bipartite graph, leveraging richer connectivity information while dampening the influence of hubs. BM25 Cosine, adapted from information retrieval, reweights ingredient occurrences by document length (recipe size) and by inverse document

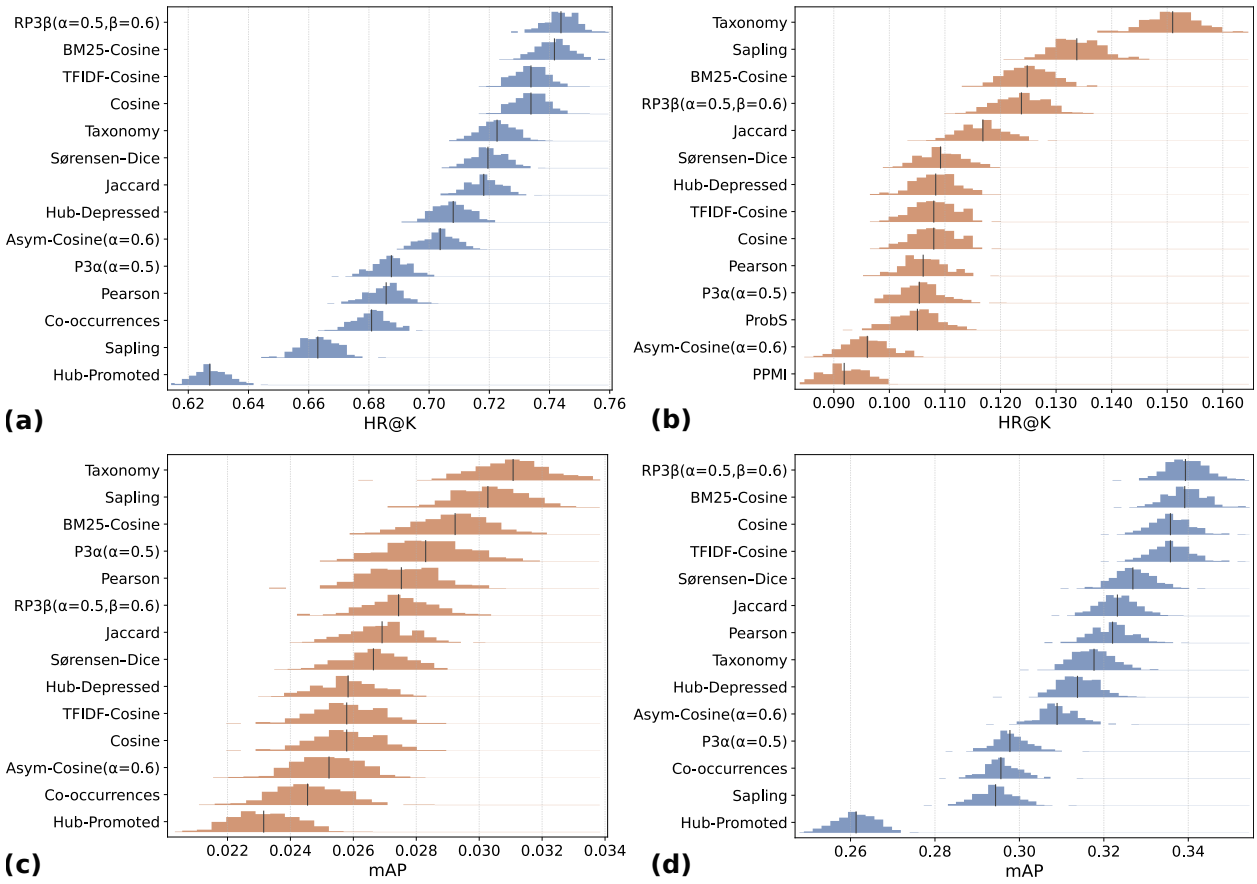


Figure 3.3: **Performance distributions across recommenders.** Ridgeline plots show the per-iteration distributions of HR@20 and mAP for recipe-based methods (a-d) and compound-based methods (b-c). Within each family, methods are ordered by their average performance, with medians marked by vertical ticks. Recipe-based similarities not only achieve higher scores on average but also exhibit narrower distributions, reflecting greater stability across train–test splits. Compound-based similarities, by contrast, show wider, noisier profiles, with only Taxonomy and Sapling emerging as relatively stronger within this weaker family.

frequency, ensuring that rare yet distinctive ingredients receive stronger weight while long recipes do not swamp shorter ones. Both methods sharpen the signal of meaningful co-usage beyond simple frequency overlap.

Taken together, these results clarify why recipe-based recommenders consistently and substantially outperform compound-based ones. Whereas flavor overlap is noisy due to the ubiquity of many molecules, co-occurrence in actual recipes naturally encodes the constraints and conventions of culinary practice. In other words, recipe-based similarities do not merely capture which ingredients *could* go together chemically, but which ingredients *do* go together in practice, balancing molecular compatibility with cultural and functional roles in the dish. This grounding in usage explains both the higher performance and the greater robustness of recipe-based recommenders across all metrics.

3.5.1 Category-restricted retrieval

A practical concern in ingredient recommendation is the prevalence of trivial substitutions. When the observed ingredients in a test recipe are, for example, three dairy items, an unconstrained recommender is likely to suggest yet another dairy ingredient—often a near-duplicate of what is already present—rather than the egg yolk, citrus, or spice that actually completes the dish. Such outcomes inflate apparent accuracy without reflecting genuine culinary compatibility.

To disentangle this effect, we adopt a *category-restricted* retrieval task. The procedure mirrors the setup of Sec. 3.3.1, with the sole difference that candidate recommendations are restricted to the *ground-truth category* of the removed ingredient $i^\star$. Thus, if the missing ingredient belongs to the category "animal product," only ingredients in this category are considered as possible candidates. This adjustment is not intended for deployment in real recommendation systems, but serves as a diagnostic to sharpen the evaluation: it removes the trivial advantage of recommending near-duplicate ingredients and yields a clearer picture of within-category pairing. We apply this restriction uniformly across all recommender families: recipe-based, compound-based, and direct baselines.

Measuring performance gains. For each recommender, we compare its performance in the category-restricted task to its performance in the unconstrained task. Specifically, for any performance metric $X \in \{\mathrm{mAP}, \mathrm{HR@20}\}$, we compute the relative percentage change

$$\Delta\% = 100 \cdot \frac{X_{\text{same-type}} - X_{\text{all}}}{X_{\text{all}}}, \tag{3.6}$$

where $X_{\text{same-type}}$ is the score under category restriction and X_{all} is the score with the full candidate set. Positive values indicate improved performance under restriction, while negative values indicate that the recommender relied on cross-category substitutions that are suppressed in this setting.

To assess uncertainty, we propagate the standard errors obtained from the 250 experimental repetitions. Let $\sigma_{\text{same-type}}$ and σ_{all} denote the standard errors of $X_{\text{same-type}}$ and X_{all},

respectively. The error on $\Delta\%$ is given by

$$\sigma_{\Delta\%} = 100 \cdot \sqrt{\left(\frac{\sigma_{\text{same-type}}}{X_{\text{all}}}\right)^2 + \left(\frac{X_{\text{same-type}}}{X_{\text{all}}^2}\,\sigma_{\text{all}}\right)^2}. \tag{3.7}$$

Visualization. We summarize these results in bar plots (Fig. 3.4), where each bar reports the average percentage change in performance across the 250 runs, with error bars reflecting $\sigma_{\Delta\%}$. To focus on the clearest signals, we display the ten compound-based recommenders with the largest average gains, the ten recipe-based recommenders with the largest average gains, and both direct baselines. Bars are ordered from highest to lowest gain, enabling a direct comparison of which recommenders benefit most from category restriction. This visualization highlights which methods are particularly prone to trivial near-duplicate substitutions, and which instead capture more robust within-category compatibilities.

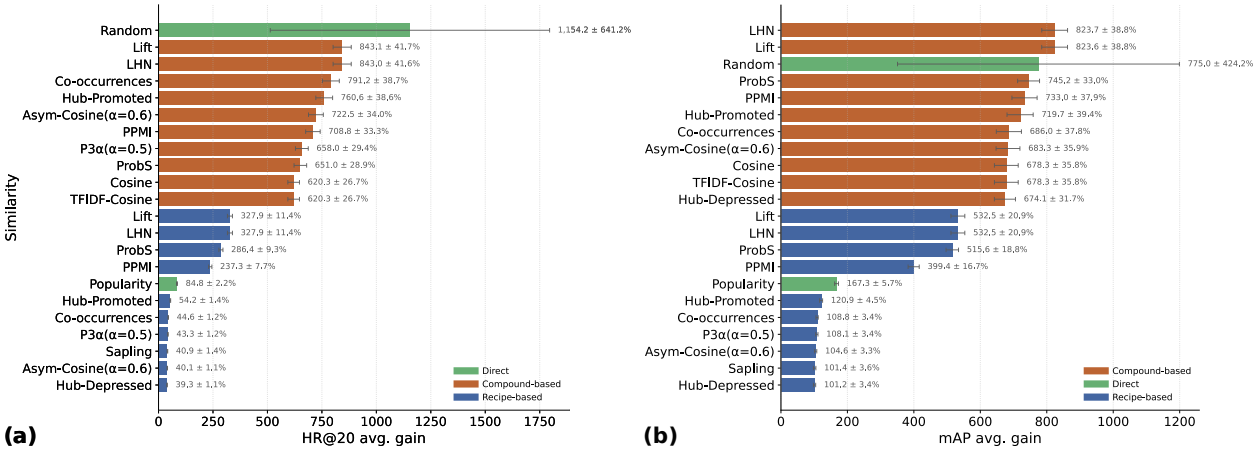


Figure 3.4: **Impact of category restriction on recommendation performance.** Bars show the percentage change in (a) HR@20 and (b) mAP when candidate recommendations are restricted to the ground-truth ingredient category, compared to the unconstrained setting. Error bars correspond to propagated standard errors (Eq. 3.7). We display the ten compound-based recommenders with the largest average gains, the ten recipe-based recommenders with the largest average gains, and both direct baselines (Random and Popularity). Recommenders are ordered from highest to lowest gain, revealing which models derive genuine benefit from within-category pairing as opposed to near-duplicate substitutions.

Figure 3.4(a–b) shows that **compound-based recommenders benefit much more from restriction than recipe-based recommenders**. Across both metrics, compound-based methods exhibit consistently larger gains with relatively small standard errors. This indicates that chemistry-driven similarities are particularly prone to trivial cross-category substitutions in the unconstrained setting, and that forcing them to compete only within the correct category reveals their limited but non-negligible ability to capture within-type affinities.

By contrast, recipe-based recommenders exhibit only modest improvements, consistent with the idea that co-occurrence in actual culinary practice already encodes type-specific

compatibilities. In other words, these methods are less reliant on trivial same-category replacements and more capable of identifying complementary roles across categories. This difference underscores the stronger contextual grounding of recipe-based similarities compared to purely molecular overlap.

The behavior of the direct baselines further clarifies these dynamics. The *Random* recommender displays the largest average gain in HR@20, as restricting the candidate pool drastically increases the chance of randomly selecting the correct ingredient. A similar effect is observed for mAP, where Random ranks among the top three methods by average gain. However, the statistical error is very large—exceeding half the mean gain in both metrics—reflecting the high variability in test-set category composition. When the removed ingredient belongs to a small category, random guessing becomes disproportionately effective, but this success is inconsistent across iterations.

The *Popularity* baseline, by contrast, shows gains that are modest compared to compound-based recommenders, confirming that its improvements are dominated by the sheer size of the restricted candidate sets rather than by any refined within-category structure. Altogether, these results indicate that category restriction sharpens the evaluation lens: it penalizes recommenders that lean heavily on cross-category substitutions, while exposing those that genuinely capture intra-category compatibility.

3.5.2 Regional breakdown

A longstanding claim in the literature is that the strength of food pairing depends on geography: the original study by Ahn et al. reported strong positive pairing in North America, but weak or negative pairing in East Asia. To test this claim within our predictive framework, we repeat the retrieval experiment separately for each world region, using the same country–region mapping as in Sec. 2.1.1. For each region, we identify the best-performing compound-based recommender in terms of average mAP and compare its performance distribution to the random baseline. This comparison is carried out both in the unrestricted retrieval task and under the category-restricted variant described in Sec. 3.5.1.

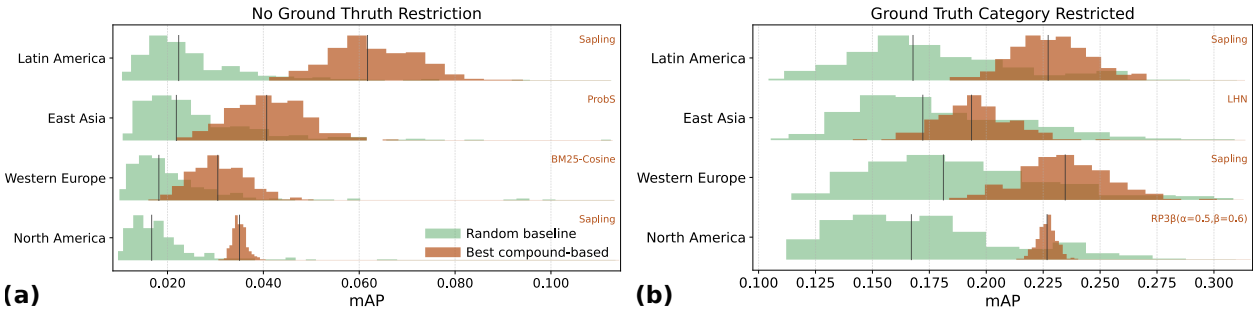


Figure 3.5: **Compound-based performance by world region.** Ridgeline plots show distributions of mAP across 250 trials for the best-performing compound-based recommender in each region (orange) compared against the random baseline (green). (a) Standard recommendation (unrestricted). (b) Category-restricted recommendation, where candidate ingredients are limited to the ground-truth type.

Figure 3.5 reports the results. In the **unrestricted case** (Fig. 3.5a), Latin America

achieves the highest average mAP, with Sapling similarity outperforming random by a clear margin. North America also shows solid improvements over random, and its distribution is notably narrow, indicating high consistency across runs even if its absolute average is not the highest. By contrast, Western Europe and East Asia display substantial overlap with the random baseline, suggesting weak or absent chemistry-driven signal.

The picture changes under the **category-restricted task** (Fig. 3.5b). Here, the Eastern European and Latin American recommenders achieve the best average performance, though their distributions are broad and still overlap with random. East Asia, in contrast, now shows the weakest results, with average mAP essentially indistinguishable from random. Once again, North America does not reach the absolute top but remains the most consistent region, with a narrow distribution centered above the random baseline. This stability across both settings suggests that North American recipes provide the most reliable ground for compound-based recommendation, even if the gains are modest in absolute terms.

Taken together, these results qualify the regional picture of chemical pairing. While Latin America and Eastern Europe can exhibit pockets of above-random predictability, their variability remains high. North America stands out not for achieving the single highest average but for producing consistently better-than-random predictions with low variance. East Asia, on the other hand, consistently fails to outperform random, in line with earlier reports of negative or negligible food pairing in this region.

Methodologically, this regional breakdown highlights both the potential and the limits of compound-based collaborative filtering. It suggests that chemistry-driven similarity can serve as a predictive probe of the food-pairing hypothesis, but its effectiveness is region-dependent and, outside of North America, often unreliable.

3.6 What drives compound-based recommendations

Global performance metrics such as mAP and HR@20 summarize accuracy but conceal the underlying mechanisms by which correct predictions arise. The modest but consistent gains of compound-based recommenders raise a key interpretive question: are these successes driven by a distributed contribution of many observed ingredients, or do they depend on a single highly similar "anchor" ingredient? To address this, we introduce a fragility analysis that quantifies how much predictions rely on the strongest individual contributor, and then complement it with an ingredient-family case study to identify which anchors are most influential.

3.6.1 Fragility check: removing the strongest anchor

We ask whether a correct (or high) rank for the true ingredient $i^\star$ is sustained by *distributed evidence* across several observed ingredients, or whether it is *fragile*, hinging on a single "anchor." Concretely, for each test recipe we (i) compute the full ranking induced by the observed ingredients; (ii) identify the single observed ingredient most similar to the target under B; (iii) recompute the scores *counterfactually* after removing only that contributor (made exact by the additivity of our scoring); and (iv) compare the target's rank before vs. after removal. Widespread upward moves (worsening) indicate anchor dependence; diagonal

persistence indicates robustness; and downward moves (improvements) can reveal hub-like effects or competition patterns.

Formally, let $M_{\text{test}} \in \{0,1\}^{n_{\text{test}} \times N}$ be the recipe×ingredient incidence matrix for the test split with the target hidden, and $B \in \mathbb{R}^{N \times N}$ the similarity matrix. Candidate scores are additive:

$$\hat{r} \;=\; M_{\text{test}} B.$$

For each test recipe α, with hidden target $i^\star(\alpha)$, we define the set of observed ingredients as

$$\mathcal{I}_\alpha = \{j : M_{\alpha j} = 1\}.$$

We identify the strongest observed contributor

$$j^*(\alpha) \;=\; \arg\max_{j \in \mathcal{I}_\alpha} B[j, i^\star(\alpha)].$$

We then compute the target's rank both in the full scoring vector

$$\text{rank}_{\text{full}}(\alpha) \;=\; \text{rank}_{\min}\big(\hat{r}_{\alpha\cdot},\, i^\star(\alpha)\big),$$

and in the counterfactual vector with $j^*(\alpha)$ removed,

$$\text{rank}_{\text{partial}}(\alpha) \;=\; \text{rank}_{\min}\big(\hat{r}_{\alpha\cdot} - B[j^*(\alpha), \cdot],\, i^\star(\alpha)\big).$$

We report results for the three strongest compound-based recommenders—Taxonomy, Sapling, and BM25-Cosine—together with their recipe-based counterparts, so as to capture a broader picture of chemistry-driven performance. Figure 3.6 shows scatter plots of $\text{rank}_{\text{full}}(\alpha)$ (x-axis) versus $\text{rank}_{\text{partial}}(\alpha)$ (y-axis) for a representative iteration. The top row corresponds to recipe-based methods, the bottom row to compound-based methods; the diagonal $y = x$ marks invariance, with points above (below) indicating worsening (improvement) after removal. Across all three similarities, removing the top contributor usually worsens the target's rank, but the effect is far more severe for compound-based recommenders. Taxonomy and BM25-Cosine display the characteristic "strip" of points above the diagonal at low initial ranks, meaning that many targets initially predicted near the top collapse once the anchor is removed. Their recipe-based counterparts lack this strip: points cluster closer to the diagonal, with only modest rank erosion among top-ranked predictions. Sapling stands out in that its recipe-based version also shows degradation among initially top-ranked targets, suggesting that this similarity induces heavier reliance on a single strong neighbor even in the recipe-based case. Another asymmetry concerns improvements: recipe-based recommenders show very few cases below the diagonal, consistent with predictions supported by multiple neighbors, while compound-based panels include a non-negligible minority of such cases, reflecting hub-like effects or signed/weaker similarities that benefit competitors more than the target. Overall, these results corroborate that chemistry-driven successes are disproportionately dependent on a single dominating ingredient, whereas recipe-based predictions degrade more gently, reflecting a more distributed support across the observed set.

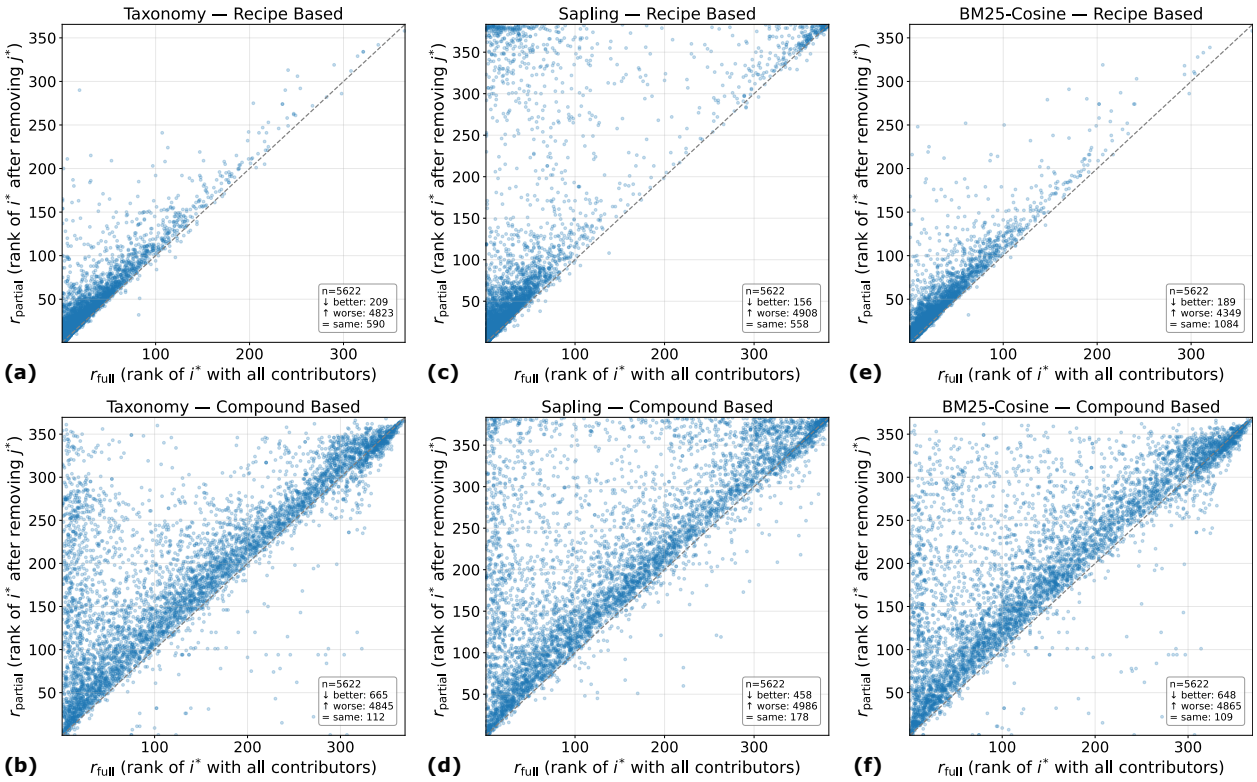


Figure 3.6: **Leave-one-influence-out fragility analysis.** Scatter plots of target rank before (x-axis) and after (y-axis) removing the most influential observed ingredient. Panels (a,c,e) show recipe-based recommenders; (b,d,f) show their compound-based counterparts. From left to right: (a,b) Taxonomy index, (c,d) Sapling, (e,f) BM25-Cosine. Across all three similarities, compound-based recommenders are markedly fragile, with strong reliance on a single anchor ingredient. Recipe-based methods degrade more gently, except in the Sapling case where top-ranked targets also suffer sensible worsening.

3.6.2 Fragility depth: beyond the top contributor

The leave-one-influence-out analysis (Sec. 3.6.1) revealed that compound-based recommenders are disproportionately fragile: removing the single strongest neighbor often causes severe rank degradation for the target, especially in cases that were initially top-ranked. A natural next step is to ask whether this fragility is confined to the very strongest contributor or whether it persists more broadly across other observed ingredients. To address this, we extend the experiment by running a *sequence of single-contributor counterfactuals*. For each test recipe we remove *exactly one* observed contributor at a time—the top-1, then the top-2, up to the top-5 most similar to the target (when available)—recompute the score vector and the target's rank, and record the resulting change. Comparing the distributions of these per-contributor rank shifts across $m = 1, \ldots, 5$ reveals whether fragility is concentrated in a single anchor or persists beyond the strongest neighbor.

Formally, for each test recipe α, with hidden target i^* and observed set $\mathcal{I}_\alpha$, we compute the full recommendation vector $s = \sum_{j \in \mathcal{I}_\alpha} B[j, :]$ and the corresponding full rank $r_{\text{full}}(\alpha)$. We then order contributors by their direct similarity to the target, $B[j, i^*]$, and denote

them $j_{(1)}, j_{(2)}, \ldots$, from strongest to weakest. For each $m \in \{1, \ldots, 5\}$ (when available) we recompute a counterfactual score vector $s^{(-j_{(m)})} = s - B[j_{(m)}, :]$, determine the new rank $\text{rank}^{(m)}_{\text{partial}}(\alpha)$, and record the rank change

$$\Delta \text{r}_m(\alpha) \;=\; \text{rank}^{(m)}_{\text{partial}}(\alpha) - \text{rank}_{\text{full}}(\alpha).$$

Figure 3.7 shows the distributions of Δr_m across test recipes, stratified by recipe size, for a single iteration of the best-performing compound-based recommender (Taxonomy) together with its recipe-based counterpart. We report the six sizes with the largest support, covering the majority of the test set. Each panel compares recipe-based (blue, left) and compound-based (orange, right) recommenders; boxplots summarize rank shifts for removal of the top-1 through top-5 contributors. The horizontal line at $\Delta r = 0$ marks stability: positive values indicate worsening when a contributor is removed, negative values improvement.

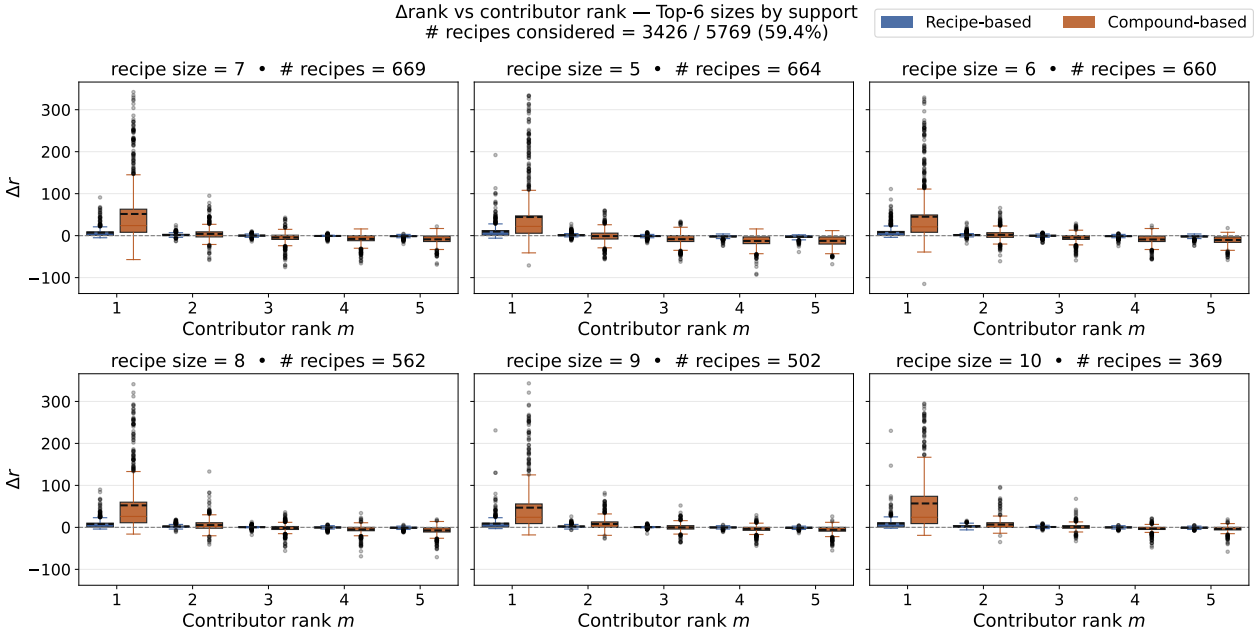


Figure 3.7: **Leave-one-contributor-out fragility across recipe sizes.** Boxplots show the distribution of rank changes Δr_m after removing the m-th most similar contributor ($m = 1, \ldots, 5$), stratified by recipe size. Results are shown for the six recipe sizes with the largest number of test recipes, with coverage and total counts reported in the figure title. Recipe-based (blue, left in each pair) and compound-based (orange, right) methods highlight their different reliance on top contributors. Numbers above panels indicate the number of recipes contributing to each size.

The results confirm that fragility is heavily concentrated in the top contributor. For recipe-based recommenders, Δr_1 is typically positive but modest, with most rank changes close to zero, and Δr_m for $m = 2, \ldots, 5$ tightly centered around zero regardless of recipe size. This indicates that predictions are supported by multiple ingredients, with only limited dependence on the single most similar neighbor. Compound-based recommenders, by contrast, show much more severe degradation: for $m = 1$, over 75% of cases yield $\Delta \text{r}_1 > 0$,

with long upper tails and occasional extreme outliers where the rank of the target collapses dramatically. These strong effects are especially pronounced for targets that were initially highly ranked, consistent with the "anchor dependence" seen in Sec. 3.6.1. For $m = 2$ through $m = 5$, the distributions are far milder and roughly centered around zero, though still broader than in the recipe-based case. Interestingly, the compound-based plots also include a minority of negative values ($\Delta r_m < 0$), indicating that in some recipes removing a contributor improves the target's rank—for example when the removed ingredient also boosted competitors more strongly than the target. Overall, these results reinforce that chemistry-driven recommenders are narrowly reliant on a single dominating contributor, while recipe-based similarities distribute support more evenly across observed ingredients.

3.6.3 Ingredient-type bridges

The fragility analyses (Secs. 3.6.1–3.6.2) showed that compound-based recommenders often rely heavily on a single anchor ingredient. A natural next question is whether these anchors tend to be *near-duplicates* of the hidden target, belonging to the same broad type category, or whether they reflect more diverse cross-type relations.

To probe this question, we examined whether the strongest observed contributor j^*—the ingredient in the recipe with the highest similarity to the hidden target i^*—shares the same type as the target. Ingredient types were taken directly from the dataset taxonomy (`type_ids`), yielding a per-recipe binary indicator of "same-type contribution." We then stratified test recipes according to the rank of the hidden target under the full model, $\text{rank}_{\text{full}}$: 1–10, 11–50, 51–200, and 201+. Within each bin we computed the prevalence of same-type contributions, and because evaluation is repeated across multiple random train–test splits, we aggregated results by reporting the mean across iterations with error bars for standard deviation. This procedure allows us to ask whether strong predictions are disproportionately driven by intra-type anchors, and how this tendency differs between recipe-based and compound-based recommenders.

The results (Fig. 3.8) reveal a sharp contrast. Compound-based recommenders display a pronounced dependence on same-type anchors, especially for high-quality predictions: when the target is ranked in the top-10, the strongest contributor shares its type in roughly 70% of cases. This prevalence declines at lower ranks, consistent with weaker or more diffuse similarity signals. Recipe-based recommenders also show an elevated prevalence of same-type anchors for top-10 predictions, but the effect is substantially weaker—around 20%—and remains nearly constant across bins.

Taken together with the fragility analyses, these results explain why compound-based methods are so vulnerable: their strongest predictions are not broadly supported by multiple ingredients, but instead hinge on trivial same-type matches that collapse once the anchor is removed. By contrast, recipe-based recommenders achieve accuracy through a more balanced mixture of same-type and cross-type signals, which distributes predictive support and makes them more robust.

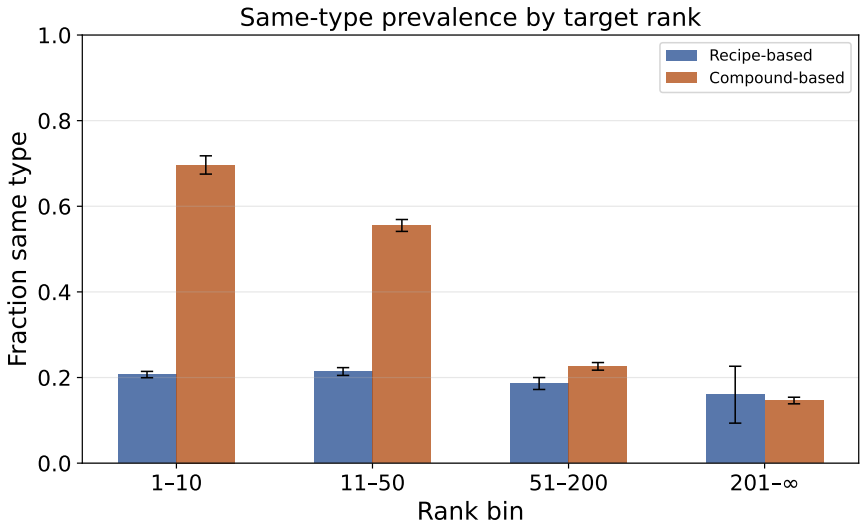


Figure 3.8: **Same-type prevalence of strongest contributors.** Fraction of cases in which the most influential observed ingredient j^* shares the same type as the hidden target i^*, stratified by target rank (r_{full}). Bars show averages across iterations for recipe-based (blue) and compound-based (orange) recommenders, with error bars indicating standard deviation. High prevalence at top ranks indicates strong reliance on intra-type anchors.

3.7 Summary and outlook

In this chapter, we examined the predictive power of different families of ingredient recommenders under a controlled retrieval task, motivated by the longstanding food pairing hypothesis. According to this hypothesis, recipes are shaped by molecular affinities between ingredients, and testing it requires going beyond descriptive overlap toward predictive evaluation. By repeatedly removing one ingredient from a recipe and assessing whether a model could recover it, we obtained a rigorous benchmark across 250 independent train–test splits, using mean Average Precision (mAP) and Hit Ratio at 20 (HR@20) to capture both ranking quality and practical hit success. Alongside recipe- and compound-based recommenders, we also included two direct baselines (Random and Popularity) as reference points. The results showed that recipe-based recommenders consistently and substantially outperformed compound-based ones in both accuracy and stability, with methods such as RP3β and BM25 Cosine gaining their advantage by incorporating second-order structure or contextual reweighting that mitigate popularity bias and highlight meaningful co-usage. Compound-based recommenders, by contrast, were hindered by noisy chemistry data dominated by ubiquitous molecules, with Taxonomy and Sapling offering only partial improvements through structural down-weighting, and the Popularity baseline even surpassing all chemistry-based methods.

Category-restricted retrieval further revealed that compound-based recommenders benefited much more from restriction, exposing their reliance on trivial cross-category substitutions, whereas recipe-based recommenders showed only modest gains, consistent with their ability to already capture within-category compatibilities. Regional analysis confirmed that chemistry-based signals are not universal: compound-based recommenders performed reli-

ably above random only in North America and, to a lesser extent, in Latin America and parts of Europe, while East Asia remained close to random even under restriction, with North America also standing out as the most stable region across repetitions.

