Supplementary Materials

Food co-purchasing networks encode geographic disparities in cardiometabolic health

Antonio Desiderio[a,b,c,*], Leonardo Rodovero[a], Sune Lehmann[a,d,*]

[a]Department of Applied Mathematics and Computer Science, Technical University of Denmark, Richard Petersens Plads, 2800, Copenhagen, Denmark.

[b]ISI Foundation, Via Chisola 5, 10126, Turin, Italy.

[c]Centro Ricerche Enrico Fermi, Via Panisperna 89a, 00184, Rome, Italy.

[d]Copenhagen Center for Social Data Science (SODAS), University of Copenhagen, Øster Farimagsgade 5, 1353, Copenhagen, Denmark.

*Corresponding authors: antde@dtu.dk and sljo@dtu.dk.

Supplementary Materials, Section 1: Dataset Information

Column	Description
Store ID	Encrypted identifier uniquely identifying each store (`store_id_encrypted`). 553 stores are assigned to a municipality.
Latitude	Geographic latitude of the store (WGS84), taken from the checkout logs.
Longitude	Geographic longitude of the store (WGS84), taken from the checkout logs.

Supplementary Table 1 Store metadata used to locate checkouts in space and to aggregate baskets by municipality.

Column	Description
Municipality	Official municipality name, aligned to StatBank nomenclature.
Population	Residential population used for per-capita downsampling of baskets.
Share under 30	Share of residents aged 0–29 years (BY2).
Share over 60	Share of residents aged 60 years and older (BY2).
Affluence index	$(N_{\geq 700\text{k}} - N_{<300\text{k}})/N_{\text{families}}$ from disposable family income (INDKF132).
Rate A10	Unique redeemers of antidiabetics (ATC A10) per resident (MEDI1, 2024).
Rate C	Unique redeemers of cardiovascular drugs (ATC C) per resident (MEDI1, 2024).
Rate C10	Unique redeemers of lipid-modifying agents (ATC C10) per resident (MEDI1, 2024).
Age stratum	For age-specific rates: 0–17, 18–29, 30–59, and ≥ 60 years.

Supplementary Table 2 2024 StatBank sociodemography and prescription redemption rates for each municipality. Age-aggregated rates use the municipal population as denominator; analyses among residents ≥ 60 years use the corresponding age-specific rates.

Column	Description
Receipt ID	Encrypted basket identifier. One ID is one checkout.
Product ID	Encrypted Product ID.
Store ID	Encrypted store at which the basket was recorded.
Date	Calendar day of the checkout.

Supplementary Table 3 Bipartite basket–product incidence. Each row is a purchased item. The bipartite graph has baskets on one side and products on the other; an edge exists if the product appears in the checkout.

Column	Description
Product ID	Encrypted product ID.
Crosswalk status	Mapped, Unmapped, or Non-Food. Nutrient joins use Mapped only.
Frida food ID	DTU Frida 5.5 FoodID.
Food group	Frida food group (DTU taxonomy, distinct from the retailer tree).
Energy	Energy (kcal per 100 g).
Macronutrients	Protein, fat, available carbohydrate, and dietary fiber (g per 100 g).
Sodium	Sodium (mg per 100 g).
Added / free sugar	Added sugar and free sugars (g per 100 g).
Saturated fat	Sum of saturated fatty acids (g per 100 g).
NOVA (FoodProX)	FoodProX NOVA group $\{1, 2, 3, 4\}$.
FScore	FoodProX continuous processing score (higher = more processed).

Supplementary Table 4 Product–nutrition link: hybrid retail-to-Frida crosswalk and Frida 5.5 fields used in the analyses.

Macro-category	Meso-categories
Fresh food	Chilled ready meals & deli; dairy, milk & eggs; meat, pork & poultry; fresh bakery & bread; fruits; vegetables & greens; seafood & fish
Frozen foods	Ice cream & frozen desserts; frozen meals & pizza; frozen meat & seafood; frozen vegetables & fruits
Pantry & packaged food	Baking supplies & flour; sauces, dips & spreads; oils, vinegars & condiments; canned goods & soups; cereals, oats & breakfast; pasta, rice & grains
Snacks & sweets	Cookies, cakes & sweet biscuits; nuts, seeds & dried fruit; energy bars & snacks; chips, crisps & salty snacks; chocolate, candy & sweets
Beverages	Wine, spirits & alcohol; juices, smoothies & soft drinks; beer & cider; coffee, tea & hot drinks; water (still & sparkling)

Supplementary Table 5 Food meso-categories used in the analyses, grouped by macro-category. This is the mapped taxonomy by the data provider.

Supplementary Materials, Section 2: Downsampling stability

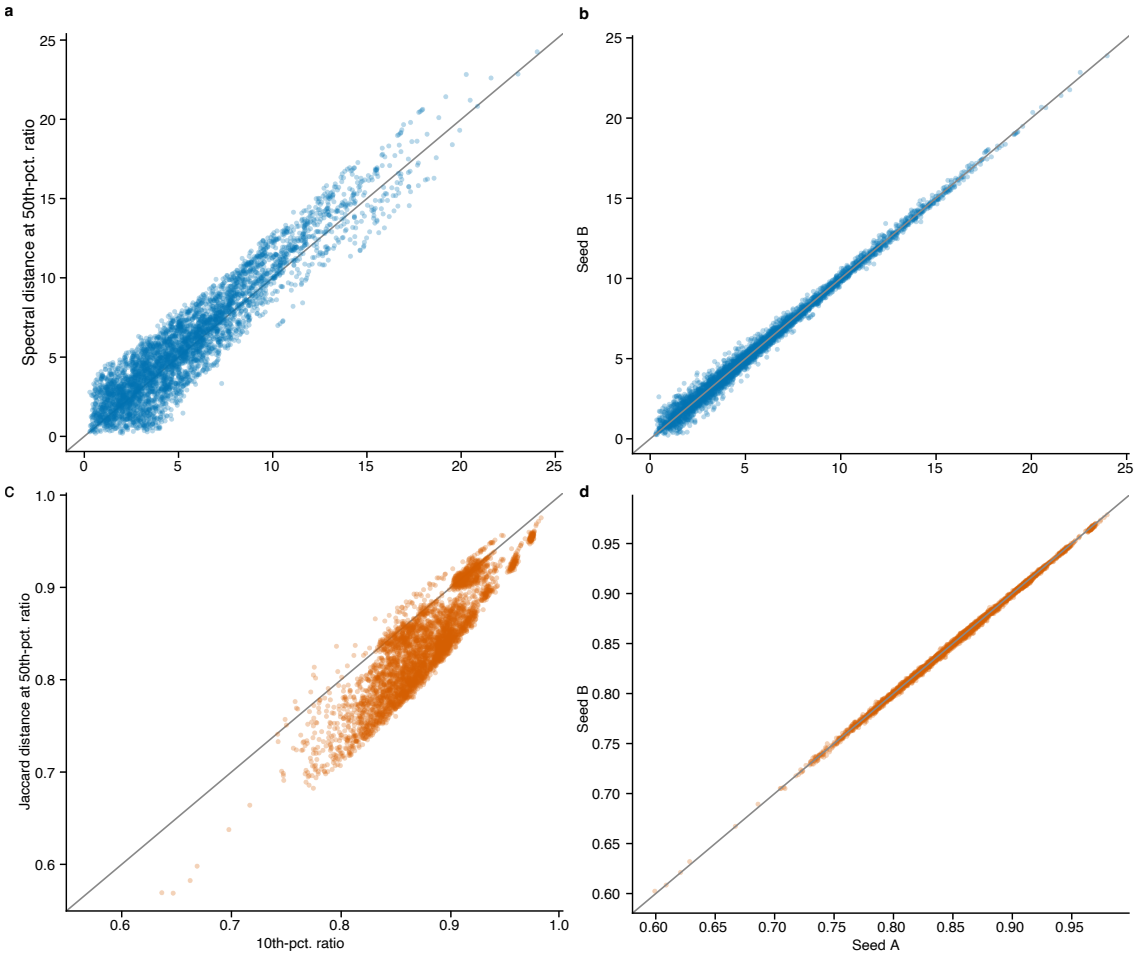
To assess the stability of the geographic pattern, we repeated the population-scaled downsampling while varying the basket-per-person ratio and the random seed used to sample receipts. Both analyses used the RNA approximation to the Bipartite Configuration Model with Bonferroni correction at $\alpha = 0.001$, matching the procedure used in the main analysis [1–3].

