A DISSERTATION

REPRESENTING BIG DATA AS NETWORKS: NEW METHODS AND INSIGHTS

Jian Xu
Department of Computer Science and Engineering
University of Notre Dame

Examination committee
Nitesh V. Chawla, *Chair*
David M. Lodge
Tijana Milenkovic
Zoltan Toroczkai

UNIVERSITY OF NOTRE DAME

May 19th, 2017

iCeNSA
Interdisciplinary Center for Network Science & Applications

Examination committee

Prof. Nitesh Chawla, *chair*

Prof. David Lodge

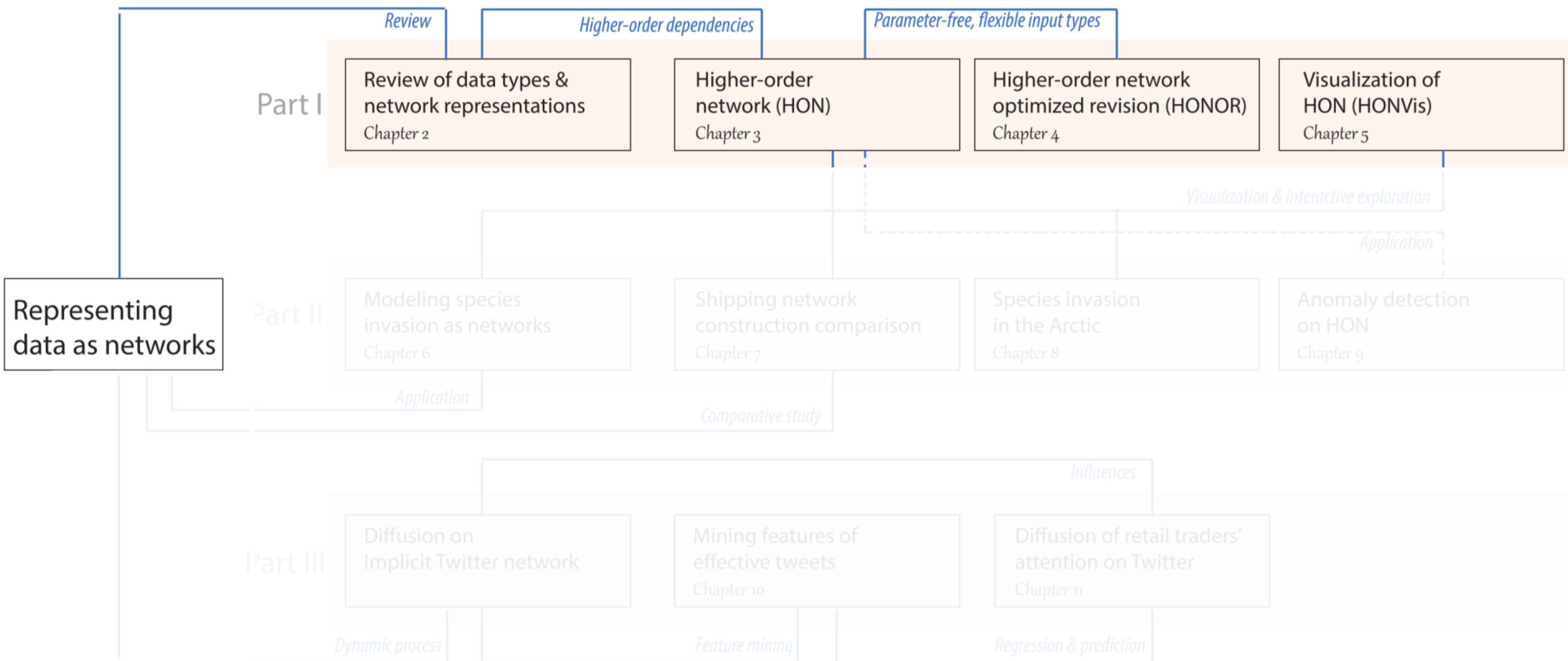
Prof. Tijana Milenkovic

Prof. Zoltan Torotzkai

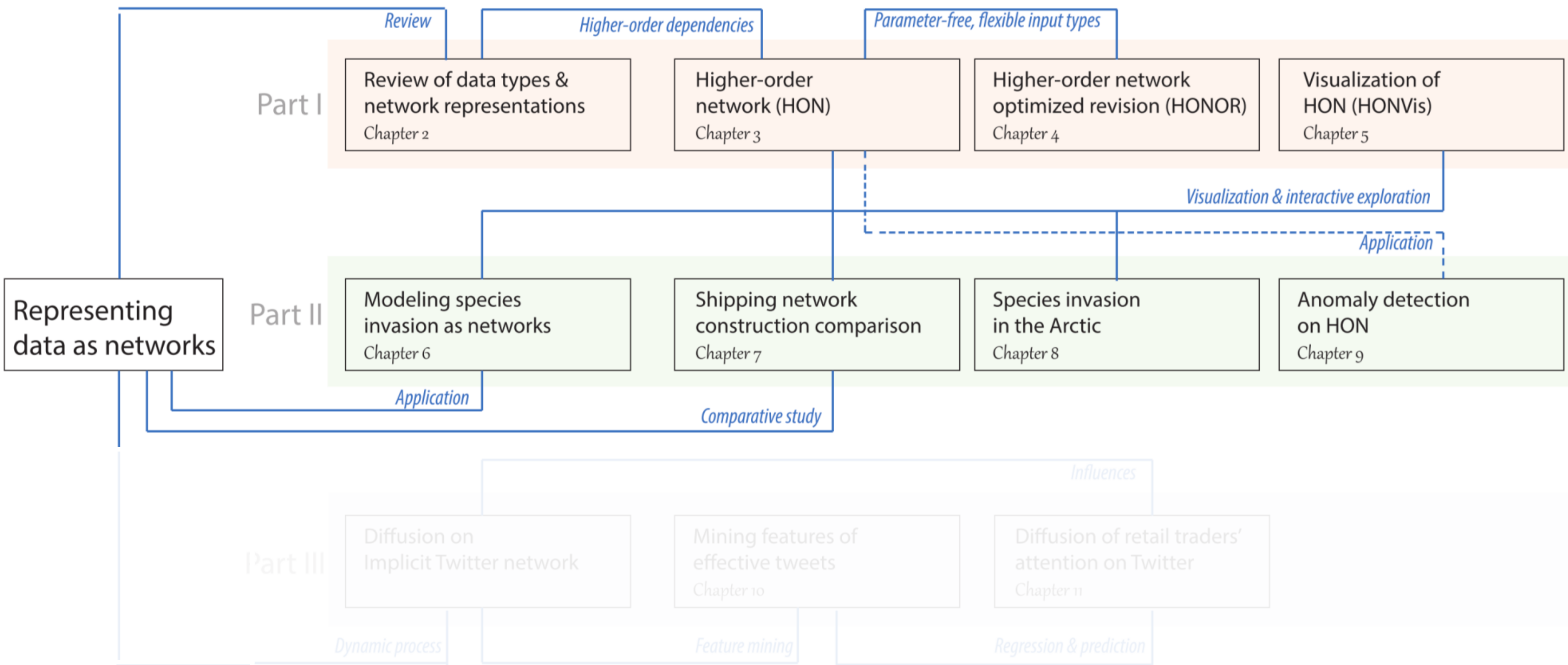
Overview

Representing
data as networks

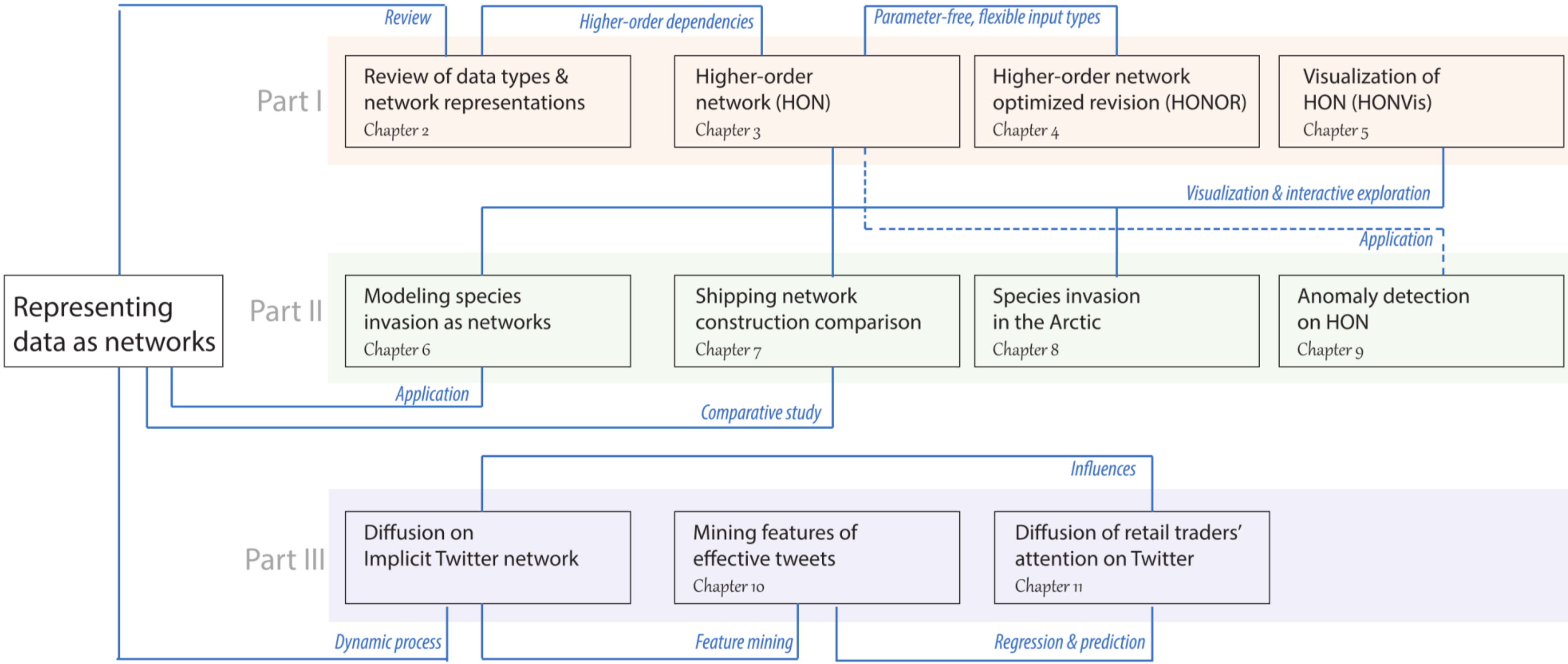
Overview


Review
Higher-order dependencies
Parameter-free, flexible input types
Part I
Review of data types & network representations
Chapter 2
Higher-order network (HON)
Chapter 3
Higher-order network optimized revision (HONOR)
Chapter 4
Visualization of HON (HONVis)
Chapter 5
Visualization & interactive exploration
Application
Representing data as networks
Part II
Modeling species invasion as networks
Chapter 6
Shipping network construction comparison
Chapter 7
Species invasion in the Arctic
Chapter 8
Anomaly detection on HON
Chapter 9
Application
Comparative study
Influences
Part III
Diffusion on Implicit Twitter network
Mining features of effective tweets
Chapter 10
Diffusion of retail traders' attention on Twitter
Chapter 11
Dynamic process
Feature mining
Regression & prediction

Overview

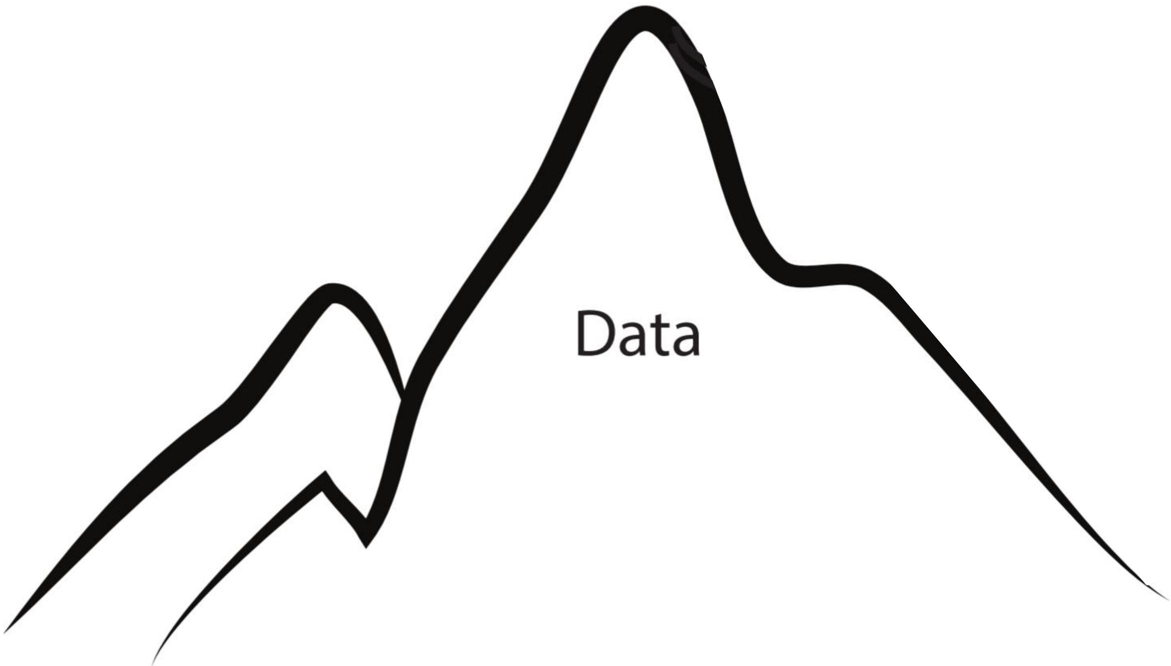

Review
Higher-order dependencies
Parameter-free, flexible input types
Part I
Review of data types & network representations
Chapter 2
Higher-order network (HON)
Chapter 3
Higher-order network optimized revision (HONOR)
Chapter 4
Visualization of HON (HONVis)
Chapter 5
Visualization & interactive exploration
Application
Representing data as networks
Part II
Modeling species invasion as networks
Chapter 6
Shipping network construction comparison
Chapter 7
Species invasion in the Arctic
Chapter 8
Anomaly detection on HON
Chapter 9
Application
Comparative study
Influences
Part III
Diffusion on Implicit Twitter network
Mining features of effective tweets
Chapter 10
Diffusion of retail traders' attention on Twitter
Chapter 11
Dynamic process
Feature mining
Regression & prediction

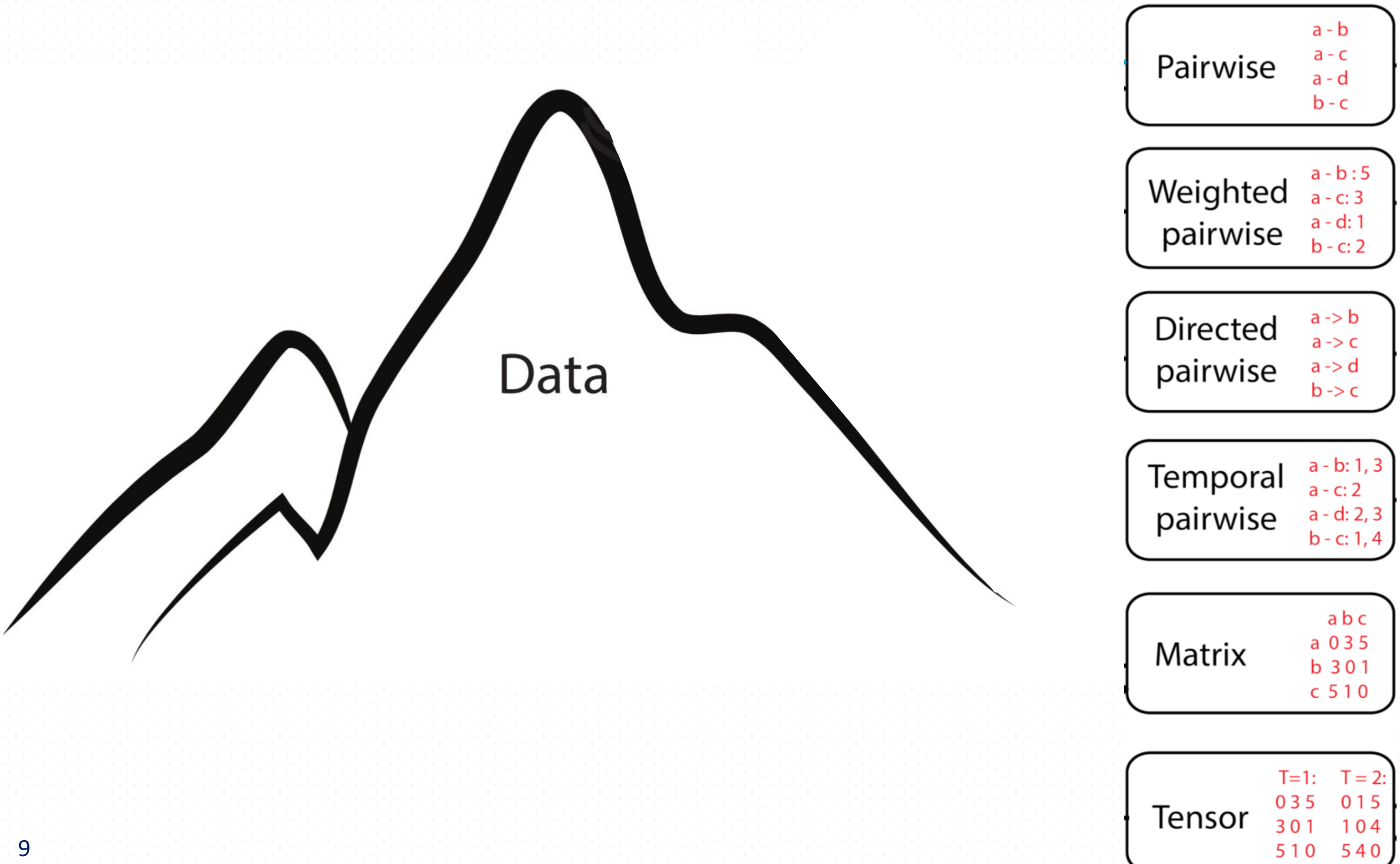
Overview

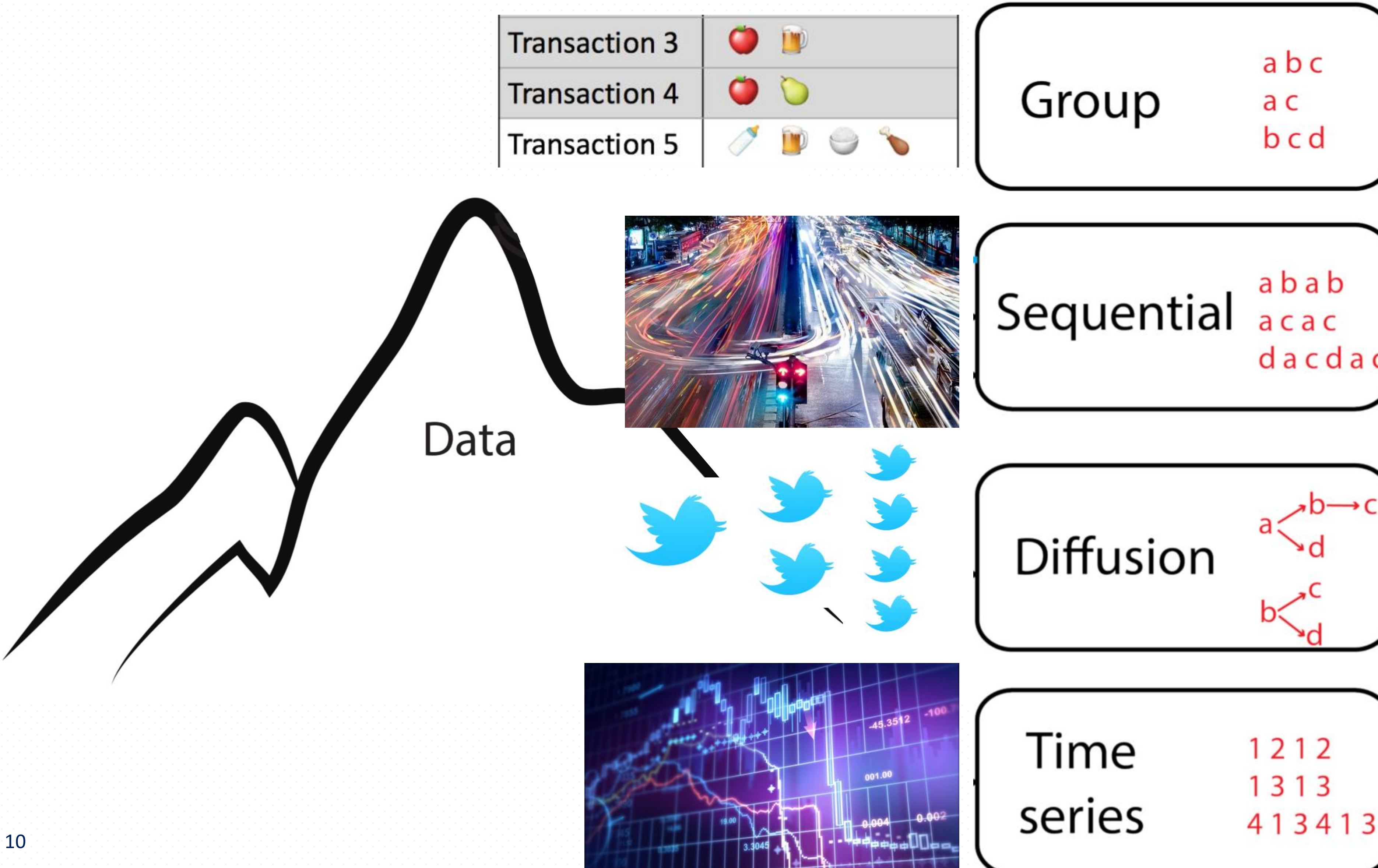

Review
Higher-order dependencies
Parameter-free, flexible input types
Part I
Review of data types & network representations
Chapter 2
Higher-order network (HON)
Chapter 3
Higher-order network optimized revision (HONOR)
Chapter 4
Visualization of HON (HONVis)
Chapter 5
Visualization & interactive exploration
Application
Representing data as networks
Part II
Modeling species invasion as networks
Chapter 6
Shipping network construction comparison
Chapter 7
Species invasion in the Arctic
Chapter 8
Anomaly detection on HON
Chapter 9
Application
Comparative study
Influences
Part III
Diffusion on Implicit Twitter network
Mining features of effective tweets
Chapter 10
Diffusion of retail traders' attention on Twitter
Chapter 11
Dynamic process
Feature mining
Regression & prediction

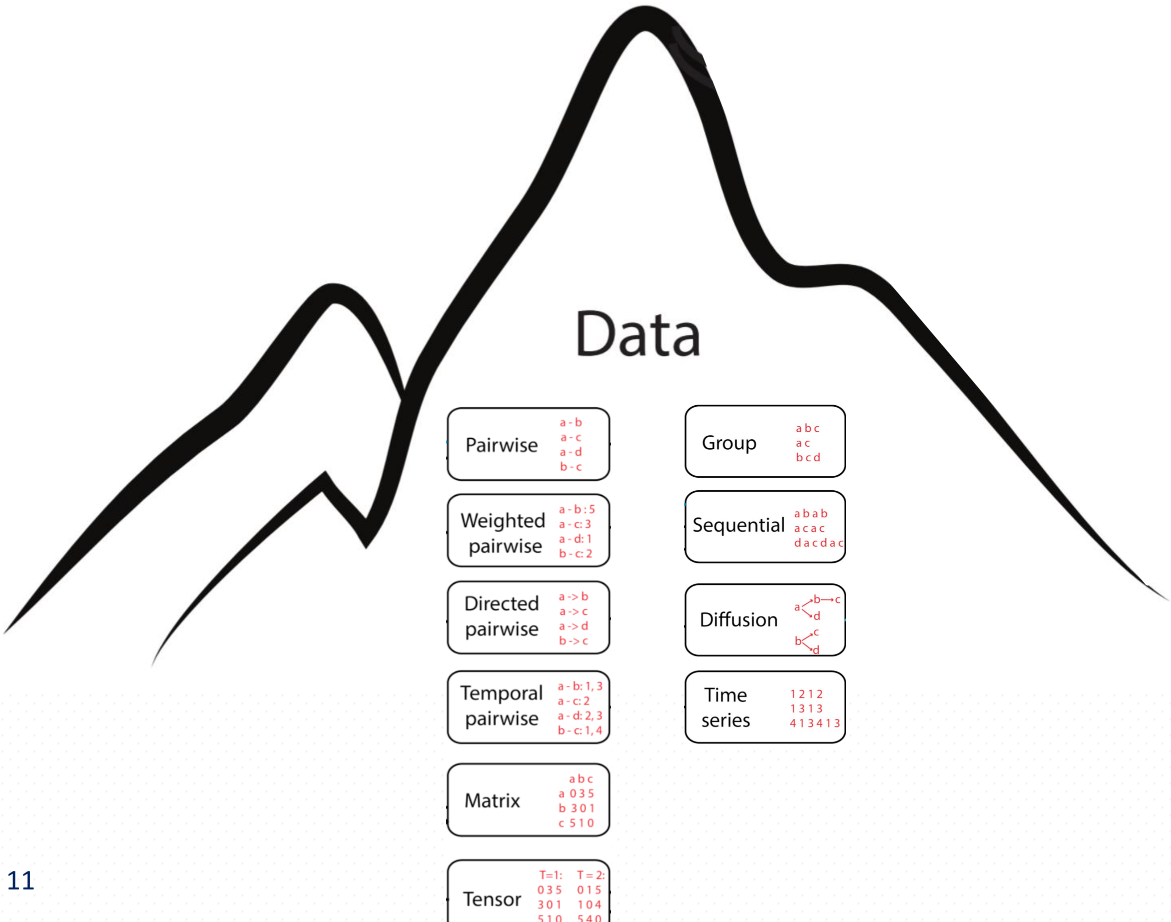
Part I

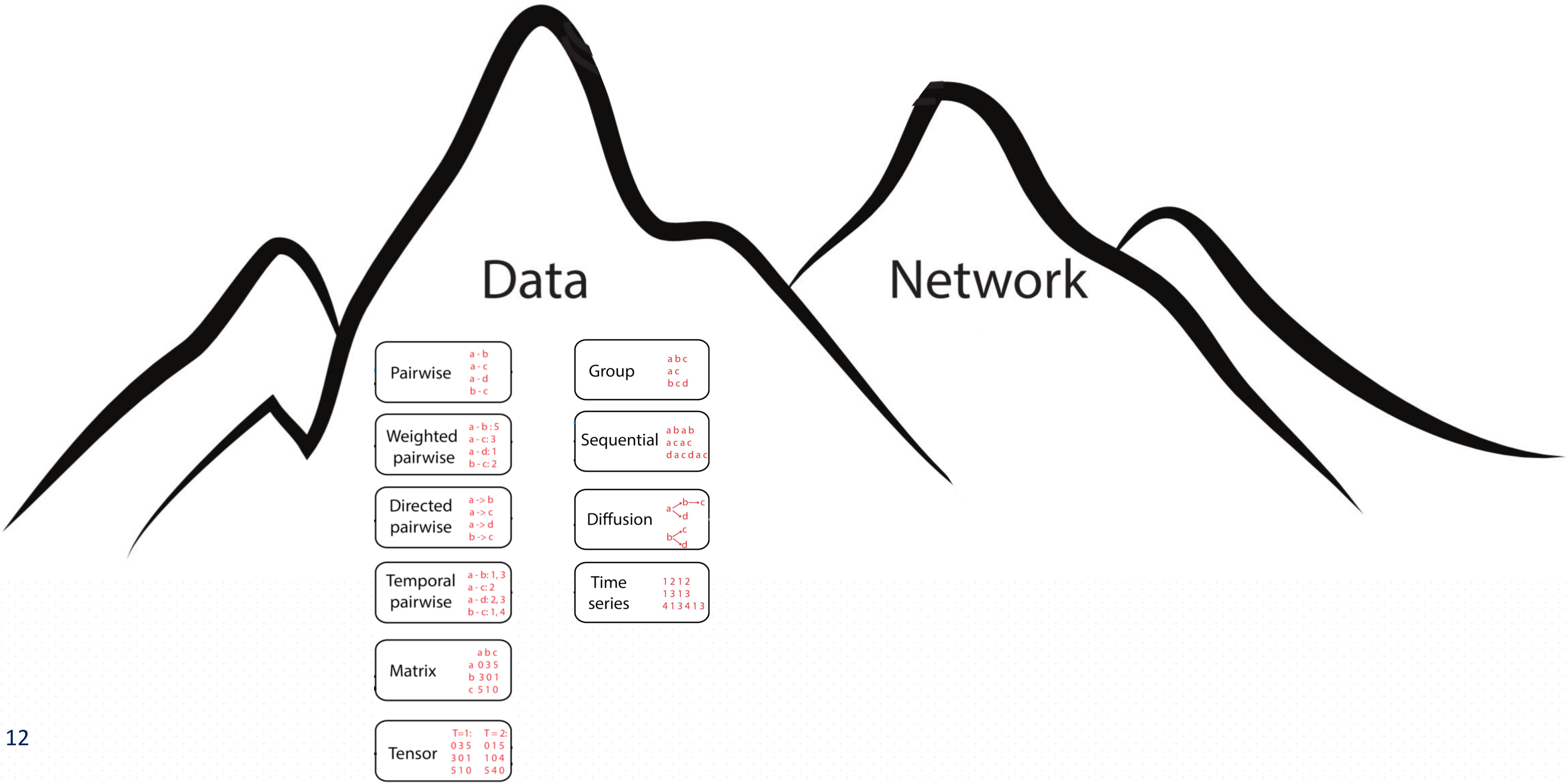
Methods to represent data as networks

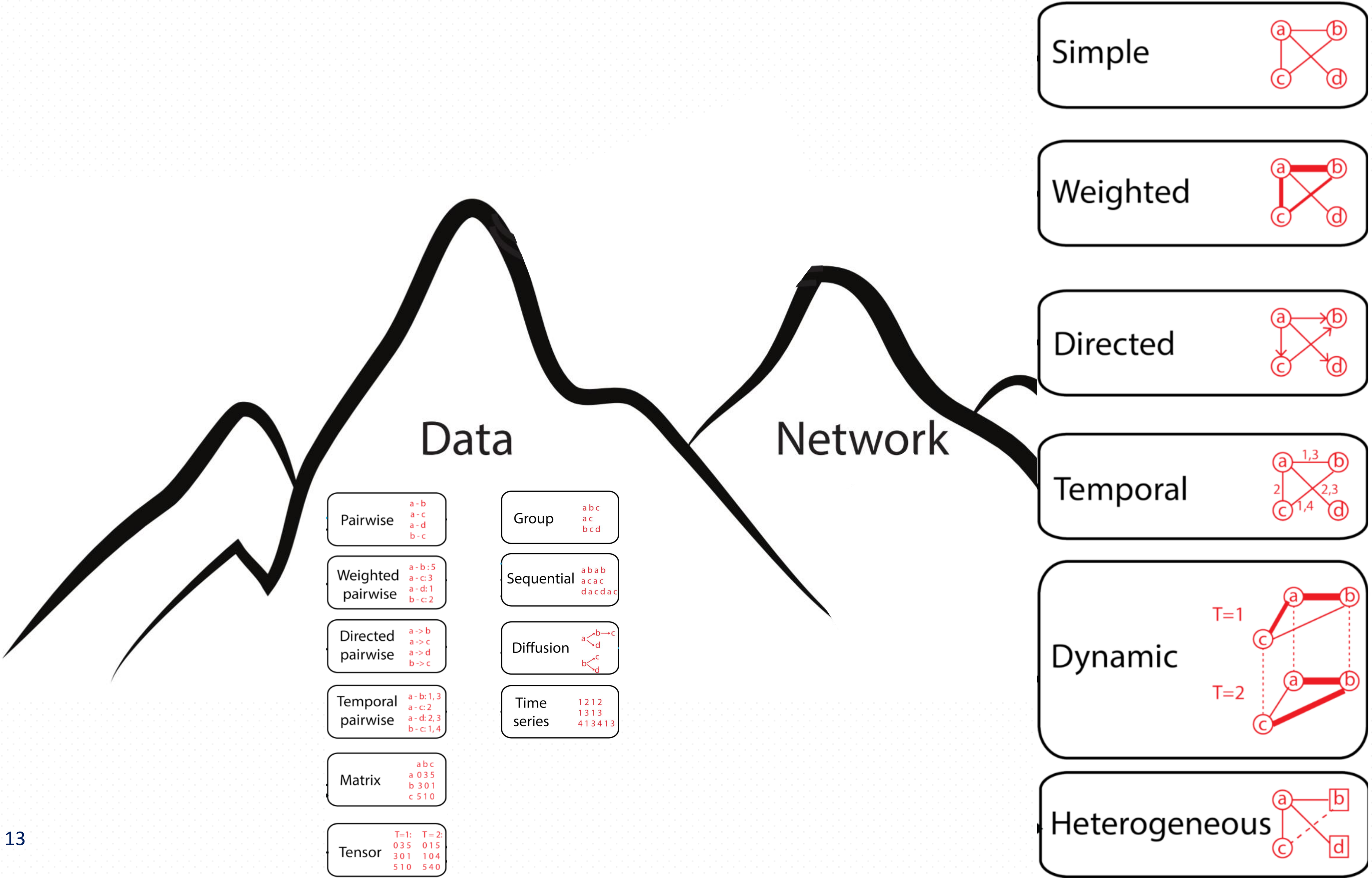

Data

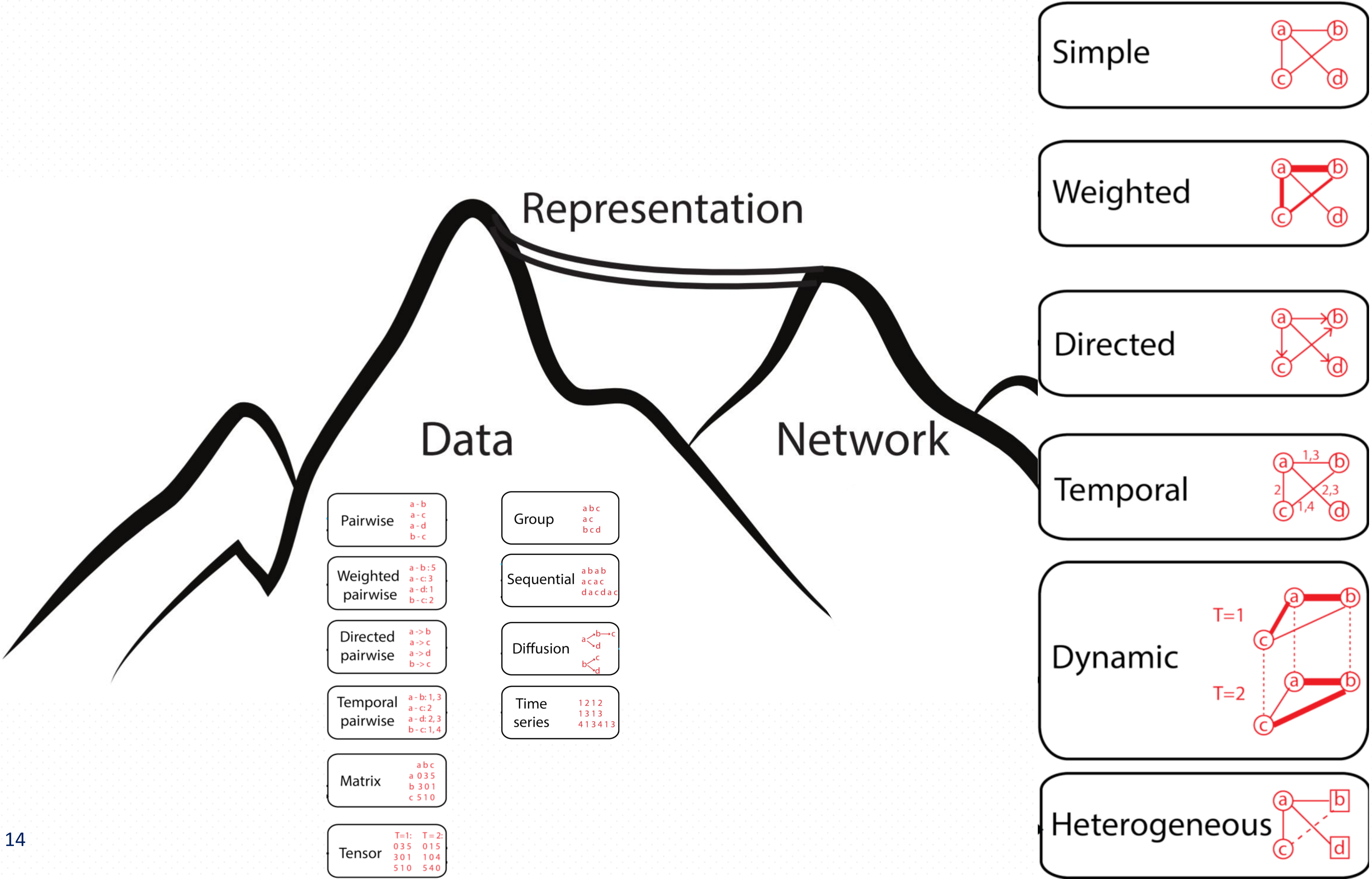

Data
Pairwise
a - b
a - c
a - d
b - c
Weighted pairwise
a - b : 5
a - c: 3
a - d: 1
b - c: 2
Directed pairwise
a -> b
a -> c
a -> d
b -> c
Temporal pairwise
a - b: 1, 3
a - c: 2
a - d: 2, 3
b - c: 1, 4
Matrix
a b c
a 0 3 5
b 3 0 1
c 5 1 0
Tensor
T=1:
0 3 5
3 0 1
5 1 0
T = 2:
0 1 5
1 0 4
5 4 0

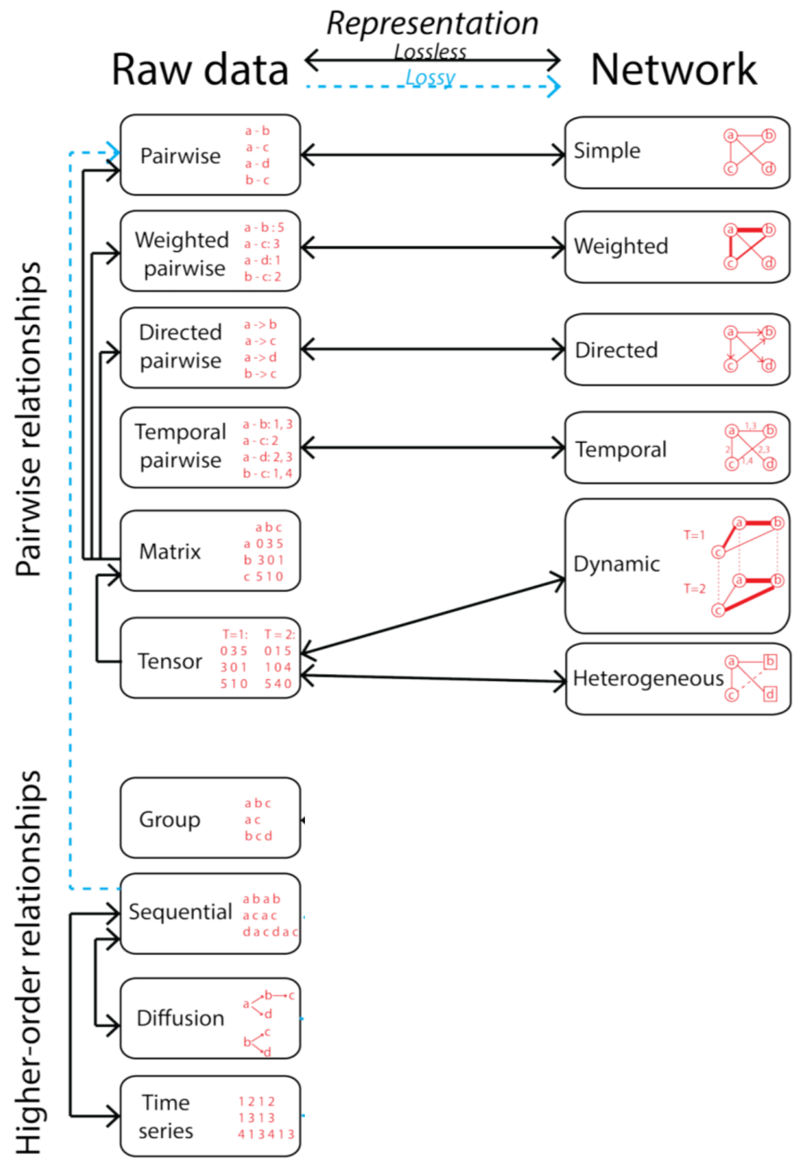

Transaction 3
Transaction 4
Transaction 5
Data
Group
a b c
a c
b c d
Sequential
a b a b
a c a c
d a c d a c
Diffusion
a
b
c
d
b
c
d
Time
series
1 2 1 2
1 3 1 3
4 1 3 4 1 3


