

CRAFT: Corpus Relatedness Analysis Using Fourier Tranforms

Abstract

A fundamental challenge in data management is the efficient discovery of term relationships from massive, unstructured text corpora, a critical first step in knowledge graph construction. This discovery task, however, faces prohibitive computational barriers: the quadratic $O(N^2)$ complexity of an all-pairs analysis and the intractability of processing the full term-document matrix. While dimensionality reduction via embeddings offers a partial solution, the resulting vector proximity often captures broad thematic similarity, failing to isolate the precise co-occurrence signals required for high-quality relation extraction.

This paper introduces **CRAFT**, a system that overcomes these limitations by re-casting term relatedness discovery as a scalable signal processing problem. CRAFT's methodology decouples the discovery process from both the term-document matrix and quadratic-time comparisons. First, it employs a **randomized Fourier transform** to sketch term occurrence signals directly into a low dimensional complex space, a process that provably preserves the inner products essential for correlation analysis without materializing the underlying matrix. Second, to break the quadratic barrier, CRAFT leverages the inherent sparsity of term relationships by formulating discovery as a **compressed sensing task**. This enables the recovery of significant correlations for any given term directly from its compressed sketch via an efficient Orthogonal Matching Pursuit algorithm, obviating the need for an all-pairs comparison. Our end-to-end implementation and comprehensive experimental evaluation show that CRAFT significantly outperforms modern baselines in both efficiency and precision, enabling high-quality relation discovery at a previously infeasible scale.

ACM Reference Format:

. 2025. CRAFT: Corpus Relatedness Analysis Using Fourier Tranforms. In . ACM, New York, NY, USA, 16 pages. <https://doi.org/10.1145/nnnnnnnn>

1 Introduction

Identifying frequently co-occurring terms in massive, unstructured text corpora is fundamental to modern data management and AI. When terms consistently appear together—what we call relatedness—they exhibit strong semantic or functional associations arising from shared topics ((e.g., “iPhone” and “Apple”), common contexts (e.g., “aspirin” and “blood clot” in medicine), or compound concepts (e.g., “machine learning”). This capability is critical for applications ranging from query expansion and trend analysis to knowledge discovery in scientific literature [10, 17, 45]. A primary

application is constructing Knowledge Graphs (KGs) in novel domains [2, 4, 22]. KGs represent information as networks of entities and relationships, powering semantic search and question-answering systems. They are increasingly vital for grounding Large Language Models (LLMs) [34, 38, 39], serving as verifiable memory sources in Retrieval-Augmented Generation (RAG) [29] frameworks to reduce hallucinations and inject domain-specific facts. However, their utility is constrained by a fundamental computational bottleneck: bootstrapping the graph from raw, unstructured text. This bottleneck centers on the candidate discovery problem: scalably identifying potentially related terms before fine-grained analysis. Current academic benchmarks [7, 37] largely sidestep this challenge, focusing instead on relation classification—assigning semantic labels to pre-identified entity pairs [28, 36]. This assumes a domain-adapted Entity Recognition tool has already identified all relevant entities, effectively bypassing the core challenge: one cannot leverage a domain-specific KG without first constructing it from source text. In real-world scenarios with novel, domain-specific corpora (scientific literature, financial reports, enterprise documents), this assumption fails. The task becomes a massive-scale filtering problem: evaluating the combinatorially vast space of potential term pairs—scaling quadratically $O(N^2)$ with vocabulary size N—to isolate the tiny fraction that are meaningfully related.

Existing methods make critical trade-offs. Brute-force pairwise analysis (e.g., Pointwise Mutual Information [11], co-occurrence graphs) generates massive noise and impractically dense outputs. Embedding-based methods [35, 40, 47] capture broad similarity rather than specific relational types, washing out faint signals. LLMs excel at focused reasoning but cannot perform exhaustive corpus-wide discovery and are prone to hallucinating unsupported relationships [23]. This gap demands a scalable “first-pass” filter for candidate discovery—one that complements LLMs in a hybrid approach: lightweight discovery at scale, followed by economical LLM-based verification using corpus evidence. We introduce **CRAFT** (Corpus Relatedness Analysis Using Fourier Transforms), a novel paradigm addressing this exact need, as shown in Figure 1. Rather than operating on sparse, high-dimensional term-document matrices, CRAFT treats term presence as signals and processes them in compact frequency representations to identify persistent co-occurrences with remarkable efficiency, formally relating co-occurrence to statistical correlations (Pearson’s coefficient). This produces a clean, high-precision candidate set ideal for KG bootstrapping, search enrichment, or LLM validation. Our key insight: the sparsity-of-effects principle—most terms interact significantly with only a small number of others.

The primary contributions of our work are as follows:

- A Novel Signal Processing Paradigm for Text Analysis: We introduce CRAFT, a framework that reinterprets the problem of term relatedness discovery in large text corpora as a signal processing task. By treating the presence of terms across documents as discrete signals and analyzing their frequency-domain representations, CRAFT enables highly efficient and

Permission to make digital or hard copies of all or part of this work for personal or
Unpublished working draft. Not for distribution.
distributed for profit or commercial advantage and that copies bear this notice and the full citation
on the first page. Copyrights for components of this work owned by others than the
author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or
republish, to post on servers or to redistribute to lists, requires prior specific permission
and/or a fee. Request permissions from permissions@acm.org.

Conference’17, Washington, DC, USA

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-x-xxxx-xxxx-x/YY/YY/MM
<https://doi.org/10.1145/nnnnnnnn.nnnnnnnn>

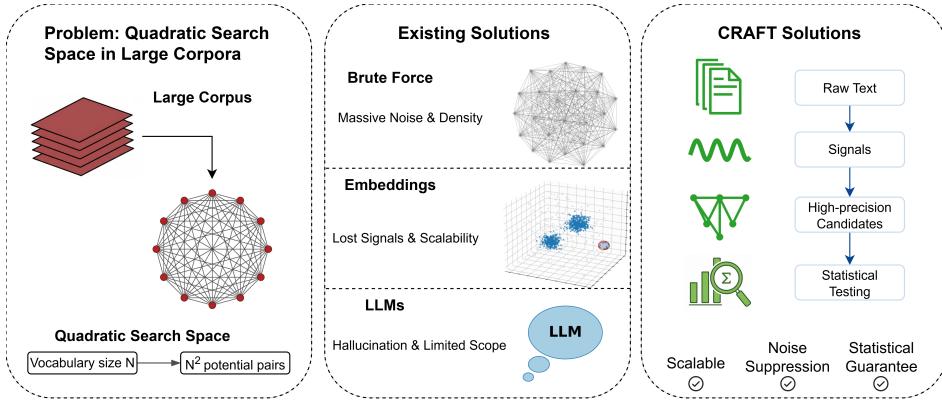
117
118
119
120
121
122
123
124
125
126
127
128
129
130

Figure 1: The candidate discovery problem and CRAFT’s paradigm shift. Given a corpus with vocabulary size N , identifying related terms requires evaluating $O(N^2)$ potential pairs. Existing methods struggle: brute-force approaches generate prohibitive noise and scale poorly; embedding methods capture similarity but miss specific relational signals; LLMs cannot exhaustively search at scale. CRAFT reframes discovery as signal processing, analyzing term co-occurrence patterns in the frequency domain to scalably isolate meaningful relationships with high precision.

scalable detection of semantic relationships without relying on pre-existing knowledge bases or expensive pairwise computations.

- Randomized Fourier Embedding with Theoretical Guarantees: We propose a Random Fourier Transform (RFT)-based dimensionality reduction technique that projects high dimensional term vectors into a low dimensional complex space. This method preserves inner products with high probability, reducing storage and computation from $O(N^2 M)$ to $O(Nk \log M)$ with $k \leq N$ (for N terms and M documents).
- Cross-Power Spectral Density (CPSD) for Correlation Estimation: We adapt Cross-Power Spectral Density (CPSD)—a classical signal processing tool—to estimate term statistical correlations in the frequency domain. CPSD not only captures the magnitude of correlation but also leverages phase information to distinguish between positive, negative, and orthogonal relationships, providing a richer and more interpretable measure of term association.
- Sparse Recovery via Compressed Sensing: We reformulate correlation recovery as a compressed sensing problem, leveraging the insight that most term relationships are sparse. We employ **Orthogonal Matching Pursuit (OMP)** [46], a computationally efficient algorithm, to recover sparse correlation vectors directly from their compressed measurements. This approach enables the identification of significant term pairs while completely avoiding the prohibitive cost of computing all $O(N^2)$ correlations.
- End-to-End Scalable System with Statistical Rigor: The full CRAFT pipeline integrates preprocessing, spectral sketching, CPSD-based correlation analysis, and sparse recovery into a cohesive system. We provide non-asymptotic error bounds, statistical significance testing with FDR control, and complexity guarantees, making the approach both practical and theoretically sound for real-world corpus-scale analysis.

- Empirical and Theoretical Superiority: We demonstrate through complexity analysis and empirical validation that CRAFT significantly outperforms existing methods in both speed and precision, enabling the discovery of high-quality relational candidates at a scale previously infeasible with state-of-the-art techniques.

The remainder of this paper is organized as follows. We begin by formalizing the term relatedness discovery problem in Section 2. We then present the core methodology of the **CRAFT** framework, detailing our randomized embedding strategy in Section 3 and our compressed sensing approach to correlation recovery in Section 4. Section 5 integrates these components into a cohesive, end-to-end system architecture. We empirically validate **CRAFT**’s performance through a comprehensive experimental study in Section 6, situate our contributions with respect to prior research in Section 7, and offer concluding remarks in Section 8.

2 Problem Formulation

Our goal is to automatically discover meaningful associations between terms by inductively analyzing a massive collection of unstructured text, such as a decade of biomedical preprints from arXiv’s q-bio section or a proprietary repository of clinical trial documents from a pharmaceutical company. Unlike methods that rely on pre-existing ontologies, this approach processes the entire corpus—examining relatedness (co-occurrence) patterns across term pairs—to uncover significant statistical relationships without prior guidance on what to search for.

The aim is to generate a ranked list of strongly associated term pairs, such as (CRISPR, Cas9, 0.98) or (ibuprofen, cyclooxygenase, 0.92), where the numerical score reflects the strength of the relationship, the Pearson correlation coefficient. This method is especially powerful for revealing non-obvious yet plausible connections—for instance, linking metabolic pathways like *succinate* signaling to immune receptors such as *GPR91*, or associating *autophagy*-related

174

175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231

genes like *ULK1* with specific cellular processes—effectively building the foundations of a novel knowledge graph directly from raw text. More formally our problem is defined as follows:

- **Input:** A corpus of documents $\mathcal{D} = \{d_1, d_2, \dots, d_M\}$ and an associated vocabulary of terms $\mathcal{V} = \{w_1, w_2, \dots, w_N\}$, where N is the vocabulary size. V is typically derived by tokenizing C .
- **Output:** A ranked list of term pairs $\ell = [(v_i, v_j, s_{ij}), \dots]$, where (v_i, v_j) is a pair of distinct terms from V and $s_{ij} \in \mathbb{R}$ is a score representing the strength of their relationship expressed in terms of the Pearson correlation coefficient. The list is ordered by s_{ij} in descending order.

The objective is to generate this list ℓ in a computationally efficient and scalable manner. The methodology for evaluating the overall quality of the generated list ℓ , will be detailed in the experiments section.

3 Large-Scale Term Correlation Detection via Randomized Nonlinear Embedding and Spectral Analysis

This section details our methodology for efficiently detecting statistically significant correlations between all pairs of terms in a large corpus. We first formalize the problem using a term-document matrix, then describe a randomized dimensionality reduction technique with theoretical guarantees, and finally present our algorithms for correlation estimation, sparse recovery, and significance testing. All proofs of theorems and lemmas are omitted due to space constraints and are available in our technical report [1].

3.1 Term-Document Matrix Construction and Statistical Formulation

We represent the corpus as a term-document matrix $\mathbf{A} \in \mathbb{R}^{N \times M}$, where each element a_{ij} quantifies the importance of term w_i in document d_j . We employ the BM25 weighting function[43]¹, a state-of-the-art ranking function robust to document length variation:

$$a_{ij} = \frac{\text{tf}(w_i, d_j) \cdot (k_1 + 1)}{\text{tf}(w_i, d_j) + k_1 \cdot \left(1 - b + b \cdot \frac{|d_j|}{\text{avgdl}}\right)} \cdot \log \left(\frac{M + 1}{\text{df}(w_i) + 0.5} \right) \quad (1)$$

where $\text{tf}(w_i, d_j)$ is the term frequency of w_i in d_j , $|d_j|$ is the length of document d_j , avgdl is the average document length in \mathcal{D} , $\text{df}(w_i)$ is the document frequency of w_i , and $k_1 = 1.2$, $b = 0.75$ are standard tuning parameters.