Within each municipality, a single receipt permutation was used to construct nested samples truncated at the 10th, 25th, and 50th percentiles of municipal coverage, corresponding to ≈ 3.80, 5.41, and 7.73, respectively. The resulting pairwise spectral-distance matrices were highly correlated, with Spearman $\rho = 0.973$ between the 10th and 25th percentiles, $\rho = 0.965$ between the 25th and 50th percentiles, and $\rho = 0.907$ between the 10th and 50th percentiles (see Figure S1 **a**). The same pattern was observed for the urban-to-peripheral axis, defined as the mean spectral distance of each of the remaining 91 municipalities from the four-city core. The rankings obtained at the three basket-density levels had Spearman correlations of $\rho = 0.989$, 0.982, and 0.955 for the same three pairwise comparisons. A median split of this axis, corresponding to the near/mid–far classification used in the main text, gave adjusted Rand indices of 0.83 between adjacent ratios and 0.68 between the 10th and 50th percentiles.

We next evaluated the effect of the random seed used for receipt sampling. An analysis of 10 municipalities, including the four cities together with rural municipalities, was repeated for 100 independent draws at the production 25th-percentile ratio. Across the 4,950 pairs of runs, the pairwise spectral-distance matrices had a mean Spearman correlation of $\rho = 0.997$ (median 0.997, range 0.991–1.000). Five independent draws on the full set of 95 municipalities gave similarly high correlations, with $\rho = 0.992$–0.996 (see Figure S1 **b**).

The corresponding urban-to-peripheral spectral axis was also highly consistent across seeds ($\rho \geq 0.996$), and the near/mid–far median split was almost identical across all replicates (ARI ~ 1). The same stability is observed for pairwise edge Jaccard similarity (see Figure S1 **c** and **d**), providing a complementary measure of the consistency of the underlying network structure. Overall, the geographic partition is consistent across the tested downsample seeds and remains similar when the basket-density ratio is varied over approximately a twofold range, with both spectral distance and edge Jaccard similarity showing consistent network structure across downsampling choices.

The processing-assortativity gradient is computed on the municipal graphs from the main analysis, obtained using the exact BiCM with Bonferroni correction ($\alpha = 0.001$), rather than on the RNA backbones used in the ratio/seed sensitivity analysis above. Reconstructing the exact backbones at the 25th and 50th coverage percentiles gives nearly identical municipal rankings with r_{FScore} [4] at q25 correlating with the main-analysis values at $\rho = +0.98$, while q25 and q50 correlate at $\rho = +0.96$. The q25 networks also retain the same ordering along the spectral axis of the main analysis ($\rho = +0.88$), with an edge-set Jaccard similarity of 0.79 relative to the main-analysis graphs. In addition, the main-analysis r_{FScore} remains correlated with the RNA spectral axis reported in this Supplementary Materials, Section 2 ($\rho = +0.79$). These results show that the urban–peripheral gradient in processing assortativity is robust to changes in basket density within the exact-BiCM construction and is not specific to the RNA sensitivity networks.

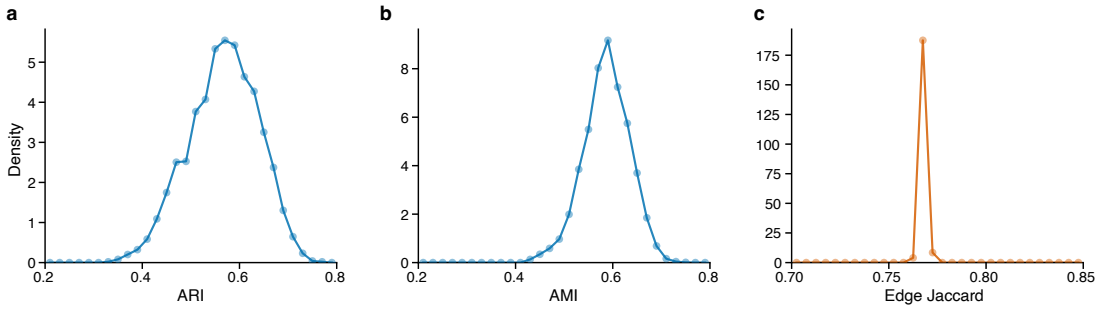


Supplementary Figure S1 Downsampling stability of municipal spectral and Jaccard geography. Pairwise distances among 95 municipalities (upper triangle of the distance matrix) after population-scaled RNA backbones (Bonferroni $\alpha = 0.001$). In all panels, the gray line is the identity $y = x$. **a)** Spectral distance (normalized Laplacian) at the 10th versus 50th percentile of municipal basket/person coverage (Spearman $\rho = 0.907$). **b)** Spectral distance for two independent draws at the 25th-percentile ratio (Spearman $\rho = 0.993$). **c)** Pairwise edge Jaccard distance at the same 10th versus 50th coverage percentiles (Spearman $\rho = 0.878$). **d)** Pairwise edge Jaccard distance for the same two independent seeds (Spearman $\rho = 0.999$). The urban–peripheral ranking is recovered under both perturbations.

We repeated the pooled global network pipeline on the ten-municipality pilot used for local downsample checks (four urban centers plus rural municipalities). For each of 100 independent receipt draws at the production 25th-percentile basket density (baskets-per-resident ≈ 5.41), we pooled downsampled baskets, fitted a global RNA BiCM with Bonferroni correction at $\alpha = 0.001$, and retained the validated edge catalog ($\sim$87 000 edges per replicate). For every unordered pair of replicates we computed the Jaccard overlap of global edge sets and the agreement between Louvain partitions on the intersection of nodes [5](Adjusted Rand Index and Adjusted Mutual Information). Across 4,950 pairs the global edge catalog was highly reproducible (Jaccard 0.768 ± 0.001; 5th–95th percentile 0.766–0.770), and community structure was consistently preserved (mean ARI 0.566, mean AMI 0.587; see Figure S2). The cardiometabolic analysis is similarly robust to the downsampling seed, as repeating the main-text construction

on two independent national redraws at the same 25th-percentile basket density yields comparable leave-one-out fits. Louvain detected ∼7 communities per replicate (mean 7.3, s.d. 0.7). Thus the downsample seed does not materially alter the global backbone; remaining variation in ARI/AMI reflects soft boundaries in the Louvain partition rather than wholesale restructuring of the catalog. We observe the same pattern on the full national pool. Comparing the main-text global network with two independent redraws at the same 25th-percentile ratio yields pairwise Jaccard 0.740 in all three pairs, and Louvain agreement at least as high as in the previous ten-municipality pilot (mean ARI 0.67, mean AMI 0.64).

The global network reported in the main text uses the RNA approximation to estimate the upper-tail probabilities of the Poisson–Binomial distributions arising from the BiCM. As a robustness check, we refitted the national pooled network using the Lugannani–Rice saddlepoint approximation [6], which estimates the tail probability from the cumulant generating function, evaluating both approximations directly on the full Bonferroni-validated graph ($\alpha = 0.001$, prior to the $z^* = 28$ percolation cutoff; 3,171 nodes and 60,126 edges). On this unfiltered backbone, the two approximations produced nearly identical leave-one-out cardiometabolic fits ($R^2_{S+I} = 0.63$, 0.93, and 0.75 under RNA versus 0.61, 0.94, and 0.76 under Lugannani–Rice for A10, C, and C10, respectively, closely matching the 0.61, 0.94, and 0.75 obtained at $z^* = 28$) and an essentially identical municipal ranking of mean NOVA 4–4 log-lift (Spearman $\rho = 0.998$).