Data
Pairwise
a - b
a - c
a - d
b - c
Weighted pairwise
a - b : 5
a - c: 3
a - d: 1
b - c: 2
Directed pairwise
a -> b
a -> c
a -> d
b -> c
Temporal pairwise
a - b: 1, 3
a - c: 2
a - d: 2, 3
b - c: 1, 4
Matrix
a b c
a 0 3 5
b 3 0 1
c 5 1 0
Tensor
T=1:
0 3 5
3 0 1
5 1 0
T = 2:
0 1 5
1 0 4
5 4 0
Group
a b c
a c
b c d
Sequential
a b a b
a c a c
d a c d a c
Diffusion
a
b
c
d
b
c
d
Time series
1 2 1 2
1 3 1 3
4 1 3 4 1 3

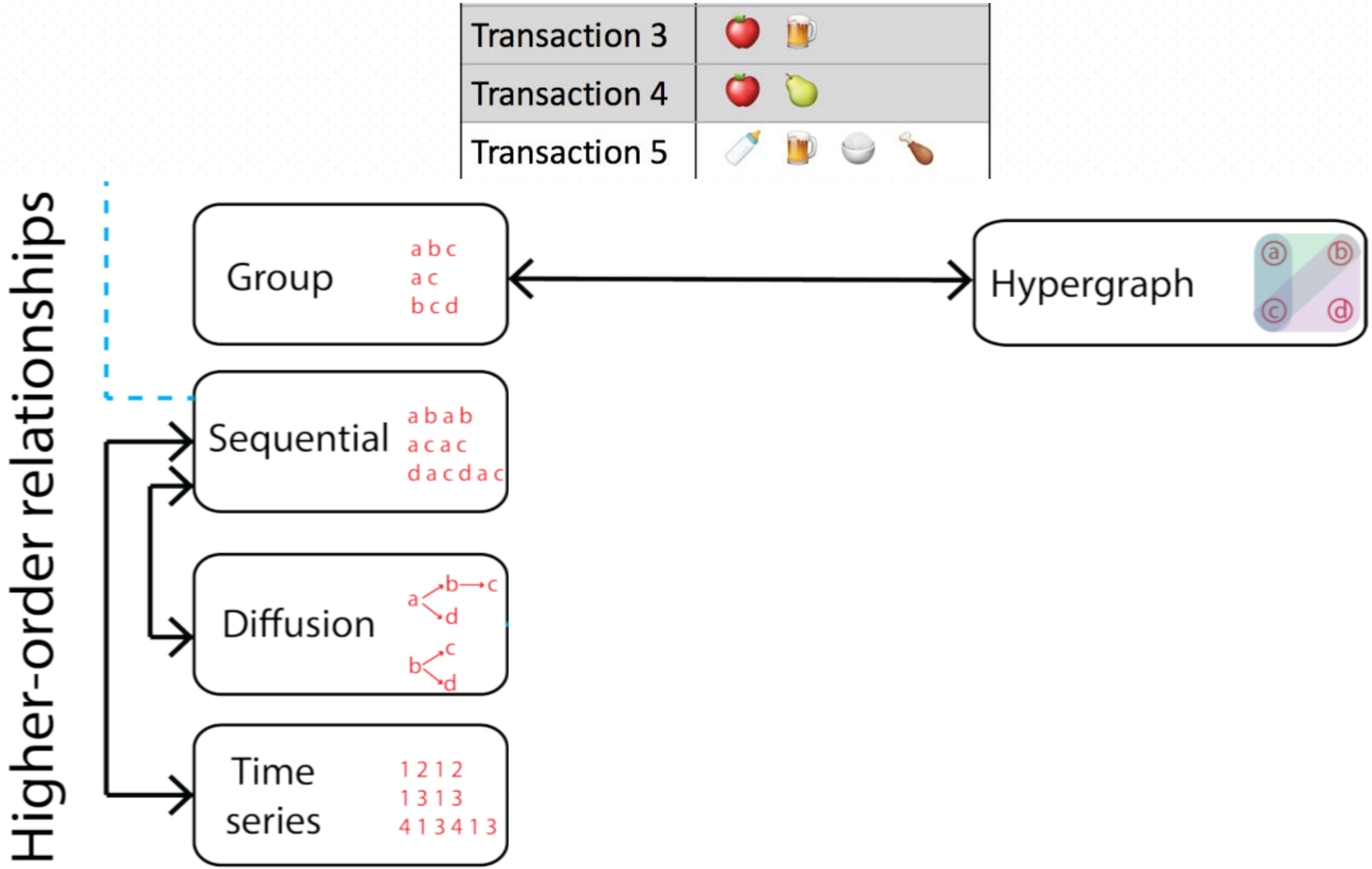

Data
Network
Pairwise
a - b
a - c
a - d
b - c
Weighted pairwise
a - b : 5
a - c: 3
a - d: 1
b - c: 2
Directed pairwise
a -> b
a -> c
a -> d
b -> c
Temporal pairwise
a - b: 1, 3
a - c: 2
a - d: 2, 3
b - c: 1, 4
Matrix
a b c
a 0 3 5
b 3 0 1
c 5 1 0
Tensor
T=1:
0 3 5
3 0 1
5 1 0
T = 2:
0 1 5
1 0 4
5 4 0
Group
a b c
a c
b c d
Sequential
a b a b
a c a c
d a c d a c
Diffusion
a
b
c
d
b
c
d
Time series
1 2 1 2
1 3 1 3
4 1 3 4 1 3


Data
Network
Pairwise
a - b
a - c
a - d
b - c
Weighted pairwise
a - b: 5
a - c: 3
a - d: 1
b - c: 2
Directed pairwise
a -> b
a -> c
a -> d
b -> c
Temporal pairwise
a - b: 1, 3
a - c: 2
a - d: 2, 3
b - c: 1, 4
Matrix
a b c
a 0 3 5
b 3 0 1
c 5 1 0
Tensor
T=1:
0 3 5
3 0 1
5 1 0
T=2:
0 1 5
1 0 4
5 4 0
Group
a b c
a c
b c d
Sequential
a b a b
a c a c
d a c d a c
Diffusion
Time series
1 2 1 2
1 3 1 3
4 1 3 4 1 3
Simple
Weighted
Directed
Temporal
1,3
2
2,3
1,4
Dynamic
T=1
T=2
Heterogeneous

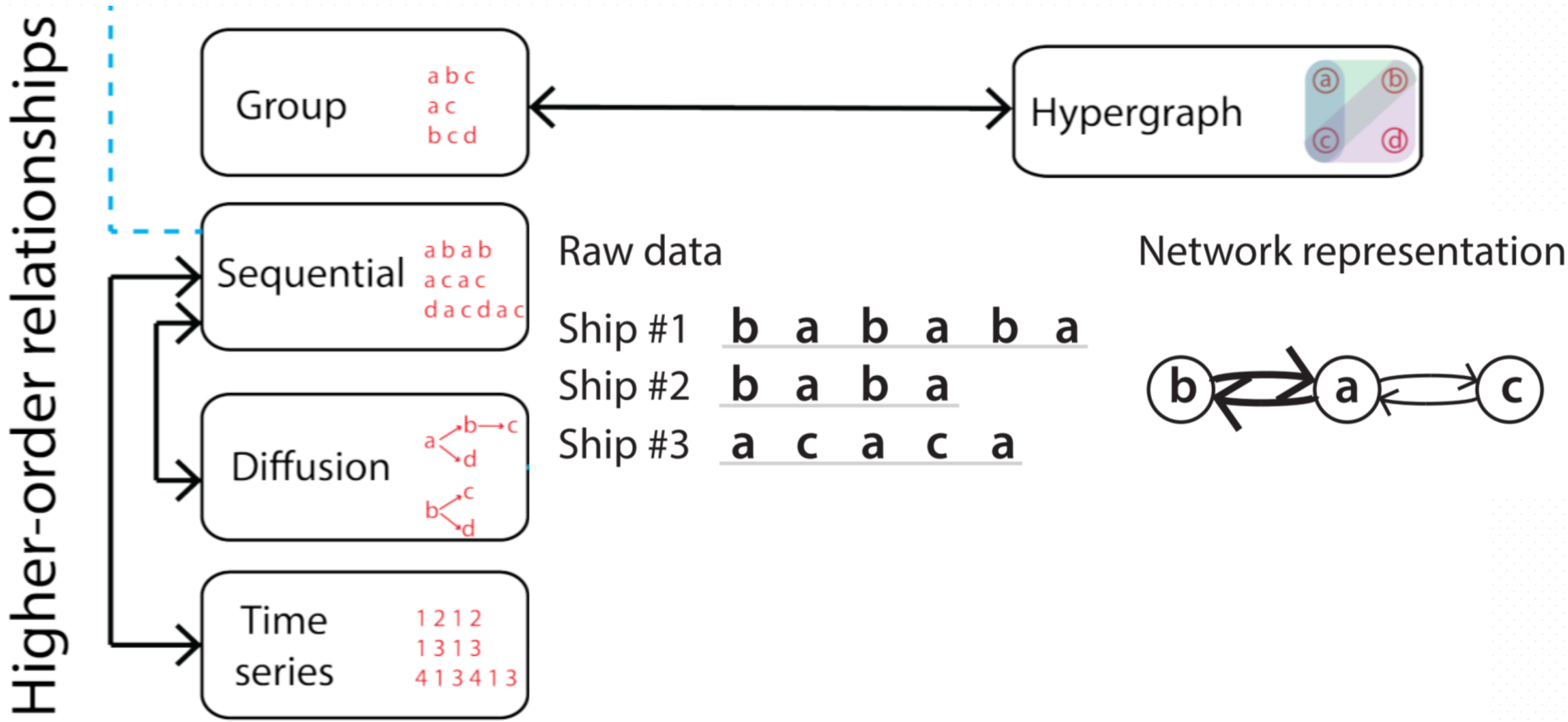

Representation
Data
Network
Pairwise
a - b
a - c
a - d
b - c
Weighted pairwise
a - b: 5
a - c: 3
a - d: 1
b - c: 2
Directed pairwise
a -> b
a -> c
a -> d
b -> c
Temporal pairwise
a - b: 1, 3
a - c: 2
a - d: 2, 3
b - c: 1, 4
Matrix
a b c
a 0 3 5
b 3 0 1
c 5 1 0
Tensor
T=1: T=2:
0 3 5 0 1 5
3 0 1 1 0 4
5 1 0 5 4 0
Group
a b c
a c
b c d
Sequential
a b a b
a c a c
d a c d a c
Diffusion
Time series
1 2 1 2
1 3 1 3
4 1 3 4 1 3
Simple
Weighted
Directed
Temporal
1,3
2
2,3
1,4
Dynamic
T=1
T=2
Heterogeneous

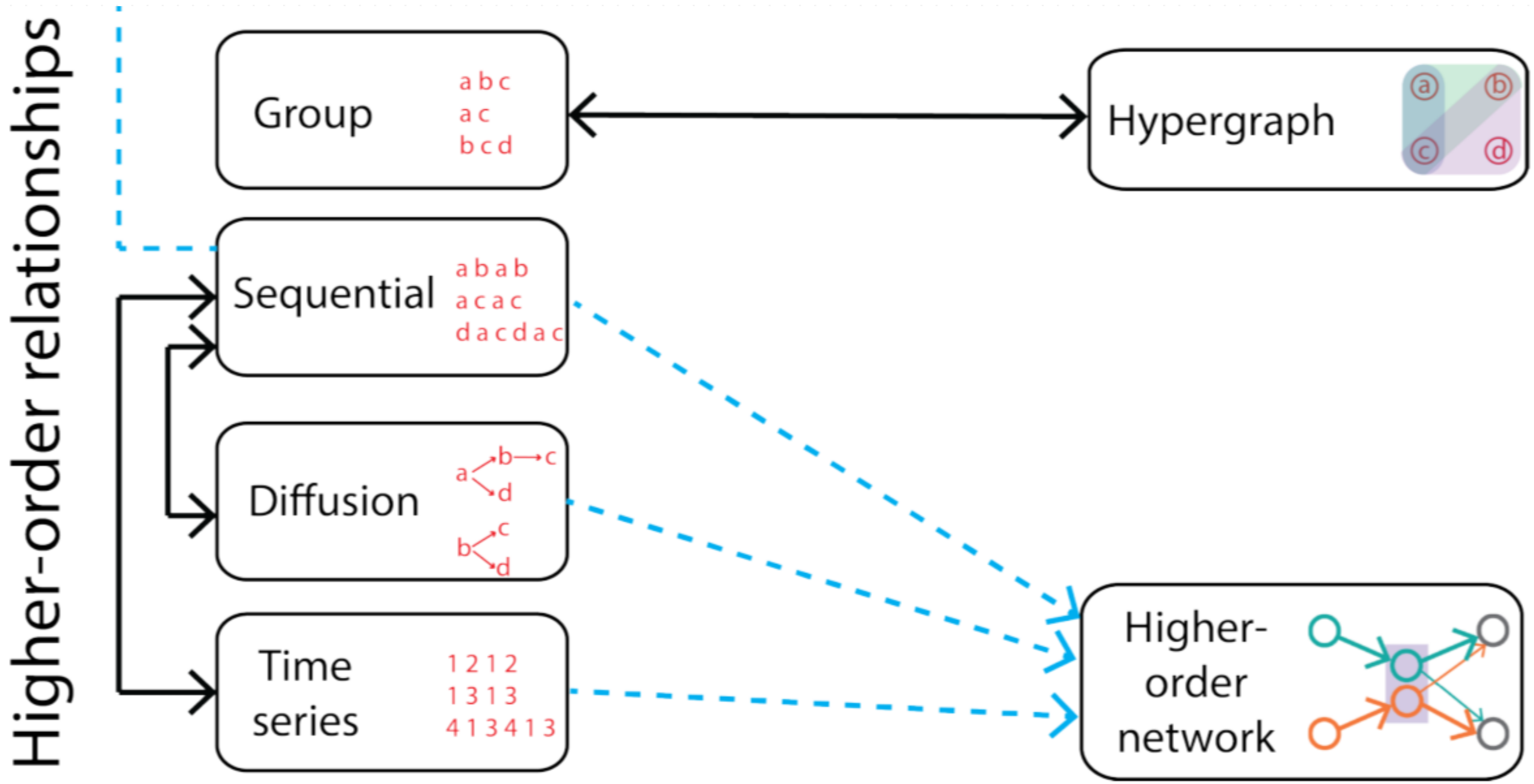


Representation
Raw data
Lossless
Lossy
Network
Pairwise relationships
Higher-order relationships
Pairwise
a - b
a - c
a - d
b - c
Weighted pairwise
a - b : 5
a - c: 3
a - d: 1
b - c: 2
Directed pairwise
a -> b
a -> c
a -> d
b -> c
Temporal pairwise
a - b: 1, 3
a - c: 2
a - d: 2, 3
b - c: 1, 4
Matrix
a b c
a 0 3 5
b 3 0 1
c 5 1 0
Tensor
T=1:
0 3 5
3 0 1
5 1 0
T = 2:
0 1 5
1 0 4
5 4 0
Group
a b c
a c
b c d
Sequential
a b a b
a c a c
d a c d a c
Diffusion
Time series
1 2 1 2
1 3 1 3
4 1 3 4 1 3
Simple
Weighted
Directed
Temporal
Dynamic
T=1
T=2
Heterogeneous



Transaction 3
Transaction 4
Transaction 5
Higher-order relationships
Group
a b c
a c
b c d
Hypergraph
a b c d
Sequential
a b a b
a c a c
d a c d a c
Diffusion
a b c d
b c d
Time series
1 2 1 2
1 3 1 3
4 1 3 4 1 3


Higher-order relationships
Group
a b c
a c
b c d
Hypergraph
a
b
c
d
Sequential
a b a b
a c a c
d a c d a c
Diffusion
a
b
c
d
b
c
d
Time
series
1 2 1 2
1 3 1 3
4 1 3 4 1 3
Raw data
Ship #1 b a b a b a
Ship #2 b a b a
Ship #3 a c a c a
Network representation
b
a
c



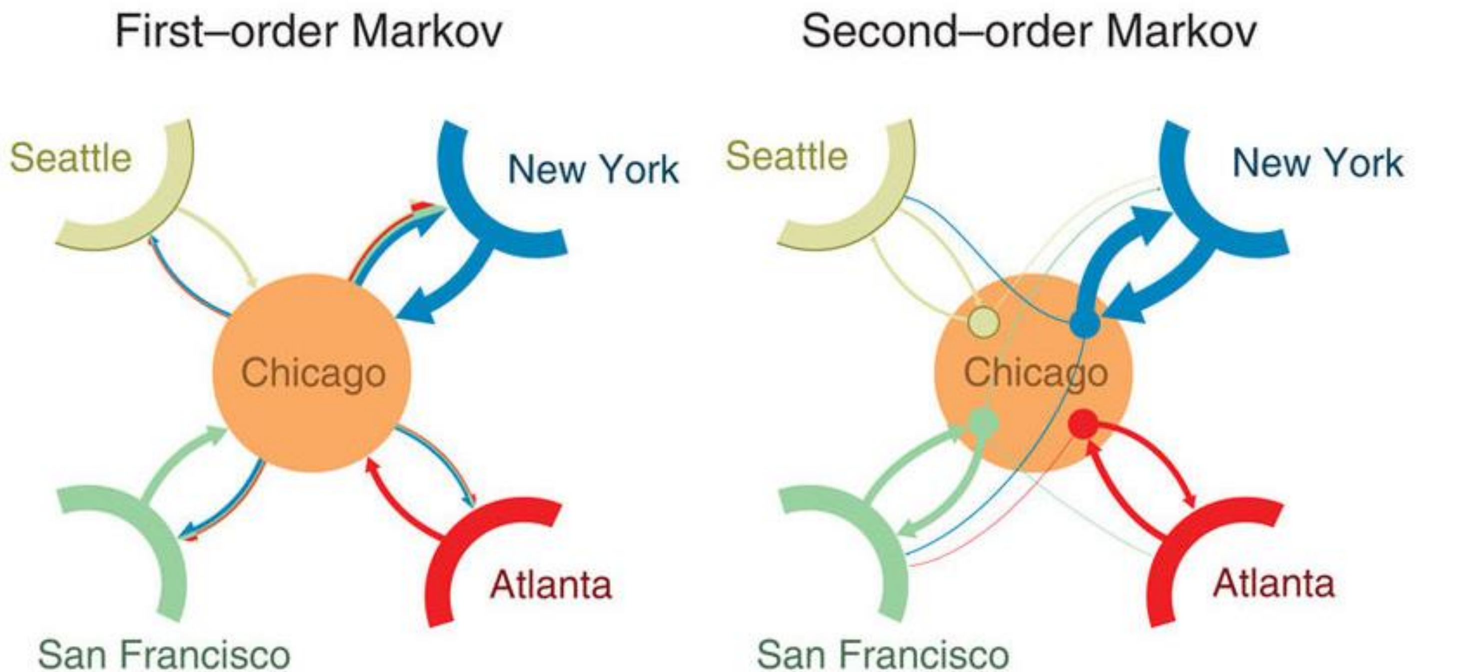
Higher-order relationships
Group
a b c
a c
b c d
Hypergraph
a
b
c
d
Sequential
a b a b
a c a c
d a c d a c
Diffusion
a
b
c
d
b
c
d
Time series
1 2 1 2
1 3 1 3
4 1 3 4 1 3
Higher-order network

Higher-order network

Representing higher-order dependencies in networks

Higher-order network

Fixed-order network

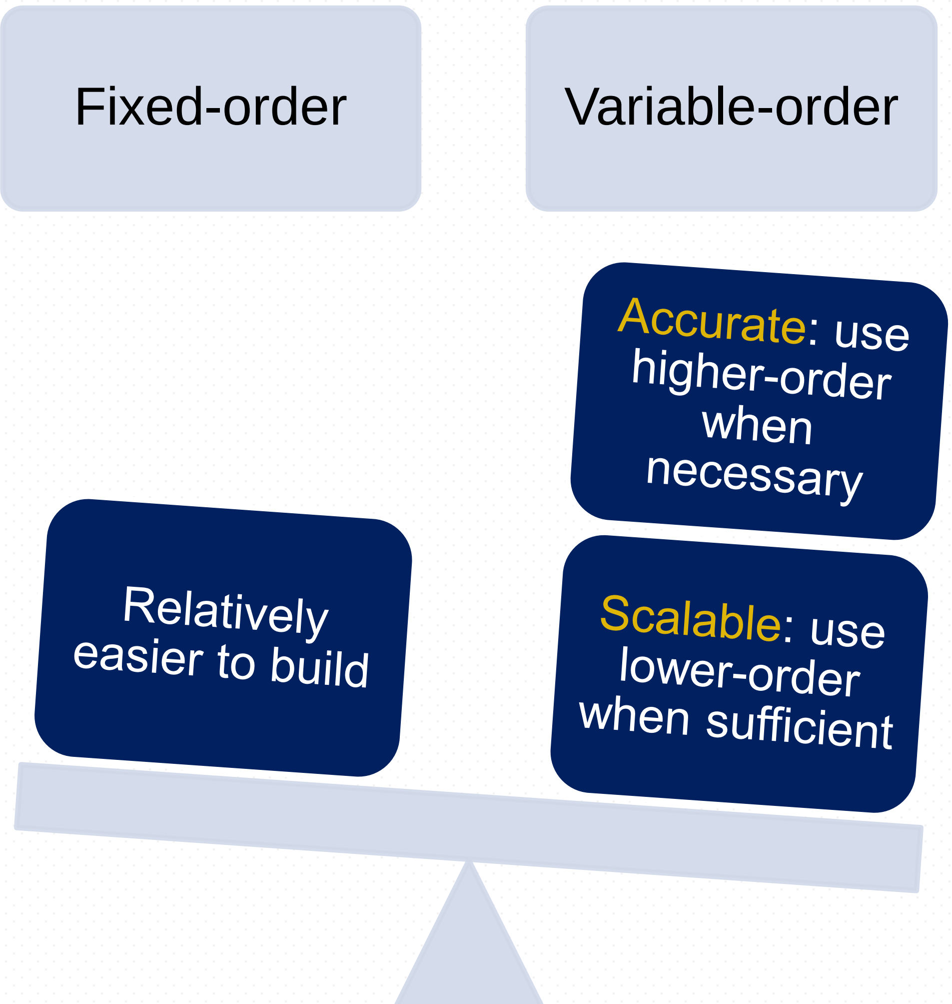


Rosvall et al. (Nature Comm. 2014)

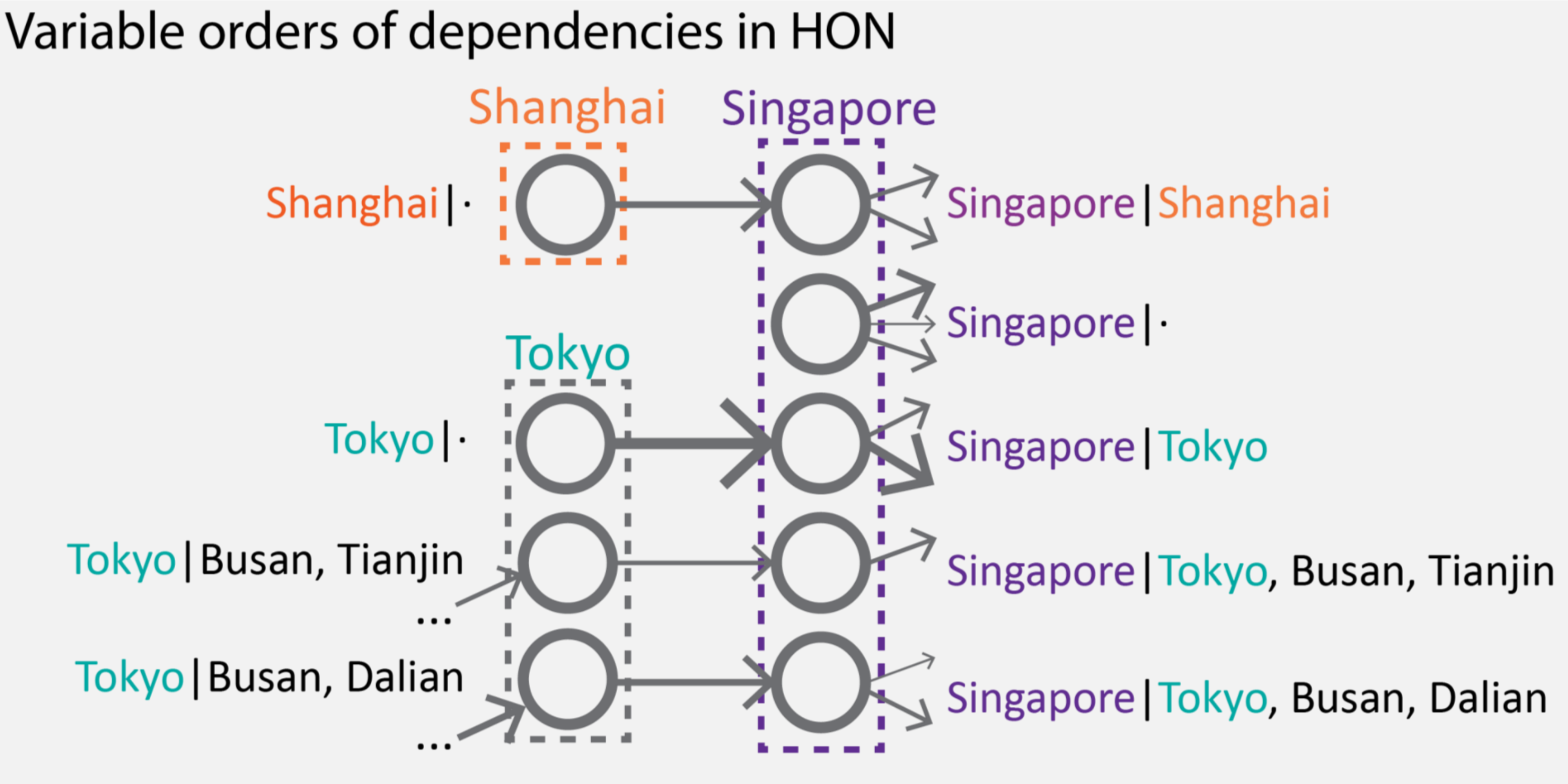
Variable orders in HON

Assuming a fixed order beyond the second order becomes impractical because *“higher-order Markov models are more complex”* due to combinatorial explosion

--- Rosvall et al. (Nature Comm. 2014)



Variable orders in HON

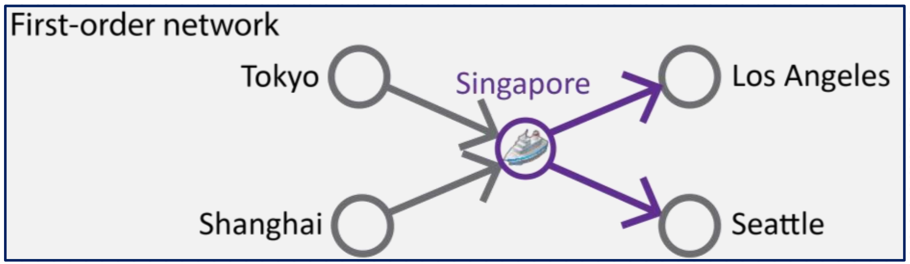


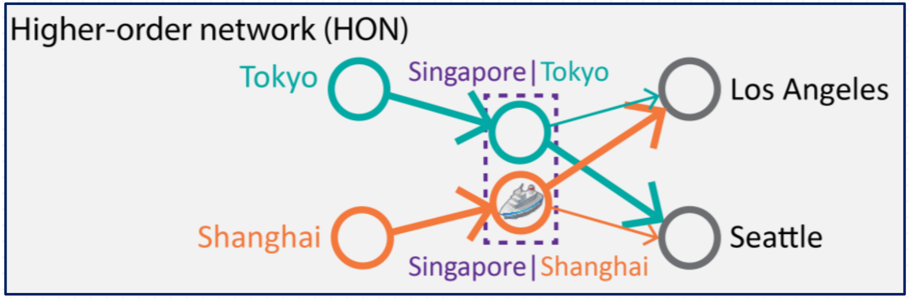
Scalable for big data

Compatible with existing tools

Conventionally: every node represents a single entity (location, state, etc.)

Now: break down nodes into higher-order nodes that carry different dependency relationships





$$P(X_{t+1}=i_{t+1} \mid X_t = i_t) = \frac{W(i_t \to i_{t+1})}{\sum_j W(i_t \to j)}$$

$$P\left(X_{t+1}=j \mid X_t = (i \mid h)\right) = \frac{W(i \mid h \to j)}{\sum_k W(i \mid h \to k)}$$

Only change the node labeling

Takeaways

Higher-order network is:

- More accurate in capturing dynamics in raw data.
- More scalable than fixed-order networks.
- Compatible with existing network algorithms.

Limitations:

- Multiple parameters: maximum order & minimum support.
- Costly to build for very high orders.

Higher-order network optimized revision

Parameter-free, scalable for arbitrarily high order

HON construction workflow

Raw data	Rule extraction	Network wiring	HON
• Sequential data	• Which nodes to split into higher-order nodes, and how high the orders are	• Connect nodes representing different orders	• Use HON like the conventional network for analyses

HON

Raw data A B C A B C A C B

HON

Raw data A B C A B C A C B

Build observation

Build all 1st-order	Build all 2nd-order	Build all 3rd-order
A -> B: 2	A|C -> B: 1	A|C.B -> B: 1
A -> C: 1	A|C -> C: 1	A|C.B -> C: 1
B -> C: 2	B|A -> C: 2	B|A.C -> C: 1
C -> A: 2	C|B -> A: 2	C|B.A -> A: 2
C -> B: 1	C|A -> B: 1	C|A.C -> B: 1

HON

Raw data

A B C A B C A C B

Build observation

Build all 1^{st}-order
A -> B: 2
A -> C: 1
B -> C: 2
C -> A: 2
C -> B: 1

Build all 2^{nd}-order
A|C -> B: 1
A|C -> C: 1
B|A -> C: 2
C|B -> A: 2
C|A -> B: 1

Build all 3^{rd}-order
A|C.B -> B: 1
A|C.B -> C: 1
B|A.C -> C: 1
C|B.A -> A: 2
C|A.C -> B: 1

Build distribution

Build all 1^{st}-order
A -> B: 0.67
A -> C: 0.33
B -> C: 1
C -> A: 0.67
C -> B: 0.33

Build all 2^{nd}-order
A|C -> B: 0.5
A|C -> C: 0.5
B|A -> C: 1
C|B -> A: 0.67
C|A -> B: 0.33

Build all 3^{rd}-order
A|C.B -> B: 0.5
A|C.B -> C: 0.5
B|A.C -> C: 1
C|B.A -> A: 0.67
C|A.C -> B: 0.33

HON

Raw data: A B C A B C A C B

Build observation	Build all 1st-order A -> B: 2 A -> C: 1 B -> C: 2 C -> A: 2 C -> B: 1	Build all 2nd-order A|C -> B: 1 A|C -> C: 1 B|A -> C: 2 C|B -> A: 2 C|A -> B: 1	Build all 3rd-order A|C.B -> B: 1 A|C.B -> C: 1 B|A.C -> C: 1 C|B.A -> A: 2 C|A.C -> B: 1
Build distribution	Build all 1st-order A -> B: 0.67 A -> C: 0.33 B -> C: 1 C -> A: 0.67 C -> B: 0.33	Build all 2nd-order A|C -> B: 0.5 A|C -> C: 0.5 B|A -> C: 1 C|B -> A: 0.67 C|A -> B: 0.33	Build all 3rd-order A|C.B -> B: 0.5 A|C.B -> C: 0.5 B|A.C -> C: 1 C|B.A -> A: 0.67 C|A.C -> B: 0.33
Rule growing	Grow all 1st-order A -> B: 0.67 A -> C: 0.33 B -> C: 1 C -> A: 0.67 C -> B: 0.33	Grow all 2nd-order A|C -> B: 0.5 A|C -> C: 0.5 B|A -> C: 1 C|B -> A: 0.67 C|A -> B: 0.33	Grow all 3rd-order A|C.B -> B: 0.5 A|C.B -> C: 0.5 B|A.C -> C: 1 C|B.A -> A: 0.67 C|A.C -> B: 0.33

HON

Raw data: A B C A B C A C B

Build observation	Build all 1st-order	Build all 2nd-order	Build all 3rd-order
	A -> B: 2	A|C -> B: 1	A|C.B -> B: 1
	A -> C: 1	A|C -> C: 1	A|C.B -> C: 1
	B -> C: 2	B|A -> C: 2	B|A.C -> C: 1
	C -> A: 2	C|B -> A: 2	C|B.A -> A: 2

$$D_{KL}(ExtDistr||Distr) \lessgtr \frac{NewOrder}{\log_2(1+\sum C[ExtSource][*])}$$

A -> C: 0.33	A|C -> C: 0.5	A|C.B -> C: 0.5
B -> C: 1	B|A -> C: 1	B|A.C -> C: 1
C -> A: 0.67	C|B -> A: 0.67	C|B.A -> A: 0.67
C -> B: 0.33	C|A -> B: 0.33	C|A.C -> B: 0.33

Rule growing	Grow all 1st-order	Grow all 2nd-order	Grow all 3rd-order
	A -> B: 0.67	A|C -> B: 0.5	A|C.B -> B: 0.5
	A -> C: 0.33	A|C -> C: 0.5	A|C.B -> C: 0.5
	B -> C: 1	B|A -> C: 1	B|A.C -> C: 1
	C -> A: 0.67	C|B -> A: 0.67	C|B.A -> A: 0.67
	C -> B: 0.33	C|A -> B: 0.33	C|A.C -> B: 0.33

HON

Raw data: A B C A B C A C B

Build observation

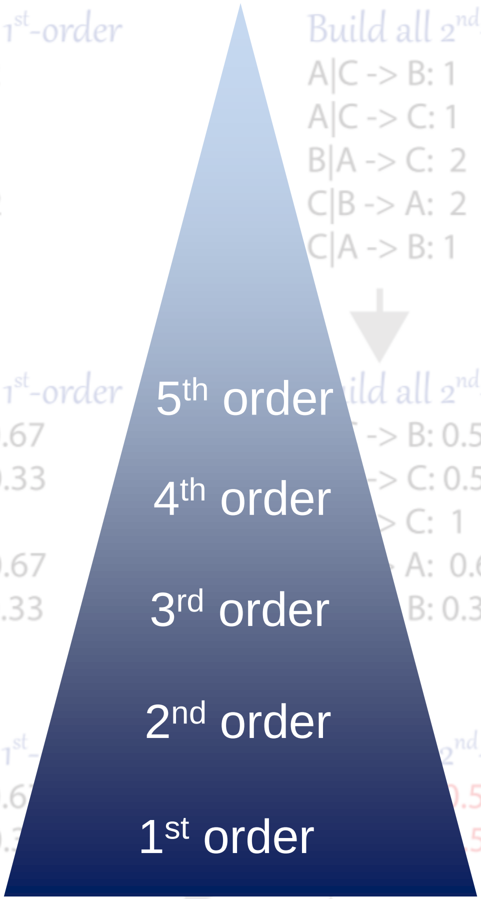
Build all 1st-order	Build all 2nd-order	Build all 3rd-order
A -> B: 2	A|C -> B: 1	A|C.B -> B: 1
A -> C: 1	A|C -> C: 1	A|C.B -> C: 1
B -> C: 2	B|A -> C: 2	B|A.C -> C: 1
C -> A: 2	C|B -> A: 2	C|B.A -> A: 2
C -> B: 1	C|A -> B: 1	C|A.C -> B: 1

Build distribution

Build all 1st-order	...ild all 2nd-order	Build all 3rd-order
A -> B: 0.67	... -> B: 0.5	A|C.B -> B: 0.5
A -> C: 0.33	...> C: 0.5	A|C.B -> C: 0.5
B -> C: 1	...> C: 1	B|A.C -> C: 1
C -> A: 0.67	... A: 0.67	C|B.A -> A: 0.67
C -> B: 0.33	... B: 0.33	C|A.C -> B: 0.33