Finally, a mechanism-level analysis clarified why compound-based successes are so fragile. Leave-one-out experiments showed that their strongest predictions almost always hinge on a single dominating "anchor" ingredient: removing it typically causes dramatic rank collapses, while contributions from other neighbors matter little. Recipe-based recommenders, in contrast, degrade more gently, reflecting distributed support from multiple ingredients. Same-type analysis further revealed that chemistry-driven anchors are overwhelmingly near-duplicates, with $\sim$70% of top-10 successes explained by ingredients of the same broad type, whereas recipe-based recommenders rely less heavily on such trivial matches ($\sim$20%) and maintain a more balanced mix of same-type and cross-type signals. Taken together, these results show that recipe-based recommenders are not only more accurate but also more robust, while compound-based methods achieve narrow, fragile wins that collapse under perturbation and rarely capture creative or non-obvious pairings.

Looking forward, a promising direction is the integration of chemical information into machine learning frameworks that can exploit richer structural signals. Graph-based recommenders such as LightGCN [He et al., 2020] illustrate how higher-order interactions between recipes and ingredients can be modeled effectively, and in this context molecular data could serve as an additional layer of enrichment to capture affinities beyond co-occurrence. At the same time, evaluation protocols should remain category-aware to avoid inflated scores from trivial substitutions, ensuring that improvements reflect genuine compatibility rather than redundancy. Expanding regional analyses will further help uncover diverse culinary logics, ultimately moving the field toward recommendation systems that move beyond near-duplicate matches to offer creative, culturally grounded ingredient suggestions that better reflect the practice of cooking.

Chapter 4

Ingredient-Type Networks as Cuisine Fingerprints

In the previous chapter, we introduced the concept of *food pairing*, the hypothesis that ingredients sharing flavor compounds tend to taste good together. This idea, first formalized and empirically tested by Ahn et al. [2011], revealed that only certain world cuisines—primarily North American and Western European—show a statistical preference for ingredient combinations with shared volatile compounds. However, as demonstrated in Chapter 3, such chemical similarity does not translate into a generally effective principle for recommending ingredients in real recipes, especially in cuisines where food pairing patterns are absent or inverted.

The food pairing hypothesis exemplifies a broader trend in contemporary food science: the focus on the chemical composition of food. Over the past decade, chemically grounded approaches have achieved remarkable success, particularly in health-related domains, nutrition science, and personalized dietary recommendations [Khoury and Ioannidis, 2014, Barabási et al., 2020]. These methods rely on a reductionist paradigm, analyzing food as a set of basic molecular constituents and reconstructing cause-effect relationships from the bottom up. While such approaches have provided critical insights into nutrition and metabolic health, they are arguably less effective in explaining the culinary logic behind ingredient use and food palatability.

Palatability is a complex phenomenon emerging from the interaction between food properties and human sensory perception. Although biological mechanisms—such as the presence of glutamate activating umami receptors [Chaudhari and Roper, 2009], or capsaicin triggering thermoreceptors [Green, 2011]—play a role in shaping taste experience, cultural context is equally important. Culinary traditions encode a variety of pairing rules, preferences, and taboos that differ dramatically across cultures. For example, mixing meat and dairy is prohibited in Jewish kosher cuisine, while it is central to iconic dishes such as cheeseburgers in the United States or beef stroganoff in Russia. Similarly, insects are an accepted protein source in many African and Southeast Asian cultures, yet are often perceived as repugnant in much of the Western world [Hartmann and Siegrist, 2015].

Food preferences are also dynamic and historically contingent. A striking example is coriander (cilantro), a herb widely used in the Mediterranean during antiquity and the Middle Ages, including in early Italian cuisine [Montanari, 2012]. Over time, its usage declined

sharply in Italy, and it is now mostly absent from mainstream Italian cooking. Moreover, many Italians report disliking its pungent aroma, which has been linked in part to genetic variation in olfactory receptors [Eriksson et al., 2012]. Environmental constraints have also historically shaped cuisines. Before globalization, geographic and climatic conditions determined the availability of ingredients. For instance, the prevalence of fermented foods in Korean cuisine—such as kimchi and doenjang—can be traced to the need for preservation in a climate with long, harsh winters and limited seasonal availability [Kim et al., 2016].

Culinary traditions can be conceptually decomposed into three primary elements: the ingredients themselves, the culinary techniques applied to them, and the recipes that emerge from their combination. To reduce the complexity of this system, one key observation proves useful: ingredients often fulfill analogous *functional roles* in recipes, regardless of their precise chemical composition. For instance, both soy sauce in East Asian cuisine and fish sauce in Southeast Asian cuisine act as umami-rich, salty flavor enhancers. Likewise, diverse fat sources—such as olive oil, butter, and lard—serve as heat conductors and carriers of fat-soluble flavors during frying or sautéing.

This insight suggests that grouping ingredients according to their *functional roles*—or more generally, by food categories—can reveal hidden organizational patterns in cuisines. Instead of analyzing ingredients as individual entities or molecules, one can examine them through the lens of *ingredient types* (e.g., Spices, Vegetables, Dairy, Cereals), which serve as proxies for culinary function. Such a coarse-grained representation not only enables meaningful cross-cultural comparisons but also abstracts away idiosyncratic differences between specific ingredients while retaining essential culinary structure.

In this chapter, we exploit this abstraction to study a global recipe dataset spanning 23 world cuisines [Bagler, 2017]. We represent each cuisine as a network of ingredient type combinations, where nodes denote types and weighted edges capture their co-use in recipes. By analyzing these networks, we uncover both universal patterns and distinctive structural fingerprints of world cuisines, identifying how cultural traditions organize and differentiate themselves through characteristic ingredient type associations.

The chapter proceeds in several steps. We begin by examining how the repertoire of recipes within a cuisine scales with its ingredient pool, revealing a tendency to reuse and recombine successful combinations rather than exhaustively exploring all possibilities. We then analyze the global hierarchy of ingredient types, identifying which categories are universally common and which remain marginal. Next, we assess the cross-cultural distinctiveness of type–type associations, contrasting globally recurrent combinations with those that are culturally specific. Finally, we show that ingredient type networks function as compact fingerprints of culinary identity: maximum spanning trees expose the minimal structural backbones of cuisines, while classification and clustering analyses demonstrate their ability to identify and group world cuisines with high accuracy.

4.1 Data and Preprocessing

To analyze cross-cultural patterns in food composition and culinary practice, we rely on *CulinaryDB* Bagler [2017], a large-scale, curated dataset of recipes collected from online sources. The dataset includes both raw and standardized ingredient information, covering

a broad spectrum of world cuisines. This section describes the structure of the dataset, the principles behind its classification system, and the preprocessing steps performed to ensure consistency, comparability, and functional interpretability.

4.1.1 Dataset Description

CulinaryDB comprises a total of 45,772 recipes, assembled by the CoSyLab research group from four major online culinary platforms: AllRecipes (16,177 recipes), Food Network (15,917), Epicurious (11,069), TarlaDalal (2,609). Each recipe is structured as a set of ingredients, and is annotated with the country or regional cuisine it belongs to. The 26 cuisines span five continents and include both single-country traditions (e.g., Italian, Thai) and broader geo-cultural regions (e.g., Central America, Scandinavia). Table 4.1 delineates the correspondence between cuisines and their respective geographical regions, providing essential context for comprehending the cultural origins and culinary traditions associated with each cuisine.

Two ingredient vocabularies accompany the dataset: one of 929 atomic (individual) ingredients, and another of 103 compound ingredients (e.g., pre-mixed spice blends). However, only 695 distinct ingredients appear in actual recipe entries. To ensure statistical reliability, we restrict our analysis to the 23 cuisines that contain at least 100 recipes. This results in a working corpus of $R = 45{,}661$ recipes and $I = 604$ unique standardized ingredients, after the preprocessing steps described below.

4.1.2 Ingredient Typing

Each ingredient is assigned to one of $T = 20$ predefined culinary *categories*, or *ingredient types*, designed to reflect shared biological origin, culinary function, and degree of processing. The types are:

Additive, Bakery, Beverages, Alcoholic Beverages, Cereals, Dairy, Essential Oils, Fish, Flowers, Fruit, Fungi, Herbs, Legumes, Maize, Meat, Nuts & Seeds, Plants, Seafood, Spices, Vegetables.

This classification, inherited from the original CulinaryDB schema, was validated through manual inspection. It follows consistent principles:

- **Biological source:** categories such as *Meat*, *Fish*, *Seafood*, *Fungi*, and *Plants* correspond to distinct biological origins.
- **Culinary function:** functional roles in cooking guide categories such as *Spices*, *Herbs*, and *Essential Oils*.
- **Processing degree:** heavily transformed items such as *Bakery*, *Dairy*, or *Beverages* are grouped accordingly.

Borderline cases are resolved by dominant culinary identity. For example, although soy sauce is a processed liquid, it is categorized under *Plants* because of its fermented soy origin and typical use as a seasoning. Similarly, honey is considered a *Plant*-based ingredient rather than an *Additive*, reflecting its culinary identity.

Table 4.1: **Cuisine-Countries mapping**

Cuisine	**Countries covered**
Indian	India, Pakistan, Bangladesh
African	Nigeria, Ethiopia, DR Congo, Tanzania, South Africa, Kenya, Uganda, Sudan, Algeria, Morocco, Angola, Ghana, Mozambique, Madagascar, Côte d'Ivoire, Cameroon, Niger, Mali, Burkina Faso, Malawi, Zambia, Chad, Somalia, Senegal, Zimbabwe, Guinea, Rwanda, Benin, Burundi, Tunisia, South Sudan, Togo, Sierra Leone, Libya, Congo, Central African Republic, Liberia, Mauritania, Eritrea, Gambia, Botswana, Namibia, Gabon, Lesotho, Guinea-Bissau, Equatorial Guinea, Mauritius, Eswatini, Djibouti, Comoros, Cabo Verde, Sao Tome & Principe, Seychelles
Chinese	China, Taiwan
Thai	Thailand
South East Asian	Cambodia, Laos, Vietnam, Philippines, Brunei, Malaysia, Singapore, Indonesia, Timor-Leste
Middle Eastern	Yemen, Oman, United Arab Emirates, Kuwait, Saudi Arabia, Jordan, Israel, Lebanon, Syria, Iraq, Iran, Turkiye, Armenia, Egypt, Qatar, Bahrain
Korean	South Korea, North Korea
Japanese	Japan
Australasian	Australia, New Zealand
Mexican	Mexico
Caribbean	Cuba, Jamaica, Haiti, Dominican Republic, Puerto Rico, Trinidad and Tobago
South American	Argentina, Bolivia, Brazil, Chile, Colombia, Ecuador, Guyana, Paraguay, Peru, Suriname, Uruguay, Venezuela
Central European	Germany, Austria, Switzerland
Greek	Greece
French	France
Misc Portuguese	Portugal
Italian	Italy
British Isles	England, Scotland, Wales, Ireland, Isle of Man, Inner and Outer Hebrides, Northern Isles
Scandinavian	Norway, Sweden, Finland, Aland Islands, Svalbard
East European	Poland, Latvia, Lithuania, Belarus, Ukraine, Moldova, Romania, Hungary, Slovakia, Czechia, Bulgaria, Russia
Spanish	Spain
American	United States of America
Canadian	Canada

4.1.3 Preprocessing Pipeline

To reduce the complexity of the dataset and improve comparability across cuisines, we applied a thorough preprocessing pipeline to the 695 ingredients initially extracted from the recipes. The goal was to merge redundant, overly specific, or inconsistently labeled ingredient entries into a more compact and semantically coherent vocabulary. As a result, we obtained a standardized set of $I = 604$ ingredients, defined according to both culinary function and ingredient identity.

The preprocessing involved the following operations:

- **Merging redundant entries:** Duplicate or near-duplicate ingredients referring to the same underlying entity were consolidated into a single standardized label. This was essential to mitigate noise arising from inconsistent naming conventions across recipes.
- **Combining ingredients with similar shape, color, or taste:** Morphological variants of staple foods—such as different types of pasta and noodles—were grouped under unified categories (e.g., `pasta`, `noodles`).
- **Merging fresh, dried, and processed forms:** Ingredients appearing in multiple preparation states were combined under a single entry. For example, `fresh peas`, `dried peas`, and `refrigerated peas` were all standardized as `peas`.
- **Generalizing ingredient subtypes:** Commercial varieties or cultivars were merged into their broader class. For instance, `cherry tomatoes`, `plum tomatoes`, and `globe tomatoes` were grouped under `tomatoes`.
- **Grouping by transformative processes:** Ingredients produced through similar culinary processes (e.g., fermentation, baking, curing) were unified under a single label representing the process outcome. Notable examples include:
  - All types of cheese (e.g., `mozzarella`, `cheddar`, `blue cheese`) → `cheese`
  - All cereal-based breads → `bread`
  - All fermented grape-based beverages (e.g., `red wine`, `white wine`, `port`) → `wine`
  - All vinegar types (e.g., `balsamic`, `apple`, `rice`, `cider`) → `vinegar`
- **Aggregating animal parts by species:** Anatomical parts from the same species were grouped into a single entry. For example, `chicken wings`, `chicken offal`, `chicken livers`, `chicken thighs`, and `chicken breast` were all merged as `chicken`.
- **Other generalizations:** Certain ingredient families were collapsed into a single entry regardless of internal diversity. For example, all types of mushroom were grouped under `mushroom`.

All preprocessing steps were conducted manually with careful domain-aware inspection to preserve culinary meaning while eliminating unnecessary granularity.

4.1.4 Formal Definitions

Let $c \in \{1, \ldots, 23\}$ denote a cuisine, and let $\mathcal{R}_c$ be the set of recipes in cuisine c. The number of recipes in cuisine c is denoted $R_c = |\mathcal{R}_c|$. Let $\mathcal{I}_c$ be the set of unique ingredients appearing in $\mathcal{R}_c$, and $I_c = |\mathcal{I}_c| \leq I$ the number of distinct ingredients in that cuisine.

Each recipe $r \in \mathcal{R}_c$ is represented as a set of ingredients, and we denote its size (i.e., the number of ingredients it contains) as n_r. Each ingredient $i \in \mathcal{I}_c$ is assigned to a type $t_i \in \{1, \ldots, T\}$. To characterize how ingredient types are used within each cuisine, we define two complementary popularity measures: one at the ingredient-token level and one at the recipe level. These allow us to quantify both how frequently a given type of ingredient appears overall in the recipes of a cuisine, and how widespread that type is across the cuisine's repertoire.

Ingredient-level popularity ($f_c(t)$). We define the **ingredient-level popularity** of type t in cuisine c, denoted $f_c(t)$, as the proportion of all ingredient tokens (i.e., individual ingredient entries across all recipes) that belong to type t. To compute $f_c(t)$, we sum the number of occurrences of all ingredients of type t across all recipes in cuisine c, and normalize by the total number of ingredients (tokens) used in that cuisine. Formally:

$$f_c(t) = \frac{\sum_{r \in \mathcal{R}_c} \sum_{i \in r} \delta(t_i, t)}{\sum_{r \in \mathcal{R}_c} n_r},$$

where $\delta(a, b)$ is the Kronecker delta function, equal to 1 when $a = b$ and 0 otherwise.

Intuitively, $f_c(t)$ estimates the probability that a randomly chosen ingredient from a recipe of cuisine c belongs to type t. This measure is sensitive to how often ingredients of a given type are used, regardless of how many recipes include them.

Recipe-level popularity ($p_c(t)$). We also define the **recipe-level popularity** of type t in cuisine c, denoted $p_c(t)$, as the fraction of recipes in which at least one ingredient of type t appears.

In other words, $p_c(t)$ measures how widespread the use of type t is across the cuisine's recipes. It is defined as:

$$p_c(t) = \frac{1}{R_c} \sum_{r \in \mathcal{R}_c} \mathbf{1}\left(\exists i \in r \mid t_i = t\right),$$

where $\mathbf{1}(\cdot)$ is the indicator function, which returns 1 if the condition inside holds, and 0 otherwise.

Put differently, $p_c(t)$ gives the probability that a randomly selected recipe from cuisine c contains at least one ingredient of type t. Unlike $f_c(t)$, which accounts for how often types are used, $p_c(t)$ captures how broadly a type is distributed across the cuisine's recipes.

These two definitions allow us to capture both ingredient abundance (how often types are used) and coverage (how widely types are used) across cuisines. They are employed in subsequent sections to identify both global trends and distinctive culinary signatures.

4.2 Statistical descriptors of world cuisines

In this section, we provide a statistical overview of the global recipe dataset described in Section 4.1, focusing on basic macroscopic descriptors of culinary diversity across cultures. Our aim is to identify large-scale patterns in ingredient usage and recipe structure that characterize and differentiate regional cuisines.

4.2.1 Cuisines as Sets of Recipes

Each cuisine in the dataset can be conceptualized as a set of recipes composed of a variable number of ingredients, each assigned to one of 20 predefined ingredient types (see Section 4.1). The number of recipes (R_c) and unique ingredients (I_c) varies substantially across cuisines, reflecting differences in cultural complexity, ecological biodiversity, and the depth of data collection from online sources. Recipe counts span nearly three orders of magnitude—from as few as 138 to over 16,000—reproducing earlier findings on culinary diversity and heterogeneity [Kinouchi et al., 2008, Ahn et al., 2011].

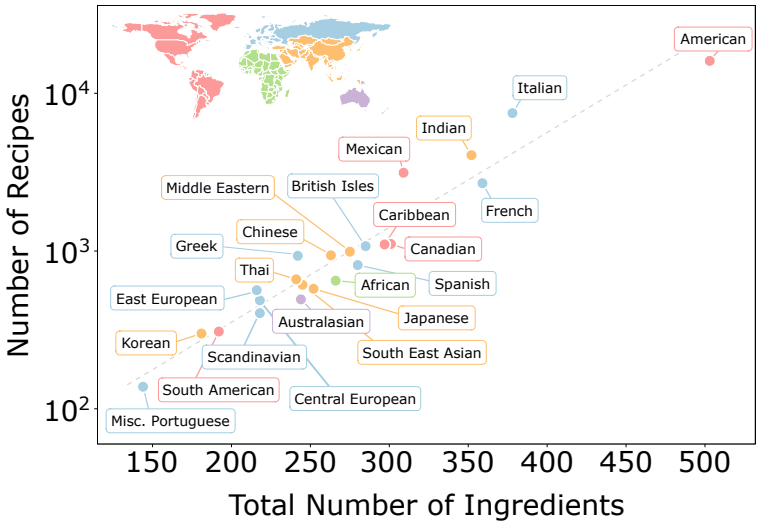


Figure 4.1: **Scaling of recipes vs total ingredients across world cuisines.** Relationship between the number of recipes and the total number of individual ingredients across various cuisines. Each data point represents a cuisine and is color-coded according to the continent affiliation, as illustrated in the world map reported in the top-left corner. A lin-log scale has been used in the plot, and the reported best-fit line highlights the exponential growth of the number of recipes with the total number of ingredients. *Figure reproduced from Caprioli et al. [2024].*

Figure 4.1 shows the scaling behavior between the number of distinct ingredients (I_c) and the number of recipes (R_c) across cuisines. While one might expect a linear relationship—i.e., more ingredients lead to proportionally more recipes—the empirical data suggest otherwise. The best-fit curve reveals an exponential growth: cuisines with richer ingredient repertoires tend to exhibit a disproportionately larger number of recipes. The relationship

is well described by $R_c \sim \exp(\gamma I_c)$, with a fitted exponent $\gamma \approx 0.1$ and a coefficient of determination $R^2 = 0.77$.

This finding underscores how culinary diversity is amplified through combinatorial possibilities. The presence of more ingredients does not merely allow more recipes but encourages the reuse and recombination of elements in increasingly complex ways.

The scaling relationship naturally partitions cuisines into two groups: those that generate more recipes than expected based on their ingredient pool (i.e., those lying above the fit line) and those that produce fewer (i.e., those lying below it). For instance, Italian cuisine stands out as an over-performing cuisine: with 378 ingredients, it yields 7,493 recipes—far exceeding the 4,300 recipes expected from the scaling law. In contrast, Portuguese cuisine exhibits underperformance: despite having 144 ingredients, it only features 138 recipes, placing it well below the scaling curve.

4.2.2 Recipe size distributions

Another important descriptor of a cuisine is its recipe size distribution, i.e. the probability $P_r(n)$ that a randomly selected recipe contains n ingredients. Figure 4.2 displays ridge plots of recipe size distributions for all 23 cuisines, grouped and color-coded by continent.

A strong continental signal emerges. European cuisines, particularly Southern and Central European traditions, tend to favor shorter recipes, with modal values ranging from 5 to 9 ingredients. In contrast, South and East Asian cuisines frequently feature larger recipes, often containing 12 or more ingredients. These patterns reflect cultural conventions in cooking practices, seasoning habits, and ingredient layering.

The recipe with the largest number of ingredients in the dataset—`Vegetable Korma`, an Indian dish—contains 31 distinct components, exemplifying the complexity often involved in traditional Indian cooking.

4.2.3 Ingredient Type Popularity

A further step toward characterizing world cuisines involves analyzing how different ingredient types are used within each culinary tradition. While previous sections have focused on the number and size of recipes and the diversity of their ingredients, here we turn to how frequently each of the $T = 20$ predefined ingredient types appears across cuisines.

We define the *popularity* of an ingredient type t in a given cuisine c, denoted $f_c(t)$, as the proportion of ingredient tokens in all recipes of c that belong to type t. This metric, introduced formally in Section 4.1, allows us to build a profile of ingredient usage for each cuisine that is independent of the cuisine's total number of recipes or its average recipe size.

Figure 4.3a presents a heatmap summarizing the ingredient type popularity profiles across all 23 cuisines. Each row corresponds to a cuisine, and each column to an ingredient type. Darker colors indicate higher values of $f_c(t)$, reflecting more frequent use of that type within the cuisine's recipe corpus. Overall, we observe a common structure across cuisines: a core group of ingredient types—including *Vegetables*, *Spices*, *Dairy*, and *Cereals*—tends to be widely used across traditions. In contrast, other types such as *Flowers*, *Essential Oils*, or *Additives* appear less frequently, leading to lighter color patterns on the right-hand side of the heatmap.

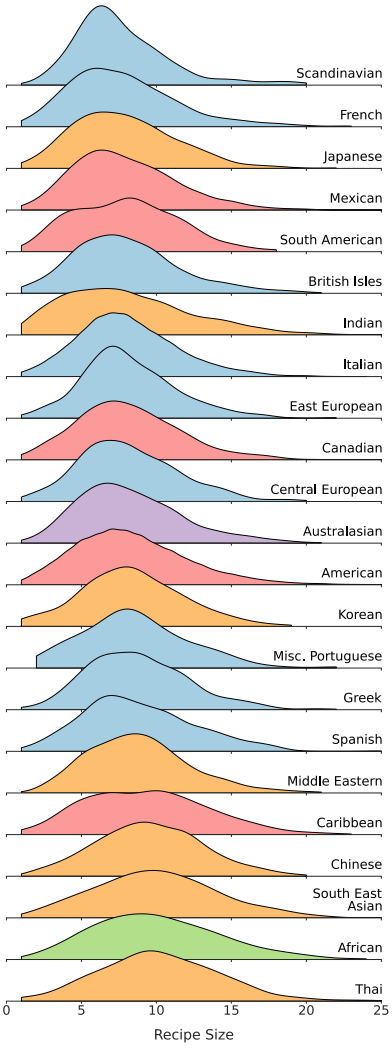
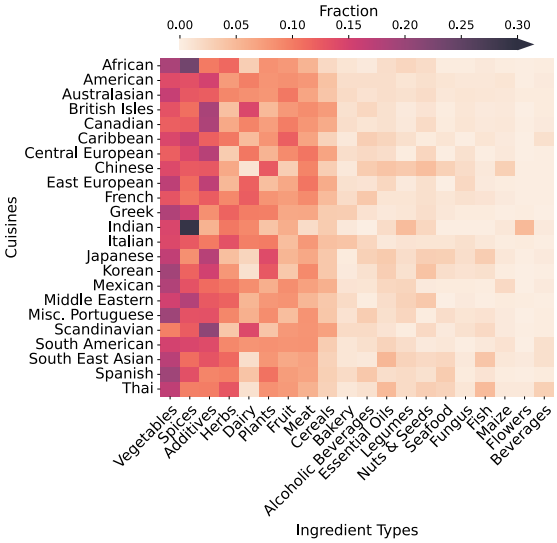
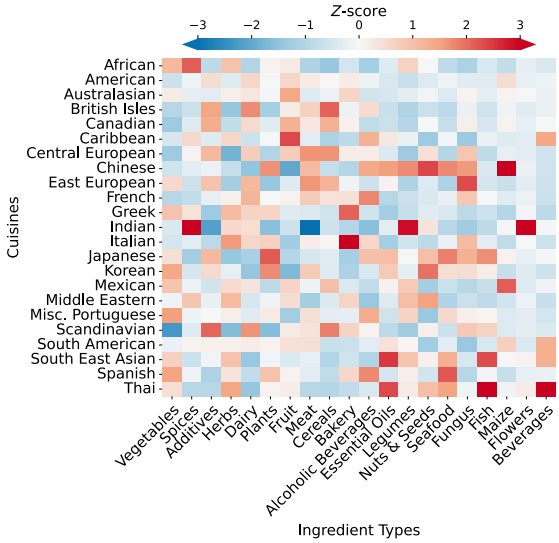


Figure 4.2: **Recipe Size Distribution.**Ridge plot showing the empirical probability distribution of recipe sizes across cuisines from different continents. Each distribution, sorted by average recipe size, corresponds to a cuisine, and is colored according to the continent affiliation (as per fig. 4.1). *Figure reproduced from Caprioli et al. [2024].*

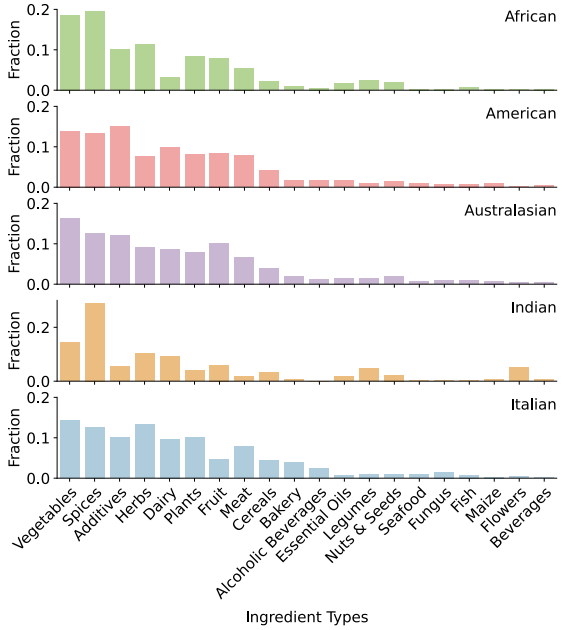


(a) **Ingredient type popularity profiles.** Each row corresponds to a cuisine, and each column to an ingredient type. Values represent $f_c(t)$ — the fraction of ingredients in the cuisine belonging to type t. Darker colors indicate higher popularity.

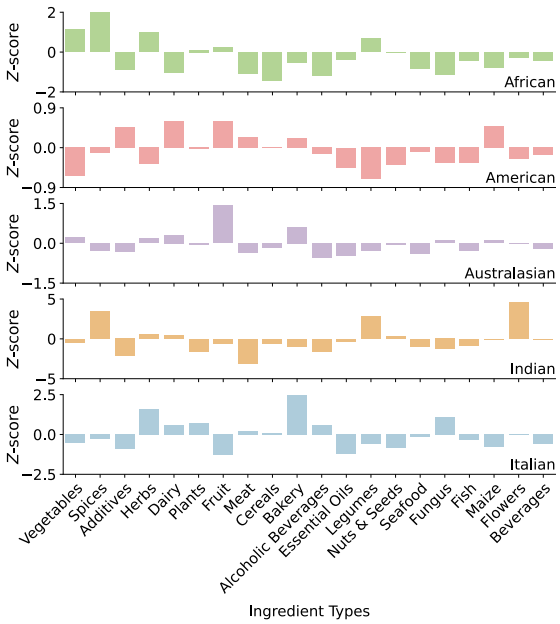


(b) **Z-score normalized ingredient usage.** Same structure as (a), but values are z-scores of popularity across cuisines. Red indicates overuse, blue underuse relative to the global average.

Figure 4.3: **Cross-cultural variation in ingredient type usage.** Comparison of absolute popularity (a) and relative over/under-use (b) of ingredient types across 23 world cuisines. *Figure reproduced from Caprioli et al. [2024].*



(a) **Ingredient type popularity for selected cuisines.** For each selected cuisine, the bars show the fraction of ingredient tokens (i.e., $f_c(t)$) that belong to each of the 20 ingredient types.



(b) **Z-score profiles for selected cuisines.** The bars represent the standardized z-score of each ingredient type within the given cuisine, highlighting distinctive overuse or underuse.

Figure 4.4: **Ingredient usage profiles in selected cuisines.** Popularity values (a) and z-score normalized values (b) for representative cuisines. *Figures reproduced from Caprioli et al. [2024].*

Despite this global pattern, clear culinary idiosyncrasies emerge. Figure 4.4a illustrates this variability in more detail, showing ingredient type popularity for five representative cuisines (we report the popularity profiles of the remaining cuisines in Fig.A.1 in Section A.1 of the Appendix). The differences highlight how culinary traditions emphasize distinct ingredient categories: for instance, *Herbs* and *Spices* are predominant in Indian and Middle Eastern cuisines, whereas Scandinavian and British Isles rely more heavily on *Dairy*, *Bakery*, and *Meat* products.

To further emphasize distinctive usage patterns, we compute the z-score of the popularity of each ingredient type across cuisines:

$$z_c(t) = \frac{f_c(t) - \mu_t}{\sigma_t},$$

where μ_t and σ_t denote the mean and standard deviation of $f_c(t)$ across all cuisines. The resulting standardized profiles, visualized in Figure 4.3b, highlight ingredient types that are used significantly more (red) or less (blue) than average in each cuisine. Distributions of z-scores for five selected cuisines are shown in Figure 4.4b.

Several culturally and geographically grounded signals are revealed by this analysis. For example, Indian cuisine displays a marked underuse of *Meat*, consistent with prevalent dietary restrictions tied to religious practices. At the same time, it overuses *Spices*, *Legumes*, and *Flowers*, the latter often appearing in the form of floral oils or extracts used in traditional cooking. In contrast, Scandinavian cuisine shows strong negative z-scores for *Vegetables*, *Plants*, and *Herbs*, a trend plausibly linked to climatic constraints, which have historically limited access to fresh produce and shaped a culinary tradition more reliant on preserved or alternative ingredients. These observations provide a quantitative basis for well-known cultural intuitions about regional food practices and reinforce the interpretability of ingredient type usage as a lens for comparing culinary systems.

Summary

At a macroscopic level, ingredient counts and recipe counts obey an exponential scaling law that partitions cuisines into combinatorially "rich" and "conservative" classes. Zooming to the recipe level, European cuisines lean toward shorter ingredient lists, Asian toward longer. Finally, type-level popularity profiles expose culturally and environmentally driven preferences. These regularities motivate a relational view: rather than studying types in isolation, we next model how types co-occur within recipes, using ingredient-type networks.

4.3 The Networks of Ingredient Combination

In previous sections, we have examined cuisines by analyzing how ingredients are used in isolation. Measures such as ingredient frequency, type popularity, and recipe size are all *zeroth-order* properties: they describe the prevalence of individual components, but do not capture how these components are brought together in recipes. Yet, the *ingredient combination patterns*, that is, the ways in which ingredients are functionally combined, are at the

core of recipe creation. Two cuisines may draw on a similar repertoire of ingredient types yet assemble them in profoundly different ways, giving rise to distinct culinary identities.

To investigate this dimension, we adopt a network-based approach that captures how ingredient types co-occur within recipes. An analysis at the level of individual ingredients would be difficult to compare across cuisines: some cuisines rely on a large and diverse ingredient set, while others use a smaller repertoire; moreover, ingredients present in one culinary tradition may be entirely absent in another. By instead working with a coarse-grained representation based on the 20 predefined ingredient types described earlier, we obtain a uniform framework that enables direct cross-cultural comparison and highlights the characteristic patterns of ingredient interaction.

In the following subsections, we introduce and analyze the networks of ingredient combinations for each cuisine. We explore both their raw structure and their distilled "backbone" representations, assess deviations from cross-cultural norms, and uncover core pairings that define each culinary tradition. We also propose minimal representations of these networks in the form of maximum spanning trees, which act as distinctive fingerprints of how ingredients are typically combined within different world cuisines.

4.3.1 Constructing the Type-Type Networks

To study how cuisines combine different classes of ingredients, we build networks that encode co-occurrence patterns at the level of ingredient types. Each cuisine is represented by a weighted, undirected graph, where nodes correspond to the $T = 20$ predefined ingredient types, and edges reflect the frequency with which pairs of types appear together in recipes.

Working at the level of ingredient types offers a twofold advantage. First, it reduces the complexity of ingredient-level graphs, which can be extremely dense and difficult to interpret visually. Second, it provides a fixed-size, uniform node set for all cuisines, enabling direct structural comparisons while retaining the essential patterns of how broad ingredient categories interact.

Formally, for each cuisine c, we map every ingredient in its recipes to its assigned type $t \in T$. Two types t and t' are connected if there exists at least one recipe containing at least one ingredient of type t and one of type t'. The edge weight $w_c(t, t')$ is defined as the total number of co-occurrences of type-t and type-t' ingredients across all recipes in c. Recipes containing multiple ingredients of the same type contribute multiplicatively to the weight.

For example, the *apple strudel* recipe from Central European cuisine contains four ingredients: apples, raisins, butter, and puff pastry. Mapping these to types yields *Fruit* (apples), *Fruit* (raisins), *Dairy* (butter), and *Bakery* (puff pastry). This recipe generates the following edges: Fruit–Fruit, Fruit–Dairy, Fruit–Bakery, and Bakery–Dairy. In particular, the edges Fruit–Dairy and Fruit–Bakery each receive a weight contribution of 2, reflecting the fact that both butter and puff pastry are paired with two distinct fruits.

More generally, for each recipe r in cuisine c, we form the multi-set [Ouvrard et al., 2018] of type pairs (t_i, t_j) for every unordered pair of ingredients i, j in r. We then compute the co-occurrence frequency $n(t_i, t_j)$ of each type pair across all recipes, and assign this value to the weight w_{t_i,t_j} of the undirected edge (t_i, t_j). In addition, each node t is enriched with an attribute v_t representing its *relative abundance*, defined as the fraction of recipes in the cuisine that contain at least one ingredient of type t.

4.3.2 Network visualization

Type–type networks are much denser than might be expected. With $T = 20$ nodes, the full graph can have up to $T(T-1)/2 = 190$ undirected edges (210 edges including self-connections). Because many recipes contain a wide range of ingredient types, real-world type–type graphs typically approach this upper bound. Across the 23 cuisines, networks exhibit an average density of 0.93, with a median density of 0.94. The highest density is observed in the American cuisine (Figure 4.5), where all possible pairs of types are connected, yielding the full set of 210 edges. The lowest density is found in Misc. Portuguese (0.74), a consequence of the limited number of recipes available. With the exception of this outlier, all cuisines have densities above 0.85, underscoring the strong tendency for type pairs to be connected.