To analyze correlation, we center each row of \mathbf{A} to obtain a matrix of deviation vectors, $\tilde{\mathbf{A}}$. The centered value \tilde{a}_{ij} for term w_i in document d_j is given by $\tilde{a}_{ij} = a_{ij} - \mu_i$, where $\mu_i = \frac{1}{M} \sum_{m=1}^M a_{im}$ is the mean score for term w_i . The Pearson correlation coefficient ρ_{ij} between terms w_i and w_j is then defined as the cosine similarity between their centered vectors $\tilde{\mathbf{a}}_i$ and $\tilde{\mathbf{a}}_j$:

$$\rho_{ij} = \frac{\langle \tilde{\mathbf{a}}_i, \tilde{\mathbf{a}}_j \rangle}{\|\tilde{\mathbf{a}}_i\|_2 \|\tilde{\mathbf{a}}_j\|_2} = \frac{\sum_{m=1}^M \tilde{a}_{im} \tilde{a}_{jm}}{\sqrt{\sum_{m=1}^M \tilde{a}_{im}^2} \sqrt{\sum_{m=1}^M \tilde{a}_{jm}^2}} \quad (2)$$

¹Any weighting function from the literature will work.

The direct computation of the full $N \times N$ correlation matrix requires $O(N^2 M)$ operations and $O(NM)$ storage, which is computationally infeasible for large N and M .

3.2 Randomized Fourier Embedding with Energy Preservation Guarantees

To overcome this limitation, we employ a randomized nonlinear embedding based on the method of Random Fourier Features (RFF) [41]. We project the high-dimensional, centered term vectors $\tilde{\mathbf{a}}_i$ into a significantly lower-dimensional complex space \mathbb{C}^k while *provably preserving* the inner product information between vectors.

Formally, for each centered term vector $\tilde{\mathbf{a}}_i \in \mathbb{R}^M$, we compute its low-dimensional complex representation $\mathbf{b}_i \in \mathbb{C}^k$ via the following linear transformation:

$$\mathbf{b}_i = \frac{1}{\sqrt{k}} \Phi \tilde{\mathbf{a}}_i \quad (3)$$

The crucial element of this embedding is the *random Fourier matrix* $\Phi \in \mathbb{C}^{K \times M}$. The entries of Φ are not learned from data but are generated stochastically. Each element is defined by:

$$\phi_{lm} = e^{-2\pi i \xi_l m / M} \quad \text{for } l = 1, \dots, k \quad \text{and } m = 1, \dots, M, \quad (4)$$

where each frequency parameter ξ_l is sampled independently and identically from a Uniform(0, 1) distribution.

Theoretical Underpinnings and Guarantees. This approach is grounded in Bochner's Theorem [31], which states that any continuous, shift-invariant kernel function (e.g., Gaussian Kernel) is the Fourier transform of a unique non-negative measure. The inner product $\mathbf{b}_i^* \mathbf{b}_j$ in the embedded space (where \mathbf{b}_i^* denotes conjugate transpose) is an unbiased estimator of a shift-invariant kernel evaluation $K(\tilde{\mathbf{a}}_i, \tilde{\mathbf{a}}_j)$ in the original space. For our specific construction, this kernel is the linear kernel, $K(\mathbf{x}, \mathbf{y}) = \mathbf{x}^\top \mathbf{y}$.

The following theorem establishes the key properties of the estimator formed by these embeddings.

THEOREM 3.1 (UNBIASED INNER PRODUCT ESTIMATOR). *Let \mathbf{b}_i and \mathbf{b}_j be the randomized Fourier embeddings of $\tilde{\mathbf{a}}_i$ and $\tilde{\mathbf{a}}_j$, respectively, as defined by Equation 3. Then, the complex inner product $\langle \mathbf{b}_i, \mathbf{b}_j \rangle = \mathbf{b}_i^* \mathbf{b}_j$ is an unbiased estimator of the true inner product:*

$$\mathbb{E}_\Phi [\mathbf{b}_i^* \mathbf{b}_j] = \tilde{\mathbf{a}}_i^\top \tilde{\mathbf{a}}_j. \quad (5)$$

Furthermore, the variance of the estimator decays linearly with the embedding dimension k :

$$\text{Var}(\mathbf{b}_i^* \mathbf{b}_j) = O(1/k). \quad (6)$$

This theorem provides the foundational pillars for our method's efficiency: unbiasedness, concentration, and complexity reduction. The **unbiasedness** of the approximated inner product ensures that its expectation is exact, which guarantees that downstream analysis—such as the eigen-decomposition of the approximated covariance matrix $\mathbf{B}^* \mathbf{B}$ —converges to the correct result in expectation, thereby delivering the *energy preservation guarantee*. Furthermore, the **concentration** property ensures that the variance of the estimator decreases linearly with the embedding dimension k ; through Johnson-Lindenstrauss-type arguments [26] and concentration inequalities, the approximation error for all pairwise inner products is tightly bounded with high probability even for a modest embedding

dimension scaling as $k \sim O(\log M)$. Finally, the drastic **complexity reduction** transforms the computational burden: instead of constructing an $N \times N$ matrix at a prohibitive cost of $O(N^2M)$, we now build a $N \times k$ matrix \mathbf{B} (where $k \ll M, N$).

This randomized embedding transforms an intractable quadratic problem into a manageable linear one, enabling the analysis of massively large datasets while providing strong theoretical guarantees on the preservation of the data's geometric structure.

A consequence of this random projection is that the Johnson-Lindenstrauss lemma holds. For a target dimension $k = O(\epsilon^{-2} \log M)$, all pairwise inner products are preserved within an ϵ -factor with high probability:

$$\mathbb{P}(|\langle \mathbf{b}_i, \mathbf{b}_j \rangle - \langle \tilde{\mathbf{a}}_i, \tilde{\mathbf{a}}_j \rangle| \geq \epsilon \|\tilde{\mathbf{a}}_i\| \|\tilde{\mathbf{a}}_j\|) \leq 2e^{-k\epsilon^2/4} \quad (7)$$

Efficient Computation using the FFT.: The operation $\tilde{\mathbf{a}}_i \Phi^\top$ is equivalent to taking a random Fourier transform of the signal $\tilde{\mathbf{a}}_i$. For each term vector $\tilde{\mathbf{a}}_i$ (a sparse vector of length M), the computational cost is $O(k \log M)$ using the Non-Uniform FFT (NUFFT) [20] or similar algorithms, where the $\log M$ factor arises from interpolating the sparse signal onto a regular grid for an FFT. The total cost for all N terms is therefore $O(Nk \log M)$.

Storage: The embedding matrix $\mathbf{B} \in \mathbb{C}^{k \times N}$ requires $O(kN)$ storage. This is a massive reduction from $O(NM)$ because $k \ll M$ (e.g., $k \sim 1000-10000$, while N can be in the billions).

3.3 Cross-Power Spectral Density Estimation with Phase Analysis

Having projected the high-dimensional, centered term vectors $\tilde{\mathbf{a}}_i$ into the low-dimensional complex space via the randomized Fourier embedding $\mathbf{b}_i = \frac{1}{\sqrt{k}} \Phi \tilde{\mathbf{a}}_i$, we now construct a similarity matrix in this embedded space. We define the *Cross-Power Spectral Density (CPSD)* matrix $\mathbf{P} \in \mathbb{C}^{N \times N}$ for the embedded terms². The entries of \mathbf{P} are given by the complex inner products of the embeddings:

$$P_{ij} = \langle \mathbf{b}_i, \mathbf{b}_j \rangle = \mathbf{b}_i^* \mathbf{b}_j \quad (8)$$

This matrix is Hermitian ($\mathbf{P} = \mathbf{P}^*$) and serves as a statistically well-founded proxy for the matrix $\mathbf{G} = \tilde{\mathbf{A}} \tilde{\mathbf{A}}^\top$.

Theoretical Justification: From CPSD to Correlation. The following theorem formalizes the connection between the CPSD matrix and the desired correlation coefficients, providing the core justification for our approach.

THEOREM 3.2 (CPSD CORRELATION ESTIMATOR). *Let \mathbf{b}_i and \mathbf{b}_j be the randomized Fourier embeddings of the centered term vectors $\tilde{\mathbf{a}}_i$ and $\tilde{\mathbf{a}}_j$, respectively. Let $P_{ij} = \mathbf{b}_i^* \mathbf{b}_j$ be their Cross-Power Spectral Density. Then, the estimator:*

$$\hat{\rho}_{ij} = \frac{\Re(P_{ij})}{\sqrt{P_{ii}} \sqrt{P_{jj}}} \quad (9)$$

is a consistent estimator for the true Pearson correlation coefficient ρ_{ij} between terms i and j . That is, $\hat{\rho}_{ij} \xrightarrow{P} \rho_{ij}$ as the embedding dimension $k \rightarrow \infty$. $\Re(P_{ij})$ takes the real part of the complex-valued inner product.

²This matrix will not be computed nor materialized in the sequel.

The Informative Role of Phase. A significant advantage of operating in the complex plane is the rich information encoded in the *phase* of the CPSD entries. For a complex value $P_{ij} = |P_{ij}| e^{i\theta_{ij}}$, the phase $\theta_{ij} = \arg(P_{ij})$ provides immediate directional insight into the relationship between terms i and j :

- $\theta_{ij} \approx 0$: The complex vectors \mathbf{b}_i and \mathbf{b}_j are nearly in phase. This indicates a *positive correlation* between the original terms; their embeddings point in roughly the same direction in \mathbb{C}^k .
- $\theta_{ij} \approx \pi (180^\circ)$: The vectors are in anti-phase. This indicates a *negative correlation* (anti-correlation); when one term is prominent, the other is likely absent.
- $\theta_{ij} \approx \pm\pi/2 (\pm 90^\circ)$: The vectors are orthogonal in the complex plane. Their real inner product $\Re(P_{ij})$ is near zero, indicating *linear independence* or orthogonality, suggesting the terms are uncorrelated.

This phase analysis offers an intuitive, geometric interpretation of the estimated correlations directly from the complex-valued \mathbf{P} matrix, often allowing for qualitative analysis without explicitly computing the normalized ratio $\hat{\rho}_{ij}$ for every pair.

Non-Asymptotic Error Bounds. The quality of the estimation is not merely asymptotic; it comes with strong, non-asymptotic probabilistic guarantees crucial for applications.

LEMMA 3.3 (ERROR BOUND FOR CORRELATION ESTIMATION). *For any pair (i, j) , any embedding dimension k , and any confidence parameter $\delta \in (0, 1)$, the error of the correlation estimator is bounded with high probability:*

$$\mathbb{P}\left(|\hat{\rho}_{ij} - \rho_{ij}| \leq C \sqrt{\frac{\log(1/\delta)}{k}}\right) \geq 1 - \delta \quad (10)$$

where C is a constant independent of N and M .

This bound, which is derivable from concentration inequalities, is critical for two reasons. First, it confirms that the estimation error decreases at a rate of $O(1/\sqrt{k})$ as the embedding dimension k increases. Second, and more importantly, it provides a rigorous mechanism for selecting the embedding dimension k based on a desired error tolerance ϵ and confidence level δ . Solving the inequality $C \sqrt{\frac{\log(1/\delta)}{k}} \leq \epsilon$ for the parameter k yields the requirement $k \geq \frac{C^2 \log(1/\delta)}{\epsilon^2}$. The logarithmic dependence on the confidence parameter δ and the inverse quadratic dependence on the error tolerance ϵ are precisely what make this randomized approach both scalable and practical for large-scale applications.

Algorithmic Implication and Complexity. While the Cross-Power Spectral Density (CPSD) matrix $\mathbf{P} = \mathbf{B} \mathbf{B}^*$ provides the theoretical foundation for correlation estimation, its explicit construction presents a significant computational bottleneck. The computation of this $N \times N$ Hermitian matrix requires calculating the inner products between all pairs of the N term embeddings, leading to a time complexity of $O(N^2k)$. This quadratic scaling in the vocabulary size N renders the direct approach infeasible for large-scale applications where N can be in the millions. In the following section, we will demonstrate how to circumvent this prohibitive cost by leveraging sparse recovery techniques, which allow us to accurately estimate

465 the correlation for each term without ever explicitly constructing
 466 the full P matrix.
 467

4 Sparse Correlation Recovery via Compressed Sensing

471 A central challenge in our setting is efficiently discovering correlations
 472 between a term and all other terms in the dataset. The naive approach of computing all pairwise correlations is computationally prohibitive for large N , scaling as $O(N^2)$. However, a key
 473 insight underpinning our method is the *sparsity-of-effects* principle: in most instances, a given term w_i interacts significantly with
 474 only a small number, $S \ll N$, of other terms. The correlation vector
 475 $\rho_i = (\rho_{i1}, \dots, \rho_{iN})^\top$ is therefore **S -sparse** or approximately
 476 sparse.
 477

478 We leverage this sparsity by reformulating the correlation recovery
 479 problem as a **Compressed Sensing (CS)** task [9, 44]. Instead of
 480 measuring all N potential correlations directly, we acquire a small
 481 number of **compressed measurements** $k \ll N$ and use convex
 482 optimization to infer the sparse correlation vector.
 483

4.1 Sparse Recovery via Orthogonal Matching Pursuit

484 For each term w_i , we recover its correlation vector by formulating
 485 the recovery as an Orthogonal Matching Pursuit (OMP) problem
 486 [18]. This formulation is amenable to an iterative solution that pro-
 487 vides a computationally efficient alternative to convex optimization.
 488 Instead of solving a single minimization problem, OMP greedily
 489 constructs the sparse solution vector $x_i \in \mathbb{R}^N$ over S steps, where
 490 S is the target sparsity.
 491

492 The algorithm maintains a residual vector r , initialized as the
 493 measurement vector z_i . In each step, it performs two key operations:
 494

- 495 (1) **Identification:** It searches for the column ψ_j in the sensing
 matrix Ψ that is most correlated with the current residual r .
- 496 (2) **Projection:** It adds this column to an active set and then
 calculates the least-squares solution for the coefficients of
 all currently active columns, ensuring an optimal fit at each
 step. The residual is then updated by subtracting this new
 fit.