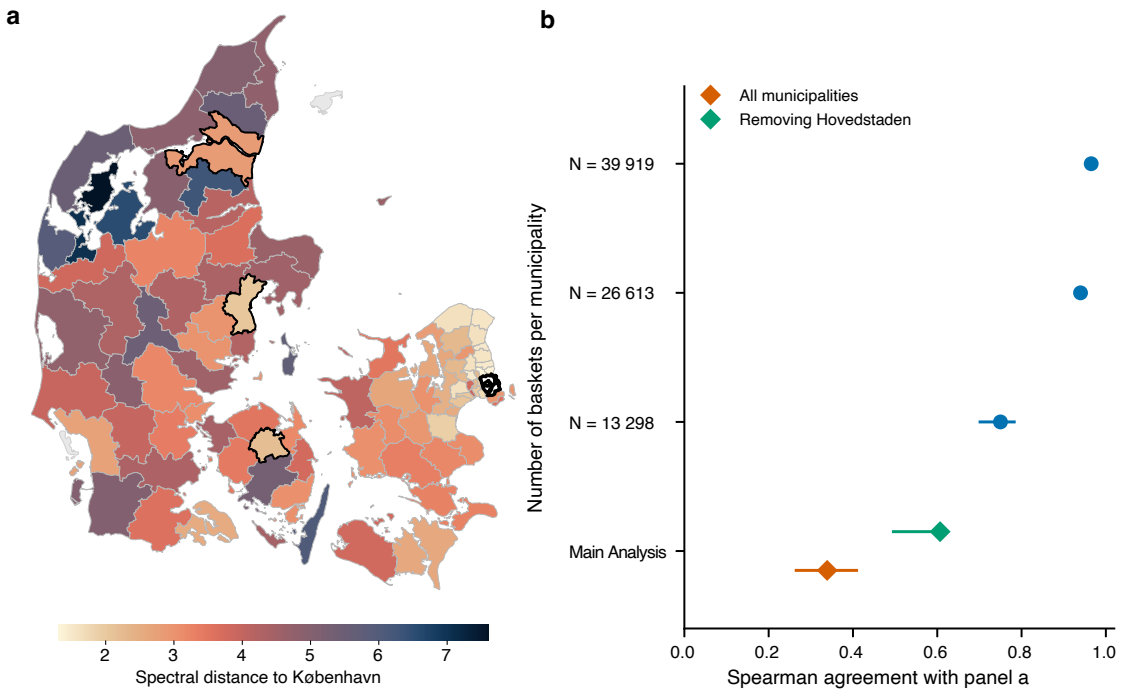


Supplementary Figure S2 Stability of pooled global network backbones and community structure across downsampling seeds. Distributions of partition concordance and edge-set overlap across all 4,950 unordered pairs of 100 independent receipt draws on the ten-municipality pilot at the 25th-percentile basket coverage (baskets-per-resident $\approx$ 5.41, RNA BiCM with Bonferroni $\alpha = 0.001$). **a)** Distribution of the Adjusted Rand Index (ARI) between Louvain community partitions (mean 0.566). **b)** Distribution of the Adjusted Mutual Information (AMI) for the same community partitions (mean 0.587). **c)** Pairwise Jaccard similarity distribution of the validated edge sets (0.768 ± 0.001, 5th–95th percentile 0.766–0.770). Variation in ARI and AMI reflects soft boundaries in Louvain modularity optimization rather than substantive changes in the validated edge catalog.

In the main analysis, each municipality is sampled in proportion to its population, so larger municipalities enter the projection with more baskets N and greater statistical power under Bonferroni validation. In a direct comparison between København and Norddjurs, equalizing sample size ($N \approx$ 190,000 baskets), applying an effect-size filter ($z/\sqrt{N}$), or permuting municipality labels across baskets within matched basket-size bins ($p < 0.01$) keeps their spectral distance well above resampling noise, whereas København remains indistinguishable from its immediate urban neighbor Frederiksberg or noticeably closer to urban centers such as Odense

and Aarhus. To apply the same control across all 95 municipalities, we fixed both the number of baskets and the basket-size distribution (number of distinct products per receipt, binned from 1 to 39 and $\geq$ 40), setting each bin to the minimum count available across municipalities to account for the fact that smaller baskets produce fewer co-purchases and sparser backbones even at equal N, yielding $N^* = 53{,}251$ baskets per municipality. We drew one hundred independent samples without replacement per municipality, projected each with the exact BiCM (Bonferroni $\alpha = 0.001$), and ranked municipalities by their mean spectral distance to København.

Fixing N^* and matching basket sizes brings 21 of the 27 Capital Region suburbs into the closest tertile to København, reflecting a metropolitan organization consistent with [7] (see Figure S3**a**). Because of this high similarity across the Capital Region, the other major urban centers (Aarhus, Odense, and Aalborg) – although still close to København and distinct from their surrounding rural municipalities – do not form a cleanly separated four-city cluster, which is why we measure spectral distance directly relative to København. The resulting ranking is stable under smaller common samples, showing Spearman correlations with the N^* ordering of $\rho = 0.97$, 0.94, and 0.75 for $3N^*/4$, $N^*/2$, and $N^*/4$ baskets, and of $\rho = 0.99$ with the effect-size filter $z/\sqrt{N} \geq 0.05$, while agreeing with the main-text ranking at $\rho = 0.61$ outside the Capital Region ($\rho = 0.34$ overall; see Figure S3**b**) and preserving a monotonic decline in edge overlap with København across fixed-N^* tertiles (0.44, 0.40, and 0.36; $\rho = -0.85$). Recomputing community intensities I_c on these fixed-N^* draws also recovers the main-text cardiometabolic fits (R^2_{S+I} of 0.59, 0.93, and 0.73 across all ages and 0.43, 0.57, and 0.47 for residents aged $\geq$ 60, with all 16 headline community coefficients keeping their sign and 13 of 16 remaining significant at $p < 0.05$). Finally, on these fixed-N^* backbones, UPFs remain peripheral (AUC < 0.5, mean 0.37) and maintain higher 4–4 than 1–1 log-lift in all 95 municipalities, while processing assortativity r_{FScore} increases with spectral distance from København both on padded spectra ($\rho = +0.38$, $p < 0.001$) and on unpadded shared-product spectra ($\rho = +0.210$, $p = 0.041$), whereas degree assortativity r_{degree} remains flat on padded spectra ($\rho = -0.11$, $p = 0.31$) and decreases on shared-product spectra ($\rho = -0.312$, $p = 0.002$), resulting in a positive gradient for excess processing assortativity $r_{\text{FScore}} - r_{\text{degree}}$ ($\rho = +0.419$, $p < 0.001$).



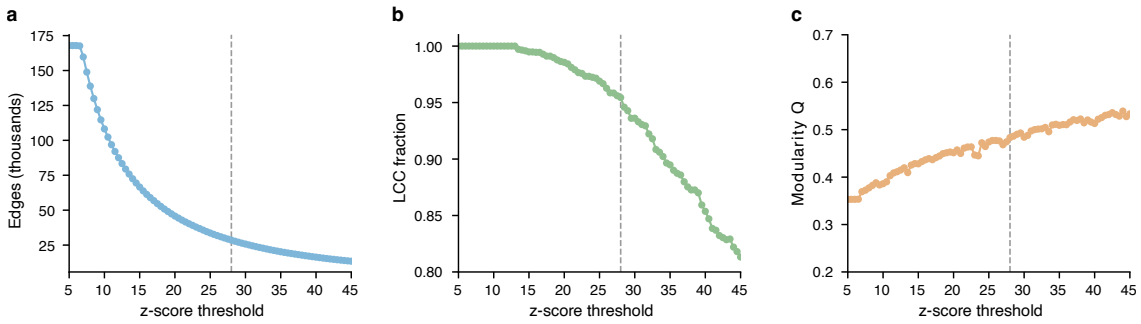
Supplementary Figure S3 Spectral distance to København at a fixed number of baskets and a fixed basket-size distribution. a) Mean spectral distance between each municipality's backbone and København's across independent draws at $N^* = 53{,}251$ baskets with identical basket-size distributions (exact BiCM, Bonferroni $\alpha = 0.001$). Lighter shades are closer to København (white); Aarhus, Odense, and Aalborg are outlined; municipalities without data are gray and Bornholm is omitted. **b)** Spearman rank correlation with **a**. Blue circles indicate the same basket-size mix drawn at smaller common N, with points showing the mean across draws and bars the leave-one-draw-out range. Diamonds indicate agreement with the population-scaled ranking in the main text, both across all municipalities and outside the Capital Region (Region Hovedstaden).

