TasteRank Explorer

Statistical Methods Primer

A Guide to the Mathematical Foundations

Jure Skarabot
April 2026

Introduction

The TasteRank Explorer uses a sequence of mathematical techniques from linear algebra, graph theory, and network science to convert wine tasting profiles into a similarity network, rank grape varieties by structural importance, and detect natural communities. This primer explains each method in the order it appears in the analytical pipeline: cosine similarity to measure resemblance between varieties, k-nearest-neighbor graph construction to build a network, eigenvector centrality to rank varieties within that network, PageRank as a robustness check, and modularity optimization to detect communities. The treatment assumes comfort with basic linear algebra (matrix multiplication, eigenvalues) but not prior exposure to network science.

Each section explains the intuition, states the formal definition, describes the algorithm used to compute it, and notes what the method reveals in the specific context of the TasteRank graph. For the full mathematical treatment with proofs and derivations, see the Technical Appendix.

1. Cosine Similarity

What it does.

Cosine similarity measures how alike two grape varieties are based on the shape of their sensory profiles, independent of overall magnitude. Each variety is represented as a 13-dimensional vector (one dimension per WSET SAT tasting attribute: color depth, aromatic intensity, floral character, and so on). Cosine similarity computes the cosine of the angle between two such vectors.

Formal definition.

Given two profile vectors p and q in $\blacksquare^{13}$, the cosine similarity is the dot product of the vectors divided by the product of their magnitudes: cos(θ) = (p · q) / (■p■ · ■q■). Because all SAT scores are non-negative integers (0–5), the result is always between 0 and 1. A score of 1 means the profiles are identical in shape (though they could differ in scale); 0 would mean the profiles share no common structure at all.

Why cosine, not Euclidean distance?

Euclidean distance measures the absolute gap between vectors, so a variety with uniformly high scores would appear “far” from one with uniformly moderate scores even if their sensory shape is identical. Cosine similarity ignores this scale difference and focuses on the relative pattern of highs and lows across dimensions. This is more appropriate for tasting profiles, where the question is whether two varieties emphasize the same sensory attributes, not whether they score at the same absolute level.

A practical subtlety.

The tannin dimension is structurally zero for most white grapes, creating a systematic difference between reds and whites. Cosine similarity handles this more gracefully than Euclidean distance: for whites, tannin contributes nothing to either the dot product or the norms, so it is effectively ignored rather than penalized. This is one reason the TasteRank network correctly identifies cross-boundary communities (C1: Light & Aromatic) that include both reds and whites with similar non-tannin profiles.

In the TasteRank graph

The full 101 × 101 cosine similarity matrix has 5,050 unique pairwise values. The mean similarity across all connected pairs is 0.983, with a narrow range of [0.909, 1.000]. This tight distribution reflects the fact that wine grape profiles share a common baseline shape—the differentiation between varieties lies in subtle but structurally meaningful differences across specific dimensions.

2. k-Nearest-Neighbor Graph Construction

What it does.

The similarity matrix is dense—every variety has a non-zero similarity to every other. A k-nearest-neighbor (kNN) graph sparsifies this by keeping only the strongest relationships: each variety is connected by an edge to its k most similar neighbors. The result is a graph where nodes are grape varieties and weighted edges represent strong sensory resemblance.

How it works.

For each variety, the algorithm sorts all other varieties by descending cosine similarity and retains the top k = 5. This produces a directed graph (variety A may include B in its top 5, but B may not include A). The graph is then symmetrized: an undirected edge is placed between A and B if either A lists B or B lists A among their five nearest neighbors. The edge weight is the cosine similarity between the pair.

Why k = 5?

The choice of k balances two competing goals. A small k produces a sparse graph that highlights only the very strongest similarities, at the risk of fragmenting the network into disconnected components. A large k produces a dense graph where every variety is connected to many others, drowning out the most meaningful relationships. At k = 5, the TasteRank graph is connected (no isolated nodes), sparse enough to reveal meaningful structure, and produces a clear community partition.

Symmetrization and degree inflation.

