Visualizing Distributed Memory Computations with Hive Plots

VizSec 2012, October 15, 2012, Seattle, Washington

Sophie Engle and Sean Whalen

Introduction

Introduction

- High performance computing environments
  - Used for scientific computing applications at several national laboratories
  - Potential for misuse by insiders and outsiders
- Anomaly detection
  - Determine normal versus abnormal behavior for these environments to prevent unauthorized use
  - Can classify codes into “computational dwarves” to determine “normal” (Asanovic 2006)

Introduction

- Several network measures can be used as features in classification
  - Time consuming to calculate these measures
  - Time consuming to compare how well these measures perform for classification
- Use visualization to choose network measures to use as classification features
  - Which measures look similar for similar codes?
  - Which measures look distinct for distinct codes?

Dataset

Original Dataset

- Data collection
  - Collected by NERSC at LBNL
  - Used IPM to monitor MPI calls between ranks *(captures communication between compute nodes)*
- Dataset contents
  - Total of 1681 IPM logs
  - Covers 29 different scientific computing codes with varying ranks, parameters, and architectures

Original Dataset

```
Src,Dst,MPICall,Bytes,Repeats,Code
0,1,29,99856,52,cactus
0,4,29,99856,52,cactus
0,0,2,4,5,cactus
0,0,2,8,7,cactus
0,1,22,599136,26,cactus
0,-1,5,0,1,cactus
0,4,22,599136,26,cactus
0,16,29,99856,52,cactus
```

Original Dataset

```
Src,Dst,MPICall,Bytes,Repeats,Code
0,1,29,99856,52,cactus
0,4,29,99856,52,cactus
0,0,2,4,5,cactus
0,0,2,8,7,cactus
0,1,22,599136,26,cactus
0,-1,5,0,1,cactus
0,4,22,599136,26,cactus
0,16,29,99856,52,cactus
```

Subset Analyzed

Code	Description	Nodes	Edges
cactus	astrophysics	64	989
ij	algebraic multi-grid	64	8596
milc	lattice gauge theory	64	1473
namd	molecular dynamics	64	8208
paratec	materials science	64	16492
superlu	sparse linear algebra	64	3239
tgyro	magnetic fusion	64	1123
vasp	materials science	64	13760

Subset Analyzed

Code	Description	Nodes	Edges
cactus	astrophysics	64	989
ij	algebraic multi-grid	64	8596
milc	lattice gauge theory	64	1473
namd	molecular dynamics	64	8208
~~paratec~~	~~materials science~~	~~64~~	~~16492~~
superlu	sparse linear algebra	64	3239
tgyro	magnetic fusion	64	1123
~~vasp~~	~~materials science~~	~~64~~	~~13760~~

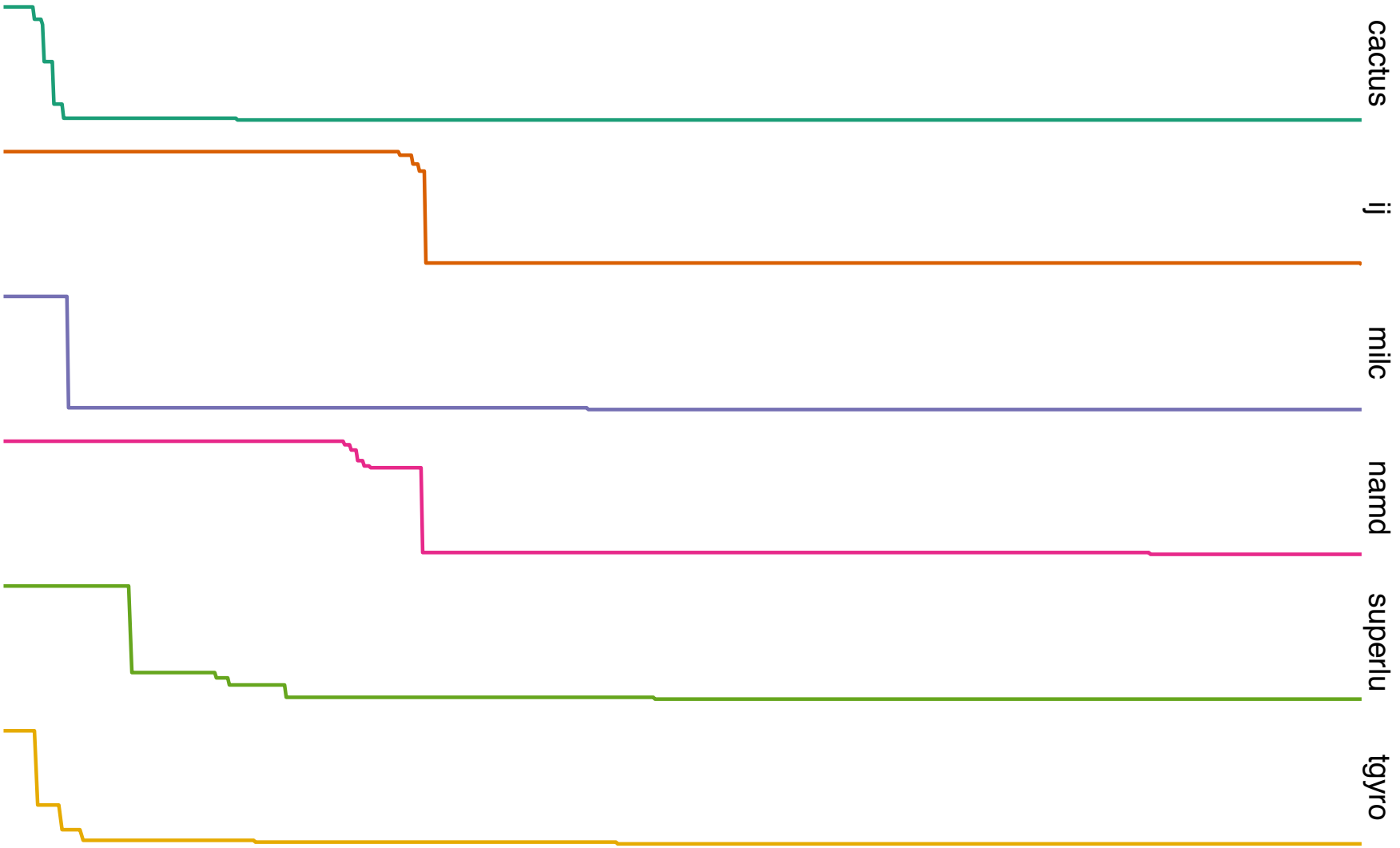
Network Measures

Measure	Description
degree	the number of adjacent edges
betweenness	number of shortest paths going through a node
closeness	measures steps required to reach every other node
eccentricity	shortest path distance from farthest node
page rank	measures relative importance of node
transitivity	probability adjacent nodes are connected (clustering coefficient)

Calculated in R using the `igraph` library.