Supplementary Materials, Section 3: National co-purchase network and communities

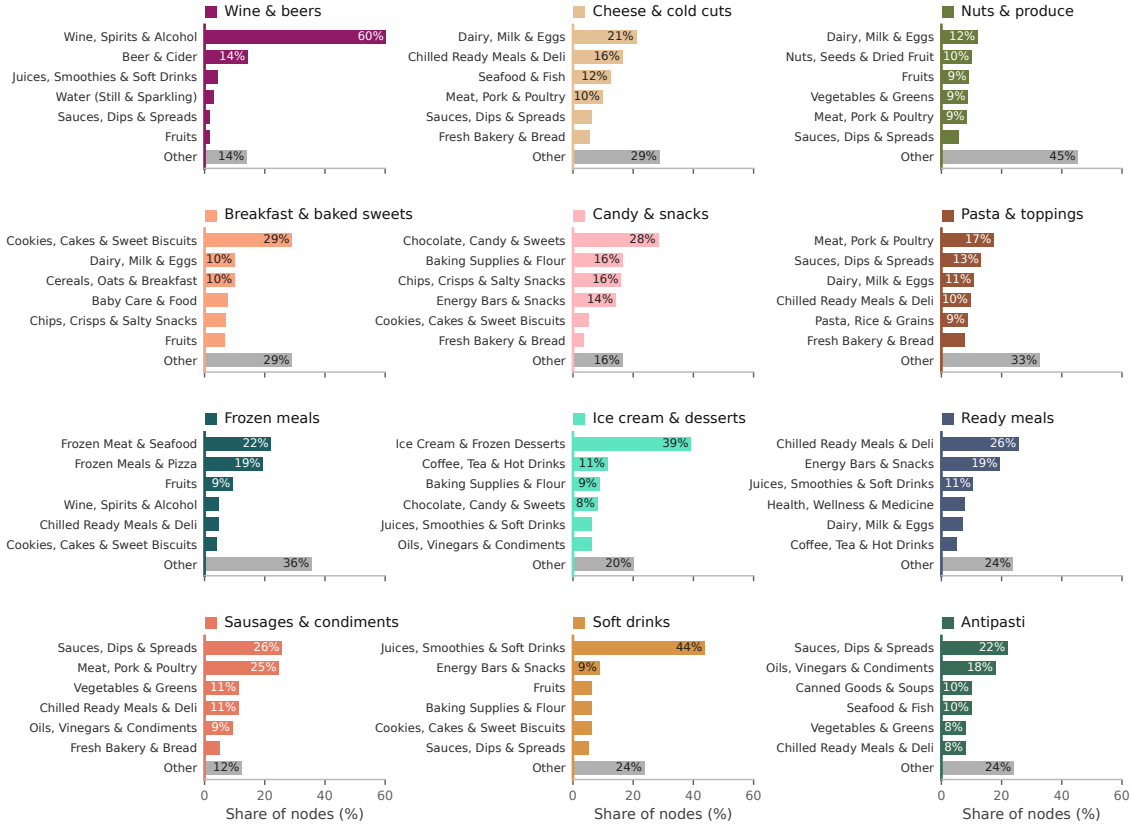
To construct the national food co-purchase network, we selected a robust threshold for downstream community detection by analyzing the network topology as a function of the BiCM-derived z-score cutoff [1, 2]. As shown in Figure S4, we tracked three key metrics: the number of preserved edges, the fraction of non-isolated nodes in the largest connected component (LCC), and the Louvain modularity Q ($\gamma = 1$).

Increasing the significance threshold leads to a monotonic, power-law-like pruning of connections, reducing the edge count from over 168,000 at $z = 5$ down to 28,435 at $z^* = 28$ (an overall reduction exceeding 80%). Importantly, the modularity profile (see Figure S4 **c**) reveals that the network already possesses an intrinsically modular organization at permissive thresholds ($Q \approx 0.4$ at $z = 10$), indicating that product clustering reflects genuine consumption patterns rather than an artifact of the filtering [5, 8]. As z increases, modularity grows steadily up to $Q \approx 0.49$ at $z^* = 28$ as spurious inter-cluster ties are systematically eliminated. In contrast, the integrity of the giant component displays a pronounced transition (see Figure S4 **b**). While the LCC remains fully intact (100% of non-isolate nodes) up to $z \approx 13$ and degrades smoothly up to $z \approx 28$ before higher thresholds induce steep percolation-like fragmentation (dropping to 81% at $z = 45$), we identified $z^* = 28$ as the optimal trade-off – the most stringent threshold preserving $\geq 95\%$ of nodes within a single coherent component.

We next characterized the composition of these communities according to the product categories of their nodes (see Figure S5). For each community, we report the share of nodes belonging to its six most frequent meso-categories, while all remaining categories are grouped as Other. The resulting partition provides the community structure used to compute municipal community intensities I_c in the cardiometabolic analysis.

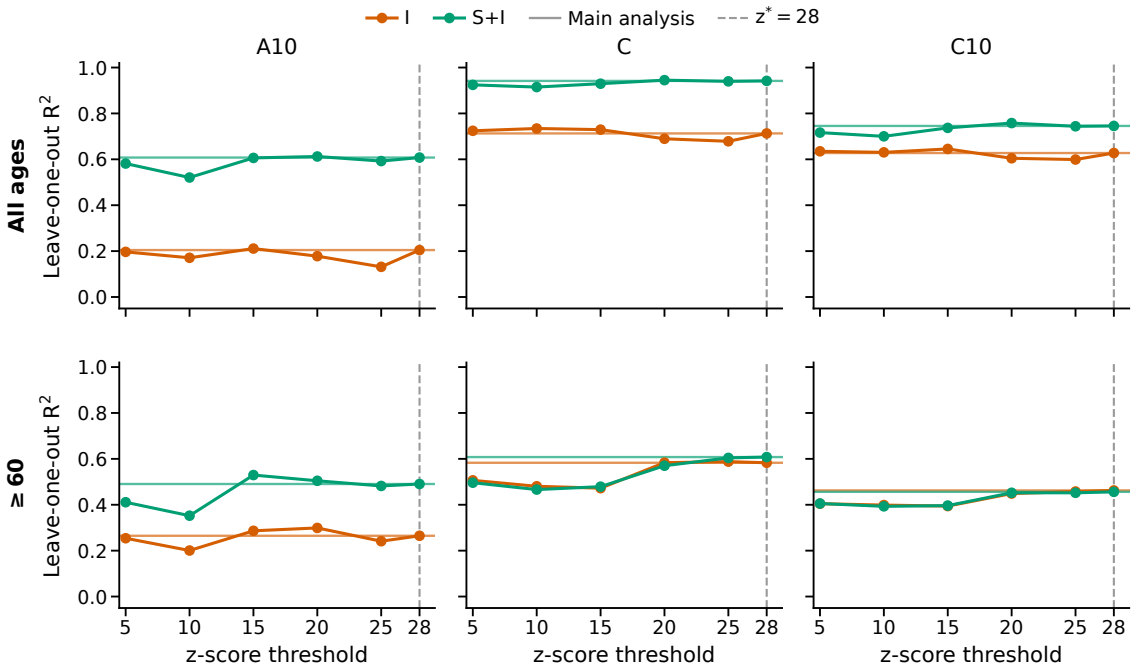


Supplementary Figure S4 Network topological properties across statistical validation thresholds. a) Number of retained edges (in thousands) as a function of the z-score cutoff. **b)** Fraction of non-isolate nodes belonging to the largest connected component (LCC). **c)** Modularity Q resulting from Louvain community detection ($\gamma = 1$). In all panels, the vertical dashed gray line marks the selected operating threshold $z^* = 28$, which balances the removal of weak cross-category ties against the onset of topological fragmentation.