505 Let us dissect the components of this process. The solution vector
 506 \hat{x}_i that serves as an estimate for the true correlation vector ρ_i is
 507 built iteratively. The **identification step** greedily finds the most
 508 significant correlation at each stage. The **projection step** is crucial
 509 as it ensures the solution remains consistent with the observed
 510 measurements z_i by minimizing the error $\|\Psi x_i - z_i\|_2$ over the
 511 set of chosen correlations. This greedy construction avoids the
 512 high computational cost of ℓ_1 -minimization while still effectively
 513 leveraging the signal's underlying sparsity.
 514

4.2 Sensing Matrix and Measurement Vector

515 The efficacy of the compressed sensing approach critically depends
 516 on the design of the **sensing matrix** Ψ and the **measurement**
 517 **vector** z_i .

518 The sensing matrix $\Psi \in \mathbb{C}^{k \times N}$ is constructed from normalized
 519 measurement vectors (computed in Section 3.2) and is formally

523 defined as:
 524

$$\Psi = \left[\frac{\mathbf{b}_1}{\sqrt{P_{11}}}, \frac{\mathbf{b}_2}{\sqrt{P_{22}}}, \dots, \frac{\mathbf{b}_N}{\sqrt{P_{NN}}} \right]^\top. \quad (11)$$

525 Each column j of Ψ corresponds to a term w_j and is given by
 526 $b_j / \sqrt{P_{jj}}$. The normalization by $\sqrt{P_{jj}}$, which approximates the stan-
 527 dard deviation of term w_j 's vector,³ is crucial as it ensures that
 528 the energy of each column is controlled—a necessary condition for
 529 the theoretical recovery guarantees. This matrix acts as a linear
 530 projection operator that maps the high-dimensional correlation
 531 vector $\rho_i \in \mathbb{R}^N$ down to the low-dimensional measurement space
 532 \mathbb{C}^k .
 533

534 The corresponding measurement vector for term w_i is defined
 535 as $z_i = b_i / \sqrt{P_{ii}} \in \mathbb{C}^k$. This vector represents the observed com-
 536 pressed measurement. The entire system is designed such that this
 537 observation approximates a linear combination of the correlations
 538 ρ_{ij} between w_i and all other terms w_j , sensed through the matrix
 539 Ψ :

$$540 z_i \approx \Psi \rho_i = \sum_{j=1}^N \rho_{ij} \cdot \frac{\mathbf{b}_j}{\sqrt{P_{jj}}}. \quad (12)$$

541 This approximation contains noise and errors inherent in the em-
 542 bedding process, which we can model as an error vector e such that
 543 $z_i = \Psi \rho_i + e$. A standard result in compressed sensing theory is that
 544 the expected energy of this stochastic error, $\|e\|_2$, scales with the
 545 number of measurements as $O(\sqrt{k})$ [18]. The OMP algorithm is
 546 robust to this error, which is explicitly accounted for in its theoretical
 547 guarantees.
 548

4.3 Theoretical Guarantee: Stable Sparse Recovery

549 The power of this greedy approach is justified by the following
 550 theorem, which provides a rigorous bound on the recovery error
 551 for OMP.
 552

553 **THEOREM 4.1 (OMP RECOVERY GUARANTEE).** *If the sensing matrix*
 554 Ψ *satisfies the **Restricted Isometry Property (RIP)** of order $S+1$ with a sufficiently small constant δ_{S+1} , then the solution \hat{x}_i produced*
 555 *by the OMP algorithm after S steps satisfies:*

$$556 \|\hat{x}_i - \rho_i\|_2 \leq C_1 \|e\|_2 + C_2 \frac{\|\rho_i - \rho_i^S\|_1}{\sqrt{S}} \quad (13)$$

557 where e is the noise vector from the measurement model, ρ_i^S is the
 558 best S -term approximation of ρ_i , and C_1, C_2 are constants.
 559

560 **Interpretation of the Guarantee:** The provided recovery guar-
 561 antee has two key interpretations. First, the role of the **Restricted**
 562 **Isometry Property (RIP)** is fundamental. This condition requires
 563 that the sensing matrix Ψ acts as a near-isometry, meaning it ap-
 564 proximately preserves the lengths of all sparse vectors. This geom-
 565 etric preservation is crucial as it ensures the greedy selections made
 566 by OMP are reliable, preventing distinct sparse vectors from being
 567 mapped to the same compressed measurement. Random matrices,
 568 including the constructed Fourier-based matrix Ψ , are known to
 569 satisfy the RIP with high probability when the number of measure-
 570 ments scales as $k = O(S \log(N/S))$ [5, 9].
 571

572 ³ b_j is zero mean and $\sigma_j^2 = \frac{1}{k} \sum_{i=1}^k |b_{ji}|^2 = P_{jj}$
 573

Second, the error bound itself decomposes into two interpretable components. The term $C_1\|\mathbf{e}\|_2$ constitutes the noise term. As established previously, the energy of the error vector $\|\mathbf{e}\|_2$ is expected to scale as $O(\sqrt{k})$, so this term quantifies how the recovery error scales with noise in the measurement process. The term $C_2\|\boldsymbol{\rho}_i - \boldsymbol{\rho}_i^S\|_1/\sqrt{S}$ is the approximation term. This term is zero if the true correlation vector $\boldsymbol{\rho}_i$ is exactly S -sparse. If $\boldsymbol{\rho}_i$ is instead *compressible* (meaning its coefficients decay rapidly, allowing for a good S -term approximation), this term remains small. This demonstrates the stability and robustness of the OMP algorithm, as the error degrades gracefully for vectors that are not perfectly sparse.

4.4 Practical Implementation

The core computational challenge is executing the OMP algorithm for each term. Naively running OMP against all N other terms is prohibitive, carrying a cost of at least $O(N^2Sk)$. To achieve feasibility, we first precompute an approximate L -Nearest Neighbor (L -NN) graph in the embedding space [19, 30, 48, 51], which allows us to restrict the recovery problem for each term to a small candidate set of $L \ll N$ neighbors.

At this stage, one might consider a simpler heuristic: computing direct correlations (e.g., dot products) with these L neighbours, which we will refer to as the Craft with dot product baseline (Craft-DP). While computationally efficient, this approach has critical drawbacks that our sparse recovery formulation overcomes. The primary advantages of using OMP on the candidate set are:

- **Denoising and Robustness:** The simple heuristic is sensitive to noise, as geometric proximity in the embedding space may not reflect true correlation. OMP's model-based approach ($\mathbf{z}_i \approx \Psi \boldsymbol{\rho}_i$) acts as a powerful denoising filter, identifying the candidates that form the most consistent and signal-rich explanation.
- **Discovery of Non-Redundant Factors:** The simple heuristic evaluates each neighbor independently, often producing a list of highly correlated, redundant terms (e.g., "database", "SQL", "RDBMS"). OMP, by contrast, considers all candidates simultaneously. It naturally produces a parsimonious result by "explaining away" the influence of other terms once a primary correlate has been chosen, thereby revealing a more fundamental set of drivers.

For the OMP algorithm on the restricted set of L neighbors, the total cost plummets to a tractable $O(NSLk)$. This is linear in the number of terms N . The final output is a sparse graph requiring $O(NL)$ storage. To empirically validate our claims, we present an experimental comparison of our proposed method against Craft-DP in section 6.

4.5 Statistical Significance Testing and Multiple Testing Correction

In large-scale correlation analyses, the sheer number of pairwise tests performed—on the order of $O(N^2)$ —inevitably leads to a high number of false positives if statistical significance is not rigorously assessed. Without proper correction, many spurious correlations may be mistakenly identified as meaningful, undermining the reliability of the results. Therefore, we incorporate a two-stage statistical testing procedure: first, we assess the significance of each individual

correlation estimate, and second, we apply a multiple testing correction to control the overall false discovery rate across the entire set of hypotheses.

For each estimated correlation $\hat{\rho}_{ij}$, we test the null hypothesis $H_0 : \rho_{ij} = 0$. We apply Fisher's z-transform to transform the raw estimates into a statistic and stabilize the variance of the correlation coefficient:

$$z_{ij} = \frac{1}{2} \log \left(\frac{1 + \hat{\rho}_{ij}}{1 - \hat{\rho}_{ij}} \right) \quad (14)$$

Under H_0 , the statistic z_{ij} is approximately normally distributed with mean 0 and variance $\frac{1}{k-3}$ (where k is the embedding dimension). The resulting test statistic $T_{ij} = z_{ij}\sqrt{k-3}$ therefore follows a standard normal distribution, $T_{ij} \sim N(0, 1)$. This allows us to compute two-tailed p -values in the standard way.

Given the $O(N^2)$ hypotheses being tested, a multiple testing correction is imperative to control the false positive rate. We control the False Discovery Rate (FDR) using the Benjamini-Hochberg procedure [6]. For a fixed term w_i , let $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(m)}$ be the sorted p -values of all its m tested correlations. We find the largest index r such that:

$$p_{(r)} \leq \frac{r}{m} \alpha \quad (15)$$

We then reject all hypotheses corresponding to $p_{(1)}, \dots, p_{(r)}$, where α is the desired significance level (e.g., 0.05). The threshold $p_{BH} = p_{(r)}$ serves as the corrected significance level for term w_i .

The final output of our algorithm is a set of significant term pairs that meet both a strength and a significance criterion:

$$E = \{(w_i, w_j) : |\hat{\rho}_{ij}| > \tau \text{ and } p_{ij} \leq p_{BH}\} \quad (16)$$

where τ is a correlation strength threshold.

Complexity Analysis. The complexity of the algorithm is dominated by the initial signal embedding stage. The randomized Fourier transform takes $O(Nk \log M)$. The subsequent stages are computationally less expensive. The sparse correlation recovery via OMP is performed for each of the N terms on a restricted candidate set of size L . With a target sparsity of S , this recovery has a total time complexity of $O(NSLk)$.

The final statistical validation phase consists of two steps. First, calculating a p-value using the Fisher z-transform for each of the (at most) S non-zero correlations per term takes $O(NS)$ time in total. Second, applying the Benjamini-Hochberg procedure requires sorting these S p-values for each of the N terms, resulting in a total cost of $O(NS \log S)$.

5 THE CRAFT SYSTEM FRAMEWORK

We now present our proposed framework, **CRAFT (Corpus Relatedness Analysis Using Fourier Transforms)**. Our core insight is to re-cast the problem of large-scale term relatedness discovery from a high-dimensional statistical challenge to a low-dimensional signal processing task. We treat the presence of a given term across the corpus's M documents as a discrete signal. Semantically related terms will exhibit correlated patterns within their respective signals. We hypothesize that these correlations can be robustly identified in the compact frequency domain, bypassing the quadratic cost of direct comparison in the original term-document space.

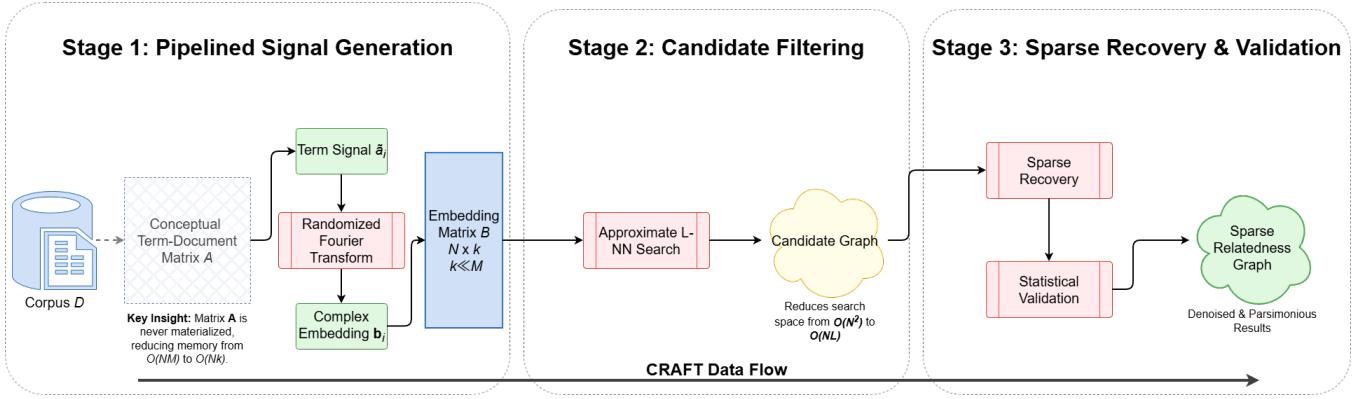


Figure 2: The CRAFT system architecture, illustrating the three main stages: 1) Pipelined Signal Generation and Embedding, 2) Candidate Filtering via Approximate k-NN, and 3) Sparse Recovery and Statistical Validation.

