2.2 Four Components in This Study

2.2.1 Kano Diagrams

Figure 1 illustrates eight Kano diagrams, where node A01 represents the primary author (i.e., the most prolific contributor), while A02 and A03 correspond to the second- and third-highest output authors, respectively. The coordinate system and visual representation are defined as follows:

- The horizontal axis indicates the number of co-authored publications with others (excluding self-pairing), reflecting the collaborative strength of the author.
- The vertical axis represents the total publication count (including both single and co-authored works), and thus, it is always greater than or equal to the horizontal axis value.

To define leadership roles, a dual-parabola boundary and a central circular "zone of indifference" are applied. Four types of leadership roles are classified based on node positioning:

When A01 exclusively occupies the upper-right quadrant (indicating both high productivity and strong collaboration), and the relative distance between A02 and A03 is considered, five leadership levels are defined:

- **8-star leadership** (Figure 1A; AAC ≥ 0.7)
- **6-star leadership** (Figure 1C; AAC ≥ 0.6)
- **5-star leadership** (Figure 1E; AAC ≥ 0.5)
- **Weak leadership** (Figure 1G; AAC < 0.5)
- **Non-leadership** (Figure 1H; node falls outside both high axes)

When multiple nodes appear in the upper-right quadrant (e.g., Figure 1B and 1D), indicating the presence of several potential leaders, they are designated as "quasi" or "sub" leaders (e.g., Sub-3-star or Sub-5-star leadership), depending on their relative positions and AAC values.

2.2.2 ACC Calculation Related to Circular Radius

The AAC is calculated using the following formulas[7,8]:

$$AAC = \frac{\gamma}{1+\gamma}, (1)$$

$$\gamma = \frac{(\frac{r1}{r2})}{(\frac{r2}{r3})}, (2)$$

Alternatively, when r1=1 (due to coordinate standardization positioning A01 at the upper-right corner), and assuming r3>0 and r2 lies between r1 and r3, γ can also be expressed as:

$$\gamma = \frac{r3}{r_2^2+r3}, (3)$$

When AAC = 0.7, the following relationship is derived:

$$r2 = \sqrt{\frac{3}{7}} \times r3,\ (4)$$

After standardizing the coordinates (with values ranging from –1 to 1 on both axes), r1, r2, and r3 represent the standardized distances of the top three members (A01, A02, and A03), respectively. The parameter γ is the odds ratio. Notably, an AAC of 0.7 corresponds to an odds ratio of approximately 2.34, indicating two-fold odds:

$$\text{Odds} = \frac{P}{(1-P)} = \frac{AAC}{(1-AAC)},(5)$$

As r2 varies from r1=1 toward r3=0.1, AAC ranges between 0 and 1 based on Equation (2). When A02 is closer to A03, AAC increases; when A02 is closer to A01, AAC decreases.

To construct matching dual parabolas and the central circular zone, the mean coordinate is used as the center, and the radius is defined based on Criterion 1:

$$\text{Criterion 1}:\begin{cases} Let{:}\,\mu = (\bar{x},\bar{y}) \\ Let{:}\,r_{init} = \sqrt{(\bar{x})^2 + (\bar{y})^2} \\ When\ AAC \geq 0.7{:} \\ r = \max(r_{init}, r1, r2) \\ \text{When AAC} < 0.7: \\ \text{r} = \min(r_{init}, \text{r2} - 0.01) \end{cases}$$

The dual parabolas intersect the x-axis and the circular boundary, extending toward the upper-right and lower-left quadrants. Differential equations (see

Supplemental Digital Content 3) confirm that AAC decreases exponentially from 1 to 0, visualized as a steep slope (Figure 1, right; threshold radius = 0.2)—analogous to a scree plot in principal component analysis, where eigenvalues decrease in size and determine the number of contributing components. Similarly, AAC values ≥ 0.7 may be used to assess unidimensional dominance and identify leadership thresholds [7,8].