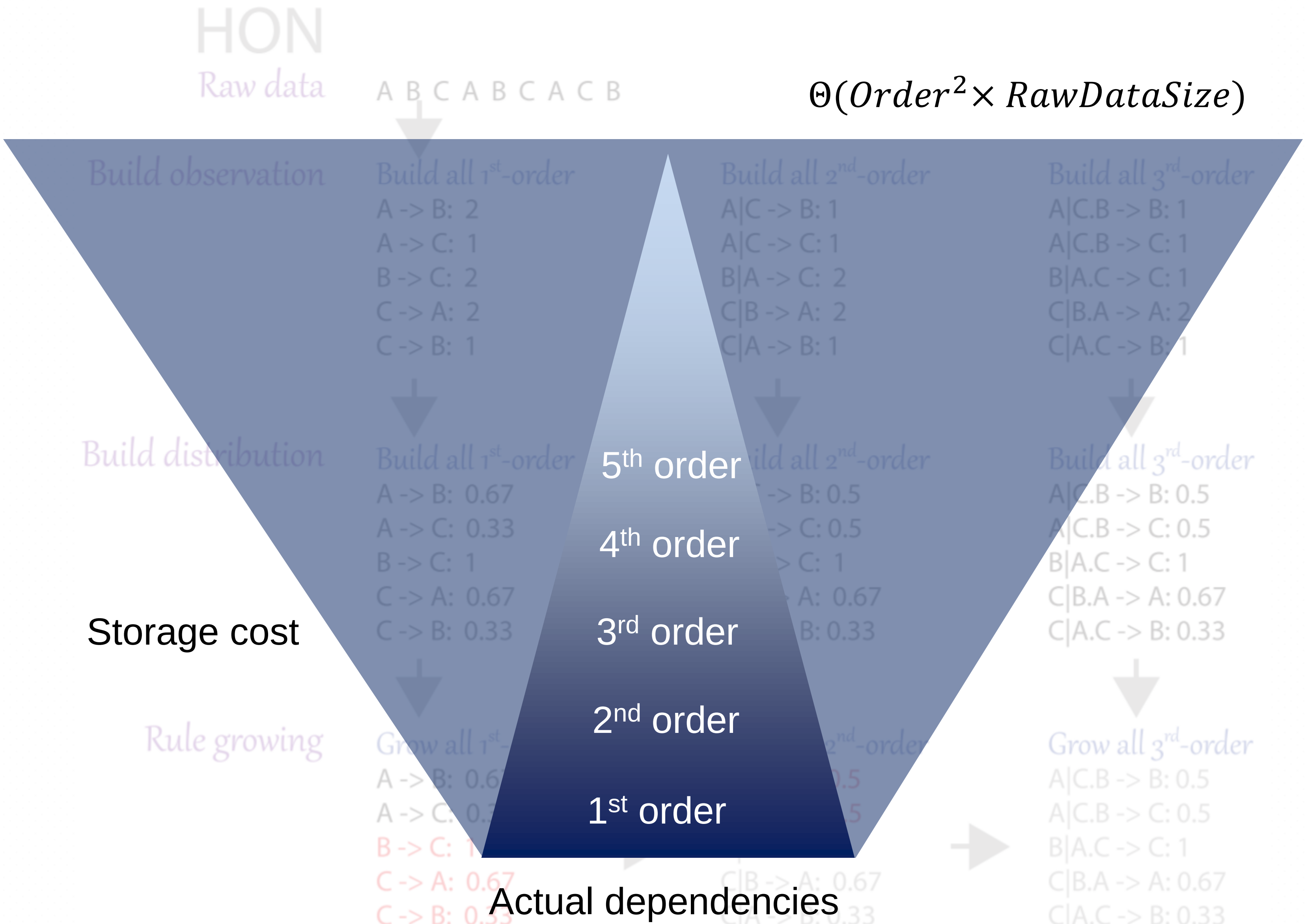
Rule growing

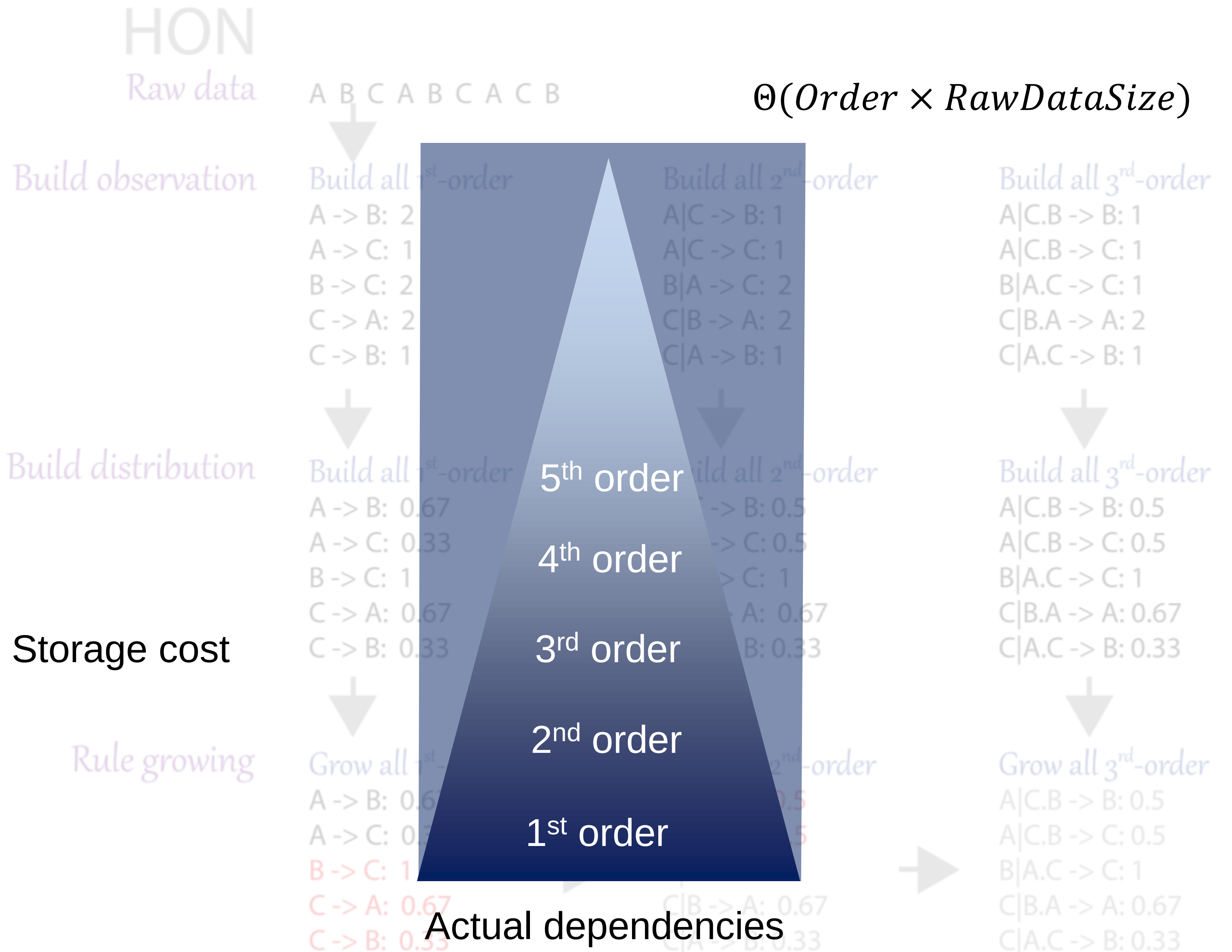
Grow all 1st-...	...2nd-order	Grow all 3rd-order
A -> B: 0.6...	...0.5	A|C.B -> B: 0.5
A -> C: 0.3...	...5	A|C.B -> C: 0.5
B -> C: 1		B|A.C -> C: 1
C -> A: 0.67	C|B -> A: 0.67	C|B.A -> A: 0.67
C -> B: 0.33	C|A -> B: 0.33	C|A.C -> B: 0.33



Actual dependencies

$\Theta(Order^2 \times RawDataSize)$





HON
Raw data
A B C A B C A C B
$\Theta(Order \times RawDataSize)$
Build observation
Build distribution
Rule growing
Storage cost
5th order
4th order
3rd order
2nd order
1st order
Actual dependencies

HON

Raw data: A B C A B C A C B

Build observation

Build all 1st-order	Build all 2nd-order	Build all 3rd-order
A -> B: 2	A|C -> B: 1	A|C.B -> B: 1
A -> C: 1	A|C -> C: 1	A|C.B -> C: 1
B -> C: 2	B|A -> C: 2	B|A.C -> C: 1
C -> A: 2	C|B -> A: 2	C|B.A -> A: 2

$$
\begin{aligned}
max(D_{KL}(ExtDistr||Distr)) &= max(\sum_{i \in Distr} P_{ExtDistr}(i) \times log_2 \frac{P_{ExtDistr}(i)}{P_{Distr}(i)}) \\
&= 1 \times log_2 \frac{1}{min(P_{Distr}(i))} + 0 + 0 + \ldots \\
&= -log_2(min(P_{Distr}(i)))
\end{aligned}
$$

A -> B: 0.67	A|C -> B: 0.5	A|C.B -> B: 0.5
A -> C: 0.33	A|C -> C: 0.5	A|C.B -> C: 0.5
B -> C: 1	B|A -> C: 1	B|A.C -> C: 1
C -> A: 0.67	C|B -> A: 0.67	C|B.A -> A: 0.67
C -> B: 0.33	C|A -> B: 0.33	C|A.C -> B: 0.33

HON

Raw data

A B C A B C A C B

Build observation

Build all 1st-order	Build all 2nd-order	Build all 3rd-order
A -> B: 2	A|C -> B: 1	A|C.B -> B: 1
A -> C: 1	A|C -> C: 1	A|C.B -> C: 1
B -> C: 2	B|A -> C: 2	B|A.C -> C: 1
C -> A: 2	C|B -> A: 2	C|B.A -> A: 2

$$-log_2(min(P_{Distr}(i))) \lessgtr \frac{NewOrder}{\log_2(1+\sum C[ExtSource][*])}$$

A -> C: 0.33	A|C -> C: 0.5	A|C.B -> C: 0.5
B -> C: 1	B|A -> C: 1	B|A.C -> C: 1
C -> A: 0.67	C|B -> A: 0.67	C|B.A -> A: 0.67
C -> B: 0.33	C|A -> B: 0.33	C|A.C -> B: 0.33

Rule growing

Grow all 1st-order	Grow all 2nd-order	Grow all 3rd-order
A -> B: 0.67	A|C -> B: 0.5	A|C.B -> B: 0.5
A -> C: 0.33	A|C -> C: 0.5	A|C.B -> C: 0.5
B -> C: 1	B|A -> C: 1	B|A.C -> C: 1
C -> A: 0.67	C|B -> A: 0.67	C|B.A -> A: 0.67
C -> B: 0.33	C|A -> B: 0.33	C|A.C -> B: 0.33

HONOR

Raw data A B C A B C A C B

HONOR

Raw data   A B C A B C A C B

Build observation   Build all 1st-order

A -> B: 2

A -> C: 1

B -> C: 2

C -> A: 2

C -> B: 1

HONOR

Raw data: A B C A B C A C B

Build observation: Build all 1^{st}-order

A -> B: 2
A -> C: 1
B -> C: 2
C -> A: 2
C -> B: 1

Build distribution: Build all 1^{st}-order

A -> B: 0.67
A -> C: 0.33
B -> C: 1
C -> A: 0.67
C -> B: 0.33

HONOR

Raw data

A B C A B C A C B

Build observation

Build all 1st-order

A -> B: 2
A -> C: 1
B -> C: 2
C -> A: 2
C -> B: 1

Build distribution

Build all 1st-order

A -> B: 0.67
A -> C: 0.33
B -> C: 1
C -> A: 0.67
C -> B: 0.33

Rule growing

Grow all 1st-order

A -> B: 0.67
A -> C: 0.33
B -> C: 1
C -> A: 0.67
C -> B: 0.33

HONOR

Raw data

A B C A B C A C B

Build observation

Build all 1st-order	Build 2nd-order on demand
A -> B: 2	A|C -> B: 1
A -> C: 1	A|C -> C: 1
B -> C: 2	B|A -> C: 2
C -> A: 2	C|B -> A: 2
C -> B: 1	C|A -> B: 1

Build distribution

Build all 1st-order	Build 2nd-order on demand
A -> B: 0.67	A|C -> B: 0.5
A -> C: 0.33	A|C -> C: 0.5
B -> C: 1	B|A -> C: 1
C -> A: 0.67	C|B -> A: 0.67
C -> B: 0.33	C|A -> B: 0.33

Rule growing

Grow all 1st-order	Grow 2nd-order on demand
A -> B: 0.67	A|C -> B: 0.5
A -> C: 0.33	A|C -> C: 0.5
B -> C: 1	B|A -> C: 1
C -> A: 0.67	C|B -> A: 0.67
C -> B: 0.33	C|A -> B: 0.33

HONOR

Raw data

A B C A B C A C B

Build observation

Build all 1st-order	Build 2nd-order on demand
A -> B: 2	A|C -> B: 1
A -> C: 1	A|C -> C: 1
B -> C: 2	B|A -> C: 2
C -> A: 2	C|B -> A: 2
C -> B: 1	C|A -> B: 1

Build distribution

Build all 1st-order	Build 2nd-order on demand
A -> B: 0.67	A|C -> B: 0.5
A -> C: 0.33	A|C -> C: 0.5
B -> C: 1	B|A -> C: 1
C -> A: 0.67	C|B -> A: 0.67
C -> B: 0.33	C|A -> B: 0.33

Rule growing

Grow all 1st-order	Grow 2nd-order on demand
A -> B: 0.67	A|C -> B: 0.5
A -> C: 0.33	A|C -> C: 0.5
B -> C: 1	B|A -> C: 1
C -> A: 0.67	C|B -> A: 0.67
C -> B: 0.33	C|A -> B: 0.33

Max divergence < threshold

Stop rule growing

Takeaways

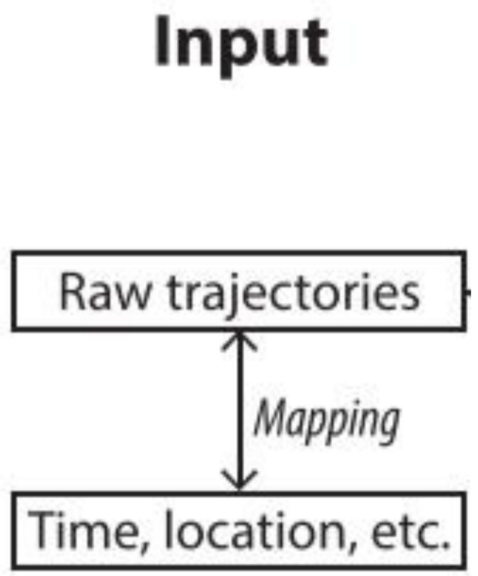
HONOR is:

- Parameter-free version of HON.
- More scalable for big data
- Supports arbitrarily high order.
- Lazy evaluation reduces actual search space.

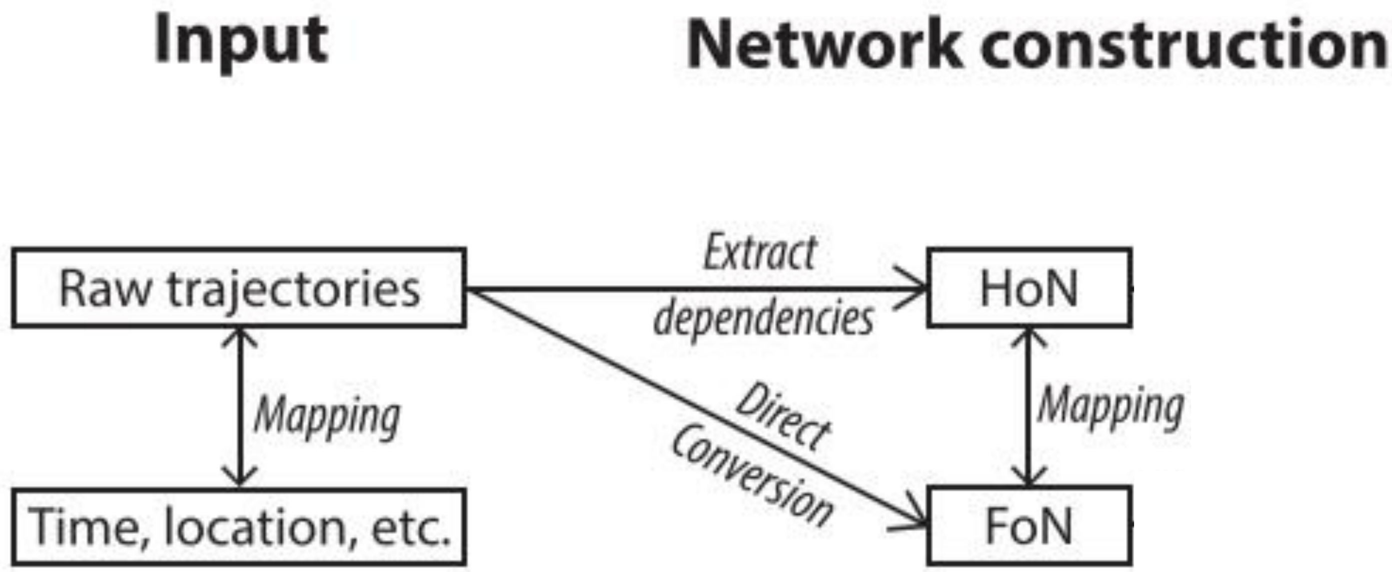
HONVis

Visualization & interactive exploration software

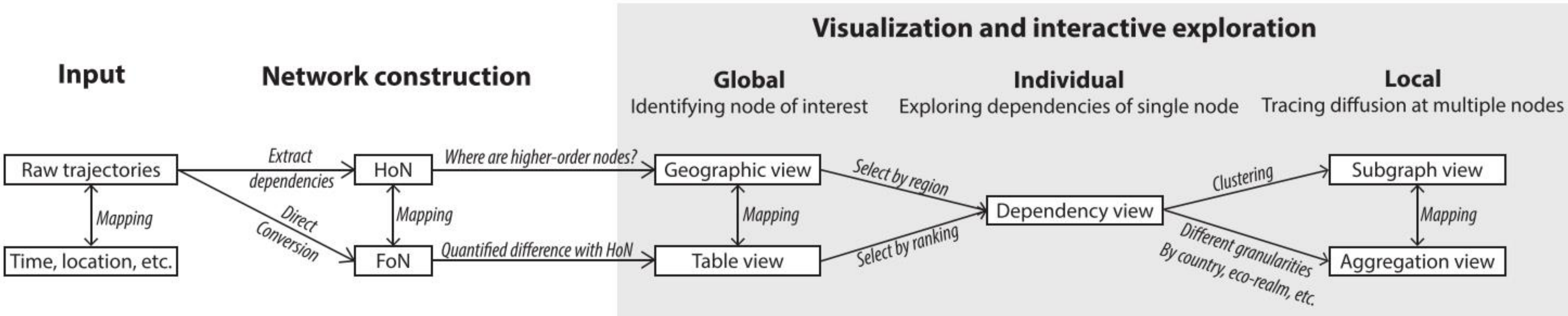
HoNVis framework



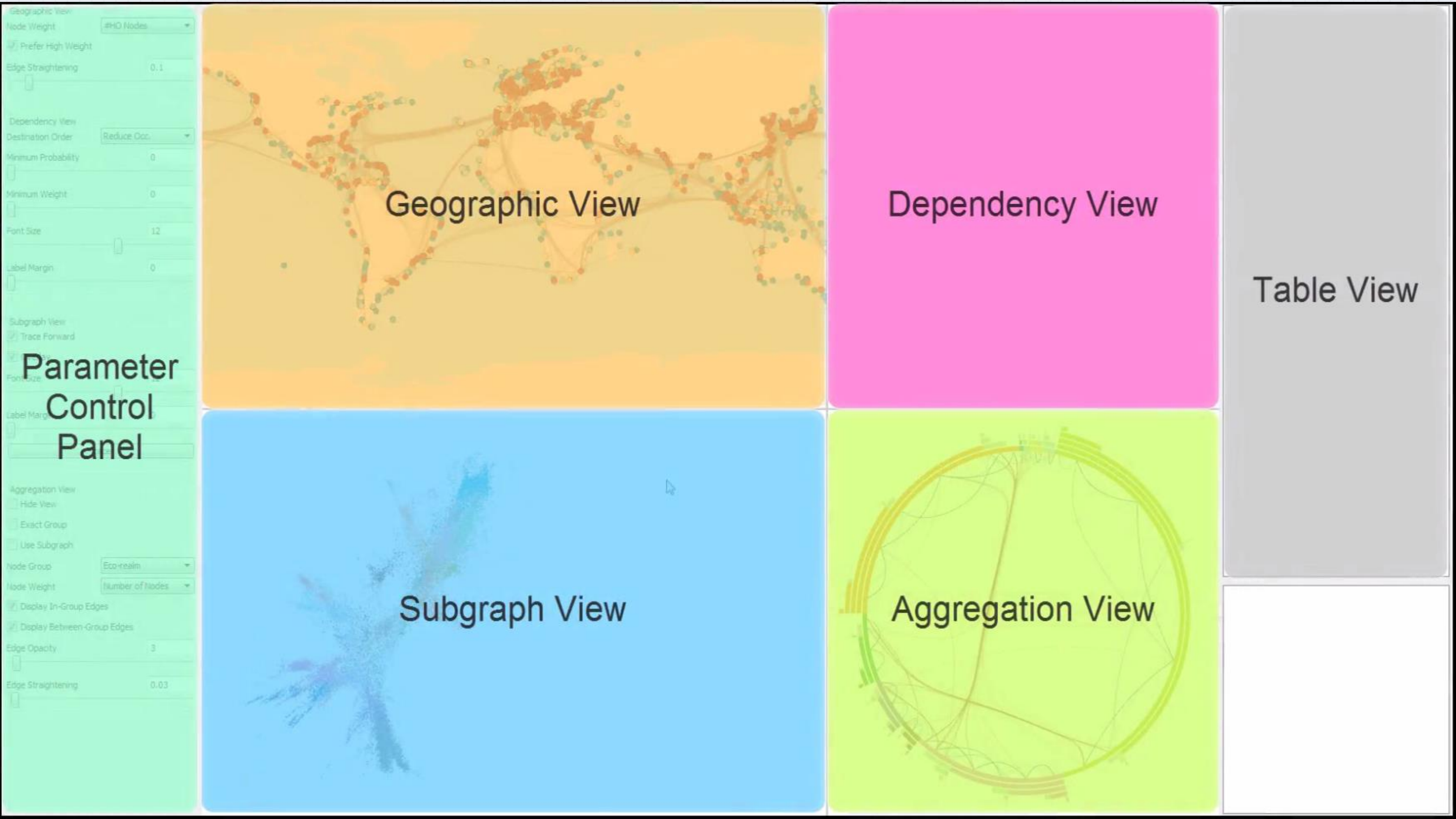
HoNVis framework



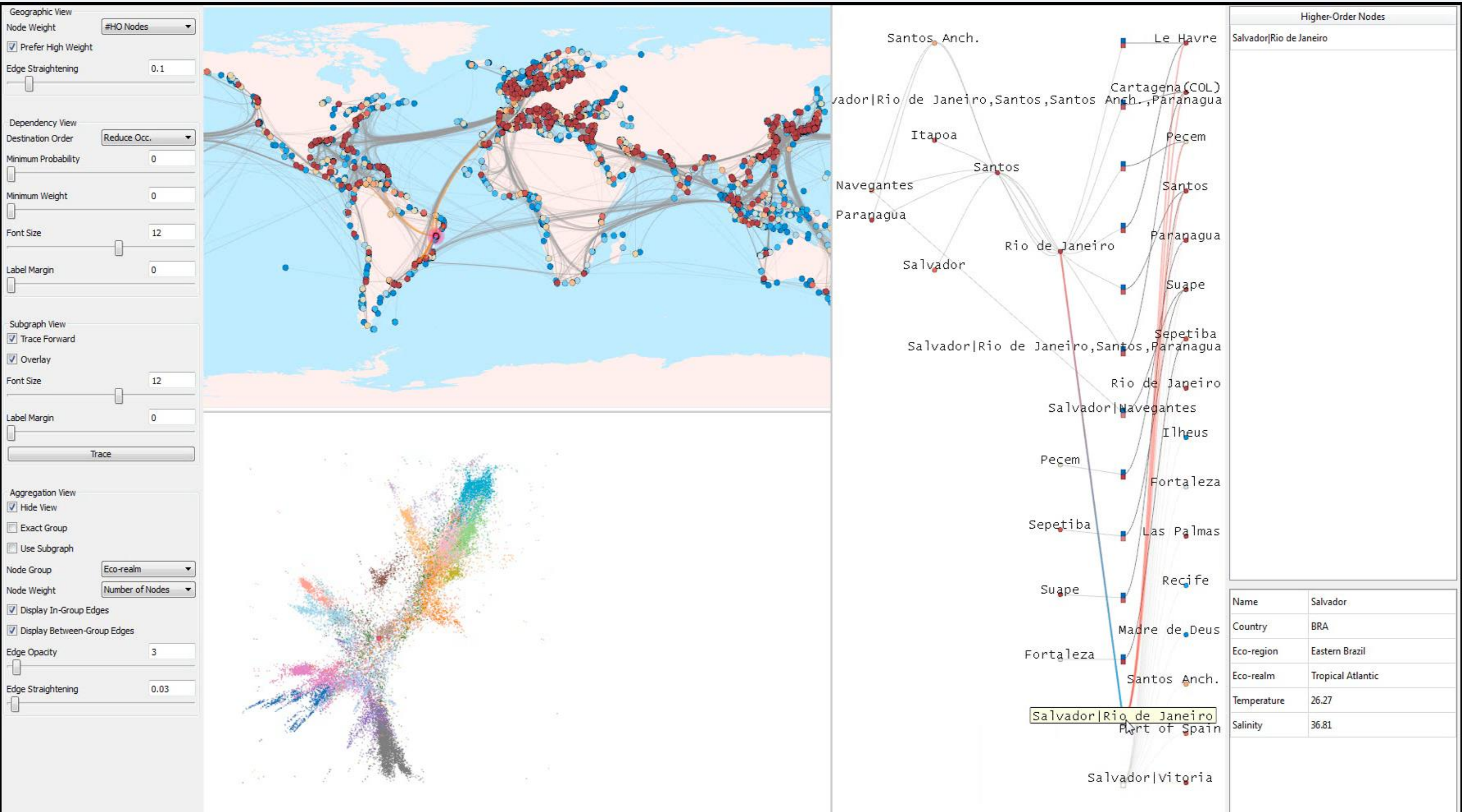
HoNVis framework



HoNVis interface



Visualization & interactive exploration



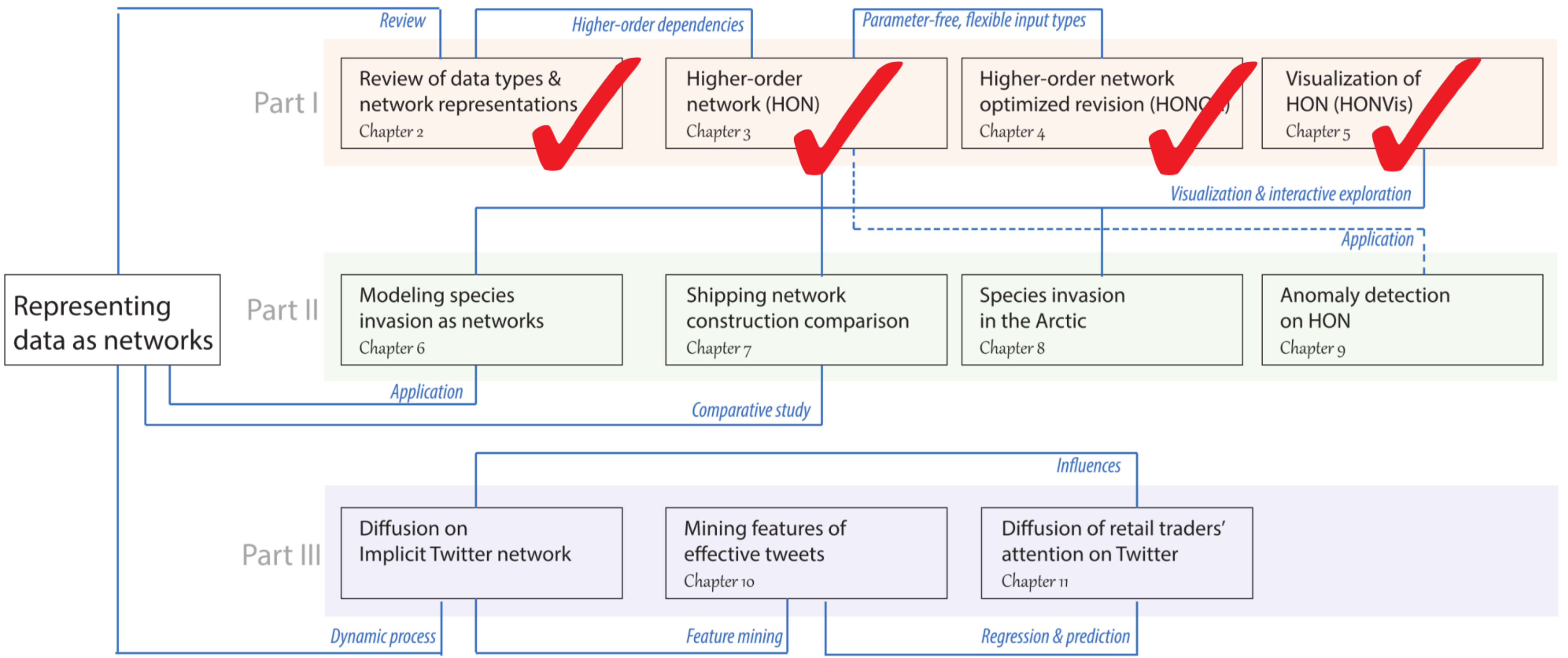
Takeaways

HONVis is:

The first visualization software for HON.

Facilitates interactive explorations.

Overview


Review
Higher-order dependencies
Parameter-free, flexible input types
Part I
Review of data types & network representations
Chapter 2
Higher-order network (HON)
Chapter 3
Higher-order network optimized revision (HON+)
Chapter 4
Visualization of HON (HONVis)
Chapter 5
Visualization & interactive exploration
Application
Representing data as networks
Part II
Modeling species invasion as networks
Chapter 6
Shipping network construction comparison
Chapter 7
Species invasion in the Arctic
Chapter 8
Anomaly detection on HON
Chapter 9
Application
Comparative study
Influences
Part III
Diffusion on Implicit Twitter network
Mining features of effective tweets
Chapter 10
Diffusion of retail traders' attention on Twitter
Chapter 11
Dynamic process
Feature mining
Regression & prediction

Part II

Insights in real-world applications

Species invasion network

Non-indigenous species risk assessment & prediction framework (NIS-RAPS)

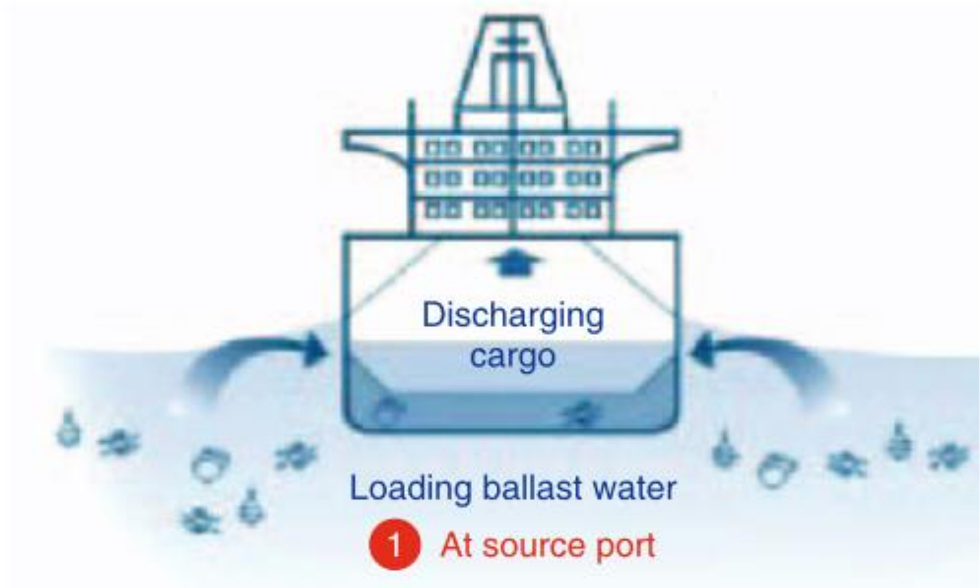
Invasive species

$120 billion / year damage & control costs

Zebra mussels @ Great Lakes
Clogging water pipes, attach to boats

Photos: Great Lakes Environmental Research Lab; TIME & LIFE Images, Getty Images

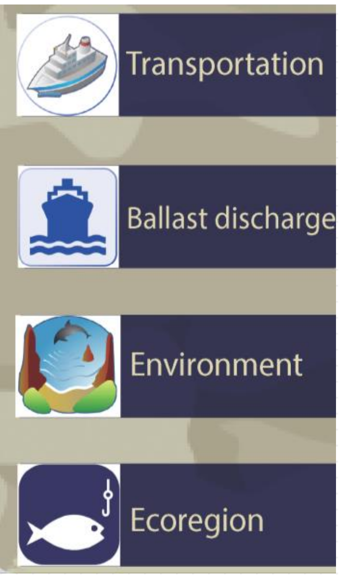
Ship-borne species invasion



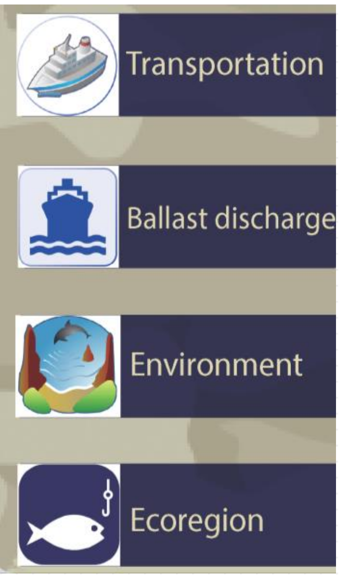
Picture from GloBallast Programme 2002

Ship-borne species invasion

Picture from GloBallast Programme 2002



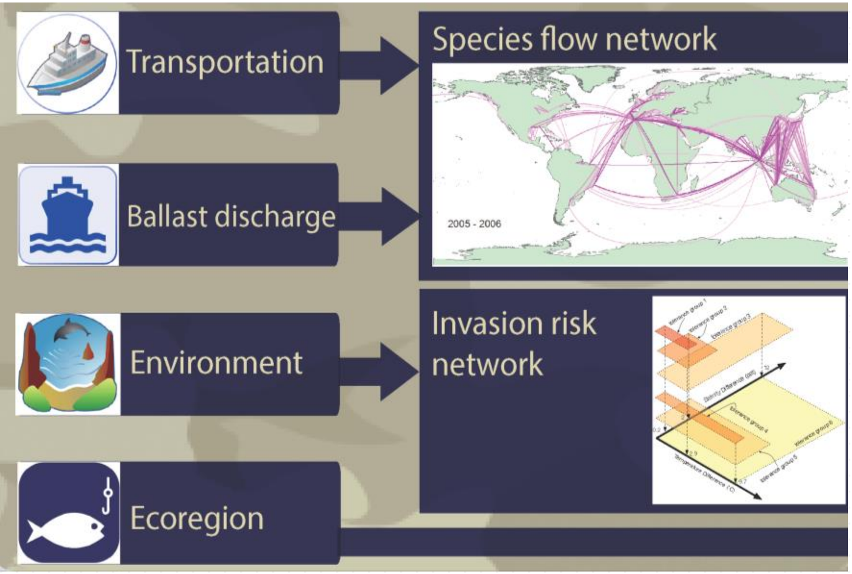
Transportation
Ballast discharge
Environment
Ecoregion

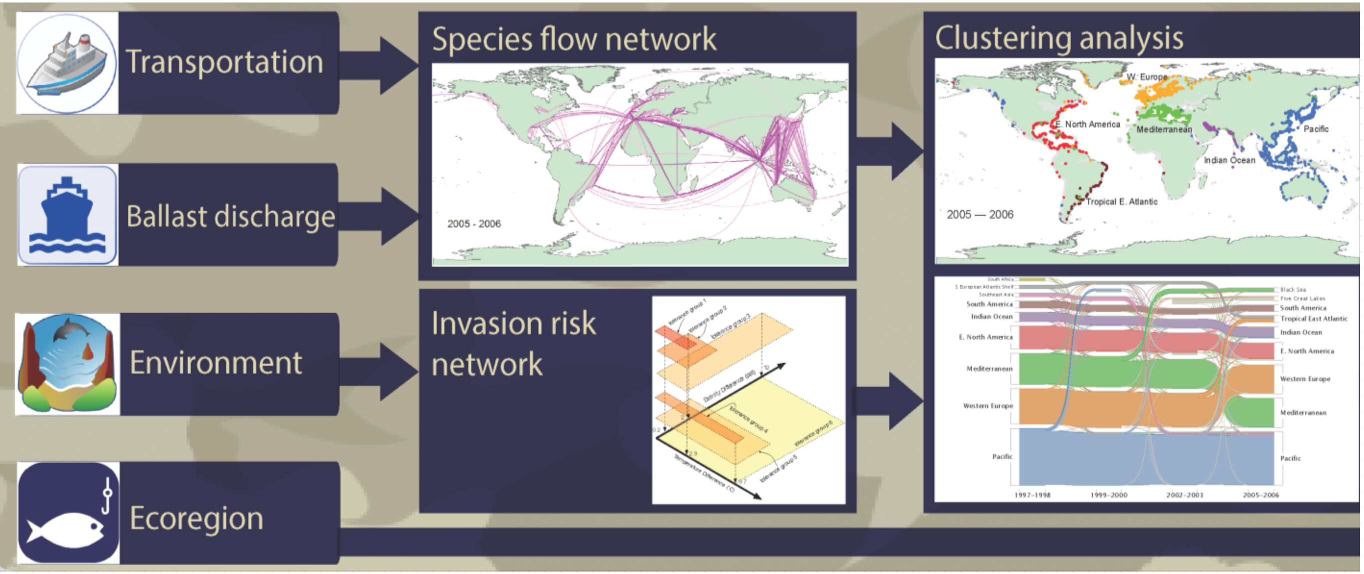


Probability of vessel v introducing species from port i to port j

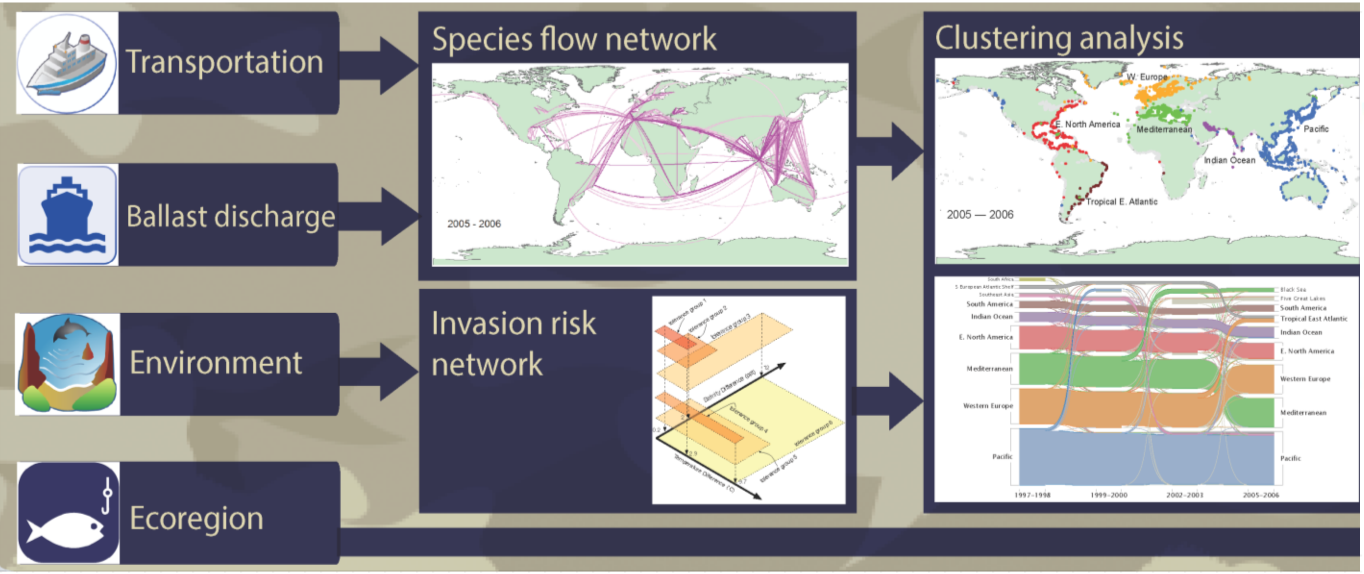
$$p_{ij}^{(v)} = \underbrace{\rho_{ij}^{(v)}}_{\text{Mgmt efficacy}} \underbrace{\left(1 - e^{-\lambda D_{ij}^{(v)}}\right)}_{\text{Ballast discharge}} \underbrace{e^{-\mu \Delta t_{ij}^{(v)}}}_{\text{Mortality}}$$

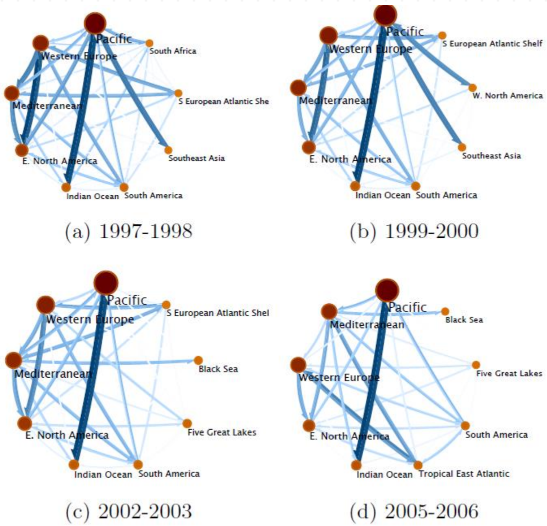
* Invasion risk estimation inspired by Seebens et al. (2013)


Transportation
Ballast discharge
Environment
Ecoregion
Species flow network
2005 - 2006
Invasion risk
network



* Clustering uses MapEquation by Rosvall et al. (2008)

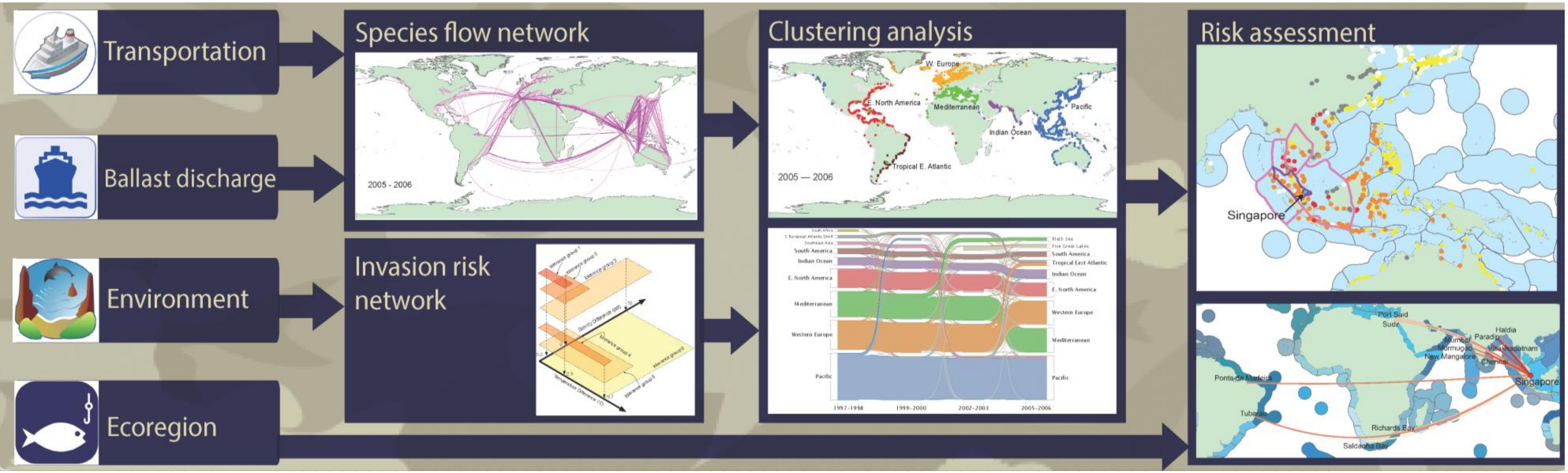




(a) 1997-1998 (b) 1999-2000

(c) 2002-2003 (d) 2005-2006

* Clustering uses MapEquation by Rosvall et al. (2008)


Transportation
Ballast discharge
Environment
Ecoregion
Species flow network
2005 - 2006
Invasion risk network
Clustering analysis
W. Europe
E. North America
Mediterranean
Pacific
Indian Ocean
Tropical E. Atlantic
2005 — 2006
Risk assessment
Singapore

Takeaways

NIS-RAPS:

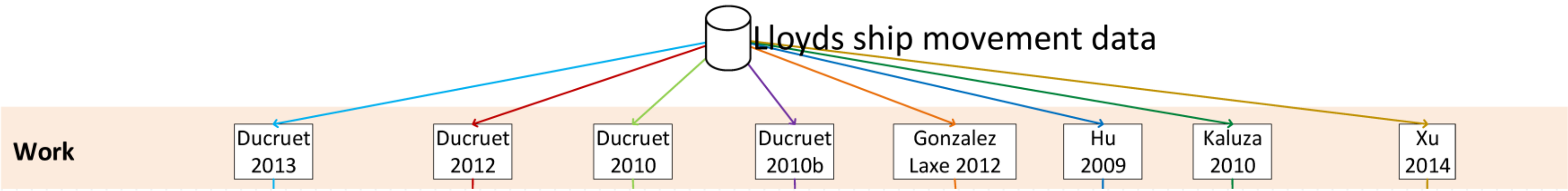
Integrates multiple sources of data.