5.1 The CRAFT Architecture

To operationalize our insight, we designed CRAFT as a scalable, three-stage data pipeline. As illustrated in Figure 2, this architecture transforms the sparse text corpus into a low-dimensional representation where we can efficiently identify a small set of candidate correlations, which are then robustly recovered and validated. The implementation of this end-to-end process is formally detailed in Algorithm 1.

Stage 1: Pipelined Signal Generation and Embedding. The first stage of CRAFT, corresponding to lines 3-9 of Algorithm 1, transforms the raw text corpus \mathcal{D} directly into a dense, low-dimensional complex embedding matrix B , where each row represents a term. This is achieved through a pipelined process that is conceptually two-part:

- (1) **Signal Definition:** For each term w_i in the vocabulary \mathcal{V} , we first define its corresponding M -dimensional signal. This process begins with standard vocabulary construction (tokenization, pruning, etc.) and uses the BM25 weighting function (Equation (1)) to define a conceptual term-document matrix A . The final signal for term w_i , denoted \tilde{a}_i , is the corresponding mean-centered row of this conceptual matrix.
- (2) **Randomized Fourier Embedding:** As each high-dimensional signal \tilde{a}_i is generated, it is immediately projected into a k -dimensional complex embedding b_i via the Randomized Fourier Transform (Equation (3)). This projection provably preserves the inner product information essential for our analysis (Theorem 3.1).

This integrated pipeline is the key to our framework's scalability. By generating and immediately embedding each signal in the loop spanning lines 5-9, the conceptual $N \times M$ term-document matrix A is **never explicitly materialized**. This approach provides a significant performance benefit, **reducing the memory footprint from a prohibitive $O(NM)$ to a tractable $O(Nk)$** , as $k \ll M$. The final $N \times k$ matrix B is constructed directly, enabling CRAFT to process massive corpora on commodity hardware.

Stage 2: Candidate Filtering via Approximate k-NN. A naive $O(N^2k)$ all-pairs comparison on the embedding matrix B is intractable. Based on the **sparsity-of-effects hypothesis**—that any term is

only meaningfully related to a few others—we reframe this as an efficient candidate filtering problem. This stage is detailed in [Lines 10-16](#) of Algorithm 1. The goal is to find, for each term w_i , a small candidate set \mathcal{L}_i of its $L \ll N$ most likely interaction partners.

This task is equivalent to a large-scale **Approximate k-Nearest Neighbor (ANN)** search. Our implementation first normalizes the embedding vectors ([line 10](#)) and then uses **Faiss** [15], a high-performance similarity search library, to build a **Hierarchical Navigable Small Worlds (HNSW)** [32] index on the normalized vectors ([line 11](#)). We chose HNSW for its exceptional query speed, which is ideal for our "build-once, query-N-times" workload. By querying this index for each of the N terms in the loop at [lines 13-16](#), we find the L nearest neighbors for each term ([line 14](#)) and generate a candidate graph that reduces the search space from $O(N^2)$ to a tractable $O(NL)$.

Stage 3: Sparse Recovery via Orthogonal Matching Pursuit. The final stage, which covers [lines 17-38](#), takes the candidate graph and recovers a sparse set of high-quality correlations for each term. Instead of convex optimization, we adopt the Orthogonal Matching Pursuit (OMP) solver, a computationally efficient greedy algorithm that is robust to the noise inherent in the embedding process.

For each term w_i , as processed in the main loop ([lines 19-38](#)), we solve for its sparse correlation vector x_i from the model $z_i \approx \Psi_{\mathcal{L}_i} x_i$. Here, z_i is the normalized embedding for w_i ([line 21](#)) and the sensing matrix $\Psi_{\mathcal{L}_i}$ is constructed from the embeddings of its L candidates ([line 22](#)), using the full transposed matrix Ψ created on [line 18](#). The `SolveOMP` function is called to recover the sparse vector of correlations $\hat{x}_i^{\mathcal{L}}$ ([line 23](#)). OMP runs for a fixed number of iterations, S (the target sparsity). In each iteration, it performs two steps: 1) **Identification**, where it selects the candidate term that is most correlated with the current residual, and 2) **Projection**, where it computes a new least-squares solution over all currently selected candidates to update the coefficients and residual.

Following recovery, the identified non-zero correlations $\hat{\rho}_{ij}$ undergo rigorous statistical validation. For each potential correlation, a p-value p_{ij} is computed via Fisher's z-transform ([line 27](#)). To account for the large number of tests, we apply the Benjamini-Hochberg procedure to control the false discovery rate at a given

Algorithm 1 The End-to-End CRAFT Pipeline

```

813
814 1: Input: Corpus  $\mathcal{D}$ , Embedding dim.  $k$ , Neighbor set size  $L$ , Sig-
815  nificance level  $\alpha$ 
816 2: Output: Set of significant correlated pairs  $E$ 
817 3:  $\mathcal{V} \leftarrow \text{BuildVocabulary}(\mathcal{D})$ 
818 4:  $\mathbf{B} \leftarrow \text{zeros}(N, k, \text{dtype}=\text{complex})$ 
819 5: for each term  $w_i \in \mathcal{V}$  (from  $i = 1$  to  $N$ ) do
820 6:    $\tilde{\mathbf{a}}_i \leftarrow \text{GenerateCenteredSignal}(\mathcal{D}, w_i)$ 
821 7:    $\mathbf{b}_i \leftarrow \text{ApplyRFT}(\tilde{\mathbf{a}}_i)$ 
822 8:    $\mathbf{B}[i, :] \leftarrow \mathbf{b}_i^\top$ 
823 9: end for
824 10:  $\mathbf{B}_{\text{norm}} \leftarrow \text{NormalizeRows}(\mathbf{B})$ 
825 11:  $\text{FaissIndex} \leftarrow \text{BuildHNSWIndex}(\mathbf{B}_{\text{norm}})$ 
826 12:  $\mathcal{L} \leftarrow \text{new map}()$ 
827 13: for each term  $w_i \in \mathcal{V}$  (from  $i = 1$  to  $N$ ) do
828 14:    $\mathcal{L}_i \leftarrow \text{FaissIndex.Search}(\mathbf{B}_{\text{norm}}[i, :], L)$ 
829 15:    $\mathcal{L}[i] \leftarrow \mathcal{L}_i$ 
830 16: end for
831 17:  $E \leftarrow \emptyset$ 
832 18:  $\Psi \leftarrow \mathbf{B}_{\text{norm}}^\top$ 
833 19: for each term  $w_i \in \mathcal{V}$  (from  $i = 1$  to  $N$ ) do
834 20:    $\mathcal{L}_i \leftarrow \mathcal{L}[i]$ 
835 21:    $\mathbf{z}_i \leftarrow \Psi[:, i]$ 
836 22:    $\Psi_{\mathcal{L}_i} \leftarrow \Psi[:, \mathcal{L}_i]$ 
837 23:    $\hat{\mathbf{x}}_i^{\mathcal{L}} \leftarrow \text{SolveOMP}(\Psi_{\mathcal{L}_i}, \mathbf{z}_i, S)$ 
838 24:    $p\_values \leftarrow []$ 
839 25:    $results \leftarrow []$ 
840 26:   for each non-zero  $\hat{\rho}_{ij}$  in  $\hat{\mathbf{x}}_i^{\mathcal{L}}$  (with index  $j \in \mathcal{L}_i$ ) do
841 27:      $p_{ij} \leftarrow \text{FisherZTest}(\hat{\rho}_{ij}, M)$ 
842 28:      $p\_values.append(p_{ij})$ 
843 29:      $results.append((i, j, \hat{\rho}_{ij}, p_{ij}))$ 
844 30:   end for
845 31:    $p_{BH} \leftarrow \text{BenjaminiHochberg}(p\_values, \alpha)$ 
846 32:   for  $(i, j, \hat{\rho}_{ij}, p_{ij}) \in results$  do
847 33:     if  $p_{ij} \leq p_{BH}$  then
848 34:        $E \leftarrow E \cup \{(w_i, w_j)\}$ 
849 35:     end if
850 36:   end for
851 37: end for
852 38: return  $E$ 
853
854

```

significance level α (**line 32**). Only those correlations whose p-values are below the corrected threshold are deemed statistically significant and are added to the final output set E (**lines 33-37**). This set of significant correlated pairs is the final output of the pipeline. This model-based recovery acts as a powerful denoising filter and produces a parsimonious set of non-redundant factors. Since this entire process is run independently for each term on its restricted candidate set, it is embarrassingly parallel and computationally tractable.

6 Experimental Evaluation

In this section, we present a comprehensive empirical evaluation of CRAFT. Our experiments are designed to answer the following key research questions:

- **Effectiveness:** How does the quality of the term correlations discovered by CRAFT compare to alternative and baseline methods?
- **Efficiency & Scalability:** How do the runtime and memory requirements of CRAFT scale with the size of the corpus and vocabulary?
- **Ablation Study:** How do the core components of CRAFT contribute to its overall performance, and how sensitive is the model to its key hyperparameters?

6.1 Experimental Setup

Our experimental design is structured to rigorously evaluate CRAFT's performance across two primary dimensions: **effectiveness** and **efficiency**. For effectiveness, we evaluate all methods under two distinct scenarios: (1) a **ranked retrieval** task where the generated candidate pairs are ordered by the similarity score, and (2) a **significance set** task where a statistical test is used to select a final set of pairs. To support this comprehensive evaluation, we employ two categories of datasets.

Benchmark Datasets for Effectiveness Evaluation. To evaluate effectiveness, we employ two standard knowledge graph construction benchmarks from the Text2KG suite [42], both of which provide ground-truth triples for each text document. Their well-structured nature and moderate scale are ideal for a precise quantitative and qualitative analysis of the discovered correlations.

- **Wikidata-TekGen:** A dataset of 13K documents synthesized from Wikidata triples. It is characterized by a broad, general-domain vocabulary of entities and relations.
- **DBpedia-WebNLG:** A corpus of 4K documents containing descriptive, well-structured sentences generated from DBpedia triples, providing a clean and varied evaluation setting.

Large-Scale Corpora for Scalability Evaluation. To address efficiency and scalability, we also test our approach on two large corpora that are representative of challenging, large-scale text. Their significant size allows us to thoroughly test the performance of CRAFT.

- **GenWiki:** A large-scale, general-domain dataset constructed from English Wikipedia[25]. Comprising over 700K articles. It was created for KG construction with golden triplet and to overcome the limitations of smaller supervised datasets for unsupervised models.
- **arXiv Corpus:** To evaluate performance on domain-specific scientific text at scale, we use a series of corpora constructed from text chunks extracted from full-text computer science papers from arXiv[12]. Characterized by a specialized vocabulary, these serve as a robust benchmark. We use three versions of increasing size: **arXiv-1M** (1 million text chunks), **arXiv-3M** (3 million chunks), and **arXiv-5M** (5 million chunks).

Compared Methods. To rigorously evaluate our framework, we compare the full CRAFT pipeline against a series of methods. To ensure a fair comparison of the embedding and recovery stages, all methods leverage the same Approximate Nearest Neighbor (ANN) search component [19, 30, 48] for candidate filtering. The following methods are evaluated:

Table 1: Effectiveness on all datasets, measured by Mean Average Precision (mAP) and F1-Score.

Method	Wikidata		DBpedia		GenWiki		arXiv-1M		arXiv-3M		arXiv-5M	
	mAP	F1	mAP	F1	mAP	F1	mAP	F1	mAP	F1	mAP	F1
RP-ANN	92.7	46.1	97.2	51.2	80.5	42.5	75.3	40.1	74.1	39.5	73.5	39.0
SVD-ANN	95.1	33.6	99.1	40.1	81.2	40.8	77.8	38.4	76.5	37.9	75.9	37.1
CRAFT-DP	99.3	80.1	99.2	78.5	83.6	80.6	82.1	78.2	80.9	77.8	80.5	77.1
CRAFT	98.7	80.5	98.1	80.3	85.0	81.3	80.9	80.5	81.1	79.9	80.1	79.2

- (1) **RP-ANN:** This method represents a standard, scalable heuristic. It first builds a term-document matrix with BM25 weights, applies a **Random Projection (RP)** [3] to reduce dimensionality, and then performs the ANN search.
- (2) **SVD-ANN:** This baseline applies a truncated **Singular Value Decomposition (SVD)** to the BM25 matrix, similar to the core of the classic Latent Semantic Analysis method, to generate dense embeddings for the subsequent ANN search.
- (3) **CRAFT-DP:** This variant first applies the CRAFT Fourier embedding and ANN filtering stages. It then isolates our final stage by replacing the OMP sparse recovery step with a simple dot product ranking on the candidate set, allowing us to quantify the benefit of OMP as part of our ablation study.
- (4) **CRAFT:** This is our full, end-to-end proposed framework, including the initial Fourier embedding, the ANN candidate filtering, and the final sparse correlation recovery via Orthogonal Matching Pursuit.