When the number of authors exceeds the number of institutions, institutional AAC tends to be lower than that of individual authors (see Supplemental Digital Content 3). Ideally, AAC ≥ 0.7 would indicate that only A01 appears in the upper-right quadrant.

Leader classification is then performed based on **Criterion 2**:

$$\text{Criterion 2}: \begin{cases} \text{Supper, if } ACC \geq 0.8 \\ \text{Quasi} - \text{Supper, if } ACC \geq 0.7 \\ Strong, if ACC \geq 0.6 \\ \text{Moderate, if } AAC \geq 0.5 \\ Weak, if\ ACC < 0.5 \\ \text{When A01 is not toppest on 2 axes,} \\ \text{Non leader} \end{cases}$$

===Figure 1 inserted here===

2.2.3 Basket Model Applied to Temporal Rectangular Data

Temporal rectangular data are frequently used in bibliometric analysis, such as country-based publications across years in Table 1. A basket model was applied to this study based on vertical co-word occurrences in columns.

===Table 1 inserted here===

The multiset of terms for column j (referred to a basket for each column) is:

$T_j = \bigcup_{i=1}^{n} \{x_i\}^{f_{ij}}$, (5)

This is a multiset where each term can appear multiple times based on its rounded score, where xi is the frequency of a term(e.g., China) in row i.

- T_j: Represents the **j-th basket** (e.g., a year, or observation unit).
- x_i: Refers to the **i-th term or item** (e.g., keyword, author, institution).
- f_{ij}: Indicates the **frequency or weight** of term x_i in basket T_j.
- $\{x_i\}^{f_{ij}}$: Denotes x_i **being repeated** f_{ij} **times** in basket T_j.
- $\bigcup$: Represents the **union** of all such weighted terms across i=1 to n.

This formula builds each basket T_j by collecting all terms x_i that appear in it, each repeated f_{ij} times, based on their frequency. This process is often implemented in R using the uncount() function to create a replicated term list suitable for constructing co-occurrence networks or correlation matrices. For example, If a column T1 includes: term A with frequency 3, and term B with frequency 2, Then:

$T_1 = \{A, A, A, B, B\}$, (6)

2.2.3.1 A single basket T_1

This basket allows pairwise co-occurrence to be computed correctly across weighted term repetitions. A single node we observed has a total count in (7):

$$Node_i = f_i = \sum_{j=1}^{1} f_{ij,}(7)$$

Edge for the node is the sum in (8) derived from (6):

$$w_i = f_i^2 + f_i \times \sum_{\substack{i \neq k \\ j=1}}^{n} f_k,$$

$$w_i = f_i \times \sum_{\substack{i \neq k \\ j=1}}^{n} f_k, \text{(8)}$$

Where i: index of the current item (row); k: index of all other items (also rows); n: total number of nodes (rows); f_i: frequency of $Node_i$ across all columns (e.g., sum of row i); in (8), the first f_i^2 represents self-contribution and the second term is so-called **co-**occurrence contribution with all other items.

We can prove that corr.$(Node_i, w_i) \approx 1$ in a single column:

Let: $f_i = Node_i$

- F=$\sum_k f_k$ is total number of term instances in a column.
- Then: $wi = f_i \cdot (F - f_i) = f_i \cdot F - f_i^2$
- $w_i \propto f_i$ (i.e., $w_i = \alpha f_i$, α is from the symbol ∝ : "is proportional to", similar to the regression line of y=a+bx) ⇒high frequency → high total edge weight, which means t**he value of w_i increases in proportion to** f_i. In other words, If fi doubles, then w_i approximately doubles. They are linearly related, but not necessarily equal — the actual relationship may be: w_i=α·f_i, where α is a

proportional constant. This explains why the Pearson correlation between them is high — often close to 1. Thus, $corr(f_i, w_i) = corr(f_i, f_i \cdot F) = 1$

The multiplication by a constant (here, F) preserves perfect linearity and correlation in terms of one item and total terms involved.

