Visualizing Ambiguity: A Grammar of Graphics Approach to Resolving Numerical Ties in Parallel Coordinate Plots

Comprehensive Exam Presentation

Denise Bradford

University of Nebraska–Lincoln

November 2025

Introduction

The Fundamental Challenge

Parallel Coordinate Plots (PCPs): Map n-dimensional data onto two-dimensional displays using parallel vertical axes (Inselberg 1985; Wegman 1990)

The Core Problem:

- Multiple observations with identical numerical values create perfectly overlapping polylines
- "Visual collision" - a single visible line may represent 1 or 1,000 observations
- No mechanism to distinguish density or trace individual observations

Critical Impact:

- Density information completely lost
- Substructure within tied groups invisible
- Exploratory analysis fundamentally compromised
- Cluster identification impossible when ties obscure structure

Research Context & Gap

Origins of Parallel Coordinates:

- Inselberg (1985): Geometric duality properties
- Wegman (1990): Statistical data analysis, high-dimensional visualization

Evolution:

- 1990s-2000s: Overplotting recognized as critical problem (Johansson and Forsell 2016)
- 2003: Hammock plots introduce alternative approach (Schonlau 2003)
- 2023: ggpcp package - Grammar of Graphics framework (VanderPlas et al. 2023)

The Persistent Gap:

- Categorical ties: Elegantly solved through hierarchical sorting
- **Numerical ties: No systematic, reproducible solution**

The ggpcp Package Foundation

Key Innovation: Separation of Concerns (VanderPlas et al. 2023)

ggpcp divides PCP creation into distinct modules:

1. **Variable Selection** (`pcp_select`): Choose and order dimensions
2. **Axis Scaling** (`pcp_scale`): Normalize or transform scales
3. **Tie Resolution** (`pcp_arrange`): Handle overlapping values

Categorical Tie-Breaking Success:

- Hierarchical sorting ("from-left", "from-right") minimizes line crossings
- Equispaced distribution ensures visibility
- Enables individual observation tracing

The Missing Piece:

- Categorical ties: ✓ Solved
- **Numerical ties: × Unsolved until now**

Key Terminology and Definitions

Numerical Ties

Definition:
Numerical ties occur when multiple observations share identical numerical values on one or more dimensions.

Sources of Ties:

- Measurement precision limitations (rounding)
- Discrete measurement instruments
- Natural clustering at specific values
- Data collection procedures

Distinction from Categorical Ties:

- Categorical: Expected within groups
- Numerical: May arise from various data characteristics
- Both require treatment but through different mechanisms

Jittering and Determinism

Jittering:
Controlled displacement technique that adds small offsets to prevent visual overlap

- Magnitude controlled by parameter ϵ (epsilon)
- Purpose: Visual separation, not data modification
- Constraint: Displacement must not mislead about underlying values

Deterministic vs. Stochastic Methods:
Deterministic methods produce identical results given identical input

- Examples: Halton sequences, Sunflower patterns
- Advantage: Scientific reproducibility
- Requirement: No random number generation

Stochastic methods incorporate randomness

- Example: Standard random jitter with `runif()` or `rnorm()`
- Disadvantage: Non-reproducible, may create artifacts

"Best For" Interpretation

Definition:

When we state a method is "best for" a scenario, we mean it optimally balances:

1. **Task Requirements**: What the analyst needs to accomplish
2. **Data Characteristics**: Properties affecting method performance
3. **Performance Criteria**: Quantitative and qualitative measures
4. **Trade-off Optimization**: Aligning method strengths with scenario needs

Example Applications:

- **Halton** is best for precision-critical work requiring mathematical guarantees
- **Sunflower** is best for general purpose use balancing uniformity and aesthetics
- **Hammock plots** are best for density visualization over individual tracing

Important Caveat:

Recommendations based on theoretical analysis, empirical evaluation, and practical experience. Users should consider their specific context.

Background and Motivation

The Problem: Visual Overlap

Three Critical Issues:

1. **Visual Collision**: Perfect overlap masks observation count
2. **Density Information Loss**: Cannot distinguish 1 from 100 observations
3. **Structural Occlusion**: Sub-clusters and patterns hidden

Impact on Analysis:

- Cluster identification compromised (Blumenschein et al. 2020)
- Outlier detection impossible
- Pattern tracing fundamentally limited

Existing Solutions: Categorical Ties in ggpcp

ggpcp's Approach:
Hierarchical sorting through `pcp_arrange(data, method, space)`:

- **"from-left"**: Tie breaks determined by values from left
- **"from-right"**: Tie breaks determined by values from right
- **space parameter**: Controls spacing between categorical levels (default: 5%)

Key Benefits:

- Reduces line crossings
- Enables observation tracing
- Provides "external cognition" reducing cognitive load

Why This Works for Categories:

- Discrete nature allows hierarchical ordering
- Equispaced distribution natural for categories
- Visual similarity to Parallel Sets when dense

Alternative Approach: Hammock Plots

Hammock Plot Strategy (Schonlau 2003; Schonlau and Yang 2024):
Uses boxes (parallelograms) where width $\propto$ number of observations

How Hammock Plots Handle Ties:

- **Aggregation through width**: Multiple tied observations $\rightarrow$ wider boxes
- **Density through visual magnitude**: Box width directly encodes frequency
- **No separation needed**: Aggregation eliminates occlusion problem

Advantages:

- Explicit density representation
- No occlusion with different frequencies
- Seamless handling of mixed variable types

Limitations:

- Loss of individual observation tracing
- Increased white space with frugal spacing
- Binning often required for continuous variables

Comparison: Hammock vs. GPCP

Feature	Hammock Plot	GPCP (ggpcp)
Between numerical variables	Constant-width boxes	Lines (overlap)
Categorical to numerical	Constant-width boxes	Triangular shapes
Individual tracing	Requires highlighting	Natural
Density visualization	Explicit (width)	Implicit (overlap)
Small datasets	Less detailed	Shows individuals
Large datasets	Clearer aggregation	Appears as areas

When Hammock Plots Excel:

- Emphasis on bivariate relationships
- Datasets with many observations per value
- Focus on aggregate patterns over trajectories

Why ggpcp Needs a Different Solution

Complementary Approaches:
Despite hammock plots' success, ggpcp requires numerical tie-breaking:

1. **Preservation of Individual Traceability**: Core ggpcp feature
2. **Grammar of Graphics Philosophy**: Position adjustment fits naturally
3. **Flexibility**: Users choose aggregation or separation based on needs
4. **Small to Medium Datasets**: Where individual observations matter

Integration Opportunity:
Future work could combine:

- Jittering for position separation (x-coordinate)
- Width encoding for density (visual weight)
- Benefits: Individual traceability + explicit density

Design Requirements

Four Core Principles

1. Determinism

Requirement: Identical input must produce identical output

- Rationale: Essential for scientific reproducibility (R. D. Peng 2011)
- Implication: Rules out standard random jitter

2. Uniformity

Requirement: Even distribution within displacement interval with minimal clustering

- Rationale: Faithful density representation, minimize artificial patterns
- Implication: Deterministic methods can guarantee uniform coverage

Four Core Principles (continued)

3. Perceptual Validity

Requirement: Balance two competing needs

- Small enough to maintain value integrity
- Large enough for visual separation
- Rationale: Users must trust displaced values (Ware 2012)

4. Scalability

Requirement: Handle 2 to 10,000+ observations per tie group efficiently

- Rationale: Real-world datasets vary enormously
- Interactive exploration requires real-time performance
- Implication: Algorithms must have favorable complexity $O(n)$ per group

Empirical Evidence and Perceptual Foundations

Perceptual Challenges in Parallel Coordinates

The Line Width Illusion:
Users perceive distance between parallel lines at right angles, not vertical distance (Hofmann and Vendettuoli 2013)

- Part of Müller-Lyer family of illusions
- Affects interpretation of separation in PCPs

Implication for Tie Resolution:

- Users judge separation based on orthogonal (perpendicular) distance
- Epsilon parameter must account for perceptual bias
- May require different values depending on line angle

The Sine Illusion:
Equal-length vertical lines in sine wave pattern appear unequal (Day and Stecher 1991)

- Lines at peaks/troughs appear longer
- Regular patterns can trigger illusion
- Sunflower's golden angle helps prevent periodic patterns

Clutter and Overplotting: The Core Problem

Severity:
Even medium-sized datasets suffer from overplotting, resulting in displays too cluttered to perceive trends or structure (Johansson and Forsell 2016)

How Ties Exacerbate the Problem:

- Without ties: Overplotting from similar ranges
- With ties: Perfect overlap of multiple observations
- Result: Complete occlusion with no frequency information

Existing Clutter Reduction Approaches:

1. **Clustering-based**: Bands, envelopes, frequency representations
   - Limitation: Loss of individual tracing
2. **Transparency/density**: Alpha blending, density plots
   - Limitation: Fails with perfect overlap (ties)
3. **Our contribution**: Deterministic jittering
   - Resolves ties before rendering
   - Preserves individual traces
   - Complements other methods

Dimension Ordering Effects

Critical Importance:
Order and arrangement of dimensions crucial for PCP effectiveness (W. Peng, Ward, and Rundensteiner 2004; Blumenschein et al. 2020)

- Similar dimensions should be adjacent
- High impact on visualization quality
- Problem is NP-complete, requires heuristics

Interaction with Tie Resolution:
Dimension ordering affects which ties become visible:

```
Ordering 1: A - B - C
  - Ties between A-B highly visible
Ordering 2: A - C - B
  - Different tie patterns emerge
```

Implication:

- Tie detection must be axis-pair specific
- Evaluation must consider multiple orderings
- Integration with existing `pcp_select()` maintains flexibility

Cluster Identification Performance

Empirical Findings:

Recent studies evaluated cluster identification in PCPs (Holten and Wijk 2010; Blumenschein et al. 2020)

- Optimal configurations depend on task type and cluster characteristics
- Reordering strategies significantly impact performance

Relevance to Tie-Breaking:

Critical question: Does tie-breaking help or hinder cluster identification?