Evaluation Metrics & Implementation.

- **Effectiveness:** We evaluate the quality of the discovered term pairs under the two scenarios based on the output format:
 - **Ranked List Evaluation:** For the ranked retrieval task, we report **Mean Average Precision (mAP)**. This metric is the mean of the Average Precision (AP) scores over a set of queries Q , defined as:

$$\text{mAP} = \frac{1}{|Q|} \sum_{q \in Q} \text{AP}(q)$$

where $\text{AP}(q)$ for a single query is the average of precision values at the position of each correct item in the ranked list. A higher mAP score indicates better performance, as it rewards models for placing correct items at the top of the list.

- **Significance Set Evaluation:** For the significance set task, we report **F1-Score**.
- A rigorous evaluation of correlation-based discovery methods requires a ground truth that is methodologically aligned with the chosen statistical metric. Accordingly, we construct our benchmark to consist exclusively of pairs that exhibit a strong statistical association as measured by the Pearson correlation coefficient. To this end, we operationally define a strong correlation as any pair with a Pearson score at or above the 95th percentile among all the possible pairs. This threshold is applied universally to ensure consistency:

- For datasets providing a golden set, we perform a refinement, retaining only those golden pairs that meet our 95th percentile criterion.
- For datasets without a golden set, this threshold is used to construct the ground truth directly.

This principled methodology guarantees that our benchmark is composed of a high-confidence set of pairs, ensuring the evaluation is a valid and precise measure of a system's ability to detect strong correlations.

- **Efficiency:** To characterize system performance, we measure both end-to-end wall-clock time and peak memory usage. Our analysis focuses on runtime as the primary efficiency metric, as the peak memory bottleneck is consistent across all methods.⁴ This bottleneck is the storage of the final dense embedding matrix of size $N \times k$, where N is the vocabulary size and k is the embedding dimension. Since this data structure is a shared requirement for all evaluated algorithms, we adopt runtime scalability as the primary axis for our efficiency evaluation.
- **Implementation:** All experiments were conducted on a server equipped with an AMD EPYC 7C13 CPU and 512 GB of RAM. Our system is implemented in Python, utilizing NumPy, SciPy, and the Faiss library for its HNSW index. For OMP, we adopted the fast algorithm from [50]. Unless otherwise noted, the default parameters for CRAFT are an embedding dimension of $k = 256$, a neighbor set size of $L = 100$ and a significance level $\alpha = 0.05$.

6.2 Effectiveness Evaluation

We evaluate effectiveness across all four datasets, with the primary results presented in Table 1. The analysis is divided into two parts, corresponding to our two evaluation scenarios: ranking quality and the quality of the final significance set.

Ranking Quality. As shown in Table 1, our CRAFT variants consistently outperform the baselines across all datasets. On well-structured benchmarks like **Wikidata** and **DBpedia**, all methods achieve high mAP scores due to the clean, template-based text, with our approach reaching a near-perfect mAP of 99.3 on **Wikidata**.

The performance gap widens on the more challenging large-scale datasets (**GenWiki** and **arXiv**). The complex syntax and noise in these corpora cause mAP scores to drop for all methods, highlighting the difficulty of the task. It is in these conditions, however, that

⁴For SVD-ANN, the full term-document matrix can be prohibitively large if densified. Standard implementations, however, typically operate directly on a sparse representation of this matrix to manage memory consumption.

1045 **CRAFT**'s robustness becomes evident. On the largest **arXiv-5M**
 1046 corpus, while baseline performance degrades significantly, CRAFT
 1047 achieves an mAP of 80.1, surpassing the strongest baseline's score of
 1048 75.9. This confirms our framework's ability to produce high-quality
 1049 rankings on both clean and real-world text.

1050 *Significance Set Quality.* The F1-Score evaluates the quality of the
 1051 final, discrete set of pairs deemed significant. As shown in Table 1,
 1052 both of our framework's variants, **CRAFT-DP** and **CRAFT**, dra-
 1053 matically outperform the baselines on this metric. The baselines
 1054 (**RP-ANN** and **SVD-ANN**) suffer from extremely low precision.
 1055 Their underlying embeddings produce dense, noisy neighborhoods
 1056 in the ANN search, resulting in an over-generation of false positive
 1057 candidates. This severely harms precision and leads to poor
 1058 F1-scores.

1059 Our framework addresses this in two stages. First, the strong per-
 1060 formance of **CRAFT-DP** indicates that the **Randomized Fourier**
 1061 **Transform (RFT)** embeddings are inherently better at separating
 1062 signal from noise. They produce a much cleaner initial candidate
 1063 set from the ANN search, leading to the first substantial jump in
 1064 F1-score over the baselines. Second, the full **CRAFT** model intro-
 1065 duces the sparse recovery stage via Orthogonal Matching Pursuit
 1066 (OMP). This acts as a powerful sparse recovery filter, taking the
 1067 already high-quality candidate set and further refining it to prune
 1068 remaining false positives. This combination—RFT for a high-quality
 1069 initial set and OMP for principled final filtering—is what allows our
 1070 framework to achieve its state-of-the-art F1-scores.

6.3 Efficiency and Scalability

1071 To evaluate the practicality of CRAFT for large-scale applications,
 1072 we analyze its computational performance on our two largest cor-
 1073 pora, **GenWiki** and **arXiv** series. We separately analyze runtime
 1074 and memory usage to provide a comprehensive picture of the sys-
 1075 tem's efficiency.

1076 We evaluate the end-to-end runtime, normalized against the
 1077 fastest baseline, **RP-ANN**. The results in Table 2 show the relative
 1078 slowdown.

1079 The performance differences are rooted in the computational
 1080 complexity of the underlying algorithms. For our **SVD-ANN** base-
 1081 line, we employ a highly efficient implementation based on **ran-**
 1082 **domized SVD**, which is a state-of-the-art approach for approx-
 1083 imating the top- k singular vectors of large matrices [21]. While
 1084 this randomized approach is significantly more scalable than a full
 1085 decomposition, its core reliance on matrix factorization still leads
 1086 to a computational bottleneck as the vocabulary and document
 1087 counts grow. This explains the significant slowdown observed for
 1088 **SVD-ANN** on the larger **arXiv** corpora.

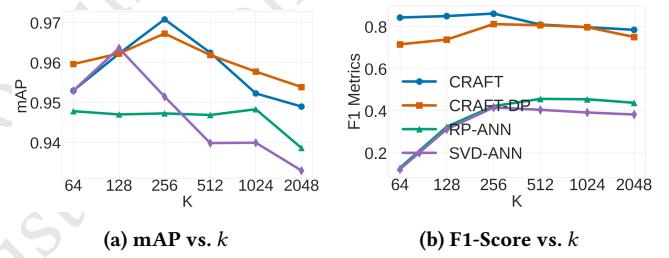
1089 In contrast, **CRAFT** is designed for efficiency. Its primary em-
 1090 bedding stage uses a randomized Fourier transform, an operation
 1091 related to the Fast Fourier Transform (FFT) with a near-linear time
 1092 complexity. The subsequent OMP recovery step, while iterative,
 1093 is only performed on small candidate sets for each term, not the
 1094 entire vocabulary. Consequently, CRAFT's overall runtime remains
 1095 highly competitive, scaling nearly as well as the much less effec-
 1096 tive RP-ANN heuristic. This result is critical, as it demonstrates
 1097 that CRAFT's superior effectiveness is achieved with only a minor
 1098 computational overhead compared to the fastest baseline.

1099 **Table 2: Relative slowdown compared to the RP-ANN baseline**
 1100 **on the large-scale datasets. A value of 1.0x means the runtime**
 1101 **is identical to RP-ANN.**

Method	GenWiki	arXiv-1M	arXiv-3M	arXiv-5M
RP-ANN	1.00x	1.00x	1.00x	1.00x
SVD-ANN	1.20x	1.52x	1.68x	1.85x
CRAFT-DP	1.06x	1.12x	1.15x	1.18x
CRAFT	1.15x	1.21x	1.22x	1.23x

6.4 Ablation Study

1102 We conduct an ablation study to analyze the individual contribu-
 1103 tions and computational cost of CRAFT's key components, as well
 1104 as to evaluate its sensitivity to its primary hyperparameters: the
 1105 embedding dimension k , the nearest neighbor list size L , and the
 1106 significance level α . This study uses a subsample of the **Genwiki**
 1107 dataset with 100K documents. Unless otherwise noted, the default
 1108 parameters are an embedding dimension of $k = 256$, a neighbour
 1109 set size of $L = 100$, and a significance level $\alpha = 0.05$.



1110 **Figure 3: Ablation study on the impact of embedding dimen-**
 1111 **sion (k) on performance.**

1112 *Sensitivity to Embedding Dimension k .* The choice of embedding
 1113 dimension k presents a trade-off between expressiveness and com-
 1114 putational cost. Our analysis shows that the effectiveness of CRAFT
 1115 is robust across a wide range of k values.

1116 As shown in Figure 3a, the ranking quality (mAP) for both
 1117 **CRAFT** and **CRAFT-DP** peaks at a moderate dimension of $k = 256$.
 1118 This behavior suggests that an embedding dimension of $k = 256$
 1119 effectively approximates the **intrinsic dimensionality** of the seman-
 1120 tic space within the Genwiki dataset for this task. Dimensions
 1121 smaller than this may be insufficient to capture the nuanced rela-
 1122 tionships between entities, leading to underfitting. Conversely,
 1123 dimensions significantly larger than the intrinsic requirement can
 1124 introduce redundancy and increase the model's capacity to overfit
 1125 to noise in the training data. This likely explains the slight, grad-
 1126 ual decrease in mAP observed for $k > 256$, as the model begins
 1127 to capture spurious correlations rather than the true underlying
 1128 data structure. In contrast, the baseline methods show either lower
 1129 peak performance (SVD-ANN) or relative insensitivity coupled with
 1130 much lower overall effectiveness (RP-ANN).

1131 For the significance set quality (F1-Score), shown in Figure 3b,
 1132 our framework demonstrates even greater stability. After an ini-
 1133 tial sharp improvement, the F1-scores for CRAFT and CRAFT-DP
 1134

remain high and plateau for all $k \geq 256$. This indicates that once the embedding dimension is sufficiently large to capture the core semantic relationships (i.e., it meets or exceeds the intrinsic dimension), the performance is not sensitive to further increases in dimensionality. This robustness is a key practical advantage, as it simplifies hyperparameter tuning.

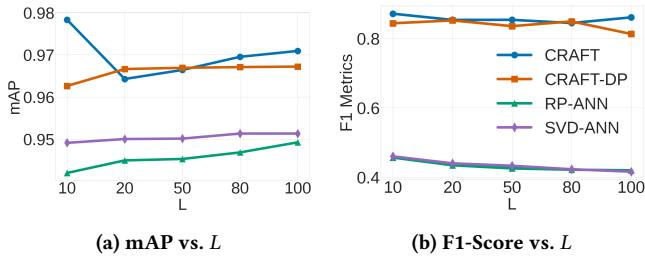


Figure 4: Ablation study on the impact of nearest neighbor list size (L) on performance.

Sensitivity to Nearest Neighbor List Size L . We also evaluate the model’s sensitivity to the size of the nearest neighbor list, L , which is a crucial parameter for the candidate generation phase. The results show that our proposed methods maintain high performance and are remarkably stable across different values of L .

As depicted in Figure 4a, the ranking performance (mAP) of the baseline methods generally improves slightly as L increases. This is an expected outcome, as a larger candidate pool raises the potential for recall. However, **CRAFT** exhibits a more complex trend, with its high performance being less dependent on a simple expansion of the candidate list. This is because its final ranking stage does not rely on a direct dot product but instead uses Orthogonal Matching Pursuit (OMP) to find a sparse approximation of the query. OMP seeks an optimal combinatorial representation, a process not guaranteed to improve monotonically with a larger, potentially less relevant, set of candidates. The consistently superior mAP of **CRAFT** and **CRAFT-DP** across all values of L underscores the effectiveness of our overall embedding and ranking strategy.

More importantly, the quality of the significance set (F1-Score) reveals a key difference between our framework and the baselines, as shown in Figure 4b. For SVD-ANN and RP-ANN, the F1-Score tends to decrease as L grows. This suggests that these methods do not produce a **well-separated similarity measure**; there is no clear distinction between the scores of truly relevant and irrelevant entities. Consequently, increasing the candidate list size introduces more borderline candidates that are incorrectly identified as significant by the statistical test, leading to an influx of **false positives** that degrades the F1-score. In stark contrast, the F1-scores for **CRAFT** and **CRAFT-DP** remain exceptionally stable and high. This demonstrates that our framework’s similarity measure is highly discriminative, allowing it to reliably identify the core set of significant neighbours regardless of the total list size. This robustness is a valuable practical feature, enabling reliable results without extensive parameter tuning.