Because the graph is symmetrized by union (not intersection), the effective degree of each node is typically higher than k. If variety A includes B in its top 5 and B also includes A, one edge results. But if A includes B and B does not include A, the edge is still created. The average degree in the TasteRank graph is 6.75, reflecting this inflation above the nominal k = 5.

In the TasteRank graph

The kNN construction produces 341 edges among 101 nodes, with a graph density of 0.068 (6.8% of all possible edges). This sparse, weighted graph is the input to all subsequent analyses—centrality, PageRank, and community detection all operate on this structure.

3. Eigenvector Centrality (TasteRank)

What it does.

Eigenvector centrality assigns a score to each node in a network that reflects not just how many connections it has, but how important its connections are. A variety earns a high TasteRank score by being similar to other varieties that are themselves highly central—a recursive definition that rewards membership in densely connected, mutually reinforcing clusters.

3.1 The eigenvalue problem

The weighted adjacency matrix W is a 101 × 101 symmetric matrix where entry W_{ij} is the cosine similarity between varieties i and j if they share an edge, and zero otherwise. Being real and symmetric, W has 101 real eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{101}$ and corresponding orthogonal eigenvectors $x_1, x_2, \ldots, x_{101}$, satisfying $W \cdot x_k = \lambda_k \cdot x_k$ for each k.

Eigenvector centrality defines the importance of each variety as its component in the leading eigenvector x_1 (the eigenvector associated with the largest eigenvalue λ_1). The TasteRank score of variety i is simply $x_{1,i}$, normalized so the eigenvector has unit length.

3.2 The recursive interpretation

Written component-wise, the eigenvector equation says that $x_i = (1/\lambda_1) \cdot \Sigma_j W_{ij} \cdot x_j$. In words: each variety's centrality is proportional to the weighted sum of its neighbors' centralities, which are defined by their neighbors' centralities, and so on to infinite depth. This circularity is not a defect—it is the defining feature. The eigenvector is the unique self-consistent solution to this infinite chain of mutual dependencies.

This is directly analogous to Google's PageRank: there is no closed-form formula for any individual page's rank. The rank is a global property of the entire network topology, and changing a single edge anywhere in the graph can in principle alter every node's score.

3.3 Why the largest eigenvalue dominates

Consider starting with any initial vector of scores z^0 and repeatedly multiplying by W. After t iterations, $z^t = W^t \cdot z^0$. Decomposing z^0 in the eigenbasis, the contribution of each eigenvector is scaled by $\lambda_k{}^t$. Since λ_1 is strictly the largest eigenvalue (guaranteed by the Perron–Frobenius theorem for connected, non-negative matrices), the ratio $|\lambda_2/\lambda_1|^t$ converges to zero exponentially. After enough iterations, only the leading eigenvector survives. This is why λ_1 alone determines the centrality ranking.

3.4 The Perron–Frobenius theorem

The theoretical guarantee behind eigenvector centrality is the Perron–Frobenius theorem, which states that for a connected graph with non-negative edge weights: (1) the largest eigenvalue λ_1 is real, positive, and strictly greater than the magnitude of all other eigenvalues; (2) the corresponding eigenvector x_1 has all positive components; and (3) x_1 is unique up to normalization. Condition (2) is essential—it ensures that centrality scores are all positive and thus interpretable as a ranking. Condition (3) ensures that the ranking is unique and does not depend on the choice of algorithm or starting point.

3.5 The spectral gap

The spectral gap $|\lambda_2/\lambda_1|$ controls both the convergence speed of the computation and the structural interpretation. A small ratio means fast convergence and a clear, dominant centrality structure. A ratio near 1 would indicate competing centers of influence with no clear hierarchy. For the TasteRank graph, the spectral gap is approximately 0.72, meaning each iteration of power iteration reduces the second-order contamination by ~28%. After 25 iterations, the residual error is below 10^{-3}; after 50, below 10^{-7}.