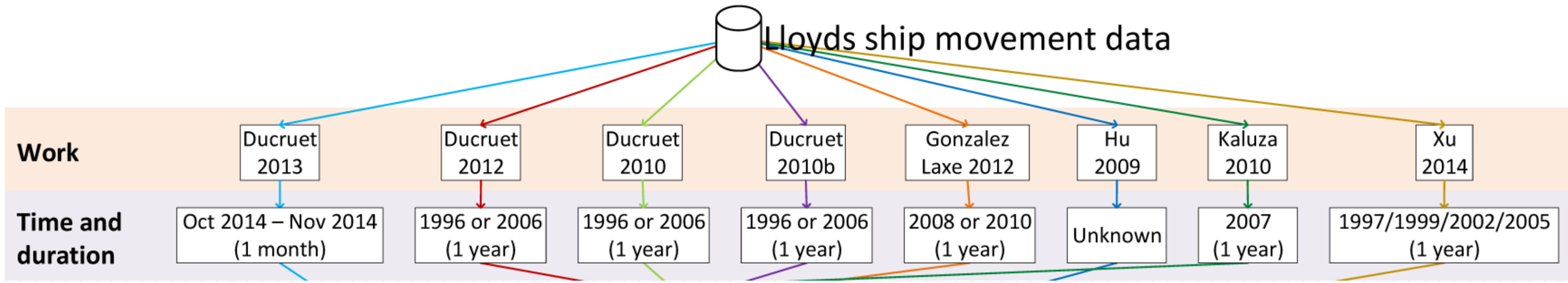
A network approach for invasive species modeling.

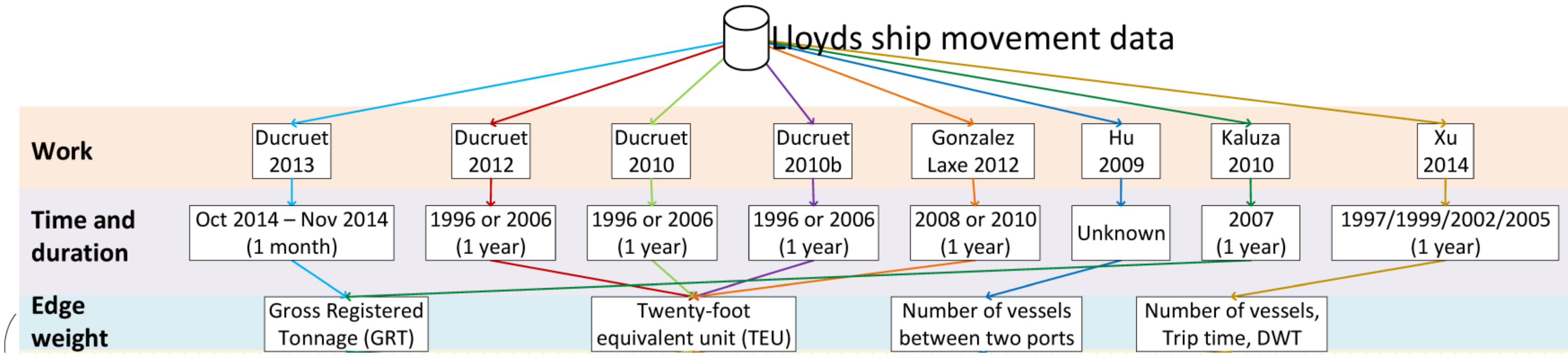
Provides insights to inform policy makers.

Shipping network construction

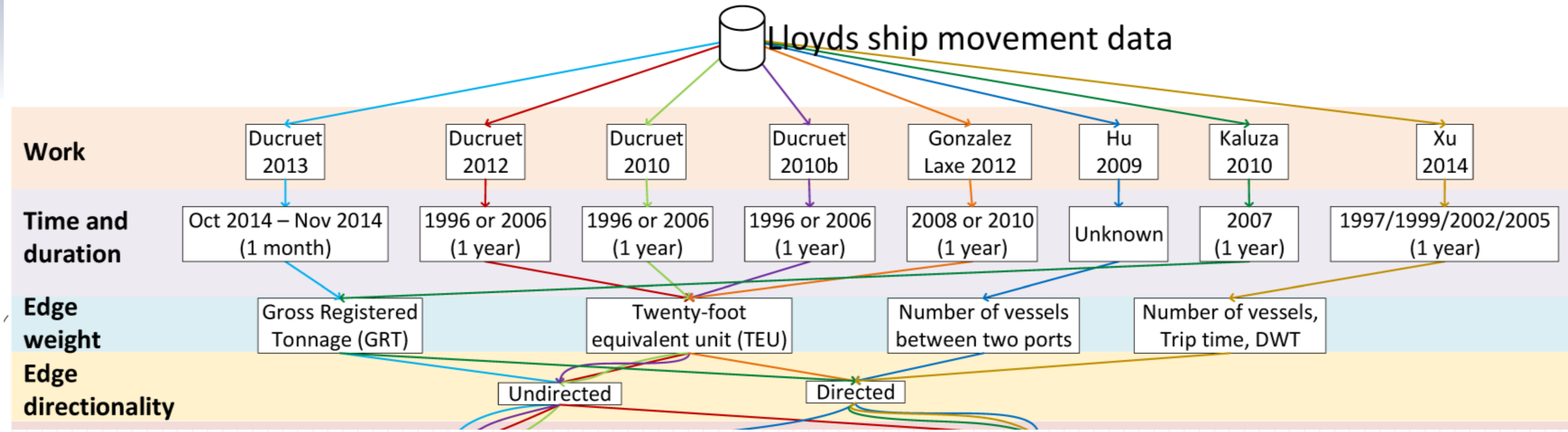
How does network construction choices
influence network properties and analysis results?

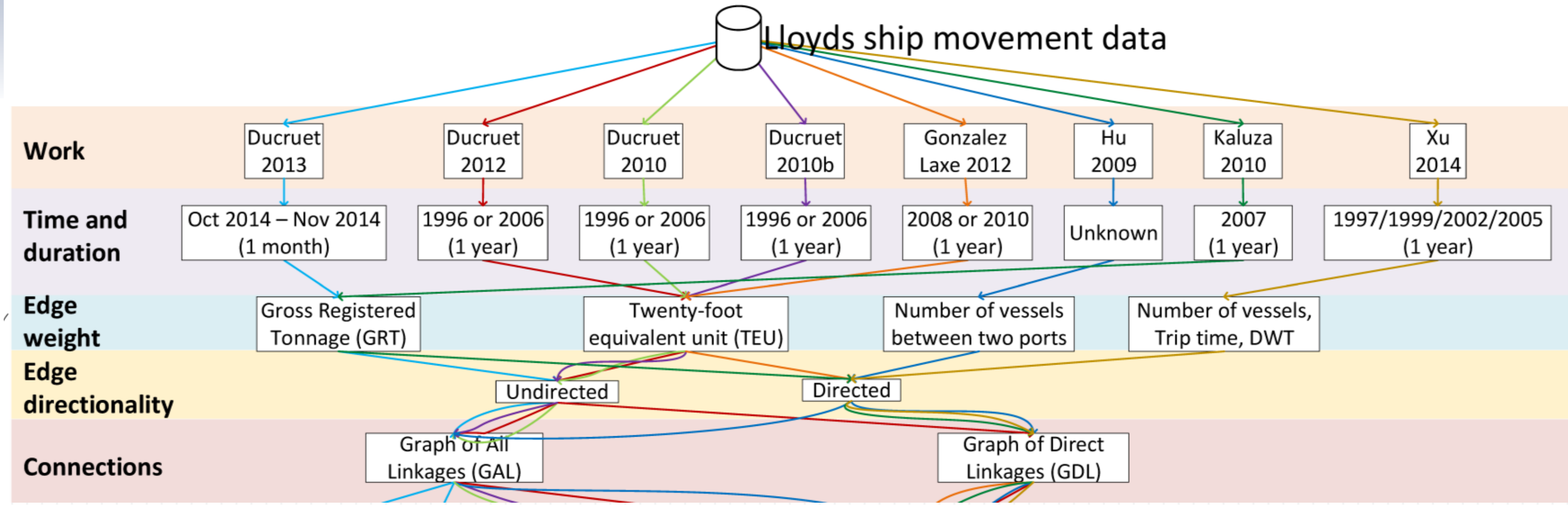

Lloyds ship movement data
Work
Ducruet 2013
Ducruet 2012
Ducruet 2010
Ducruet 2010b
Gonzalez Laxe 2012
Hu 2009
Kaluza 2010
Xu 2014


Lloyds ship movement data
Work
Ducruet 2013
Ducruet 2012
Ducruet 2010
Ducruet 2010b
Gonzalez Laxe 2012
Hu 2009
Kaluza 2010
Xu 2014
Time and duration
Oct 2014 – Nov 2014 (1 month)
1996 or 2006 (1 year)
1996 or 2006 (1 year)
1996 or 2006 (1 year)
2008 or 2010 (1 year)
Unknown
2007 (1 year)
1997/1999/2002/2005 (1 year)

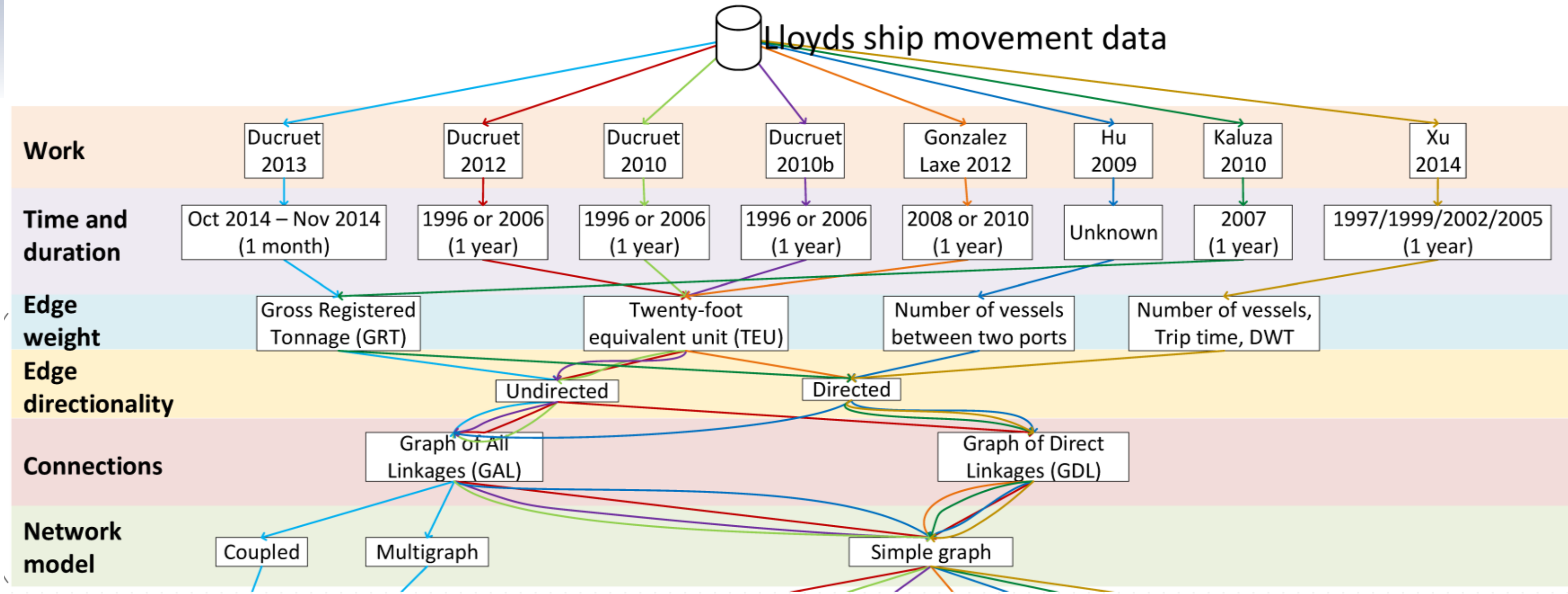


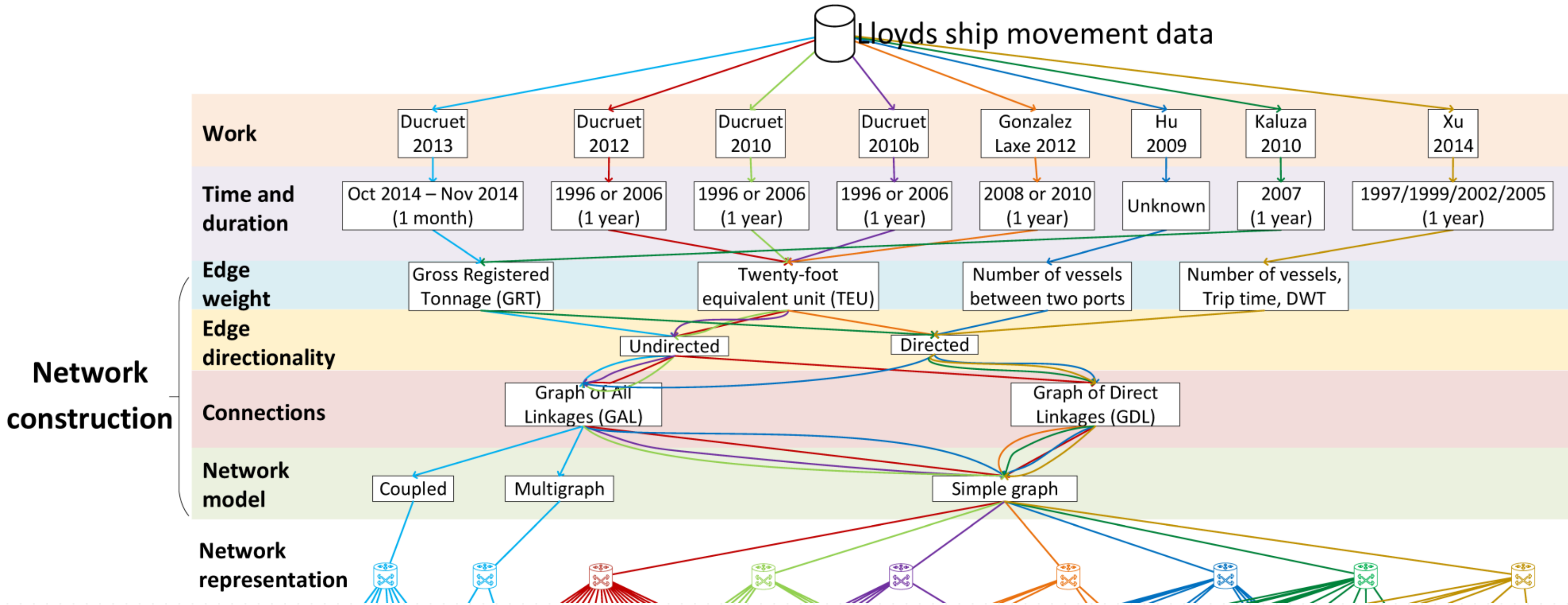
Lloyds ship movement data
Work
Ducruet 2013
Ducruet 2012
Ducruet 2010
Ducruet 2010b
Gonzalez Laxe 2012
Hu 2009
Kaluza 2010
Xu 2014
Time and duration
Oct 2014 – Nov 2014 (1 month)
1996 or 2006 (1 year)
1996 or 2006 (1 year)
1996 or 2006 (1 year)
2008 or 2010 (1 year)
Unknown
2007 (1 year)
1997/1999/2002/2005 (1 year)
Edge weight
Gross Registered Tonnage (GRT)
Twenty-foot equivalent unit (TEU)
Number of vessels between two ports
Number of vessels, Trip time, DWT


Lloyds ship movement data
Work
Ducruet 2013
Ducruet 2012
Ducruet 2010
Ducruet 2010b
Gonzalez Laxe 2012
Hu 2009
Kaluza 2010
Xu 2014
Time and duration
Oct 2014 – Nov 2014 (1 month)
1996 or 2006 (1 year)
1996 or 2006 (1 year)
1996 or 2006 (1 year)
2008 or 2010 (1 year)
Unknown
2007 (1 year)
1997/1999/2002/2005 (1 year)
Edge weight
Gross Registered Tonnage (GRT)
Twenty-foot equivalent unit (TEU)
Number of vessels between two ports
Number of vessels, Trip time, DWT
Edge directionality
Undirected
Directed

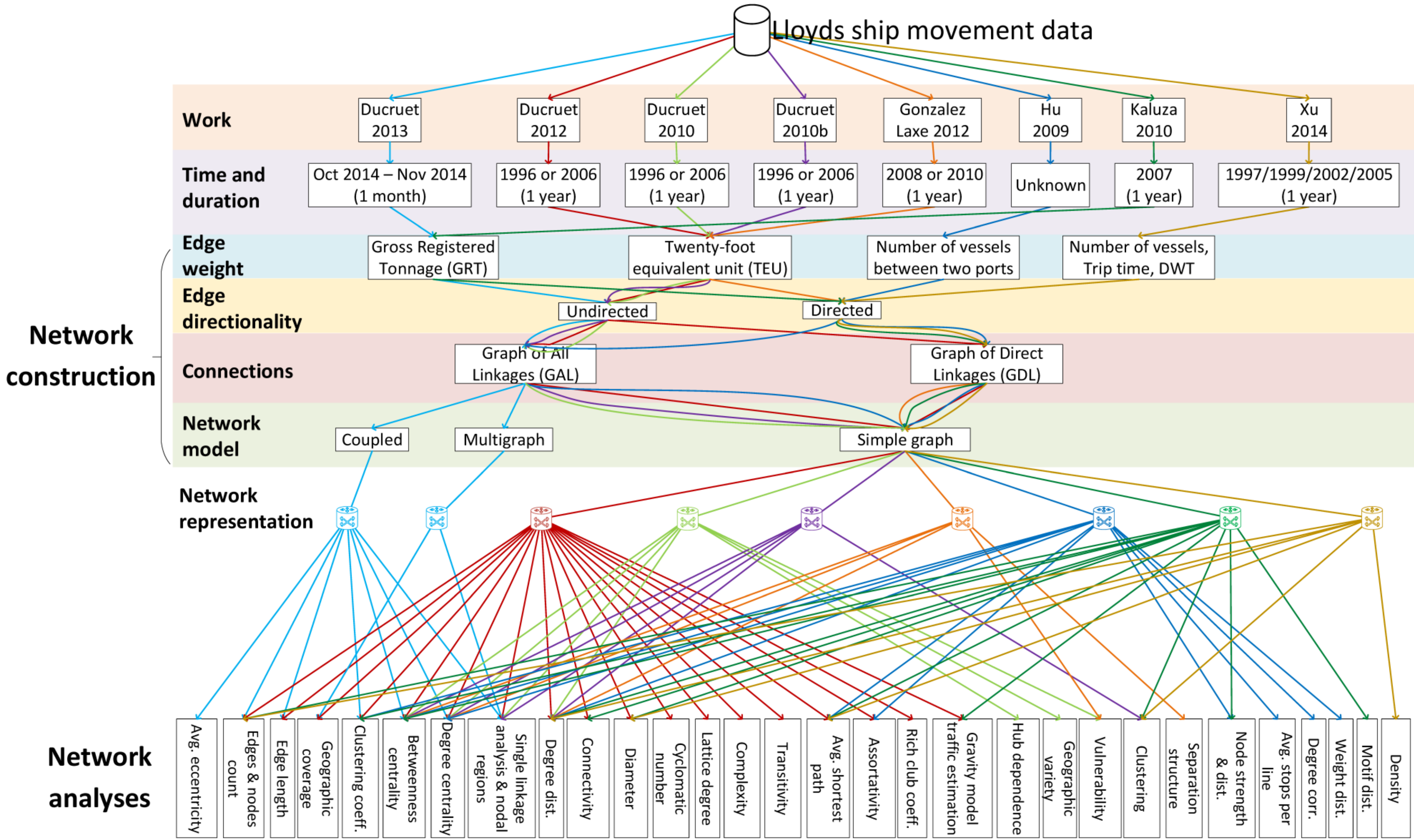


Lloyds ship movement data
Work
Ducruet 2013
Ducruet 2012
Ducruet 2010
Ducruet 2010b
Gonzalez Laxe 2012
Hu 2009
Kaluza 2010
Xu 2014
Time and duration
Oct 2014 – Nov 2014 (1 month)
1996 or 2006 (1 year)
1996 or 2006 (1 year)
1996 or 2006 (1 year)
2008 or 2010 (1 year)
Unknown
2007 (1 year)
1997/1999/2002/2005 (1 year)
Edge weight
Gross Registered Tonnage (GRT)
Twenty-foot equivalent unit (TEU)
Number of vessels between two ports
Number of vessels, Trip time, DWT
Edge directionality
Undirected
Directed
Connections
Graph of All Linkages (GAL)
Graph of Direct Linkages (GDL)

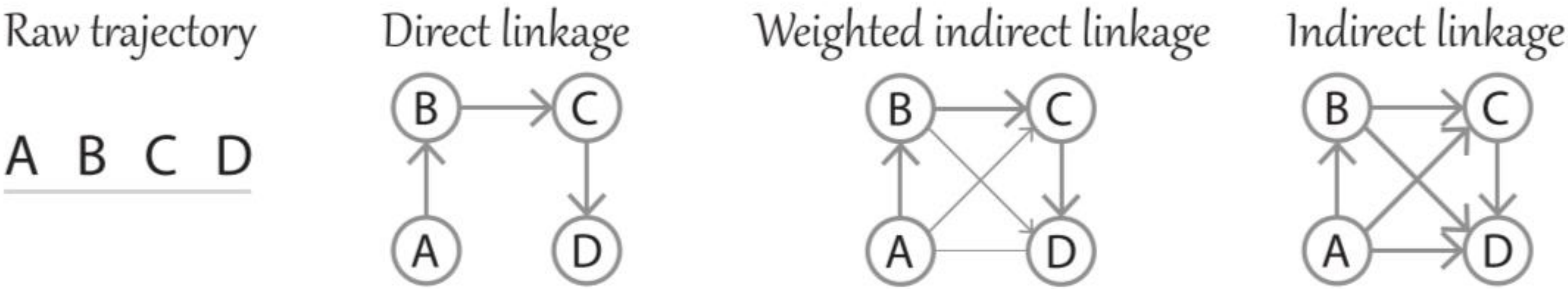

Lloyds ship movement data
Work
Ducruet 2013
Ducruet 2012
Ducruet 2010
Ducruet 2010b
Gonzalez Laxe 2012
Hu 2009
Kaluza 2010
Xu 2014
Time and duration
Oct 2014 – Nov 2014 (1 month)
1996 or 2006 (1 year)
1996 or 2006 (1 year)
1996 or 2006 (1 year)
2008 or 2010 (1 year)
Unknown
2007 (1 year)
1997/1999/2002/2005 (1 year)
Edge weight
Gross Registered Tonnage (GRT)
Twenty-foot equivalent unit (TEU)
Number of vessels between two ports
Number of vessels, Trip time, DWT
Edge directionality
Undirected
Directed
Connections
Graph of All Linkages (GAL)
Graph of Direct Linkages (GDL)
Network model
Coupled
Multigraph
Simple graph



Lloyds ship movement data
Work
Ducruet 2013
Ducruet 2012
Ducruet 2010
Ducruet 2010b
Gonzalez Laxe 2012
Hu 2009
Kaluza 2010
Xu 2014
Time and duration
Oct 2014 – Nov 2014 (1 month)
1996 or 2006 (1 year)
1996 or 2006 (1 year)
1996 or 2006 (1 year)
2008 or 2010 (1 year)
Unknown
2007 (1 year)
1997/1999/2002/2005 (1 year)
Network construction
Edge weight
Gross Registered Tonnage (GRT)
Twenty-foot equivalent unit (TEU)
Number of vessels between two ports
Number of vessels, Trip time, DWT
Edge directionality
Undirected
Directed
Connections
Graph of All Linkages (GAL)
Graph of Direct Linkages (GDL)
Network model
Coupled
Multigraph
Simple graph
Network representation


Lloyds ship movement data
Work
Ducruet 2013
Ducruet 2012
Ducruet 2010
Ducruet 2010b
Gonzalez Laxe 2012
Hu 2009
Kaluza 2010
Xu 2014
Time and duration
Oct 2014 – Nov 2014 (1 month)
1996 or 2006 (1 year)
1996 or 2006 (1 year)
1996 or 2006 (1 year)
2008 or 2010 (1 year)
Unknown
2007 (1 year)
1997/1999/2002/2005 (1 year)
Network construction
Edge weight
Gross Registered Tonnage (GRT)
Twenty-foot equivalent unit (TEU)
Number of vessels between two ports
Number of vessels, Trip time, DWT
Edge directionality
Undirected
Directed
Connections
Graph of All Linkages (GAL)
Graph of Direct Linkages (GDL)
Network model
Coupled
Multigraph
Simple graph
Network representation
Network analyses
Avg. eccentricity
Edges & nodes count
Edge length
Geographic coverage
Clustering coeff.
Betweenness centrality
Degree centrality
Single linkage analysis & nodal regions
Degree dist.
Connectivity
Diameter
Cyclomatic number
Lattice degree
Complexity
Transitivity
Avg. shortest path
Assortativity
Rich club coeff.
Gravity model traffic estimation
Hub dependence
Geographic variety
Vulnerability
Clustering
Separation structure
Node strength & dist.
Avg. stops per line
Degree corr.
Weight dist.
Motif dist.
Density

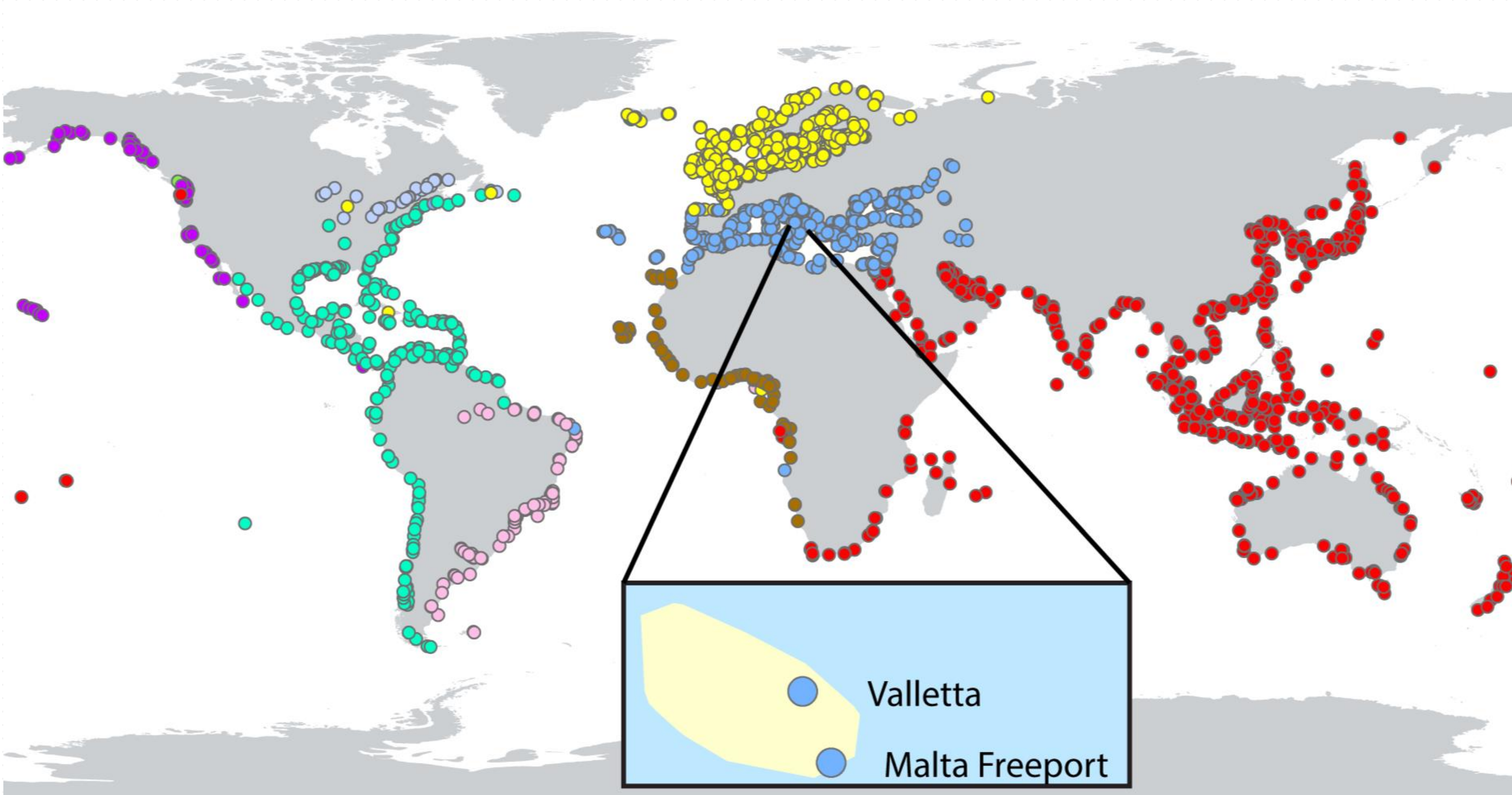
Network linkage mechanisms



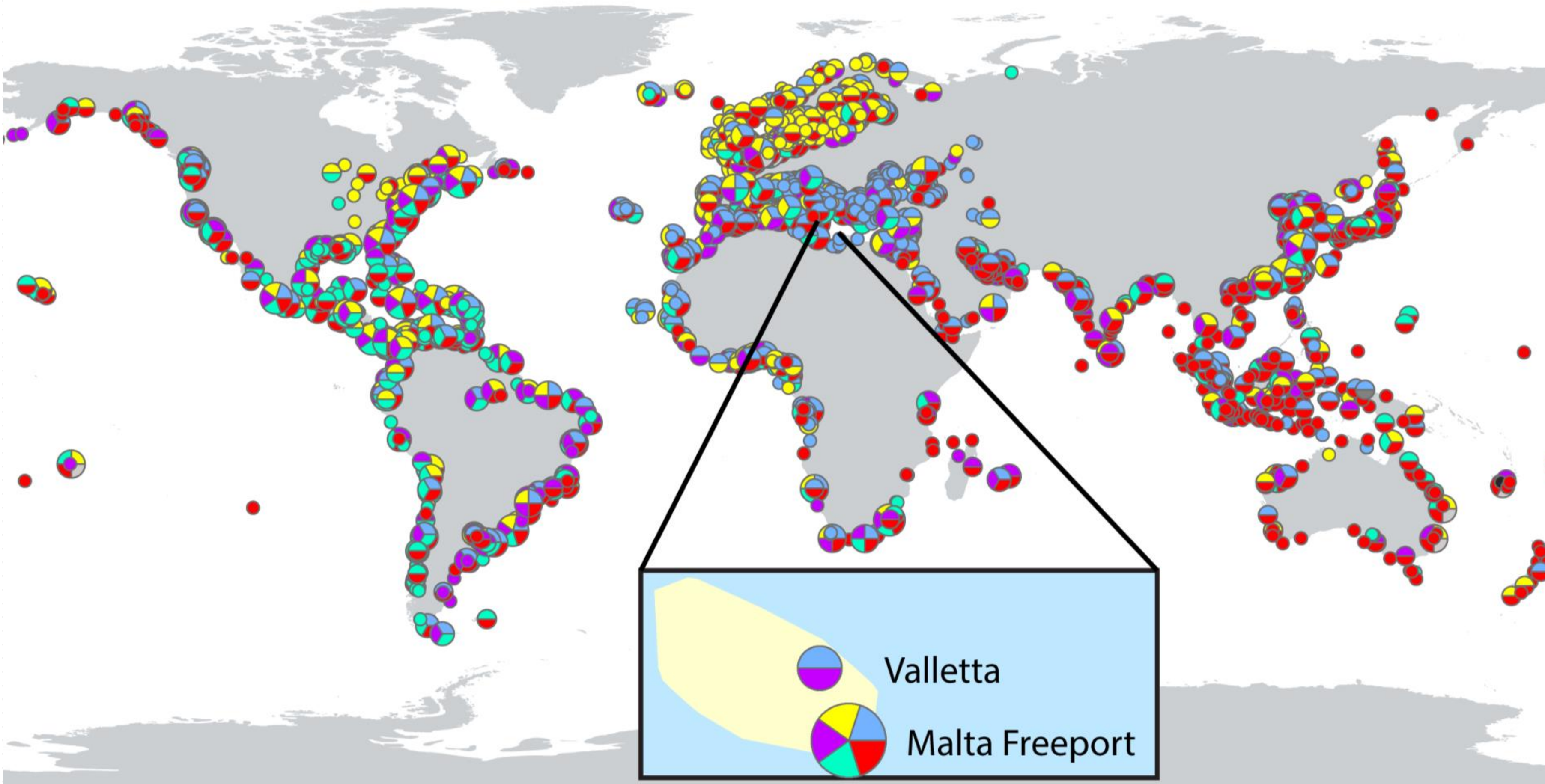
Network linkage mechanisms

	Direct Linkage		Indirect Linkage
num_of_nodes	3.60E+3	=	3.60E+3
num_of_edges	**1.32E+5**	<	**7.37E+5**
density	**2.05E-2**	<	**1.14E-1**
average_degree	**7.35E+1**	<	**4.10E+2**
highest_degree	1.28E+3	<	2.42E+3
generalized_clustering_coefficient	**5.48E-1**	<	**7.23E-1**
transitivity	**2.96E-1**	<	**4.96E-1**
avg_shortest_path	**2.65**	>	**2.04**
diameter	**8**	>	**5**
radius	4	>	3

Clustering: first-order network



Clustering: higher-order network



Takeaways

Global shipping traffic is:

- Imbalanced in directionality – directed network
- Unevenly distributed shipping frequency & traffic – weighted network
- Higher-order movement patterns – higher-order network
- Other important factors include
  - Linkage mechanisms, time window, seasonality, evolution

Considerations when representing shipping traffic as network, or reporting analysis results

Species invasion in the Arctic

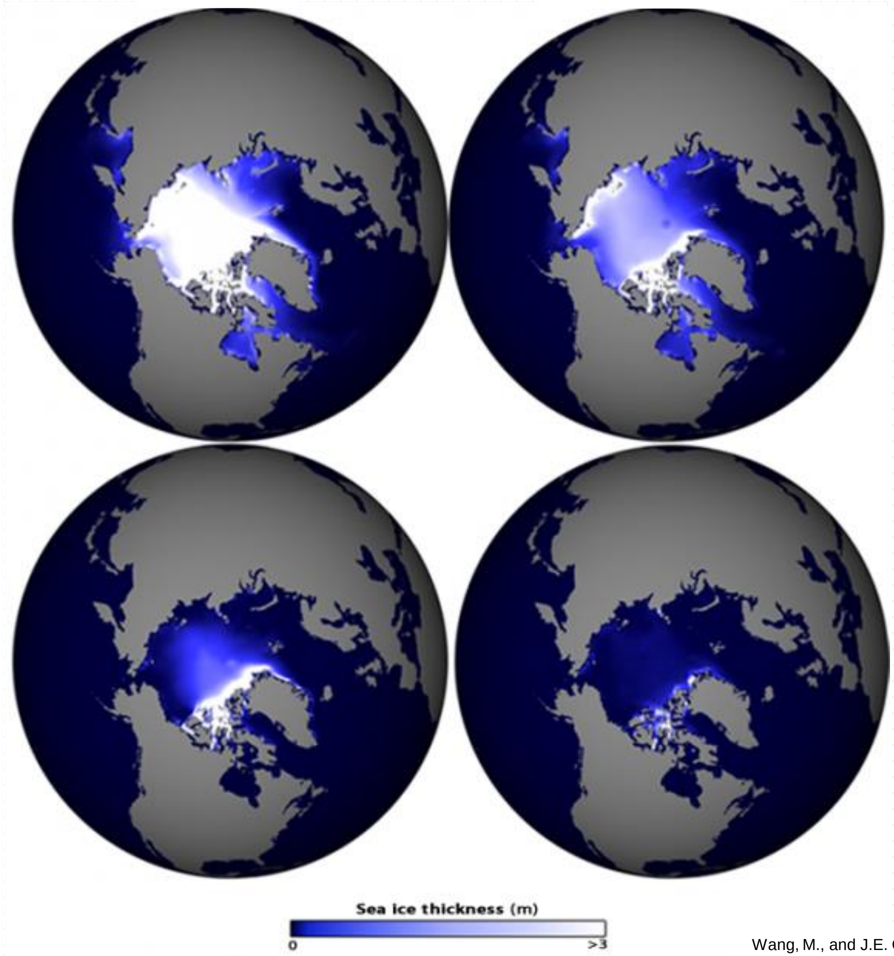
Introduction to Arctic ports & diffusion among Arctic ports

The melting Arctic sea ice

March September

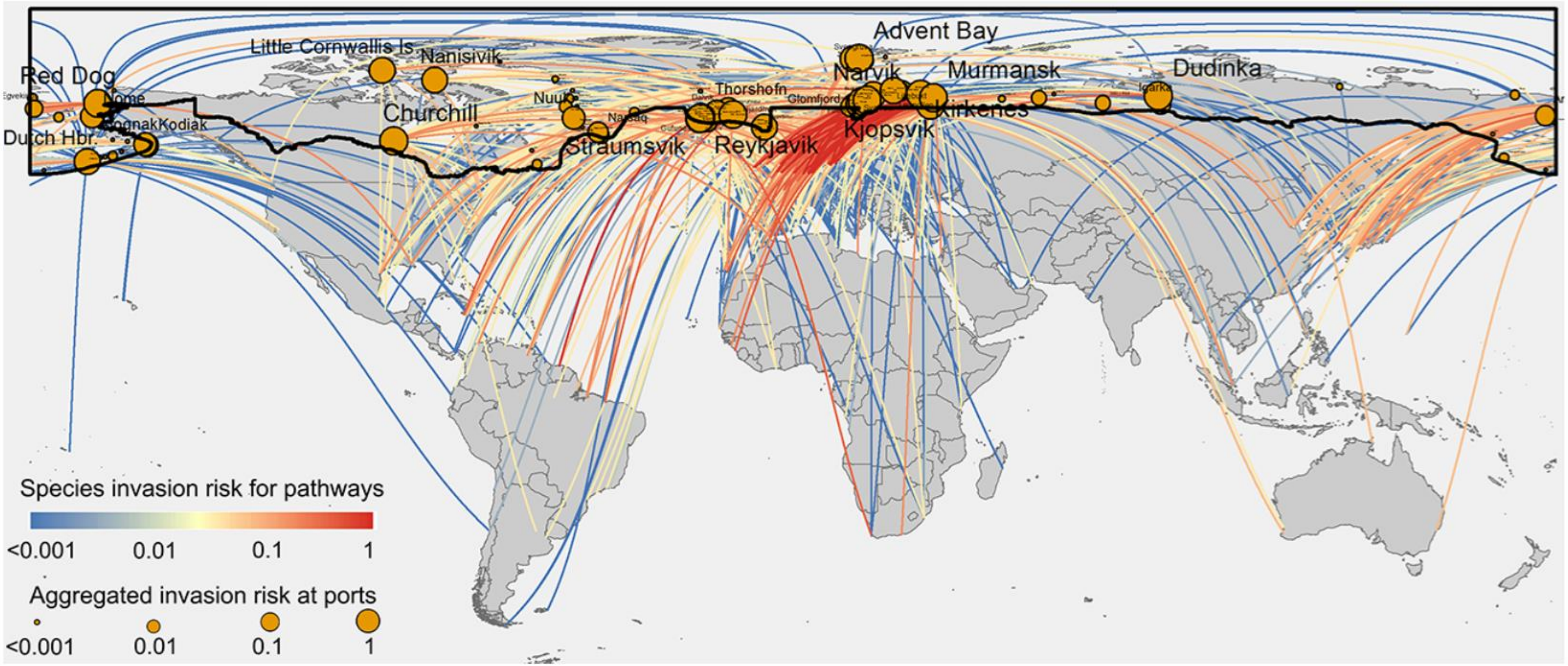
2009

2039

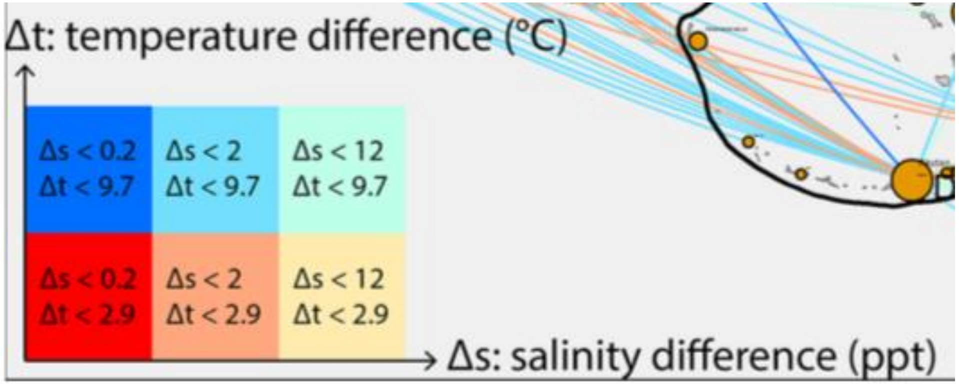


Wang, M., and J.E. Overland (2009): A sea ice free summer Arctic within 30 years? Geophys. Res. Lett., 36, L07502, doi: 10.1029/2009GL037820.