Motivation

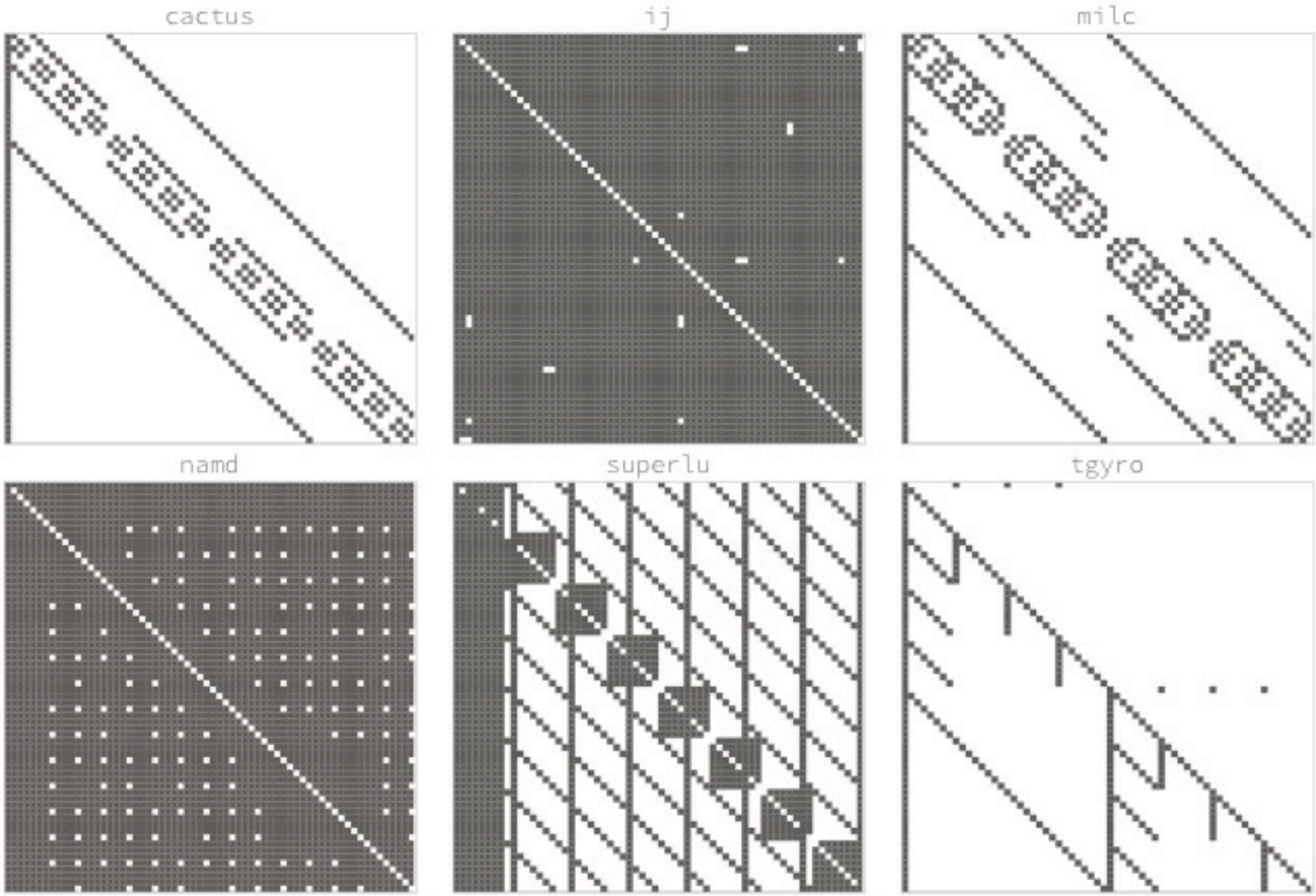
Traditional Degree CCDF Plot



Traditional Degree CCDF Plot

- Pros:
  - Able to compare individual metrics across datasets
  - Simple approach, widely used
- Cons:
  - Contains no information on topology
  - Lines look visually similar, may not be appropriate for generating visual signatures

Adjacency Matrices



Adjacency Matrices

- Pros:
  - Comparable across datasets
  - Easy to see communication patterns
  - Many distinct codes look distinct
- Cons:
  - No information on metrics needed for classification

Issues Identified

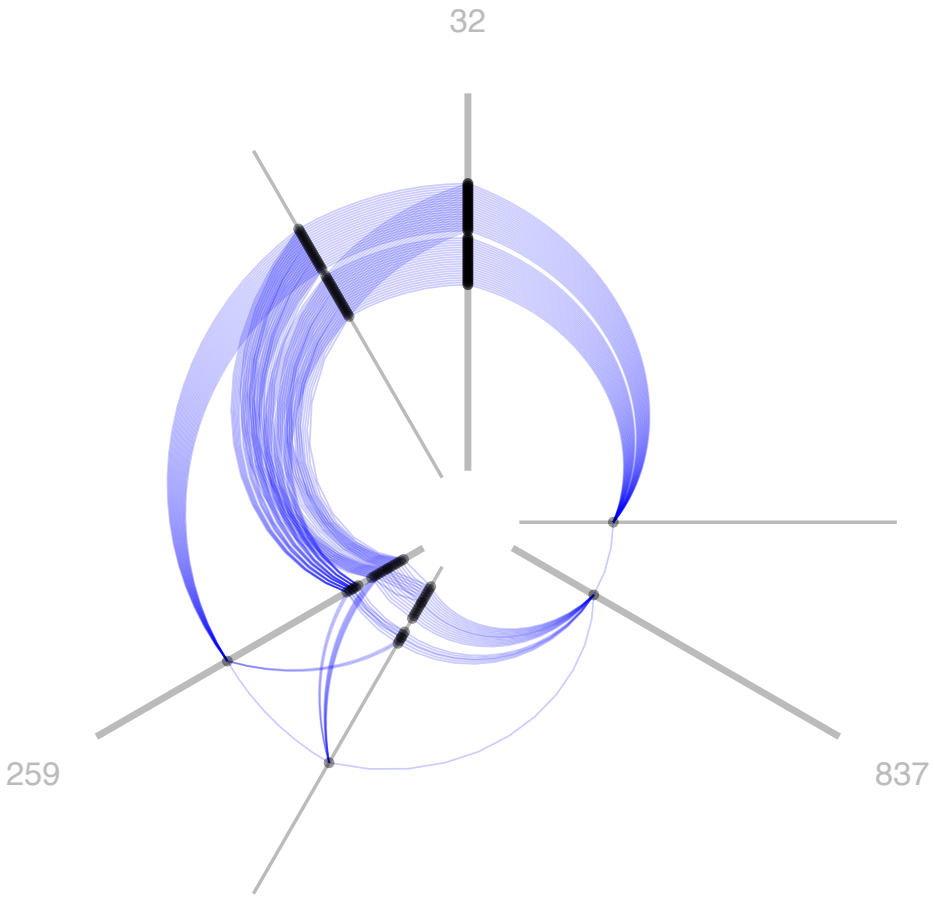
- Traditional CCDF plot does not convey any information about network topology
- Traditional adjacency matrices do not convey any information about network properties
- Traditional network layout algorithms are not repeatable or comparable across networks