Potential Positive Effects:

- Reveals hidden clusters within tied groups
- Improves cluster boundary visibility
- Enables size estimation

Potential Negative Effects:

- Displacement might obscure tight clusters
- Cognitive load from visual complexity
- Poor jittering could suggest false clusters

Task-Dependent Performance

General Finding:
PCP effectiveness varies by task complexity (Johansson and Forsell 2016)

Task Categories for Evaluation:

Task	Difficulty	Expected Impact
Density Estimation	Simple	Large improvement with clear separation
Cluster Identification	Medium	Reveals structure, sensitive to method
Outlier Detection	Complex	Essential for individual line tracing

Integration with Existing Evidence:

- Use validated tasks from literature
- Control for confounds (ordering, size, design)
- Measure multiple outcomes (accuracy, time, confidence)
- Compare against baselines (No Jitter, Random Jitter)

Practical Design Recommendations

Current Best Practices (Johansson and Forsell 2016; Blumenschein et al. 2020):

1. Manage visual clutter
2. Optimize dimension ordering
3. Consider perceptual factors
4. Support interaction
5. Use appropriate encodings

Our Extension - Adding a Sixth Principle:

6. **Resolve numerical ties deterministically:**
   - Apply jittering to prevent perfect overlap
   - Use low-discrepancy methods (Halton/Sunflower)
   - Integrate with existing capabilities
   - Maintain reproducibility

Implementation in ggpcp:

All six principles addressed through integrated approach

Mathematical Framework

Displacement Constraint

For each tied value v with n_{ties} observations:
Distribute points within displacement interval:

$$\left[v - \frac{\epsilon}{2}, v + \frac{\epsilon}{2}\right]$$

Key Parameter:

- ϵ: maximum displacement magnitude
- Typically 0.05–0.10 of axis range
- User-adjustable based on data characteristics and perceptual requirements

Optimization Goals:

1. Maximize minimum inter-point distance (prevent collision)
2. Minimize visual artifacts (avoid clustering and false patterns)
3. Maintain deterministic reproducibility (enable verification)
4. Achieve uniform coverage (faithfully represent density)

Three Deterministic Jittering Methods

Method Comparison Overview

Method	Theoretical Basis	Dimension	Scaling	Best For
Halton	Quasi-random sequences (Halton 1960)	Pure 1D	Constant	Uniform distributions
Sunflower	Phyllotaxis (Vogel 1979)	2D $\rightarrow$ 1D	Sublinear $(\sqrt{n})$	Aesthetic + performance
Intelligent	Golden ratio direct application	Hybrid 1D	Linear	Research comparison

All methods provide:

- Deterministic, reproducible output
- Theoretically-grounded distributions
- Computational efficiency $O(n)$ or $O(n \log n)$

Method 1: Sunflower Jitter - Biological Inspiration

Phyllotaxis: Nature's optimal packing solution
Sunflower seed arrangements follow evolutionary optimization (Vogel 1979)

The Golden Angle: $137.508° = 360° \times (2 - \phi)$
where $\phi = \frac{1+\sqrt{5}}{2} \approx 1.618$ is the golden ratio

Why This Angle Works:

- Most "irrational" number in continued fraction sense
- Ensures no radial alignment even after hundreds of iterations
- Optimal space-filling property

$$\phi = 1 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{1 + \cdots}}}$$

Sunflower Jitter - Mathematical Formulation

For observation j in tie group of size n_{ties}:

$$\text{angle}_j = (j-1) \times 137.508°$$

$$\text{radius}_j = \epsilon \times \sqrt{\frac{j-1}{n_{\text{ties}}}}$$

$$\text{displacement}_j = \text{radius}_j \times \cos(\text{angle}_j)$$

Key Features:

- **Square root scaling**: Maintains constant density as radius increases
- **Cosine projection**: Maps 2D polar $\rightarrow$ 1D linear displacement
- **Spiral structure**: Preserved even in 1D projection
- **Biomimetic**: Leverages millions of years of natural selection

Sunflower Jitter - Why Square Root Scaling?