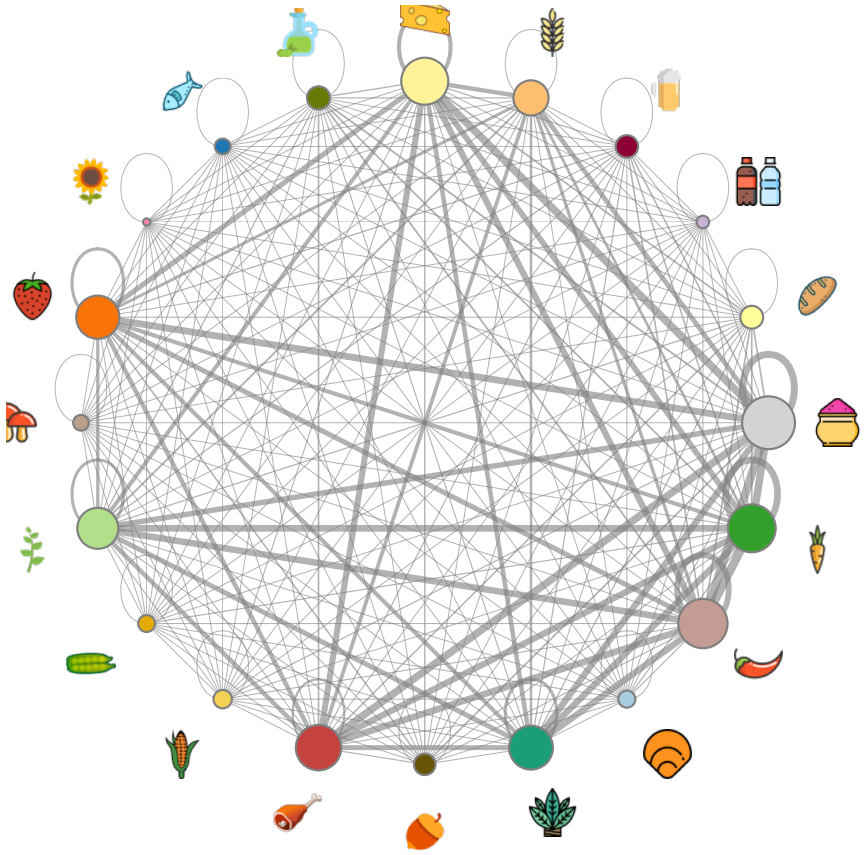


Figure 4.5: Unfiltered type–type graph of the American cuisine. Node sizes are proportional to relative abundance v_t, and edge thicknesses to co-occurrence weights w_{t_i,t_j}. he complete connectivity of the American type graph exemplifies the need for filtering methods to make structural patterns visible.

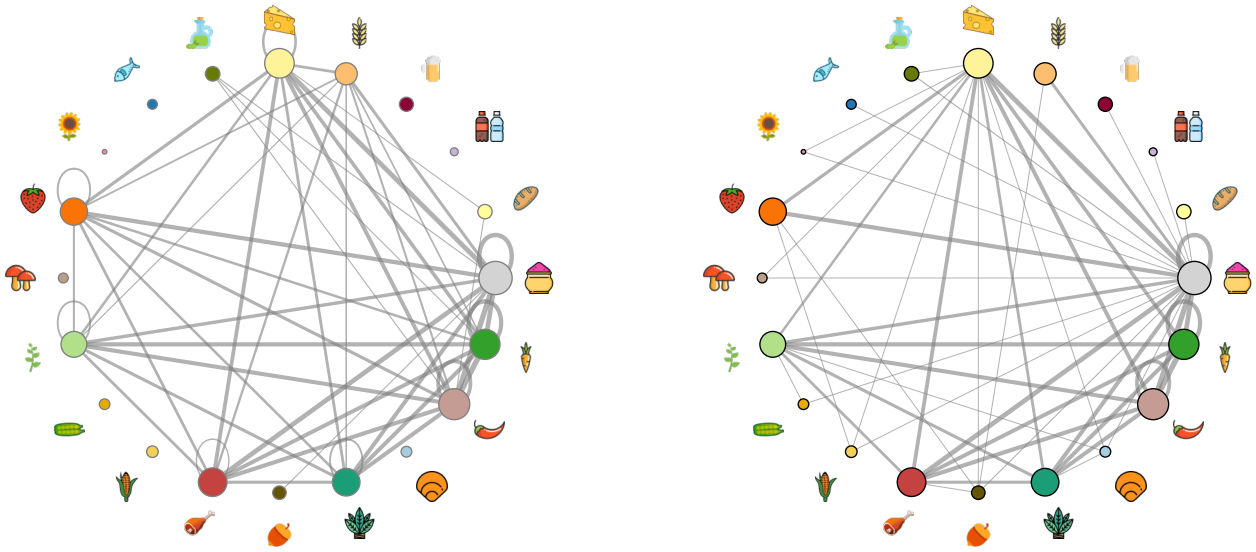
This near-complete connectivity makes it difficult to identify dominant pairings or meaningfully compare structures across cuisines. To address this, we apply two complementary filtering strategies that reduce network density while preserving the most informative associations.

Filtering strategy 1: fixed edge count. The first approach consists of retaining a fixed number K of the strongest edges (those with the largest weights) in each network. This strategy has two advantages: (i) it highlights the most frequently used type pairings within a cuisine; and (ii) it enables visual comparisons between cuisines while keeping the number of displayed edges constant. In particular, by enforcing the same number of retained edges across all cuisines, this approach ensures that observed structural differences are attributable to culinary choices rather than to trivial differences in graph density or size. A limitation of this method, however, is that it may discard important edges in very dense cuisines, while in sparse cuisines it may overemphasize relatively weak pairings simply to reach the fixed quota.

Filtering strategy 2: backbone extraction. The second approach applies a statistical backbone extraction method, specifically the disparity filter [Serrano et al., 2009]. This method assesses the statistical significance of each edge weight relative to a null model in which a node's total strength is uniformly distributed among its edges. Only edges with a p-value below a chosen threshold are retained. Unlike the fixed-K approach, this method preserves edges that are unexpectedly strong given a node's connectivity, even if their absolute weight is modest. This helps reveal culturally distinctive pairings that may not be among the globally most common. A limitation of the backbone approach is that it may remove globally frequent pairings that, although not statistically exceptional, are nevertheless central to a cuisine's identity. In addition, results can be sensitive to the choice of significance threshold, with too strict or too lenient values leading to under- or over-pruning. Thus, while backbone filtering is effective in uncovering *surprisingly strong associations*, it may obscure some of the *universally common pairings* that play an equally important role in defining culinary practices.

In practice, these two strategies produce noticeably different visualizations. For example, in the American cuisine (Fig. 4.6), the fixed-K representation highlights only the strongest type–type pairings, whereas the backbone filtering recovers a richer set of weaker but statistically meaningful links that would otherwise be discarded. Both filtering strategies are therefore complementary: the fixed-K approach emphasizes the most common type pairings in each cuisine, while the backbone method highlights unexpectedly strong pairings relative to the cuisine's overall structure. In addition to edge weights, the networks also encode information at the node level through the relative abundance v_t of each type. This attribute interacts differently with the two filtering strategies. In the fixed-K approach, highly abundant types (large v_t) are more likely to dominate the retained edges, as their ubiquity increases the chance of co-occurrence with many others. By contrast, the backbone method emphasizes pairings that are disproportionately strong relative to expectations, which can elevate less abundant types if they consistently co-occur with specific partners. This contrast between "absolute prominence" and "relative distinctiveness" provides two complementary perspectives on the combinatorial logic of different cuisines.

In the next subsection, we extend this comparison across all cuisines and quantify deviations from cross-cultural norms.



(a) Fixed-K network (American cuisine). (b) Backbone network (American cuisine).

Figure 4.6: **Comparison of filtering strategies for the American cuisine**. The fixed-K network (a) showcases only the strongest links, producing a compact view where weaker associations are suppressed. The backbone network (b), in contrast, retains many lower-weight links that are statistically significant given local connectivity, bringing to light finer-grained structures that are hidden by rank-based filtering. Node sizes are proportional to relative abundance v_t, and edge thicknesses to co-occurrence weights w_{t_i,t_j}.

4.4 Cross-cultural analysis of ingredient type usage

Having introduced the filtering strategies that allow us to highlight meaningful structures in type–type networks, we now turn to a comparative analysis across cuisines. This section develops along four complementary directions: (i) visual inspection of backbone-filtered networks, (ii) assessment of the global prevalence of ingredient types, (iii) identification of types that are statistically distinctive to each cuisine, and (iv) analysis of the strongest type pairings that shape culinary identities.

4.4.1 Visual analysis of filtered type–type networks

We begin with a qualitative comparison of filtered type–type networks. Building on the filtering strategies introduced in §4.3.2, we consider two complementary views: (i) retaining a fixed number of top-weight edges per cuisine (fixed-K) and (ii) extracting the statistical backbone via the disparity filter. As discussed earlier, the former foregrounds the *most frequent* pairings in each cuisine under a controlled edge budget, while the latter highlights *surprisingly strong* pairings relative to local connectivity patterns.

Fixed-K networks: most frequent pairings. Figure 4.7 displays the fixed-K networks (top 50 edges by weight) for eight representative cuisines. Enforcing the same edge count across panels makes structural differences attributable to culinary organization rather than

graph size.

A striking resemblance emerges among the American, Australasian, and Canadian networks: all three retain broad, multi-hub backbones dominated by *Vegetables*, *Spices*, and *Additives*, with extensive cross-type connectivity. A plausible interpretation is that these cuisines, shaped by a common Anglosaxon cultural basis and reinforced by large-scale immigration, integrate diverse culinary influences into a comparatively homogenized pairing structure.[1]

By contrast, Indian and Japanese networks exhibit more distinctive structures. Indian cuisine is dominated by *Spices* and vegetarian-related types, while Japanese cuisine shows a balanced core involving *Vegetables*, *Seafood*, and *Essential oils*. Notably, all dairy-related connections are pruned in the Japanese case, whereas they persist in every other cuisine shown. This reflects the limited diffusion of dairy in Japan, consistent with the high prevalence of lactose intolerance in East Asian populations.

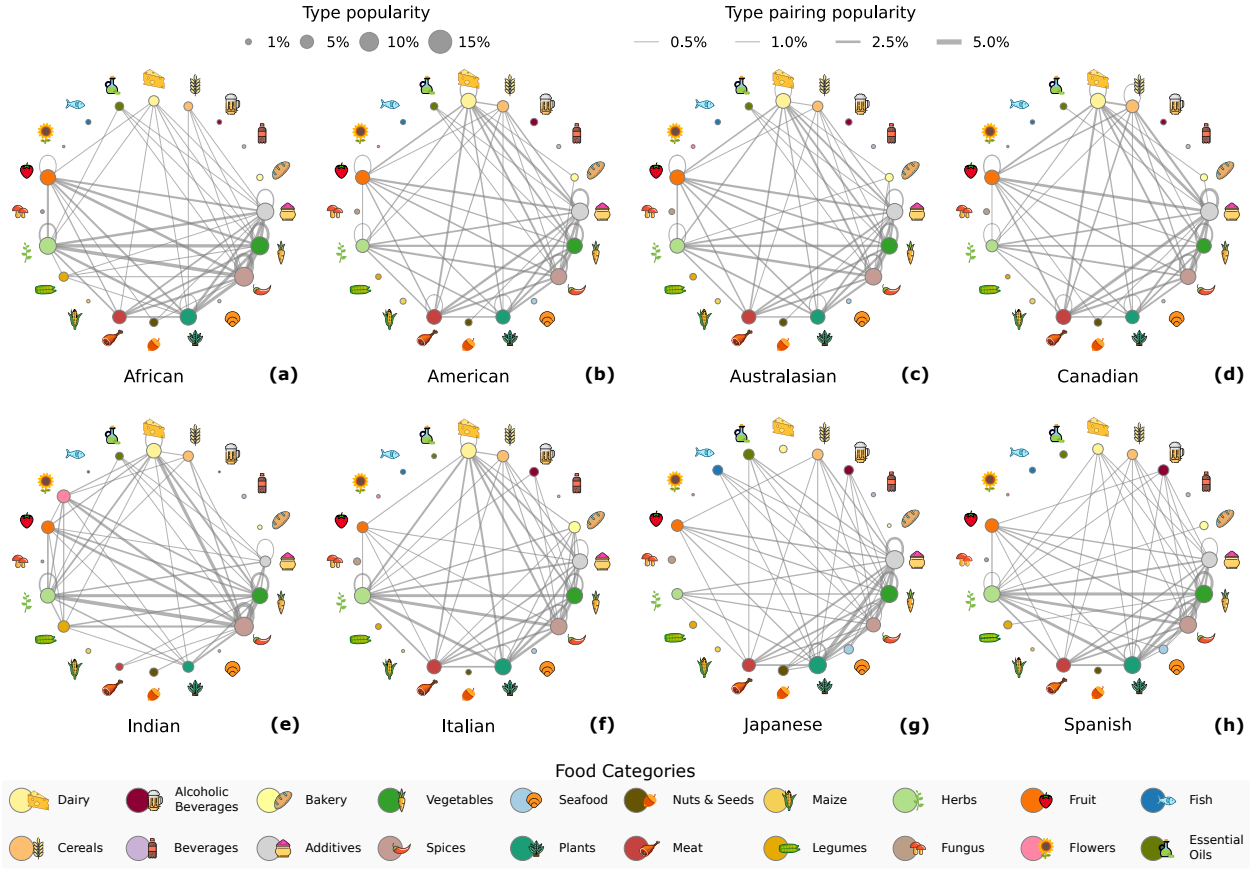


Figure 4.7: **Fixed-K type–type networks (top 50 edges).** Node size: relative abundance v_t; edge thickness: co-occurrence weight w_{t_i,t_j}. Using a common edge budget foregrounds the most frequent pairings per cuisine while enabling controlled visual comparison across panels.

[1] We return to this point quantitatively below by comparing edge-weight profiles and their dispersion across cuisines.

Backbone networks: disproportionately strong pairings. Figure 4.8 shows disparity-filter backbones for the same set of cuisines (for a display of all cuisines see Fig. A.3 in Section A.2 of the Appendix). Unlike fixed-K, the number of retained edges adapts to local heterogeneity; edges are preserved when their weight is unexpectedly large given a node's connectivity.

Backbones sharpen cuisine-specific signatures. In Indian cuisine, *Spices* remain a dominant hub, but *Dairy* emerges with the highest number of retained links, underscoring its central role when connections are evaluated by relative strength rather than absolute frequency. The Italian and American backbones both exhibit balanced structures with multiple hubs (*Vegetables*, *Dairy*, *Herbs*), though with subtle differences in bridging patterns. The Japanese backbone reveals a compact, marine- and plant-centered core, but also retains many additive-related edges. This likely reflects the filter's emphasis on relative weight: additives frequently co-occur with diverse types, so their edges survive pruning despite being modest in absolute frequency.

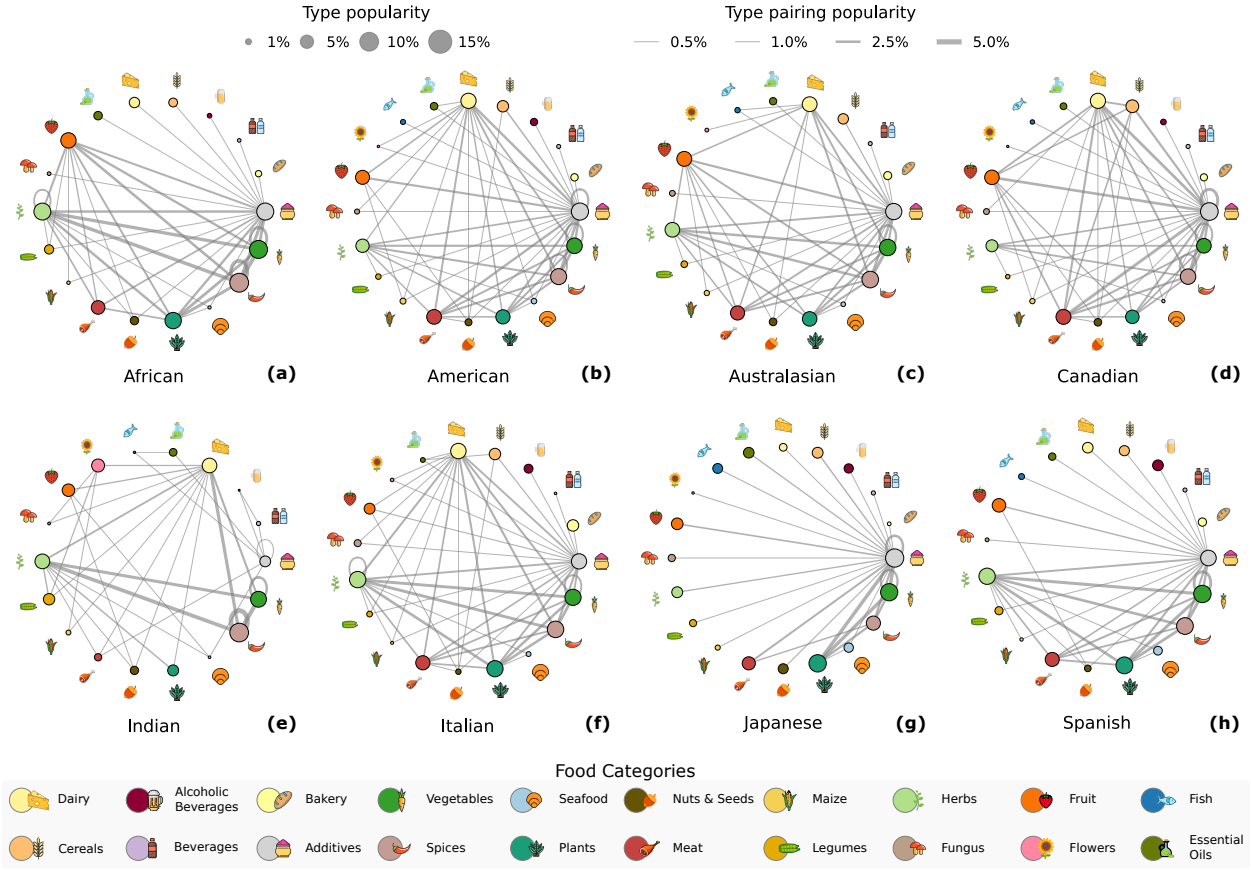


Figure 4.8: **Backbone type–type networks (disparity filter).** Edges are retained when their weight is statistically significant relative to a null in which a node's strength is uniformly distributed among its incident edges.

Complementarity. Side-by-side, the two filters offer different lenses: fixed-K reflects the *absolute* hierarchy of pairings (most frequent edges), while the backbone emphasizes *relative*

salience (edges unusually strong given local mixing). In practice, frequent "workhorse" pairings (e.g., *Vegetables* with *Spices/Additives*) dominate fixed-K panels across many cuisines, whereas backbones surface cuisine-specific associations that would otherwise be visually drowned by ubiquitous high-weight edges.

4.4.2 Cross-cultural distinctiveness of type usage

So far, our analysis has focused on the organization of type–type pairings within cuisines, centered on a visual exploration of two different filtered network representations. We now shift perspective to the prevalence of the ingredient pairings themselves, asking whether there exist combinations that are globally central and others that remain more marginal. One of our aims is to reconstruct a unified picture of world cuisine by comparing how frequently different culinary traditions employ the same type pairs. This global baseline can then be used as a reference point when assessing the distinctive features of individual cuisines.

To address this question, we leverage the weighted networks introduced earlier. For each type pair in a given cuisine, we assess its significance by comparing its weight to the corresponding distribution of weights for the same pair across all other cuisines. In this way, we evaluate how strongly a given pairing is used relative to a global baseline. The degree of deviation is quantified using the z-score of the edge weight, which measures whether a type combination is over- or under-represented compared to the worldwide norm. Figure 4.9 reports the distribution of standardized z-scores of type pairings across cuisines. Each point corresponds to the relative usage of a given type pair in a specific cuisine, expressed as its deviation from the global mean. The resulting spread highlights both the universal role of common types such as *Spices* and *Vegetables*, and the more selective prominence of other types, such as *Dairy* or *Seafood*, whose representation varies strongly across regions.

The comparison across cuisines reveals distinct patterns. Some cuisines exhibit long right tails in their z-score distributions, indicating the presence of type pairings that are employed far more frequently than expected from the global baseline, despite being relatively rare worldwide. This trend is particularly evident in several Asian cuisines: for instance, Indian cuisine shows a strong emphasis on the *Flowers–Legumes* combination, while Thai cuisine gives unusual prominence to *Beverages–Fish*. These examples illustrate how culturally specific traditions can elevate niche pairings into consistent structural roles, creating strong local signatures.

In contrast, cuisines such as American, Canadian, and Australasian display much narrower distributions, with most z-scores concentrated within the interval $[-1, +1]$. This pattern suggests a more uniform and balanced use of ingredient type combinations, with relatively few pairs standing out as exceptional. As noted before in 4.4.1 when comparing filtered graphs, a plausible explanation lies in the historical trajectories of these cuisines, which have been shaped by successive waves of immigration and cultural mixing during the twentieth century. As diverse culinary influences blended together, the distinctive characteristics of individual traditions may have averaged out, leading to a more homogenized repertoire where no single combination dominates disproportionately.

This contrast between long-tailed and narrow z-score distributions highlights the dual role of ingredient types in global gastronomy. Some categories, such as *Spices* and *Vegetables*, underpin a universal baseline, appearing consistently across most cuisines. Others, however,

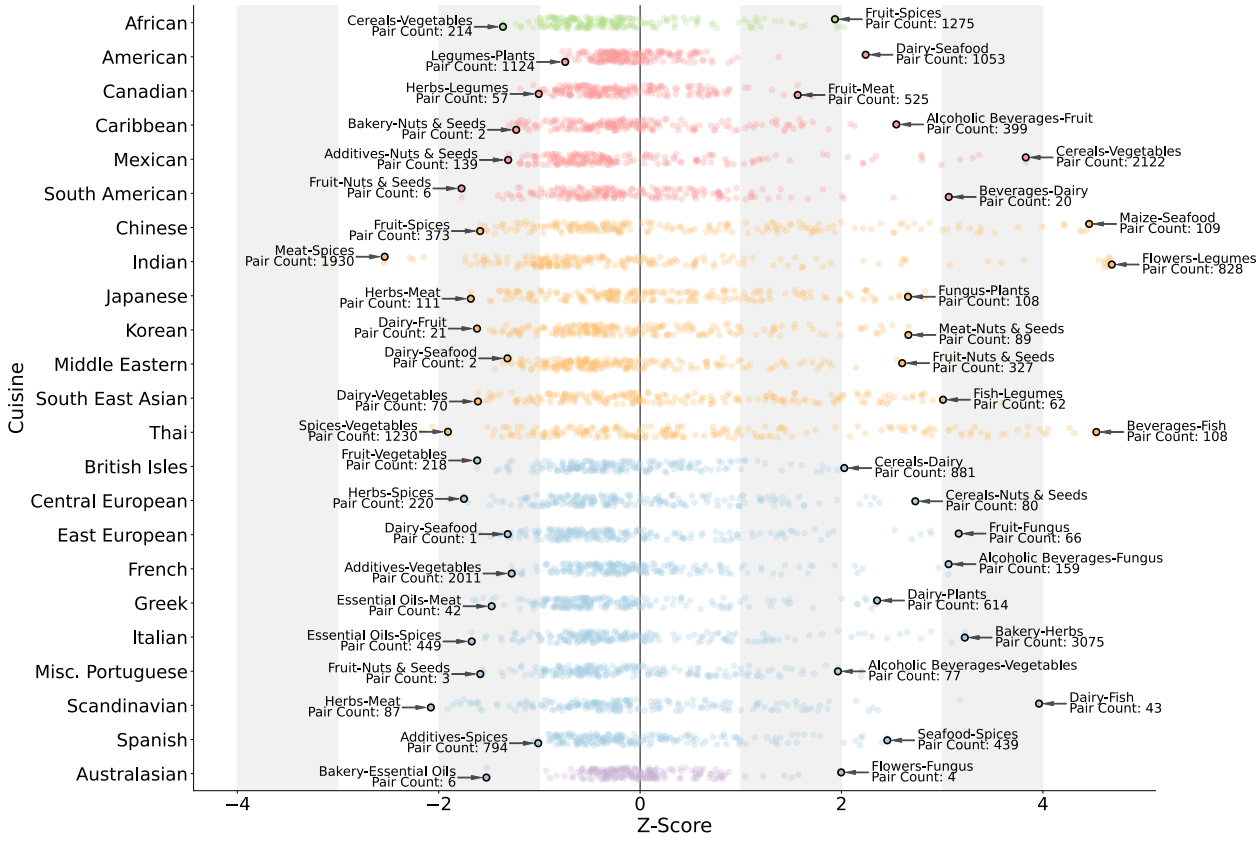


Figure 4.9: **Cross-cultural distinctiveness of type usage** Distribution of standardized z-scores across cuisines. Each strip corresponds to a cuisine, with points indicating the z-score of a given type pair relative to the global distribution. Two pairs are highlighted in each cuisine: one with a high z-score and one with a low z-score, representing significant positive and negative deviations. Long-tailed distributions reveal culturally distinctive types, whereas narrow bands indicate more universal prevalence. *Figure reproduced from Caprioli et al. [2024].*

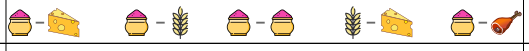
acquire meaning only in specific cultural contexts, where they become central despite their global rarity. Recognizing these divergent patterns is essential, as it sets the stage for the next step: examining the absolute prevalence of type–type associations across cuisines, and how a handful of dominant pairings recur worldwide while others remain tightly bound to particular cultural traditions. Recognizing these divergent patterns is essential, as it sets the stage for a complementary view: beyond relative distinctiveness, which pairings carry most of the total co-occurrence mass, and which associations recur most consistently across cuisines?

Table 4.2 reports, for each cuisine, the five heaviest edges together with the cumulative fractions explained by the top 5 and top 50 edges. Two regularities emerge. First, in most cuisines the top 5 edges—roughly 3% of all possible pairs—already account for about 15% of the total weight. Second, expanding to the top 50 links (around 25% of pairs) captures nearly 75% of the total weight. East Asian cuisines constitute the main exception to both patterns, exhibiting substantially lower fractions. Their pairing weights are spread more evenly across the network, indicating a less concentrated reliance on a small set of high-frequency associations.

To identify pairings that are *universally* prominent (as opposed to cuisine-specific), Table 4.3 (left) ranks type pairs by their *average* weight across cuisines. The leading positions are consistently occupied by pairings among *Spices*, *Vegetables*, and *Additives*. Indeed, nine of the top ten associations involve either *Spices* or *Vegetables*, confirming their foundational role in recipe construction worldwide. The sole exception, *Additives–Meat*, still aligns with widely shared preparation and preservation practices.

Conversely, Table 4.3 (right) ranks type pairs by their *variability* (standard deviation of weights across cuisines). Here, *Dairy*-involving pairings dominate four of the top five slots, reflecting the uneven cultural embedding of dairy: they are staple components in many European traditions yet far less common in East Asian cuisines. Taken together, these two rankings delineate a compact picture: some pairings form a global core (*Spices*, *Vegetables*, *Additives*), while others (notably those involving *Dairy*) act as markers of regional identity.

Finally, comparing these absolute summaries with the distinctiveness patterns from the previous subsection yields a coherent picture: cuisines with long-tailed z-score distributions (e.g., several Asian traditions) not only feature distinctive over- and under-represented pairings, but also tend to distribute their total weight across a broader set of edges; conversely, cuisines with more compact z-score spreads (e.g., American, Canadian, Australasian) concentrate a larger share of mass in a relatively small set of high-frequency associations. Together, the relative and absolute views reveal how cuisines balance a universal backbone of type associations with culturally distinctive deviations.

Cuisine	Top 5 Edges	Top 5 Weight Fraction	Top 50 Weight Fraction
African		0.17	0.80
American		0.12	0.74
Australasian		0.13	0.73
British Isles		0.16	0.78
Canadian		0.14	0.78
Caribbean		0.13	0.75
Central European		0.15	0.78
Chinese		0.11	0.65
East European		0.15	0.80
French		0.12	0.74
Greek		0.15	0.82
Indian		0.18	0.78
Italian		0.14	0.76
Japanese		0.15	0.68
Korean		0.16	0.77
Mexican		0.14	0.78
Middle Eastern		0.14	0.78
Misc. Portuguese		0.15	0.75
Scandinavian		0.17	0.75
South American		0.13	0.77
South East Asian		0.12	0.68
Spanish		0.15	0.71
Thai		0.10	0.62

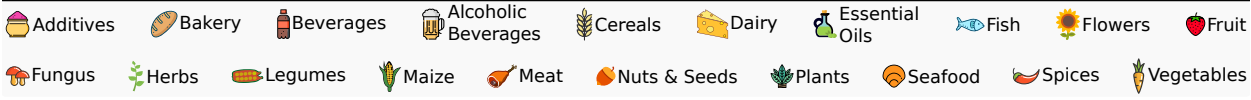


Table 4.2: **Dominant type–type associations by cuisine.** For each of the 23 cuisines, the five heaviest type–type pairs are listed, together with the cumulative fractions explained by the top 5 and top 50 links. In most cuisines, the top 5 (about 3% of links) explain $\sim 15\%$ of total weight and the top 50 (about 25%) explain $\sim 75\%$; East Asian cuisines depart from this pattern with lower concentration.

Top pairs by average weight (global core)			Top pairs by variability (standard deviation)		
Type Pair	**Avg**	**SD**	**Type Pair**	**SD**	**Avg**
Spices–Vegetables	0.027	0.006	Additives–Dairy	0.013	0.020
Additives–Spices	0.025	0.004	Cereals–Dairy	0.012	0.014
Additives–Vegetables	0.023	0.005	Dairy–Meat	0.011	0.016
Additives–Meat	0.022	0.007	Additives–Cereals	0.010	0.017
Herbs–Vegetables	0.022	0.008	Dairy–Spices	0.009	0.016
Vegetables–Vegetables	0.022	0.005	Additives–Additives	0.008	0.020
Herbs–Spices	0.021	0.008	Herbs–Spices	0.008	0.021
Plants–Vegetables	0.021	0.006	Cereals–Meat	0.008	0.013
Meat–Spices	0.020	0.004	Herbs–Vegetables	0.008	0.022
Plants–Spices	0.020	0.006	Additives–Meat	0.007	0.022

Table 4.3: **Recurrent and variable type–type associations across cuisines.** (**Left**) Top ten pairs by *average* weight across cuisines (global core). (**Right**) Top ten pairs by *weight variability* (standard deviation), highlighting culturally sensitive associations (notably *Dairy*-involving pairs).

4.5 Ingredient type networks as fingerprints of cuisines

The preceding analyses have shown that ingredient type networks encode both universal patterns and culture-specific deviations. Universal patterns are represented by type pairs such as *Spices–Vegetables* or *Additives–Spices*, which appear with consistent frequency across many cuisines and thus constitute a global baseline. By contrast, culture-specific deviations are reflected in the prominence of pairings that are locally important yet globally rare, such as *Flowers–Legumes* in Indian cuisine or *Beverages–Fish* in Thai cuisine. Together, these findings suggest that ingredient type networks capture both the shared foundations and the unique signatures of world culinary traditions. We now turn to their potential as *fingerprints* of culinary identity: compact yet distinctive representations that capture the essence of a cuisine. We explore this idea along two complementary directions. First, we construct maximum spanning trees (MSTs) of the networks, which provide minimal but highly informative backbones. Second, we test whether these representations can be used to identify and cluster cuisines through classification experiments.

4.5.1 Maximum spanning trees: minimal structural backbones

To reduce network complexity while preserving its strongest associations, we construct maximum spanning trees (MSTs) of the type–type graphs(see Section 2.2.2 and Latora et al. [2017]). An MST connects all nodes in a tree structure that maximizes the sum of retained weights, thereby producing a connected minimal graph that highlights the most salient associations. These trees can be regarded as compact "fingerprints" of cuisines, easier to interpret and compare than the full networks.

Figure 4.10 shows MSTs for six representative cuisines (see Fig. A.4 in Section A.2 of the Appendix for all cuisines). First of all, we notice that all MSTs are organized around one or two highly central nodes, identified by their high closeness centrality. At the same

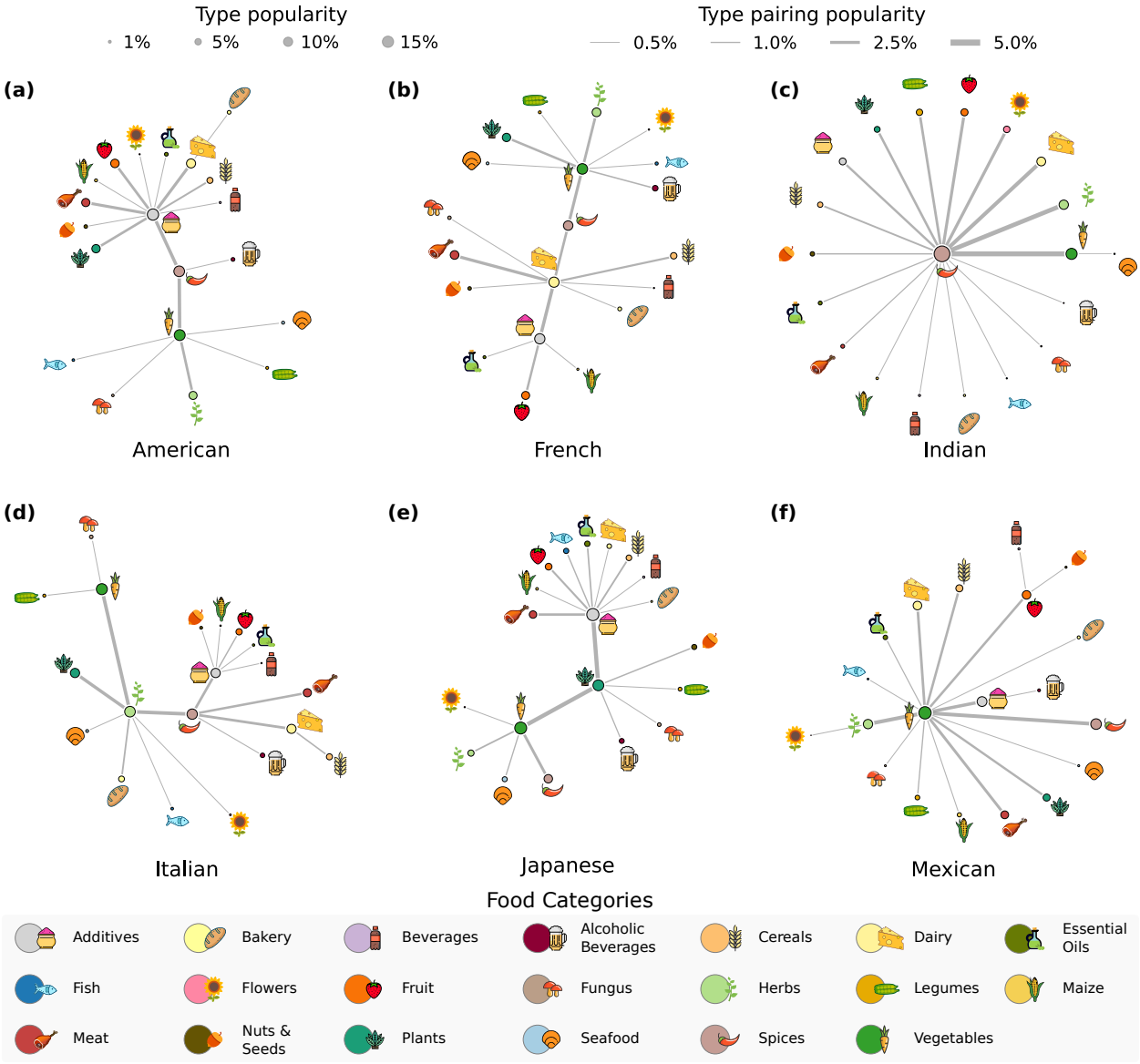


Figure 4.10: **Maximum spanning trees of type–type networks.** Examples for six representative cuisines. The trees highlight the most important associations while preserving connectivity. Distinctive patterns emerge across cuisines: Indian (**c**) shows a star-like structure with *Spices* at the hub, reflecting their centrality; Mexican (**e**) is similarly organized around *Vegetables*; French (**b**) emphasizes *Dairy* as a central connector to *Meat* and *Fungi*; Italian (**d**) features dual hubs, *Herbs* and *Spices*, a configuration shared only with Thai cuisine (see Fig. A.4). American cuisine (**a**) has a more distributed layout with the largest diameter, while Central European cuisine (see Fig. A.4) shows the longest diameter overall. *Figure reproduced from Caprioli et al. [2024].*

time, clear distinctions emerge, underscoring the effectiveness of this method in highlighting culinary diversity that would otherwise remain obscured in the fully connected graphs.