Supplementary Figure S5 Product category composition of national dietary co-purchasing communities. Breakdown of food categories across the 12 major dietary communities identified via Louvain modularity optimization on the validated national co-purchasing backbone ($z^* = 28.0$). Bars represent the percentage share of product nodes belonging to each food category within each community. Minor categories are aggregated under "Other" and displayed in gray.

To assess whether the cardiometabolic association is tied to the cutoff, we repeated the I_c construction at different z-score thresholds, recomputing Louvain communities on the LCC at each threshold rather than imposing the z^* partition (see Figure S6). Leave-one-out R^2 for I and $S+I$ recovers the main-analysis values for A10, C and C10 in both the age-aggregated and ≥ 60 specifications, indicating that the result at $z^* = 28$ is robust to the choice of threshold. Repeating that evaluation on the LCC at $z \in \{10, 15, 20, 25, 28\}$, 4–4 mean log-lift exceeded 1–1 in all 95 municipalities at every threshold, and the Spearman association with spectral distance remained negative and significant for $z \geq 15$ (ρ from -0.34 to -0.58, $p < 0.001$), as in the main text.



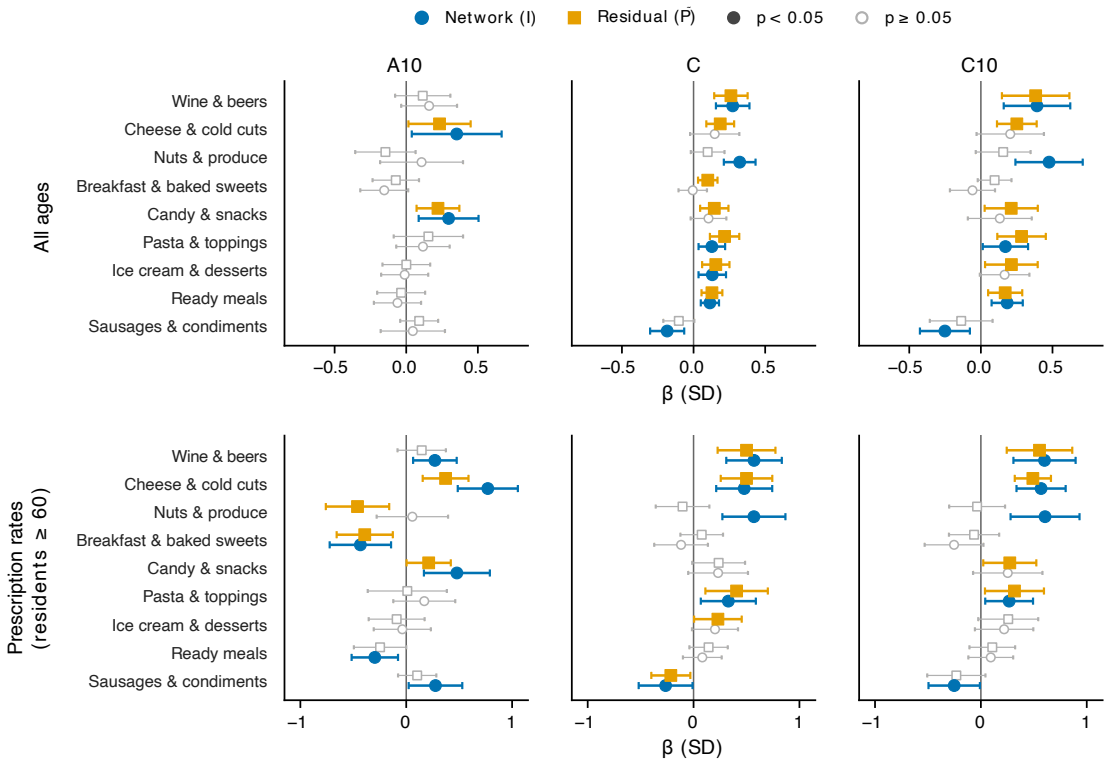
Supplementary Figure S6 Cardiometabolic I_c fits across z-score thresholds. Leave-one-out Ridge R^2 for community intensities I (orange) and the joint model $S+I$ (green) across z-score thresholds, with communities re-derived at each cutoff using Louvain on the LCC. Results are shown for age-aggregated rates (top) and rates among residents ≥ 60 years (bottom) across A10, C, and C10. Horizontal lines indicate the main-analysis R^2 at $z^* = 28$, while the vertical dashed line marks $z^* = 28$.

Supplementary Materials, Section 4: Community-level regression coefficients and spatial robustness

To identify the dietary communities contributing to the block-level association between local co-purchasing structure and cardiometabolic outcomes, we fitted one-community-at-a-time OLS models of municipal prescription rates on the sociodemographic block S and a single community-level co-purchasing feature. Predictors were standardized, and heteroskedasticity-robust HC3 standard errors were used throughout. For each community, we considered either the mean intra-community log-lift I_c or the residualized co-purchase measure $\widetilde{P}$. We considered nine of the twelve major Louvain communities identified in the national backbone [5]; frozen meals, soft drinks, and antipasti were excluded because fewer than 50 local intra-community edges were available for these categories in certain municipalities.

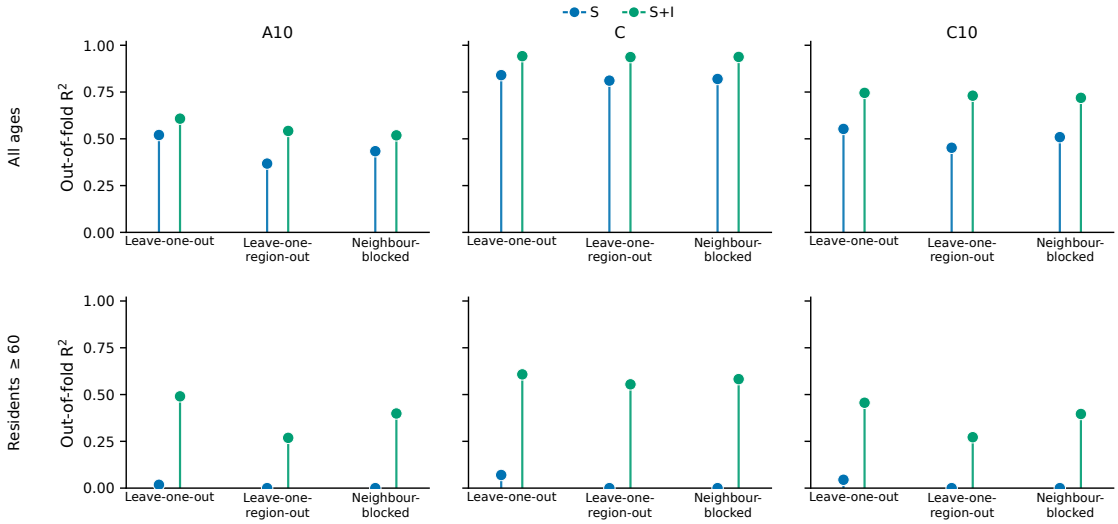
The resulting standardized regression coefficients are shown in Figure S7. For age-aggregated rates, the clearest associations are observed for A10. Candy & snacks and Cheese & cold cuts show positive associations with antidiabetic prescription rates for both measures of co-purchasing cohesion [9, 10], with coefficients of approximately $\beta = 0.30$ and 0.35 for I_c, and $\beta = 0.22$ and 0.23 for $\widetilde{P}$, respectively. Wine & beers is positively associated with both cardiovascular (C) and lipid-modifying (C10) rates, with I_c coefficients of approximately $\beta = 0.27$ and 0.39, respectively, and the associations remain positive for $\widetilde{P}$. Nuts & produce shows comparatively large positive I_c coefficients for C and C10, but these associations are not retained for $\widetilde{P}$, indicating that their apparent relationship with cardiometabolic rates is largely accounted for by local product prevalence.