Geometric Justification:

In a 2D disk:

- Circumference at radius r: $2\pi r$
- Number of points at radius r_j: proportional to j
- For constant area density: Need $\frac{dN}{dA} = \text{constant}$

Mathematical Derivation:

$$A \propto r^2 \implies N \propto r^2 \implies r \propto \sqrt{N}$$

Therefore: $r \propto \sqrt{j}$ balances linear growth in points with radial expansion

Distribution Quality:

- Near-optimal minimum separation (Vogel 1979)
- Low-discrepancy properties in 2D
- Aesthetically consistent across scales
- Progressively validated through evolution

Sunflower Jitter - Applications Beyond PCPs

Proven Applications:

- **Point cloud sampling**: Uniform distribution on discs/spheres
- **Sphere packing**: Near-optimal packing density
- **Texture synthesis**: Organic, non-repetitive patterns
- **Computer graphics**: Stratified sampling for ray tracing
- **Quasi-Monte Carlo methods**: Numerical integration

Why It Works Broadly:

The golden angle property—maximal incommensurability—creates optimal spacing in any radial system

Method 2: Halton Jitter - Beyond Pseudo-Randomness

The Random Number Problem:
Pseudo-random numbers inevitably cluster (birthday paradox)

Example: 100 random points on $[0, 1]$:

- Expected maximum gap: 0.05
- Expected minimum gap: 0.0001
- Creates misleading visual artifacts
- Density perception unreliable

Halton's Solution (Halton 1960):
Place each new point to systematically fill largest gaps using van der Corput sequence

- Deterministic (not random)
- Low-discrepancy (uniform space-filling)
- Number-theoretically constructed using prime bases
- Mathematically guaranteed coverage

Van der Corput Sequence Construction

Algorithm (base 2):

1. Take integer index i
2. Convert to binary
3. Reverse the binary digits
4. Interpret as binary fraction

i	Binary	Reversed	Decimal h_i
0	0	0	0.0
1	1	1	0.5
2	10	01	0.25
3	11	11	0.75
4	100	001	0.125
5	101	101	0.625

Pattern: Each point bisects the largest remaining gap

Halton Jitter - Mathematical Formulation

For observation i in tie group:

$$h_i = \text{VanDerCorput}(i, \text{base} = 2)$$

$$\text{displacement}_i = \epsilon \times (h_i - 0.5)$$

Centering around 0.5 creates symmetric bidirectional displacement

Theoretical Guarantees - Discrepancy Theory:
Star discrepancy measures uniformity (Niederreiter 1992):

$$D_n^* = \sup_{I \subseteq [0,1]} \left| \frac{\#\{x_i \in I\}}{n} - |I| \right|$$

- Random sequences: $D_n = O(n^{-1/2})$
- Halton sequences: $D_n = O(n^{-1} \log n)$ $\leftarrow$ near-optimal
- Optimal lower bound: $D_n = \Omega(n^{-1} \log n)$

Halton Jitter - Applications

Widely Used in:

- **Quasi-Monte Carlo integration**: Better convergence than random sampling
- **Computer graphics**: Anti-aliasing, global illumination
- **Ray tracing**: Sample generation for realistic rendering
- **Numerical analysis**: Multidimensional quadrature
- **Machine learning**: Hyperparameter search spaces

Higher-Dimensional Extensions:

For 2D applications (e.g., scatter plots):

- $x_i = \text{VanDerCorput}(i, 2)$ (base 2 for x-axis)
- $y_i = \text{VanDerCorput}(i, 3)$ (base 3 for y-axis)

Different prime bases for each dimension maintain low-discrepancy

Method 3: Intelligent Jitter - Novel Exploration

Design Motivation:

Research question: Can we apply golden ratio directly in 1D rather than through angular spacing?

Mathematical Formulation:

For observation j in tie group of size n_{ties}:

$$\text{angle}_j = (j-1) \times 2\pi \times 0.618$$

$$\text{displacement}_j = \epsilon \times \cos(\text{angle}_j) \times \frac{j-1}{n_{\text{ties}}}$$

Key Distinctions from Sunflower:

- Angle: Golden ratio $\times$ 2π (224.4°) vs. Golden angle (137.5°)
- Scaling: **Linear** (j/n) vs. Square root $(\sqrt{j/n})$
- Projection: 1D cosine modulation vs. 2D spiral $\rightarrow$ 1D

Intelligent Jitter - Failure Analysis

Initial Hypothesis:

Linear scaling would create "progressive reveal":

- Early observations: Small displacement (near true value)
- Later observations: Larger displacement (fill space)
- Intuitive interpretation: "early arrivals" cluster, "late arrivals" spread

Empirical Reality: Three Critical Problems

Problem 1: Excessive Displacement

For $n = 100$:

- Observation 1: 0% of ϵ
- Observation 50: 49% of ϵ
- Observation 100: 99% of $\epsilon \leftarrow$ Near boundary!