2.2.3.2 Generalized Expression for Multiple Baskets

$$Node_i = \sum_{j=1}^{L} f_{ij,}(7)$$

Edge for the node is the sum in (8) derived from (6):

$$w_i = \sum_{j=1}^{L} (f_{ij} \times \sum_{\substack{i \neq k \\ k=1}}^{n} f_{kj})$$

$$= \sum_{j=1}^{L} f_{ij} \times (\sum_{k=1}^{n} f_{kj} - f_i) = \sum_{j=1}^{L} f_{ij} \times F_j - f_i^2 = f_i \times F_j - f_i^{\,2}, (8)$$

So unless all F_j are equal (i.e., each basket has the same total size), $w_i \ not \propto f_i$ perfectly. However, if the basket sizes F_j are roughly constant as vertical columns shown , then $w_i \approx \bar{F} \times \mathrm{f}_i$, where $\bar{F}$=$\frac{1}{L}\sum_{j=1}^{L} F_j$

The Pearson correlation between w_i andf_i is defined as:

$corr(f_i, w_i) = \frac{Cov(f_i, w_i)}{\sigma_{fi} \times \sigma_{wi}}$

If $w_i \approx \bar{F} \times \mathrm{f}_i(or\ w_i = \alpha f_i\ )$, Then

Covariance: Cov(f_i, w_i)= $\mathrm{Cov}(f_i, \alpha f_i) = \alpha \times \mathrm{Var}(f_i) =\approx \bar{F} \times \mathrm{Var}(f_i)$

Standard deviations: $\sigma_{wi} \approx \bar{F} \times \sigma_{fi}$

So, $corr(f_i, w_i) = \frac{\bar{F} \times \mathrm{Var}(f_i)}{\sigma_{fi} \times \overline{(\bar{F}} \times \sigma_{fi})} = \frac{\bar{F} \times \mathrm{Var}(f_i)}{\bar{F} \times \sigma_{fi}^2} = 1$

Thus, under mild conditions — especially if basket sizes F_j are not wildly different — the total edge weight w_i is approximately linear in f_i , and therefore:

$$corr(f_i, w_i) \approx 1, and\ corr(f_i, F) \leq 1$$

As such, equality holds when all baskets have equal size, and less than 1 when there is variability in basket sizes. Otherwise, $Cov(f_i, w_i) \approx 0$ makes $corr(f_i, w_i) \approx 0$ and lone wolves unusually exist in network.

2.2.3.3 Top-n Selection Within the Multi-Basket Model

In each basket T_j, only the top n terms (based on frequency or score) are retained—similar to performance evaluations where the highest and lowest judge scores are excluded to ensure fairness. The top-n selection strategy in this study emphasizes high-ranking leaders by assigning greater weight to dominant terms within each basket.

$f_i^{(top)}$=$\sum_{J=1}^{l} f_{iJ}^{(top)}$ means total frequency of term i under top-n filtering.

$w_i^{(top)}$denotes edge weight of term i under top-n selection.

$$f_i^{(top)}=\begin{cases} f_{ij}, if\ i \in Top_n(T_j) \\ 0, \qquad otherwise \end{cases}$$

From (8), $w_i^{(top)} = \sum_{j=1}^{L} f_{ij}^{(top)} \times \sum_{k=1}^{n} f_{kj}^{(top)}$=$\sum_{j=1}^{L} f_{ij}^{(top)} \times F_j^{top}$

This mirrors the full-basket formula, but uses sparse(partial) selection. F_j^{top} denotes total weight in basket j after top-n selection. In this study, top-n/2 is applied.

- Terms that frequently appear among the top-n across baskets will have higher $f_i^{(top)}$ and higher $w_i^{(top)}$.
- Terms never ranked in the top-n will have $f_i^{(top)}$=0and contribute nothing to co-occurrence.
- The top-n strategy effectively filters noise and emphasizes strong signal terms — those dominant within each basket.