Hive Plots

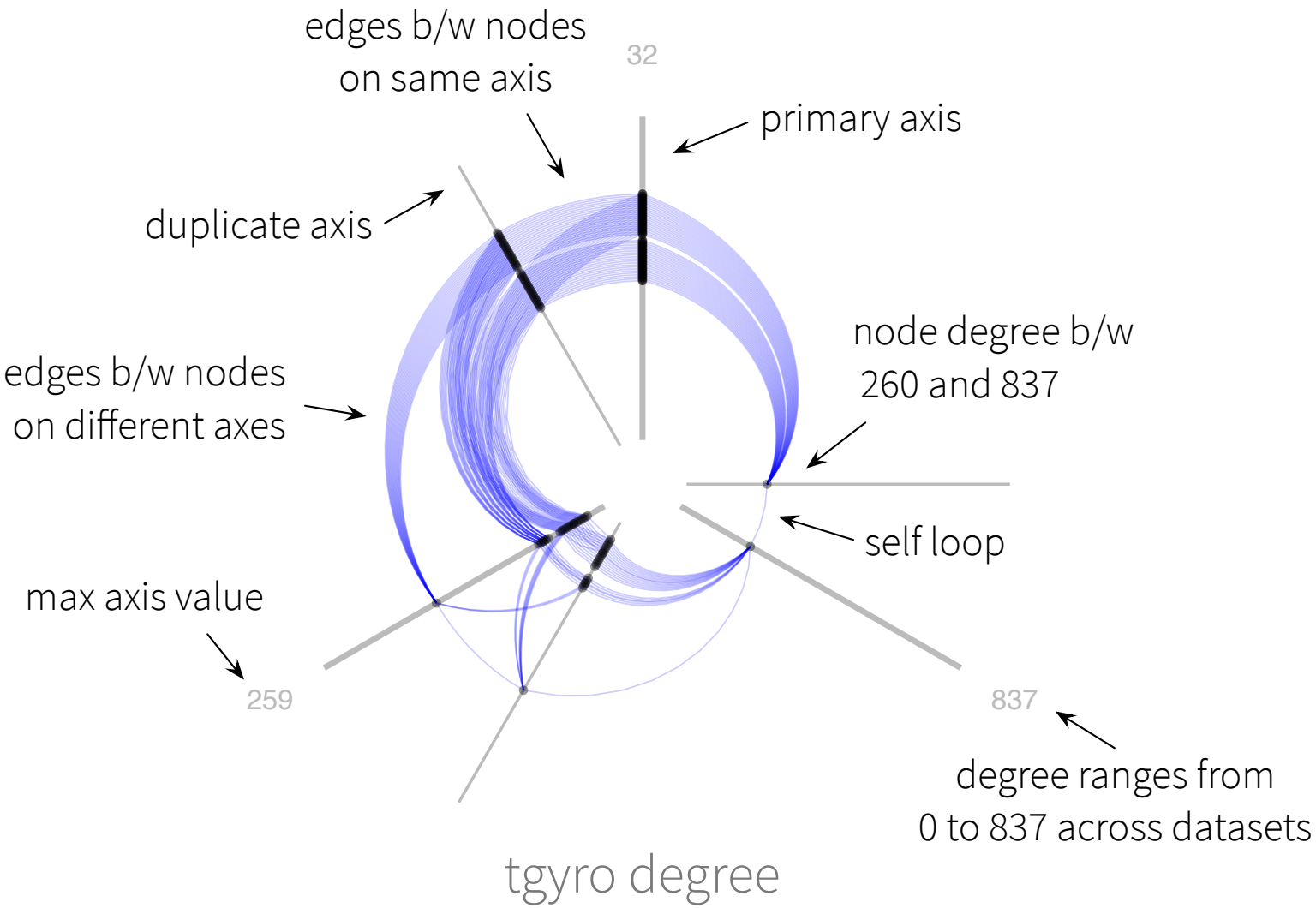
Introduction to Hive Plots

- What are hive plots?
  - Network layout algorithm using network properties for consistent node placement
  - A radially-arranged parallel coordinate plot
- Why use hive plots?
  - Repeatable, comparable network layouts
  - Integration of network properties with topology