3.6 Power iteration

The standard algorithm for computing the leading eigenvector is power iteration. Starting from a uniform vector, the algorithm repeatedly multiplies by W and normalizes: $x^{t+1} = W \cdot x^t / \blacksquare W \cdot x^t \blacksquare$. This converges to x_1 because the repeated multiplication amplifies the leading eigenvector component and suppresses all others (as described above). The eigenvalue is recovered as the Rayleigh quotient: $\lambda_1 = x^T W x$. In the TasteRank graph, convergence to tolerance 10^{-6} is achieved in 35–45 iterations.

In the TasteRank graph

Sagrantino ranks #1 (TasteRank = 0.3020) because it is connected to Nero d'Avola, Lagrein, Montepulciano, Petite Sirah, and Petit Verdot—all top-10 varieties themselves embedded in the dense Mediterranean red cluster (C0). The recursive amplification through this mutually reinforcing core pushes it to the top. Conversely, Riesling ranks #77 because its neighbors (Albariño, Furmint, Sauvignon Blanc) are themselves peripheral, and the recursive logic works in reverse.

4. PageRank

What it does.

PageRank, developed by Brin and Page (1998) for ranking web pages, is a modified form of eigenvector centrality that includes a “teleportation” mechanism to prevent excessive concentration of importance within a single dense cluster. In TasteRank, it serves as a robustness check on the eigenvector centrality ranking.

4.1 The random surfer model

Imagine a “random taster” navigating the grape similarity network. At each step, the taster either (a) with probability α, follows an edge to a similar variety, choosing among neighbors proportionally to cosine similarity weight; or (b) with probability $1 – \alpha$, teleports to a completely random variety. The PageRank of each variety is the long-run fraction of time the random taster spends there.

Formally, the transition is governed by the matrix $G = \alpha \cdot M + (1 – \alpha) \cdot (1/n) \cdot J$, where M is the row-normalized adjacency matrix (each row sums to 1) and J is the all-ones matrix. The PageRank vector π is the stationary distribution: $G^T \cdot \pi = \pi$. With the standard damping factor $\alpha = 0.85$, the random taster follows edges 85% of the time and teleports 15% of the time.

4.2 How it differs from eigenvector centrality

The teleportation mechanism ensures that every variety receives a baseline importance floor of $(1 - \alpha)/n$, regardless of its network position. This prevents the extreme concentration that can occur with pure eigenvector centrality in networks with dense subgraphs. PageRank also rewards bridge varieties—nodes connecting multiple communities—because the random walk must pass through them when transitioning between clusters.

The two measures agree strongly overall (Spearman rank correlation $\rho \approx 0.92$) but diverge in informative ways. Varieties deep inside a dense cluster have higher eigenvector centrality (recursive amplification) but slightly lower PageRank (teleportation redistributes some importance outward). Bridge varieties show the opposite pattern. The most striking example is Sangiovese: TasteRank rank 40 vs. PageRank rank ~8, because it bridges the Mediterranean red cluster (C0) and mid-weight structured reds (C2).

Interpreting divergences

> *When PageRank >> TasteRank for a variety, it is acting as a structural bridge between communities. When TasteRank >> PageRank, the variety is embedded deep in a dense cluster core. The gap between the two measures is itself a diagnostic for network position.*

5. Modularity and Community Detection

What it does.

Community detection partitions the graph into groups where within-group connectivity is stronger than would be expected by chance. The result is a set of clusters that reflect natural groupings in the data—in this case, families of grape varieties with mutually similar sensory profiles.

5.1 The modularity function

The modularity function Q measures the quality of a partition. For each pair of nodes in the same community, it compares the actual edge weight to the expected weight under a null model (the configuration model, which randomly rewires edges while preserving each node's total connectivity). The formula sums this excess across all same-community pairs and normalizes by total edge weight.

Formally: $Q = (1/2m) \cdot \Sigma_{ij} [W_{ij} - (s_i \cdot s_j / 2m)] \cdot \delta(c_i, c_j)$, where m is total edge weight, s_i is the weighted degree of node i, c_i is its community assignment, and δ is the Kronecker delta (1 if same community, 0 otherwise). A partition with $Q > 0.3$ is generally considered to have significant community structure.