The Indian tree appears as a star-like graph, with *Spices* at its core, a configuration consistent with previous studies highlighting the centrality of spices in Indian cooking Jain and Bagler [2015]. The Mexican tree is similarly centralized, but around *Vegetables*. By contrast, French and Italian cuisines exhibit more balanced structures: French trees emphasize *Dairy* (especially cheese) and its connections to *Meat* and *Fungi*, reflecting the traditional role of cheese in dishes emphasizing umami flavors; The Italian MST (4.10(d)) is unique in its dual centrality, where *Herbs* and *Spices* act as joint hubs. This feature is otherwise found only in Thai cuisine (see Fig. A.4 in Section A.2 of the Appendix), pointing to a distinctive culinary strategy in which aromatic herbs structure recipes alongside spices.

A simple quantitative measure that exemplifies structural differences across MSTs is their *diameter*, which quantifies the length of the longest shortest path between any two ingredient types. Small diameters correspond to highly centralized networks, where one or two dominant types connect efficiently to all others, whereas larger diameters reflect more distributed organizations in which central types are less dominant and ingredient repertoires more diversified. Indian cuisine presents the smallest diameter, corresponding for example to the three links connecting *Seafood* to *Meat* via *Spices*. By contrast, American cuisine has a diameter of five, connecting *Fungi* to *Bakery* through *Vegetables*, *Spices*, *Additives*, and *Dairy*. The largest diameter among the cuisines studied occurs in Central European cuisine (see Fig. A.4 in Section A.2 of the Appendix), where two distant types are separated by six steps.

These differences in diameter point to distinct pairing logics of cuisines. Small diameters, as in Indian cuisine, indicate the presence of one or two dominant ingredient types (e.g. *Spices*) that are combined almost ubiquitously to all others, suggesting that these types function as universal anchors across recipes. Larger diameters, as in American or Central European cuisines, reflect a more decentralized structure in which no single category dominates; instead, ingredient types are arranged into more weakly connected groups, mirroring the broader and more heterogeneous repertoires of these traditions.

The strong contrasts observed in maximum spanning trees suggest that networks of ingredient type associations can serve as effective *fingerprints* of culinary traditions. To test this idea more rigorously, we ask whether such representations allow us to identify a cuisine from a sample of its recipes and to cluster cuisines into meaningful groups.

4.5.2 Classification: identifying cuisines from networks

We frame cuisine recognition as a supervised classification problem. Each cuisine is represented in three alternative ways: (i) a vector of ingredient type frequencies, (ii) its full ingredient type network, and (iii) the maximum spanning tree of that network. These representations provide inputs of increasing complexity, ranging from simple counts to structured topological signatures.

To construct the training data, we repeatedly sample subsets of recipes across cuisines in a balanced fashion. In each trial, we reserve 85% of the recipes of each cuisine for training and use the remaining 15% for testing. From the training portion, 75 recipes per cuisine are randomly drawn and converted into the three feature representations, producing one labeled

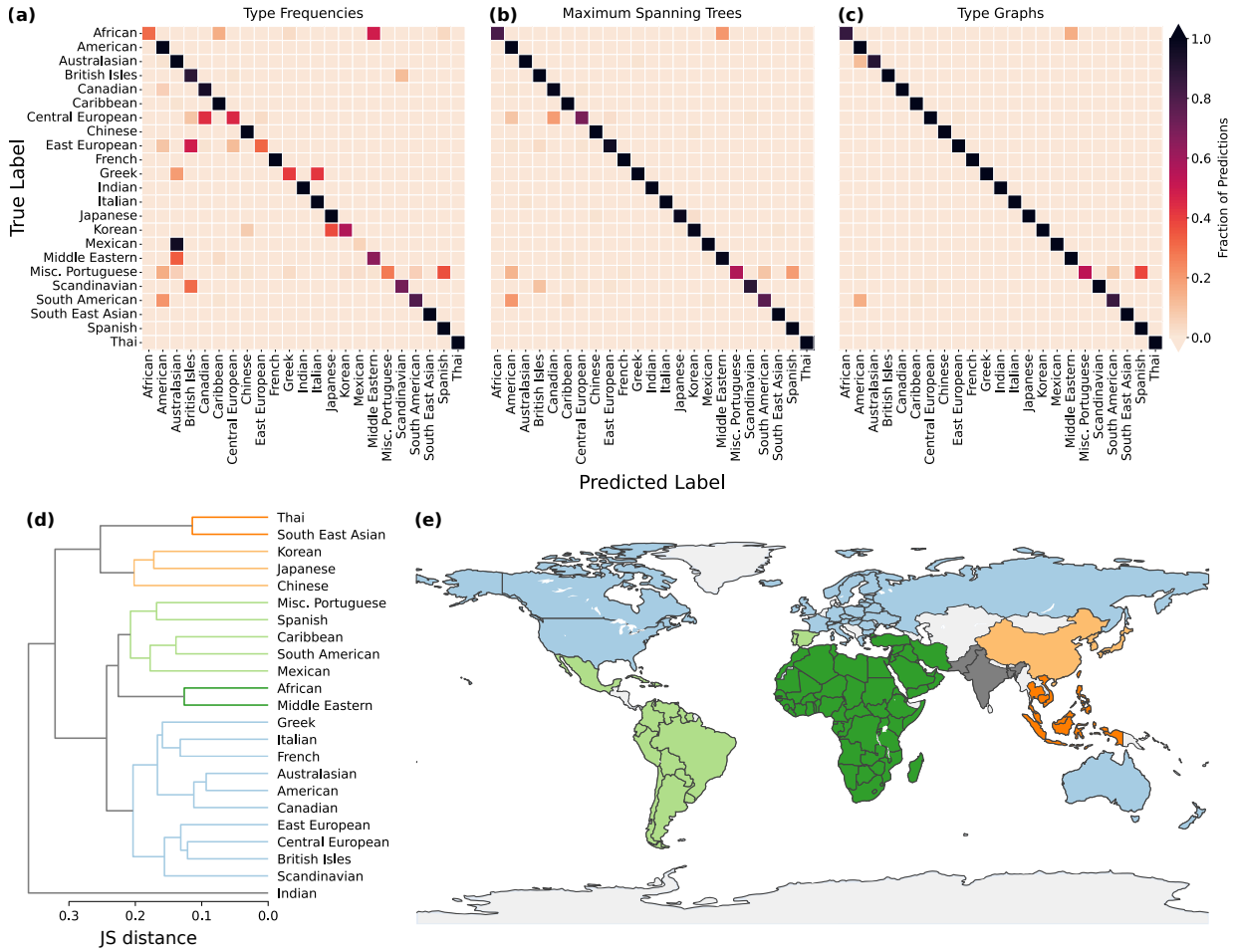


Figure 4.11: **Identification and clustering of world cuisines**. (a–c) Confusion matrices of SVM classifiers trained on (a) ingredient frequency vectors, (b) MSTs of ingredient type networks, and (c) full ingredient type networks. Graph-based representations substantially outperform frequency-based ones. (d) Hierarchical clustering of type networks using Jensen–Shannon distance (average linkage), with six clusters identified at cut distance 0.21. (e) Geographical mapping of the clusters, highlighting both regional and cultural ties (e.g. the uniqueness of Indian cuisine, Iberian–Latin American connections, and Asian subclusters). *Figure reproduced from Caprioli et al. [2024].*

datapoint per representation. Repeating this process 100 times per cuisine yields balanced training sets of 2,300 datapoints (100 samples $\times$ 23 cuisines) for each representation. The held-out 15% provide the test set, containing one datapoint per cuisine. This procedure is repeated 250 times in total, yielding a robust estimate of classification performance.

For each representation, we train a Support Vector Machine (SVM) classifier implemented with `scikit-learn`'s `SVC` [scikit-learn developers, 2025]. We use the radial basis function (RBF, Gaussian) kernel with `gamma=scale` and default regularization parameter $C = 1.0$. Multiclass discrimination across the 23 cuisines is handled by the one-vs-one strategy built into `SVC`. The same specification is trained separately on the three feature sets (type frequency, MST, and full networks). We report the average accuracy across all repetitions, with the standard deviation capturing variability due to sampling. This setup allows us to directly compare the predictive power of the three representations.

Results are summarized in Figure 4.11(a–c), where we report the confusion matrices for the three feature representations. Models trained on type frequencies achieve an average accuracy of 0.79 ± 0.07, but show substantial confusion between culturally or geographically close cuisines (e.g., Greek vs. Italian, Korean vs. Japanese). Incorporating network structure substantially improves recognition: MST-based models reach an average accuracy of 0.87 ± 0.06, while full networks achieve 0.95 ± 0.05. Even though MSTs prune nearly 90% of edges, their performance remains very strong, underscoring their value as compact fingerprints. Full networks, however, resolve nearly all ambiguities, with the only notable exception being the Misc. Portuguese cuisine, which has the fewest recipes.

Overall, these results confirm that networks of ingredient type combinations carry rich information about culinary identity, far beyond what is captured by frequency counts alone.

4.5.3 Clustering world cuisines

Beyond classification, ingredient type networks can also be used to group–or classify– cuisines in an unsupervised way. To do so, we perform a hierarchical clustering using the weighted adjacency matrices of the ingredient type graphs. To build the dendrogram we use the average linkage method and quantify how "distant" cuisines are from each other based on Jensen–Shannon distance. The resulting dendrogram (Fig. 4.11d) identifies six main groups at a cut distance of 0.21. Mapping these clusters geographically (Fig. 4.11e) reveals both expected and surprising patterns. A first cluster uniquely isolates Indian cuisine, consistent with its highly distinctive use of Spices and other types. Asian cuisines form two latitude-dependent clusters, while European, Mediterranean, and New World cuisines intermingle in larger groups that reflect long-term cultural and colonial histories. For example, Spanish and Portuguese cuisines cluster with Latin American ones, while North African and Middle Eastern cuisines cluster with the Iberian peninsula, a likely echo of medieval Islamic influence.

To complement the hierarchical clustering, we also employ dimensionality reduction to obtain a geometric view of the proximity relations among cuisines. Specifically, we perform a two-dimensional t-SNE [van der Maaten and Hinton, 2008] analysis on the weighted adjacency matrices of the ingredient type graphs, exploring different values of the perplexity parameter. Figure 4.12 reports the embeddings obtained for perplexities 5, 10, and 15. Across all values, consistent patterns emerge: cuisines from Northern Europe and from East Asia form compact and well-separated clusters, appearing at opposite ends of a culinary

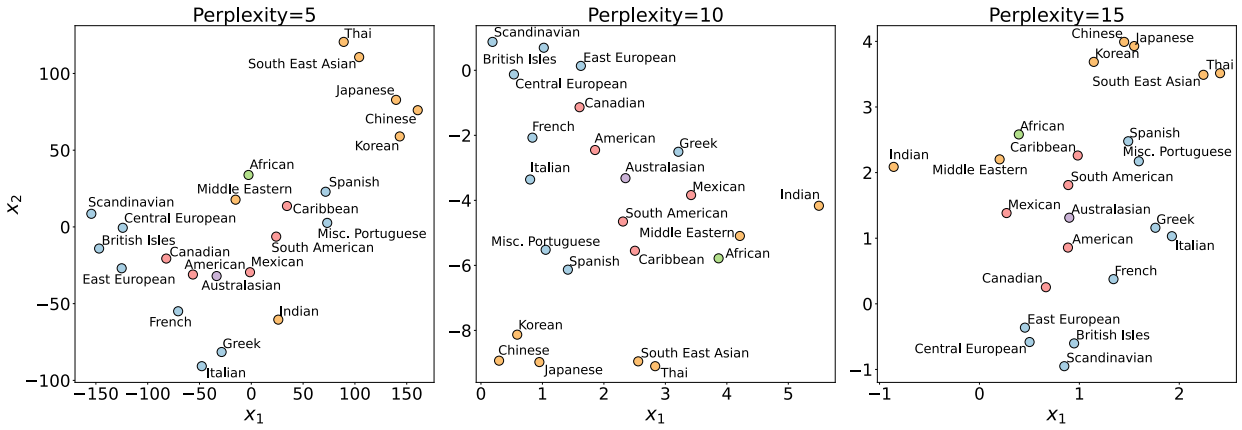


Figure 4.12: **2d t-SNE plots of ingredient type graphs**. Each plot represents the 2 dimensional embedding of the weighted adjacency matrices of type graphs for all 23 cuisines. (a) Perplexity 5, (b) Perplexity 10, (c) Perplexity 15. Cuisines are colored according to their continent affiliation for additional context. Asian and Northern European cuisines exhibit clustering across all perplexity levels, indicating significant similarities in their ingredient usage. *Figure reproduced from Caprioli et al. [2024].*

continuum. Other traditions, particularly Mediterranean and New World cuisines, occupy intermediate positions, bridging between these extremes.

These projections are broadly compatible with the hierarchical clustering results (Fig. 4.11d–e), but emphasize the idea of continuity rather than discrete grouping. In the t-SNE embedding, cuisines are not only clustered but also arranged along a gradient that reflects their degree of similarity in ingredient type usage. This gradient suggests that world cuisines do not simply partition into a small number of isolated families, but instead span a structured continuum of culinary strategies, with geographical proximity playing a major role in shaping their relative positions.

Taken together, the classification and clustering analyses demonstrate that ingredient type networks act as robust and distinctive fingerprints of world cuisines: they not only allow high-accuracy recognition of individual traditions, but also reveal the cultural and geographical ties that structure global gastronomy.

4.6 Summary and outlook

In this chapter we investigated the structural organization of world cuisines through the lens of a novel network-based approach, shedding light on both general trends and distinctive features of culinary traditions. We began by considering the recipe space within a single cuisine, showing that the number of recipes scales exponentially with the number of ingredients. This observation highlights the vast discrepancy between the potential combinatorial space of ingredients and the much smaller number of realized dishes: modern cuisines explore only a minute fraction of the possible combinations, favoring variations on established templates rather than entirely novel constructions. Such limitations suggest a principle of least energy and resonate with evolutionary models of gastronomy, where high-fitness ingredients

gradually replace lower-fitness ones [Kinouchi et al., 2008, Nazari et al., 2024]. Functional constraints such as antimicrobial properties have also been proposed as drivers of ingredient selection Billing and Sherman [1998], Sherman and Hash [2001], further emphasizing that the recipe space is shaped by both cultural and biological pressures.

Turning to ingredient usage at the level of types, we uncovered a global hierarchy of prevalence. Categories such as *Vegetables*, *Spices*, and *Additives* dominate across virtually all cuisines, reflecting their nutritional and culinary versatility, while others such as *Flowers*, *Beverages*, or *Maize* play only marginal roles. These regularities point to shared functional needs and constraints in cooking, while still leaving room for strong cultural variation. A z-score analysis of type–type pairings revealed precisely this duality: a global core of associations used comparably across cuisines, and culturally distinctive pairs that mark specific traditions (e.g. *Dairy* in European cuisines versus its relative absence in East Asian ones). Complementary analyses of absolute prevalence showed that a handful of type pairs account for a disproportionate share of recipe structure, with East Asian cuisines again standing out for their more evenly distributed usage.

To investigate these regularities systematically, we introduced the networks of ingredient type combinations, or *ingredient type graphs*, where nodes correspond to categories of ingredients and weighted edges capture their co-occurrence in recipes. These networks provide a natural representation of cuisines: they distill the complex space of individual recipes into relational structures that emphasize how ingredient types interact. By comparing these graphs across cuisines, we gained a unified framework for detecting both universal building blocks and culturally specific strategies of pairing.

In the second half of the chapter, we advanced the idea that type networks act as *fingerprints* of culinary identity. By extracting maximum spanning trees, we constructed minimal backbones that highlight the most salient associations while preserving connectivity. These trees revealed striking contrasts across cuisines: from the spice-centered star of Indian cuisine to the vegetable-based hub of Mexican cuisine, or the dual aromatic hubs of Italian and Thai traditions. Structural differences were further reflected in network diameters, distinguishing cuisines organized around a few dominant types from those with more decentralized repertoires.

We then tested the fingerprinting power of these networks quantitatively. In supervised classification experiments, graph-based representations substantially outperformed simple frequency vectors, with full networks achieving near-perfect recognition accuracy. Even their sparsified maximum spanning trees retained high predictive power, demonstrating that the essential identity of a cuisine is encoded in its type associations. Unsupervised clustering and dimensionality reduction further confirmed these insights: hierarchical trees and 2D t-SNE embeddings both revealed meaningful groupings of cuisines, shaped by cultural, geographical, and historical ties, while also isolating highly distinctive outliers such as India.

Taken together, these results establish ingredient type networks as compact yet discriminative signatures of world cuisines. They capture both universal building blocks and culturally specific strategies, explain why only a small portion of the recipe space is realized, and reveal the relational patterns that bind or separate culinary traditions. In the next chapter, we build on this foundation to move beyond description and examine the generative mechanisms that drive the evolution of recipes and cuisines over time.

Chapter 5

Mapping Influence: The Network of Chefs

The preceding chapters traced a progression from molecular principles of ingredient compatibility (Chapter 3), through cultural fingerprints of ingredient-type usage (Chapter 4). In this final chapter, the scope expands from *what* is cooked to *who* cooks, and to the relational structures through which culinary ideas propagate. Throughout history, cuisines have been decisively shaped by the creative interventions of chefs. Auguste Escoffier's codification of haute cuisine at the turn of the twentieth century professionalized the kitchen and established new culinary standards [Grabowska, 2025]. Later, stylistic revolutions such as nouvelle cuisine and molecular gastronomy redefined what counted as legitimate culinary practice, positioning chefs as cultural innovators [Leschziner, 2007]. More recently, René Redzepi and the New Nordic Cuisine movement exemplify how reinterpretations of terroir and heritage can transform regional foodways into globally influential gastronomic paradigms [Gyimóthy and Mykletun, 2009]. Such figures do not merely produce dishes; they reconfigure repertoires of ingredients, techniques, and meanings, leaving legacies that extend far beyond their own restaurants. In this sense, chefs can be understood as pivotal agents of gastronomic change. Yet such change does not occur in a vacuum or as the work of isolated genius. It unfolds within professional and cultural milieus, sustained by relationships of mentorship, collaboration, and stylistic exchange. Chefs are therefore not only creative actors, but also nodes in complex networks whose embeddedness shapes the evolution of cuisines as decisively as ingredients themselves.

Scholarly work on chef careers emphasizes that professional identity in gastronomy is forged through relationships as much as through individual artistry. Mentorship and apprenticeship are consistently identified as the backbone of culinary development: Iomaire [2008] describes mentoring as the primary model for nurturing talent in kitchens, while Bressan et al. [2023] highlight knowledge transfer as a relational process embedded in institutions and professional communities. These ties extend beyond simple dyads. As Feeney and Bozeman [2008] argue in their broader theory of careers, it is *networked mentorship structures*—ensembles of teachers, peers, rivals, and collaborators—that provide resilience and opportunity.

The culinary field is thus characterized by what Borkenhagen [2018] terms a "social structure of knowledge," comparable to laboratories in the sciences, where tacit expertise is

transmitted through close contact and shared practice. In haute cuisine especially, success depends on embedding oneself within constellations of influence: Elbasha and Baruch [2022] show that mobility between institutions, access to prominent mentors, and collaboration across elite kitchens are key drivers of global recognition. At the same time, studies underscore persistent asymmetries. Albors-Garrigós et al. [2020] demonstrate how gendered access to mentorship shapes career trajectories, while Iomaire and Allen [2017] identify structural barriers that differentiate the pathways of head chefs across cultural contexts.

Against this backdrop, the present chapter constructs a network perspective on culinary influence. Drawing on the *Atlante dei Cuochi* [Vignoli and Bolasco, 2023] as a curated primary source, enriched with open knowledge bases, we model chefs as interconnected agents whose careers are defined not only by individual achievements but by their embeddedness in historical and professional genealogies. By analyzing these relational structures, the chapter aims to map how culinary ideas travel across generations, regions, and institutions, and to show how the creative evolution of cuisine is inseparable from the networks of chefs themselves.

We begin by tracing the temporal patterns of influence across generations, then use centrality measures to identify key mentors, heirs, and ties that sustain connectivity. We next examine how national traditions and stylistic schools shape community structure, before extracting a backbone in the form of a maximum spanning tree that distills the dense network into a genealogical map. Together, these perspectives provide a multi-dimensional view of how haute cuisine evolves through the interplay of individuals, relationships, and regions.

5.1 The *Atlante dei Cuochi* source

5.1.1 Introduction to the source

This chapter draws on *L'atlante dei cuochi. I miti, i maestri, gli allievi protagonisti della storia della cucina* (hereafter, *the Atlante*, or *Chef Atlas*), an Italian-language illustrated volume by Enrico Vignoli and Marco Bolasco, published by Slow Food Editore in 2023.[1] The book presents a curated, international survey of chefs and their relationships, conceived explicitly to map genealogies of mentorship and influence in contemporary and historical gastronomy. The volume profiles 257 chefs, ranging from canonical figures such as Auguste Escoffier to contemporary innovators like Ferran Adrià and René Redzepi.

Each entry in the *Atlante* corresponds to a single chef (see Fig. 5.2). The page layout follows a consistent format: it includes the chef's name, a stylized sketch of their face, birth year and birthplace, and a short biographical essay outlining their career trajectory and significance in culinary history. Crucially for the present work, the lower-left section of each page encodes genealogical ties. It displays two columns of names: the column on the left lists chefs who influenced the profiled chef, while the column on the right lists those influenced by them. This bidirectional encoding of influence—predecessors on one side, successors on the other—provides the raw material for constructing a directed network of culinary relationships.

[1] ISBN 978-88-8499-776-0. See publisher page: `https://giunti.it/products/latlante-dei-cuochi-bolasco-marco-9788884997760`

Beyond these individual profiles, the *Atlante* also includes a large infographic that renders the "family tree" of cuisine as a genealogical *star-map*. This visualization clusters chefs into constellations and schools, connecting masters to pupils and highlighting stylistic currents across generations. The *star-map* offers a global overview of the field, while the page-level columns provide a systematic, fine-grained basis for extracting network edges.

Although the scope is international, the *Atlante* devotes significant attention to Western and European fine-dining chefs—a factor to consider when interpreting the network's global structure.



Figure 5.1: **Cover of *L'Atlante dei Cuochi*** (Slow Food Editore, 2023). Image reproduced for illustrative purposes.

In the context of this thesis, the *Atlante* complements recipe-centric datasets by shifting the analytical focus from *what* is cooked (ingredients, recipes) to *who* cooks and *how* culinary ideas propagate—through mentorship, lineage, and institutional affiliation. It thus provides a historically grounded foundation from which to construct the chef-influence directed layer and seed enrichment with open knowledge bases (Wikidata, Wikipedia) for additional metadata such as nationality, restaurants, active years, and awards.

5.1.2 Scope and coverage

Overall scale. The *Atlante* profiles a total of 257 chefs. While the volume provides a carefully curated overview of the field, its coverage is by no means exhaustive—it is not intended to catalog every notable culinary figure, which would be both impossible and impractical. Achieving unbiased sampling of chefs across time, geography, and culinary domains presents inherent challenges: the field is vast, constantly evolving, and shaped by asymmetries in



Figure 5.2: Representative biographical page for Albert Adrià and Ferran Adrià from *L'Atlante dei Cuochi*. Each entry includes a portrait sketch, birthplace and birth year, a short biography, and—crucially—the two columns of influence ties at the lower left: on the left, chefs who influenced the profiled figure; on the right, chefs influenced by them. Illustration by Martina Dirce Carcano.

media visibility, linguistic access, and archival availability. Instead, the inclusion criteria appear to rest on the authors' curatorial expertise. As food journalists and historians, Enrico Vignoli and Marco Bolasco selected chefs whose mentorship networks, stylistic legacies, or institutional prominence made them representative of major currents in global gastronomy. We rely on their domain knowledge and editorial judgment as the basis for this dataset.

Geographic representation. Although the *Atlante* features chefs from multiple continents and traditions, it exhibits a clear emphasis on *Western and European fine-dining*, reflecting both the authors' cultural context and the institutional prominence of those chefs. To provide a visual summary, Figure 5.3 presents a world map with shaded markers indicating the birthplaces or primary nationalities of the 257 chefs profiled. The map shows a dense clustering of chefs in France, Italy, and Spain, with additional concentrations in the United States and Japan. By contrast, large parts of Africa, South Asia, and Southeast Asia appear nearly absent. This suggests that subsequent network metrics such as centrality or connectivity will be disproportionately shaped by European and transatlantic lineages. While this imbalance reflects the cultural standpoint of the *Atlante*, it is analytically useful insofar as Western fine dining institutions have historically set global benchmarks for recognition and mobility. Capturing their mentorship genealogies therefore allows us to trace many of the most central nodes in the worldwide structure of culinary influence, even if other gastronomic traditions remain under-represented.

Temporal span. The volume spans chefs active across more than a century of culinary history—from early 20th-century pioneers (e.g., Auguste Escoffier, active c. 1900–1920) to contemporary innovators (e.g., Ferran Adrià, René Redzepi). The distribution displays a

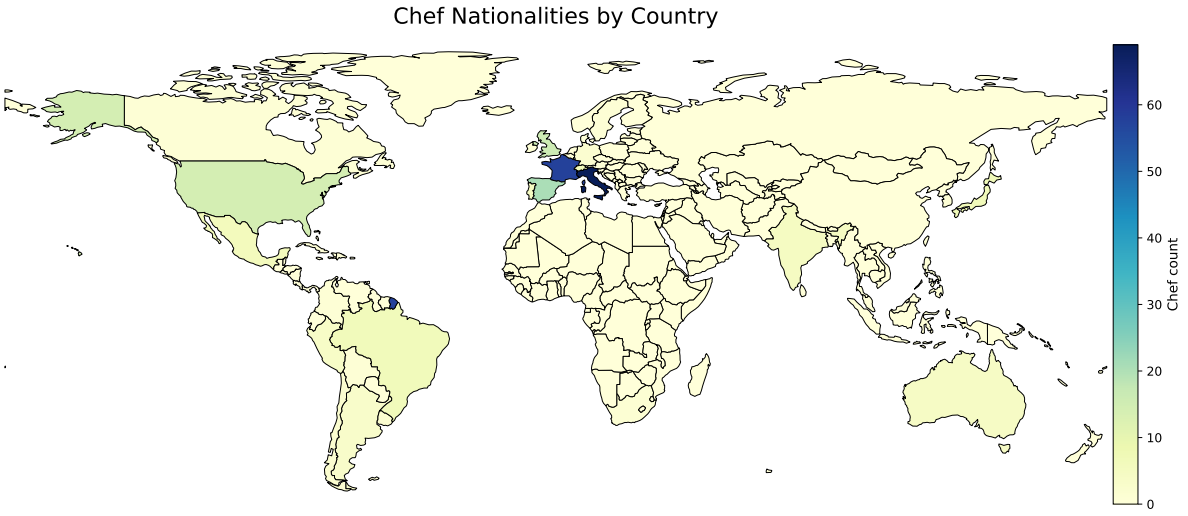


Figure 5.3: Geographic distribution of chefs in the *Atlante*, by country of origin or nationality (257 chefs).

pronounced recency bias, with the majority of chefs having had careers in the late 20th or early 21st centuries. To illustrate this, Figure 5.4 shows a histogram of the chefs' birth years (or, when birth year is unavailable, career-peak decade).

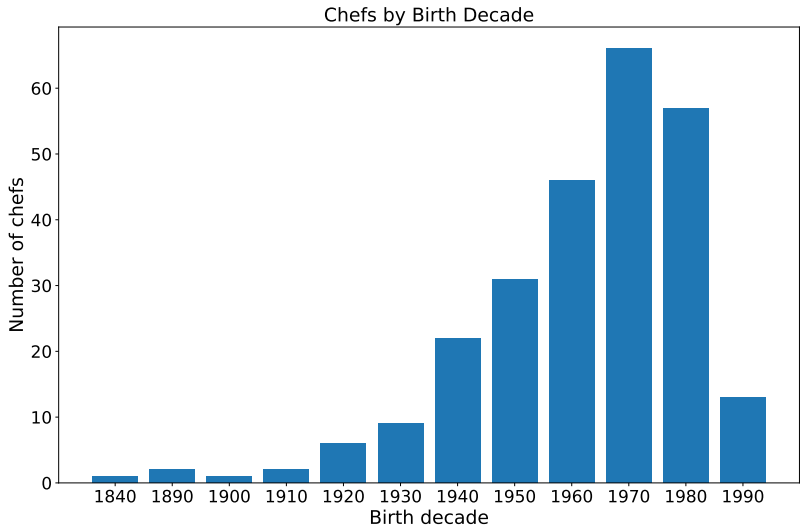


Figure 5.4: Temporal distribution of chefs in the *Atlante* by birth year or main career era (257 chefs), showing a recency bias.

The histogram reveals distinct generational waves: an early cohort of foundational figures around 1900–1930, a surge in post-war decades (1950s–1970s), and a marked expansion from the 1980s onward coinciding with the globalization of fine dining and the rise of molecular gastronomy. This temporal clustering foreshadows the formation of influential "schools" of chefs in the network, where younger generations are densely connected to a smaller set of pioneering mentors.

5.2 Data-curation pipeline

To transform the *Atlante dei Cuochi* into a structured dataset, we extracted the network of relationships encoded in the volume's biographical entries and genealogical diagrams. Each page of the *Atlante* presents a chef's profile alongside their connections to other figures (see Fig. 5.2). By digitizing this structure, we obtained a directed influence network suitable for quantitative analysis. The construction process involved two main stages: (i) network extraction from the *Atlante*, and (ii) validation and classification of links using external sources.

5.2.1 Network extraction from the *Atlante*

The extraction procedure was as follows:

- Each page of the *Atlante* corresponding to a chef was photographed at high resolution.
- Using an image recognition pipeline, we automatically extracted:
  - the chef's name,
  - the names appearing in the upward column (indicating predecessors or mentors),
  - the names in the downward column (indicating pupils, collaborators, or stylistic successors).
- We constructed a directed network in which each chef node is connected by an outward edge to all names in the downward column, and by an inward edge from all names in the upward column.
- We performed manual checks to correct OCR errors and misspellings, and ensured link consistency across pages.
- Finally, each chef was assigned a unique integer identifier to enable computational processing.

5.2.2 Link validation and classification

While the *Atlante* provides a first curated genealogy, we enriched and validated the extracted ties by consulting online sources such as biographies, interviews, and press coverage. Each directed edge was reviewed and mapped into one of three categories:

Working together A confirmed relationship established through direct professional collaboration, such as training, apprenticeship, or shared employment in the same kitchen.

Inspiration without collaboration A stylistic or creative influence explicitly acknowledged, but without evidence of direct work experience.

Uncertain Connections reported in the *Atlante* for which no external validation could be found.

The distribution of ties across these categories reflects both the predominance of direct professional connections and the residual importance of less easily documented influences. Out of 417 total links, *Working together* accounts for the vast majority (363 ties, 87.1%), while *Inspiration without collaboration* represents 23 ties (5.5%), and *Uncertain* connections amount to 31 ties (7.4%).

The *Uncertain* category deserves special attention. Its presence does not imply that the ties are spurious. On the contrary, it reflects the limits of our validation strategy, which relies exclusively on explicit, recorded evidence. As culinary experts, the authors of the *Atlante* are able to evaluate chefs' cuisine, techniques, and professional activities, and to infer legitimate genealogical ties from stylistic analysis and insider knowledge. Thus, while such links remain unconfirmed in our classification, they likely capture real channels of influence that are invisible to purely documentary methods.

This distinction highlights the complementary nature of expert curation and computational validation: the former can access tacit dimensions of influence (stylistic resemblance, culinary ethos), while the latter provides replicable and transparent criteria based on written records. By retaining all three categories, we preserve this epistemic layering in the dataset, enabling later analyses to differentiate between validated collaborations, acknowledged inspirations, and expert-inferred ties.

5.2.3 Metadata enrichment

In addition to extracting relational ties, we enriched each chef node with biographical metadata. The enrichment process was carried out in two complementary steps:

1. **Wikidata extraction.** We first queried Wikidata for each chef name (using canonical and variant spellings) and recorded the available fields: birth date, death date (if applicable), place of birth (city and country), and nationality.

2. **Manual completion.** The Wikidata pipeline did not cover the entire dataset and frequently contained missing or incomplete records. For all unmatched chefs, and for missing fields among matched chefs, we manually researched online sources such as interviews, biographies, press articles, and institutional pages.