Understanding Hive Plots



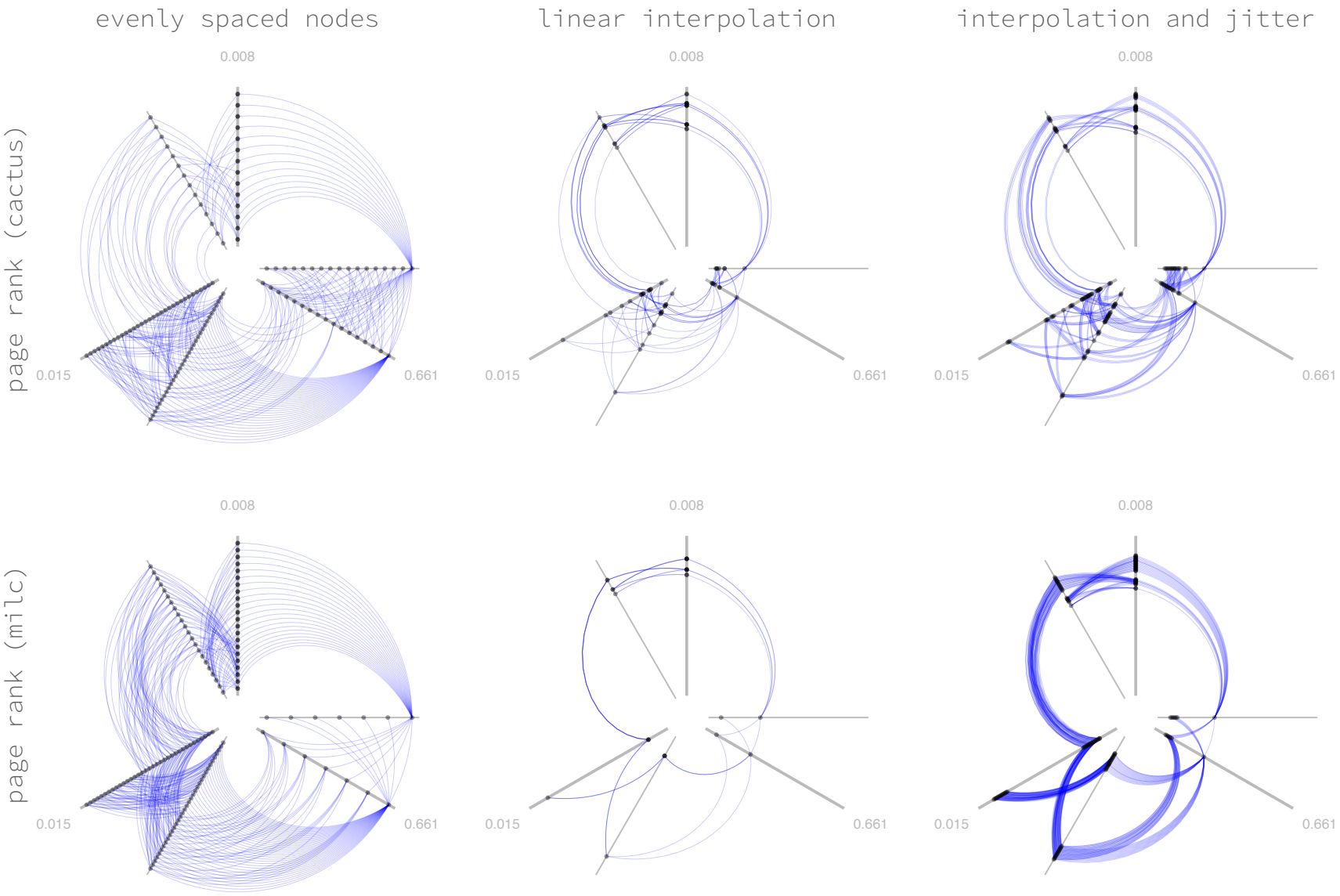
Understanding Hive Plots



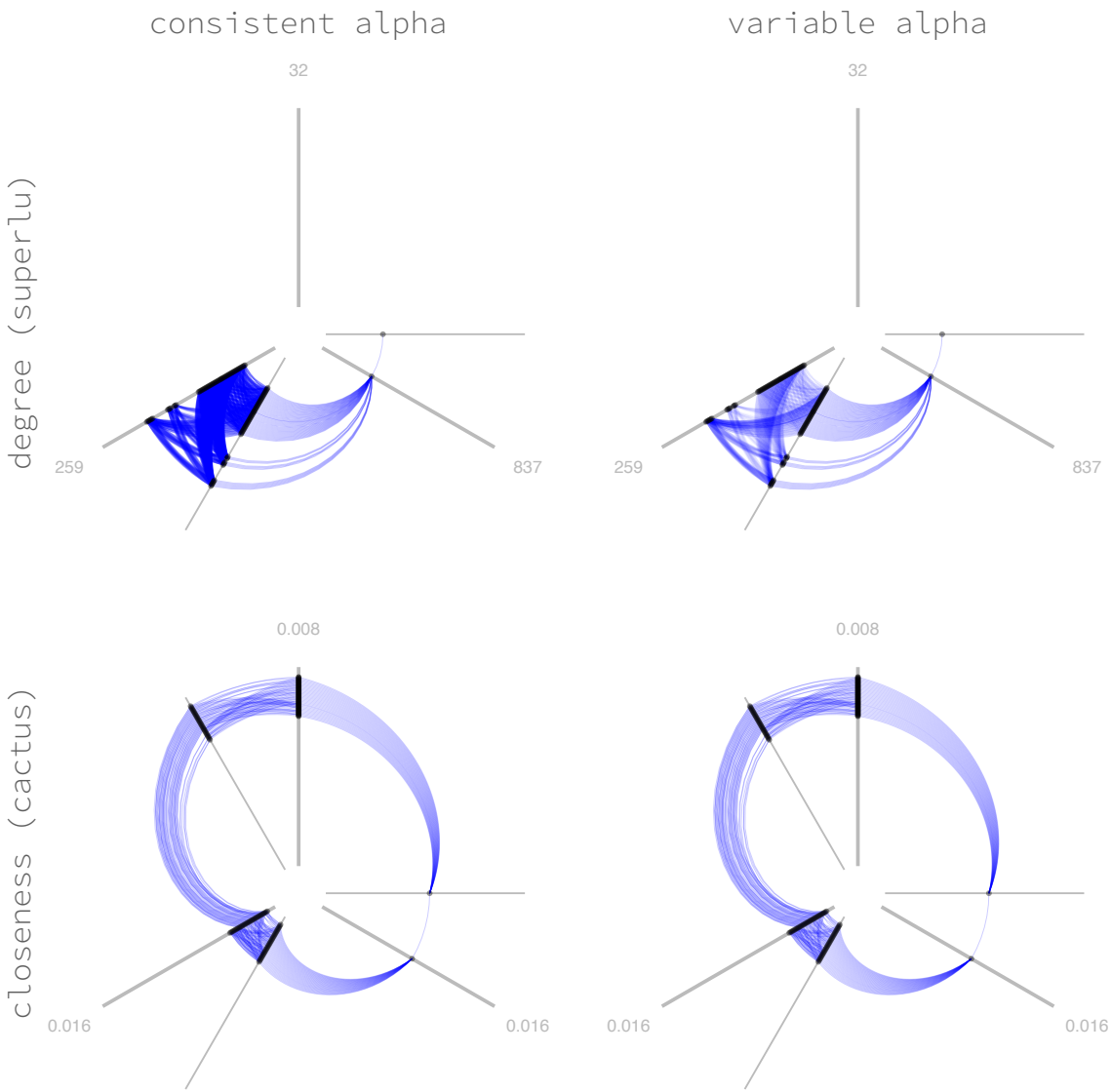
Implementation

- Existing implementations exist
  - `JHive` (Java)
  - `HiveR` (R)
  - `HiveGraph` (webapp)
  - Prototypes in Perl and `d3.js`
- Custom implementation in R and **`ggplot2`**
  - Implements **grammar of graphics** (Wilkinson)
  - **Polar plots** to create hive plots
  - **Facets** to create hive panels*
  - *Non-interactive*