The age-specific analysis shows a broadly similar pattern despite the reduced sociodemographic adjustment. In particular, Candy & snacks and Cheese & cold cuts remain positively associated with A10, while Wine & beers remains positively associated with C and C10. Several coefficients are larger in magnitude than in the age-aggregated analysis, consistent with the substantially lower explanatory power of the sociodemographic baseline in this age group. Overall, the community-level analysis therefore identifies a small set of dietary communities whose co-purchasing cohesion is consistently associated with cardiometabolic outcomes, while also distinguishing associations that persist after accounting for local product prevalence from those that do not. Because the HC3 standard errors used in the regression models of the main analysis address heteroskedasticity but not residual spatial autocorrelation, we assessed Moran's I on OLS residuals using row-standardized 5-nearest-neighbor weights on municipal centroids and 999 permutations, retaining the same complete-case samples as in the main analysis. For age-aggregated rates, the rates themselves are spatially clustered for A10, C, and C10 ($I = 0.36$, 0.45, and 0.42; all $p_{\text{sim}} = 0.001$). Adjusting for S reduces this autocorrelation but does not completely remove it ($I = 0.27$, 0.35, and 0.34; all $p_{\text{sim}} = 0.001$). Adding the nine community intensities I_c reduces Moran's I to 0.19, 0.03, and 0.15 ($p = 0.002$, 0.214, and 0.005), leaving no evidence of residual spatial dependence for C. Among residents aged ≥ 60, residual clustering remains high after S ($I = 0.54$, 0.35, and 0.43), while $S{+}I$ lowers these values ($I = 0.19$, 0.11, and 0.25). Sociodemographic covariates alone are therefore insufficient to explain spatial correlation between neighboring municipalities; community-level indicators absorb spatial structure that demographic factors leave unexplained.



Supplementary Figure S7 Community-level associations between food co-purchasing structure and cardiometabolic prescription rates. Standardized OLS regression coefficients (β, in outcome standard deviations) from municipality-level models including the sociodemographic block S and one community-level co-purchasing measure at a time. From left to right, panels show associations with antidiabetic (A10), cardiovascular (C), and lipid-modifying (C10) prescription rates, for age-aggregated rates (top) and residents aged $\geq$ 60 (bottom). Blue points show mean intra-community log-lift (I_c), while orange squares show residualized co-purchase ($\widetilde{P}$); horizontal bars indicate HC3 robust 95% confidence intervals. Filled symbols indicate $p < 0.05$ while open symbols $p \geq 0.05$.

Ordinary leave-one-out Ridge can leak spatial information if neighboring municipalities stay in the training set. We therefore repeated the leave-one-out Ridge models of the main analysis with leave-one-region-out and with neighbor-blocked leave-one-out, which holds out municipality i and its five nearest neighbors and predicts only i. Figure S8 reports out-of-fold R^2 for S and $S+I$, showing that the increment after S remains in both cases. For age-aggregated rates, the residual shares after S (A10/C/C10) are 18/63/43% (leave-one-out), 15/65/43% (neighbor-blocked), and 28/67/51% (leave-one-region-out), while among residents aged $\geq$ 60, S stays near zero out of fold whereas $S+I$ remains large (neighbor-blocked residual shares: 45/60/45%). Taken together, these results demonstrate that the predictive power of community intensities is robust to spatial dependency, absorbing municipal-level spatial structure without relying on local information leakage.

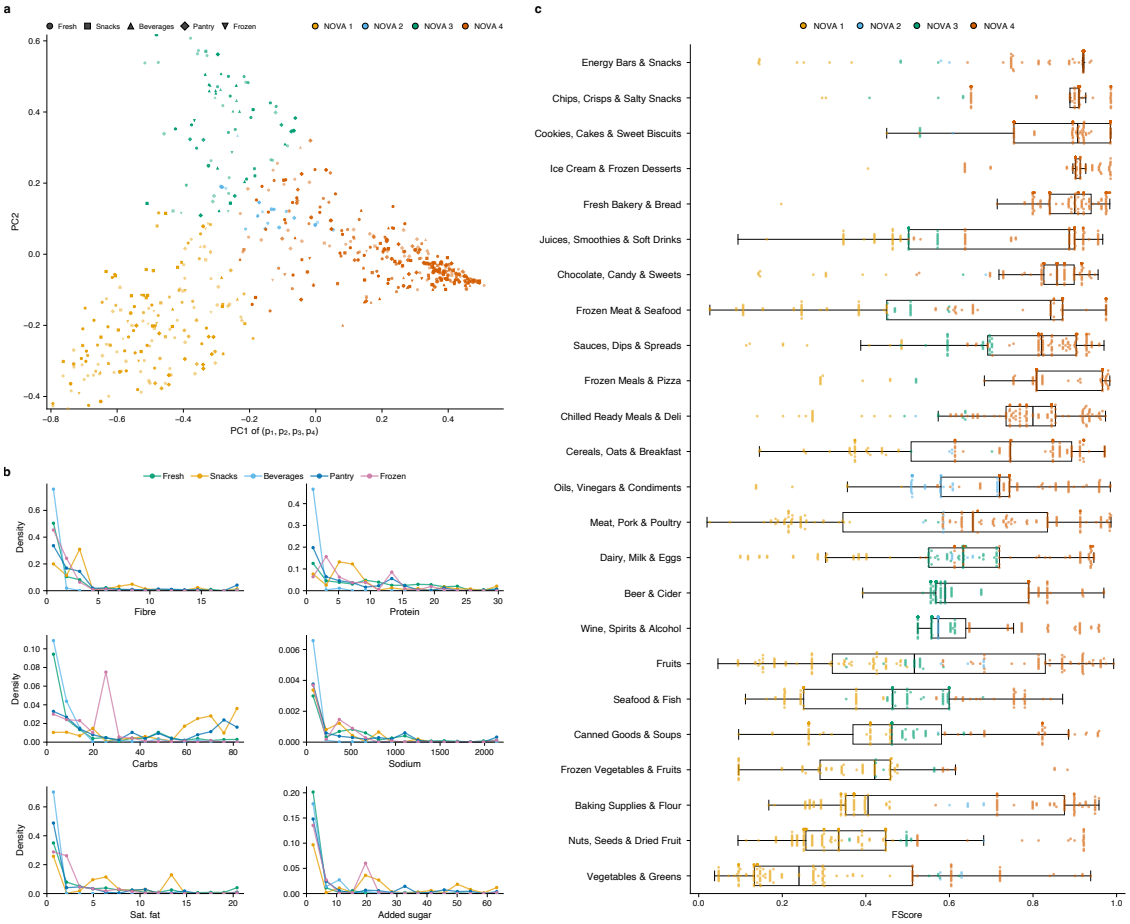


Supplementary Figure S8 Out-of-fold predictive performance across spatial cross-validation schemes. Out-of-fold coefficient of determination (R^2) from Ridge regression models predicting municipal prescription rates using sociodemographic covariates alone (S) and including the nine community intensities ($S + I$). From left to right, panels show models for antidiabetic (A10), cardiovascular (C), and lipid-modifying (C10) prescription rates, for age-aggregated rates (top) and residents aged ≥ 60 (bottom). Within each panel, we compare ordinary leave-one-out, neighbor-blocked leave-one-out (holding out municipality i and its five nearest neighbors while predicting only i), and leave-one-region-out schemes; bars report models with S alone versus $S + I$.