This two-step process ensured complete coverage across all 257 chefs. We successfully filled *all nationality fields*. In cases of dual nationality, we elected a primary nationality, privileging the dominant affiliation (for example, favoring larger European countries when relevant). Birth dates were recorded with the highest available precision. When the full date (day, month, year) was available, it was stored in the format `YYYY-MM-DDT00:00:00Z` (e.g., `1974-04-25T00:00:00Z`). When only the year was available, we used January 1 as a placeholder (e.g., `1974-01-01T00:00:00Z`). Because our analyses focus on the level of birth years, this approximation does not affect results.

The resulting metadata enables us to situate chefs not only in professional genealogies, but also in time and space, linking the influence network to broader cultural and historical dynamics.

5.3 Descriptive Network Statistics

5.3.1 Global overview

The chef influence network extracted from the *Atlante* comprises 257 nodes (chefs) and 414 directed edges, corresponding to ties of professional collaboration, stylistic inspiration, or unconfirmed genealogical influence (see Section 5.2.2). Its density is very low ($d \approx 0.006$), consistent with the sparse character of social lineage networks. The network is not strongly connected, but its *largest weakly connected component* includes 225 chefs, indicating that most chefs are indirectly linked through chains of influence. Fig. 5.5 provides a visualization of this component, with nodes sized in proportion to out-degree (number of chefs influenced) and labeled only when total degree exceeds 5. Edge colors reflect the type of influence: grey for *Working together*, green for *Inspiration without collaboration*, and orange for *Uncertain*. The global statistics are summarized in Tab. 5.1.

Statistic	**Value**
Nodes	257
Edges	414
Density	0.0063
Strongly connected	False
Largest weakly connected component size	225
Average in-degree	1.62
Average out-degree	1.62

Table 5.1: **Global statistics** of the chef influence network.

5.3.2 Degree distributions and inequality

Fig. 5.6(a) and Fig. 5.6(b) show the probability mass functions (PMFs) of in-degree and out-degree on a log–lin scale. Both distributions decay rapidly, consistent with the fact that only a few chefs accumulate more than a handful of ties. This pattern reflects a strongly right-skewed structure: the majority of chefs are linked to one or two others, while a limited set of figures connect to substantially more. For example, the maximum in-degree is 7 (against an average of 1.6), while the maximum out-degree is 20. Rather than exhibiting heavy-tailed or scale-free behavior, the network shows an exponential-like decay with moderate hubs, which is typical of genealogical and apprenticeship systems where social and biographical constraints limit the maximum degree.

This asymmetry between in-degree and out-degree is also theoretically expected. The limited range of in-degree values reflects the natural constraints of a chef's formative years: one can usually work under only a handful of mentors, often for extended periods of time, before transitioning into leadership positions that end the apprenticeship phase. By contrast, out-degree is shaped by the cumulative span of a chef's career. Once established as head chef, an individual may train dozens of apprentices, host stagiaires, or inspire successive generations. In practice, this means that while in-degree is capped by biographical limits,

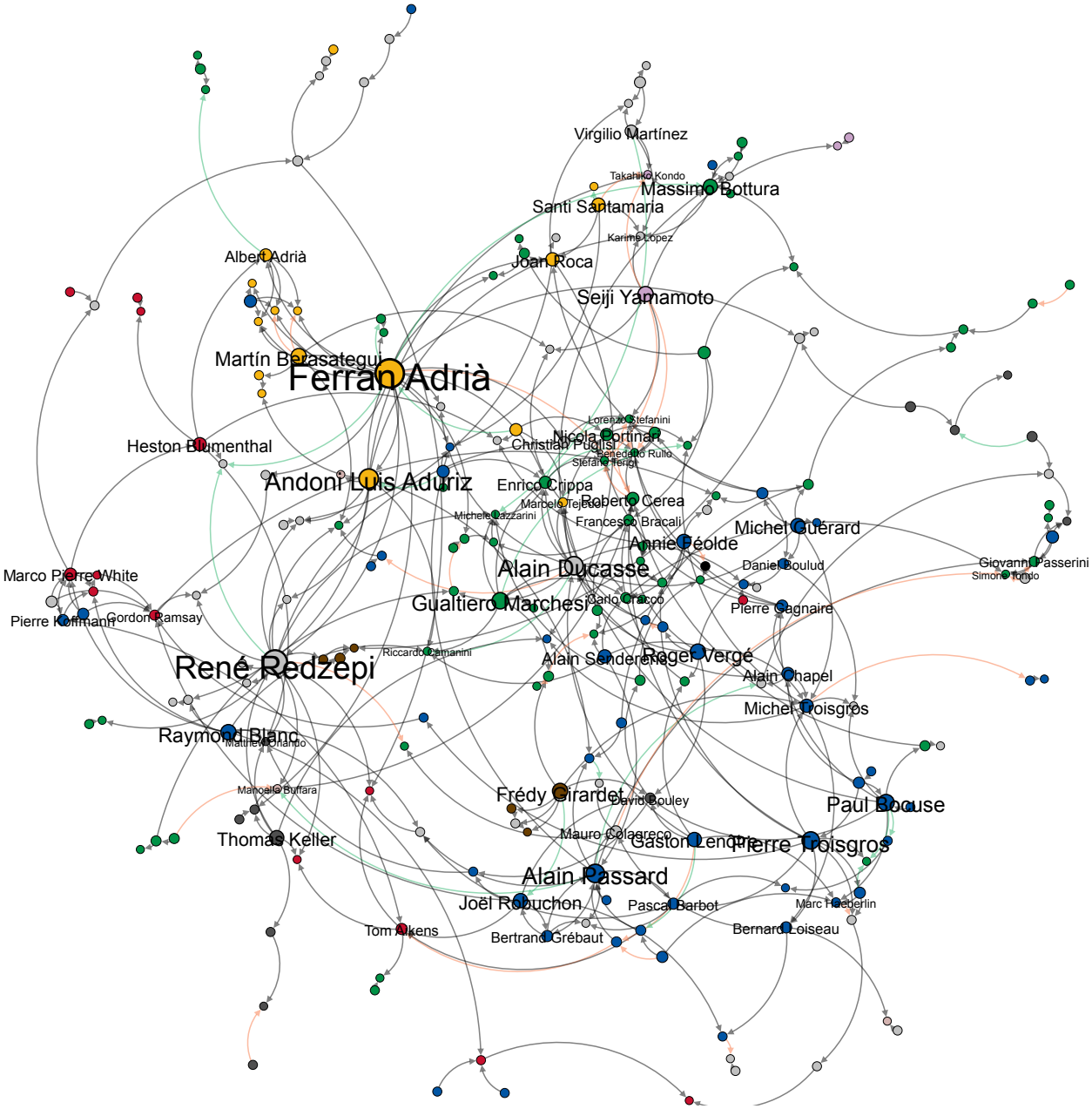


Figure 5.5: **Global chef influence network.** Visualization of the largest weakly connected component (225 chefs). Nodes are scaled by out-degree; names are displayed only for chefs with total degree greater than 5. Node colors represent the chef's nationality; edge colors indicate the type of influence: grey for *Working together*, green for *Inspiration without collaboration*, orange for *Uncertain*.

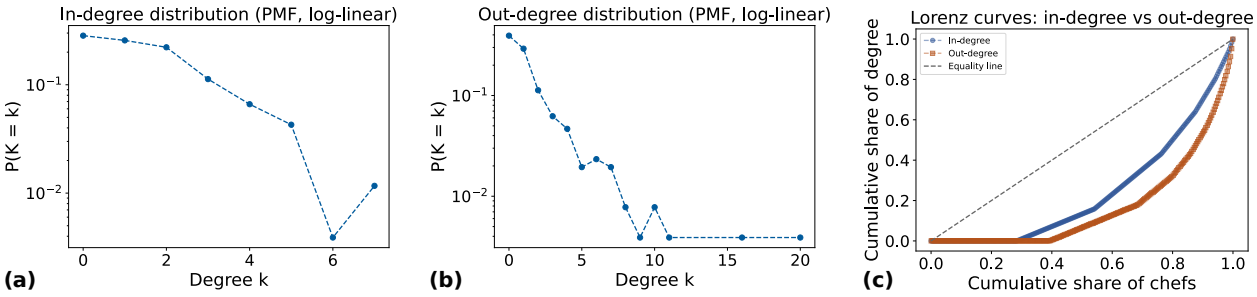


Figure 5.6: **Degree distributions and inequality.** (a) In-degree distribution (PMF, log–lin). (b) Out-degree distribution (PMF, log–lin). (c) Lorenz curves for in- and out-degree compared to the line of equality. The plots show that most chefs maintain only a handful of ties, while a minority concentrate a disproportionate share of connections, especially in terms of out-degree.

out-degree can grow over decades and may vary widely depending on a chef's institutional prominence, geographic mobility, and capacity to sustain large brigades. This explains why the maximum observed out-degree (20) is substantially higher than the maximum in-degree (7), even in a relatively small network.

This asymmetry is not unique to gastronomy. A closely related pattern appears in advisor–advisee genealogies. Analyses of the Mathematics Genealogy Project highlight that the *in-degree* of scholars is overwhelmingly equal to one (most PhD holders have a single advisor), while the *out-degree* (number of advisees) is markedly more dispersed, with a right-skewed distribution plotted on a log–log scale [Semenov et al., 2020]. This mirrors what we observe in the chef network: formative ties are necessarily few, whereas opportunities to transmit influence can accumulate and vary widely across a career.

Fig. 5.6(c) provides a complementary view by plotting Lorenz curves for in-degree and out-degree relative to the line of equality. The curves reveal that ties are unevenly distributed: while most chefs account for very few links, a limited number contribute disproportionately to the overall connectivity of the network. This inequality is more pronounced for out-degree (mentorship or influence transmitted) than for in-degree (mentorship received), consistent with the biographical constraints discussed above. Quantitatively, the Gini coefficient for in-degree is 0.51, while for out-degree it reaches 0.66, confirming that opportunities to transmit influence are concentrated in the hands of a relatively small set of chefs. Thus, even in the absence of extreme hubs or heavy-tailed scaling, the network exhibits a clear hierarchical profile: most chefs transmit influence to very few, while a limited number anchor disproportionately large genealogies.

5.3.3 Temporal patterns of influence

The temporal dimension of the network clarifies how influence is transmitted across generations. We examine four complementary aspects: (a) the demographic distribution of chefs by birth cohort, (b) the timing of ties in terms of age gaps, (c) correlations between birth year and network centrality, and (d) the genealogical reach of chefs as measured by *generational depth*, defined as the length of the longest directed path rooted in each chef. We complement

these quantitative patterns with a temporal visualization of the network, which makes the chronological flow of influence directly observable. Together, these perspectives show how mentorship and inspiration follow a strong life-cycle logic, with older chefs acting as sources of influence and younger chefs as receivers.

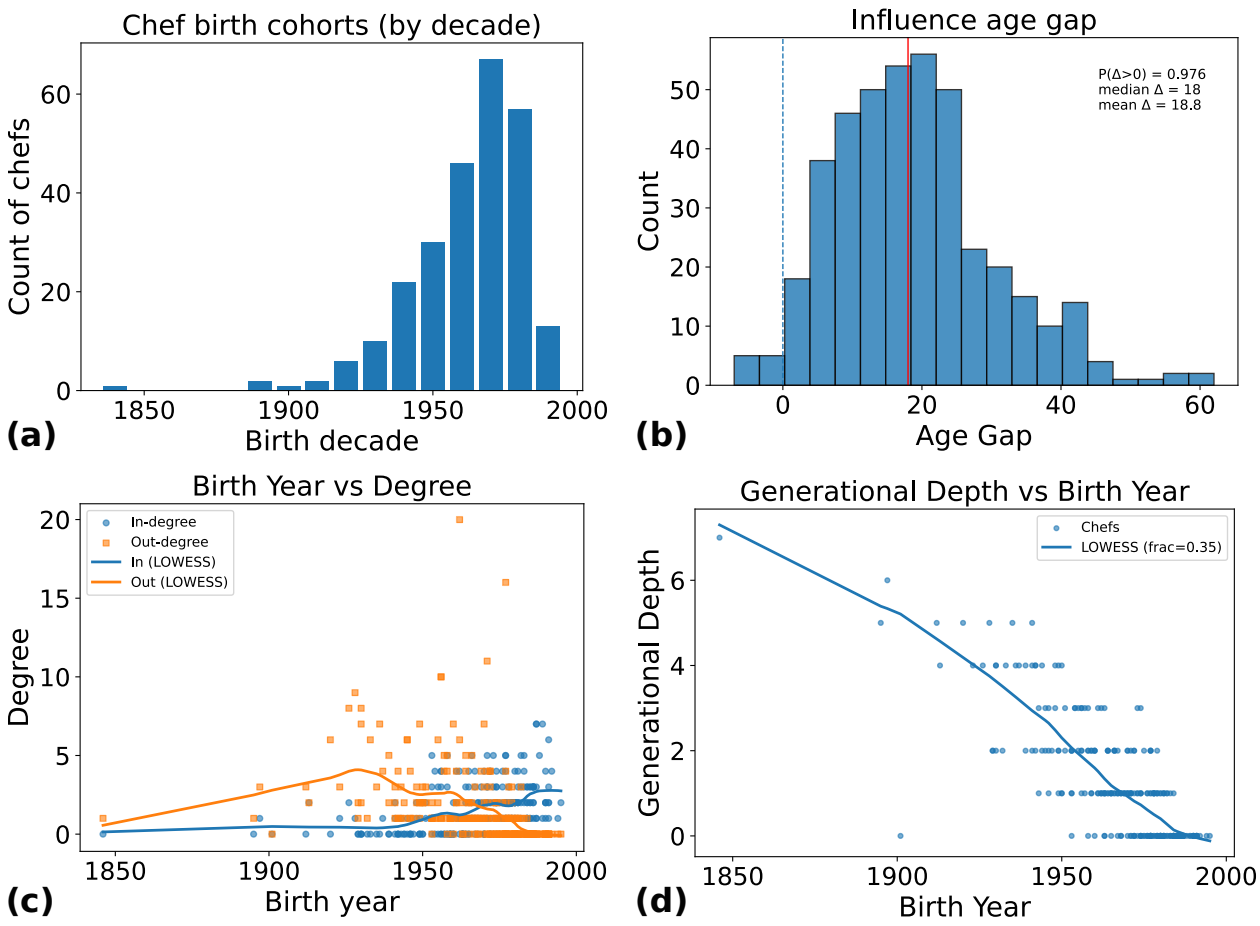


Figure 5.7: **Temporal patterns of influence.** (a) Distribution of chefs by decade of birth, highlighting the concentration in the 1960s–1980s cohorts. (b) Histogram of age gaps between source and target of ties. The blue dashed line marks zero (younger-to-older ties), while the red solid line marks the mean gap (18.8 years). The inset reports the share of ties with positive gap ($P(\Delta > 0)$), the median gap, and the mean gap. (c) Birth year versus in-degree (blue) and out-degree (orange), with LOWESS smoothing. Out-degree peaks among the 1940s cohorts and declines in later generations, while in-degree rises steadily toward the youngest chefs. (d) Generational depth—the length of the longest directed path originating from a chef—plotted against birth year, showing a decreasing trend: earlier-born chefs anchor deeper genealogical trees, whereas younger chefs remain closer to the leaves.

Birth cohorts are heavily concentrated in the post-war period: 66% of chefs were born between the 1960s and 1980s (170 of 257), and 78% between the 1950s and 1980s (200 of 257). This strong recency bias reflects the *Atlante*'s focus on modern fine dining, where most influential careers emerged after World War II. It also implies that older pioneers are comparatively few in number, a fact to keep in mind when interpreting centrality measures.

Fig. 5.7(a) visualizes this cohort bulge by decade of birth.

Edge-level timing further illustrates this generational structure. Fig. 5.7(b) shows the histogram of age gaps between connected chefs. The distribution is centered well above zero, with a mean gap of 18.8 years (red line) and a median of 18 years. The vast majority of ties (97.6%) point from older to younger chefs, consistent with a generational flow of influence, while only a small minority (2.4%) cross in the opposite direction (blue dashed line). A pronounced long tail shows that in some cases mentorship and inspiration span more than half a century. The most extreme example is Gualtiero Marchesi (b. 1930), linked to Francesco Capuzzo Dolcetta (b. 1992) across 62 years via inspiration, and to Michele Lazzarini (b. 1991) across 61 years through direct collaboration. Such wide gaps emphasize the longevity of influence exerted by senior chefs, sustaining continuity across successive cohorts.

The relationship between cohort and network position is made clearer in Fig. 5.7(c), which plots in-degree and out-degree against birth year. In-degree increases with birth year (Spearman's $\rho = 0.484$), showing that younger chefs tend to accumulate influence from their seniors. Conversely, out-degree decreases with birth year ($\rho = -0.67$), indicating that earlier-born chefs are more likely to act as mentors and hubs of influence. The fitted LOWESS (locally weighted scatterplot smoothing with a tricube kernel, Cleveland [1979]) curves reveal that out-degree peaks among the 1940s cohorts—corresponding to chefs who came of age during the rise of Nouvelle Cuisine—while in-degree rises steadily among younger cohorts, who are still in the process of building their own careers. This asymmetry highlights a generational gradient: influence flows downward through the hierarchy of age.

Finally, Fig. 5.7(d) considers *generational depth*. The fitted LOWESS curve (locally weighted scatterplot smoothing) highlights a clear decreasing trend. Earlier-born chefs anchor deep genealogical trees, standing at the root of long chains of mentorship, while more recent cohorts show shallower depth, consistent with their position as recipients still in the early stages of lineage formation.

Figure 5.8 provides a visual counterpart to the statistical patterns. By anchoring chefs to their birth years on the horizontal axis, the layout makes explicit how influence flows forward in time, with edges running predominantly from left to right. The community bands show that most ties remain embedded within stylistic or institutional clusters, yet a number of cross-band connections bridge different traditions and cohorts. A more detailed analysis of the network's communities and their culinary significance is presented in Section 2.2.5.

Taken together, these results show that the network's directionality largely reflects the generational transmission of knowledge. Earlier-born chefs function as genealogical anchors spanning multiple cohorts, while later-born chefs appear primarily as receivers of influence. This temporal ordering mirrors other apprenticeship-based systems, where lineage structures arise from mentor–disciple ties. Two caveats are worth noting.

First, older cohorts are under-represented in the data due to limited documentation and the *Atlante*'s focus on postwar gastronomy, which may understate the role of early pioneers whose influence was real but less systematically recorded. Second, career length amplifies the out-degree of senior chefs: those with decades-long trajectories have naturally had more opportunities to train apprentices and build collaborations. As a result, out-degree reflects not only structural centrality but also career duration. These limitations suggest that temporal comparisons should be read with caution, since observed patterns blend genuine

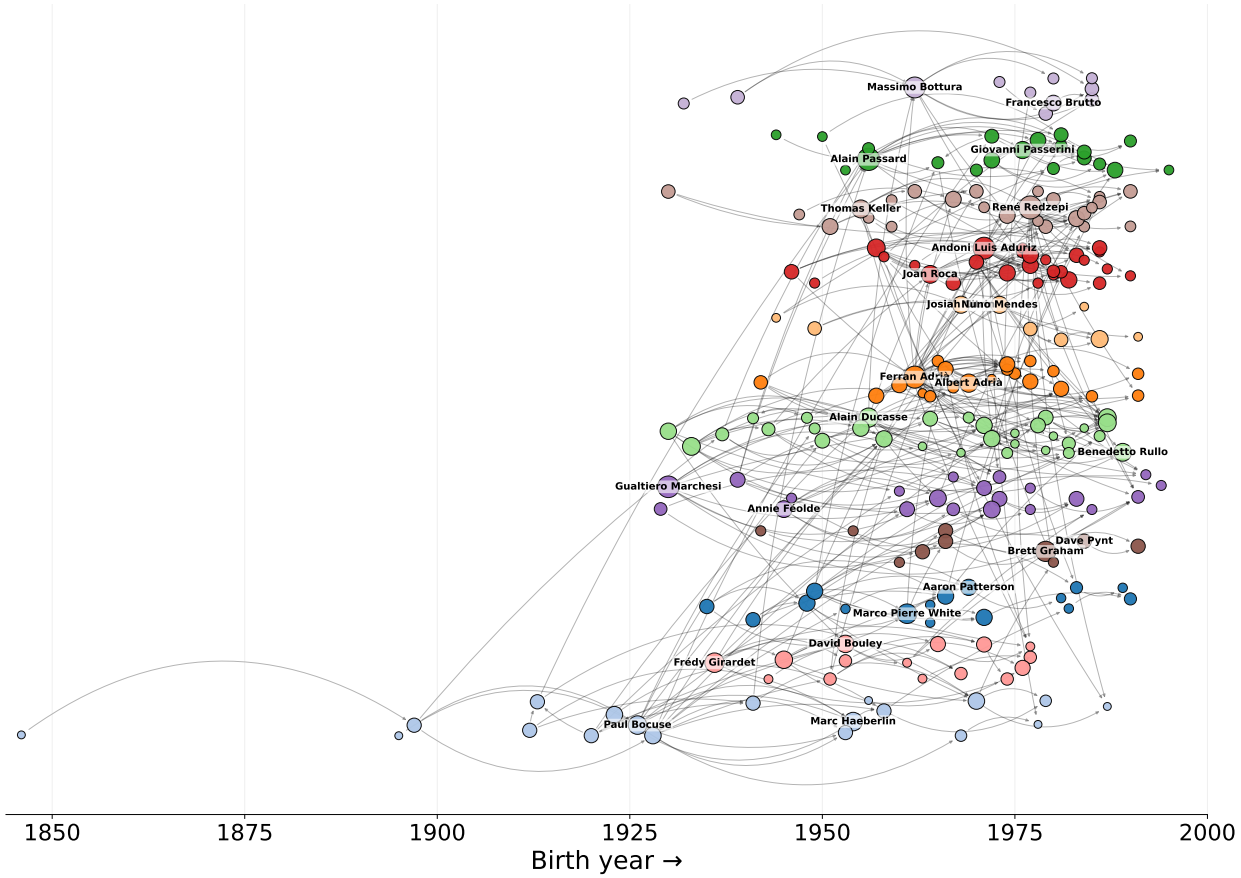


Figure 5.8: **Temporal layout of the chef influence network.** Visualization of the largest weakly connected component, with each node placed at its birth year on the horizontal axis and assigned to a horizontal "band" corresponding to a community detected via Louvain modularity on the undirected projection. Within bands, nodes are arranged using a spring layout with small vertical jitter and collision avoidance. Directed edges are drawn as light curved lines, bending upward or downward depending on relative vertical order. Node size reflects within-community prominence (z^w-score of within-module degree), and color encodes community identity. The two most prominent hubs in each community are labeled. The layout highlights how influence flows across generations (x-axis) while remaining structured within community bands (y-axis).

generational effects with artifacts of documentation and life-cycle span.

5.3.4 Central figures in the influence network

Who are the most central figures in the chef influence network? Centrality measures provide complementary answers, each capturing a different aspect of prominence. The most natural starting point is *degree centrality*, which counts the number of ties adjacent to a chef. In a directed network, this splits into *in-degree* (number of incoming ties) and *out-degree* (number of outgoing ties).

Table 5.2: **Top chefs by in- and out-degree centrality.** In-degree measures how many mentors or predecessors point to a chef; out-degree measures how many successors or mentees a chef has influenced.

	Top In-Degree			**Top Out-Degree**		
Rank	Chef	Birth	Degree	Chef	Birth	Degree
1	Lorenzo Stefanini (Italy)	1987	7	Ferran Adrià (Spain)	1962	20
2	Benedetto Rullo (Italy)	1989	7	René Redzepi (Denmark)	1977	16
3	Stefano Terigi (Italy)	1987	7	Andoni Luis Aduriz (Spain)	1971	11
4	Michele Lazzarini (Italy)	1991	6	Alain Ducasse (France)	1956	10
5	Matthew Orlando (USA)	1977	5	Alain Passard (France)	1956	10
6	Simone Tondo (Italy)	1988	5	Pierre Troisgros (France)	1928	9
7	David Bouley (USA)	1953	5	Paul Bocuse (France)	1926	8
8	Gordon Ramsay (UK)	1966	5	Gualtiero Marchesi (Italy)	1930	8
9	Karime Lopez (Mexico)	1982	5	Martín Berasategui (Spain)	1960	7
10	Manoella Buffara (Brazil)	1983	5	Seiji Yamamoto (Japan)	1970	7

- A high **in-degree** indicates that many predecessors are linked to a chef. These chefs are widely acknowledged as heirs within the system, accumulating recognition through multiple mentorships or inspirations.
- A high **out-degree** identifies chefs who influenced many others. These figures function as transmitters of knowledge, training numerous protégés or shaping international trends.

Table 5.2 lists the chefs with the highest in- and out-degrees. The contrast is striking, and consistent with the generational patterns identified in Sec. 5.3.3. The in-degree ranking is dominated by younger figures, mostly born in the 1980s and 1990s (e.g., Lorenzo Stefanini, Benedetto Rullo, Michele Lazzarini), who are heavily connected to older mentors and thus appear as receivers of accumulated influence. By contrast, the out-degree ranking is led by senior innovators such as Ferran Adrià (20 outgoing ties), René Redzepi (16), and Alain Ducasse (10), chefs who trained large cohorts or catalyzed new culinary movements. French pioneers like Paul Bocuse and Pierre Troisgros also appear prominently, reflecting their foundational role in transmitting the principles of Nouvelle Cuisine. In sum, in-degree

highlights chefs positioned as *heirs* within the system, whereas out-degree captures those who acted as *sources*, shaping subsequent generations.

While in- and out-degree provide the most immediate measure of chef prominence—respectively as receivers and sources of influence—other centrality measures capture complementary dimensions of importance in the network:

Betweenness centrality. Betweenness measures how often a node lies on the shortest paths connecting pairs of other nodes [Newman, 2018]. A chef with high betweenness is not necessarily the one with the largest number of ties, but rather the one strategically positioned to channel influence between otherwise distant groups. In culinary terms, these are brokers who bridge national cuisines or stylistic schools: chefs who trained abroad and brought techniques back home, or figures who mentored across different traditions. Their importance lies in enabling the circulation of knowledge between communities that would otherwise remain disconnected.

PageRank centrality. PageRank extends the idea of in-degree by weighting incoming ties according to the importance of their sources [Page et al., 1999]. In practice, being endorsed by an already prestigious mentor counts more than being endorsed by a peripheral figure. Thus, a chef may have only a few incoming ties, but if those ties come from highly influential chefs, their PageRank will be high. This measure highlights chefs embedded in elite genealogical chains, whose prestige derives not only from the quantity of their mentors but also from the quality of their lineage. It captures a recursive notion of influence: the prestige of a chef is amplified by the prestige of those who trained or inspired them.

Information Centrality Global efficiency measures how easily information flows through a network, defined as the average inverse shortest path length between all pairs of nodes [Latora and Marchiori, 2007]:

$$E(G) \;=\; \frac{1}{|V|(|V|-1)} \sum_{i \neq j} \frac{1}{d_{ij}},$$

where d_{ij} is the shortest path length between i and j. For each chef v, we define the *efficiency loss* as

$$\Delta E(v) \;=\; E(G) - E(G \setminus v),$$

that is, the drop in global efficiency after removing v and all its incident ties. Chefs with large information centrality are those whose absence makes the network substantially less navigable, increasing the distance between otherwise connected regions. They act as structural linchpins: their removal not only cuts direct ties but also undermines the cohesion of the system as a whole. In contrast to degree or PageRank—which emphasize visibility and prestige—efficiency loss highlights chefs who sustain the global backbone of culinary influence.

The three rankings in Table 5.3 illustrate how different notions of centrality bring distinct chefs to the fore. Betweenness centrality points to two well-known figures of Nouvelle Cuisine, Alain Ducasse and Alain Passard, who often appear on paths linking otherwise separate

Table 5.3: **Top chefs by alternative centrality measures.** The three rankings highlight complementary roles: betweenness emphasizes brokers who connect traditions, PageRank rewards chefs embedded in prestigious lineages, and efficiency loss identifies systemic anchors whose removal would disrupt global cohesion.

Betweenness	**PageRank**	**Efficiency Loss**
Alain Ducasse (FR, 1956)	Jordan Bailey (UK, 1990)	Ferran Adrià (ES, 1962)
Alain Passard (FR, 1956)	Isaac McHale (UK, 1980)	René Redzepi (DK, 1977)
Fernand Point (FR, 1897)	Brett Graham (AU, 1979)	Massimo Bottura (IT, 1962)
Massimo Bottura (IT, 1962)	Antonio Biafora (IT, 1984)	Brett Graham (AU, 1979)
Marco Pierre White (UK, 1961)	Chiara Pavan (IT, 1985)	Alain Passard (FR, 1956)
Michel Troisgros (FR, 1958)	Jake Kellie (AU, 1991)	Alain Ducasse (FR, 1956)
Alain Chapel (FR, 1937)	Jeremiah Stone (US, 1984)	Frédy Girardet (CH, 1936)
Ferran Adrià (ES, 1962)	Gianluca Gorini (IT, 1983)	Michel Troisgros (FR, 1958)
Pierre Troisgros (FR, 1928)	Francesco Brutto (IT, 1980)	Andoni Luis Aduriz (ES, 1971)
Paul Bocuse (FR, 1926)	Manoella Buffara (BR, 1983)	Thomas Keller (US, 1955)

parts of the network. The ranking also includes earlier pioneers: four of the top ten chefs were born before 1950 (Fernand Point, Pierre Troisgros, Paul Bocuse, Alain Chapel). This pattern reflects the structural logic of betweenness: chefs active in earlier generations are more likely to lie on the shortest paths that connect historical pioneers to later heirs. Their prominence comes less from the number of direct ties than from their role as intermediaries along genealogical chains. In this sense, figures of the Nouvelle Cuisine era remain visible as brokers in the historical transmission of techniques and styles.

PageRank, by contrast, elevates younger chefs such as Jordan Bailey and Isaac McHale, whose ties to prestigious mentors increase their standing beyond what raw degree would suggest. A notable feature is that the top ten PageRank chefs are all born after 1980. This outcome is consistent with two structural features of the data. First, the network is *nearly acyclic in time*: as shown in Sec. 5.3.3, 97% of edges run from older to younger chefs. Because PageRank propagates prestige along the direction of edges, endorsement tends to flow forward in time. Prestigious, older mentors (high out-degree) distribute their PageRank to those they influenced; younger chefs then accumulate prestige without feeding it back upstream, since backward edges are rare. Second, many of these younger chefs have relatively few outgoing ties, so their incoming prestige is not heavily diluted through redistribution. In this setting, PageRank answers the question "*who is endorsed by prestigious mentors?*" rather than "*who trained many protégés?*" — which helps explain why recent heirs embedded in elite genealogies (e.g., Bailey, McHale, Pavan, Lazzarini) rank highly.

Two clarifications help interpret this result. (i) Teleportation ($\alpha = 0.85$) prevents rank from collapsing entirely onto sinks, but does not remove the directional bias: in a graph

where influence runs older → younger, PageRank still favors downstream nodes that receive links from high-status sources. (ii) If edge directions were reversed (i.e., edges from mentee to mentor), PageRank would instead highlight senior figures; with the present definition, it emphasizes lineage prestige among heirs. Taken together, PageRank complements degree centrality: out-degree highlights chefs who acted as *sources* of influence, while PageRank highlights chefs whose position reflects the prestige of their predecessors.

Information centrality identifies chefs whose removal would reduce the overall efficiency of the network. The top of the ranking includes Ferran Adrià, René Redzepi, and Massimo Bottura, chefs whose careers combined high visibility with connections across different parts of the network. Removing such figures would increase distances between groups and make the network less cohesive. Chefs with high information centrality can therefore be seen as important channels of culinary knowledge, sustaining the accessibility of the network as a whole.

Interestingly, several of these chefs are also widely recognized as innovators.. Ferran Adrià pioneered molecular gastronomy and experimental techniques at El Bulli [Encyclopaedia Britannica Editors, 2024a]. René Redzepi shaped the New Nordic movement through Noma, redefining locality and seasonality [Nordic Council of Ministers, 2004]. Massimo Bottura introduced a playful, narrative-driven approach to Italian cuisine at Osteria Francescana [Michelin Guide Editors, 2024b]. Alain Passard (ranked 5th) revolutionized fine dining with vegetable-centered menus at L'Arpège [Michelin Guide Editors, 2024a], while Alain Ducasse (6th) created a global empire that spread French haute cuisine internationally [Encyclopaedia Britannica Editors, 2024b]. Frédy Girardet (7th) was celebrated for elevating nouvelle cuisine in Switzerland, and Michel Troisgros (8th) carried forward one of the foundational French lineages while also pushing modernist refinements. Interestingly, Brett Graham (4th) ranks highly on both information centrality and PageRank (3rd). The latter reflects his training under prestigious mentors, while the former suggests that his career—rooted in the UK but consolidated in Australia—links otherwise distant segments of the network: Graham is notable for his role in adapting fine dining to the Australian context, bridging culinary worlds.

Thus, information centrality does not merely highlight structural brokers, but tends to align with chefs who introduced influential ideas and styles. Their systemic importance lies not only in holding the network together, but also in sustaining the diffusion of innovations across its structure.

5.3.5 Central ties in the influence network

While node centrality highlights the chefs who occupy prominent positions in the network—whether as influential mentors who serve as channels of culinary knowledge or as heirs who accumulate prestige from their predecessors—not all of a chef's relationships contribute equally to shaping the network. Some ties play a disproportionately important role in sustaining genealogical continuity and cohesion. Edge-level centrality measures capture this dimension by identifying mentorships or inspirational links whose presence strengthens connectivity across generations and traditions, and whose absence would fragment the flow of influence.

Table 5.4: **Top 10 edges by *edge betweenness* — directed vs. undirected (LCC).** Type codes: W = working together; I = inspiration; U = uncertain.

Directed				**Undirected projection**			
Source	Target	T	Betw.	Chef 1	Chef 2	T	Betw.
A. Escoffier	Fernand Point	I	0.001 508	Brett Graham	Isaac McHale	W	0.058 443
Alain Chapel	Alain Ducasse	W	0.001 369	Ferran Adrià	René Redzepi	W	0.057 016
Fernand Point	Alain Chapel	W	0.001 190	René Redzepi	Isaac McHale	W	0.051 530
Juan M. Arzak	Ferran Adrià	I	0.001 151	Ferran Adrià	R. Cerea	U	0.047 640
Fernand Point	P. Troisgros	W	0.001 151	Ferran Adrià	Nuno Mendes	W	0.045 657
Marco P. White	H. Blumenthal	W	0.000 972	Ferran Adrià	M. Bottura	I	0.043 776
Alain Ducasse	M. Bottura	W	0.000 972	Alain Ducasse	M. Bottura	W	0.040 568
M. Troisgros	Bernard Loiseau	W	0.000 694	P. Troisgros	R. Cerea	W	0.038 148
Gaston Boyer	Alain Passard	W	0.000 665	Raymond Blanc	R. Camanini	W	0.034 168
Gérard Boyer	Alain Passard	W	0.000 665	Ferran Adrià	Joan Roca	W	0.033 727

Edge betweenness centrality. Edge betweenness counts how often a tie lies on shortest paths between pairs of chefs [Girvan and Newman, 2002]. High-betweenness edges act as bridges across genealogical branches. In the directed network (left panel of Table 5.4), the top links trace a classic chronological chain—Escoffier → Point, Point → Chapel, Chapel → Ducasse—alongside cross-generational inspirations (Arzak → Adrià) and formative apprenticeships (White → Blumenthal). These ties sit on many directed routes that carry influence forward in time, so they are repeatedly traversed by shortest paths.