Why Intelligent Jitter Fails

Problem 2: Artificial Stratification

Linear scaling creates visible "layers" that don't represent data structure

- Visual appearance suggests distinct sub-groups
- These "clusters" are algorithmic artifacts
- Misrepresents uniform density as stratified distribution

Problem 3: Perceptual Distortion

Users interpret visual patterns as data patterns:

- Large gaps appear meaningful (but are artifacts)
- Density gradients suggest ordering (observations are exchangeable)
- Boundary concentration implies separation (all values are tied)

Root Cause:

Linear scaling violates uniformity and perceptual validity principles

Value of This Negative Result

Scientific Contributions:

1. **Design Pattern to Avoid**
   - Lesson: Linear displacement scaling creates misleading stratification
   - Implication: Future methods should use constant or sublinear scaling
2. **Golden Ratio Not Universal**
   - Lesson: Works in specific geometric contexts, not universally
   - Implication: Biomimetic approaches require careful adaptation
3. **Importance of Scaling Function**
   - Lesson: Scaling function as critical as distribution algorithm
   - Implication: Must consider angular distribution AND radial scaling together
4. **Empirical Validation Essential**
   - Lesson: Theoretical elegance practical effectiveness
   - Implication: User studies necessary even for mathematically motivated methods

Comparative Analysis

Dimensional Analysis

Method	Approach	Dimension	Projection
Halton	Pure 1D sequence	1D	None
Sunflower	2D spiral	2D $\rightarrow$ 1D	Cosine
Intelligent	1D with 2D-inspired modulation	Hybrid	Cosine

Scaling Behavior

Method	Scaling Function	Growth Rate	At $n = 50$, $j = 25$
Halton	Uniform distribution	Constant	$ 0.5 $
Sunflower	$\sqrt{j/n}$	Sublinear	$ 0.7 $
Intelligent	j/n	Linear	$ 0.5 $

Distribution Quality Metrics:

- **Minimum separation**: Halton guaranteed $O(1/n)$, highly predictable
- **Discrepancy**: Halton $O(n^{-1} \log n)$ near-optimal
- **Visual clustering**: Halton minimal, Sunflower slight central concentration, Intelligent severe stratification

Use Case Recommendations

Halton: Best for...

- Precision-critical applications (scientific publications)
- Maximum uniformity requirements
- Mathematical rigor and provable guarantees
- Large datasets with efficient performance

Sunflower: Best for...

- General purpose use (good balance of properties)
- Aesthetic presentations
- Exploratory analysis
- User preference (often preferred in studies)

Intelligent: Best for...

- Methodological research (comparison baseline)
- Educational examples (teaching what NOT to do)
- **DO NOT use for production visualizations**

Integration with ggpcp Package

Architectural Integration

ggpcp's Three Core Modules:

1. **Variable Selection** (`pcp_select`): Choose and order dimensions
2. **Axis Scaling** (`pcp_scale`): Normalize or transform scales
3. **Tie Resolution** (`pcp_arrange`): Handle overlapping values ← **EXTENDED**

Our Contribution:

Extends `pcp_arrange` to handle numerical ties alongside existing categorical tie-breaking

Design Philosophy:

- Maintains backward compatibility
- Follows Grammar of Graphics principles
- Integrates seamlessly with existing workflow

Proposed ggpcp Implementation

Function Signature:

```
pcp_arrange(
  data,
  method = c("from-left", "from-right",
             "halton", "sunflower", "intelligent"),
  space = 0.05,
  epsilon = NULL,
  numeric_ties = TRUE
)
```

New Parameters:

- `method`: Now includes "halton", "sunflower", "intelligent"
- `epsilon`: Maximum displacement for numerical ties
  - `NULL` (default): Auto-determined as 0.05 $\times$ axis range
  - Numeric value: User-specified displacement
- `numeric_ties`: Whether to apply jittering (default: `TRUE`)

Example Usage

```r
library(ggpcp)
library(dplyr)

# Basic usage with Sunflower (recommended default)
iris_plot <- iris %>%
  pcp_select(1:5) %>%
  pcp_scale(method = "uniminmax") %>%
  pcp_arrange(method = "sunflower") %>%
  ggplot() +
  geom_pcp()

# Halton for maximum uniformity
iris_halton <- iris %>%
  pcp_select(Sepal.Length:Species) %>%
  pcp_arrange(
    method = "halton",
    epsilon = 0.08,
    numeric_ties = TRUE
  ) %>%
  ggplot() +
  geom_pcp(aes(color = Species))
```