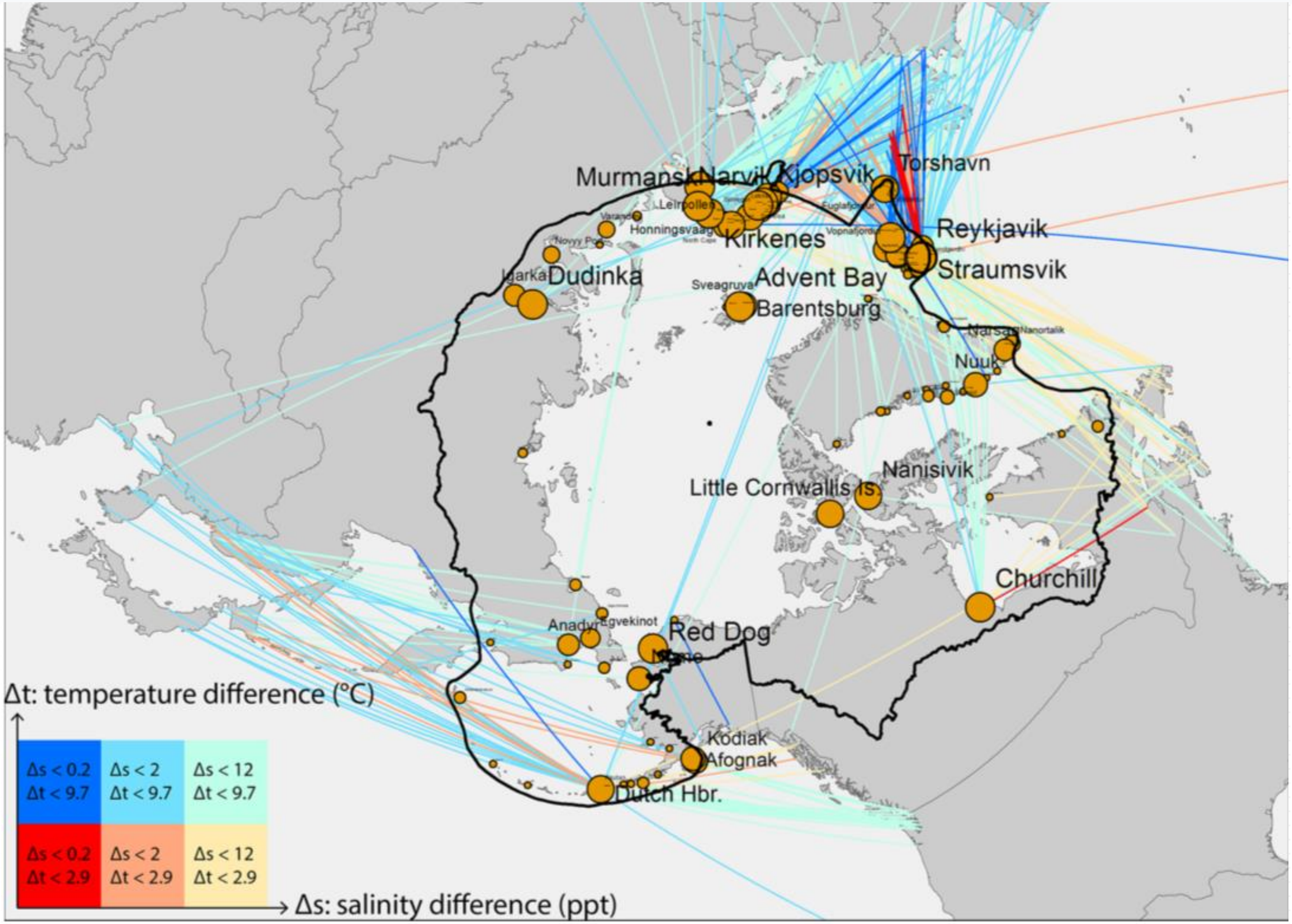
Species introduction pathways to the Arctic



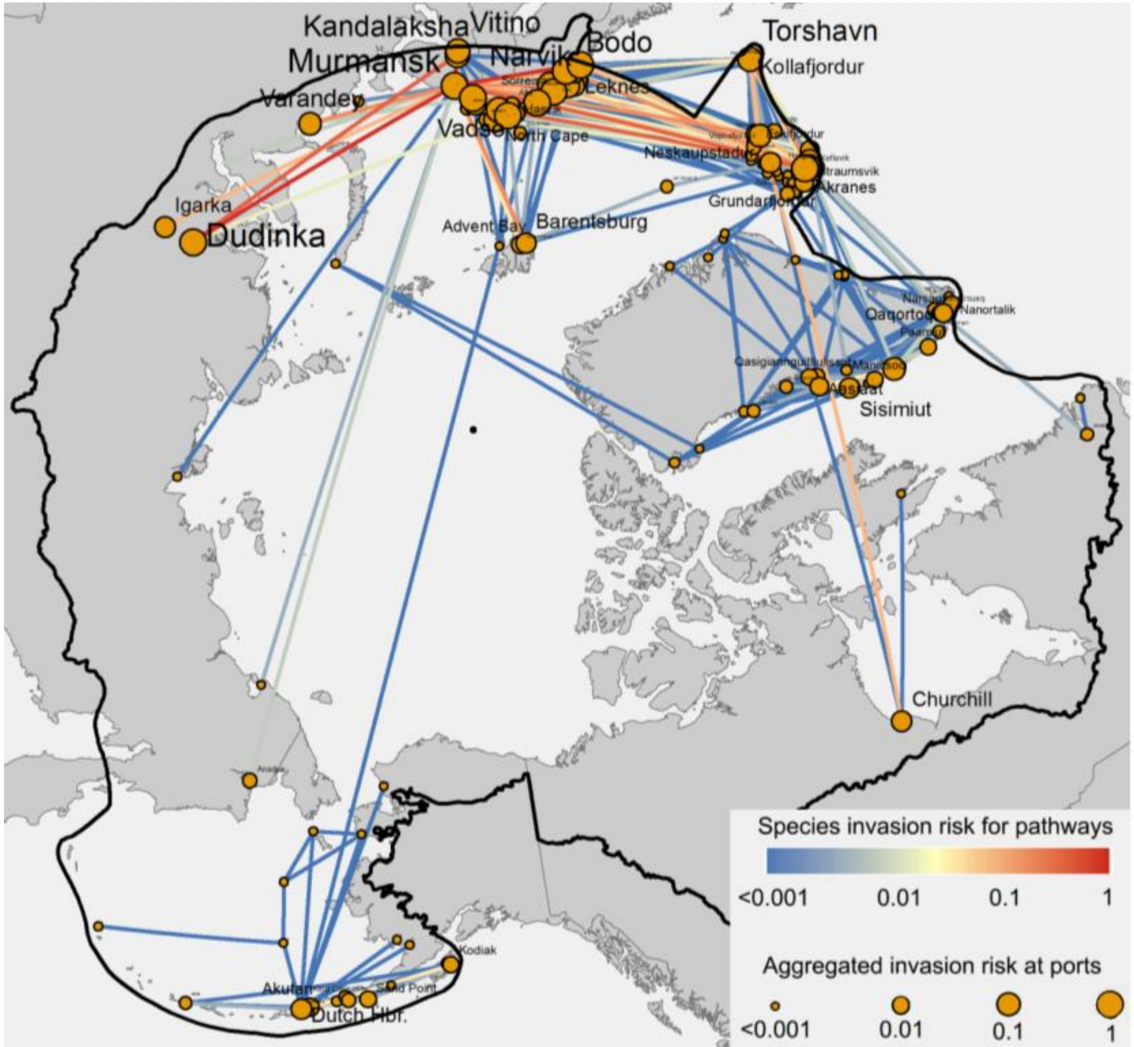
Environmental tolerance

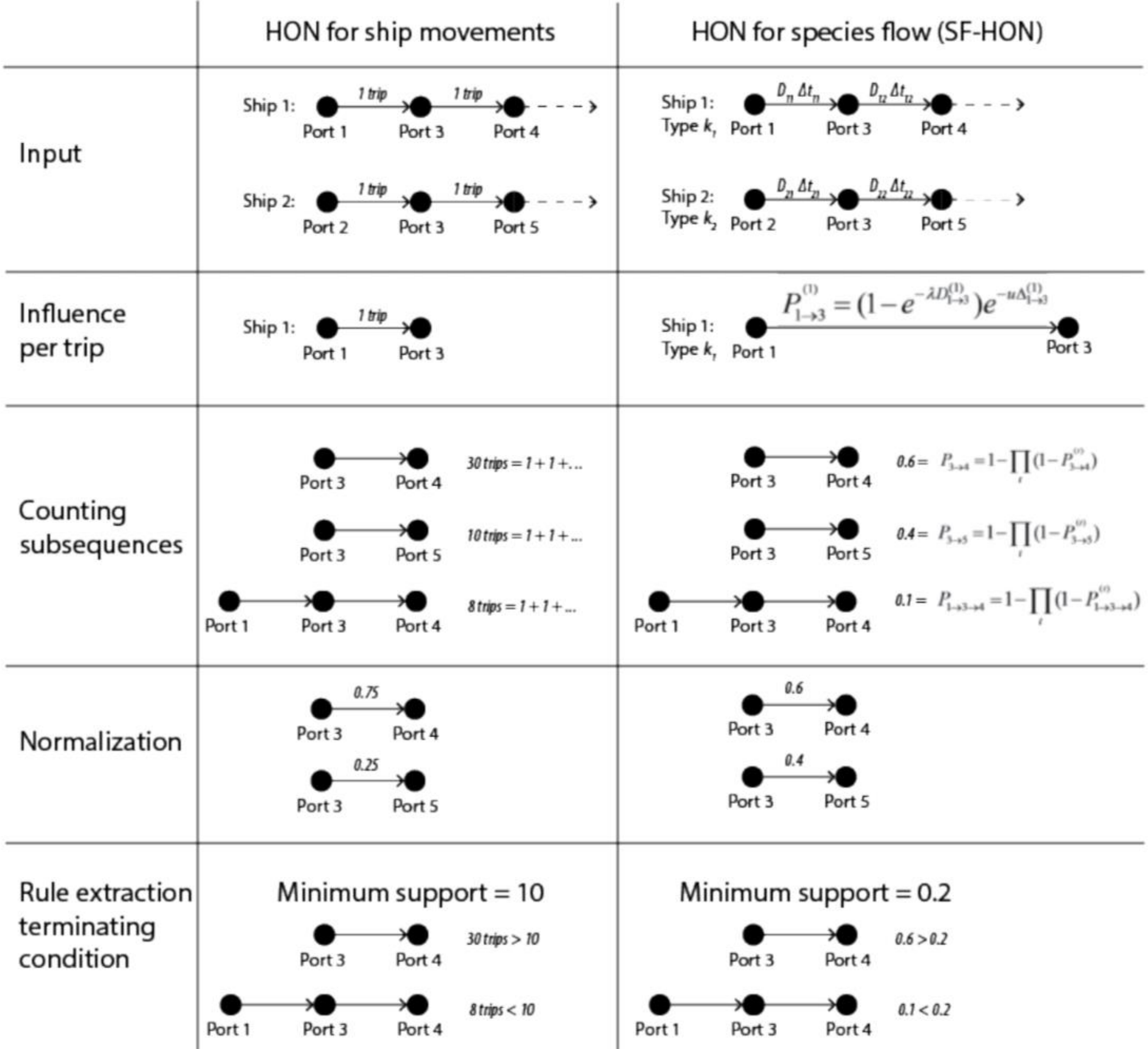


Environmental tolerance

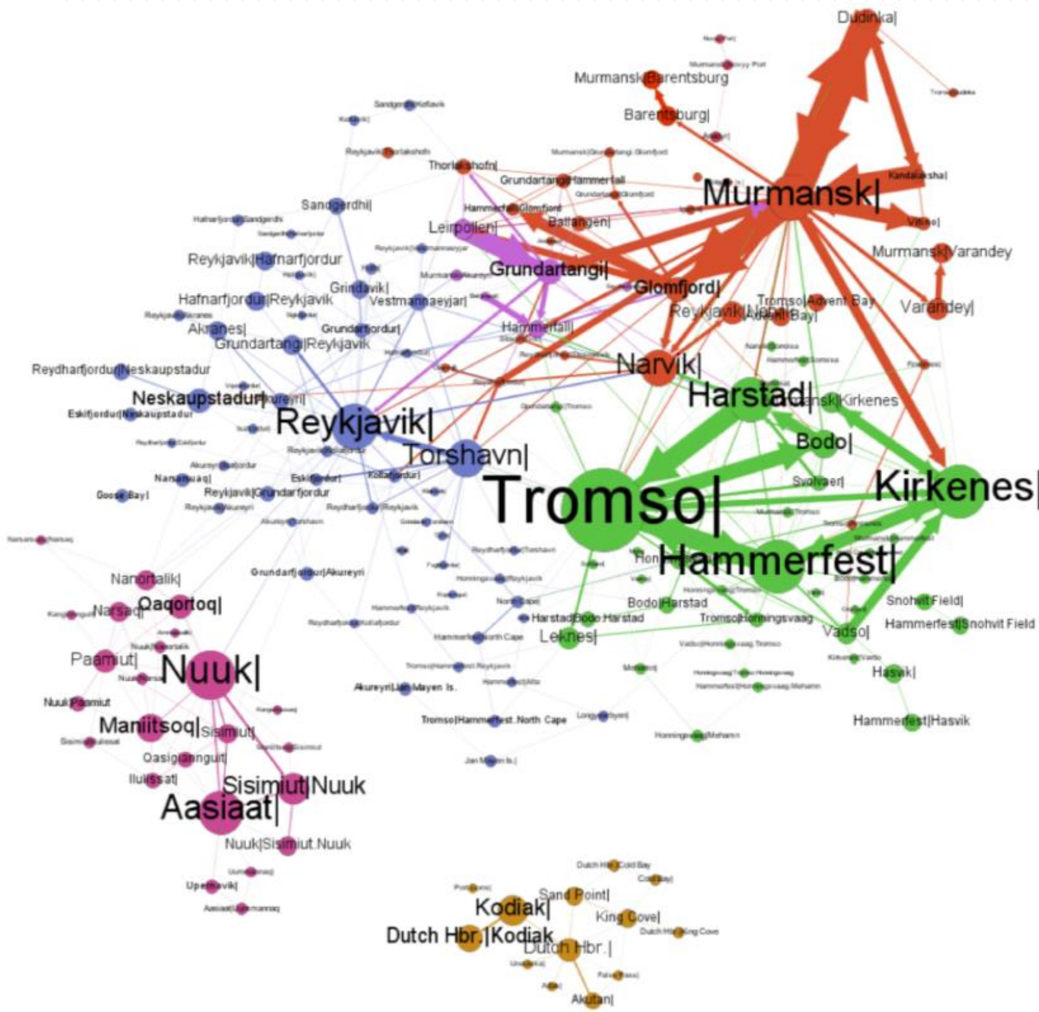


Species diffusion within the Arctic

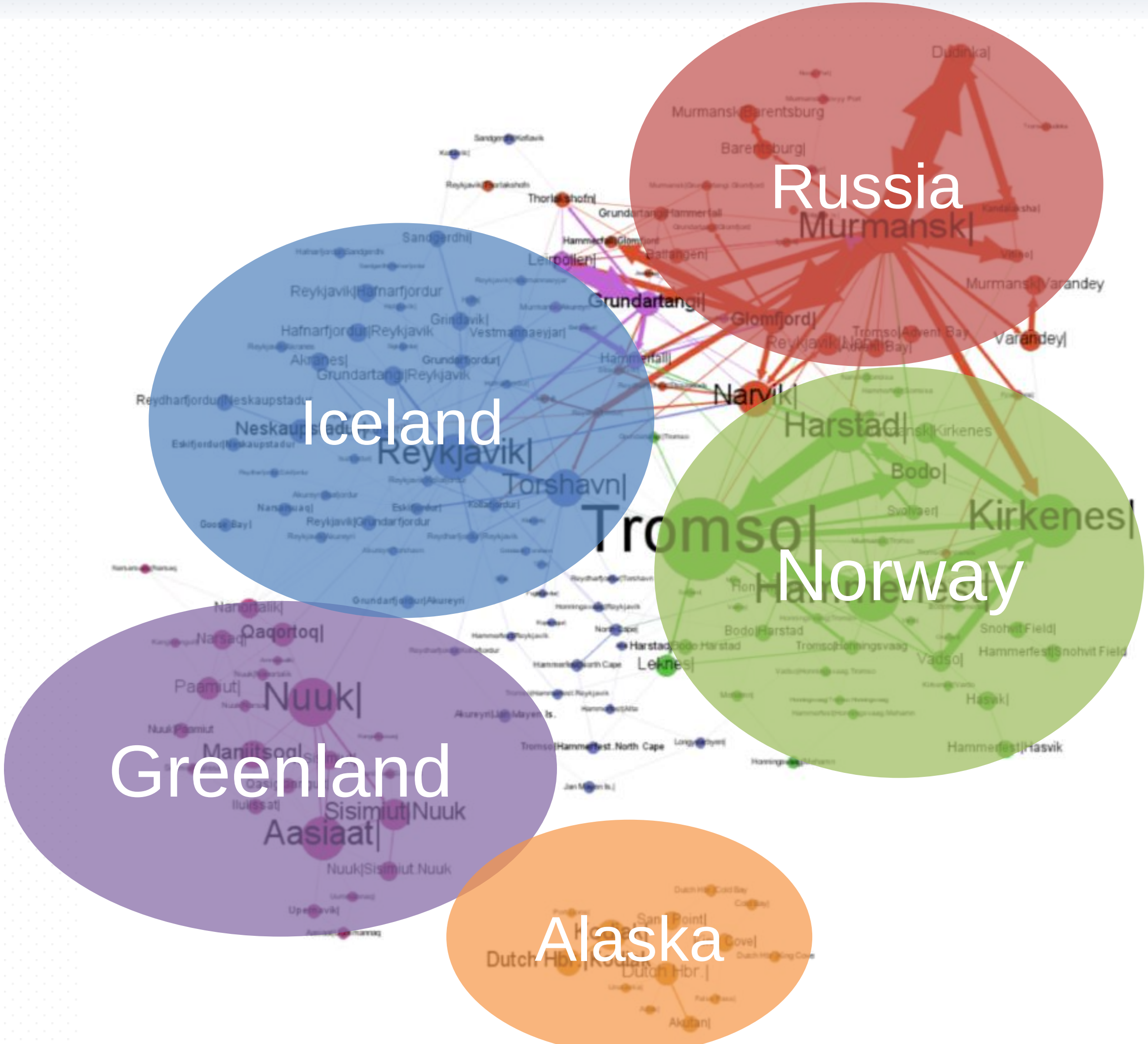



HON for ship movements
HON for species flow (SF-HON)
Input
Influence per trip
Counting subsequences
Normalization
Rule extraction terminating condition
Ship 1:
Ship 2:
Type k_1
Type k_2
Port 1
Port 2
Port 3
Port 4
Port 5
1 trip
$D_{11}\,\Delta t_{11}$
$D_{12}\,\Delta t_{12}$
$D_{21}\,\Delta t_{21}$
$D_{22}\,\Delta t_{22}$
$P^{(1)}_{1\to3} = (1 - e^{-\lambda D^{(1)}_{1\to3}})e^{-u\Delta^{(1)}_{1\to3}}$
30 trips = 1 + 1 + ...
10 trips = 1 + 1 + ...
8 trips = 1 + 1 + ...
$0.6 = P_{3\to4} = 1 - \prod_i (1 - P^{(i)}_{3\to4})$
$0.4 = P_{3\to5} = 1 - \prod_i (1 - P^{(i)}_{3\to5})$
$0.1 = P_{1\to3\to4} = 1 - \prod_i (1 - P^{(i)}_{1\to3\to4})$
0.75
0.25
0.6
0.4
Minimum support = 10
Minimum support = 0.2
30 trips > 10
8 trips < 10
0.6 > 0.2
0.1 < 0.2

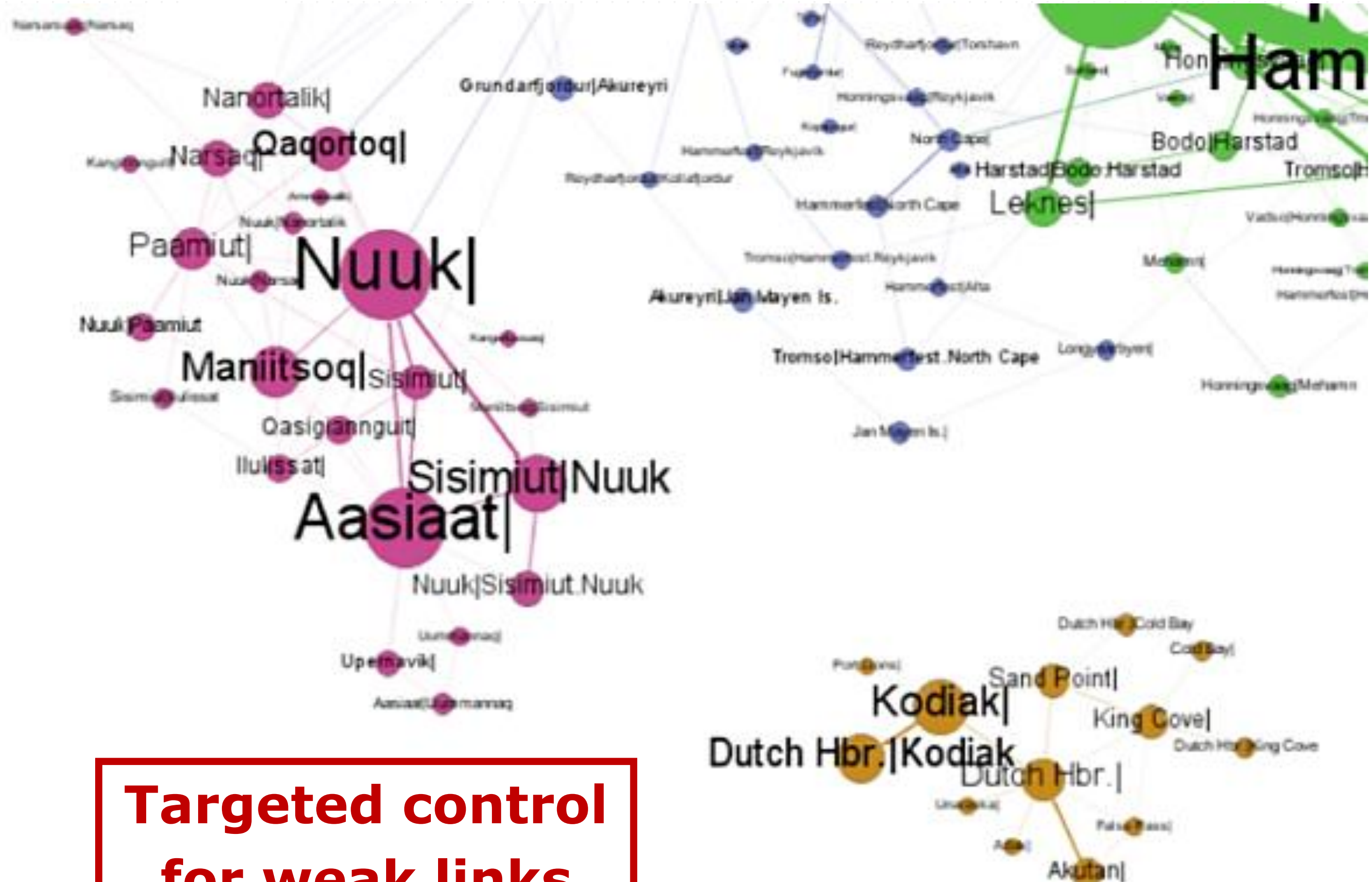
Species flow higher-order network



Species flow higher-order network



Species flow higher-order network



Targeted control for weak links

Case studies

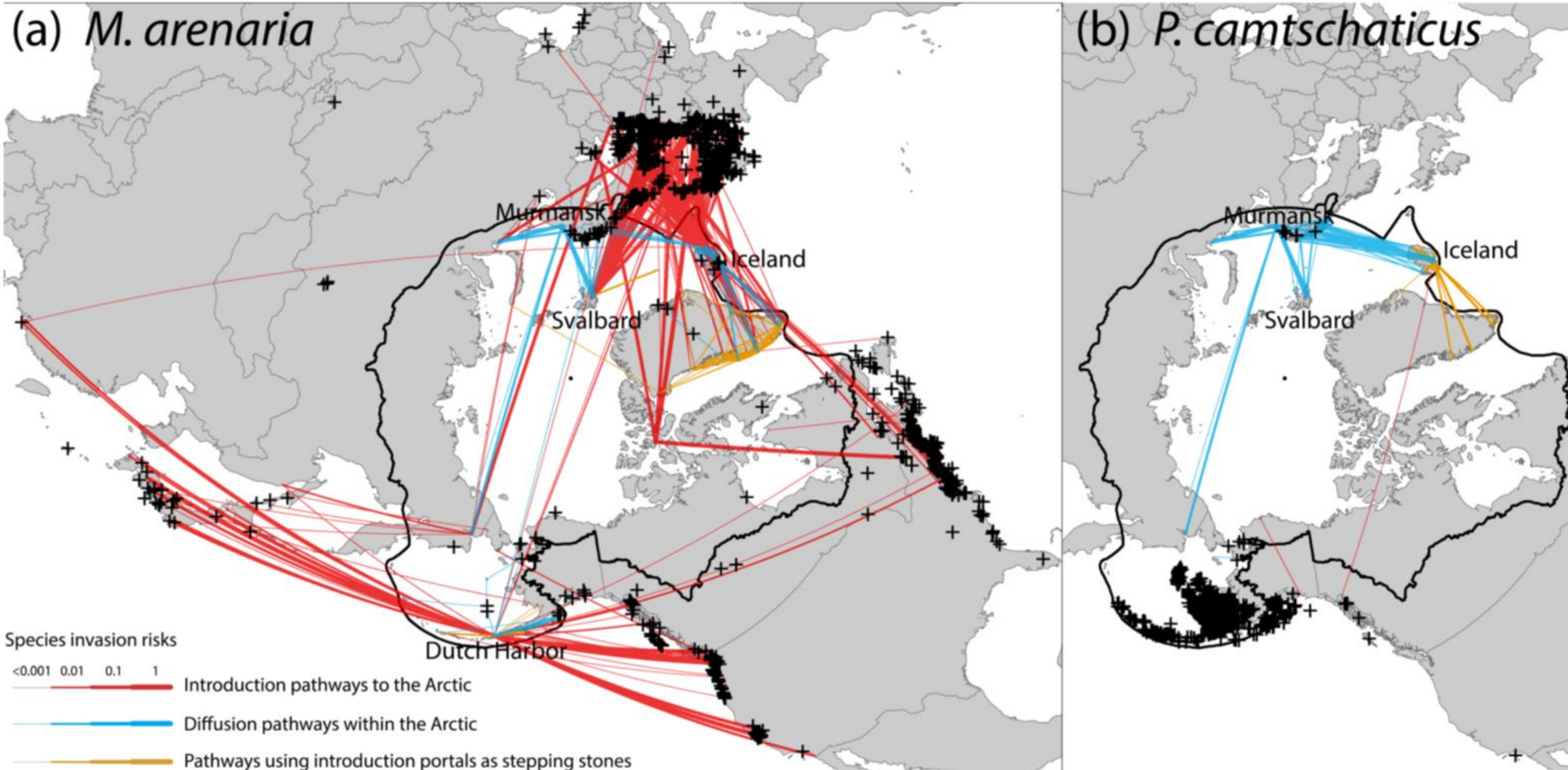
Soft shell clam

Scientific name: *M. arenaria*
Temperature tolerance: -2 – 18 ℃
Salinity tolerance: 28 – 35 PSU

Red king crab

Scientific name: *P. camtschaticus*
Temperature tolerance: -2 – 18 ℃
Salinity tolerance: 28 – 35 PSU

Case studies



Risk projection

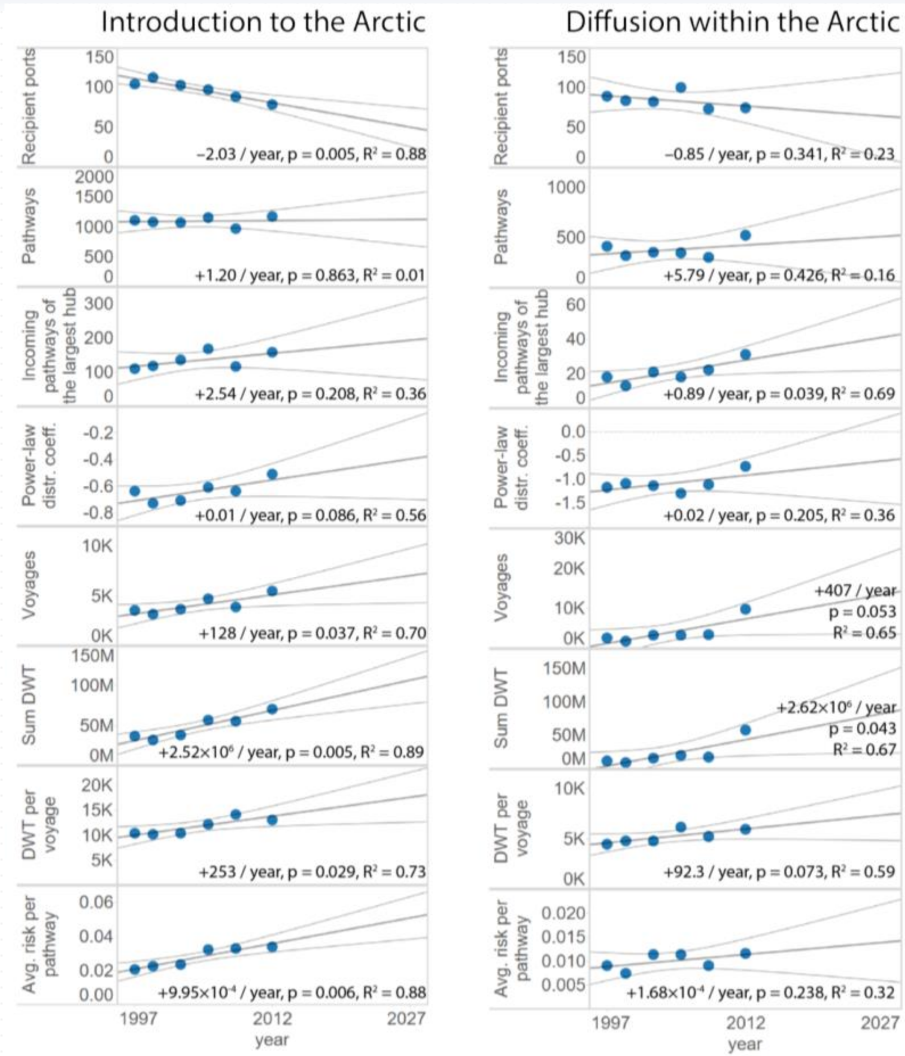
Decreasing recipient ports

Same # pathways

Increased connections at hubs

“Rich gets richer”

Emergence of hubs



Risk projection

Decreasing recipient ports

Same # pathways

Increased connections at hubs

“Rich gets richer”