Implementation



Implementation



Hive Plot References

Hive Plots—Rational Approach to Visualizing Networks

by Martin Krzywinski, Inanc Birol, Steven JM Jones and Marco A Marra

in Briefings in Bioinformatics, volume 13, issue 5, pages 627–644, 2012

Hive Plots: Rational Network Visualization—Farewell to Hairballs

by Martin Krzywinski at http://www.hiveplot.com online

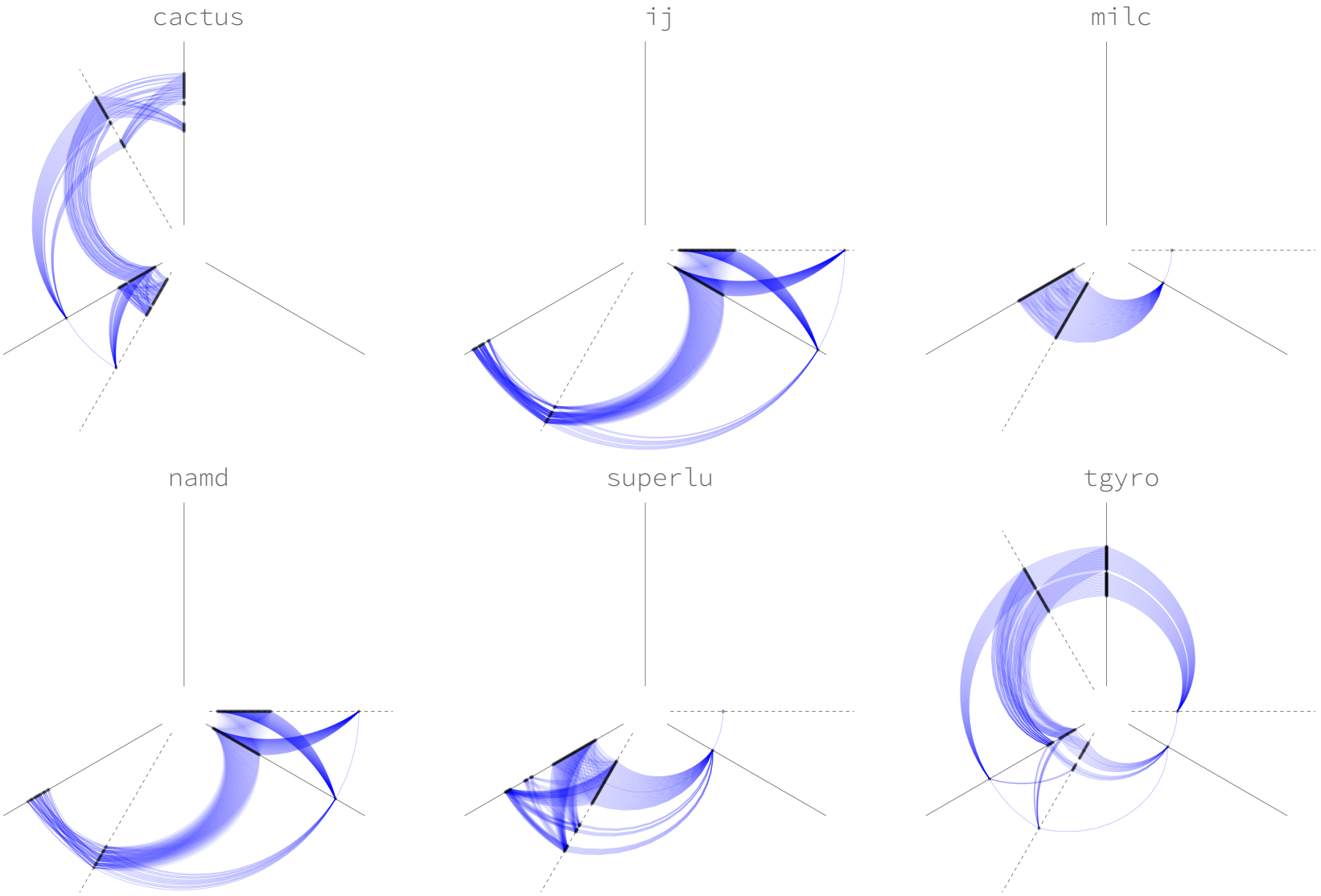
Getting Into Visualization of Large Biological Data Sets

by Martin Krzywinski, Inanc Birol, Steven Jones, Marco Marra

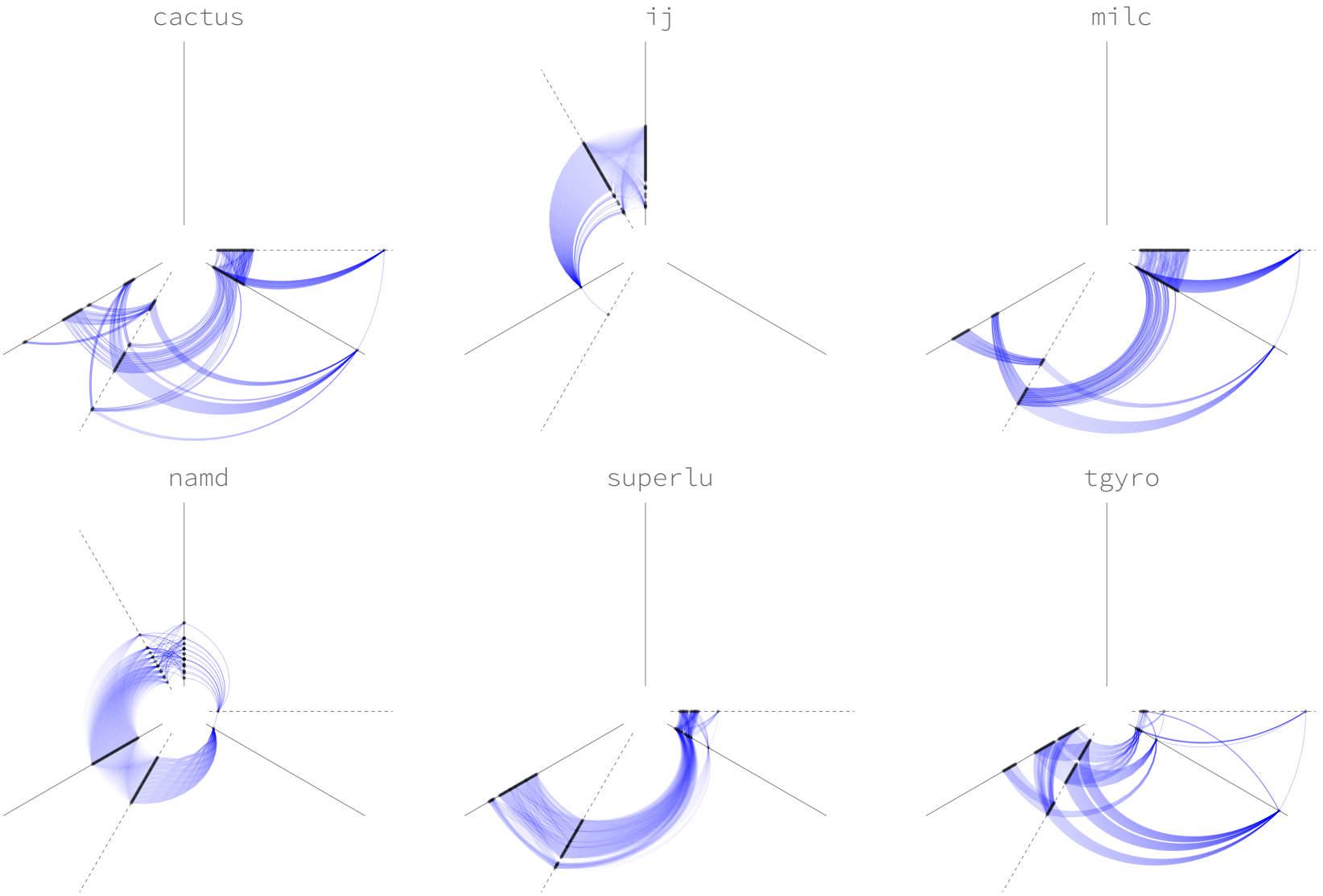
in BioVis 2012 Posters, 2nd floor foyer, Sunday 8:30am – Monday 5:55pm