Supplementary Materials, Section 5: Structural Validation of Food Mapping and Processing Scores

To quantify the performance of individual mapping tiers, we benchmarked each step independently against Open Food Facts NOVA classifications [11]. Direct FoodEx2 linkage [12] ($n = 783$) achieved the highest fidelity (69.1% overall accuracy, UPF $F_1 = 0.846$, and 79.7% on NOVA 1 vs. 4), but covered only approximately one-quarter of the retail catalog. Unconstrained string matching alone reduced performance to 61.7% overall accuracy ($F_1 = 0.763$; 69.2% on NOVA 1 vs. 4), primarily due to token collisions across processing states (e.g., canned fish matching raw fish, chocolate dairy snacks matching buttermilk via the shared token "milk", or snack bars matching peanut butter). When mapped Frida foods and local on-pack nutrient profiles were compositionally close (top distance tercile), categorical NOVA agreement reached 87.9% with a Pearson correlation of $r = 0.70$ on the continuous processing score, confirming that generic Frida proxies preserve the product-level processing signal.

After mapping, each retail product inherits the complete Frida nutrient profile, the FoodProX class probability simplex (p_1, p_2, p_3, p_4), and the continuous processing metric *F*Score [4, 13] ($n = 8{,}672$ food products mapped to 729 distinct Frida reference foods, excluding non-food lines). Principal component analysis of (p_1, p_2, p_3, p_4) reveals a clear separation between minimally processed items (NOVA 1) and ultra-processed formulations (NOVA 4), forming two distinct clusters at opposite ends of the first principal component (see Figure S9**a**). Fresh commodities (vegetables, fresh meat, produce, seafood) concentrate at the minimally processed pole, whereas confectioneries, savory snacks, and industrial bakery products anchor the ultra-processed extreme; frozen and packaged pantry goods bridge the intermediate space. Nutrient distributions systematically reflect these processing differences (see Figure S9**b**), with fresh categories supplying the bulk of dietary fiber and protein, while snacks and sweets exhibit marked shifts toward saturated fats and added sugars. Ranking store subcategories by median *F*Score establishes a coherent industrial processing gradient (see Figure S9**c**) ranging from raw produce (*F*Score ≈ 0.24) to industrial chips and snack bars (*F*Score > 0.90). While categories such as potato chips and commercial bakery items display tight within-group uniformity, others – notably breakfast cereals and dairy – exhibit broad compositional heterogeneity. This empirical ordering aligns with US grocery benchmarks [14], with particularly strong concordance among the highest-ranked snack, bakery, sweet, and frozen-dessert categories and broadly consistent ordering of staple, dairy, meat, and produce categories toward the middle and lower ends, confirming the structural validity of the crosswalk.



Supplementary Figure S9 Nutritional composition and industrial processing landscape of the mapped retail catalog. **a)** Principal component analysis (PCA) of FoodProX predicted class probabilities (p_1, p_2, p_3, p_4) across mapped retail products (n = 8,672). Points are colored by assigned FoodProX NOVA class (1: unprocessed/minimally processed to 4: ultra-processed), with marker shapes denoting catalog macro-families (Fresh Food, Snacks & Sweets, Beverages, Pantry & Packaged Food, Frozen Foods). The primary axis separates minimally processed foods (NOVA 1, left, orange) from ultra-processed formulations (NOVA 4, right, red) along a continuous processing gradient. **b)** Empirical distributions of key nutrient densities (dietary fiber, protein, available carbohydrates, sodium, saturated fat, and added sugar in g or mg per 100 g edible portion of the mapped Frida food) across the five catalog macro-families (colors). Fresh foods exhibit high protein and fiber densities, whereas snacks and sweets show pronounced skews toward added sugar and saturated fat. **c)** Store subcategories ($n \geq 50$ mapped products) ranked by median continuous processing score ($FScore \in [0, 1]$). The ranking spans from raw vegetables and dried nuts to industrial bakery goods, cookies, and crisps, with tightly constrained dispersion in confectioneries contrasting with wide heterogeneity in breakfast cereals and dairy.

Supplementary Materials, Section 6: Basket Compositional Context of Co-Purchasing Ties

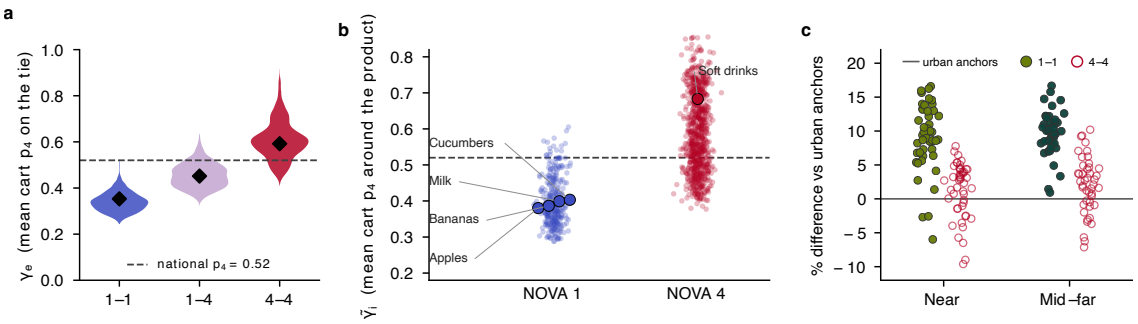
To determine whether the validated co-purchasing backbone is supported by mixed baskets or by mutually exclusive transaction types (such as separate fresh-food versus ultra-processed receipts), we evaluated the compositional context of validated product pairs. All evaluations are conducted on the national Bipartite Configuration Model skeleton used in the main text [2] (see Methods and Supplementary Materials, Section 3). We avoid splitting receipts into boolean subsets (such as checkouts with versus without NOVA-4 items), reconstructing separate null models on sub-graphs, or computing likelihood differences ($\Delta \log L$) on arbitrary transaction subsets, as such stratified constructions conflate sample size N, degree sequences, and basket sizes k, confounding genuine topological association with marginal size distributions.

Let I_r denote the set of mapped brand $\times$ micro-category products in receipt r, $k_r = |I_r|$ the basket size, $n_{4,r}$ the count of ultra-processed items (NOVA 4 via FoodProX [4, 11]), and $p_4(r) = n_{4,r}/k_r$ the local UPF share. For any validated skeleton edge $e = (i, j)$, let $R_e = \{r : i, j \in I_r, k_r \geq 3\}$ denote the set of supporting receipts containing both items with a minimum checkout size of three items. We define the tie carriage γ_e as the expected UPF share among baskets supporting edge e, $\gamma_e = \mathbb{E}\left[p_4(r) \mid r \in R_e\right]$. By construction, γ_e is directly comparable to aggregate purchasing baselines, where the item-weighted national UPF share is $p_4 = 0.52$ and the mean UPF share across mixed baskets is 0.49. We compute national values by pooling municipal receipts weighted by supporting transaction volume $N_{3,e} = |R_e|$ rather than equal municipal weights, retaining edges with $N_{3,e} \geq 50$ and both endpoints FoodProX-mapped to ensure estimation stability (82.7% of total edges).

Across 48.7 million food receipts, 47.9% contain both a NOVA-1 and a NOVA-4 item. These mixed carts account for 76.6% of all mapped retail volume, and 83.3% of receipts containing at least one NOVA-1 item also contain at least one NOVA-4 item. Because the endpoints of edge $e = (i, j)$ enter $p_4(r)$ directly, the class distributions of γ_e are partially separated by arithmetic constraints: $p_4 \geq 0$ for 1–1 pairs, $p_4 \geq 1/k_r$ for 1–4 pairs, and $p_4 \geq 2/k_r$ for 4–4 pairs. The substantive result is that minimally processed pairs are not confined to fresh-food baskets, with $\bar{\gamma}_{11} = 0.353$ (95% bootstrap CI: $[0.342, 0.364]$), a value that sits well within mixed carts rather than isolated trips (see Supplementary Figure S10**a**). In contrast, mixed and ultra-processed pairs yield $\bar{\gamma}_{14} = 0.452$ and $\bar{\gamma}_{44} = 0.592$, showing that UPF–UPF ties concentrate in already UPF-dense baskets.