Sensitivity to Significance Level α . Finally, we analyze the framework’s sensitivity to the significance level α , which is the threshold

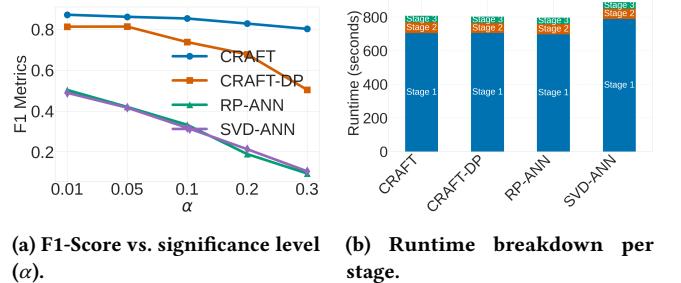


Figure 5: Ablation study on the impact of significance level (α) and a component-wise runtime analysis.

used in the final statistical testing phase to determine the significance set. As this test is only applied to generate the final set, this hyperparameter exclusively impacts the **F1-Score**.

Figure 5a shows the F1-Score as a function of α . For all methods, the F1-Score decreases as α becomes less strict (i.e., increases). This is because a larger α lowers the bar for statistical significance, leading to a sharp increase in the number of **false positives** being included in the final set. This influx of false positives severely degrades precision, which in turn lowers the F1-Score.

Notably, **CRAFT** demonstrates the highest F1-score and greatest robustness across all tested values, followed by **CRAFT-DP**, with both methods significantly outperforming the baselines. The superior stability of **CRAFT**, particularly as the significance threshold loosens, indicates that its OMP-based similarity measure is more discriminative than the direct dot-product approach in **CRAFT-DP**. This allows the statistical test to effectively identify true positives, which have a strong signal, without wrongly including a large number of false positives. In contrast, the performance of SVD-ANN and RP-ANN degrades rapidly, suggesting their underlying similarity scores are less reliable and not well-separated, causing the statistical test to fail.

Component-wise Runtime Analysis. To understand the computational costs of our framework, we analyze the runtime of each major stage for this 100k dataset, as shown in Figure 5b. The process is divided into three stages: **Stage 1** (Embedding Generation), **Stage 2** (Candidate Filtering), and **Stage 3** (Sparse Recovery/reranking with significance Testing).

The analysis reveals that Stage 1 is the dominant computational bottleneck across all methods, consuming the vast majority of the total processing time. This is expected, as generating high-dimensional embeddings are inherently expensive operations.

Among the evaluated methods, **SVD-ANN** exhibits the highest overall runtime, driven primarily by a more costly initial indexing phase in Stage 1. In contrast, **CRAFT**, **CRAFT-DP**, and **RP-ANN** demonstrate comparable total runtimes. Notably, while **CRAFT** employs a more sophisticated Orthogonal Matching Pursuit algorithm in its ranking stage (Stage 3), its runtime overhead is minimal compared to the simpler dot-product re-ranking in **CRAFT-DP** and the baselines.

This result is significant: it shows that the substantial improvements in ranking quality (mAP) and significance set identification (F1-Score) delivered by **CRAFT** are achieved without a meaningful

penalty in computational cost. The primary expense lies in the embedding and indexing stage, which is similar across most methods, while our more effective ranking and testing stages add negligible overhead.

1281

1282

1283

7 Related Work

The fundamental task of identifying meaningful relationships between terms in a large corpus has been a long-standing challenge in natural language processing and information retrieval.

1287

1288

1289

1290

1291

1292

1293

1294

1295

1296

1297

1298

1299

1300

1301

1302

1303

1304

1305

1306

1307

1308

1309

1310

1311

1312

1313

1314

1315

1316

1317

1318

1319

1320

1321

1322

1323

1324

1325

1326

1327

1328

1329

1330

1331

1332

1333

1334

Statistical Association Measures. A foundational approach relies on computing statistical measures from a term co-occurrence matrix. Classic techniques include Pointwise Mutual Information (PMI) [11], which measures the deviation of the observed co-occurrence probability from statistical independence, and the Chi-Squared (χ^2) test [33], which assesses the significance of the co-occurrence. While intuitive and computationally straightforward for a single pair, a critical limitation is their inherent quadratic scaling; exhaustively evaluating all possible term pairs across a large vocabulary of size N requires $O(N^2)$ computations, which is prohibitive for large-scale applications. This scalability challenge is a classic data management problem, previously highlighted by the database community for data mining[16, 49]. Furthermore, PMI is known to be biased towards infrequent events, overestimating the importance of rare term pairs, while χ^2 can be difficult to interpret as a strength-of-association metric. In contrast, the Pearson Correlation Coefficient offers a more suitable measure for our goal of identifying terms with consistently covarying frequencies. It is normalized, providing a intuitive bounded score between -1 and 1 that indicates both the strength and direction (positive or negative) of a linear relationship. Unlike PMI, it is less sensitive to sparse counts and effectively captures broader, more proportional co-occurrence patterns, making it a robust choice for filtering widely correlated entity pairs.

Graph-Based Methods. To capture global and indirect relationships, one could model the corpus as a graph where nodes represent terms and edges are weighted by an association measure like PMI. Algorithms such as PageRank [8] or HITS [27] can then be used to identify central or authoritative entities within this network. The primary challenge for these methods is their extreme computational and memory overhead, as constructing and processing a dense term-term graph for a large vocabulary is often intractable.

Neural and Embedding-Based Methods. Modern approaches leverage distributional semantics, learning dense vector representations for terms from their contexts. Models like Word2Vec [35] and GloVe [40] infer relationships via cosine similarity in the embedding space. More recently, large-scale pre-trained language models like BERT [14] have achieved state-of-the-art performance on supervised relation extraction tasks. However, the computational cost of training embeddings or performing inference with large models for every term pair is prohibitive for exhaustive, discovery-oriented analysis over massive corpora. Our work aims to circumvent this bottleneck by providing an ultra-efficient, lightweight candidate generation engine, enabling the application of these powerful but expensive models on a focused, high-quality subset of potential term pairs.

Latent Semantic Analysis (LSA). LSA [13] applies Singular Value Decomposition (SVD) to a term-document matrix to project

terms into a latent topic space. While effective, the computational complexity of performing a full SVD on a large matrix is often too expensive for truly large-scale data. Moreover the nature of the associations discovered by LSA are fairly different. Pairs discovered by LSA have high cosine similarity in the reduced space and the measure has no universal scale. In contrast Pearson correlation is strictly between -1 and 1 and universal. The LSA relationships are strictly symmetric as opposed to Pearson correlation that discovers directional relationships (positive, negative). Finally relationships discovered by LSA are based on shared latent content as opposed to Pearson correlation in which pairs have a precise, directional linear relationship where one variable changes predictably with the other.

Random Projections. A highly scalable alternative is to use random projections, which are guaranteed to preserve pairwise distances between points with high probability under the Johnson-Lindenstrauss lemma [24, 26]. This involves projecting the high dimensional data onto a random lower-dimensional subspace. One could directly apply random projections to the term vectors derived from an inverted index. While feasible, this approach still requires associating N vectors, resulting in an $O(N^2)$ operation, which entails significant overheads. Our method avoids this quadratic bottleneck entirely by operating directly on the compressed frequency domain using sparse recovery techniques. Furthermore, standard random projections are designed to preserve Euclidean geometry, whereas our use of the Random Fourier Transform (RFT) is explicitly designed to create a compact sketch that preserves the frequency-domain correlations and cross-power spectral properties that are central to efficiently estimating Pearson correlation. This provides a direct and computationally advantageous link between the time-domain (token occurrences) and the desired association metric.

8 Conclusion

Discovering related terms in massive, unstructured text corpora is a fundamental challenge for applications like knowledge graph construction. We present **CRAFT**, a system that addresses this problem by reframing it through the lens of signal processing.

CRAFT makes three core technical contributions. First, we develop a **randomized Fourier embedding** to efficiently compress the sparse term-document matrix into a dense representation, with theoretical guarantees on preserving the inner products crucial for correlation analysis. Second, we adapt **Cross-Power Spectral Density (CPSD)** to robustly estimate statistical correlations in this compact frequency domain. Third, we formulate term discovery as a **sparse recovery** problem, leveraging Orthogonal Matching Pursuit (OMP) to identify the most significant relationships without materializing the full pairwise correlation matrix.

Our comprehensive experimental evaluation confirms that CRAFT significantly outperforms state-of-the-art baselines, achieving superior precision with near-linear scalability. This establishes CRAFT as a highly efficient and accurate first-pass filter for candidate discovery, providing a scalable foundation to bootstrap downstream tasks or complement the reasoning of large language models.