Increasing voyages

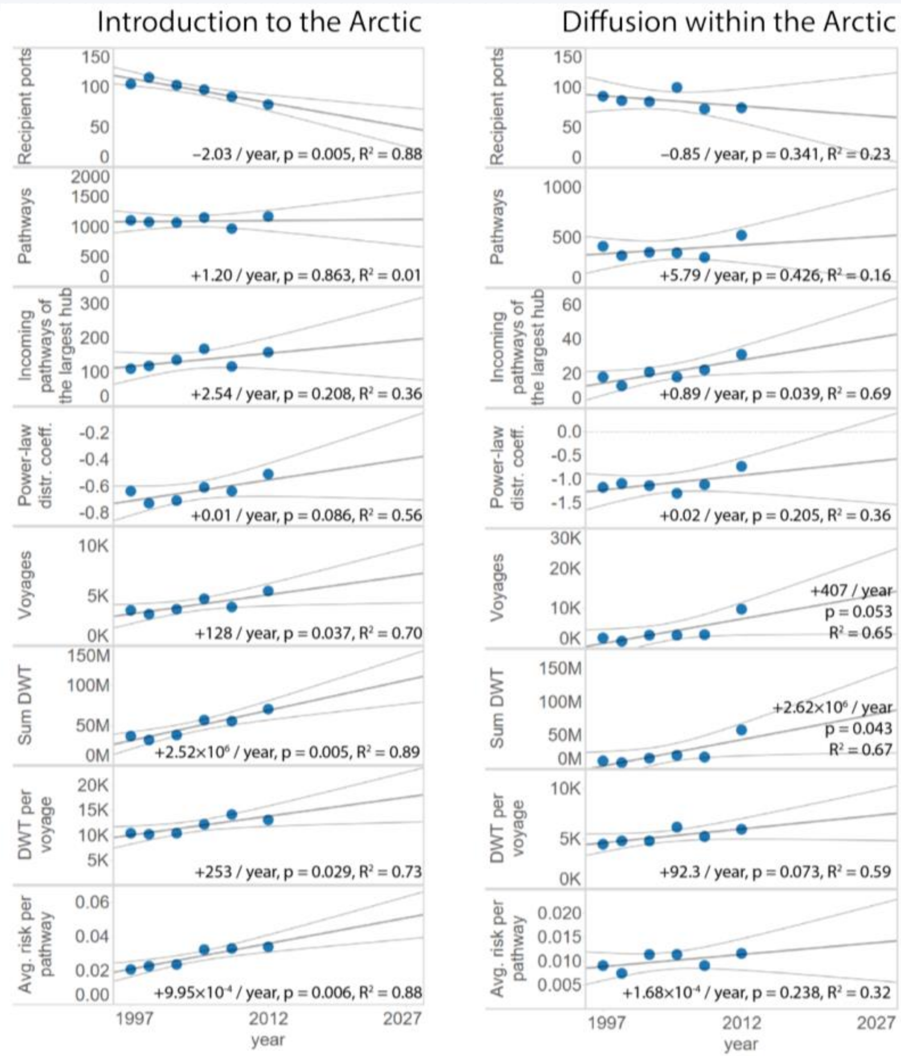
Increasing sum of shipping

Increasing average ship size

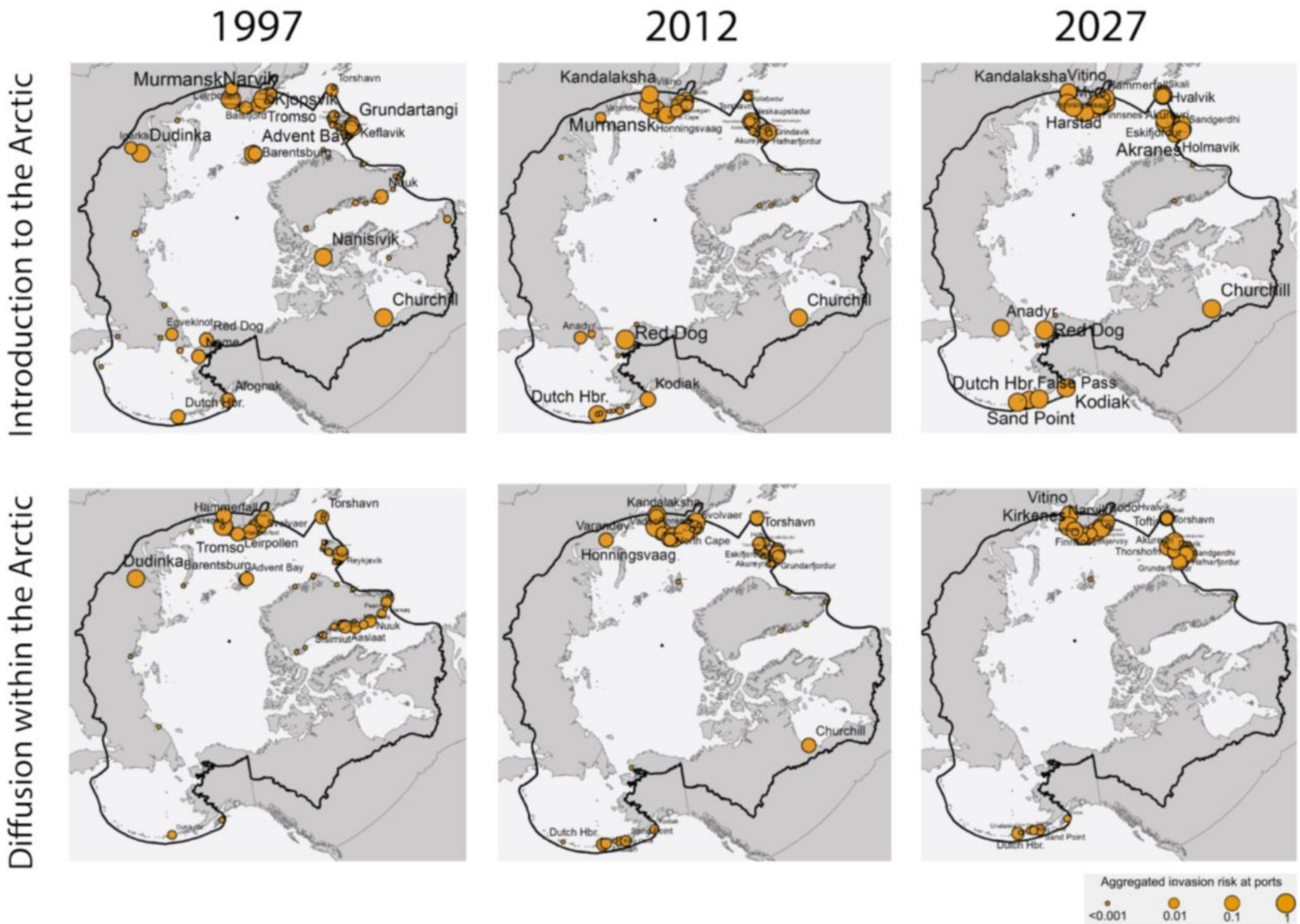
Increasing risk per pathway

Emergence of hubs

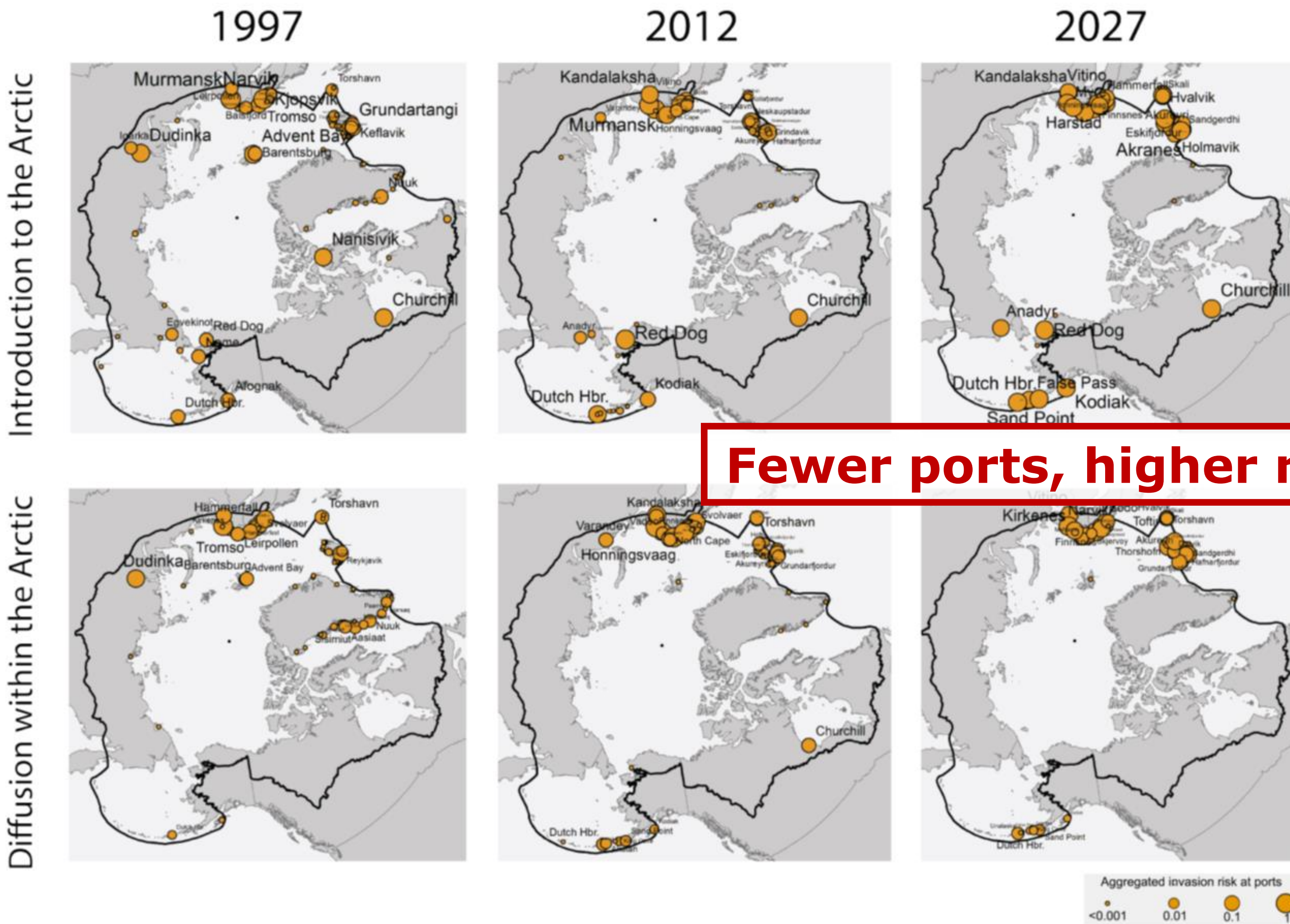
Increasing risk



Risk projection



Risk projection

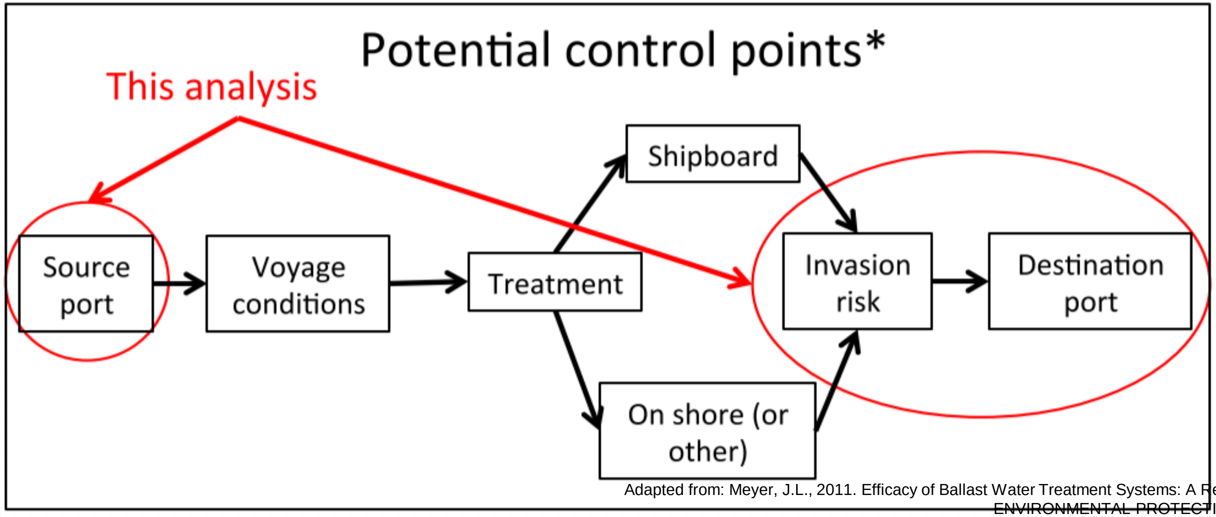


Fewer ports, higher risks

Takeaways

Arctic species invasion

- Increasing risk
- Emergence of hubs
- Targeted controls



Adapted from: Meyer, J.L., 2011. Efficacy of Ballast Water Treatment Systems: A Report by the EPA Science Advisory Board (No. EPA-SAB-11-009). ENVIRONMENTAL PROTECTION AGENCY WASHINGTON DC SCIENCE ADVISORY BOARD.

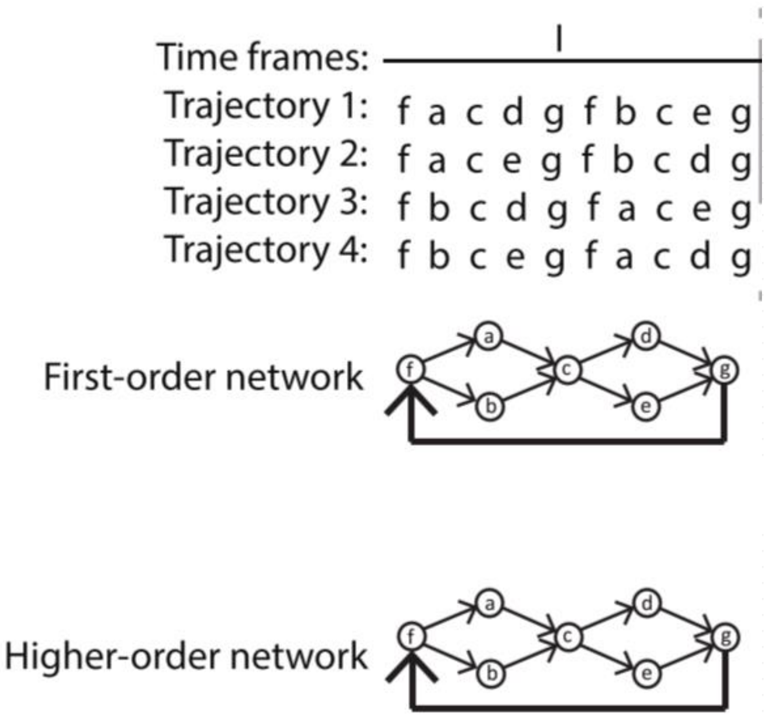
Anomaly detection

Unveiling higher-order anomalies with HON

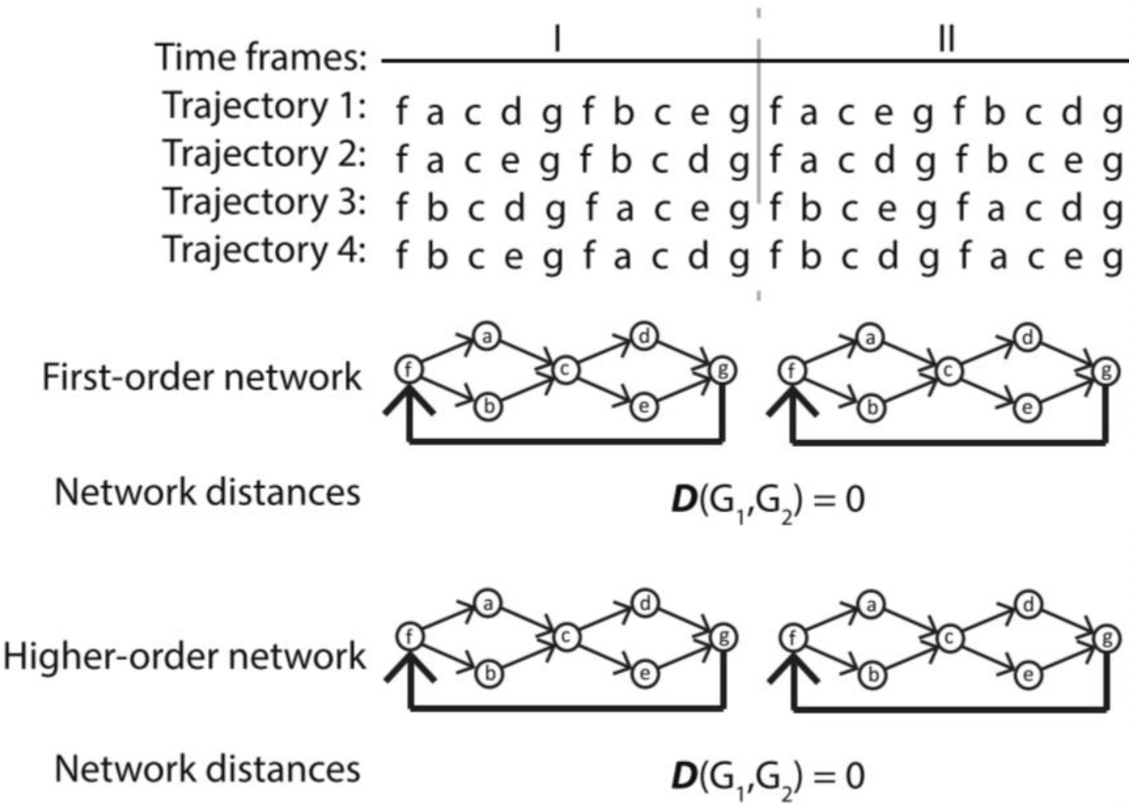
Anomaly detection with dynamic network

Time frames:										
Trajectory 1:	f	a	c	d	g	f	b	c	e	g
Trajectory 2:	f	a	c	e	g	f	b	c	d	g
Trajectory 3:	f	b	c	d	g	f	a	c	e	g
Trajectory 4:	f	b	c	e	g	f	a	c	d	g

Anomaly detection with dynamic network

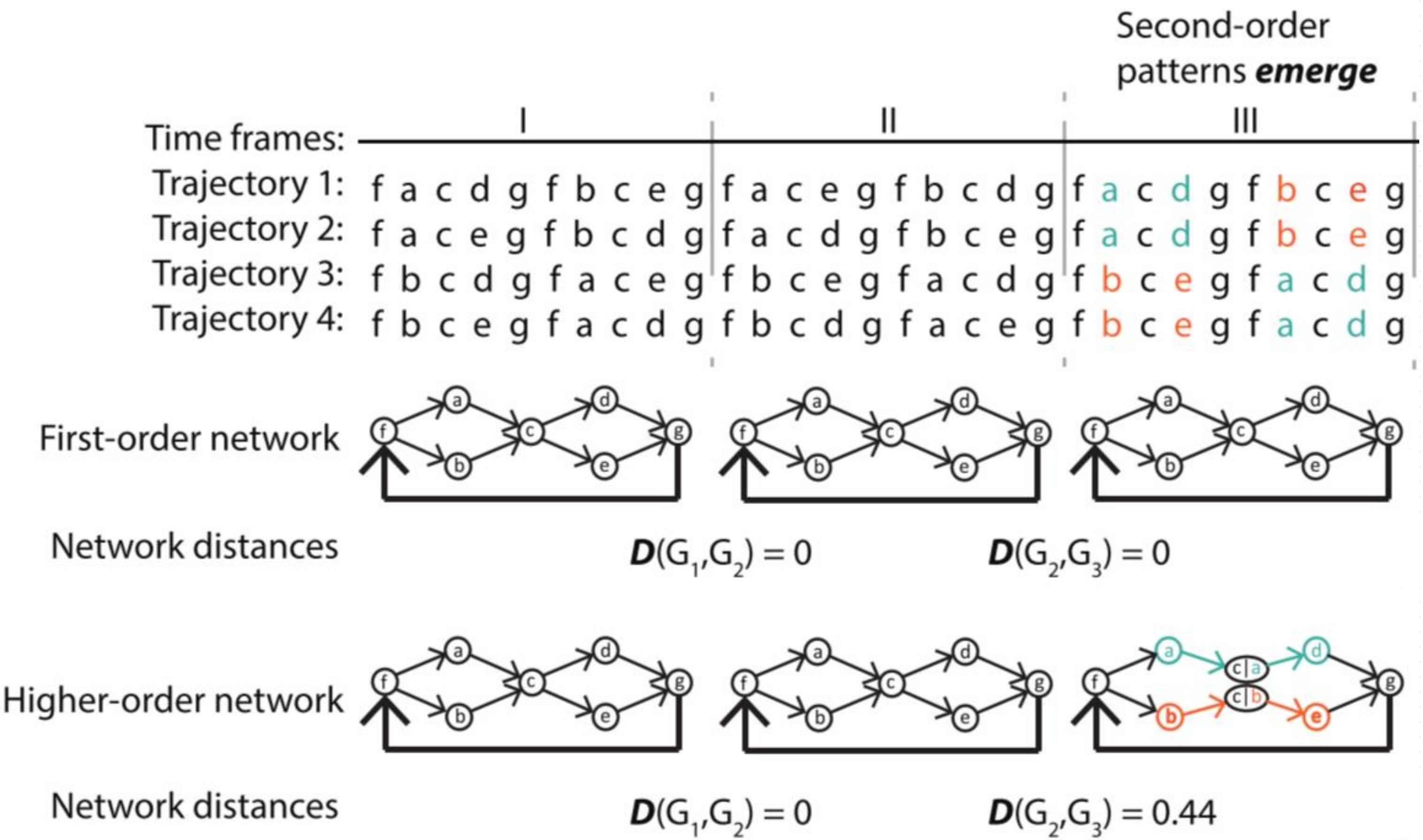


Anomaly detection with dynamic network

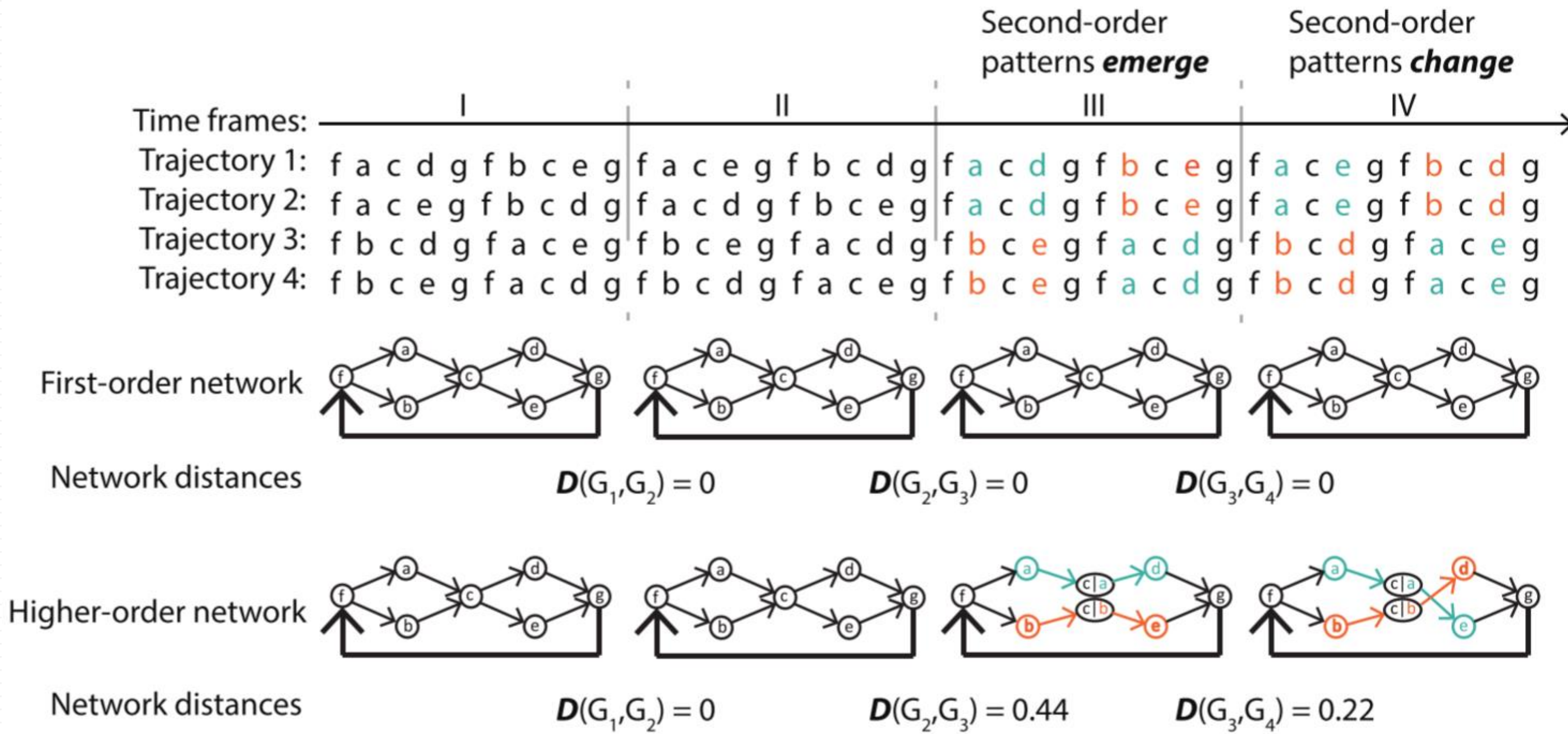


$$\mathcal{D}(G,H) = \frac{\sum\limits_{u,v \in V} \frac{|w_E^G(u,v) - w_E^H(u,v)|}{max(w_E^G(u,v) - w_E^H(u,v))}}{|E_G \cup E_H|}$$

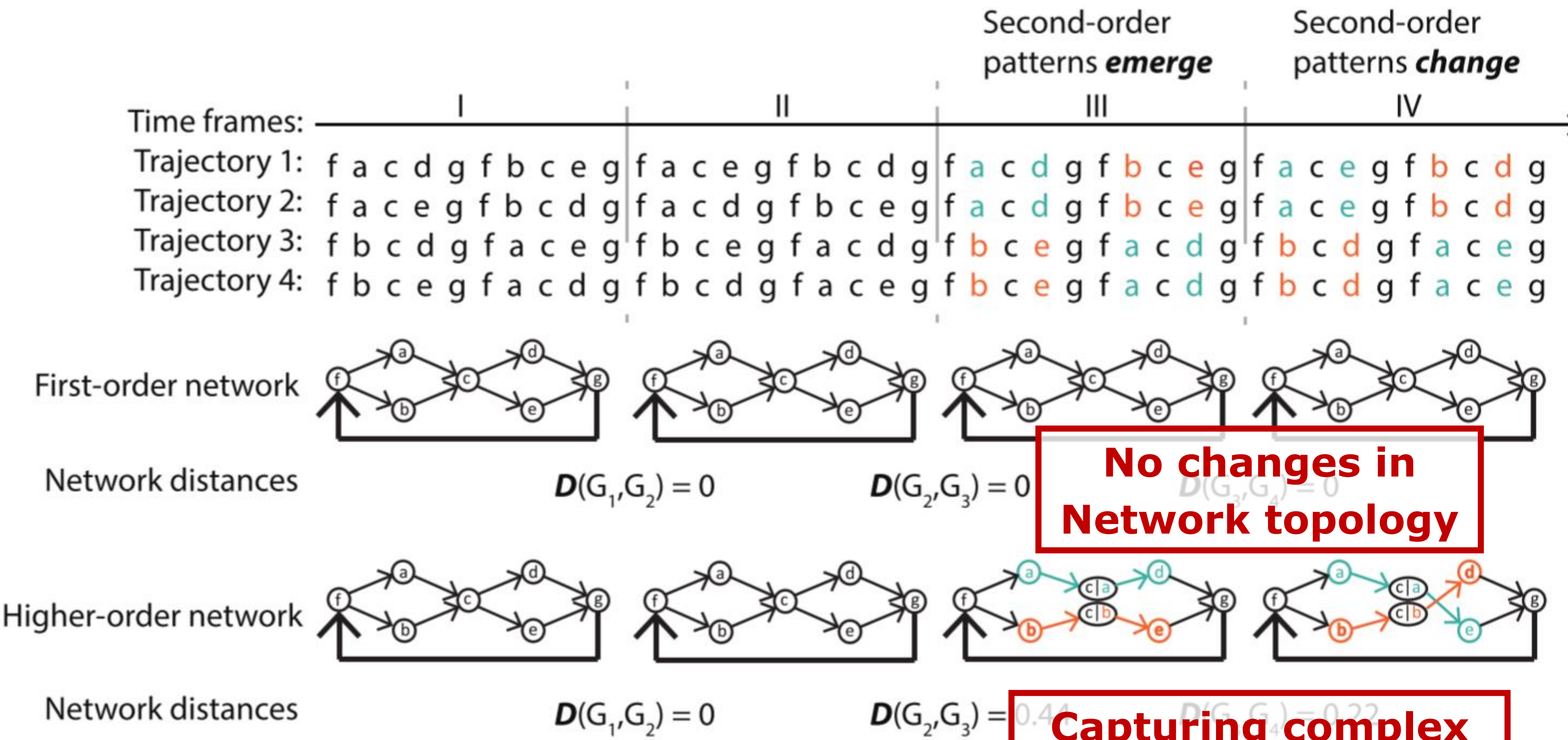
Anomaly detection with dynamic network



Anomaly detection with dynamic network



Anomaly detection with dynamic network



No changes in Network topology

Capturing complex anomalous patterns

Synthetic data with 11 billion movements

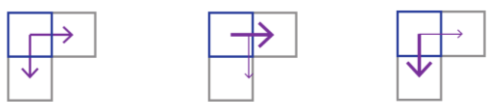
100,000 ships, each moving 100 steps;
11 scenarios, each repeating 100 times;
Total: 11,000,000,000 movements

First order

t = [1, 100]
Random walking right and down

t = [101, 200]
Add first order
@ cell 00, 03, 06

t = [201, 300]
Change first order
@ cell 00, 03, 06

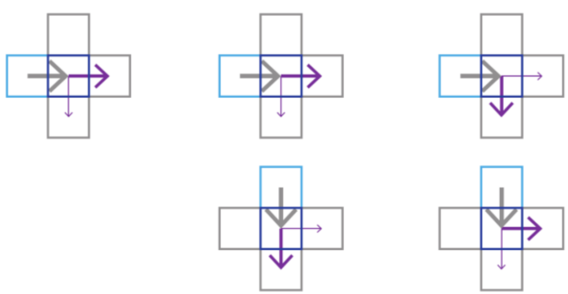


Second order

t = [301, 400]
Add second order
@ cell 28

t = [401, 500]
Add second orders
@ cell 31, 35

t = [501, 600]
Change second orders
@ cell 31, 35

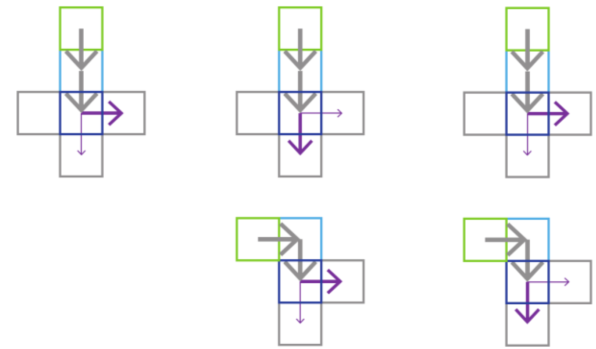


Third order

t = [601, 700]
Add third order
@ cell 81

t = [701, 800]
Add third orders
@ cell 84, 87

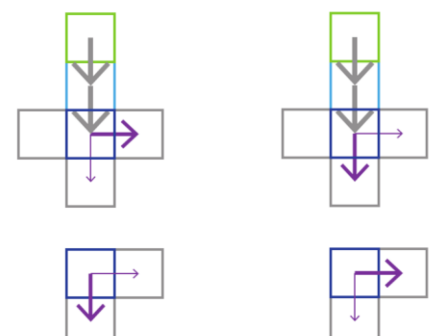
t = [801, 900]
Change third orders
@ cell 84, 87



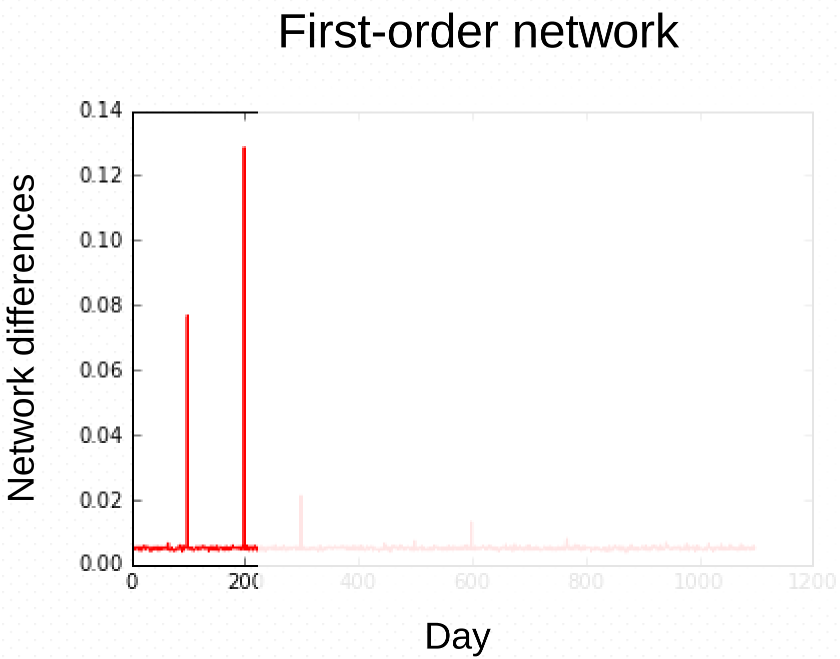
Mixed order

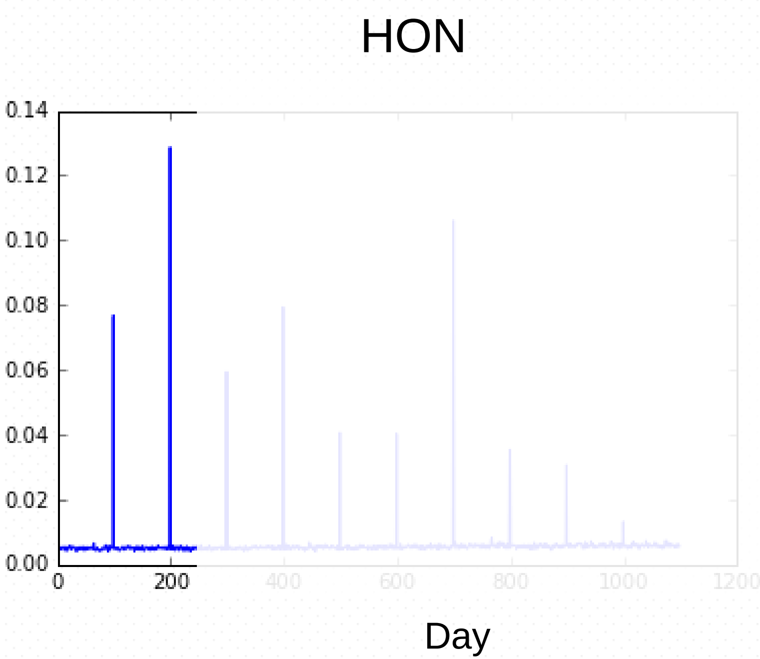
t = [901, 1000]
Add mixed orders
@ cell 59

t = [1001, 1100]
Change mixed orders
@ cell 59



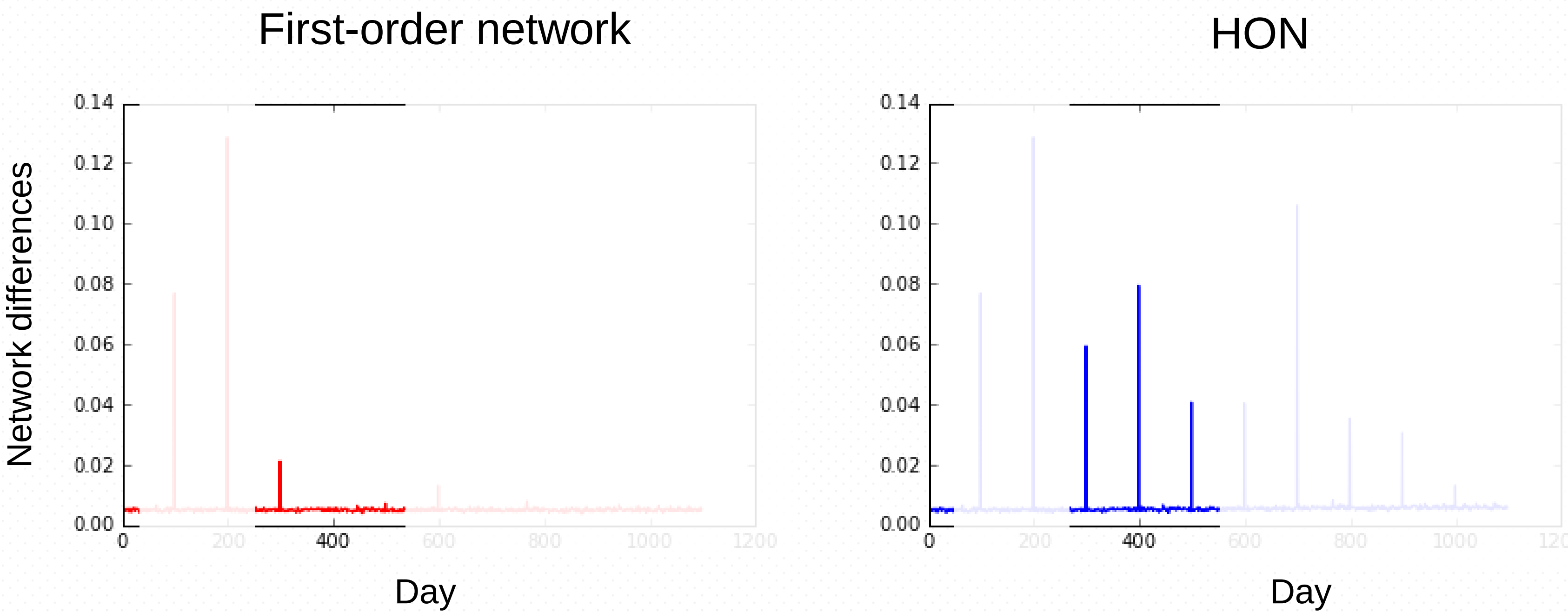
Higher-order anomalies captured by HON





Injection and alternation of 1st order dependencies

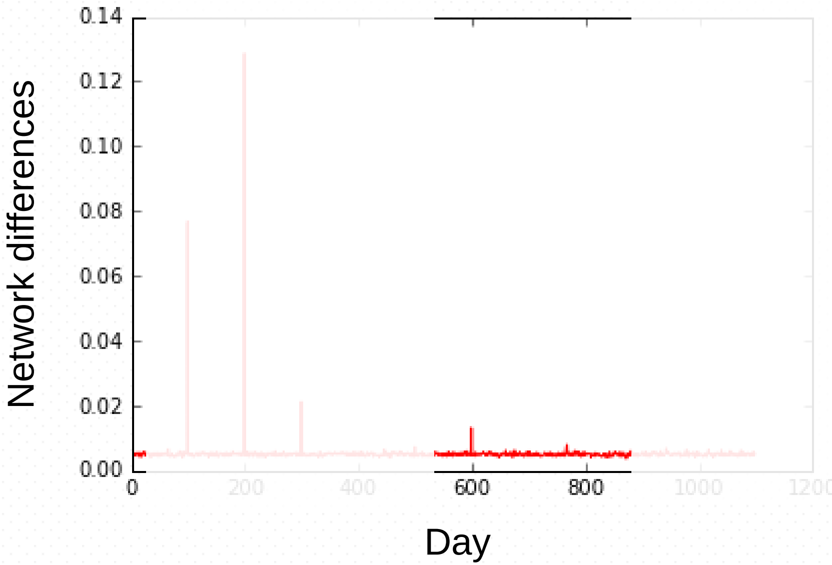
Higher-order anomalies captured by HON



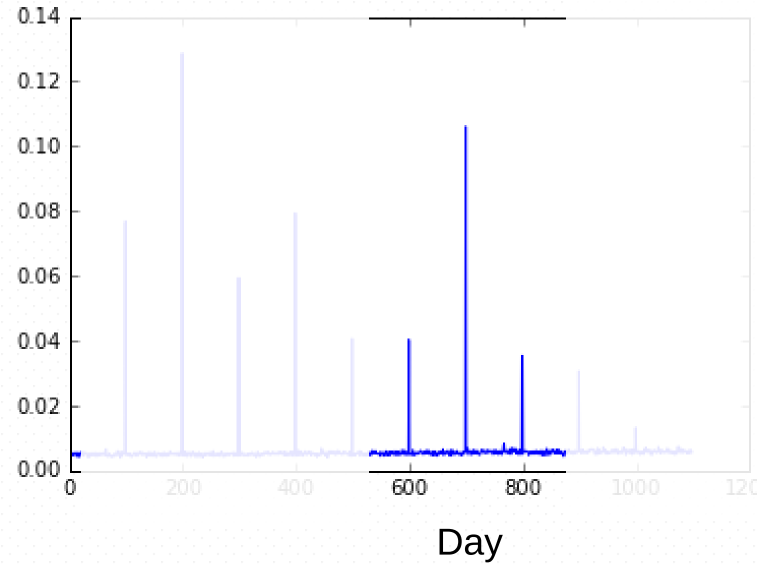
Injection and alternation of 2nd order dependencies