To probe the role of direction, we recomputed edge betweenness on the undirected projection (right panel of Table 5.4). The ranking shifts markedly toward dense avant-garde and contemporary clusters: Adrià–Redzepi, Adrià–Bottura, Adrià–Mendes, Adrià–Roca, and Brett Graham–Isaac McHale top the list, with additional UK/Italian links (e.g., Raymond Blanc–Riccardo Camanini). In an undirected setting, multiple alternative routes within tightly knit schools reinforce one another, pushing intra-cluster ties upward; in the directed setting, the asymmetry imposed by time accentuates long genealogical chains. Some relationships are robust across both views (e.g., Ducasse–Bottura), but the emphasis differs: directed betweenness foregrounds intergenerational pathways, whereas undirected betweenness elevates hubs that stitch together cohesive, contemporaneous groups.

Edge information centrality. Edge information centrality quantifies how much the removal of a tie reduces the network's overall efficiency [Latora and Marchiori, 2007]. In the directed network (left panel of Table 5.5), the ranking again emphasizes the canonical French lineage: Escoffier → Point, Point → Chapel, Chapel → Ducasse, together with the inspirational link Arzak → Adrià and formative mentorships such as Marco Pierre White → Blumenthal. These are edges whose absence would disproportionately impede information flow along historical genealogical paths.

On the undirected projection (right panel of Table 5.5), however, the picture changes. Here the top edges are not the classic French lineages but rather ties within contemporary clusters: Keller–Colicchio in the U.S., Albert Adrià–Paolo Griffa across Spain and Italy, Brett Graham–Jake Kellie in the UK–Australia axis, and Redzepi–Ambrosino in the New

Table 5.5: **Top 10 edges by *information centrality* — directed vs. undirected (LCC).** Type codes: W = working together; I = inspiration; U = uncertain.

Directed				Undirected projection			
Source	Target	T	Info.	Chef 1	Chef 2	T	Info.
Juan M. Arzak	Ferran Adrià	I	0.029 321	T. Keller	T. Colicchio	W	0.019 416
Alain Chapel	Alain Ducasse	W	0.027 035	Albert Adrià	Paolo Griffa	I	0.019 222
A. Escoffier	Fernand Point	I	0.026 379	Brett Graham	Jake Kellie	W	0.016 863
M. P. White	H. Blumenthal	W	0.021 806	Liam Tomlin	Brett Graham	W	0.016 863
Alain Ducasse	M. Bottura	W	0.021 636	René Redzepi	M. Ambrosino	W	0.016 231
Fernard Point	Alain Chapel	W	0.018 186	M. Troisgros	Anne-S. Pic	U	0.014 787
M. Troisgros	Bernard Loiseau	W	0.015 878	B. Loiseau	Alex Atala	W	0.013 700
G. Lenôtre	Alain Ducasse	W	0.014 079	Tom Aikens	Marco Martini	W	0.013 691
B. Loiseau	M. Colagreco	W	0.013 422	H. Blumenthal	James Lowe	W	0.013 581
Fernand Point	P. Troisgros	W	0.013 041	S. Yamamoto	T. Kawada	W	0.012 986

Nordic circle. The undirected measure elevates connections that strengthen cohesion within stylistic schools and geographic clusters, rather than those that bridge across generations.

Taken together, the two views show complementary facets of influence. Directed information centrality highlights the indispensability of long historical mentorship chains for maintaining genealogical depth, while undirected information centrality foregrounds ties that hold present-day schools together. Both perspectives converge on the idea that a handful of relationships are disproportionately important, but differ in whether they spotlight intergenerational bridges or intra-school cohesion.

Summary

The descriptive analysis shows that influence among chefs is distributed in a markedly asymmetric way, both structurally and temporally. Most chefs establish only a handful of ties, while a minority accumulate disproportionately many, especially as mentors. This inequality reflects career dynamics: formative years permit only a few mentors, whereas decades of leadership allow some chefs to train or inspire dozens. Temporally, nearly all ties flow from older to younger chefs, with a typical generational gap of about 18 years, producing a hierarchical structure where senior figures anchor deeper genealogical trees and younger chefs accumulate influence as heirs.

Centrality measures add further nuance. Degree centrality distinguishes between chefs who receive influence (high in-degree, typically younger cohorts) and those who transmit it (high out-degree, often established mentors and innovators). Betweenness centrality highlights earlier pioneers, especially from the Nouvelle Cuisine era, whose careers continue to mediate long genealogical paths. PageRank favors recent heirs embedded in prestigious lineages, reflecting the directional and nearly acyclic nature of the network: prestige flows forward in time, amplifying the standing of younger chefs endorsed by renowned mentors. Information centrality, in turn, identifies figures whose removal would reduce the overall cohesion of the network, a group that overlaps substantially with chefs recognized as innovators. In this sense, chefs like Ferran Adrià, René Redzepi, Massimo Bottura, and Alain

Passard are central not only structurally but also in shaping new culinary styles.

Edge-level measures complement this picture by showing that not all relationships carry the same weight. Some ties—such as Alain Chapel's mentorship of Alain Ducasse, or Juan Mari Arzak's inspiration of Ferran Adrià—stand out as disproportionately important for sustaining genealogical continuity. Others, like the symbolic link between Escoffier and Fernand Point, illustrate how influence can bridge distinct culinary eras. Comparing directed and undirected versions of edge centrality further reveals a duality: in the directed network, classical mentorship chains dominate, while in the undirected projection, redundancy within avant-garde clusters (e.g., around Ferran Adrià or René Redzepi) boosts centrality. This contrast underscores that edge centrality captures both genealogical depth and intra-school cohesion, depending on whether temporal directionality is enforced.

Together, these perspectives show that prominence in the chef network cannot be reduced to a single dimension: different measures highlight heirs, mentors, brokers, and systemic anchors, offering complementary views on how influence is organized and sustained.

5.4 Communities of Influence

Collaboration networks frequently exhibit homophily by nationality: geographic and institutional proximity lowers the barriers to working together, while cultural and linguistic commonality enhances trust and communication [Ponds et al., 2007, Hoekman et al., 2010]. In the culinary domain—where training typically occurs within kitchens, schools, and domestic networks—these mechanisms suggest that chefs are more likely to mentor or collaborate with others from the same country. Beyond structural proximity, additional forces reinforce nationality-aligned ties: awareness through domestic media and guides; professional opportunities via local contacts; training pipelines anchored in national institutions; and linguistic, regulatory, and cultural continuity in culinary traditions. These factors together make it plausible that working relationships will cluster along national boundaries, even as some chefs transcend them.

5.4.1 Nationality and assortativity

To move from qualitative expectations to quantitative evidence, we examine the alignment of ties with nationality in the directed chef influence network. Let $G = (V, E)$ be the directed graph, where an edge $u \to v$ indicates that chef u influenced chef v. Each chef v is associated with a categorical nationality $s(v) \in \mathcal{C}$.

The distribution of ties across nationalities is captured by the *nationality mixing matrix* $E = \{e_{ij}\}$, defined as

$$e_{ij} \;=\; \frac{1}{|E|}\,\#\{\, u \to v \in E :\; s(u) = i,\; s(v) = j \,\}, \qquad \sum_{i,j} e_{ij} = 1. \tag{5.1}$$

Here, e_{ij} is the fraction of edges originating from nationality i and pointing to nationality j. Diagonal elements ($i = j$) measure the prevalence of *domestic* ties, while off-diagonal

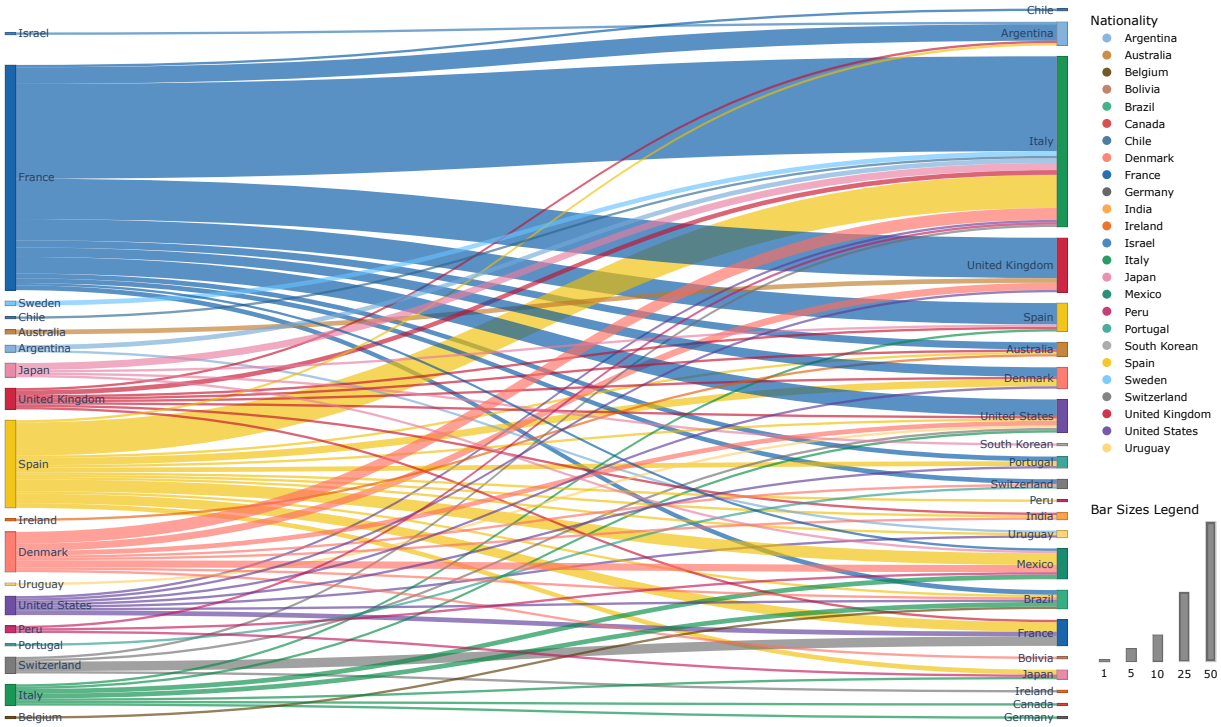


Figure 5.9: **Cross-national flows of influence between chefs.** Alluvial diagram of mentorship and collaboration ties running between chefs of different nationalities (same-nationality ties excluded). Stream width is proportional to the number of directed edges from source nationality (left) to target nationality (right). The visualization highlights France as a dominant exporter of influence, Italy as a major receiver, and Spain as a bridge toward Latin America and Northern Europe, while other traditions appear more domestically oriented.

elements trace the extent and direction of *cross-national* influence flows. Row sums

$$a_i = \sum_j e_{ij}$$

give the fraction of all outgoing ties emitted by chefs of nationality i, while column sums

$$b_j = \sum_i e_{ij}$$

give the fraction of incoming ties received by chefs of nationality j.

From the mixing matrix we obtain a scalar summary of homophily by nationality: the directed assortativity coefficient [Newman, 2003],

$$r_{\text{dir}} \;=\; \frac{\sum_i e_{ii} \;-\; \sum_i a_i b_i}{1 \;-\; \sum_i a_i b_i}. \tag{5.2}$$

The numerator compares the observed fraction of within-nationality ties ($\sum_i e_{ii}$) to the expected fraction under independence ($\sum_i a_i b_i$), while the denominator rescales to the maximum possible deviation.

Empirical results. In our data ($|E| = 414$), the observed share of same-nationality ties is

$$p_{\text{same,obs}} = \sum_i e_{ii} = 0.507,$$

whereas the expected baseline is

$$p_{\text{same,exp}} = \sum_i a_i b_i = 0.164.$$

This yields an excess of $\Delta = 0.343$, a lift of 3.09, and a directed assortativity of

$$r_{\text{dir}} = 0.410.$$

Substantively, this means that chefs of the same nationality are linked about three times more often than expected by chance, even after adjusting for differences in how often nationalities appear as sources or targets. Because the baseline accounts for compositional imbalance, the excess reflects genuine structuring: culinary influence flows disproportionately within national contexts.

At the same time, the assortativity coefficient in Eq. 5.2 collapses all international ties into a single scalar, and therefore cannot reveal *which* cross-national connections are most prominent. To examine these patterns, we turn back to the nationality mixing matrix defined in Eq. 5.1.

Whereas the coefficient relied on its normalized form, here we focus on the off-diagonal entries expressed in absolute counts, which capture the number of ties flowing from nationality i to nationality j. The resulting structure, visualized in Fig. 5.9, makes the asymmetries of international exchange explicit. French chefs dominate as exporters, shaping

both domestic successors and foreign traditions. Italian chefs, by contrast, appear primarily as receivers, with more than half of their incoming cross-national ties originating from France—underscoring a privileged channel of influence between the two countries. Similarly, the United Kingdom shows a strong dependence on French chefs, reflecting the historic role of French haute cuisine in shaping modern British fine dining. Spanish chefs, led by figures such as Ferran Adrià, act as catalysts of innovation into Latin America and Northern Europe, consolidating Spain's role as a global avant-garde hub. Denmark, anchored by René Redzepi and the New Nordic movement, exerts influence across continents, with ties extending not only to neighboring European countries but also to Asia and the Americas. Other national traditions appear less influential overall in this dataset, contributing comparatively fewer cross-national ties than the major exporters identified above. Together, these cross-border flows extend the assortativity analysis by showing not only the strength of nationality-based homophily but also the main channels through which influence moves between countries. At the same time, nationality alone does not fully explain the structure of the network: many chefs are embedded in stylistic schools or mentorship lineages that cut across borders. Exploring these additional principles of organization motivates the community analysis that follows.

5.4.2 Community Structure

To move beyond individual chefs and dyadic ties, we investigate how influence organizes into broader groups of chefs. In network terms, such groups correspond to *communities*—sets of nodes that are more densely connected internally than to the rest of the graph [Fortunato, 2010]. Detecting these meso–level structures helps assess whether the field of gastronomy fragments into identifiable schools, lineages, or institutional traditions.

Methodological Setup

We restrict attention to the strongest ties: *working–together* relationships (apprenticeships, collaborations, or direct professional interactions). Edges classified as *inspiration without collaboration* or *uncertain* are excluded, as they capture more diffuse or unverifiable links. Including them would risk diluting the signal of sustained social relationships that community detection is designed to capture.

From this filtered subnetwork, we analyze the largest weakly connected component, which comprises 213 chefs and 346 directed edges (344 when symmetrized). Because our goal is to identify cohesive groups rather than trace the precise direction of mentorship, we treat edges as undirected. This choice is common in practice; while directed modularity exists [Leicht and Newman, 2008], its interpretation for apprenticeship–style ties is less straightforward than the undirected case.

Communities are detected using the Louvain algorithm for modularity maximization [Blondel et al., 2008], via the `python-louvain` implementation.[2] Modularity quantifies how much denser ties are within communities than expected under a null model preserving the degree sequence [Newman, 2006]. There is no universal threshold for "high" modularity—values

[2] `https://github.com/taynaud/python-louvain`.

depend on network size, degree heterogeneity, and method—but, in empirical social and collaboration networks, values around $Q \gtrsim 0.5$–0.6 are typically interpreted as strong structure [see discussion and caveats in Fortunato, 2010]. We verified that the partitions were stable across multiple Louvain runs, mitigating sensitivity to the algorithm's stochasticity. This approach yields interpretable and robust groupings of chefs whose working relationships form significantly denser clusters than chance alone would predict.

Global results

Applying the Louvain algorithm to the undirected projection of the working–together subnetwork yields a modularity of $Q = 0.633$. Such a high modularity indicates strong community structure: chefs linked by working relationships are much more likely to cluster into distinct groups than to spread evenly across the network. In substantive terms, this means that professional collaborations and apprenticeships are highly assortative, concentrating within relatively cohesive groups rather than dispersing across the field.

The partition contains 13 distinct communities. This number is neither so large as to suggest fragmentation into isolated dyads, nor so small as to imply one dominant bloc. Instead, the network resolves into clusters consistent with the idea that haute cuisine is organized into recognizable schools, genealogies, or stylistic movements.

Fig. 5.10 illustrates these clusters, with node colors corresponding to community membership. At a broad level, three major clusters dominate the landscape: the classic traditions of French haute cuisine and Nouvelle Cuisine (pink); a global avant-garde bridge connecting New Nordic, the *elBulli* movement, and related experimental cuisines (green); and the British modern haute lineage rooted in the Roux brothers (light blue).

The first cluster (35 chefs) is anchored by figures such as Paul Bocuse, Alain Ducasse, Michel Guérard, and Gaston Lenôtre. It represents the lineage of French haute cuisine and Nouvelle Cuisine, radiating internationally through disciples such as Daniel Boulud and Heinz Beck [Rao et al., 2003].

The second cluster (32 chefs) centers on René Redzepi, Andoni Luis Aduriz, and Joan Roca. It maps onto the avant-garde international movement, combining New Nordic, Spanish molecular gastronomy, and American experimental fine dining. Prominent chefs in this cluster include Grant Achatz, Gaggan Anand, and Thomas Keller, underscoring its role as a transnational bridge of contemporary innovation.

The third cluster (20 chefs) is dominated by Albert and Michel Roux, Marco Pierre White, Gordon Ramsay, Heston Blumenthal, and Pierre Koffmann. This community corresponds to the UK-based lineage rooted in the Roux brothers, extending into the modern British scene and shaping the evolution of contemporary British haute cuisine.

Other medium-sized clusters capture distinct traditions: Massimo Bottura's transcontinental tree (turquoise), linking Italy, Japan, and Latin America and exemplifying globalization of haute cuisine; and the Italian haute cuisine lineage centered on Gualtiero Marchesi (ruby red), regarded as the father of modern Italian gastronomy.

Smaller but coherent groups highlight stylistic niches such as fire-based cooking (Bittor Arginzoniz, Dave Pynt) or cosmopolitan hubs of contemporary fine dining (Andreas Caminada, Nuno Mendes, Wolfgang Puck). A particularly notable community is the cluster centered on Ferran Adrià, together with his close circle of *elBulli* collaborators, which stands

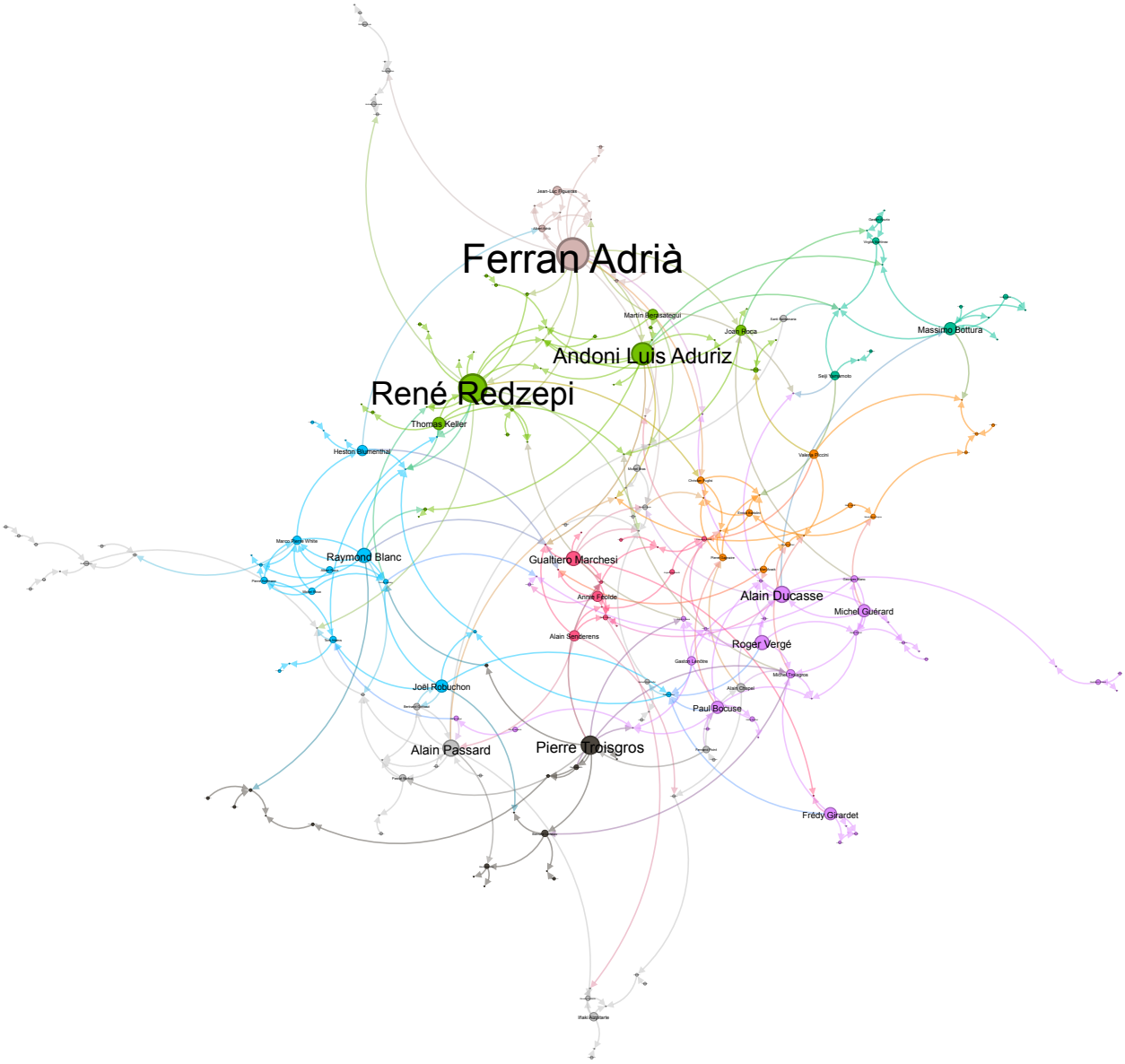


Figure 5.10: **Communities in the working–together network.** Visualization of the largest weakly connected component (213 chefs, 344 undirected edges). Nodes are colored by Louvain community, sized by out-degree, and labeled when total degree exceeds 5. The modularity of the partition is $Q = 0.633$, indicating strong community structure.

out as a laboratory of experimental innovation within the broader avant-garde movement [Capdevila et al., 2018].

Community detection thus recovers the socio-historical structure of modern gastronomy: from the dominant genealogies of French, British, Italian, *elBulli*, and New Nordic traditions to the smaller stylistic niches that enrich the field.

Nationality alignment

How strongly do these communities align with nationality? The global work layer shows moderate nationality assortativity ($r_{\text{dir}} = 0.392$, Eq. 5.2), confirming that chefs are more likely to collaborate with compatriots. Within individual communities, however, assortativity is lower on average ($\bar{r}_w = 0.179$), suggesting that while national homophily is present, it does not fully explain community boundaries.

We quantify this more directly using information–theoretic measures of similarity between partitions. Normalized mutual information (NMI) captures how much knowing a chef's community reduces uncertainty about their nationality, while Adjusted mutual information (AMI) corrects for agreement expected by chance [Danon et al., 2005, Vinh et al., 2010]. In our data, NMI = 0.372 and AMI = 0.215. These values indicate a nontrivial but far from perfect alignment: nationality matters, but communities also capture other organizing forces, such as stylistic schools, mentorship lineages, and transnational avant-garde movements.

To assess whether this alignment exceeds what is expected by chance given the global homophily of the network, we compare to a constrained rewired-graph null model (Appendix C). This null preserves both the degree sequence and the nationality mixing matrix ((5.1)) while randomizing higher-order wiring. Across 500 null realizations, the observed NMI and AMI fall comfortably within null expectations, as does within-community assortativity ($\bar{r} = 0.179$). In other words, once overall nationality preferences are fixed, communities do not display additional nationality alignment. By contrast, cohesion measures deviate modestly from the null: embeddedness is significantly higher ($p \approx 0.022$), with coverage and conductance also trending in the cohesive direction.

These results show that while nationality shapes the chef network at a global level, community formation cannot be reduced to national blocs. Instead, the modular structure reflects stylistic and institutional traditions that cut across borders, from avant-garde movements to genealogical lineages, underscoring the multi-dimensional organization of modern haute cuisine.

5.4.3 Backbone of influence

In Section 5.3.5 we examined the centrality of individual ties, showing that some relationships play a disproportionate role in sustaining network cohesion. Here we use those same edge–centrality measures to construct an interpretable *backbone* of the chef influence network. Specifically, we build a *maximum spanning tree* (MaxST) of the largest weakly connected component, weighting edges by their *directed information centrality*. This approach retains a spanning set of ties that maximizes overall structural informativeness, providing a reduced but meaningful genealogy of modern gastronomy.

For each directed edge in the network we compute its information centrality, defined as the relative drop in global efficiency if that edge is removed (see Sec. 5.3.5). These scores are min–max normalized, and the network is projected to an undirected form. When both directions exist between a pair of chefs, the higher of the two centrality scores is retained as the undirected weight. On this weighted projection, we compute the maximum spanning tree using standard algorithms (Kruskal/Prim with deterministic tie–breaking). By construction, the tree includes every chef in the component but only $|V| - 1$ edges, prioritizing those ties that most enhance global efficiency. For visualization, each undirected tree edge is oriented post hoc: we prefer the original direction when unique, otherwise fall back on the chronological rule *older* $\rightarrow$ *younger*. The tree is rooted at the oldest chef to preserve a generational layout.

It is important to stress that the MaxST should not be interpreted as a strict genealogy: edges do not all flow downward in time, and many genuine ties are omitted by the tree constraint. Its usefulness lies elsewhere. First, it highlights the pioneering anchors of modern gastronomy: following the root Auguste Escoffier and his successor Fernand Point, one immediately encounters the central figures of twentieth–century French cuisine such as the Troisgros family, Paul Bocuse, and Alain Chapel. Second, the tree reveals how the network organizes into distinct branches with different characteristics. For example, the lineage radiating from Ferran Adrià links chefs from multiple continents, illustrating the global reach of the Spanish avant–garde, while the branch associated with Alain Ducasse contains a high concentration of Italian protégés, reflecting his role in shaping Italy's contemporary fine dining. Third, the backbone makes visible points of confluence where major schools intersect. Massimo Bottura, for instance, sits at the junction of the classical French lineage of Ducasse and the experimental tradition of Adrià, embodying the synthesis of heritage and innovation.

The extracted backbone reflects both substantive and methodological choices. Because the MaxST necessarily contains only a single path between any two chefs, many valid influences are pruned away. Weight ties can in principle yield multiple alternative trees, though deterministic tie–breaking ensures reproducibility and sensitivity checks confirm robustness. Finally, since edge weights are derived from efficiency-based centrality, the backbone emphasizes structural importance rather than frequency or prestige: a relatively modest tie may appear if it is crucial to holding together distant parts of the network.

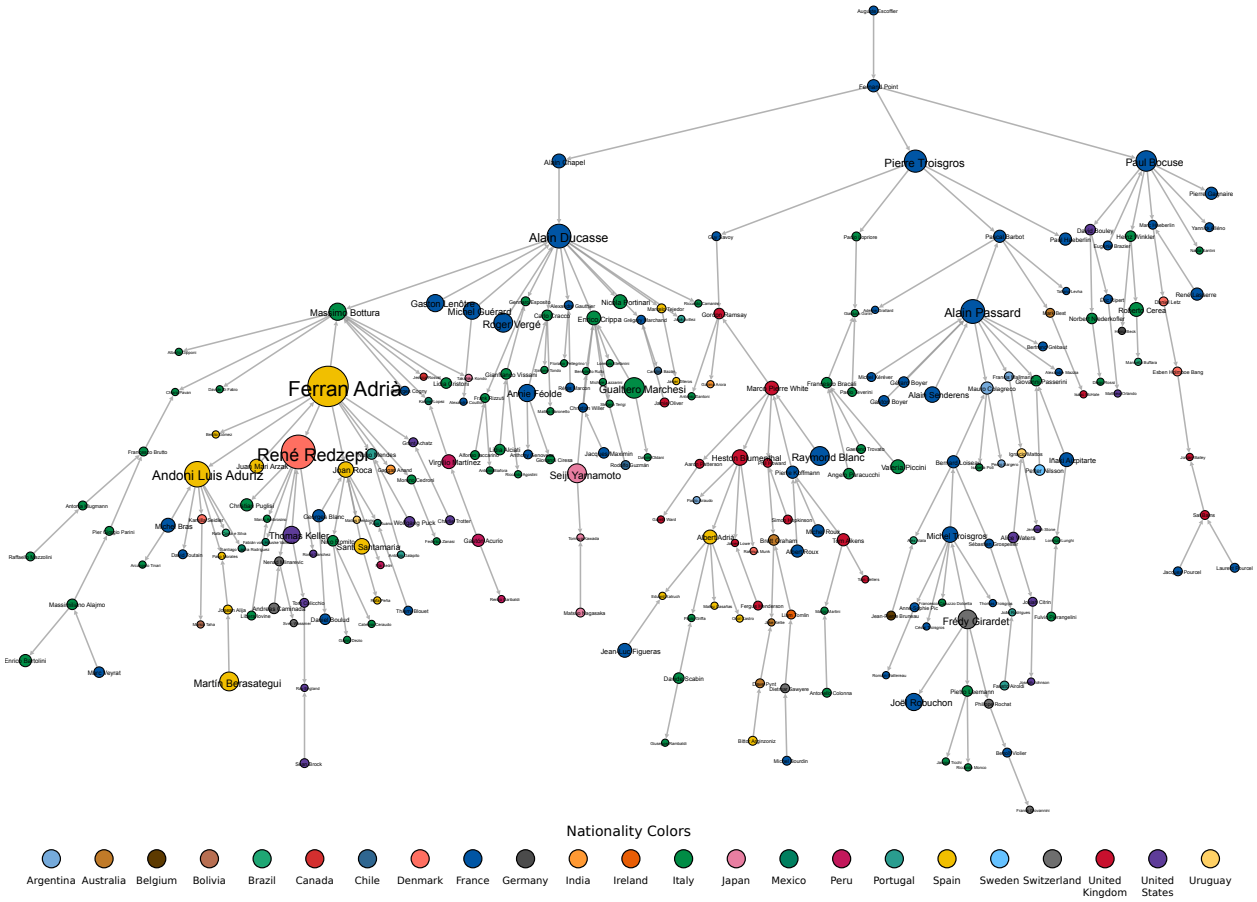


Figure 5.11: **Maximum centrality spanning tree of the chef influence network.** Shown is the maximum spanning tree (rooted at the oldest chef) extracted from the largest weakly connected component using *edge information centrality* as edge weight (cf. Section 2.2.2). Node color encodes nationality (see legend at the bottom). Chef names are displayed as node labels, and both node and label sizes are proportional to each chef's out-degree in the original directed network (number of downstream influence links). Edges are oriented for visualization by preferring the original direction when unique, otherwise from older to younger based on birth year.

5.5 Motif analysis of influence patterns

While centrality and community structure highlight prominent chefs and cohesive schools, they do not capture the finer building blocks of influence. As discussed in Section 2.2.6, network motifs are small subgraph patterns that occur more frequently than expected under appropriate null models, providing a complementary lens on the local architecture of influence [Milo et al., 2002]. Here, they serve as a lens on the architecture of influence, allowing us to test whether chefs tend to form linear genealogies, concentrate around pivotal figures, or branch outward into new schools. To evaluate whether the observed frequencies of motifs are meaningful, we compare them against a degree-preserving rewiring null model.

Degree-preserving rewiring null model. To assess whether observed motif patterns arise from genuine structural tendencies rather than from trivial degree heterogeneity, we employ a degree–preserving rewiring procedure as a null model. The algorithm repeatedly selects two directed edges $(u \to v)$ and $(x \to y)$ at random and swaps their endpoints to produce $(u \to y)$ and $(x \to v)$, provided this does not create self-loops or duplicate edges. Iterating this *double-edge swap* many times effectively randomizes the wiring of the network while preserving every node's in- and out-degree.

This construction generates an ensemble of randomized graphs that respect the local degree constraints of the empirical network, thereby controlling for the heterogeneous popularity of chefs as mentors (out-degree) or as stylistic influences (in-degree). Any systematic deviations between observed motif counts and the null distribution can thus be interpreted as evidence of non-random higher-order organization (e.g., clustering, hierarchy, or stylistic closure). Operationally, we generate B rewired graphs $G^{(1)}, \ldots, G^{(B)}$ from the largest weakly connected component of the directed network, using a burn-in of αm successful swaps and thinning of βm successful swaps between samples (with $m = |E|$). For each motif type k (triads or 4-node motifs), let $N_k(G)$ denote its count in the observed network and $N_k^{(b)} = N_k(G^{(b)})$ its count in the b-th null. We summarize departures from the null by

$$\rho_k \;=\; \frac{N_k(G)}{\mu_k}, \qquad z_k \;=\; \frac{N_k(G) - \mu_k}{\sigma_k}, \quad \text{with} \quad \mu_k = \tfrac{1}{B}\sum_{b=1}^{B} N_k^{(b)}, \;\; \sigma_k^2 = \tfrac{1}{B-1}\sum_{b=1}^{B}(N_k^{(b)} - \mu_k)^2,$$

and estimate Monte Carlo p-values with the plus-one rule (right-tailed for over-representation, left-tailed for under-representation, or two-sided as appropriate).

Operationally, we generate $B = 1000$ rewired graphs $G^{(1)}, \ldots, G^{(B)}$ from the largest weakly connected component of the directed network. Each chain is initialized with a burn-in of $\alpha m = 50m$ successful swaps and subsequently thinned by $\beta m = 50m$ swaps between retained samples ($m = |E|$ is the number of edges). Random seeds were fixed for reproducibility.

Triadic motifs. To interpret the results, we rely on the standard triad census (see Section 2.2.6 for a complete description of all triad types). Applying this framework to triads, we find that most patterns occur at frequencies consistent with degree–preserving randomness. The key exception is the **transitive triple** (030T), which is strongly over-represented ($z \approx 5.1$, $p < 0.001$). This motif corresponds to feed–forward chains $A \to B$, $B \to C$,

complemented by a direct $A \to C$ tie. In genealogical terms, it captures cases where a grand-mentor, mentor, and pupil are all connected, reinforcing hierarchical closure across generations. Its excess prevalence reflects the strongly layered, temporally ordered nature of the chef network, in which senior figures often influence not only their direct apprentices but also later generations.