Mixed Categorical and Numerical Ties

```
# Handle both categorical and numerical ties
mixed_data %>%
  pcp_select(cat1, num1, cat2, num2) %>%
  pcp_arrange(
    method = "sunflower",  # Applied to numerical
    space = 0.05           # Applied to categorical
  )

# Method comparison
library(patchwork)

p_none <- iris %>% pcp_select(1:4) %>%
  pcp_arrange(method = "none") %>% plot_pcp()

p_halton <- iris %>% pcp_select(1:4) %>%
  pcp_arrange(method = "halton") %>% plot_pcp()

p_sunflower <- iris %>% pcp_select(1:4) %>%
  pcp_arrange(method = "sunflower") %>% plot_pcp()

(p_none | p_halton | p_sunflower) +
  plot_annotation(
```

Documentation Requirements

Function Documentation:

- Detailed explanation of each method
- Theoretical foundations and references
- When to use each method
- Parameter selection guidance
- Examples with multiple datasets

Vignettes:

1. "Handling Numerical Ties in ggpcp"
2. "Comparing Tie-Breaking Methods"
3. "Advanced Tie Resolution"
4. "Theory of Deterministic Jittering"

Visual Indicators:

Optional indicators showing:

- Which axes have tie-breaking applied
- Magnitude of epsilon used
- Number of tied observations per group

Research Questions and Methodology

Primary Research Question

How can the formal structure of the Grammar of Graphics be extended to systematically incorporate and evaluate methods for resolving numerical ties in parallel coordinate plots, and what is the quantifiable impact of these methods on the accuracy and efficiency of visual data analysis?

RQ1: Theory

How can the management of numerical ties be most effectively and coherently formalized within the layered grammar of graphics, building on the established ggpcp framework?

Methodology:

- Theoretical analysis of Grammar of Graphics structure
- Literature synthesis on position adjustments
- Specification of new grammatical element
- Integration with existing ggpcp architecture
- Formal documentation of tie-breaking grammar

Deliverables:

- Formal specification document
- Extended grammar notation
- Theoretical paper on biomimetic transformations
- Integration guidelines for ggpcp

RQ2: Methodology

What are the optimal algorithmic criteria for ordering and spacing tied data points to maximize visual clarity while preserving underlying data properties?

Methodology:

- Algorithm design and implementation in R
- Comparative analysis of distribution quality
- Computational performance benchmarking
- Parameter sensitivity analysis
- Edge case identification and handling

Evaluation Metrics:

- Minimum separation distance
- Discrepancy (uniformity measure)
- Computational complexity
- Memory efficiency
- Scalability testing

RQ3: Perception

How do different visualization strategies for numerical ties affect an analyst's ability to perform key visual tasks?

Study Design:

- Type: Within-subjects repeated measures
- Participants: 30-50 analysts (mixed expertise)
- Methods: No jitter, Random, Halton, Sunflower, Intelligent
- Tasks:
  - Density estimation
  - Cluster identification
  - Outlier detection
  - Pattern tracing

Dependent Variables:

1. Accuracy (absolute error from ground truth)
2. Completion time (seconds)
3. Confidence (self-reported 1-10 scale)
4. Preference (comparative ranking)

RQ3: Expected Hypotheses

Statistical Analysis:

- Repeated-measures ANOVA
- Bonferroni post-hoc tests
- Effect size calculations (partial η^2)
- Correlation analysis (accuracy vs. confidence)

Expected Hypotheses:

- **H1**: Halton and Sunflower $>$ Random $>$ No Jitter (accuracy)
- **H2**: Halton Sunflower $<$ Random $<$ No Jitter (time)
- **H3**: Sunflower Halton $>$ Random $>$ No Jitter (preference)
- **H4**: Intelligent performs poorly across all metrics

RQ4: Practice

Can a set of evidence-based heuristics be developed to guide practitioners in selecting the most appropriate numerical tie-breaking method for their specific data context?

Methodology:

- Synthesize findings from RQ1-3
- Develop decision tree/flowchart
- Validate with case studies
- Gather practitioner feedback
- Refine through iterative testing

Case Study Domains:

1. Bioinformatics: Gene expression data
2. Finance: Market data with discrete prices
3. Engineering: Sensor data with limited precision
4. Social Science: Survey responses with Likert scales
5. Climate Science: Model ensemble outputs

Implementation Roadmap

Phase 1: Algorithm Refinement (Winter 2025)

Tasks:

1. Finalize all three jittering implementations
2. Develop adaptive epsilon selection
3. Optimize computational performance
4. Complete test suite with edge cases
5. Benchmark against large datasets

Deliverables:

- Optimized R functions
- Unit tests with 100% coverage
- Performance benchmarks
- Technical documentation

Phase 2: ggpcp Integration (Spring 2026)

Tasks:

1. Extend `pcp_arrange()` function
2. Implement automatic tie detection
3. Add epsilon auto-determination
4. Create visual indicators
5. Write package vignettes
6. Prepare for CRAN submission

Deliverables:

- Updated ggpcp package
- Comprehensive documentation
- Three tutorial vignettes
- Package ready for CRAN