1335

1336

1337

1338

1339

1340

1341

1342

1343

1344

1345

1346

1347

1348

1349

1350

1351

1352

1353

1354

1355

1356

1357

1358

1359

1360

1361

1362

1363

1364

1365

1366

1367

1368

1369

1370

1371

1372

1373

1374

1375

1376

1377

1378

1379

1380

1381

1382

1383

1384

1385

1386

1387

1388

1389

1390

1391

1393 References

- [1] 2025. CRAFT: Corpus Relatedness Analysis Using Fourier Tranforms. <https://anonymous.4open.science/r/CRAFT-FC97>.
- [2] Hassan Abdallah, Béatrice Markhoff, and Arnaud Soulet. 2025. Ranking Indicator Discovery from Very Large Knowledge Graphs. *Proc. VLDB Endow.* 18, 4 (May 2025), 1183–1195. doi:10.14778/3717755.3717775
- [3] Dimitris Achlioptas. 2003. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *J. Comput. System Sci.* 66, 4 (2003), 671–687. doi:10.1016/S0022-0002(03)00025-4 Special Issue on PODS 2001.
- [4] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. DBpedia: a nucleus for a web of open data. In *Proceedings of the 6th International The Semantic Web and 2nd Asian Conference on Asian Semantic Web Conference* (Busan, Korea) (ISWC'07/ASWC'07). Springer-Verlag, Berlin, Heidelberg, 722–735.
- [5] Richard Baraniuk, Mark Davenport, Ronald DeVore, and Michael Wakin. 2008. A Simple Proof of the Restricted Isometry Property for Random Matrices. *Constructive Approximation* 28 (12 2008), 253–263. doi:10.1007/s00365-007-9003-x
- [6] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 57, 1 (1995), 289–300. <http://www.jstor.org/stable/2346101>
- [7] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2* (Lake Tahoe, Nevada) (NIPS'13). Curran Associates Inc., Red Hook, NY, USA, 2787–2795.
- [8] Sergey Brin. 1998. The PageRank citation ranking: bringing order to the web. *Proceedings of ASIS*, 1998 98 (1998), 161–172.
- [9] Emmanuel Candès, Justin Romberg, and Terence Tao. 2004. Robust Uncertainty Principles: Exact Signal Reconstruction from Highly Incomplete Frequency Information. arXiv:math/0409186 [math.NA] <https://arxiv.org/abs/math/0409186>
- [10] Claudio Carpineto and Giovanni Romano. 2012. A Survey of Automatic Query Expansion in Information Retrieval. *ACM Comput. Surv.* 44, 1, Article 1 (Jan. 2012), 50 pages. doi:10.1145/2071389.2071390
- [11] Kenneth Ward Church and Patrick Hanks. 1990. Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics* 16, 1 (1990), 22–29. <https://aclanthology.org/J90-1003/>
- [12] Colin B. Clement, Matthew Bierbaum, Kevin P. O’Keeffe, and Alexander A. Alemi. 2019. On the Use of ArXiv as a Dataset. arXiv:1905.00075 [cs.IR] <https://arxiv.org/abs/1905.00075>
- [13] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science* 41, 6 (1990), 391–407. doi:10.1002/(SICI)1097-4571(199009)41:6<391::AID-AS1>3.0.CO;2-9
- [14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs.CL] <https://arxiv.org/abs/1810.04805>
- [15] Matthijs Douze, Alexandre Guizheng, Chengqi Deng, Jeff Johnson, Gergely Szilvassy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The Faiss library. (2024). arXiv:2401.08281 [cs.LG]
- [16] Ahmed El-Kishky, Yanglei Song, Chi Wang, Clara R. Voss, and Jiawei Han. 2014. Scalable topical phrase mining from text corpora. *Proc. VLDB Endow.* 8, 3 (Nov. 2014), 305–316. doi:10.14778/2735508.2735519
- [17] Usama M. Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. 1996. *From data mining to knowledge discovery: an overview*. American Association for Artificial Intelligence, USA, 1–34.
- [18] Simon Foucart and Holger Rauhut. 2013. *A Mathematical Introduction to Compressive Sensing*. Birkhäuser Basel.
- [19] Cong Fu, Chao Xiang, Changxu Wang, and Deng Cai. 2019. Fast approximate nearest neighbor search with the navigating spreading-out graph. *Proc. VLDB Endow.* 12, 5 (Jan. 2019), 461–474. doi:10.14778/3303753.3303754
- [20] Leslie Greengard and June-Yub Lee. 2004. Accelerating the Nonuniform Fast Fourier Transform. *SIAM Rev.* 46, 3 (2004), 443–454. arXiv:<https://doi.org/10.1137/S003614450343200X> doi:10.1137/S003614450343200X
- [21] Nathan Halko, Per-Gunnar Martinsson, and Joel A. Tropp. 2010. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. arXiv:0909.4061 [math.NA] <https://arxiv.org/abs/0909.4061>
- [22] Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia D’Amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann. 2021. Knowledge Graphs. *Comput. Surveys* 54, 4 (July 2021), 1–37. doi:10.1145/3447772
- [23] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *Comput. Surveys* 55, 12 (March 2023), 1–38.
- [24] Jiawei Jiang, Fangcheng Fu, Tong Yang, and Bin Cui. 2018. SketchML: Accelerating Distributed Machine Learning with Data Sketches. In *Proceedings of the 2018 International Conference on Management of Data* (Houston, TX, USA) (SIGMOD ’18). Association for Computing Machinery, New York, NY, USA, 1269–1284. doi:10.1145/3183713.3196894
- [25] Zhijing Jin, Qipeng Guo, Xipeng Qiu, and Zheng Zhang. 2020. GenWiki: A Dataset of 1.3 Million Content-Sharing Text and Graphs for Unsupervised Graph-to-Text Generation. In *Proceedings of the 28th International Conference on Computational Linguistics*, Donia Scott, Nuria Bel, and Chengqing Zong (Eds.). International Committee on Computational Linguistics, Barcelona, Spain (Online), 2398–2409. doi:10.18653/v1/2020.coling-main.217
- [26] William B Johnson, Joram Lindenstrauss, et al. 1984. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary mathematics* 26, 189–206 (1984), 1.
- [27] Jon M. Kleinberg. 1999. Authoritative sources in a hyperlinked environment. *J. ACM* 46, 5 (Sept. 1999), 604–632. doi:10.1145/324133.324140
- [28] Shantanu Kumar. 2017. A Survey of Deep Learning Methods for Relation Extraction. arXiv:1705.03645 [cs.CL] <https://arxiv.org/abs/1705.03645>
- [29] Patrick Lewis, Ethan Perez, Aleksandr Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) (NIPS ’20). Curran Associates Inc., Red Hook, NY, USA, Article 793, 16 pages.
- [30] Kejing Lu, Mineichi Kudo, Chuan Xiao, and Yoshiharu Ishikawa. 2021. HVS: hierarchical graph structure based on voronoi diagrams for solving approximate nearest neighbor search. *Proc. VLDB Endow.* 15, 2 (Oct. 2021), 246–258. doi:10.14778/3489496.3489506
- [31] P. M. Mäkilä, J. R. Partington, and T. Norlander. 1998. Bounded Power Signal Spaces for Robust Control and Modeling. *SIAM Journal on Control and Optimization* 37, 1 (1998), 92–117. arXiv:<https://doi.org/10.1137/S0363012997316664> doi:10.1137/S0363012997316664
- [32] Yu A. Malkov and D. A. Yashunin. 2020. Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs. *IEEE Trans. Pattern Anal. Mach. Intell.* 42, 4 (April 2020), 824–836. doi:10.1109/TPAMI.2018.2889473
- [33] Christopher D Manning and Hinrich Schütze. 1999. Foundations of statistical natural language processing.
- [34] Stefano Marchesin and Gianmaria Silvello. 2025. Credible Intervals for Knowledge Graph Accuracy Estimation. *Proc. ACM Manag. Data* 3, 3, Article 142 (June 2025), 26 pages. doi:10.1145/3725279
- [35] Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*. <http://arxiv.org/abs/1301.3781>
- [36] Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, Keh-Yih Su, Jian Su, Janyce Wiebe, and Haizhou Li (Eds.). Association for Computational Linguistics, Suntec, Singapore, 1003–1011. <https://aclanthology.org/P09-1113/>
- [37] Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. 2011. A three-way model for collective learning on multi-relational data. In *Proceedings of the 28th International Conference on International Conference on Machine Learning* (Bellevue, Washington, USA) (ICML ’11). Omnipress, Madison, WI, USA, 809–816.
- [38] Reham Omar, Omij Mangukya, and Essam Mansour. 2025. Dialogue Benchmark Generation from Knowledge Graphs with Cost-Effective Retrieval-Augmented LLMs. *Proc. ACM Manag. Data* 3, 1, Article 31 (Feb. 2025), 26 pages. doi:10.1145/3709681
- [39] Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024. Unifying Large Language Models and Knowledge Graphs: A Roadmap. *IEEE Transactions on Knowledge and Data Engineering* 36, 7 (July 2024), 3580–3599. doi:10.1109/tkde.2024.3352100
- [40] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Alessandro Moschitti, Bo Pang, and Walter Daelemans (Eds.). Association for Computational Linguistics, Doha, Qatar, 1532–1543. doi:10.3115/v1/D14-1162
- [41] Ali Rahimi and Benjamin Recht. 2007. Random Features for Large-Scale Kernel Machines. In *Advances in Neural Information Processing Systems*, J. Platt, D. Koller, Y. Singer, and S. Roweis (Eds.), Vol. 20. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2007/file/013a006f03dbc5392effeb8f18fd4755-Paper.pdf
- [42] Diego Rincon-Yanez and Sabrina Senatore. 2022. FAIR Knowledge Graph construction from text, an approach applied to fictional novels. In *Proceedings of the 1st International Workshop on Knowledge Graph Generation From Text and the 1st International Workshop on Modular Knowledge co-located with 19th Extended Semantic Conference (ESWC 2022)*. CEUR-WS, Hersonissos, Greece, 94–108.

- 1509 http://ceur-ws.org/Vol-3184/TEXT2KG_Paper_7.pdf 1567
- 1510 [43] Stephen Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Frame- 1568
- 1511 work: BM25 and Beyond. *Found. Trends Inf. Retr.* 3, 4 (April 2009), 333–389. 1569
doi:10.1561/1500000019
- 1512 [44] Florin Rusu and Alin Dobro. 2007. Statistical analysis of sketch estimators. In 1570
Proceedings of the 2007 ACM SIGMOD International Conference on Management of 1571
Data (Beijing, China) (*SIGMOD '07*). Association for Computing Machinery, New 1572
York, NY, USA, 187–198. doi:10.1145/1247480.1247503
- 1513 [45] Chuan Shi, Yitong Li, Jiawei Zhang, Yizhou Sun, and Philip S. Yu. 2017. A Survey 1573
of Heterogeneous Information Network Analysis. *IEEE Trans. on Knowl. and 1574*
Data Eng. 29, 1 (Jan. 2017), 17–37. doi:10.1109/TKDE.2016.2598561
- 1514 [46] Joel A. Tropp and Anna C. Gilbert. 2007. Signal Recovery From Random 1575
Measurements Via Orthogonal Matching Pursuit. *IEEE Transactions on Information 1576*
Theory 53, 12 (2007), 4655–4666. doi:10.1109/TIT.2007.909108
- 1515 [47] Feiyu Wang, Qizhi Chen, Yuanpeng Li, Tong Yang, Yaofeng Tu, Lian Yu, and Bin 1577
Cui. 2023. JoinSketch: A Sketch Algorithm for Accurate and Unbiased Inner- 1578
Product Estimation. *Proc. ACM Manag. Data* 1, 1, Article 81 (May 2023), 26 pages. 1579
doi:10.1145/3588935
- 1516 [48] Mengzhao Wang, Xiaoliang Xu, Qiang Yue, and Yuxiang Wang. 2021. A com- 1580
prehensive survey and experimental comparison of graph-based approximate 1581
nearest neighbor search. *Proc. VLDB Endow.* 14, 11 (July 2021), 1964–1978. 1582
doi:10.14778/3476249.3476255
- 1517 [49] Hao Yan, Shuming Shi, Fan Zhang, Torsten Suel, and Ji-Rong Wen. 2010. Efficient 1583
term proximity search with term-pair indexes. In *Proceedings of the 19th ACM 1584*
International Conference on Information and Knowledge Management (Toronto, 1585
ON, Canada) (*CIKM '10*). Association for Computing Machinery, New York, NY, 1586
USA, 1229–1238. doi:10.1145/1871437.1871593
- 1518 [50] Huiyuan Yu, Jia He, and Maggie Cheng. 2025. Fast Orthogonal Matching Pursuit 1587
through Successive Regression. arXiv:2404.00146 [cs.CV] <https://arxiv.org/abs/2404.00146>
- 1519 [51] Xi Zhao, Yao Tian, Kai Huang, Bolong Zheng, and Xiaofang Zhou. 2023. Towards 1588
Efficient Index Construction and Approximate Nearest Neighbor Search in High- 1589
Dimensional Spaces. *Proc. VLDB Endow.* 16, 8 (April 2023), 1979–1991. doi:10. 1590
14778/3594512.3594527
- 1520
- 1521
- 1522
- 1523
- 1524
- 1525
- 1526
- 1527
- 1528
- 1529
- 1530
- 1531
- 1532
- 1533
- 1534
- 1535
- 1536
- 1537
- 1538
- 1539
- 1540
- 1541
- 1542
- 1543
- 1544
- 1545
- 1546
- 1547
- 1548
- 1549
- 1550
- 1551
- 1552
- 1553
- 1554
- 1555
- 1556
- 1557
- 1558
- 1559
- 1560
- 1561
- 1562
- 1563
- 1564
- 1565
- 1566

1625 Appendix: Detailed Proofs of Theorems

1626 This appendix provides comprehensive and detailed proofs for the
 1627 key theorems stated in Section 3, ensuring mathematical rigor and
 1628 completeness.

1629 Proof of Theorem 3.1: Inner Product Preservation

1630 We restate Theorem 3.1 for clarity.

1633 **THEOREM (THEOREM 3.1: INNER PRODUCT PRESERVATION).** For
 1634 any two term vectors $\tilde{\mathbf{a}}_i, \tilde{\mathbf{a}}_j \in \mathbb{R}^M$, their randomized Fourier embed-
 1635 dings $\mathbf{b}_i, \mathbf{b}_j \in \mathbb{C}^k$, defined by $\mathbf{b}_i = \frac{1}{\sqrt{k}} \Phi \tilde{\mathbf{a}}_i$ where $\phi_{lm} = e^{-2\pi i \xi_l m / M}$
 1636 and $\xi_l \sim \text{Uniform}(0, 1)$, satisfy:

- 1637 (1) $\mathbb{E}[\langle \mathbf{b}_i, \mathbf{b}_j \rangle] = \langle \tilde{\mathbf{a}}_i, \tilde{\mathbf{a}}_j \rangle$
 1638 (2) $\text{Var}(\langle \mathbf{b}_i, \mathbf{b}_j \rangle) = \frac{1}{k} \left(\|\tilde{\mathbf{a}}_i\|^2 \|\tilde{\mathbf{a}}_j\|^2 - 2 \sum_{m=1}^M \tilde{a}_{im}^2 \tilde{a}_{jm}^2 \right)$

1640 **PROOF.** Let $X = \langle \mathbf{b}_i, \mathbf{b}_j \rangle = \mathbf{b}_j^* \mathbf{b}_i = \sum_{l=1}^k \overline{b}_{j,l} b_{i,l}$.