Higher-order anomalies captured by HON

First-order network



HON

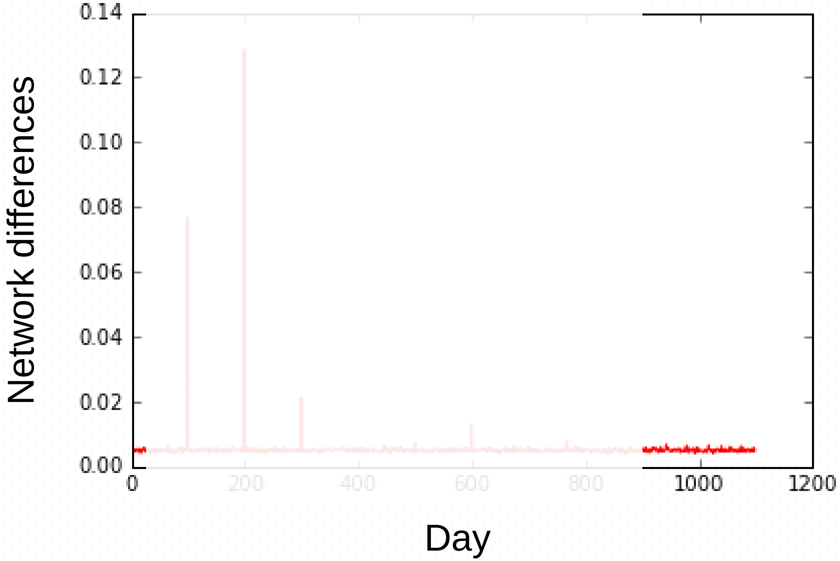


Injection and alternation of 3rd order dependencies

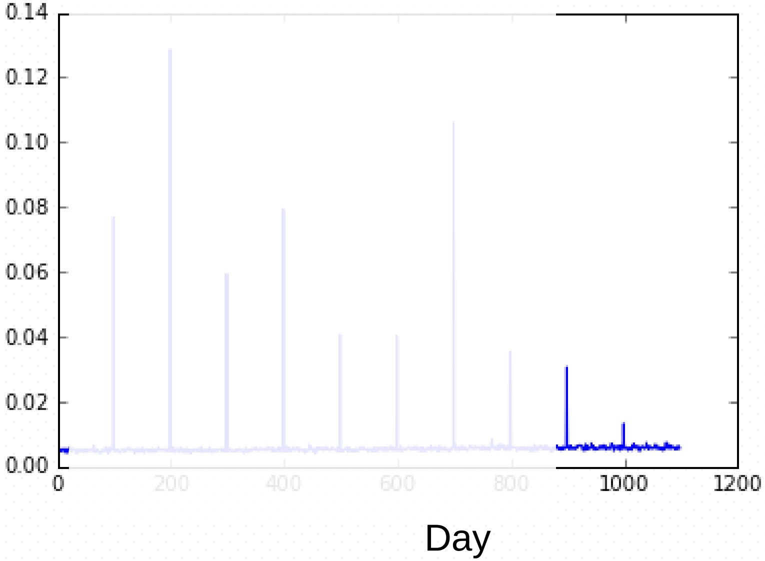
Higher-order anomalies captured by HON

Fails to capture certain anomalies

First-order network



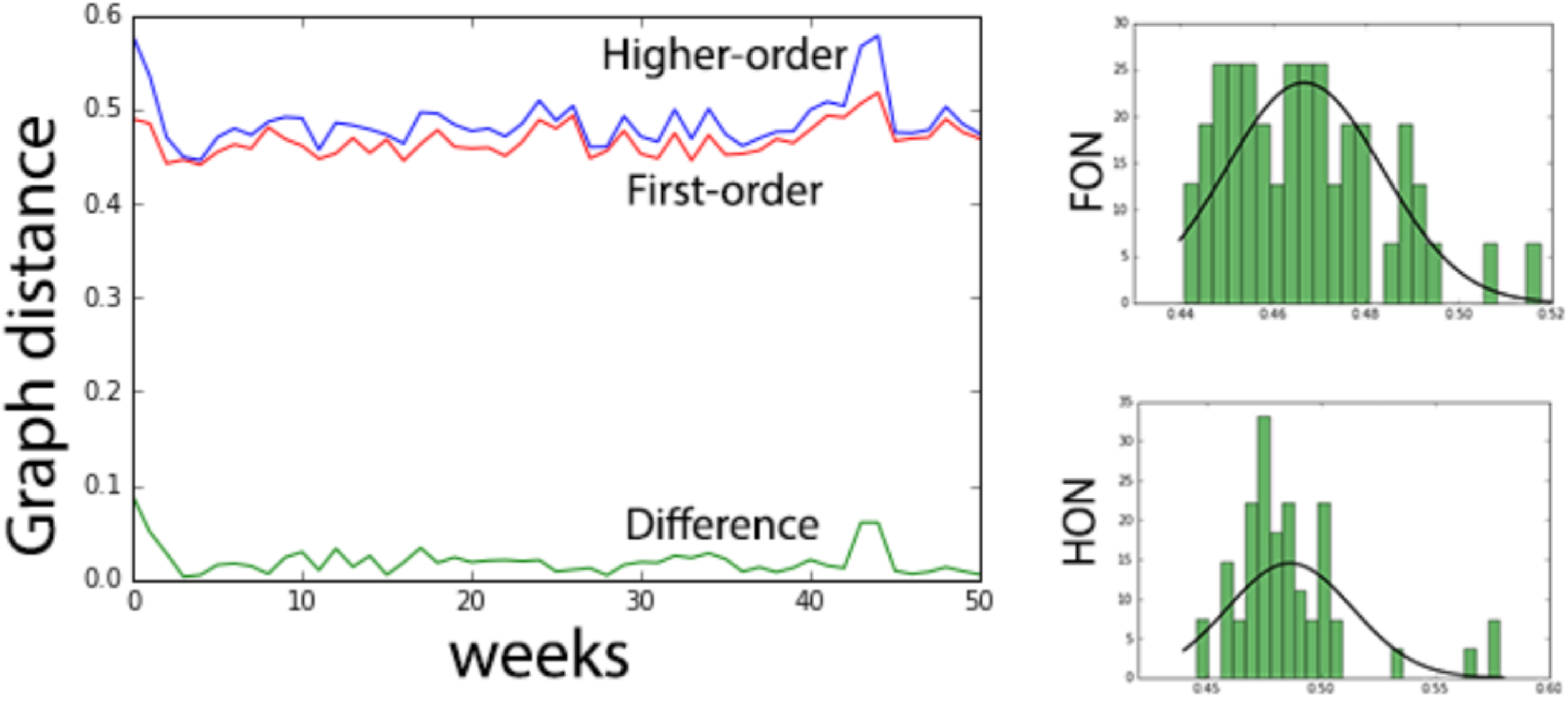
HON



Injection and alternation of higher-order dependencies

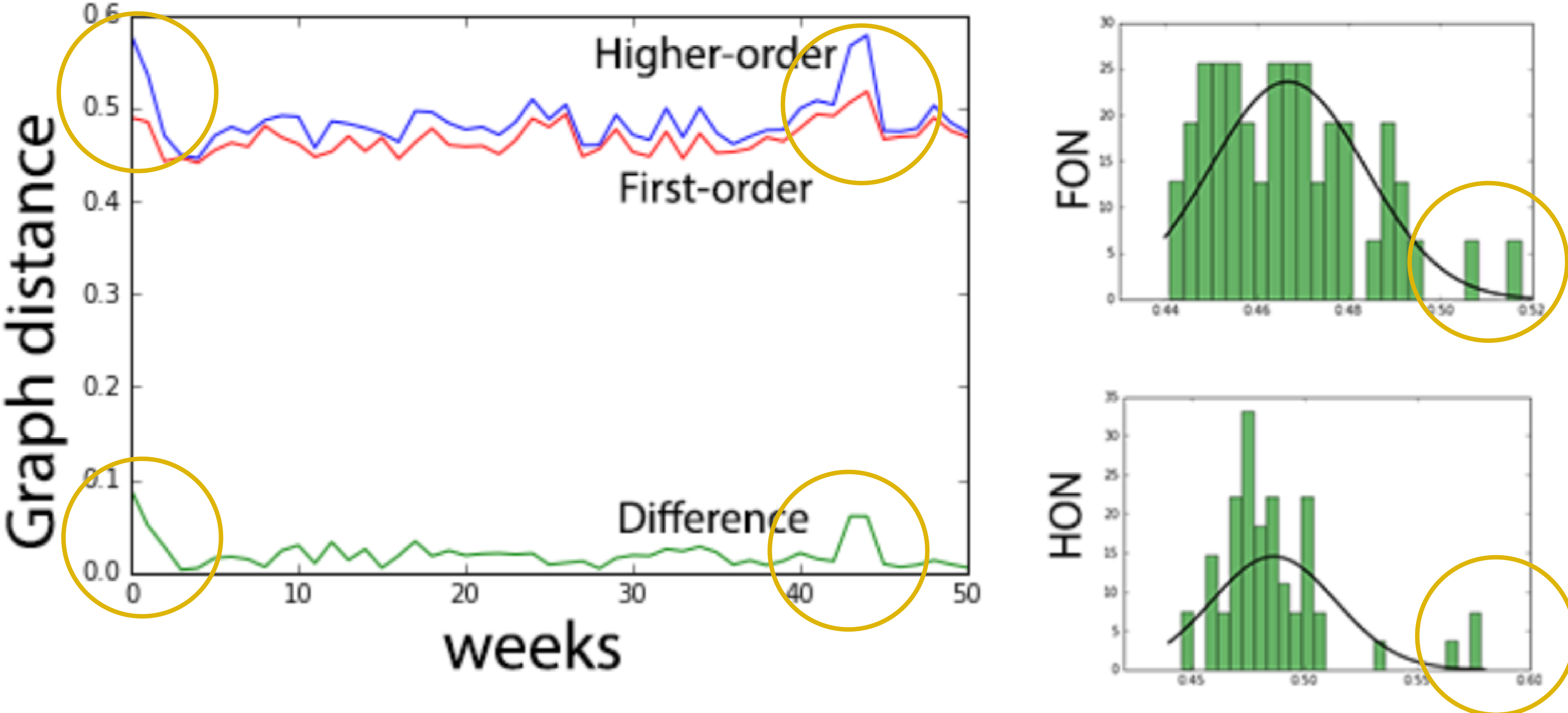
Higher-order anomalies captured by HON

Porto Taxi GPS trajectory data, 1 year



Higher-order anomalies captured by HON

Amplifying anomalous signals

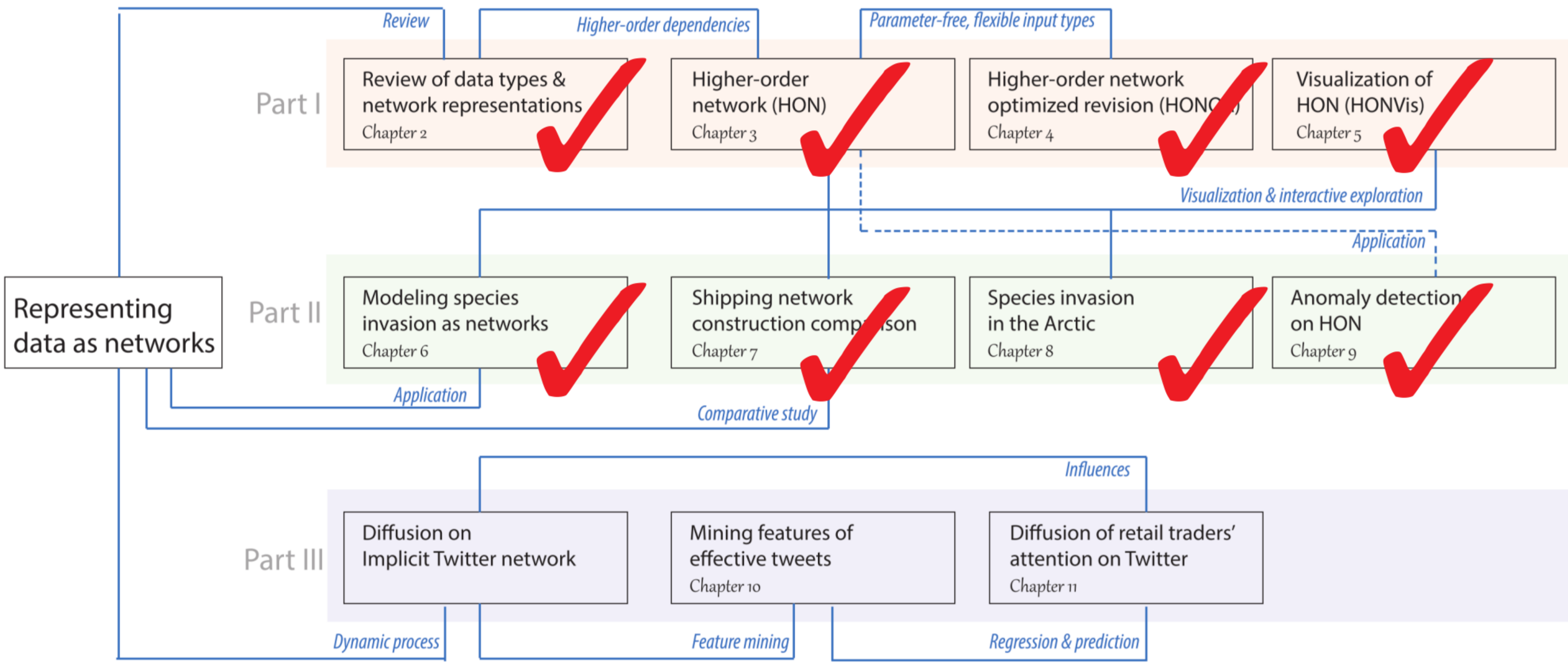


Takeaways

Anomaly detection on dynamic HON

- Unveils higher-order anomalies that are otherwise ignored
- Amplifies anomaly signals

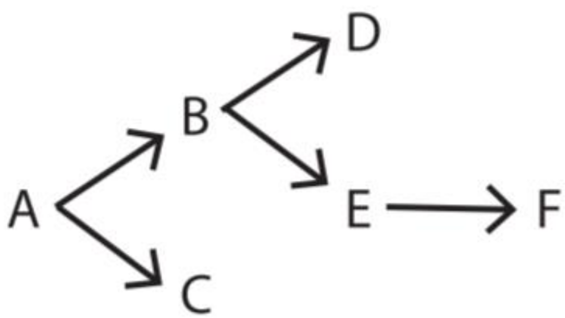
Overview


Review
Higher-order dependencies
Parameter-free, flexible input types
Part I
Review of data types & network representations
Chapter 2
Higher-order network (HON)
Chapter 3
Higher-order network optimized revision (HON+)
Chapter 4
Visualization of HON (HONVis)
Chapter 5
Visualization & interactive exploration
Application
Representing data as networks
Part II
Modeling species invasion as networks
Chapter 6
Shipping network construction comparison
Chapter 7
Species invasion in the Arctic
Chapter 8
Anomaly detection on HON
Chapter 9
Application
Comparative study
Influences
Part III
Diffusion on Implicit Twitter network
Mining features of effective tweets
Chapter 10
Diffusion of retail traders' attention on Twitter
Chapter 11
Dynamic process
Feature mining
Regression & prediction

Discussions

Flexible inputs

Raw diffusion data

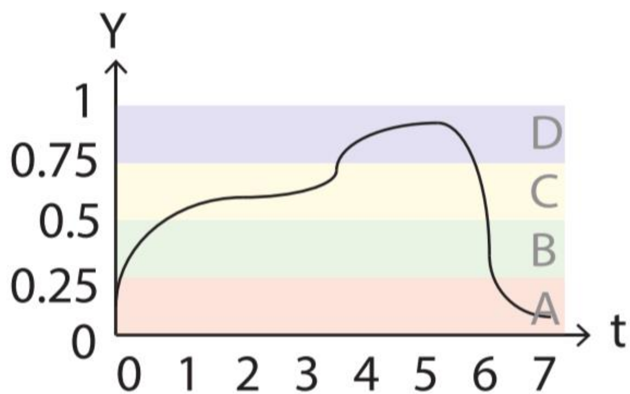


Observations

A B
A C
B D
B E
E F
A B D
A B E
B E F
A B E F

Flexible inputs

Raw time series data



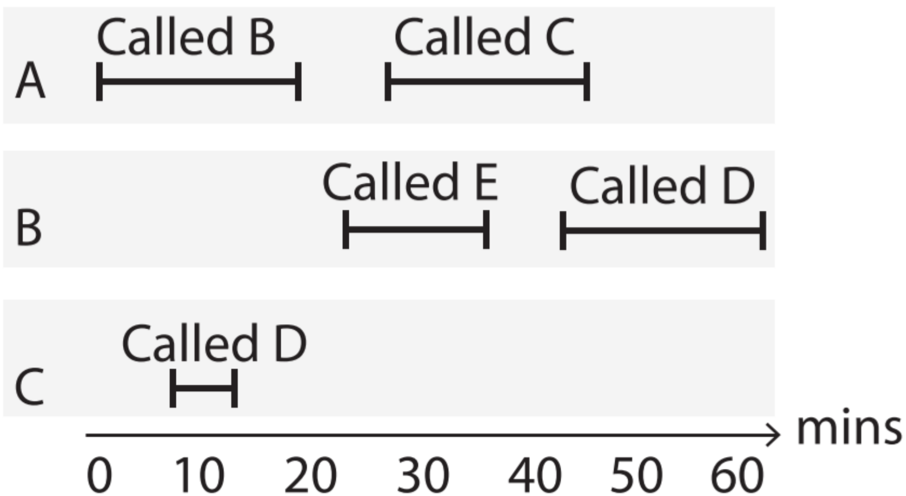
A C ~~C~~ ~~C~~ D ~~D~~ B A

Observations

A C
C D
D B
B A
A C D
C D B
D B A
A C D B
C D B A
A C D B A

Flexible inputs

Raw pairwise interaction temporal data



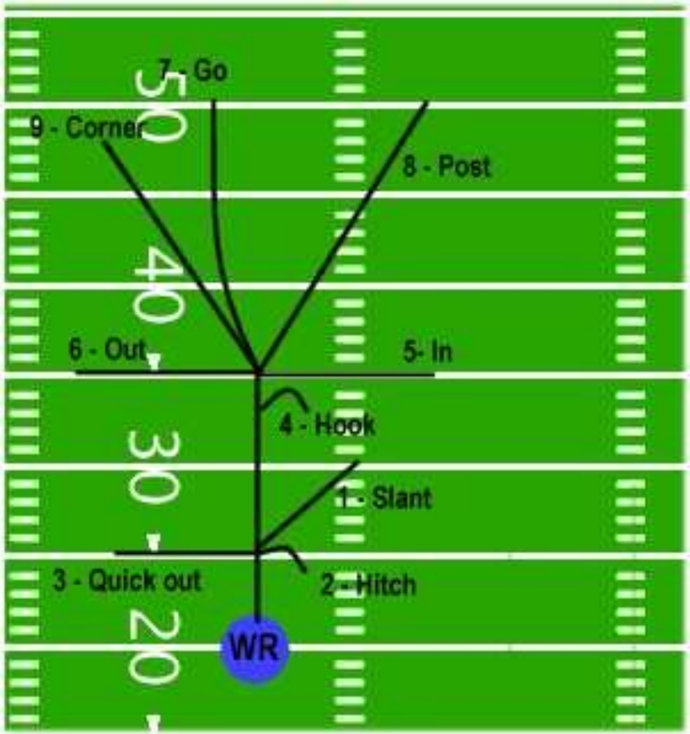
Observations

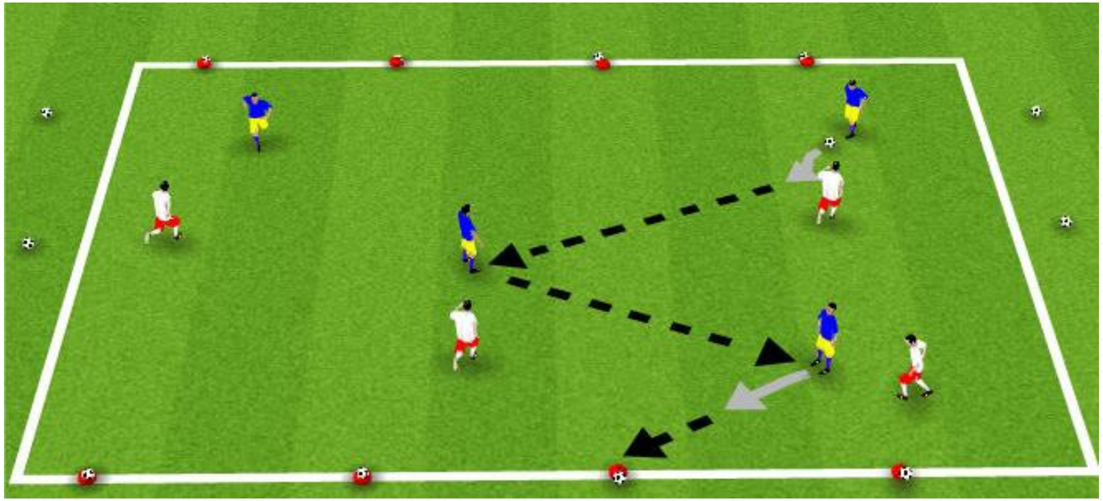
A B
A C
B E
B D
C D
C B
A B E

Varieties of data

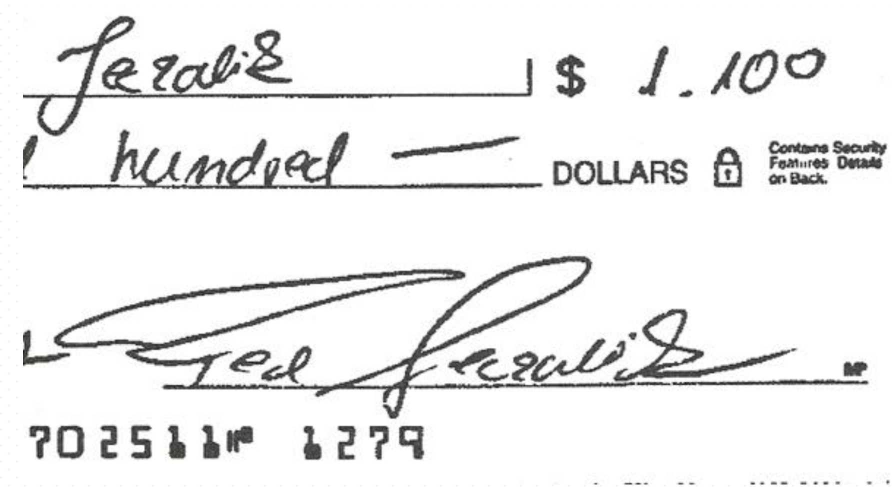
Application fields	Input Trajectories	Nodes	Edges
Transportation	Ship trajectories	Ports	Ship traffic
Computer network	Clickstreams	Web pages	Web traffic
Human interactions	Phone call or message cascades	People	Information flow
Human behavior	Human movements	POIs	Traffic
Healthcare	Patient records	Diseases	Disease evolutions
NLP	Sentences	Words	# word pairs

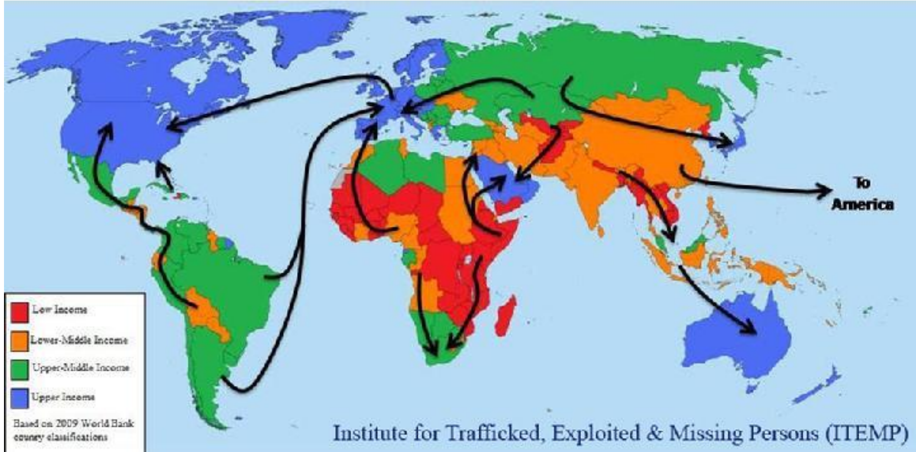
Other potential applications

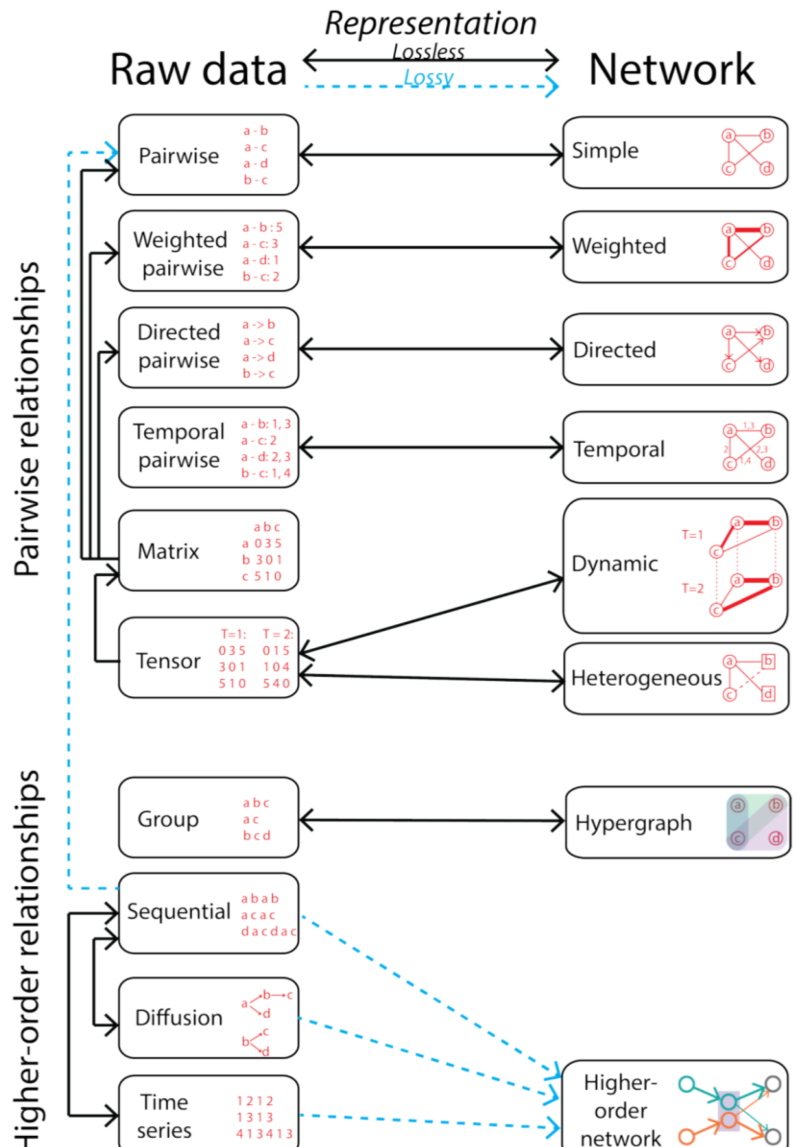


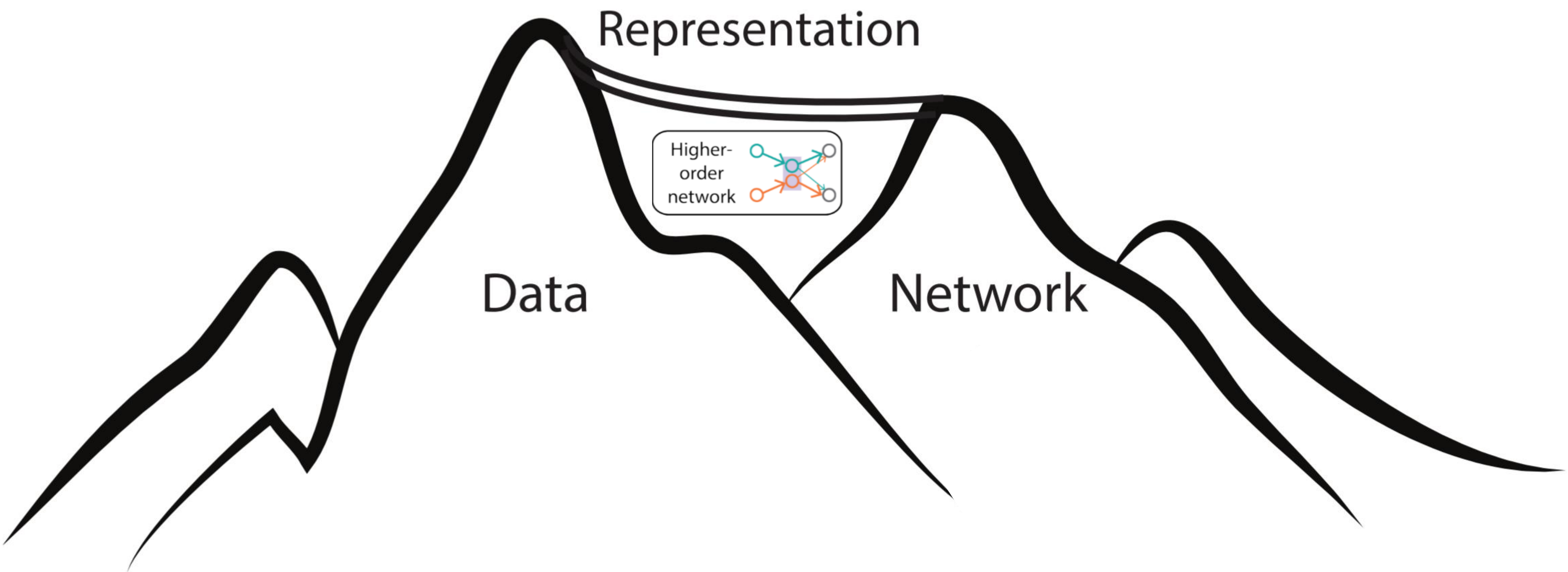


Other potential applications






Representation
Lossless
Lossy
Raw data
Network
Pairwise relationships
Higher-order relationships
Pairwise
a - b
a - c
a - d
b - c
Weighted pairwise
a - b : 5
a - c: 3
a - d: 1
b - c: 2
Directed pairwise
a -> b
a -> c
a -> d
b -> c
Temporal pairwise
a - b: 1, 3
a - c: 2
a - d: 2, 3
b - c: 1, 4
Matrix
a b c
a 0 3 5
b 3 0 1
c 5 1 0
Tensor
T=1:
0 3 5
3 0 1
5 1 0
T = 2:
0 1 5
1 0 4
5 4 0
Group
a b c
a c
b c d
Sequential
a b a b
a c a c
d a c d a c
Diffusion
Time series
1 2 1 2
1 3 1 3
4 1 3 4 1 3
Simple
Weighted
Directed
Temporal
Dynamic
T=1
T=2
Heterogeneous
Hypergraph
Higher-order network


Representation
Higher-
order
network
Data
Network

Research outputs

As leading student author:

- HON, published @ *Science advances*: **Jian Xu**, Thanuka L. Wickramarathne, and Nitesh V. Chawla. "Representing higher-order dependencies in networks." 2, no. 5 (2016): e1600028.
- HoNVis, published @ *IEEE PacificVis*: Jun Tao, **Jian Xu**, Chaoli Wang, and Nitesh V. Chawla. "HoNVis: Visualizing and Exploring Higher-Order Networks."
- HoNVis, demo published @ *IEEE IoTDI*: Jian Xu, Jun Tao, Nitesh V. Chawla and Chaoli Wang. “Visual Analytics of Higher-order Dependencies in Sensor Data”
- Species invasions, published @ *ACM SIGKDD*: **Jian Xu**, Thanuka L. Wickramarathne, Nitesh V. Chawla, Erin K. Grey, Karsten Steinhaeuser, Reuben P. Keller, John M. Drake, and David M. Lodge. "Improving management of aquatic invasions by integrating shipping network, ecological, and environmental data: data mining for social good.”

- Retail diffusion: under review @ *Journal of Management Science*: Nitesh Chawla, Zhi Da, **Jian Xu**, and Mao Ye. *Catching fire: the diffusion of retail attention on twitter*.
- Effective tweeting: under review @ *ASONAM*: **Jian Xu**, Nitesh Chawla.

- Arctic species invasion: to submit to *Nature Communications*. Jian Xu, Salvatore Curasi, Erin Grey, Nitesh Chawla and David Lodge. “Species introduction and diffusion in the Arctic through global shipping: risk assessment and projection”
- Anomaly detection with HON: to submit to *ICDM*. Jian Xu, Nitesh Chawla.

Research outputs

Other published collaborative work:

- Structural diversity, published @ *ACM SIGKDD*: Yuxiao Dong, Reid A. Johnson, **Jian Xu** and Nitesh V. Chawla. “Structural Diversity and Homophily: A Study Across More Than One Hundred Big Networks”

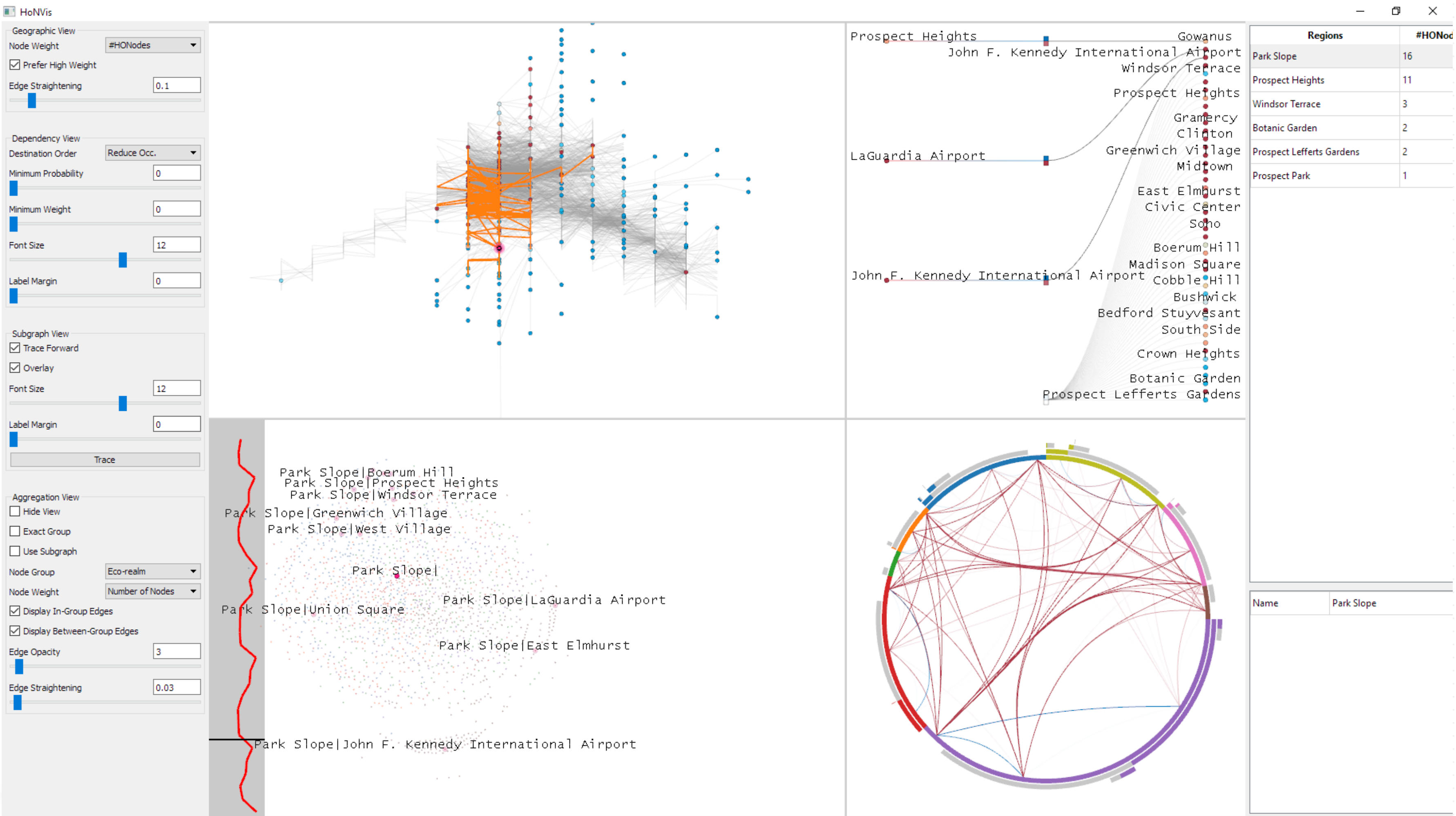
- Temporal motifs, published @ *IEEE Transaction on Systems, Man and Cybernetics*. Zhang, Yi-Qing, Xiang Li, **Jian Xu**, and Athanasios V. Vasilakos. "Human interactive patterns in temporal networks."

Other work in progress:

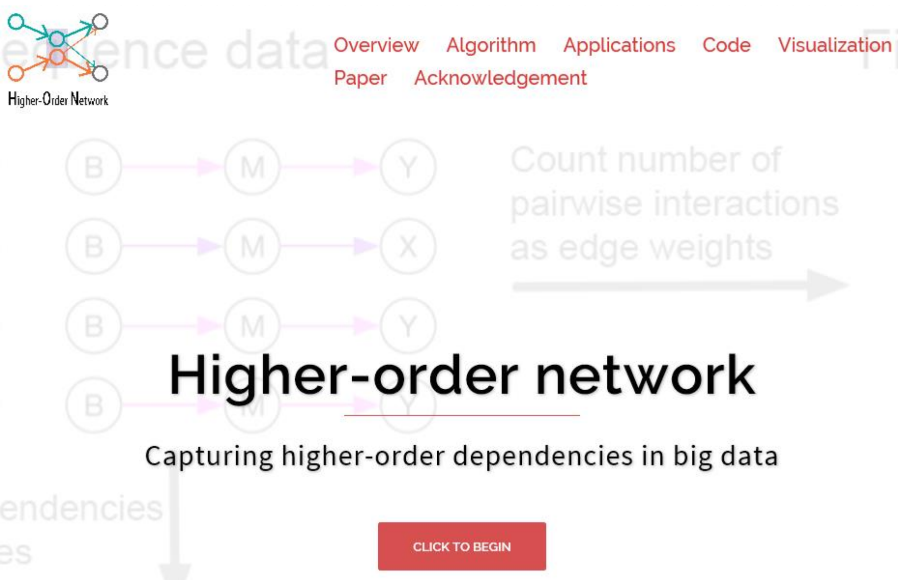
- HONVis extension: adding the time dimension, and the anomaly detection module.

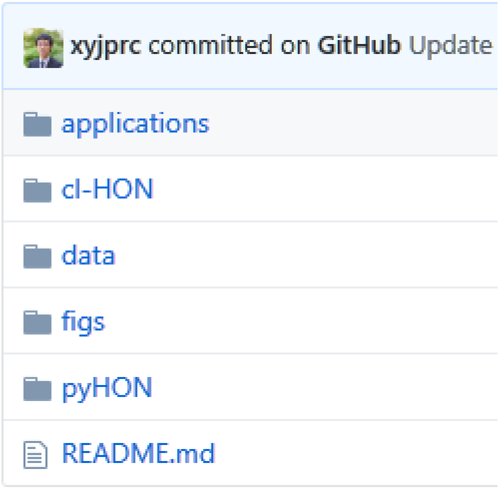
- Comparative analysis of different network representations of global shipping.

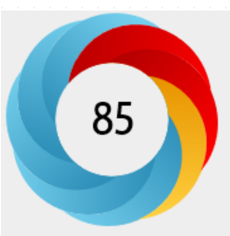
HoNVis for dynamic HON & anomaly detection



For the community







In the top 5% of all research outputs scored by Altmetric

High Attention Score compared to outputs of the same age (97th percentile)

Acknowledgements: committee

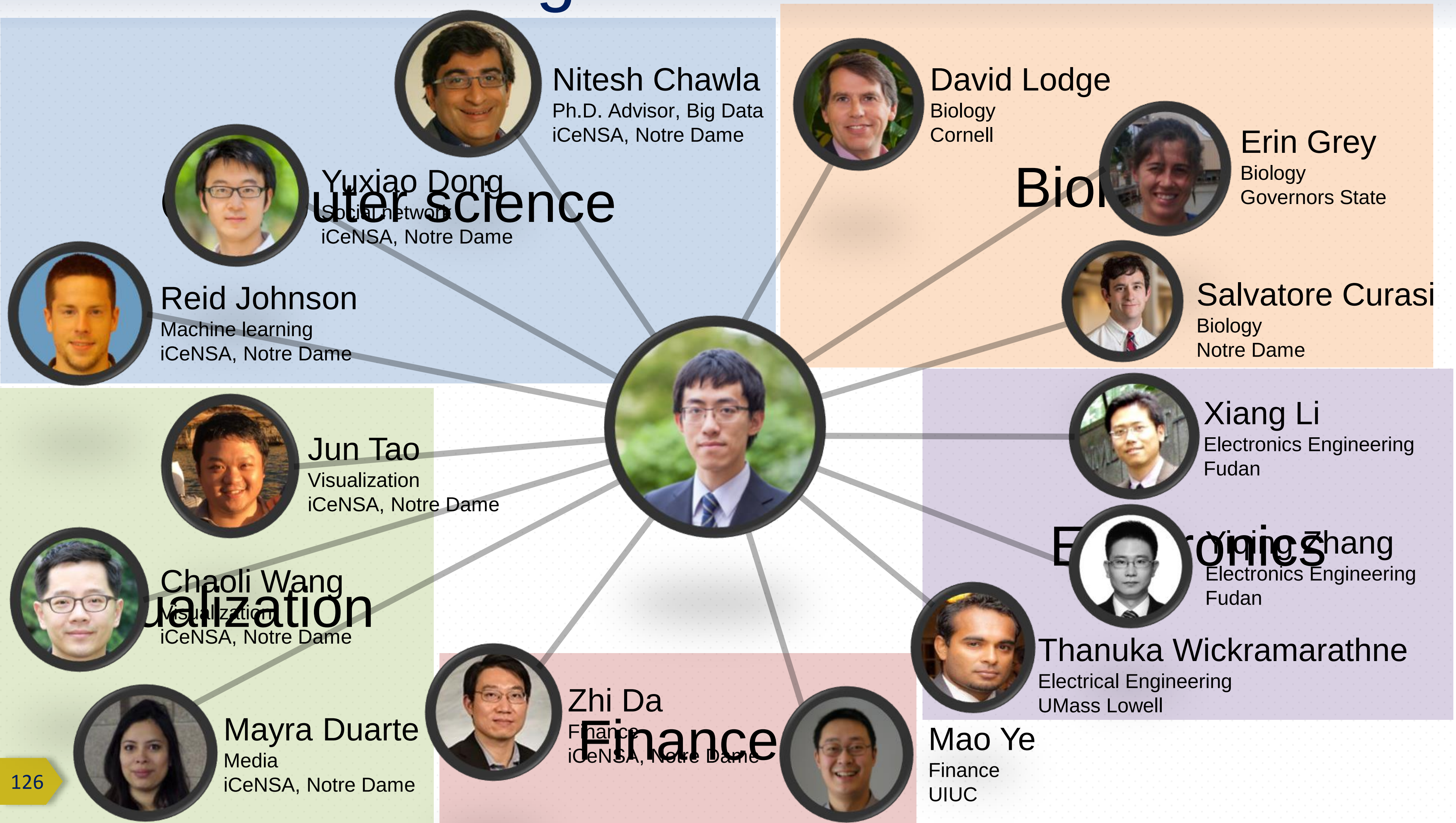
Prof. Nitesh Chawla, *chair*

Prof. David Lodge

Prof. Tijana Milenkovic

Prof. Zoltan Torotzkai

Acknowledgements: collaborators



Acknowledgements: friends

Acknowledgements: Funding

NSF

ARL

NETWORK
SCIENCE
CTA

Thank you!