Therefore, the structure of $w_i^{(top)}$ preserves proportionality to $f_i^{(top)}$ (just like in the full model), but on a filtered dataset. If a term is frequently top-ranked, then:

$$w_i^{(top)} \approx f_i^{(top)} \times \overline{F^{(top)}} \Rightarrow \text{corr}(f_i^{(top)}, w_i^{(top)}) \approx 1$$

So: top-n selection doesn't break the correlation structure, but sharpens it by focusing on dominant contributors. In the multi-basket model, the top-n selection strategy creates a truncated co-occurrence network where only dominant terms in each basket are considered. Despite this reduction, the edge weight $w_i^{(top)}$ remains proportional to $f_i^{(top)}$, preserving the high correlation corr($w_i^{(top)}, w_i^{(top)}$)≈1. This confirms that the top-n strategy is valid and effective for highlighting leading terms in bibliometric and co-word network models.

All bubbles in the Kano diagram will appear along a trajectory from the lower-left to the upper-right corner when the basket model is combined with a Top-n/2(i.e., a half of terms) selection strategy.

The R script[22] to draw the KDAAC is provided with an MP4 video[23]

2.2.3.4 Traditional Co-Word Analysis

In contrast, the traditional pairwise relationships between first and corresponding authors are organized into a matrix $[c_{ij}]$ (Criterion 1), where each cell c_{ij} represents the number of co-authorships. The diagonal elements indicate self-relations.

In the Kano diagram, the x-axis values $\boldsymbol{w_i}$ and y-axis values $\boldsymbol{w_i^*}$ exhibit a linear relationship, as shown in the final row of Criterion 3.

$$
Criterion\ 3\text{:}\begin{cases}
[f_{ij}]\text{: Co} - \text{authorship frequency matrix} \\
\text{n: Number of nodes} \\
w_i\text{: Total co} - \text{authorships of node } \mathbf{i} \text{ excluding self} \\
w_i^*\text{: Total co} - \text{authorships including self} \\
f_i\text{: Self} - \text{occurrence} \\
w_i = \sum_{\substack{i\neq j\\ j=1}}^{n} c_{ij} \\
w_i^* = \sum_{j=1}^{n} c_{ij} = f_i + w_i \\
w_i^* \geq w_i \\
When\ w_i \propto fi \text{ : a proportional relationship holds} \\
w_i = \alpha \times f_i \text{ : then a linear relationship exists} \\
w_i^* = f_i + w_i = f_i + \alpha \times f_i \\
\text{Otherwise, } corr(f_i\,, w_i) \text{is weak, but} \\
corr(f_i\,, w_i^*) \text{ is strong:} \\
\text{it suggests the presence of ``}\mathbf{lone - wolf}\text{'' } \mathbf{researcher} \\
\text{(i.e., authors who are both first and corresponding authors} \\
\text{in most of their publications)}
\end{cases}
$$

The way to execute the traditional

The R script[24] to draw the KDAAC under the traditional co-word network analysis is also provided.

2.2.4 FLCA Algorithm and Sankey Diagrams

Using the follower-leading clustering algorithm(FLCA)[25], followers with lower $\boldsymbol{w_i^*}$ values are associated only with the one leader with whom they have the strongest connection. A node becomes a leader if it has multiple followers, unless it is the overall top node (i.e., node A01 as described in Section 2.2.2).

Sankey diagrams applied to bibliographical studies[26-29] uses a flow-based representation to emphasize primary relationships while maintaining a clear and structured view of collaborations between elements through the FLCA[25] and Sankeymatic software[30]. Node size corresponds to publication count; edge thickness represents the strength of collaboration or co-occurrence. Colors indicate distinct clusters generated by FLCA. The left-to-right layout reflects the rank ordering of node prominence.

2.3 Method for Drawing the Kano Diagram

To draw the Kano diagram, the user must paste the rectangular data and run the provided R scripts [22, 24]. Pressing the **Enter** key will immediately generate the Kano diagram, as demonstrated in the video tutorial provided at the link[23].

According to Criterion 1, only the Top 1 node (A01) will lie outside the circle if AAC ≥ 0.7; otherwise, multiple nodes may appear outside the circle.