1641 **Part 1: Expectation.** We begin by analyzing a single component
 1642 l of the sum. By definition:

$$\begin{aligned} b_{i,l} &= \frac{1}{\sqrt{k}} \sum_{m=1}^M \phi_{lm} \tilde{a}_{im} = \frac{1}{\sqrt{k}} \sum_{m=1}^M e^{-2\pi i \xi_l m / M} \tilde{a}_{im}, \\ \overline{b}_{j,l} &= \frac{1}{\sqrt{k}} \sum_{n=1}^M \overline{\phi_{ln}} \tilde{a}_{jn} = \frac{1}{\sqrt{k}} \sum_{n=1}^M e^{2\pi i \xi_l n / M} \tilde{a}_{jn}. \end{aligned}$$

1644 Their product is:

$$\overline{b}_{j,l} b_{i,l} = \frac{1}{k} \sum_{m=1}^M \sum_{n=1}^M e^{-2\pi i \xi_l (m-n) / M} \tilde{a}_{im} \tilde{a}_{jn}.$$

1645 Taking the expectation with respect to the random variable ξ_l :

$$\begin{aligned} \mathbb{E}_{\xi_l} [\overline{b}_{j,l} b_{i,l}] &= \frac{1}{k} \sum_{m=1}^M \sum_{n=1}^M \tilde{a}_{im} \tilde{a}_{jn} \mathbb{E}_{\xi_l} [e^{-2\pi i \xi_l (m-n) / M}] \\ &= \frac{1}{k} \sum_{m=1}^M \sum_{n=1}^M \tilde{a}_{im} \tilde{a}_{jn} \int_0^1 e^{-2\pi i t (m-n) / M} dt. \end{aligned}$$

1646 The value of the integral is:

$$\int_0^1 e^{-2\pi i t (m-n) / M} dt = \begin{cases} 1 & \text{if } m = n, \\ 0 & \text{if } m \neq n. \end{cases}$$

1647 Thus, the double sum collapses to the case $m = n$:

$$\mathbb{E}_{\xi_l} [\overline{b}_{j,l} b_{i,l}] = \frac{1}{k} \sum_{m=1}^M \tilde{a}_{im} \tilde{a}_{jm} = \frac{1}{k} \langle \tilde{\mathbf{a}}_i, \tilde{\mathbf{a}}_j \rangle.$$

1648 Since the ξ_l for $l = 1, \dots, k$ are independent and identically dis-
 1649 tributed (i.i.d.), the expectation of the full inner product is:

$$\begin{aligned} \mathbb{E}[X] &= \mathbb{E} \left[\sum_{l=1}^k \overline{b}_{j,l} b_{i,l} \right] \\ &= \sum_{l=1}^k \mathbb{E}_{\xi_l} [\overline{b}_{j,l} b_{i,l}] \\ &= \sum_{l=1}^k \frac{1}{k} \langle \tilde{\mathbf{a}}_i, \tilde{\mathbf{a}}_j \rangle = \langle \tilde{\mathbf{a}}_i, \tilde{\mathbf{a}}_j \rangle. \end{aligned}$$

1650 This proves the first statement.

1652 **Part 2: Variance.** The variance of the complex random variable
 1653 X is defined as $\text{Var}(X) = \mathbb{E}[|X|^2] - |\mathbb{E}[X]|^2$. We already have
 1654 $|\mathbb{E}[X]|^2 = |\langle \tilde{\mathbf{a}}_i, \tilde{\mathbf{a}}_j \rangle|^2$.

1655 We now compute $\mathbb{E}[|X|^2]$:

$$\mathbb{E}[|X|^2] = \sum_{l=1}^k \sum_{l'=1}^k \mathbb{E} \left[\overline{b}_{j,l} b_{i,l} \overline{b}_{j,l'} b_{i,l'} \right].$$

1656 We analyze this double sum by considering the cases where $l = l'$
 1657 and $l \neq l'$.

1658 **Case 1: $l = l'$.** (k terms). The contribution from these terms is:

$$\begin{aligned} S_1 &= k \cdot \mathbb{E}_{\xi_l} [|b_{i,l}|^2 |b_{j,l}|^2] \\ &= \frac{1}{k} \left(\|\tilde{\mathbf{a}}_i\|^2 \|\tilde{\mathbf{a}}_j\|^2 + |\langle \tilde{\mathbf{a}}_i, \tilde{\mathbf{a}}_j \rangle|^2 - 2 \sum_{m=1}^M \tilde{a}_{im}^2 \tilde{a}_{jm}^2 \right). \end{aligned}$$

1659 **Case 2: $l \neq l'$.** ($k(k-1)$ terms). By independence of ξ_l and $\xi_{l'}$, the
 1660 contribution is:

$$S_2 = k(k-1) \cdot \mathbb{E} \left[\overline{b}_{j,l} b_{i,l} \right] \mathbb{E} \left[b_{j,l'} \overline{b}_{i,l'} \right] = \frac{k-1}{k} |\langle \tilde{\mathbf{a}}_i, \tilde{\mathbf{a}}_j \rangle|^2.$$

1661 Combining the contributions:

$$\begin{aligned} \mathbb{E}[|X|^2] &= S_1 + S_2 \\ &= \frac{1}{k} \left(\|\tilde{\mathbf{a}}_i\|^2 \|\tilde{\mathbf{a}}_j\|^2 + |\langle \tilde{\mathbf{a}}_i, \tilde{\mathbf{a}}_j \rangle|^2 - 2 \sum_{m=1}^M \tilde{a}_{im}^2 \tilde{a}_{jm}^2 \right) \\ &\quad + \frac{k-1}{k} |\langle \tilde{\mathbf{a}}_i, \tilde{\mathbf{a}}_j \rangle|^2 \\ &= \frac{1}{k} \|\tilde{\mathbf{a}}_i\|^2 \|\tilde{\mathbf{a}}_j\|^2 + |\langle \tilde{\mathbf{a}}_i, \tilde{\mathbf{a}}_j \rangle|^2 - \frac{2}{k} \sum_{m=1}^M \tilde{a}_{im}^2 \tilde{a}_{jm}^2. \end{aligned}$$

1662 Finally, we obtain the variance:

$$\begin{aligned} \text{Var}(X) &= \mathbb{E}[|X|^2] - |\mathbb{E}[X]|^2 \\ &= \frac{1}{k} \|\tilde{\mathbf{a}}_i\|^2 \|\tilde{\mathbf{a}}_j\|^2 + |\langle \tilde{\mathbf{a}}_i, \tilde{\mathbf{a}}_j \rangle|^2 \\ &\quad - \frac{2}{k} \sum_{m=1}^M \tilde{a}_{im}^2 \tilde{a}_{jm}^2 - |\langle \tilde{\mathbf{a}}_i, \tilde{\mathbf{a}}_j \rangle|^2 \\ &= \frac{1}{k} \left(\|\tilde{\mathbf{a}}_i\|^2 \|\tilde{\mathbf{a}}_j\|^2 - 2 \sum_{m=1}^M \tilde{a}_{im}^2 \tilde{a}_{jm}^2 \right). \end{aligned}$$

1663 This proves the second statement. \square

1664 Proof of Theorem 3.2: CPSD Correlation Estimator

1665 **THEOREM (THEOREM 3.2: CPSD CORRELATION ESTIMATOR).** The
 1666 normalized real part of the CPSD provides a consistent estimator for
 1667 the Pearson correlation:

$$\hat{\rho}_{ij} = \frac{\Re(P_{ij})}{\sqrt{P_{ii}} \sqrt{P_{jj}}} \xrightarrow{p} \rho_{ij} \quad \text{as } k \rightarrow \infty.$$

1668 **PROOF.** Let $U_k = \Re(P_{ij})$, $V_k = P_{ii}$, and $W_k = P_{jj}$. Our es-
 1669 timator is a function of these variables, $\hat{\rho}_{ij} = g(U_k, V_k, W_k) =$
 $U_k / (\sqrt{V_k} \sqrt{W_k})$.

1670 From Theorem 3.1, we have $\mathbb{E}[P_{ij}] = \langle \tilde{\mathbf{a}}_i, \tilde{\mathbf{a}}_j \rangle$ and $\text{Var}(P_{ij}) =$
 $O(1/k)$. By Chebyshev's inequality, this implies that P_{ij} converges

1671

1672

1673

1674

1675

1676

1677

1678

1679

1680

1681

1682

1683

1684

1685

1686

1687

1688

1689

1690

1691

1692

1693

1694

1695

1696

1697

1698

1699

1700

1701

1702

1703

1704

1705

1706

1707

1708

1709

1710

1711

1712

1713

1714

1715

1716

1717

1718

1719

1720

1721

1722

1723

1724

1725

1726

1727

1728

1729

1730

1731

1732

1733

1734

1735

1736

1737

1738

1739

1740

1741 in probability to its mean, $P_{ij} \xrightarrow{P} \langle \tilde{\mathbf{a}}_i, \tilde{\mathbf{a}}_j \rangle$. The same holds for the
 1742 diagonal terms, $P_{ii} \xrightarrow{P} \|\tilde{\mathbf{a}}_i\|^2$ and $P_{jj} \xrightarrow{P} \|\tilde{\mathbf{a}}_j\|^2$.

1743 The function $g(u, v, w)$ is continuous for $v > 0, w > 0$, and the
 1744 functions $\Re(\cdot)$ and $\sqrt{\cdot}$ are also continuous. By the continuous mapping
 1745 theorem, the sequence of random vectors (U_k, V_k, W_k) converges in probability to the constant vector
 1746 $(\langle \tilde{\mathbf{a}}_i, \tilde{\mathbf{a}}_j \rangle, \|\tilde{\mathbf{a}}_i\|^2, \|\tilde{\mathbf{a}}_j\|^2)$. Applying the continuous function g to this sequence, we get:
 1747

$$1748 \hat{\rho}_{ij} = g(U_k, V_k, W_k) \xrightarrow{P} g(\langle \tilde{\mathbf{a}}_i, \tilde{\mathbf{a}}_j \rangle, \|\tilde{\mathbf{a}}_i\|^2, \|\tilde{\mathbf{a}}_j\|^2) = \frac{\langle \tilde{\mathbf{a}}_i, \tilde{\mathbf{a}}_j \rangle}{\|\tilde{\mathbf{a}}_i\| \|\tilde{\mathbf{a}}_j\|} = \rho_{ij}.$$

1751 This proves that $\hat{\rho}_{ij}$ is a consistent estimator for ρ_{ij} . \square

1753 Justification for Theorem 4.1: OMP Sparse 1754 Recovery Guarantee

1755 **THEOREM (THEOREM 4.1: OMP RECOVERY GUARANTEE).** *If the*
 1756 *sensing matrix Ψ satisfies the Restricted Isometry Property (RIP) of*
 1757 *order $S + 1$ with a sufficiently small constant δ_{S+1} , then the solution*
 1758 *$\hat{\mathbf{x}}_i$ produced by the Orthogonal Matching Pursuit (OMP) algorithm*
 1759 *after S steps from the measurement model $\mathbf{z}_i = \Psi \boldsymbol{\rho}_i + \mathbf{e}$ satisfies:*

$$1760 \|\hat{\mathbf{x}}_i - \boldsymbol{\rho}_i\|_2 \leq C_1 \|\mathbf{e}\|_2 + C_2 \frac{\|\boldsymbol{\rho}_i - \boldsymbol{\rho}_i^S\|_1}{\sqrt{S}},$$

1761 where \mathbf{e} is the noise vector, $\boldsymbol{\rho}_i^S$ is the best S -term approximation of $\boldsymbol{\rho}_i$,
 1762 and C_1, C_2 are constants.

1763 **JUSTIFICATION.** The guarantee for the OMP algorithm relies on
 1764 two key properties of the compressed sensing framework:

1765 **1. RIP of the Sensing Matrix Ψ :** The matrix Ψ is constructed
 1766 by normalizing the columns of the embedding matrix \mathbf{B} (i.e., $\psi_j =$
 1767 $\mathbf{b}_j / \|\mathbf{b}_j\|$). It is a well-established property that random matrices,
 1768 including those constructed from orthonormal systems like the
 1769 Fourier system, satisfy the RIP with high probability. This property
 1770 ensures that Ψ approximately preserves the lengths of all sparse
 1771 vectors, which is the crucial geometric condition that allows greedy
 1772 algorithms like OMP to succeed. For the RIP to hold, the number of
 1773 measurements k must be sufficiently large relative to the sparsity
 1774 level S , typically on the order of $k \gtrsim S \log(N/S)$.

1775 **2. Stable Recovery Guarantee for OMP:** When the sensing
 1776 matrix Ψ satisfies the RIP, the greedy nature of the OMP algorithm is
 1777 guaranteed to provide a stable and robust recovery of the underlying
 1778 signal $\boldsymbol{\rho}_i$. The algorithm works by iteratively identifying the column
 1779 of Ψ most correlated with the current residual and projecting the
 1780 measurement onto the space spanned by the selected columns. The
 1781 error bound reflects the algorithm's robustness to both noise and
 1782 signals that are not perfectly sparse:

- 1783 • **Noise Term ($C_1 \|\mathbf{e}\|_2$):** This term bounds the error's dependence
 1784 on the energy of the measurement noise \mathbf{e} . In our case,
 1785 this noise arises from the approximation error inherent in
 1786 the random Fourier embedding.
- 1787 • **Approximation Term ($C_2 \frac{\|\boldsymbol{\rho}_i - \boldsymbol{\rho}_i^S\|_1}{\sqrt{S}}$):** This term accounts
 1788 for the fact that the true correlation vector $\boldsymbol{\rho}_i$ may not be
 1789 perfectly sparse, but only *compressible* (i.e., its sorted coeffi-
 1790 cients decay rapidly). The recovery error degrades gracefully
 1791 with how well the signal can be approximated by its largest
 1792 S components.

1793 The constants C_1, C_2 depend only on the RIP constant δ_{S+1} . By en-
 1794 suring our embedding dimension k is sufficiently large, we meet the
 1795 conditions for this theorem, enabling reliable recovery of the sig-
 1796 nificant correlations for each term via the computationally efficient
 1797 OMP algorithm. \square

1798

1799

1800

1801

1802

1803

1804

1805

1806

1807

1808

1809

1810

1811

1812

1813

1814

1815

1816

1817

1818

1819

1820

1821

1822

1823

1824

1825

1826

1827

1828

1829

1830

1831

1832

1833

1834

1835

1836

1837

1838

1839

1840

1841

1842

1843

1844

1845

1846

1847

1848

1849

1850

1851

1852

1853

1854

1855