By contrast, the **in-star** motif (021U)—two mentors converging on a single chef—is significantly under-represented ($z \approx -3.0$, $p \approx 0.007$). This suggests that chefs rarely accumulate many direct mentors, consistent with career trajectories in which formative influence is concentrated in one or two long-term relationships rather than dispersed across multiple teachers.

All other triad types, including reciprocal dyads (e.g. 111U/111D), cycles (030C), and more complex mixed configurations (120*, 201), occur at levels indistinguishable from the null. This absence of reciprocal and cyclic triads further underscores the near-acyclic temporal organization of the network: influence flows predominantly forward in time, leaving little room for feedback loops or mutual ties between contemporaries.

Four-node motifs. Beyond triads, we also consider a set of four-node patterns that capture branching and convergence of influence. Figure 5.12 illustrates the motif types included in our analysis, following the notation of Milo et al. [2004].

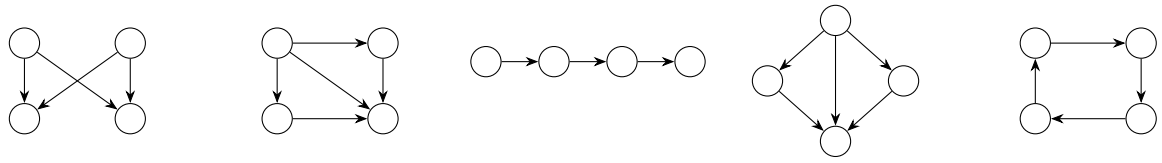


Figure 5.12: Directed 4-node motifs considered in the analysis. BiFan: two sources jointly targeting two sinks; FFBox: two feed-forward paths (A→B→D and A→C→D) plus direct A→D; P4: directed path of length 3; C4: directed 4-cycle; TT4: transitive tournament (acyclic complete orientation) on 4 nodes.

Among 4-node configurations, the most striking result is the massive over-representation of the **bi-fan** motif ($z! \approx !14.6$, $p! < !0.001$). This structure consists of two mentors simultaneously influencing two pupils, forming a rectangular pattern of cross-connections. Its abundance highlights the institutional dimension of culinary training: influential chefs often shape multiple apprentices at once, while their pupils in turn share common teachers. Such "teaching pairs" and "sibling cohorts" create dense micro-communities that reinforce stylistic continuity across multiple careers.

Other 4-node motifs show no significant deviation from the null. Path structures (P_4), cyclic squares (C_4), and transitive tetrahedra (TT_4) occur at frequencies indistinguishable from degree-constrained randomness, while the feed-forward box ($FFBox$) appears slightly enriched but not significantly so ($p! \approx !0.09$). The null results for these more complex motifs

suggest that higher-order organization in the chef network is not characterized by long acyclic chains or closed cycles, but rather by clustered broadcasting patterns captured by the bi-fan.

5.6 Summary and Outlook

In this chapter, we have examined the network of influence among chefs as documented in the *Atlante*, reconstructing the relational fabric through which culinary knowledge, mentorship, and stylistic innovation circulate. By treating chefs as nodes and their ties of collaboration or inspiration as directed edges, we have mapped the structural and temporal patterns that underlie the genealogy of modern gastronomy.

Our analysis has revealed several consistent features. First, influence is distributed in a markedly asymmetric way: while most chefs establish only a handful of ties, a minority accumulate disproportionately many, especially as mentors. This asymmetry is reinforced by temporal dynamics: nearly all ties flow from older to younger chefs, with a generational gap of roughly two decades, producing a layered structure in which senior figures anchor deep lineages and younger chefs inherit accumulated prestige.

Second, by applying a range of centrality measures, we have highlighted the multiple dimensions of prominence. Degree centrality distinguishes influential mentors from highly connected heirs, betweenness identifies brokers bridging genealogical branches, PageRank emphasizes the cumulative prestige of embedded heirs, and information centrality pinpoints chefs whose presence is vital for the cohesion of the network. Edge-level centralities complement this picture, showing that not all relationships carry the same weight: certain mentorships and inspirational ties emerge as disproportionately important conduits linking traditions and sustaining connectivity.

Third, we have shown that nationality exerts a strong structuring force. Chefs are more likely to collaborate with compatriots than chance would predict, yet cross-national flows reveal clear asymmetries: France dominates as an exporter of influence, Italy and the United Kingdom appear as major receivers, Spain consolidates its role as an avant-garde hub, and Denmark extends the reach of the New Nordic movement globally. Community detection further confirms that the network is organized into cohesive clusters, reflecting both national traditions and transnational schools, from the classic French and Italian lineages to the British modern haute scene, the experimental *elBulli* group, and the New Nordic movement.

Finally, by extracting a backbone in the form of a maximum spanning tree weighted by information centrality, we have distilled the dense network into a genealogy of indispensable ties. This representation foregrounds culinary pioneers, clarifies the branching structure of stylistic traditions, and highlights key points of confluence where lineages intersect, such as Massimo Bottura's position at the crossroads of French mentorship and Spanish innovation.

Two caveats must be acknowledged. The *Atlante* focuses primarily on postwar gastronomy, leaving earlier pioneers under-documented and covering only a fraction of chefs active today. Not all ties are equally verifiable, especially those classified as "uncertain," which require further corroboration. These limitations are difficult to overcome but must be borne in mind when interpreting the findings.

Building on the present work, several extensions are possible. One important step would be to expand the network: while it is practically impossible to reconstruct all active and inactive chefs in history, the dataset could be enlarged by starting from the existing nodes and systematically searching for missing ties, particularly across underrepresented regions and eras. Another promising line of development concerns the refinement of influence cate-

gories. The current three classes (*working together*, *inspiration*, *uncertain*) are intentionally broad, but a more fine-grained taxonomy could differentiate apprenticeships, long-term collaborations, short-term stages, and indirect influences, thereby offering a richer account of the diversity of professional relationships in gastronomy. Finally, influence does not operate on a single plane, and extending the analysis to multilayer and temporal networks would open new perspectives. Enriching the dataset with additional relational layers—such as the restaurants where chefs trained or worked—would allow us to capture the institutional dimension of influence, while temporal extensions would make it possible to follow how lineages, schools, and institutions rise, decline, and interact across decades.

Together, these developments would advance our understanding of how haute cuisine evolves not only through individual creativity but also through the relational and institutional structures that connect chefs across generations and traditions.

Chapter 6

Conclusions and Future Work

This thesis set out to characterize how culinary systems are structured across different scales, from molecules to cuisines to chefs, using a computational and network-theoretic lens. The motivation has been twofold: to shed light on the organizing principles of gastronomy and to contribute methodological tools to the emerging field of computational gastronomy. Drawing on large-scale datasets and methods from network science, we developed a multi-level framework that reconstructs culinary practice as the interplay of molecular feasibility, cultural organization, and human transmission.

6.1 Summary of contributions

At the molecular level, we asked whether chemistry provides a universal principle for ingredient pairing. By testing the "food pairing hypothesis" in a predictive framework, where ingredient recommendation was treated as a retrieval task, we contrasted flavor-based similarity derived from shared compounds with recipe-based similarity derived from empirical co-use. The results show that chemistry can occasionally aid creativity—particularly within certain categories or cuisines—but that its predictive power is limited. Recipe-based similarity consistently outperforms, pointing to cultural practice rather than molecular compatibility as the dominant organizing force.

At the cultural level, we investigated how cuisines are structured in terms of ingredient usage. Representing cuisines as networks of ingredient types revealed not only differences in popularity and recipe sizes but also distinctive "structural fingerprints" that capture culinary identity. These networks expose characteristic associations that differentiate cuisines, and their distilled backbones in the form of maximum spanning trees summarize the essential building blocks. Both full and simplified representations enable robust classification of cuisines and uncover geo-cultural clusters, showing that cuisines can be meaningfully compared and grouped through their network organization.

At the human level, we turned to chefs and their institutions, treating them as nodes in a network of influence defined by mentorship and inspiration. The resulting network reconstructs the genealogical fabric of modern gastronomy. Analysis reveals strong asymmetries, where a minority of chefs exert disproportionate influence across generations. Temporal dynamics reinforce these asymmetries, with ties flowing predominantly from older to younger

chefs across a generational gap of about two decades. Centrality measures highlight multiple pathways to prominence—mentorship, brokerage, cumulative prestige, or structural indispensability—while nationality emerges as a key axis of organization. The network is assortative by national origin, with clear asymmetries: France dominates as an exporter of influence, Italy as a receiver, and Spain as a global avant-garde hub. Community detection reveals cohesive clusters aligned with national and stylistic schools, from classical lineages to the experimental Catalan scene. Finally, by extracting a maximum spanning tree weighted by information centrality, we distilled the dense fabric into a genealogical backbone that foregrounds pivotal figures and points of contact, offering a global picture of the evolution of haute cuisine.

Taken together, these three investigations frame gastronomy into a conceptual division of mechanisms operating at different scales. Chemistry provides a pool of feasible combinations, but by itself it does not explain culinary practice. Cultural traditions filter and stabilize subsets of these possibilities, producing recognizable fingerprints of cuisines. Human networks—through mentorship, influence, and institutions—transmit, recombine, and transform these cultural repertoires across time and space. In this way, the multi-scale framework proposed here presents culinary creativity as the outcome of chemical possibilities, cultural selection, and human transmission, each contributing in complementary ways.

Beyond substantive insights, this thesis contributes methods of general value. At the molecular level, it introduced a controlled predictive evaluation for ingredient recommendation that contrasts chemical and empirical similarity. At the cultural level, it presented a scalable framework for constructing and analyzing ingredient-type networks, including the use of maximum spanning trees as compact backbones. At the human level, it combined centrality analysis with efficiency-based spanning-tree extraction to map genealogical backbones of influence. Together, these approaches expand the methodological repertoire of computational gastronomy and provide benchmarks for future work.

6.2 Limitations and Future Work

Several limitations should be acknowledged. First, while large-scale datasets have made computational gastronomy feasible, their structure poses challenges. Public recipe repositories are designed for human use rather than systematic analysis: formats vary widely, ingredient lists are inconsistently written, and metadata on geography, time period, or preparation style is often missing. The standardization performed on the two datasets used in this thesis alleviates some of these problems, but structural biases embedded in the data cannot be entirely removed. These challenges create barriers to integrating heterogeneous sources and ultimately constrain the scope of what can be captured computationally.

Second, the mapping of raw ingredients to standardized categories, necessary for large-scale analysis, introduces a degree of arbitrariness. Many ingredients resist neat categorization, and subtle differences in classification can affect observed patterns. This process inevitably introduces noise, though efforts were made to minimize it through systematic preprocessing.

Third, chemical compound datasets remain incomplete and provide only a static snapshot of ingredients. Crucially, they do not account for the transformations that occur during cook-

ing—such as Maillard reactions, caramelization, or enzymatic changes—that fundamentally alter flavor and taste perception. This means that the molecular dimension is approximated rather than fully captured.

Finally, modeling choices impose their own constraints. Analyses based on ingredient co-occurrence or network measures necessarily reflect the scope and biases of the available data, both temporally and geographically. Recipe datasets used in this thesis rely on public repositories, which are uneven in coverage: some world regions and cuisines (such as Western European or East Asian) are richly represented, while others (such as African or indigenous culinary traditions) appear only sparsely. Large datasets that could partially mitigate these gaps do exist, but they are typically unstructured and lack reliable standardization. As a result, the structural "fingerprints" extracted from ingredient networks are weighted toward those culinary traditions that are more thoroughly documented online. Temporal scope is another limitation: the recipe datasets mainly capture contemporary or recent culinary practice, leaving historical cuisines underrepresented, while the chef influence network focuses primarily on the genealogies of modern haute cuisine from the mid-20th century onward. These temporal and geographical imbalances mean that the detected patterns should be interpreted as characterizing the documented portion of world gastronomy rather than its full diversity.

The limitations identified above point toward several promising directions for future research, spanning both data and modeling.

On the data side, a first challenge lies in improving the scope and structure of available resources. Public recipe repositories have enabled large-scale analysis, but they remain uneven in coverage and inconsistent in format. While vast unstructured collections exist that could compensate for regional or cultural imbalances, they require systematic cleaning, parsing, and enrichment with reliable metadata such as geography, chronology, and preparation style. Advances in natural language processing and ontology-driven annotation may help transform these raw sources into standardized datasets suitable for comparative study.

At the level of ingredient representation, classification remains a delicate task. Mapping raw ingredients into standardized categories inevitably introduces arbitrariness, as many foods resist neat boundaries. Future work could refine this process by combining culinary taxonomies, chemical composition, and machine learning to generate multidimensional representations that capture flavor, texture, and functional role within recipes. Such richer ontologies would reduce noise and allow for analyses that are more faithful to culinary practice.

A parallel need arises on the molecular dimension. Existing compound datasets are incomplete and static, neglecting transformations that occur during cooking—such as Maillard reactions, caramelization, or enzymatic processes—that fundamentally alter taste. Integrating chemical reaction data and sensory studies would bring the molecular layer closer to the reality of flavor perception, enabling more nuanced models of ingredient compatibility.

Alongside these data improvements, new modeling strategies are essential. Current approaches largely treat cuisines and influence networks as static, but culinary traditions are dynamic systems shaped by migration, technological change, and shifting ingredient availability. Developing temporal network models would allow us to track the evolution of cuisines, the rise and fall of ingredients, and the diffusion of stylistic schools across generations.

Moreover, most network analyses focus on pairwise relations, capturing co-occurrence

between ingredient pairs. This abstraction is useful but incomplete, as recipes are inherently higher-order objects where multiple ingredients interact simultaneously. Extending analyses to hypergraphs and other higher-order structures would better capture these multi-ingredient complexes, revealing deeper organizational principles behind recipe construction.

Finally, future work should strengthen the connection between computational models and culinary practice. Sensory validation, chef collaborations, and applications to recipe recommendation or adaptation would provide benchmarks that move beyond abstract measures of accuracy. Such collaborations would help assess whether computational gastronomy can capture not only patterns in data but also the lived creativity of cooking and eating.

Appendices

Appendix A

Ingredient Type Network Figures

In the main text, we presented illustrative examples of popularity profiles, of backbones of ingredient type networks and their maximum spanning trees (MSTs) for a subset of cuisines. Here, we provide the complete set of visualizations for all 23 world cuisines considered in this study. These supplementary figures allow for direct comparison across cuisines and serve as reference material for readers interested in the detailed structural organization of ingredient type combinations.

A.1 Ingredient popularity and z-score profiles

Figure A.1 reports the ingredient type frequency distributions for the 18 cuisines not shown in the main text. Together with the examples already discussed in Section 4.2.3, these profiles highlight both the universal hierarchy of types (e.g., *Vegetables*, *Spices*, *Additives*) and the culturally specific prominence of others (e.g., *Dairy* in Northern Europe).

Figure A.2 presents the standardized z-scores corresponding to the frequency profiles in Figure A.1. These z-scores indicate whether the prevalence of each type in a given cuisine is above or below the global average, thus revealing its relative cultural significance.

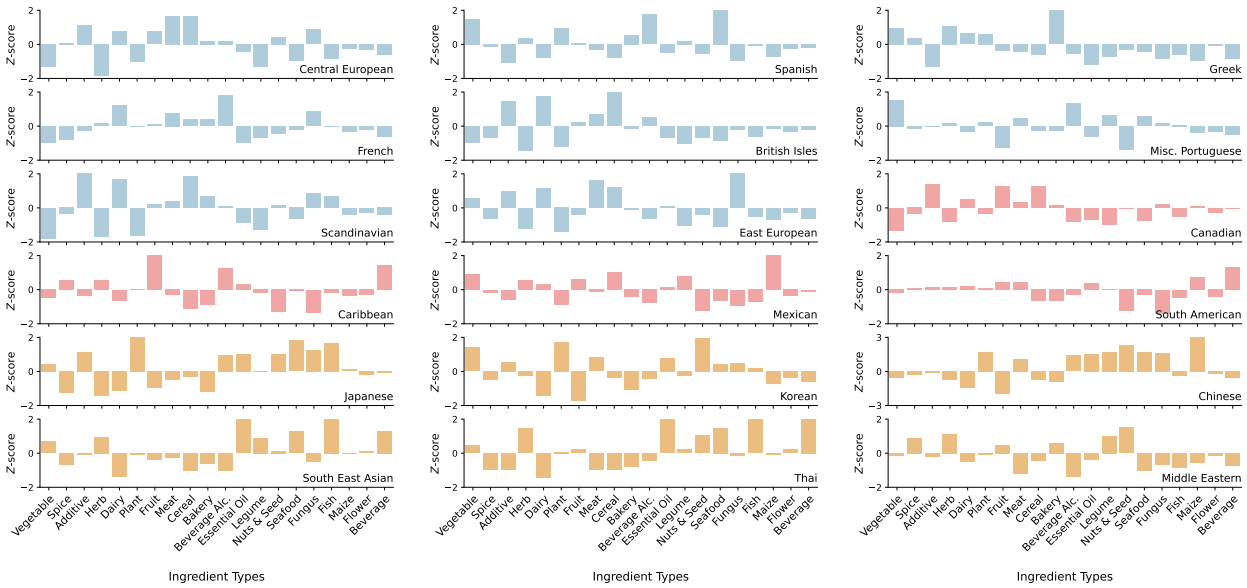


Figure A.1: **Ingredient type popularity profiles for 18 additional cuisines.** Relative frequencies of ingredient types, complementing the examples shown in Section 4.2.3.

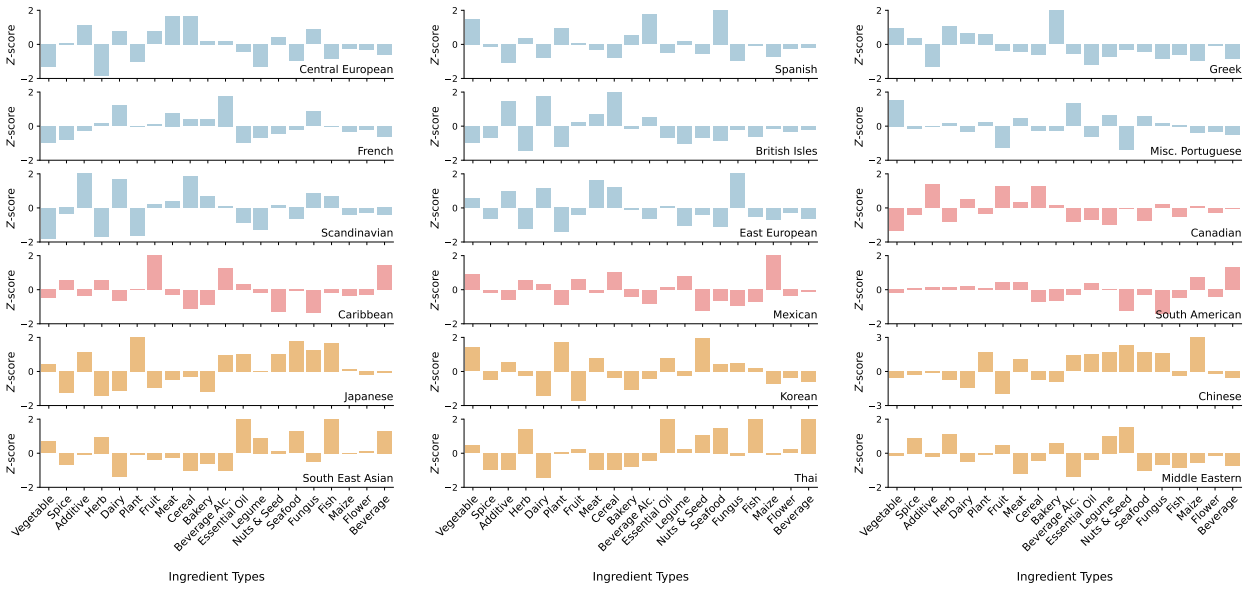


Figure A.2: z**-score profiles of ingredient type popularity.** Standardized z-scores corresponding to the frequency distributions reported in Figure A.1, highlighting over- and under-represented ingredient types relative to the global baseline.

A.2 Backbone networks

Figure A.3 reports the filtered ingredient type networks for all 23 cuisines. As discussed in Section 4.4.1, these networks retain only the strongest associations between ingredient types, highlighting the backbones that structure recipes in each culinary tradition.

A.3 Maximum spanning trees (MSTs)

In Section 4.5, we introduced maximum spanning trees as minimal structural backbones that serve as fingerprints of cuisines. Figure A.4 shows the MSTs for all 23 cuisines, complementing Figure 4.10 in the main text. These trees reveal distinctive centralities and diameters across culinary traditions, from the spice-centered star-like structure of Indian cuisine to the more distributed layouts observed in American and Central European cuisines.



Figure A.3: **Networks of ingredient combinations.** Ingredient combination networks for all 23 analyzed cuisines, filtered using the disparity filter method introduced by Serrano et al. [2009]. Refer to Fig. 4.8 for details on the realization.

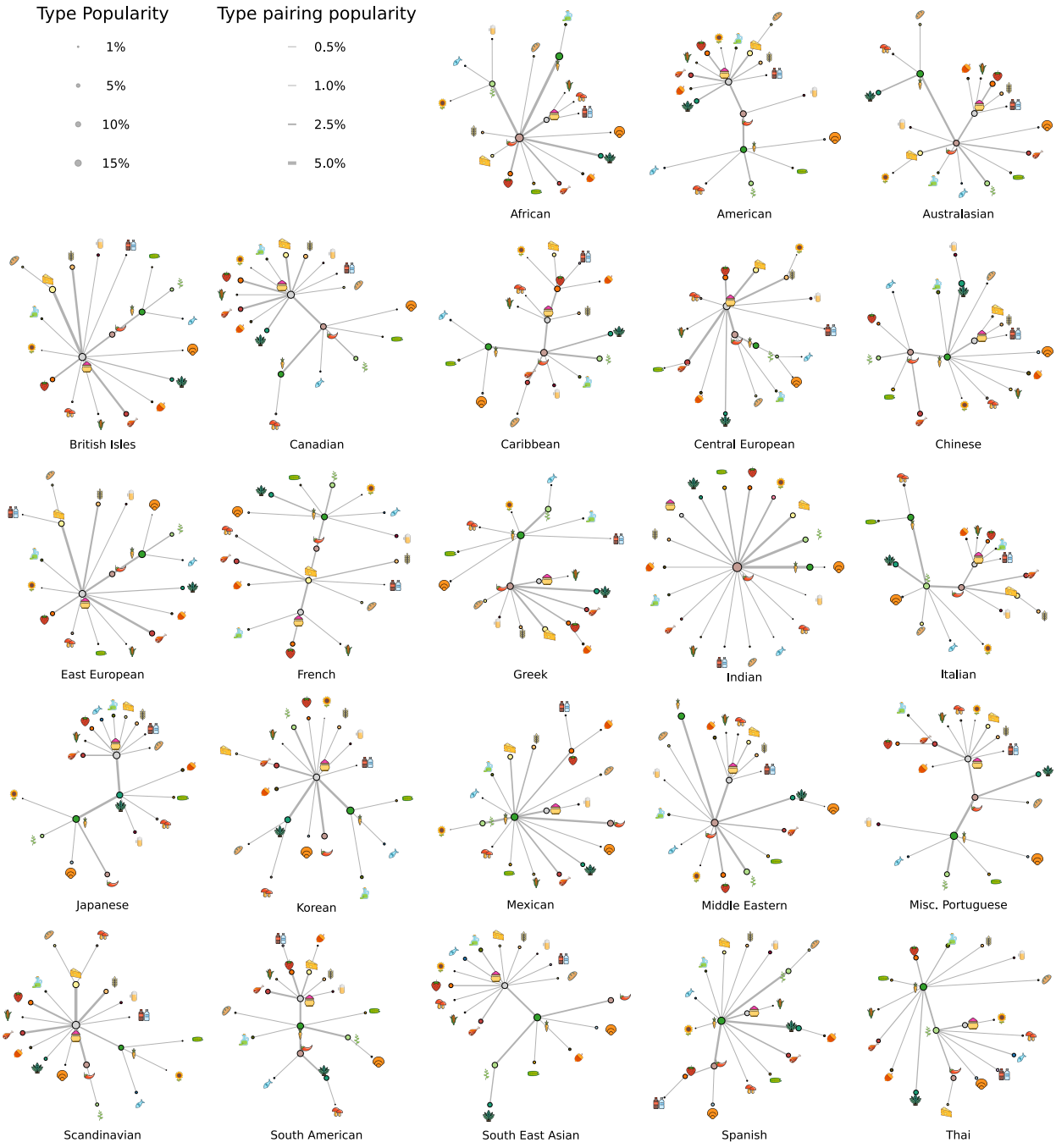


Figure A.4: **Maximum Spanning Trees of Ingredient Combination Networks**. Maximum Spanning Trees of the Ingredient Type Graphs for all 23 cuisines. Details on the realization are reported in Fig. 4.10 of the main text.

Appendix B

Extended Centrality Rankings

In the main text we focused on the most prominent chefs as ranked by different centrality measures, highlighting contrasts between receivers and sources of influence, brokers, and structural linchpins. Here we provide the extended top-20 rankings for each measure—betweenness, PageRank, and information centrality—together with normalized centrality values (Table B.1).

These rankings illustrate complementary aspects of centrality in the chef influence network. Betweenness emphasizes historically foundational figures such as Alain Ducasse, Alain Passard, and Fernand Point, who connected multiple lineages. PageRank highlights younger chefs embedded in prestigious lineages (e.g., Jordan Bailey, Isaac McHale, Chiara Pavan), reflecting how the measure weights endorsement by already central actors. Information Centrality underscores systemic importance, with Ferran Adrià, René Redzepi, and Massimo Bottura standing out as structural linchpins whose removal would most diminish global cohesion.

Table B.1: **Top 20 chefs by betweenness, PageRank, and Information Centrality.** Values are normalized centrality scores. Nationality codes: FR = France, IT = Italy, ES = Spain, UK = United Kingdom, US = United States, DK = Denmark, CH = Switzerland, AU = Australia, BR = Brazil, AR = Argentina, JP = Japan, PT = Portugal, IN = India.

Betweenness			**PageRank**			**Information Centrality**		
Rank	Chef	Value	Rank	Chef	Value	Rank	Chef	Value
1	Alain Ducasse (FR, 1956)	0.0021	1	Jordan Bailey (UK, 1990)	0.0144	1	Ferran Adrià (ES, 1962)	0.0668
2	Alain Passard (FR, 1956)	0.0014	2	Isaac McHale (UK, 1980)	0.0123	2	René Redzepi (DK, 1977)	0.0593
3	Fernand Point (FR, 1897)	0.0011	3	Brett Graham (AU, 1979)	0.0113	3	Massimo Bottura (IT, 1962)	0.0496
4	Massimo Bottura (IT, 1962)	0.0010	4	Antonio Biafora (IT, 1984)	0.0111	4	Brett Graham (AU, 1979)	0.0341
5	Marco Pierre White (UK, 1961)	0.0009	5	Chiara Pavan (IT, 1985)	0.0109	5	Alain Passard (FR, 1956)	0.0272
6	Michel Troisgros (FR, 1958)	0.0009	6	Jake Kellie (AU, 1991)	0.0105	6	Alain Ducasse (FR, 1956)	0.0255
7	Alain Chapel (FR, 1937)	0.0009	7	Jeremiah Stone (US, 1984)	0.0099	7	Frédy Girardet (CH, 1936)	0.0255
8	Ferran Adrià (ES, 1962)	0.0009	8	Gianluca Gorini (IT, 1983)	0.0092	8	Michel Troisgros (FR, 1958)	0.0246
9	Pierre Troisgros (FR, 1928)	0.0009	9	Francesco Brutto (IT, 1980)	0.0088	9	Andoni Luis Aduriz (ES, 1971)	0.0239
10	Paul Bocuse (FR, 1926)	0.0008	10	Manoella Buffara (BR, 1983)	0.0086	10	Thomas Keller (US, 1955)	0.0232
11	Bernard Loiseau (FR, 1951)	0.0008	11	Esben Holmboe Bang (DK, 1982)	0.0081	11	Albert Adrià (ES, 1969)	0.0222
12	Heston Blumenthal (UK, 1966)	0.0007	12	Antonio Ziantoni (IT, 1986)	0.0080	12	Tom Aikens (UK, 1970)	0.0184
13	René Redzepi (DK, 1977)	0.0006	13	Michele Lazzarini (IT, 1991)	0.0080	13	Joan Roca (ES, 1964)	0.0183
14	Mauro Colagreco (AR, 1978)	0.0006	14	Benedetto Rullo (IT, 1989)	0.0080	14	Bernard Loiseau (FR, 1951)	0.0181
15	David Bouley (US, 1953)	0.0005	15	Stefano Terigi (IT, 1987)	0.0080	15	Heston Blumenthal (UK, 1966)	0.0176
16	Andoni L. Aduriz (ES, 1971)	0.0005	16	Lorenzo Stefanini (IT, 1987)	0.0080	16	Gualtiero Marchesi (IT, 1930)	0.0165
17	Joan Roca (ES, 1964)	0.0005	17	Simone Tondo (IT, 1988)	0.0079	17	Grégory Marchand (FR, 1978)	0.0159
18	Tom Aikens (UK, 1970)	0.0004	18	Frank Rizzuti (IT, 1967)	0.0075	18	Michele Lazzarini (IT, 1991)	0.0159
19	Enrico Crippa (IT, 1971)	0.0004	19	Sven Wassmer (CH, 1986)	0.0073	19	Nuno Mendes (PT, 1973)	0.0154
20	Albert Adrià (ES, 1969)	0.0004	20	Garima Arora (IN, 1986)	0.0073	20	Seiji Yamamoto (JP, 1970)	0.0153

Appendix C

Community–Nationality Alignment

To evaluate whether the observed alignment between communities and nationality exceeds what can be explained by degree heterogeneity and global homophily, we construct a rewired-graph null model. The idea is to generate alternative networks that preserve key low-order properties of the data—namely the degree sequence, the global nationality mixing matrix, and overall connectivity—while randomizing higher-order wiring. By repeatedly refitting Louvain communities on these randomized graphs and comparing resulting statistics to the empirical ones, we obtain Monte Carlo benchmarks against which to assess the significance of our findings.

C.1 Null Model

Let $G_0 = (V, E_0)$ be the undirected projection of the largest connected component of the work layer, with $|V| = n$ nodes and $|E_0| = m$ edges. Each chef $v \in V$ carries a categorical attribute $A(v) \in \mathcal{A}$ (here, nationality), and has degree

$$d(v) = \deg_{G_0}(v).$$

For each unordered label pair (a, b) with $a \leq b$, we record the observed *mixing counts*:

$$M_{ab} \;=\; \big|\{\{u, w\} \in E_0 : \{A(u), A(w)\} = \{a, b\}\}\big|.$$

We then define the constrained ensemble

$$\Omega_{\text{conn}}(d, M) \;=\; \Big\{ G = (V, E) : \begin{array}{ll} \deg_G(v) = d(v) & \forall v \in V, \\ \#\{\{u, w\} \in E : \{A(u), A(w)\} = \{a, b\}\} = M_{ab} & \forall a \leq b, \\ G \text{ connected} & \end{array} \Big\}.$$

This set consists of all connected graphs on V that preserve both the degree sequence and the full nationality mixing matrix of G_0.

Null hypothesis. Under the null hypothesis, community structure is assumed to emerge solely from degree heterogeneity and the observed global distribution of national ties. Formally, we define the null distribution as follows:

1. Sample a graph $G^{(b)}$ *uniformly* from $\Omega_{\text{conn}}(d, M)$.
2. Sample a Louvain seed s_b from a distribution S (e.g., uniform over integers).
3. Compute the partition $\pi^{(b)} = \text{Louvain}_\gamma(G^{(b)}; s_b)$ at resolution γ.
4. Evaluate test statistics $T^{(b)} = T(G^{(b)}, \pi^{(b)}, A)$.

The observed network G_0 with its partition $\pi^\star$ produces

$$T_{\text{obs}} = T(G_0, \pi^\star, A),$$

which is evaluated against the null distribution $T^{(b)}$ via Monte Carlo simulation.

C.2 Algorithmic implementation

Exact uniform sampling from $\Omega_{\text{conn}}(d, M)$ is computationally infeasible. Instead, we approximate it using a constrained *double-edge-swap* Markov chain. This procedure preserves the degree sequence and the full mixing matrix M exactly, while enforcing connectivity at every step. Each rewired graph is then subjected to Louvain community detection, and the resulting partitions are used to compute the evaluation statistics.

MCMC kernel: mixing-preserving double-edge swap

Algorithm 1: Mixing-preserving edge swap with connectivity enforcement

Partition edges into buckets $\mathcal{B}_{ab} = \{\{x, y\} \in E : \{A(x), A(y)\} = \{a, b\}\}$;
Sample a bucket (a, b) with $|\mathcal{B}_{ab}| \geq 2$;
Pick two disjoint edges $e_1 = \{x, y\}, e_2 = \{u, v\} \in \mathcal{B}_{ab}$;
Form candidate rewirings: *straight* $\{x, u\}, \{y, v\}$ or *cross* $\{x, v\}, \{y, u\}$;
Select an orientation that preserves the multiset of label pairs of $\{e_1, e_2\}$;
Reject if a self-loop or multi-edge would be created;
Tentatively apply the swap; accept iff the resulting graph remains connected;
otherwise revert;

This kernel guarantees that:

- the degree sequence $d(v)$ is preserved,
- the full nationality mixing matrix M_{ab} (and thus global assortativity) is fixed,
- the graph remains connected throughout the chain.

Burn-in and thinning

Let $m = |E|$ be the number of edges. To reduce dependence on the initial graph and mitigate autocorrelation, we proceed as follows:

- **Burn-in:** perform αm successful swaps before recording any samples (with $\alpha = 50$ in practice).
- **Thinning:** between two recorded samples, perform βm additional successful swaps (with $\beta = 10$).
- **Sampling:** collect B rewired graphs $G^{(1)}, \ldots, G^{(B)}$ after thinning.

Community refit and statistics

For each sampled graph $G^{(b)}$:

- Run the Louvain algorithm at resolution γ, initialized with a random seed s_b, to obtain a partition $\pi^{(b)}$.
- Compute the evaluation statistics, including:
  1. Alignment with nationality: $\mathrm{NMI}(\pi^{(b)}, A)$ and $\mathrm{AMI}(\pi^{(b)}, A)$.
  2. Within-community assortativity:
  $$\bar{r}^{(b)} \;=\; \frac{\sum_q |E(G^{(b)}[q])|\, r_q}{\sum_q |E(G^{(b)}[q])|},$$
  where r_q is the assortativity of community q.
  3. Partition shape measures: number of communities, coverage, average conductance, size inequality (Gini, entropy), purity, and embeddedness.