Phase 3: User Study (Spring-Summer 2026)

Tasks:

1. Obtain IRB approval (early Spring)
2. Develop study materials
3. Recruit participants
4. Conduct study sessions
5. Analyze results
6. Write empirical paper

Deliverables:

- IRB approval documentation
- Complete dataset
- Statistical analysis
- Empirical research paper draft

Phase 4: Case Studies & Writing (Summer 2026)

Tasks:

1. Apply to diverse real-world datasets
2. Gather practitioner feedback
3. Develop decision heuristics
4. Write dissertation chapters
5. Integrate all components
6. Prepare defense presentation

Deliverables:

- Five domain case studies
- Practitioner's guide
- Complete dissertation draft
- Defense presentation

Phase 5: Final Review and Defense (May-July 2026)

Tasks:

1. Committee review of dissertation
2. Incorporate feedback
3. Final revisions
4. Defense rehearsals
5. Dissertation defense

Deliverables:

- Final dissertation
- Successful defense
- Submitted for graduation

Expected Outcomes

Theoretical Contributions

1. **Grammar of Graphics Extension**
   - Formal specification of biomimetic transformations
   - Integration of natural optimization principles
   - New category of position adjustments
2. **Cross-Domain Algorithm Adaptation**
   - Phyllotaxis $\rightarrow$ data visualization
   - Quasi-random sequences $\rightarrow$ statistical graphics
   - Demonstrates value of interdisciplinary approaches
3. **Negative Result Documentation**
   - Intelligent jitter failure analysis
   - Design patterns to avoid
   - Methodological lessons for future research

Methodological Contributions

1. **Three Novel Algorithms**
   - Halton jitter for PCPs
   - Sunflower jitter for PCPs
   - Intelligent jitter (with failure analysis)
2. **Comparative Framework**
   - Systematic evaluation criteria
   - Quantitative metrics
   - Perceptual assessment methods
3. **Implementation Quality**
   - Production-ready R code
   - Comprehensive testing
   - Extensive documentation

Practical Contributions

1. **ggpcp Package Enhancement**
   - Complete tie-handling solution
   - Categorical + numerical ties
   - Unified grammar interface
2. **User Guidance**
   - Evidence-based selection heuristics
   - Interactive decision tools
   - Tutorial materials
3. **Real-World Impact**
   - Improved exploratory data analysis
   - More accurate pattern detection
   - Better density visualization

Empirical Contributions

1. **User Study Results**
   - Quantitative performance data
   - Perceptual effectiveness measures
   - Preference rankings
2. **Case Study Collection**
   - Diverse domain applications
   - Best practices examples
   - Common pitfall documentation
3. **Benchmark Dataset**
   - Performance comparisons
   - Scalability testing
   - Reference implementations

Broader Implications

Beyond Parallel Coordinates

The methods generalize to other visualization contexts:

1. 2D Scatter Plots
Problem: Overplotting with tied values
Solution: Full 2D Sunflower and 2D Halton

2. Time Series Visualization
Problem: Multiple series with identical values at time points
Solution: Vertical displacement using deterministic methods

3. Network Visualization
Problem: Node positioning with spatial constraints
Solution: Optimal space-filling using golden angle principles

General Principle

Wherever random jitter is currently used, deterministic low-discrepancy alternatives should be considered.

Benefits:

- Reproducibility for scientific publications
- Better distribution quality
- Elimination of clustering artifacts
- Theoretical guarantees on uniformity

Broader Impact:
Establishes design patterns applicable across visualization domains and establishes principles for perceptually-valid position adjustments

Timeline

Timeline to Dissertation Defense

Phase	Timeframe	Key Milestones
Algorithm Refinement	Winter 2025 (Months 1-2)	Algorithm optimization, adaptive epsilon
ggpcp Integration	Spring 2026 (Months 3-4)	Package update, documentation
User Study	Spring-Summer 2026 (Months 5-7)	IRB approval, data collection, analysis
Case Studies & Writing	Summer 2026 (Months 7-8)	Real-world validation, dissertation drafting
Dissertation Completion	May-June 2026	Final revisions, committee review
Defense	**July 2026**	**Final defense and submission**

Conclusion

Summary of Contribution

The Problem:
Numerical ties in parallel coordinate plots create severe visual occlusion, preventing:

- Density visualization
- Individual observation tracing
- Cluster identification
- Pattern detection

The Solution:
Systematic approach using three deterministic jittering methods:

- **Halton**: Quasi-random sequences with mathematical guarantees
- **Sunflower**: Biomimetic approach leveraging natural optimization
- **Intelligent**: Exploratory method demonstrating negative results

The Impact:
Complete framework for tie resolution in PCPs, extending Grammar of Graphics and enabling reproducible, high-quality visualizations

Key Findings

Evidence-Based Conclusions:

Low-discrepancy methods superior

- Halton and Sunflower outperform random jitter
- Uniform distribution = faithful density representation
- Mathematical guarantees translate to perceptual benefits

Determinism essential

- Reproducibility in scientific visualization (R. D. Peng 2011)
- Predictable, interpretable results
- Eliminates artifacts from stochasticity

Linear scaling problematic

- Intelligent jitter demonstrates design failure
- Excessive displacement distorts perception
- Lesson: Scaling function as critical as distribution algorithm

Questions for Discussion

Open Questions for Committee:

1. Should adaptive epsilon be user-overrideable or always automatic?
2. Should package default to Sunflower or Halton?
3. Additional task types or datasets for user study?
4. Should dissertation include 2D scatter plot extension?
5. Publication strategy - single comprehensive vs. multiple focused papers?

Acknowledgments

Thank you:

- Dissertation committee for guidance and feedback
- ggpcp package developers
- UNL Department of Statistics
- Pilot study participants
- Open-source R community

Contact Information:

- Email: denise.bradford@huskers.unl.edu
- GitHub: https://github.com/drbradford12/Dissertation-Data

Questions?

References

References i

Blumenschein, Michael, Xuan Zhang, David Pomerenke, Daniel A. Keim, and Johannes Fuchs. 2020. "Evaluating Reordering Strategies for Cluster Identification in Parallel Coordinates." *Computer Graphics Forum* 39 (3): 537–49. https://doi.org/10.1111/cgf.14000.

Day, Ross H., and Erica J. Stecher. 1991. "The Sine Illusion." *Perception & Psychophysics* 49 (4): 333–39. https://doi.org/10.3758/BF03205988.

Halton, John H. 1960. "On the Efficiency of Certain Quasi-Random Sequences of Points in Evaluating Multi-Dimensional Integrals." *Numerische Mathematik* 2 (1): 84–90. https://doi.org/10.1007/BF01386213.

Hofmann, Heike, and Marie Vendettuoli. 2013. "Common Angle Plots as Perception-True Visualizations of Categorical Associations." *IEEE Transactions on Visualization and Computer Graphics* 19 (12): 2297–2305. https://doi.org/10.1109/TVCG.2013.140.

Holten, Danny, and Jarke J. van Wijk. 2010. "Evaluation of Cluster Identification Performance for Different PCP Variants." *Computer Graphics Forum* 29 (3): 793–802. https://doi.org/10.1111/j.1467-8659.2009.01666.x.

Inselberg, Alfred. 1985. "The Plane with Parallel Coordinates." *The Visual Computer* 1 (2): 69–91. https://doi.org/10.1007/BF01898350.

Johansson, Jimmy, and Camilla Forsell. 2016. "Evaluation of Parallel Coordinates: Overview, Categorization and Guidelines for Future Research." *IEEE Transactions on Visualization and Computer Graphics* 22 (1): 579–88. https://doi.org/10.1109/TVCG.2015.2466992.

Niederreiter, Harald. 1992. *Random Number Generation and Quasi-Monte Carlo Methods*. Philadelphia, PA: Society for Industrial; Applied Mathematics. https://doi.org/10.1137/1.9781611970081.

Peng, Roger D. 2011. "Reproducible Research in Computational Science." *Science* 334 (6060): 1226–27. https://doi.org/10.1126/science.1213847.

Peng, Wei, Matthew O Ward, and Elke A Rundensteiner. 2004. "Clutter Reduction in Multi-Dimensional Data Visualization Using Dimension Reordering." In *IEEE Symposium on Information Visualization*, 89–96. IEEE.

Schonlau, Matthias. 2003. "The Hammock Plot: Visualizing Mixed Categorical and Numerical Data." In *Proceedings of the American Statistical Association*.

Schonlau, Matthias, and Rosie Yuyan Yang. 2024. "Hammock Plots: Visualizing Categorical Data Beyond Parallel Coordinates." *Journal of Computational and Graphical Statistics*.

VanderPlas, Susan, Yawei Ge, Antony Unwin, and Heike Hofmann. 2023. "Penguins Go Parallel: A Grammar of Graphics Framework for Generalized Parallel Coordinate Plots." *Journal of Computational and Graphical Statistics* 32 (4): 1405–20. https://doi.org/10.1080/10618600.2023.2181762.

References ii

Vogel, Helmut. 1979. "A Better Way to Construct the Sunflower Head." *Mathematical Biosciences* 44 (3-4): 179–89. https://doi.org/10.1016/0025-5564(79)90080-4.

Ware, Colin. 2012. *Information Visualization: Perception for Design*. 3rd ed. Waltham, MA: Morgan Kaufmann.

Wegman, Edward J. 1990. "Hyperdimensional Data Analysis Using Parallel Coordinates." *Journal of the American Statistical Association* 85: 664–75.