To test whether tie carriage reflects genuine dyadic structure or simply the marginal composition of the linked products, we computed a leave-two-out contrast Δ_e against a municipality-specific exact-k baseline ($\bar{\Delta}_{11} = -0.123$, $\bar{\Delta}_{44} = +0.117$). Because Δ_e is almost entirely accounted for by an additive combination of marginal product means [$R^2(\Delta_e \sim \frac{u_i + u_j}{2}) = 0.95$, with Spearman rank correlations to mean basket size $\bar{k}$ of 0.00, -0.37, and -0.67 within 1–1, 1–4, and 4–4 pairs], it captures node-level composition rather than residual tie context. Projecting carriage directly to the product level via the incident $N_{3,e}$-weighted mean, $\tilde{\gamma}_i = \sum_{e \in \partial i} N_{3,e}\gamma_e / \sum_{e \in \partial i} N_{3,e}$, shows that core minimally processed staples cluster tightly near $\tilde{\gamma}_i \approx 0.38$–$0.40$, whereas branded carbonated soft drinks reach 0.68 (Supplementary Figure S10**b**). Spatially, municipal mean $\bar{\gamma}$ displays a discrete level shift between the urban core and the rest of the country rather than a continuous gradient (Supplementary Figure S10**c**). Relative to the urban core, municipal $\bar{\gamma}_{11}$ is approximately 9–10%

higher in Near and Mid–far municipalities, whereas $\bar{\gamma}_{44}$ remains within about 2% of that urban baseline. This divergence represents a baseline offset defined by the urban core rather than a continuous gradient along the spectral axis, with neither class showing a significant rank correlation with spectral distance ($\rho = 0.17$ for 1–1; $\rho = 0.02$ for 4–4; $p > 0.05$).



Supplementary Figure S10 Basket compositional context of BiCM-validated co-purchasing ties. a) Edge-level carriage γ_e, defined as the mean basket UPF share p_4 across transactions supporting each edge. Distributions are shown across processing pairs: minimally processed (NOVA 1–1, blue), mixed (NOVA 1–4, purple), and ultra-processed (NOVA 4–4, red). Black diamonds indicate $N_{3,e}$-weighted class means ($\bar{\gamma}_{11} = 0.353$, $\bar{\gamma}_{14} = 0.452$, $\bar{\gamma}_{44} = 0.592$). The horizontal dashed line marks the item-weighted national baseline ($p_4 = 0.52$). **b)** Product-level carriage $\bar{\gamma}_i$, computed as the $N_{3,e}$-weighted mean of incident γ_e values across NOVA-1 and NOVA-4 products. Common Staples (apples, bananas, milk, cucumbers) sit near 0.38–0.40, while ultra-processed formulations (e.g., carbonated soft drinks) reach 0.68. **c)** Percentage difference in municipal mean tie carriage $\bar{\gamma}$ relative to the urban core (horizontal line at zero), grouped into Near and Mid–far municipalities (with filled circles NOVA 1–1 and open circles NOVA 4–4). The peripheral shift reflects a discrete level shift in 1–1 ties (+9 to +10%) rather than a continuous spectral gradient.

References

[1] Saracco, F., Di Clemente, R., Gabrielli, A., Squartini, T.: Randomizing bipartite networks: the case of the World Trade Web. Scientific Reports **5**(1), 10595 (2015) https://doi.org/10.1038/srep10595 . Accessed 2026-08-03

[2] Saracco, F., Straka, M.J., Clemente, R.D., Gabrielli, A., Caldarelli, G., Squartini, T.: Inferring monopartite projections of bipartite networks: an entropy-based approach. New Journal of Physics **19**(5), 053022 (2017) https://doi.org/10.1088/1367-2630/aa6b38 . Accessed 2026-08-03

[3] Cimini, G., Squartini, T., Saracco, F., Garlaschelli, D., Gabrielli, A., Caldarelli, G.: The statistical physics of real-world networks. Nature Reviews Physics **1**(1), 58–71 (2019) https://doi.org/10.1038/s42254-018-0002-6 . Accessed 2026-08-03

[4] Menichetti, G., Ravandi, B., Mozaffarian, D., Barabási, A.-L.: Machine learning prediction of the degree of food processing. Nature Communications **14**(1), 2312 (2023) https://doi.org/10.1038/s41467-023-37457-1 . Accessed 2026-08-03

[5] Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment **2008**(10), 10008

(2008) https://doi.org/10.1088/1742-5468/2008/10/P10008 . Accessed 2026-08-03

[6] Lugannani, R., Rice, S.: Saddle point approximation for the distribution of the sum of independent random variables. Advances in Applied Probability **12**(2), 475–490 (1980) https://doi.org/10.2307/1426607 . Accessed 2026-09-05

[7] Desiderio, A., Galdeman, A., Bäuerlein, F., Lehmann, S.: Mapping regional disparities in discounted grocery products. npj Science of Food **10**(1), 112 (2026) https://doi.org/10.1038/s41538-026-00764-0 . Accessed 2026-09-30

[8] Rosvall, M., Bergstrom, C.T.: Maps of random walks on complex networks reveal community structure. Proceedings of the National Academy of Sciences **105**(4), 1118–1123 (2008) https://doi.org/10.1073/pnas.0706851105 . Accessed 2026-08-03

[9] Lane, M.M., Gamage, E., Du, S., Ashtree, D.N., McGuinness, A.J., et al.: Ultra-processed food exposure and adverse health outcomes: umbrella review of epidemiological meta-analyses. BMJ, 077310 (2024) https://doi.org/10.1136/bmj-2023-077310 . Accessed 2026-08-03

[10] Micha, R., Wallace, S.K., Mozaffarian, D.: Red and Processed Meat Consumption and Risk of Incident Coronary Heart Disease, Stroke, and Diabetes Mellitus: A Systematic Review and Meta-Analysis. Circulation **121**(21), 2271–2283 (2010) https://doi.org/10.1161/CIRCULATIONAHA.109.924977 . Accessed 2026-08-03

[11] Monteiro, C.A., Cannon, G., Levy, R.B., Moubarac, J.-C., Louzada, M.L., Rauber, F., Khandpur, N., Cediel, G., Neri, D., Martinez-Steele, E., Baraldi, L.G., Jaime, P.C.: Ultra-processed foods: what they are and how to identify them. Public Health Nutrition **22**(5), 936–941 (2019) https://doi.org/10.1017/S1368980018003762 . Accessed 2026-08-03

[12] European Food Safety Authority (EFSA): The food classification and description system FoodEx 2 (revision 2). EFSA Supporting Publications **12**(5) (2015) https://doi.org/10.2903/sp.efsa.2015.EN-804 . Accessed 2026-08-31

[13] Poulsen, A., Langwagen, M., Jacobsen, J., Christensen, T.: Dataset from The Danish Food Composition Database, Frida version 5.5. Technical University of Denmark. Artwork Size: 28769892 Bytes Pages: 28769892 Bytes (2026). https://doi.org/10.11583/DTU.29500682 . https://data.dtu.dk/articles/dataset/Dataset_from_The_Danish_Food_Composition_Database_Frida_version_4_1/29500682 Accessed 2026-08-29

[14] Ravandi, B., Ispirova, G., Sebek, M., Mehler, P., Barabási, A.-L., Menichetti, G.: Prevalence of processed foods in major US grocery stores. Nature Food **6**(3), 296–308 (2025) https://doi.org/10.1038/s43016-024-01095-7 . Accessed 2026-08-03