5.2 The null model

The null model deserves emphasis because it defines what “expected by chance” means. The configuration model preserves the degree sequence—each node retains its total edge weight—but randomizes which specific nodes are connected. Under this model, the expected edge weight between nodes i and j is $s_i \cdot s_j / 2m$. Nodes with high total connectivity are expected to share more edge weight simply because they are well-connected, not because they are in the same community. Modularity measures the excess above this baseline, isolating the signal from the noise.

5.3 The greedy algorithm

The Clauset–Newman–Moore algorithm is an agglomerative method that starts with every node in its own singleton community (101 communities of one) and iteratively merges the pair of communities whose merger produces the largest increase in Q. The sequence of merges defines a dendrogram (a hierarchical tree), and the partition with the highest Q across the entire merge history is selected as the final community structure.

This greedy approach does not guarantee the globally optimal partition (modularity optimization is NP-hard), but it is fast and produces results that are competitive with more expensive methods. For the TasteRank graph, the algorithm was validated against the Louvain method and spectral bisection, all three producing partitions with similar membership and comparable modularity scores.

5.4 Connection to the eigenvalue spectrum

There is a deep connection between community structure and the eigenvalues of the modularity matrix B (where $B_{ij} = W_{ij} - s_i s_j/2m$). The number of positive eigenvalues of B provides an upper bound on the number of meaningful communities. For the TasteRank graph, the first six eigenvalues of B are positive and clearly separated from the bulk spectrum, consistent with the six-community partition. Spectral methods can also be used directly for community detection by partitioning nodes based on the signs of eigenvector components—the second eigenvector of the adjacency matrix W approximately separates reds from whites, reflecting the strongest structural divide in the network.

In the TasteRank graph

> *Six communities are detected with modularity Q ≈ 0.41, indicating strong structure. The partition tracks sensory logic rather than geography: Nerello Mascalese clusters with Pinot Noir in C2 (mid-weight structured reds), not with its Sicilian neighbor Nero d'Avola in C0 (Mediterranean reds). Riesling and Assyrtiko join the cross-boundary C1 cluster rather than the mineral-white C4, because their extreme acidity and complexity scores align them with light aromatic varieties.*

6. The Pipeline in Summary

The five methods form a coherent analytical pipeline, each building on the output of the previous step:

1. Cosine similarity converts 101 sensory profile vectors into a 101 × 101 similarity matrix, measuring the shape-based resemblance between every pair of varieties.

2. kNN graph construction sparsifies the dense similarity matrix into a weighted network of 341 edges, retaining only the strongest relationships (k = 5 per variety, symmetrized).

3. Eigenvector centrality ranks varieties by their recursive structural importance in this network—varieties connected to other highly central varieties score highest. This is the TasteRank score.

4. PageRank provides a robustness check with a damping mechanism that prevents excessive concentration. Divergences between the two rankings identify bridge varieties and cluster cores.

5. Modularity optimization partitions the network into six communities of varieties with mutually similar profiles, revealing natural grape families that track sensory logic rather than geography.

The mathematical thread connecting these steps is the weighted adjacency matrix W, which encodes the network's structure. Eigenvector centrality uses the leading eigenvector of W; PageRank uses a damped, row-normalized version of W; and modularity detection uses a null-model–corrected version of W. The same matrix, analyzed from three different perspectives, yields the TasteRank ranking, the robustness diagnostic, and the community partition.

References

Bonacich, P. (1987). Power and centrality: A family of measures. *American Journal of Sociology*, 92(5), 1170–1182.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1–7), 107–117.

Clauset, A., Newman, M. E. J., & Moore, C. (2004). Finding community structure in very large networks. *Physical Review E*, 70(6), 066111.

Fruchterman, T. M. J., & Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software: Practice and Experience*, 21(11), 1129–1164.

Meyer, C. D. (2000). *Matrix Analysis and Applied Linear Algebra*. SIAM. [Perron–Frobenius theorem, spectral theory.]

Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23), 8577–8582.

Newman, M. E. J. (2010). *Networks: An Introduction*. Oxford University Press. [Comprehensive reference for eigenvector centrality, modularity, and spectral methods.]