Initial Results

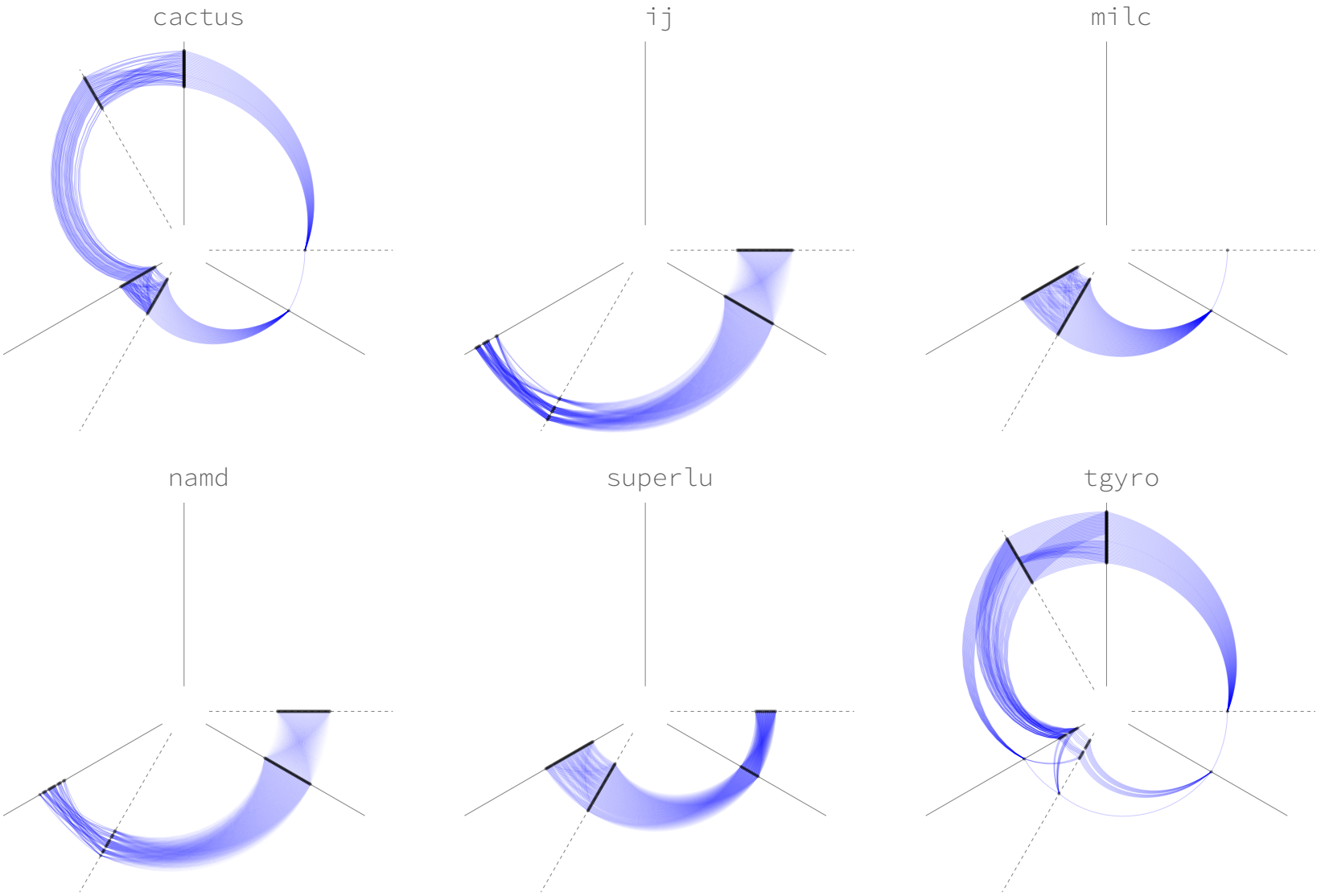
Degree



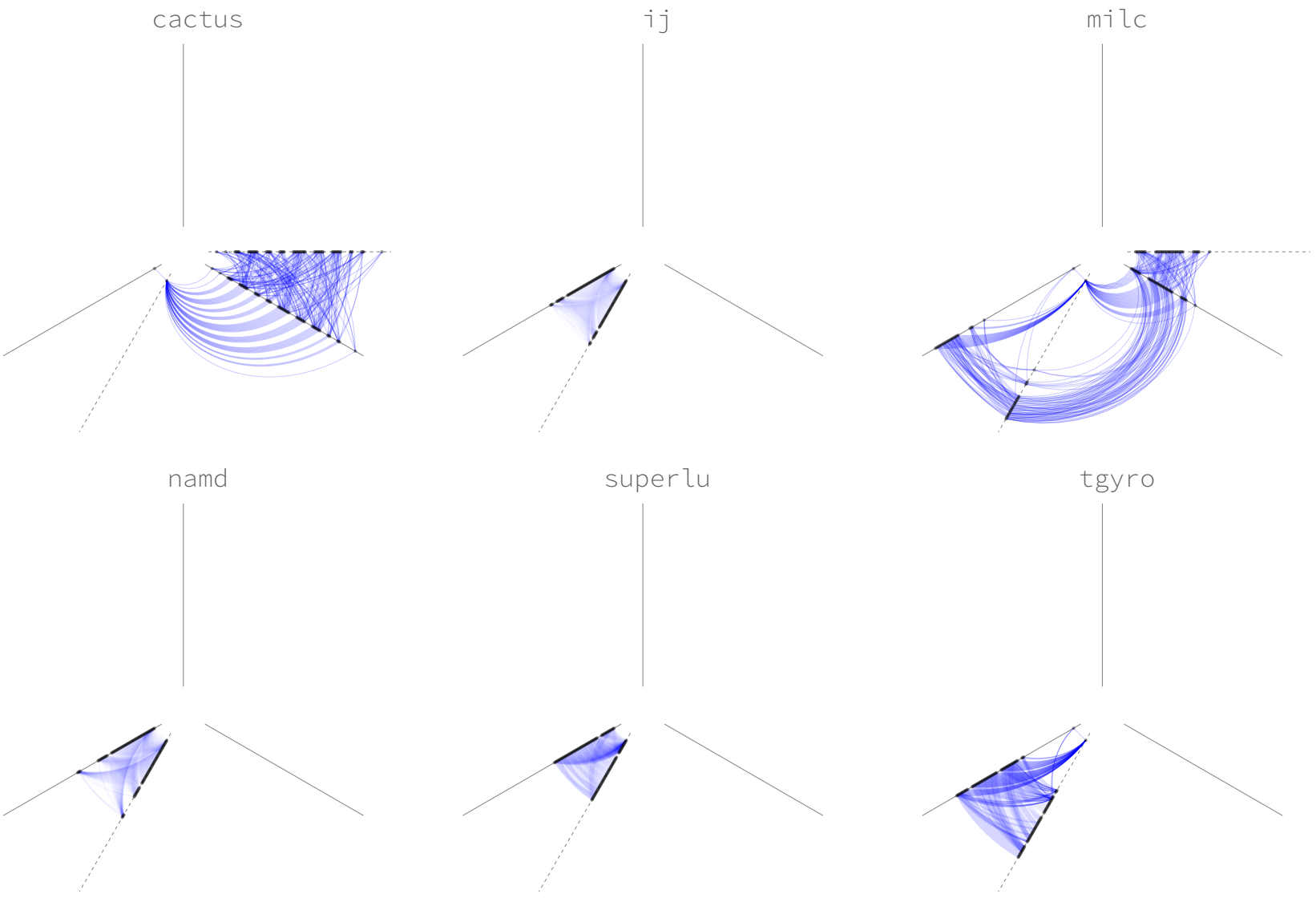
Betweenness



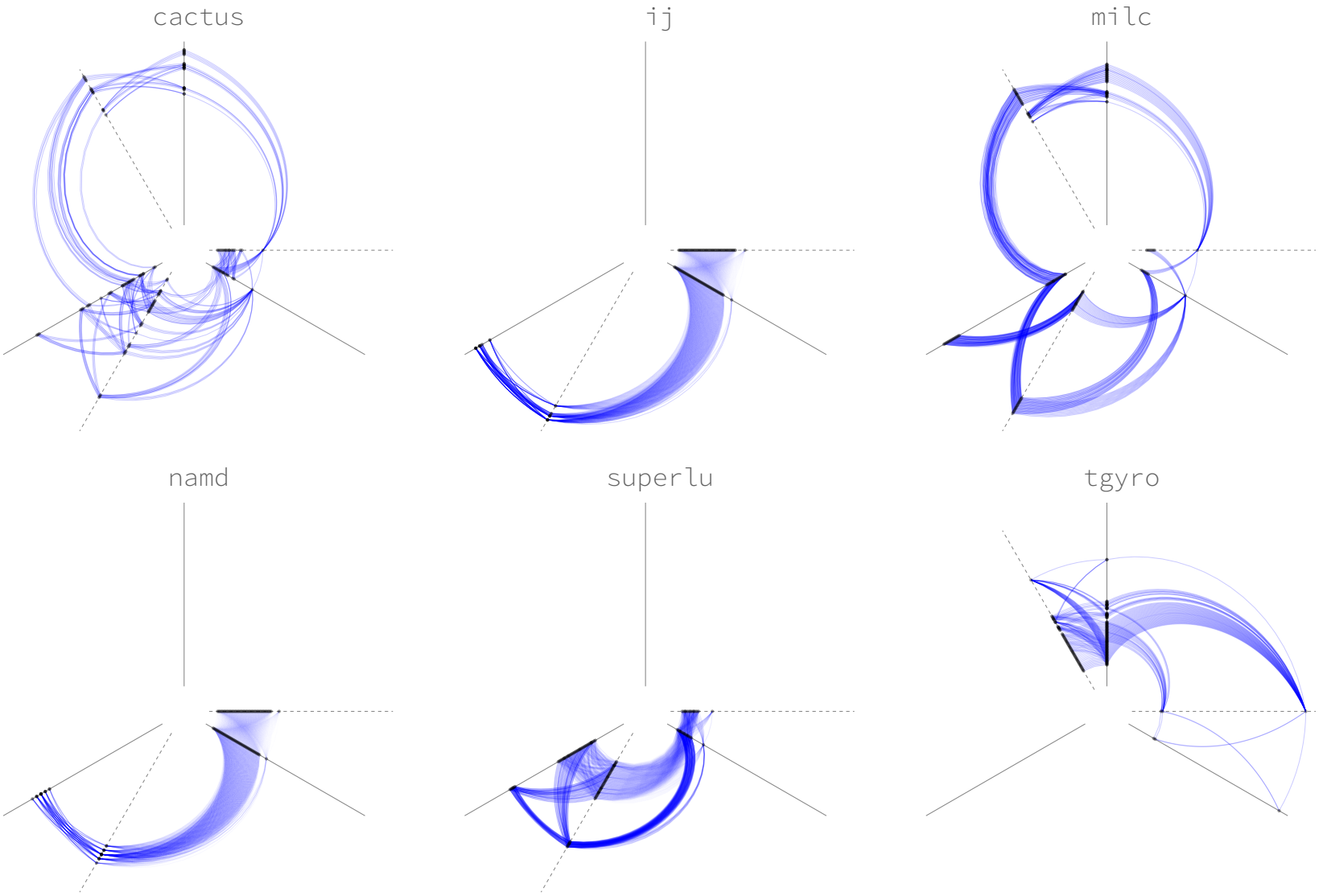
Closeness



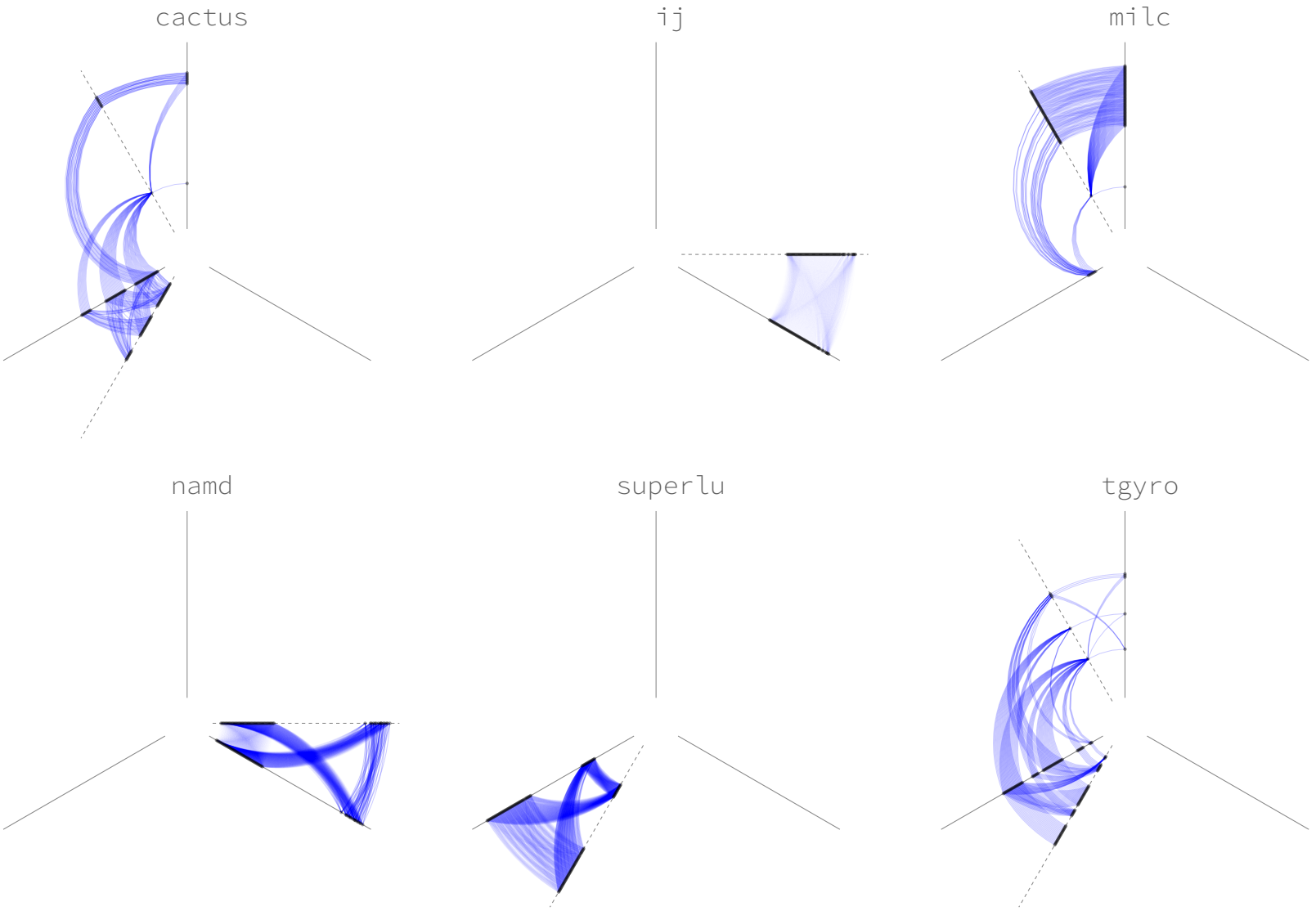
Eccentricity



Page Rank



Transitivity



Visually Distinct

	cactus	ij	milc	namd	superlu	tygro
degree	X		X		X	X
betweenness	X	X	X	X	X	X
closeness	X		X		X	X
eccentricity	X		X			X
page rank	X		X		X	X
transitivity	X	X	X	X	X	X

Visually Distinct

	cactus	ij	milc	namd	superlu	tygro
degree	X		X		X	X
betweenness	X	X	X	X	X	X
closeness	X		X		X	X
eccentricity	X		X			X
page rank	X		X		X	X
transitivity	X	X	X	X	X	X

Next Steps

Next Steps

- Improve hive plot visualizations
  - Explore variable-length axes
  - Explore better axes assignment
- Incorporate more information from data set
  - Multiple-edge connections
  - Type of IPM calls
  - Amount of data transmitted