Inference

From the B rewired samples we obtain a distribution of statistics $\{T^{(b)}\}$. For each statistic, we estimate the null mean and variance as

$$\hat{\mu} = \tfrac{1}{B}\sum_{b=1}^{B} T^{(b)}, \qquad \hat{\sigma}^2 = \tfrac{1}{B-1}\sum_{b=1}^{B}\big(T^{(b)} - \hat{\mu}\big)^2,$$

and compute a standardized z-score

$$z = \frac{T_{\mathrm{obs}} - \hat{\mu}}{\hat{\sigma}},$$

where T_{obs} is the observed statistic.

Monte Carlo p-values are derived in two forms: *one-sided*, testing whether the observed statistic exceeds the null distribution,

$$\hat{p}_{\rightarrow} = \frac{1 + \sum_{b=1}^{B} \mathbf{1}\{T^{(b)} \geq T_{\mathrm{obs}}\}}{B+1},$$

and *two-sided*, testing whether it deviates in either direction,

$$\hat{p}_{\leftrightarrow} = \frac{1 + \sum_{b=1}^{B} \mathbf{1}\{|T^{(b)} - \hat{\mu}| \geq |T_{\mathrm{obs}} - \hat{\mu}|\}}{B+1}.$$

Remarks. By fixing the full nationality mixing matrix M, the null preserves global assortativity exactly. Any excess alignment or cohesion in the observed network therefore reflects higher-order topology rather than overall homophily. Connectivity is enforced to ensure comparability with the observed largest connected component. Burn-in and thinning reduce autocorrelation in the Markov chain, while varying Louvain seeds incorporates the algorithm's stochasticity into the null distribution.

C.3 Test statistics

For each observed partition $\pi^\star$ and each null partition $\pi^{(b)}$, we compute a set of statistics that capture two dimensions: (i) how strongly communities align with nationality labels, and (ii) how cohesive and balanced the communities are as subgraphs.

(1) Alignment with nationality. Let $A : V \to \mathcal{A}$ be the nationality label function and $\pi : V \to \{1, \ldots, K\}$ a community assignment.

- **Normalized mutual information (NMI).**

$$\mathrm{NMI}(\pi, A) \;=\; \frac{2\, I(\pi; A)}{H(\pi) + H(A)},$$

  where $I(\pi; A)$ is the mutual information between π and A, and H denotes Shannon entropy.

- **Adjusted mutual information (AMI):** AMI corrects NMI for agreement that would be expected by chance [Vinh et al., 2010]. Formally,

$$\mathrm{AMI}(\pi, A) \;=\; \frac{I(\pi; A) - \mathbb{E}[I(\pi; A)]}{\frac{1}{2}\big(H(\pi) + H(A)\big) - \mathbb{E}[I(\pi; A)]},$$

  where $\mathbb{E}[I(\pi; A)]$ is the expected mutual information under a random labeling with fixed community and nationality marginals.

(2) Within-community nationality assortativity. For a community q with induced subgraph $G[q]$, let r_q be the categorical assortativity coefficient [Newman, 2003] of edges inside $G[q]$ with respect to nationality. We aggregate across communities by an edge-weighted average:

$$\bar{r}(\pi) \;=\; \frac{\sum_q |E(G[q])| \cdot r_q}{\sum_q |E(G[q])|}.$$

(3) Partition size and shape.

- **Number of communities:**

$$C(\pi) = \big|\{q : \exists v,\ \pi(v) = q\}\big|.$$

- **Coverage:** fraction of edges lying within communities,

$$\mathrm{cov}(\pi) = \frac{\sum_q |E(G[q])|}{|E|}.$$

- **Conductance:** for each community q,

$$\phi(q) = \frac{|\{(u,v) \in E : u \in q, v \notin q\}|}{\min(\mathrm{vol}(q), \mathrm{vol}(V \setminus q))},$$

  where $\mathrm{vol}(S) = \sum_{u \in S} d(u)$. We report the average $\bar{\phi}(\pi)$, weighted by edge count.

- **Size inequality:** Let $\{n_q\}$ be community sizes. We define Gini coefficient

$$G = \frac{\sum_i \sum_j |n_i - n_j|}{2K \sum_i n_i},$$

  and entropy

$$H_{\text{size}} = -\sum_{q=1}^{K} \frac{n_q}{n} \log \frac{n_q}{n}.$$

(4) Purity and embeddedness.

- **Purity:** for each community q, let p_q be the majority nationality share. Overall purity is

$$\mathrm{purity}(\pi) = \sum_{q=1}^{K} \frac{|E(G[q])|}{|E|} p_q.$$

- **Embeddedness:** for each node v, we define node embeddedness as

$$\epsilon(v) = \frac{k_{\text{in}}(v)}{k(v)},$$

  where $k_{\text{in}}(v)$ counts edges internal to $\pi(v)$. The overall embeddedness is the edge-weighted average

$$\bar{\epsilon}(\pi) = \frac{1}{|E|} \sum_{v \in V} k(v)\, \epsilon(v).$$

Together, these statistics disentangle two dimensions: whether communities are *nationality-driven*, and whether they are *structurally cohesive* subgraphs. Under the null experiments, each statistic can be benchmarked against its empirical distribution $\{T^{(b)}\}$.

Summary

Table C.1 summarizes the invariants, randomized aspects, and evaluation statistics for this null model.

Table C.1: Summary of the rewired-graph refit null model.

Aspect	**Preserved (invariants)**	**Randomized / evaluated**
Graph structure		
	• Degree sequence $d(v)$ • Connectivity of G • Global nationality mixing matrix M_{ab}	• Higher-order wiring of edges • Local clustering/topology
Community detection		
	• Louvain resolution parameter γ	• Louvain seed (stochasticity) • Resulting partition $\pi^{(b)}$
Evaluation statistics	–	
		• Alignment: NMI, AMI • Within-community assortativity $\bar{r}$ • Partition shape: # communities, coverage, conductance, size Gini/entropy • Purity, embeddedness • Stability across seeds

C.4 Results

The rewired-graph null confirms that the strong nationality signal observed in the raw data can be fully explained by global homophily. Once degree sequence and the full nationality mixing matrix are preserved, the observed values of NMI, AMI, and within-community assortativity fall well within null expectations (Table C.2). By contrast, cohesion-related measures deviate modestly: embeddedness is significantly higher than expected, and coverage and conductance trend in the cohesive direction. This indicates that communities are not simply artifacts of nationality composition, but also capture genuine structural clustering.

Table C.2: Rewired-graph null results on the work-layer LCC ($|V| = 213$, $|E| = 344$). Null preserves degree sequence, connectivity, and the full nationality mixing matrix; Louvain refit per sample. Right-tailed tests used for NMI/AMI and within-community assortativity $\bar{r}$; two-sided tests for the other measures.

Statistic	**Observed**	$\boldsymbol{\mu}_{\text{null}}$	$\boldsymbol{\sigma}_{\text{null}}$	$\boldsymbol{z}$	$\boldsymbol{p}$
Within-comm. assortativity $\bar{r}$	0.179	0.229	0.039	−1.290	0.898
NMI(π, Nat)	0.372	0.384	0.023	−0.510	0.693
AMI(π, Nat)	0.215	0.239	0.027	−0.910	0.808
# communities	13.000	11.660	1.020	1.320	0.333
Coverage	0.735	0.708	0.014	1.920	0.058
Average conductance	0.262	0.287	0.014	−1.840	0.056
Size Gini	0.251	0.185	0.040	1.650	0.112
Size entropy	2.459	2.393	0.089	0.740	0.441
Purity (edge-weighted)	0.472	0.485	0.029	−0.450	0.649
Embeddedness (edge-weighted)	0.825	0.801	0.011	2.260	0.022

Notes: (i) Right-tailed p for $\bar{r}$, NMI, AMI (testing "greater than null"); two-sided p for all others (agnostic direction). (ii) $z = (T_{\text{obs}} - \mu_{\text{null}})/\sigma_{\text{null}}$. (iii) Embeddedness shows a significant excess over the null; coverage and conductance are near-significant in the cohesive direction.

Bibliography

Yaser S. Abu-Mostafa, Malik Magdon-Ismail, and Hsuan-Tien Lin. *Learning From Data: A Short Course*. AMLBook, 2012. ISBN 978-1-60049-006-4.

Lada A. Adamic and Eytan Adar. Friends and Neighbors on the Web. *Social Networks*, 25 (3):211–230, 2003. doi: 10.1016/S0378-8733(03)00009-1.

Gediminas Adomavicius and Alexander Tuzhilin. Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6):734–749, 2005. doi: 10.1109/TKDE.2005.99.

Charu C. Aggarwal. *Recommender Systems: The Textbook*. Springer, Cham, 2016. ISBN 978-3-319-29657-9. doi: 10.1007/978-3-319-29659-3.

Yong-Yeol Ahn, Sebastian E. Ahnert, James P. Bagrow, and Albert-László Barabási. Flavor Network and the Principles of Food Pairing. *Scientific Reports*, 1(1):196, 2011. doi: 10.1038/srep00196.

Sebastian E. Ahnert. Network Analysis and Data Mining in Food Science: The Emergence of Computational Gastronomy. *Flavour*, 2(4):1–7, 2013. doi: 10.1186/2044-7248-2-4.

Laura M. Aiello, Rossano Schifanella, Daniele Quercia, and Lucia Del Prete. Large-scale and High-resolution Analysis of Food Purchases and Health Outcomes. *EPJ Data Science*, 8 (1):14, 2019. doi: 10.1140/epjds/s13688-019-0191-y.

Ibn Sayyár al Warrāq. *Annals of the Caliphs' Kitchens: Ibn Sayyar al-Warrāq's Tenth-Century Baghdadi Cookbook*. Brill, 2007. ISBN 978-90-04-15867-2.

Giambattista Albora, Lavinia Rossi Mori, and Andrea Zaccaria. Sapling Similarity: A performing and interpretable memory-based tool for recommendation. *Knowledge-Based Systems*, 275:110659, 2023. doi: 10.1016/j.knosys.2023.110659.

José Albors-Garrigós, Maissa Haddaji, and Purificación García-Segovia. Gender Differences in the Evolution of Haute Cuisine Chef's Career. *Journal of Culinary Science & Technology*, 18(3):193–218, 2020. doi: 10.1080/15428052.2019.1640156.

AllRecipes. Allrecipes. `https://www.allrecipes.com/`.

Apicius. *Apicius: The Roman Cookery Book*. Dover Publications, 2006. ISBN 0245522689.

Ganesh Bagler. Culinarydb: A repository of structured recipe data from world cuisines, 2017. URL `https://cosylab.iiitd.edu.in/culinarydb/`. Accessed: 2025-08-19; dataset of recipes and ingredient categories from 22+ geo-cultural world regions.

Albert-László Barabási. *Network Science*. Cambridge University Press, Cambridge, UK, 2016. ISBN 978-1-107-07626-6.

Albert-László Barabási. The Unmapped Chemical Complexity of Our Diet. *Nature Food*, 1 (1):33–37, 2020. doi: 10.1038/s43016-019-0005-1.

Albert-László Barabási, Giulia Menichetti, and Joseph Loscalzo. The Unmapped Chemical Complexity of Our Diet. *Nature Food*, 1(1):33–37, 2020. doi: 10.1038/s43016-019-0005-1.

Emiliano López Barrera and Thomas Hertel. Solutions to the Double Burden of Malnutrition Also Generate Health and Environmental Benefits. *Nature Food*, 4:616–624, 2023. doi: 10.1038/s43016-023-00798-7.

Jennifer Billing and Paul W. Sherman. Antimicrobial Functions of Spices: Why Some Like It Hot. *The Quarterly Review of Biology*, 73(1):3–49, 1998. doi: 10.1086/420058.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, New York, NY, 2006. ISBN 978-0-387-31073-2.

Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast Unfolding of Communities in Large Networks. *Journal of Statistical Mechanics: Theory and Experiment*, (10):P10008, 2008. doi: 10.1088/1742-5468/2008/10/P10008.

Stephen P. Borgatti. Centrality and Network Flow. *Social Networks*, 27(1):55–71, 2005. doi: 10.1016/j.socnet.2004.11.008.

Christopher R. Borkenhagen. *Scientific Knowledge and Social Structure in the Culinary Arts and Mathematical Finance*. PhD thesis, University of Chicago, 2018. URL `https://knowledge.uchicago.edu/record/1369`.

Jean Bottéro. *The Oldest Cuisine in the World: Cooking in Mesopotamia*. University of Chicago Press, Chicago, 1995. ISBN 978-0-226-06735-3.

Alessandro Bressan, Abel Duarte Alonso, and Oanh Thi Kim Vu. The Importance of Knowledge Management in Gastronomy: A Chef's Future Career, the Next Generations of Chefs and Trends. *International Journal of Contemporary Hospitality Management*, 35(2):655–673, 2023. doi: 10.1108/IJCHM-03-2022-0391.

Sergey Brin and Lawrence Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998. doi: 10.1016/S0169-7552(98)00110-X.

Alberto Capatti and Massimo Montanari. *Italian Cuisine: A Cultural History*. Columbia University Press, 2003. ISBN 978-0-231-12232-0.

Ignasi Capdevila, Patrick Cohendet, and Laurent Simon. From a Local Community to a Global Influence: How elBulli Restaurant Created a New Epistemic Movement in the World of Haute Cuisine. *Industry and Innovation*, 25(5):526–549, 2018. doi: 10.1080/13662716.2017.1327844.

Claudio Caprioli, Saumitra Kulkarni, Federico Battiston, Iacopo Iacopini, Andrea Santoro, and Vito Latora. The Networks of Ingredient Combination in Cuisines Around the World. *arXiv preprint*, arXiv:2408.15162, 2024. doi: 10.48550/arXiv.2408.15162. URL `https://arxiv.org/abs/2408.15162`.

Davide Cassi. Science and Cooking: The Era of Molecular Cuisine. *EMBO Reports*, 12(3):191–196, 2011. doi: 10.1038/embor.2011.18.

Nirupa Chaudhari and Stephen D. Roper. The taste of umami. *The Journal of Physiology*, 587(Pt 19):4359–4373, 2009. doi: 10.1113/jphysiol.2009.179705.

William S. Cleveland. Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association*, 74(368):829–836, 1979. doi: 10.1080/01621459.1979.10481038.

Corinna Cortes and Vladimir Vapnik. Support-Vector Networks. *Machine Learning*, 20(3):273–297, 1995. doi: 10.1007/BF00994018.

CoSyLab. Cosylab — complex systems laboratory. URL `https://cosylab.iiitd.edu.in/`.

Alfred W. Crosby. *The Columbian Exchange: Biological and Cultural Consequences of 1492*. Greenwood Publishing Group, Westport, CT, 1972. ISBN 0837158214.

Leon Danon, Albert Díaz-Guilera, Jordi Duch, and Alex Arenas. Comparing Community Structure Identification. *Journal of Statistical Mechanics: Theory and Experiment*, (09):P09008, 2005. doi: 10.1088/1742-5468/2005/09/P09008.

Christian Desrosiers and George Karypis. A Comprehensive Survey of Neighborhood-Based Recommendation Methods. In Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor, editors, *Recommender Systems Handbook*, pages 107–144. Springer, Boston, MA, 2011. doi: 10.1007/978-0-387-85820-3_4.

Jared Diamond. *Guns, Germs, and Steel: The Fates of Human Societies*. W. W. Norton & Company, New York, 1997. ISBN 0-393-03891-2.

Reinhard Diestel. *Graph Theory*, volume 173 of *Graduate Texts in Mathematics*. Springer, Berlin, Heidelberg, 5 edition, 2017. ISBN 978-3-662-53621-6. Electronic edition available at `https://diestel-graph-theory.com/`.

Tamer Elbasha and Yehuda Baruch. Becoming Top Global Chef: What Does It Take to Become a Highly Successful Entrepreneurial Careerist? *International Journal of Contemporary Hospitality Management*, 34(7):2559–2578, 2022. doi: 10.1108/IJCHM-07-2021-0940.

Encyclopaedia Britannica Editors. Ferran Adrià. Encyclopaedia Britannica Online, 2024a. URL https://www.britannica.com/biography/Ferran-Adria. Accessed: 2025-09-19.

Encyclopaedia Britannica Editors. Alain Ducasse. Encyclopaedia Britannica Online, 2024b. URL https://www.britannica.com/biography/Alain-Ducasse. Accessed: 2025-09-26.

Epicurious. Epicurious. https://www.epicurious.com/.

Nicholas Eriksson, Sarah Wu, Chuong B. Do, Amy K. Kiefer, Joyce Y. Tung, Joanna L. Mountain, Uta Francke, Sheri Sloan, and David A. Hinds. A genetic variant near olfactory receptor genes influences cilantro preference. *Flavour*, 1(1):22, 2012. doi: 10.1186/2044-7248-1-22.

Mary K. Feeney and Barry Bozeman. Mentoring and Network Ties. *Human Relations*, 61 (2):233–254, 2008. doi: 10.1177/0018726708098081.

Giovanni Fenaroli. *Fenaroli's Handbook of Flavor Ingredients*. Chemical Rubber Company, 1971. ISBN 0878195335.

Food Network. Food network. https://www.foodnetwork.com/.

Santo Fortunato. Community Detection in Graphs. *Physics Reports*, 486(3–5):75–174, 2010. doi: 10.1016/j.physrep.2009.11.002.

Santo Fortunato and Marc Barthélemy. Resolution Limit in Community Detection. *Proceedings of the National Academy of Sciences*, 104(1):36–41, 2007. doi: 10.1073/pnas.0605965104.

Linton C. Freeman. Centrality in Social Networks: Conceptual Clarification. *Social Networks*, 1(3):215–239, 1979. doi: 10.1016/0378-8733(78)90021-7.

Michelle Girvan and M. E. J. Newman. Community Structure in Social and Biological Networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002. doi: 10.1073/pnas.122653799.

Mausam Goel and Ganesh Bagler. Computational Gastronomy: A Data Science Approach to Food. *Journal of Biosciences*, 47(1):12, 2022. doi: 10.1007/s12038-021-00248-1.

Zuzanna Grabowska. *O zmianach w kuchni i restauratorstwie dokonanych przez Auguste Escoffier na przełomie XIX i XX w.* PhD thesis, Jagiellonian University, 2025. URL https://ruj.uj.edu.pl/handle/item/554070.

Barry G. Green. Chemesthesis and the Chemical Senses as Components of a "Chemofensor Complex". *Chemical Senses*, 37(3):201–206, 2011. doi: 10.1093/chemse/bjr119.

Szilvia Gyimóthy and Reidar J. Mykletun. Scary Food: Commodifying Culinary Heritage as Meal Adventures in Tourism. *Journal of Vacation Marketing*, 15(3):259–273, 2009. doi: 10.1177/1356766709104271.

Christina Hartmann and Michael Siegrist. The Psychology of Eating Insects: A Cross-Cultural Comparison Between Germany and China. *Food Quality and Preference*, 44: 148–156, 2015. doi: 10.1016/j.foodqual.2015.04.013.

Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20)*, pages 639–648, 2020. doi: 10.1145/3397271.3401063.

Jarno Hoekman, Koen Frenken, and Robert J. W. Tijssen. Changing Spatial Patterns of Scientific Collaboration within Europe. *Research Policy*, 39(5):662–673, 2010. doi: 10.1016/j.respol.2010.01.012.

Paul W. Holland and Samuel Leinhardt. Local Structure in Social Networks. *Sociological Methodology*, 7:1–45, 1976. doi: 10.2307/270703.

Máirtín Mac Con Iomaire. Understanding the Heat — Mentoring: A Model for Nurturing Culinary Talent. *Journal of Culinary Science & Technology*, 6(1):43–62, 2008. doi: 10.1080/15428050701884196.

Máirtín Mac Con Iomaire and Helen Allen. Secrets of a Head Chef: Exploring Factors Influencing Success in Irish Kitchens. *Journal of Culinary Science & Technology*, 15(3): 187–222, 2017. doi: 10.1080/15428052.2016.1225538.

Paul Jaccard. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37:547–579, 1901. doi: 10.5169/seals-266450.

Anupam Jain and Ganesh Bagler. Analysis of Food Pairing in Regional Cuisines of India. *PLOS ONE*, 10(10):e0139539, 2015. doi: 10.1371/journal.pone.0139539.

Muin J. Khoury and John P. A. Ioannidis. Medicine. Big Data Meets Public Health. *Science*, 346(6213):1054–1055, 2014. doi: 10.1126/science.aaa2709.

Seunghyeon Kim, Jaeyun Sung, Mathias Foo, Yong-Su Jin, and Pan-Jun Kim. Uncovering the Nutritional Landscape of Food. *PLOS ONE*, 10(4):e0118697, 2015. doi: 10.1371/journal.pone.0118697.

Soon Hee Kim, Myung Sunny Kim, Myoung Sook Lee, Yong Soon Park, Hae Jeong Lee, Soon ah Kang, Hyun Sook Lee, Kyung-Eun Lee, Hye Jeong Yang, Min Jung Kim, Young-Eun Lee, and Dae Young Kwon. Korean Diet: Characteristics and Historical Background. *Journal of Ethnic Foods*, 3(1):26–31, 2016. doi: 10.1016/j.jef.2016.03.002.

Osame Kinouchi, Rosa W. Diez-Garcia, Adriano J. Holanda, Pedro Zambianchi, and Antonio C. Roque. The Non-equilibrium Nature of Culinary Evolution. *New Journal of Physics*, 10(7):073020, 2008. doi: 10.1088/1367-2630/10/7/073020.

Joseph B. Kruskal. On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem. *Proceedings of the American Mathematical Society*, 7(1):48–50, 1956. doi: 10.1090/S0002-9939-1956-0078686-7.

Solomon Kullback and Richard A. Leibler. On Information and Sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86, 1951. doi: 10.1214/aoms/1177729694. URL `https://doi.org/10.1214/aoms/1177729694`.

Vito Latora and Massimo Marchiori. Efficient Behavior of Small-World Networks. *Physical Review Letters*, 87(19):198701, 2001. doi: 10.1103/PhysRevLett.87.198701.

Vito Latora and Massimo Marchiori. A Measure of Centrality Based on Network Efficiency. *New Journal of Physics*, 9:188, 2007. doi: 10.1088/1367-2630/9/6/188.

Vito Latora, Vincenzo Nicosia, and Giovanni Russo. *Complex Networks: Principles, Methods and Applications*. Cambridge University Press, Cambridge, UK, 2017. ISBN 978-1-107-10318-4.

Antoine-Laurent Lavoisier. *Traité élémentaire de chimie: présenté dans un ordre nouveau et d'après les découvertes modernes*. Chez Cuchet, Paris, 1789. Volumes I–III.

Elizabeth A. Leicht and M. E. J. Newman. Community Structure in Directed Networks. *Physical Review Letters*, 100(11):118703, 2008. doi: 10.1103/PhysRevLett.100.118703.

Elizabeth A. Leicht, Petter Holme, and M. E. J. Newman. Vertex Similarity in Networks. *Physical Review E*, 73(2):026120, 2006. doi: 10.1103/PhysRevE.73.026120.

Vanessa Leschziner. *Recipes for Success: Culinary Styles and Professional Careers in Haute Cuisine*. PhD thesis, Rutgers University, 2007. URL `https://rucore.libraries.rutgers.edu/rutgers-lib/23483`.

David Liben-Nowell and Jon Kleinberg. The Link-Prediction Problem for Social Networks. *Journal of the American Society for Information Science and Technology*, 58(7):1019–1031, 2007. doi: 10.1002/asi.20591.

Harold McGee. *On Food and Cooking: The Science and Lore of the Kitchen*. Scribner, New York, 1984. ISBN 978-0-684-80001-1.

Giulia Menichetti and Albert-László Barabási. Nutrient Concentrations in Food Display Universal Behaviour. *Nature Food*, 3(5):375–382, 2022. doi: 10.1038/s43016-022-00511-0.

Giulia Menichetti, Babak Ravandi, Dariush Mozaffarian, and Albert-László Barabási. Machine Learning Prediction of the Degree of Food Processing. *Nature Communications*, 14:2345, 2023. doi: 10.1038/s41467-023-37457-1.

Menupan. Menupan. `https://www.menupan.com/`.

Michelin Guide Editors. L'Arpège — Michelin Guide. Michelin Guide Official Listing, 2024a. URL `https://guide.michelin.com/hu/en/ile-de-france/paris/restaurant/arpege`. Accessed: 2025-09-19.

Michelin Guide Editors. Osteria Francescana — Michelin Guide. Michelin Guide Official Listing, 2024b. URL https://guide.michelin.com/en/emilia-romagna/modena/restaurant/osteria-francescana. Accessed: 2025-09-19.

Reuven Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon. Network Motifs: Simple Building Blocks of Complex Networks. *Science*, 298(5594): 824–827, 2002. ISSN 0036-8075. doi: 10.1126/science.298.5594.824.

Ron Milo, Shalev Itzkovitz, Nadav Kashtan, Raphael Levitt, Shai Shen-Orr, Inbal Ayzenshtat, Michal Sheffer, and Uri Alon. Superfamilies of evolved and designed networks. *Science*, 303(5663):1538–1542, 2004. doi: 10.1126/science.1089167.

Massimo Montanari. *Medieval Tastes: Food, Cooking, and the Table*. Columbia University Press, New York, 2012. ISBN 9780231160841.

Vazrick Nazari, Antonella Pasqualone, Andrea Pieroni, Valentina Todisco, Sofia Belardinelli, and Telmo Pievani. Evolution of the italian pasta ripiena: the first steps toward a scientific classification. *Discover Food*, 4(1):57, 2024.

M. E. J. Newman. Assortative Mixing in Networks. *Physical Review Letters*, 89(20):208701, 2002. doi: 10.1103/PhysRevLett.89.208701.

M. E. J. Newman. Modularity and Community Structure in Networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006. doi: 10.1073/pnas.0601602103.

M. E. J. Newman. *Networks (2nd Edition)*. Oxford University Press, Oxford, UK, 2 edition, 2018. ISBN 978-0-19-252749-3.

M. E. J. Newman and Michelle Girvan. Finding and Evaluating Community Structure in Networks. *Physical Review E*, 69(2):026113, 2004. doi: 10.1103/PhysRevE.69.026113.

Mark E. J. Newman. Mixing Patterns in Networks. *Physical Review E*, 67(2):026126, 2003. doi: 10.1103/PhysRevE.67.026126.

Nordic Council of Ministers. New Nordic Food Manifesto. Policy Document, Nordic Council of Ministers, 2004. URL https://www.norden.org/en/information/new-nordic-food-manifesto. Accessed: 2025-09-19.

Xavier Ouvrard, Jean-Marie Le Goff, and Stephane Marchand-Maillet. Adjacency and tensor representation in general hypergraphs. part 2: Multisets, hb-graphs and related e-adjacency tensors. *arXiv preprint arXiv:1805.11952*, 2018.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, 1999. URL http://ilpubs.stanford.edu:8090/422/. Previous version appeared as Technical Report 1997-007.

Giovanni Palermo, Claudio Caprioli, and Giambattista Albora. Food Pairing Unveiled: Exploring Recipe Creation through Recommender Systems. *arXiv preprint*, arXiv:2406.15533, 2024. doi: 10.48550/arXiv.2406.15533. URL `https://arxiv.org/abs/2406.15533`.

Roderik Ponds, Frank van Oort, and Koen Frenken. The Geographical and Institutional Proximity of Research Collaboration. *Papers in Regional Science*, 86(3):423–443, 2007. doi: 10.1111/j.1435-5957.2007.00126.x.

R. C. Prim. Shortest Connection Networks and Some Generalizations. *Bell System Technical Journal*, 36(6):1389–1401, 1957. doi: 10.1002/j.1538-7305.1957.tb01515.x.

Hayagreeva Rao, Bernard Monin, and Rodolphe Durand. Institutional change in toque ville: Nouvelle cuisine as an identity movement in french gastronomy. *American Journal of Sociology*, 108(4):795–843, 2003. doi: 10.1086/367917.

Erzsebet Ravasz, Anna L. Somera, Dale A. Mongru, Zoltan N. Oltvai, and Albert-László Barabási. Hierarchical Organization of Modularity in Metabolic Networks. *Science*, 297 (5586):1551–1555, 2002. doi: 10.1126/science.1073374.

Francesco Ricci, Lior Rokach, and Bracha Shapira, editors. *Recommender Systems Handbook*. Springer, Cham, Switzerland, 3 edition, 2022. ISBN 978-1-0716-2196-7. doi: 10.1007/978-1-0716-2197-4.

S. E. Robertson and Hugo Zaragoza. The Probabilistic Relevance Framework: BM25 and Beyond. 3(4), 1009. ISSN 1554-0669. doi: 10.1561/1500000019.

Gerard Salton and Michael J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, 1983. ISBN 0070544840.

scikit-learn developers. Support vector classification (svc) — scikit-learn documentation. `https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html`, 2025. Accessed: 2025-10-20.

Robert Sedgewick and Kevin Wayne. 4.3 minimum spanning trees. Algorithms, Part I Lecture notes / Algs4, 2025. `https://algs4.cs.princeton.edu/43mst/`.

Alexander Semenov, Alexander Veremyev, Alexander Nikolaev, Eduardo L. Pasiliao, and Vladimir Boginski. Network-based Indices of Individual and Collective Advising Impacts in Mathematics. *Computational Social Networks*, 7(1):1–15, 2020. doi: 10.1186/s40649-019-0075-0.

M. Ángeles Serrano, Marián Boguñá, and Alessandro Vespignani. Extracting the Multiscale Backbone of Complex Weighted Networks. *Proceedings of the National Academy of Sciences*, 106(16):6483–6488, 2009. doi: 10.1073/pnas.0808904106.

Upendra Shardanand and Pattie Maes. Social Information Filtering: Algorithms for Automating "Word of Mouth". In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '95)*, pages 210–217. ACM, 1995. doi: 10.1145/223904.223931.

Paul W. Sherman and Geoffrey A. Hash. Why Vegetable Recipes Are Not Very Spicy. *Evolution and Human Behavior*, 22(3):147–163, 2001. doi: 10.1016/S1090-5138(00)00068-4.

Tiago Simas, Michal Ficek, Albert Diaz-Guilera, Pere Obrador, and Pablo R. Rodríguez. Food-Bridging: A New Network Construction to Unveil the Principles of Cooking. *Frontiers in ICT*, 4:14, 2017. doi: 10.3389/fict.2017.00014.

Bruce D. Smith. *The Emergence of Agriculture*. Scientific American Library, New York, 1998. ISBN 9780716750550.

Thorvald Sørensen. A Method of Establishing Groups of Equal Amplitude in Plant Sociology Based on Similarity of Species and Its Application to Analyses of the Vegetation on Danish Commons. *Biologiske Skrifter*, 5(4):1–34, 1948.

Rebecca L. Spang. *The Invention of the Restaurant: Paris and Modern Gastronomic Culture*. Harvard University Press, 2000. ISBN 978-0-674-00064-3.

TarlaDalal. Tarladalal. `https://www.tarladalal.com/`.

Chun-Yuen Teng, Yu-Ru Lin, and Lada A. Adamic. Recipe Recommendation Using Ingredient Networks. In *Proceedings of the 4th Annual ACM Web Science Conference (WebSci '12)*, pages 298–307. ACM, 2012. doi: 10.1145/2380718.2380757.

Hervé This. *Molecular Gastronomy: Exploring the Science of Flavor*. Columbia University Press, 2006. ISBN 978-0-231-13312-8.

Vincent A. Traag, Ludo Waltman, and Nees Jan van Eck. From Louvain to Leiden: Guaranteeing Well-Connected Communities. *Scientific Reports*, 9:5233, 2019. doi: 10.1038/s41598-019-41695-z.

Christoph Trattner and David Elsweiler. Food Recommender Systems: Important Contributions, Challenges and Future Research Directions, 2017. URL `https://arxiv.org/abs/1711.02760`.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

Enrico Vignoli and Marco Bolasco. *L'Atlante dei Cuochi: I miti, i maestri, gli allievi protagonisti della storia della cucina*. Slow Food Editore, 2023. ISBN 978-8884997760.

Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance. *Journal of Machine Learning Research*, 11:2837–2854, 2010. ISSN 1532-4435. URL `https://jmlr.org/papers/v11/vinh10a.html`.

Nicola Della Volpe, Lorenzo Mainetti, Alessio Martignetti, Andrea Menta, Riccardo Pala, Giacomo Polvanesi, Azzurra Ragone, Claudio Vairo, Fernando B. Pérez Maurera, Cesare Bernardis, and Maurizio Ferrari Dacrema. Lightweight Model for Session-Based

Recommender Systems with Seasonality Information in the Fashion Domain. In *Proceedings of the RecSys Challenge 2022 (RecSysChal '22)*, page 18–23. ACM, 2022. doi: 10.1145/3556702.3556829.

Justus von Liebig. *Researches on the Chemistry of Food*. Taylor & Walton, London, 1847.

Lifeng Wang, Qing Meng, and Chun-Hsien Su. From Food Supplements to Functional Foods: Emerging Perspectives on Post-Exercise Recovery Nutrition. *Nutrients*, 16(23):4081, 2024. doi: 10.3390/nu16234081.

Richard W. Wrangham, James H. Jones, Greg Laden, David Pilbeam, and NancyLou Conklin-Brittain. The Raw and the Stolen: Cooking and the Ecology of Human Origins. *Current Anthropology*, 40(5):567–594, 1999. doi: 10.1086/300083.

Andrea Zaccaria, Matthieu Cristelli, Andrea Tacchella, and Luciano Pietronero. How the Taxonomy of Products Drives the Economic Development of Countries. *PLOS ONE*, 9 (12):e113770, 2014. doi: 10.1371/journal.pone.0113770.

Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. Deep Learning Based Recommender System: A Survey and New Perspectives. *ACM Computing Surveys*, 52(1):1–38, 2019. doi: 10.1145/3285029.

Tao Zhou, Jie Ren, Matúš Medo, and Yi-Cheng Zhang. Bipartite Network Projection and Personal Recommendation. *Physical Review E*, 76(4):046115, 2007. doi: 10.1103/PhysRevE.76.046115.

Tao Zhou, Linyuan Lü, and Yi-Cheng Zhang. Predicting Missing Links via Local Information. *The European Physical Journal B*, 71(4):623–630, 2009. doi: 10.1140/epjb/e2009-00335-8.

Irit Zohar, Nira Alperson-Afil, Naama Goren-Inbar, Marion Prévost, Thomas Tütken, Guy Sisma-Ventura, Israel Hershkovitz, and Jens Najorka. Evidence for the Cooking of Fish 780,000 Years Ago at Gesher Benot Ya'aqov, Israel. *Nature Ecology & Evolution*, 6: 900–909, 2022. doi: 10.1038/s41559-022-01910-z.