Next Steps

- Feature identification
  - Compare hive plots for more distinct codes
  - Compare hive plots for similar codes
  - Identify features that visually distinguish codes
- Classification and anomaly detection
  - Determine if features identified by visualization lead to better classifiers and anomaly detection

Conclusion

Summary

- Motivation and goals
  - Improve anomaly detection in HPC environments
  - Improve classification of HPC codes
  - Use exploratory visualization for feature selection

Summary

- Motivation and goals
  - Improve anomaly detection in HPC environments
  - Improve classification of HPC codes
  - **Use exploratory visualization for feature selection**
- Initial results
  - Hive plots allow visual comparison of HPC codes
  - Some features distinguish distinct HPC codes

References

Hive Plots—Rational Approach to Visualizing Networks

by Martin Krzywinski, Inanc Birol, Steven JM Jones and Marco A Marra

in Briefings in Bioinformatics, volume 13, issue 5, pages 627–644, 2012

Network-Theoretic Classification of Parallel Computation Patterns

by Sean Whalen, Sophie Engle, Sean Peisert, and Matt Bishop

in International Journal of High Performance Computing Applications (IJHPCA), volume 26, number 2, pages 159–169, May 2012

Multiclass Classification of Distributed Memory Parallel Computations

by Sean Whalen, Sean Peisert, and Matt Bishop

to appear in Pattern Recognition Letters (PRL), 2012

Contact Information

Sophie Engle

University of San Francisco
Department of Computer Science

sjengle@cs.usfca.edu ◆ http://sjengle.cs.usfca.edu

Sean Whalen

Mount Sinai School of Medicine
Institute for Genomics and Multiscale Biology

shwhalen@cs.columbia.edu ◆ http://node99.org

Questions?