Jian Xu

UNIVERSITY OF NOTRE DAME

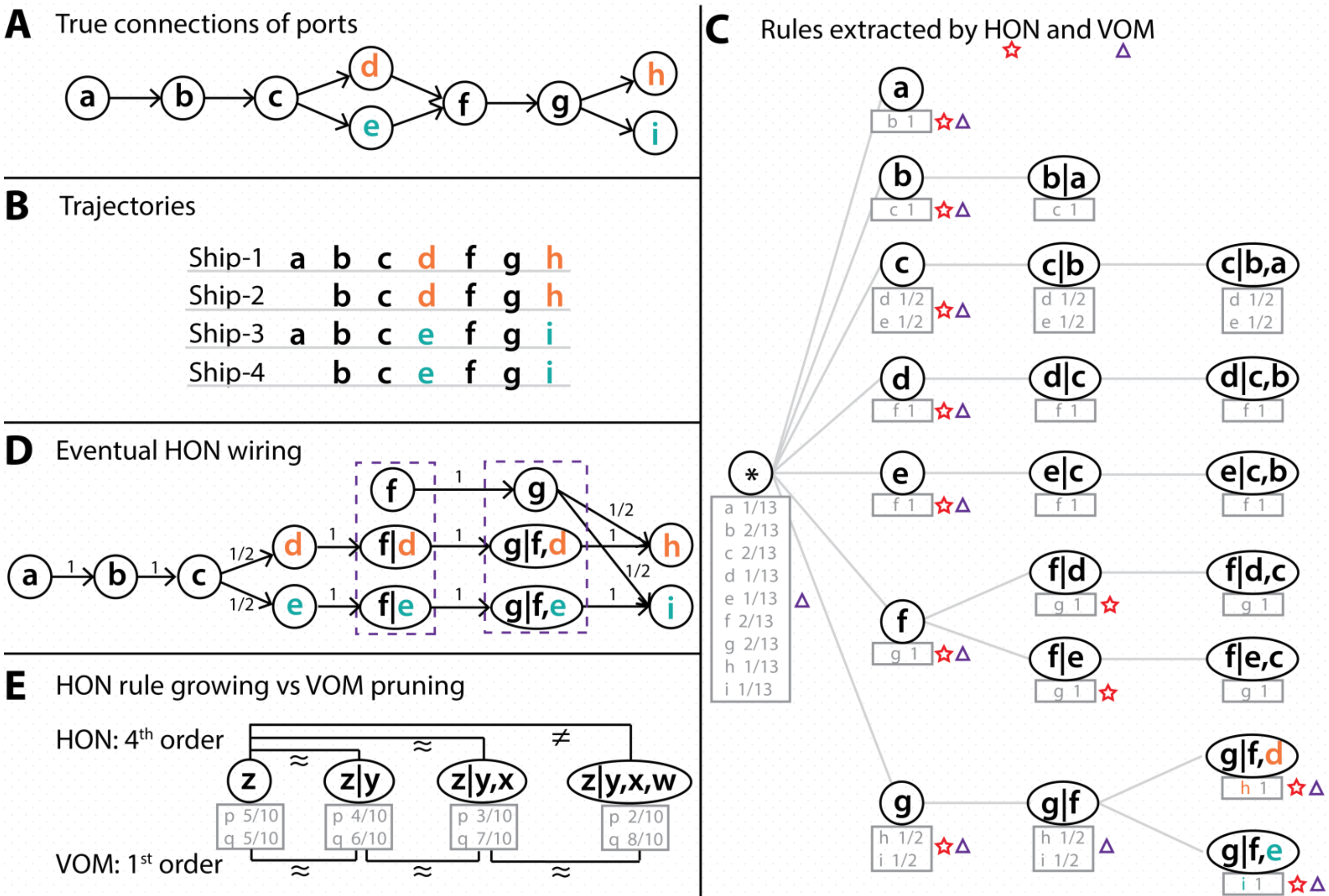
iCeNSA
Interdisciplinary Center for Network Science & Applications

Appendix

References

- Rosvall, Martin, Alcides V. Esquivel, Andrea Lancichinetti, Jevin D. West, and Renaud Lambiotte. "Memory in network flows and its effects on spreading dynamics and community detection." *Nature communications* 5 (2014).
- Chierichetti, Flavio, Ravi Kumar, Prabhakar Raghavan, and Tamas Sarlos. "Are web users really markovian?." In *Proceedings of the 21st international conference on World Wide Web*, pp. 609-618. ACM, 2012.
- Pons, Pascal, and Matthieu Latapy. "Computing communities in large networks using random walks." *J. Graph Algorithms Appl.* 10, no. 2 (2006): 191-218.
- Page, Lawrence, Sergey Brin, Rajeev Motwani, and Terry Winograd. "The PageRank citation ranking: bringing order to the web." (1999).
- Gleich, David F., Lek-Heng Lim, and Yongyang Yu. "Multilinear PageRank." *SIAM Journal on Matrix Analysis and Applications* 36, no. 4 (2015): 1507-1541.
- Akoglu, Leman, Mary McGlohon, and Christos Faloutsos. "Anomaly detection in large graphs." In *In CMU-CS-09-173 Technical Report*. 2009.
- Seebens, H., M. T. Gastner, and B. Blasius. "The risk of marine bioinvasion caused by global shipping." *Ecology Letters* 16, no. 6 (2013): 782-790.
- Rosvall, Martin, and Carl T. Bergstrom. "Maps of random walks on complex networks reveal community structure." *Proceedings of the National Academy of Sciences* 105, no. 4 (2008): 1118-1123.
- Ducruet, César. "Network diversity and maritime flows." *Journal of Transport Geography* 30 (2013): 77-88.
- Ducruet, César, Sung-Woo Lee, and Adolf KY Ng. "Centrality and vulnerability in liner shipping networks: revisiting the Northeast Asian port hierarchy." *Maritime Policy & Management* 37, no. 1 (2010): 17-36.
- Ducruet, César, Céline Rozenblat, and Faraz Zaidi. "Ports in multi-level maritime networks: evidence from the Atlantic (1996–2006)." *Journal of Transport Geography* 18, no. 4 (2010): 508-518.
- Ducruet, César, and Theo Notteboom. "The worldwide maritime network of container shipping: spatial structure and regional dynamics." *Global Networks* 12, no. 3 (2012): 395-423.
- Kaluza, Pablo, Andrea Kölzsch, Michael T. Gastner, and Bernd Blasius. "The complex network of global cargo ship movements." *Journal of the Royal Society Interface* 7, no. 48 (2010): 1093-1103.
- Hu, Yihong, and Daoli Zhu. "Empirical analysis of the worldwide maritime transportation network." *Physica A: Statistical Mechanics and its Applications* 388, no. 10 (2009): 2061-2071.

Comparison with VOM



Scalability

Scalability

Network representation (global shipping data)	Number of edges	Number of nodes	Network density	Clustering time (mins)	Ranking time (s)
Conventional first-order	31,028	2,675	4.3×10^{-3}	4	1.3
Fixed second-order	116,611	19,182	3.2×10^{-4}	73	7.7
HON, max order two	64,914	17,235	2.2×10^{-4}	45	4.8
HON, max order three	78,415	26,577	1.1×10^{-4}	63	6.2
HON, max order four	83,480	30,631	8.9×10^{-5}	67	7.0
HON, max order five	85,025	31,854	8.4×10^{-5}	68	7.6

Scalability

Network representation (global shipping data)	Number of edges	Number of nodes	Network density	Clustering time (mins)	Ranking time (s)
Conventional first-order	31,028	2,675	4.3×10^{-3}	4	1.3
Fixed second-order	116,611	19,182	3.2×10^{-4}	73	7.7
HON, max order two	64,914	17,235	2.2×10^{-4}	45	4.8
HON, max order three	78,415	26,577	1.1×10^{-4}	63	6.2
HON, max order four	83,480	30,631	8.9×10^{-5}	67	7.0
HON, max order five	85,025	31,854	8.4×10^{-5}	68	7.6

Scalability

Network representation (global shipping data)	Number of edges	Number of nodes	Network density	Clustering time (mins)	Ranking time (s)
Conventional first-order	31,028	2,675	4.3×10^{-3}	4	1.3
Fixed second-order	116,611	19,182	3.2×10^{-4}	73	7.7
HON, max order two	64,914	17,235	2.2×10^{-4}	45	4.8
HON, max order three	78,415	26,577	1.1×10^{-4}	63	6.2
HON, max order four	83,480	30,631	8.9×10^{-5}	67	7.0
HON, max order five	85,025	31,854	8.4×10^{-5}	68	7.6

Scalability

Network representation (global shipping data)	Number of edges	Number of nodes	Network density	* Clustering time (mins)	** Ranking time (s)
Conventional first-order	31,028	2,675	4.3×10^{-3}	4	1.3
Fixed second-order	116,611	19,182	3.2×10^{-4}	73	7.7
HON, max order two	64,914	17,235	2.2×10^{-4}	45	4.8
HON, max order three	78,415	26,577	1.1×10^{-4}	63	6.2
HON, max order four	83,480	30,631	8.9×10^{-5}	67	7.0
HON, max order five	85,025	31,854	8.4×10^{-5}	68	7.6

* Using MapEquation with 1000 iterations
** Using PageRank

Goals

How shall we represent such big data derived from complex system as networks, and accurately capture higher-order dependencies?

- Network representation
  - Accuracy
  - Compatibility
  - Scalability

Higher-order dependencies revealed by HON

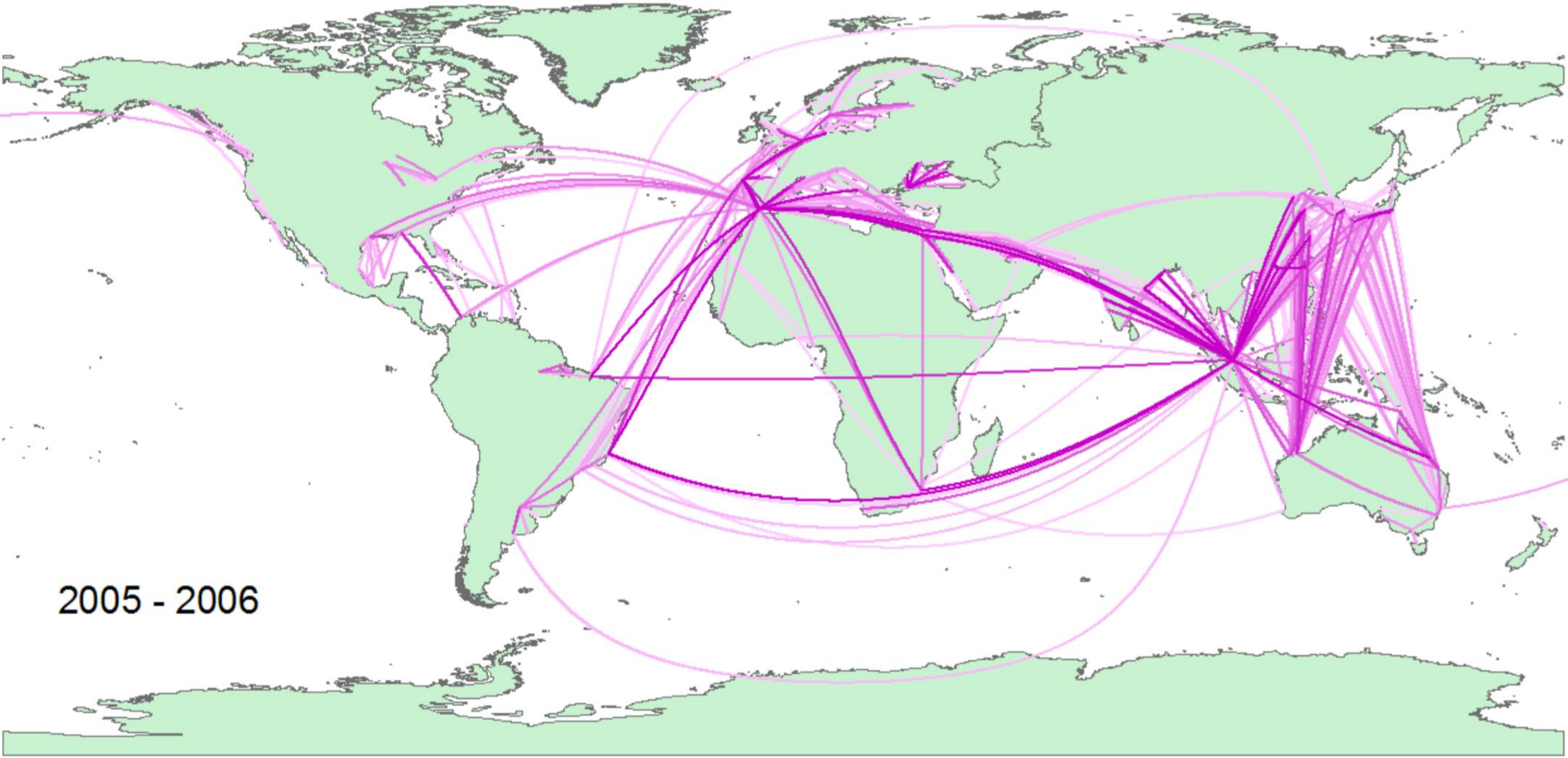
Data	# Records	Inject known variable-order dependencies
Synthetic	10,000,000	10 second-order 10 third-order 10 fourth-order

- **Effectiveness**: correctly captures all 30 of the higher-order dependencies
- **Accuracy**: does not extract false dependencies beyond the fourth order
- **Compactness**: determines that all other dependencies are first-order

Clustering: higher-order network

- 45% of ports belong to more than one cluster
- 44 ports (1.7% of all) belong to five clusters
  - New York, Shanghai, Hong Kong, Gibraltar, Hamburg, etc.
- Panama Canal belongs to six clusters
- Highlighting ports that may be invaded by species from multiple regions

Ship-borne species diffusion pathways



Ranking on clickstream network

Pages that gain PageRank scores	ΔPageRank	Pages that lose PageRank scores	ΔPageRank
South Bend Tribune - Home.	+0.0119	KTUU - Home.	−0.0057
Hagerstown News / **obituaries** - Front.	+0.0115	KWCH - Home.	−0.0031
South Bend Tribune - **Obits** - 3rd Party.	+0.0112	Imperial Valley Press - Home.	−0.0011
South Bend Tribune / sports / notredame - Front.	+0.0102	Hagerstown News / sports - Front.	−0.0005
Aberdeen News / news / **obituaries** - Front.	+0.0077	Imperial Valley Press / classifieds / topjobs - Front.	−0.0004
WDBJ7 - Home.	+0.0075	Gaylord - Home.	−0.0004
KY3 / **weather** - Front.	+0.0075	WDBJ7 / weather / web-cams - Front.	−0.0004
Hagerstown News - Home.	+0.0072	KTUU / about / **meetnewsteam** - Front.	−0.0003
Daily American / lifestyle / **obituaries** - Front.	+0.0054	Smithsburg man faces more charges following salvag	−0.0003
WDBJ7 / **weather** / closings - Front.	+0.0048	KWCH / about / station / **newsteam** - Front.	−0.0003
WSBT TV / **weather** - Front.	+0.0041	South Bend Tribune / sports / highschoolsports - Front.	−0.0003
Daily American - Home.	+0.0036	Hagerstown News / opinion - Front.	−0.0002
WDBJ7 / **weather** / radar - Front.	+0.0036	WDBJ7 / news / **anchors-reporters** - Front.	−0.0002
WDBJ7 / **weather** / 7-day-planner - Front.	+0.0031	Petoskey News / news / obituaries - Front.	−0.0002
WDBJ7 / **weather** - Front.	+0.0019	KWCH / news - Front.	−0.0002

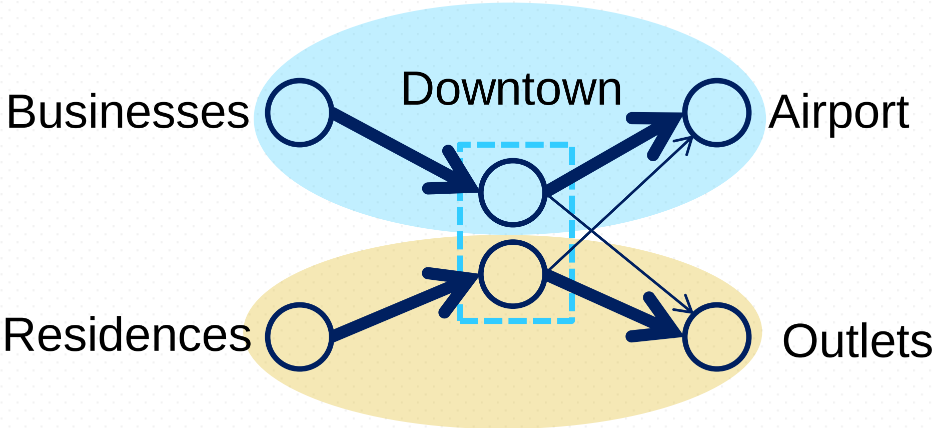
Ranking on clickstream network

Pages that gain PageRank scores	ΔPageRank	Pages that lose PageRank scores	ΔPageRank
South Bend Tribune - Home.	+0.0119	KTUU - Home.	−0.0057
Hagerstown News / **obituaries** - Front.	+0.0115	KWCH - Home.	−0.0031
South Bend Tribune - **Obits** - 3rd Party.	+0.0112	Imperial Valley Press - Home.	−0.0011
South Bend Tribune / sports / notredame - Front.	+0.0102	Hagerstown News / sports - Front.	−0.0005
Aberdeen News / news / **obituaries** - Front.	+0.0077	Imperial Valley Press / classifieds / topjobs - Front.	−0.0004
WDBJ7 - Home.	+0.0075	Gaylord - Home.	−0.0004
KY3 / **weather** - Front.	+0.0075	WDBJ7 / weather / web-cams - Front.	−0.0004
Hagerstown News - Home.	+0.0072	KTUU / about / **meetnewsteam** - Front.	−0.0003
Daily American / lifestyle / **obituaries** - Front.	+0.0054	Smithsburg man faces more charges following salvag	−0.0003
WDBJ7 / **weather** / closings - Front.	+0.0048	KWCH / about / station / **newsteam** - Front.	−0.0003
WSBT TV / **weather** - Front.	+0.0041	South Bend Tribune / sports / highschoolsports - Front.	−0.0003
Daily American - Home.	+0.0036	Hagerstown News / opinion - Front.	−0.0002
WDBJ7 / **weather** / radar - Front.	+0.0036	WDBJ7 / news / **anchors-reporters** - Front.	−0.0002
WDBJ7 / **weather** / 7-day-planner - Front.	+0.0031	Petoskey News / news / obituaries - Front.	−0.0002
WDBJ7 / **weather** - Front.	+0.0019	KWCH / news - Front.	−0.0002

No changes
to the ranking algorithm

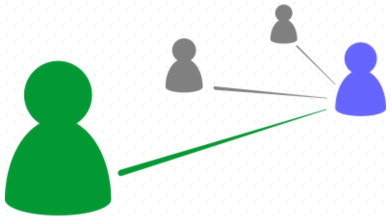
Interdisciplinary applications

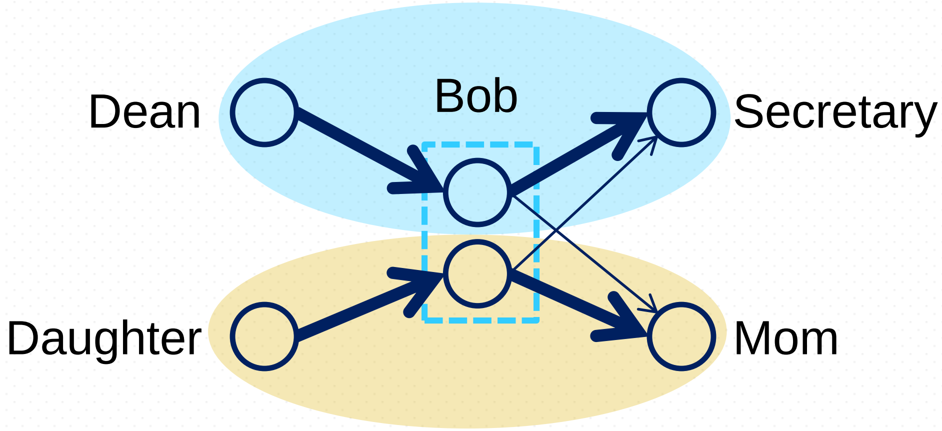
Urban planning
&
Event detection



Interdisciplinary applications

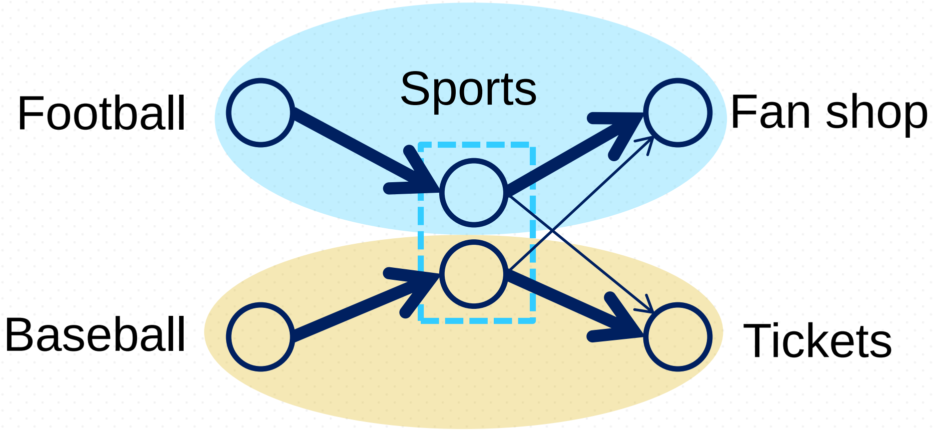
Social network
&
Information diffusion





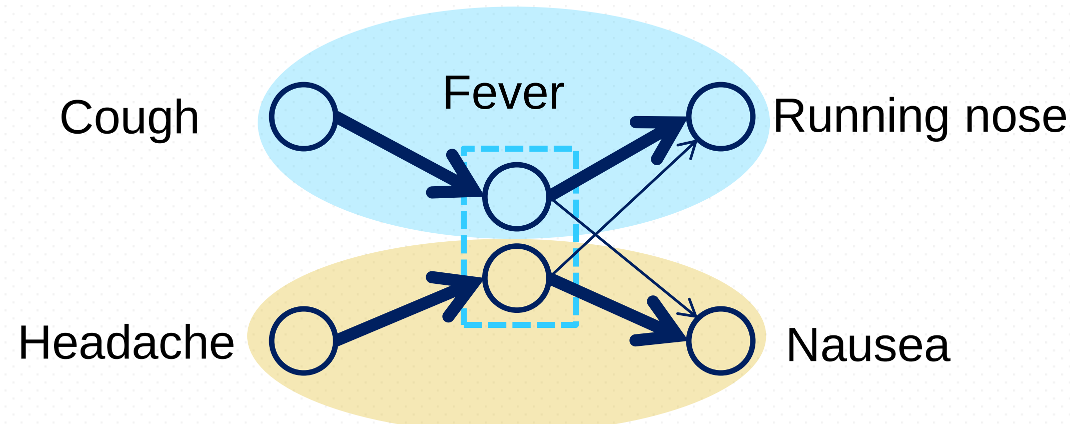
Interdisciplinary applications

Web optimization,
advertising,
network security

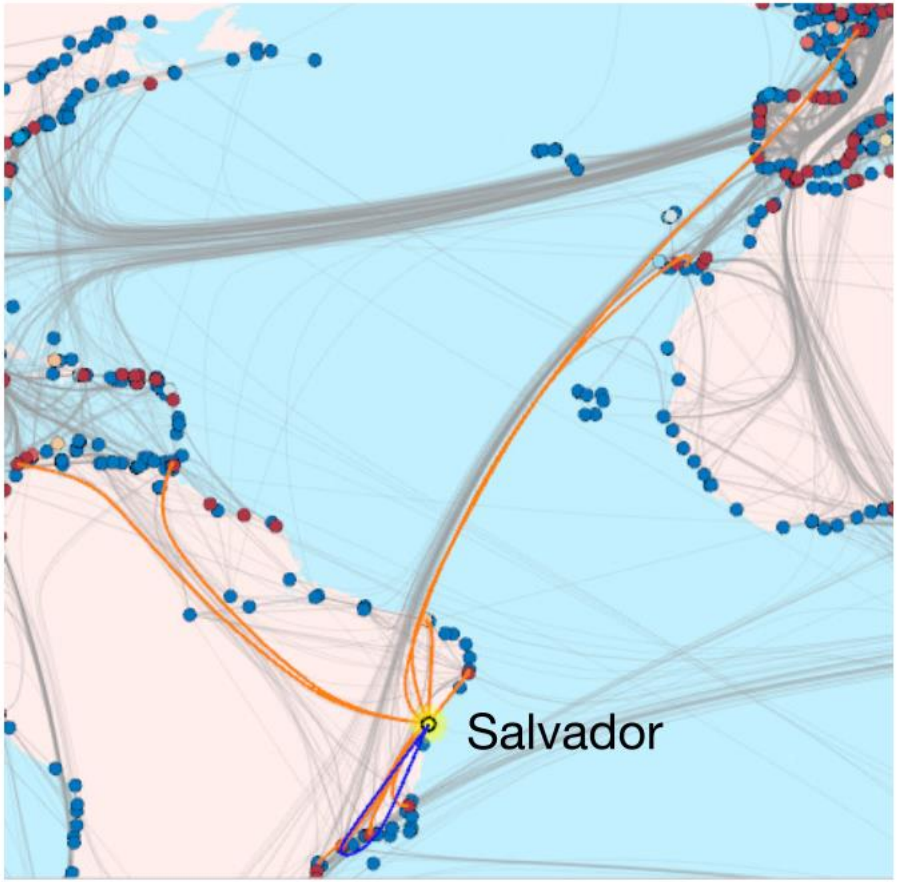


Interdisciplinary applications

Healthcare,
Epidemics monitoring,
Gene tech

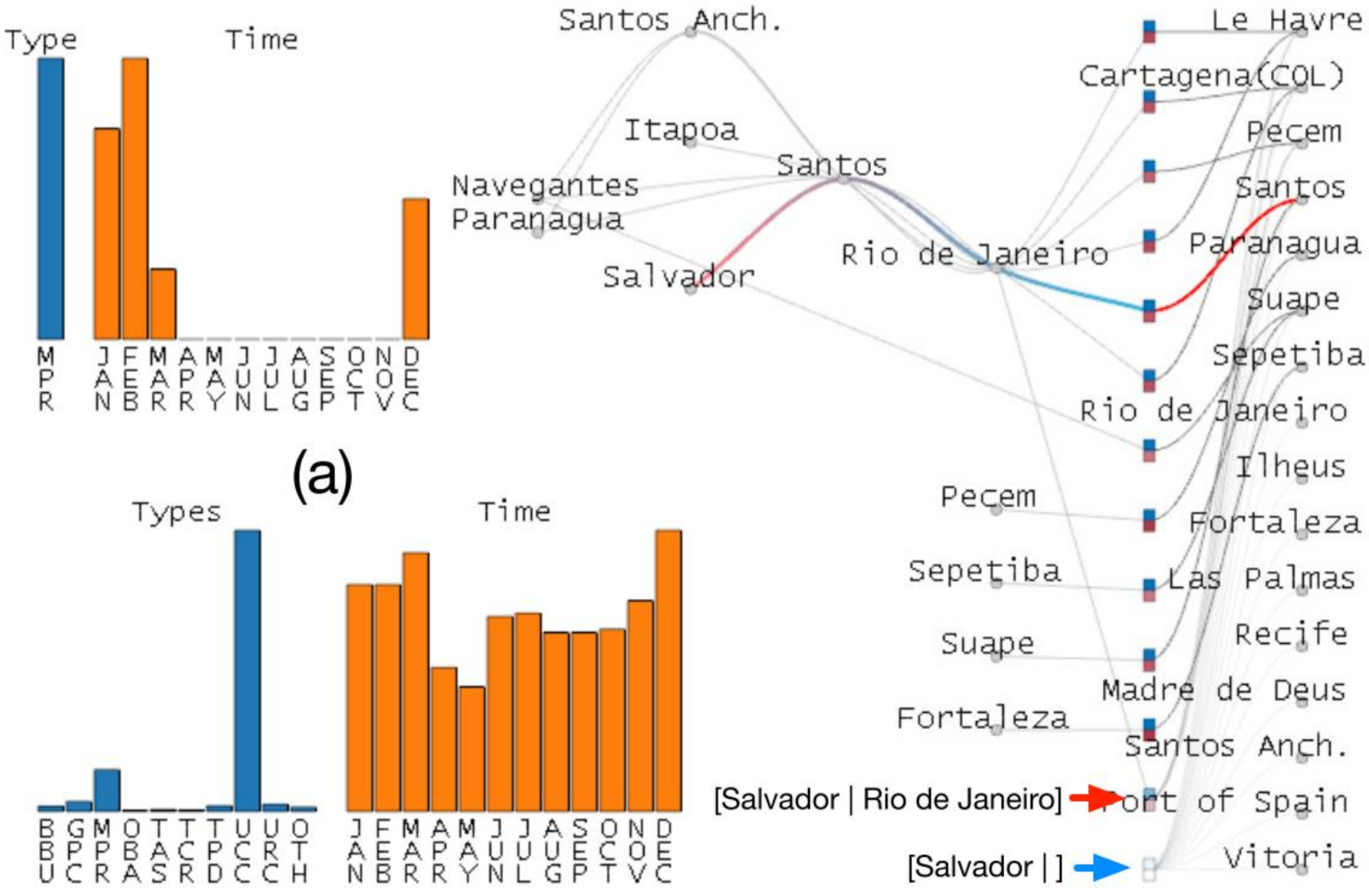


Explore geographically & rank by features

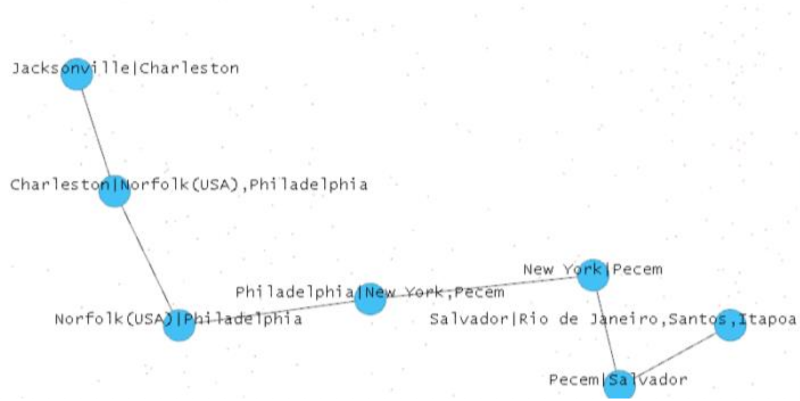


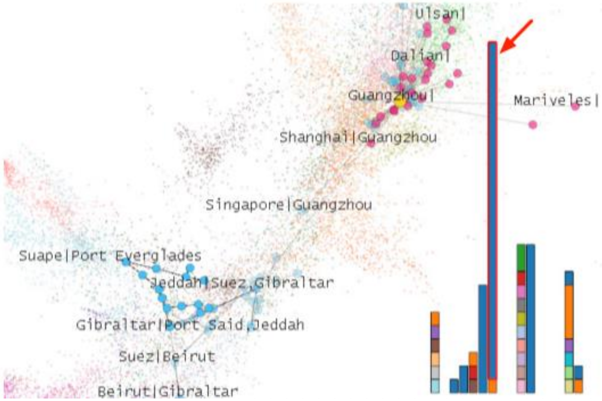
Ports	#HO Nodes
Suape	19
Vitoria	19
Salvador	13
Tubarao	6
Praia Mole	5
Portocel	2
Ponta do Ubu	2
Aratu	2
Recife	2
Madre de Deus	1
Cabedelo	1
Ilheus	1
Maceio	1
Jubarte Field	1

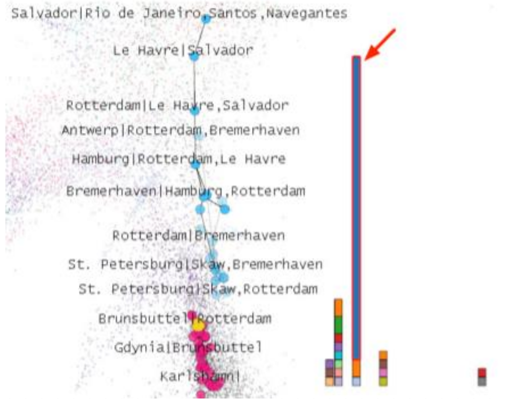
View dependencies & underlying metadata



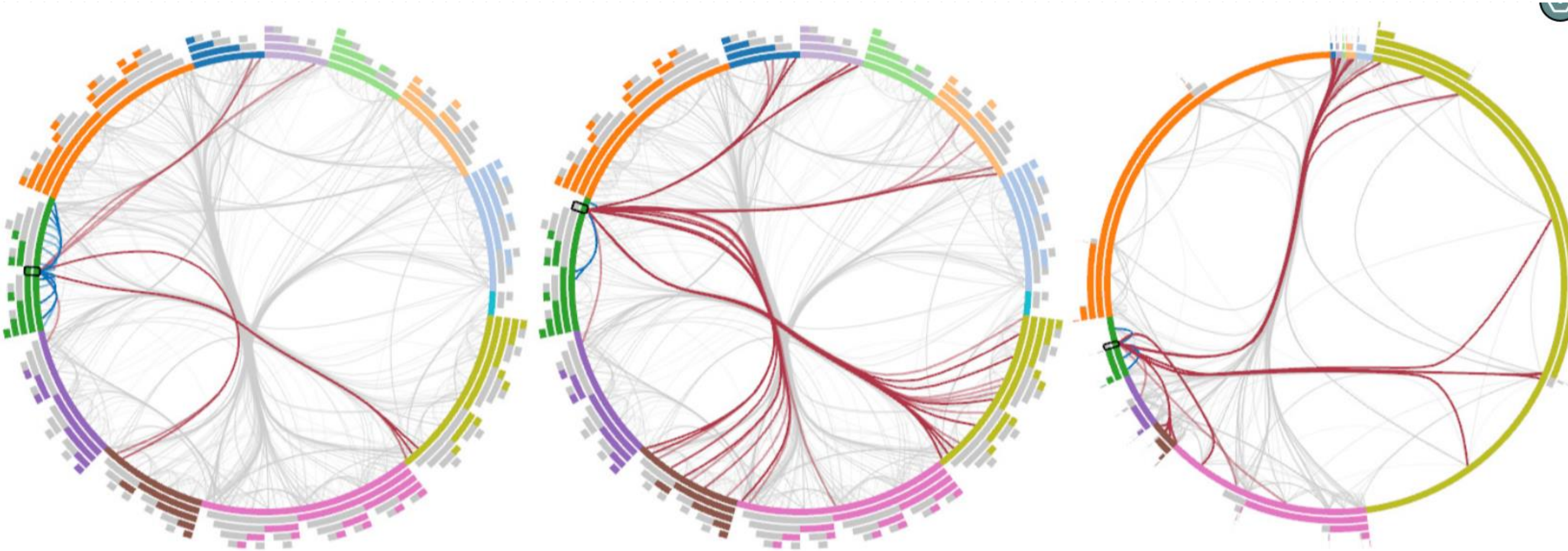
Track diffusion on the network







Aggregate at different granularities



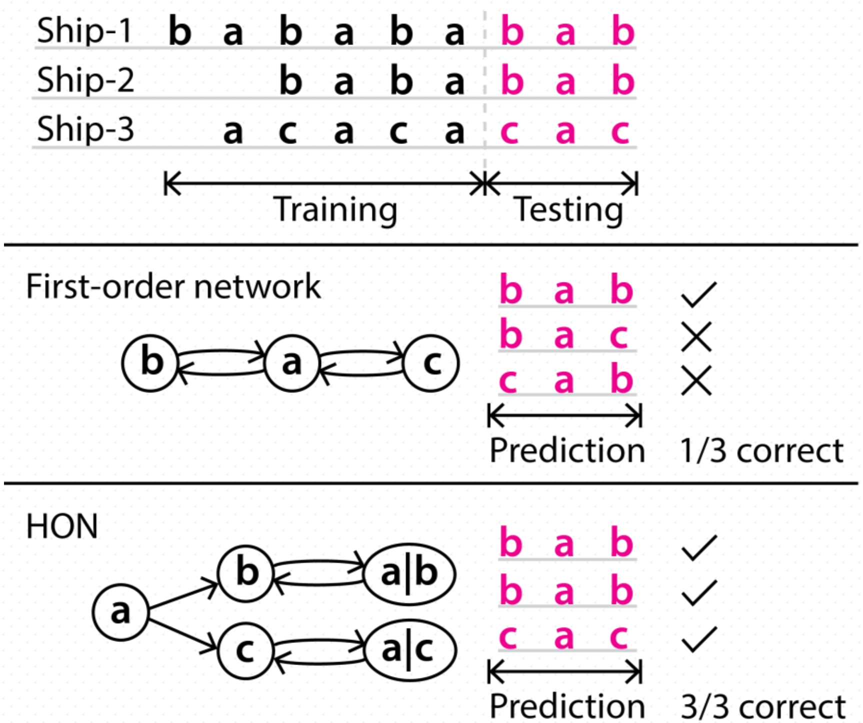
Clustering

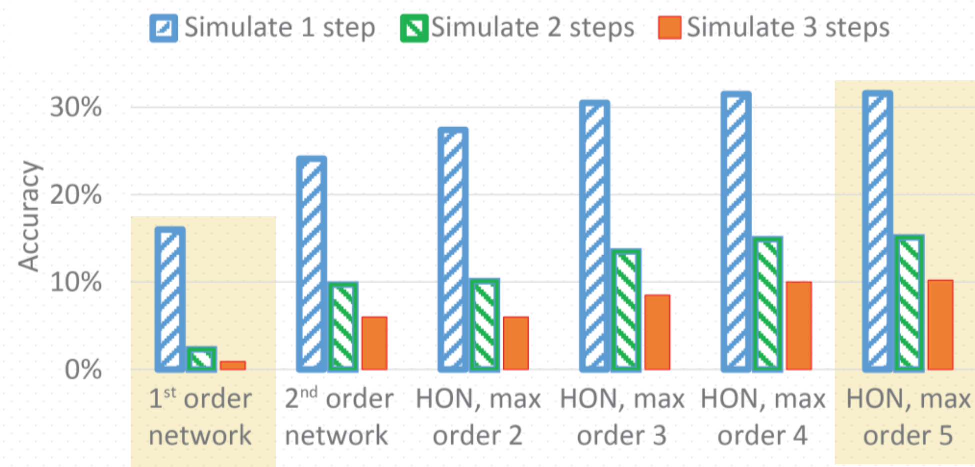
- *Walktrap*: "Random walks on a graph tend to get 'trapped' into densely connected parts corresponding to communities." (Pons & Latapy 2006)

Ranking

- *PageRank*: "The simplified version corresponds to the standing probability distribution of a random walk on the graph of the Web." (Page et al. 1999)

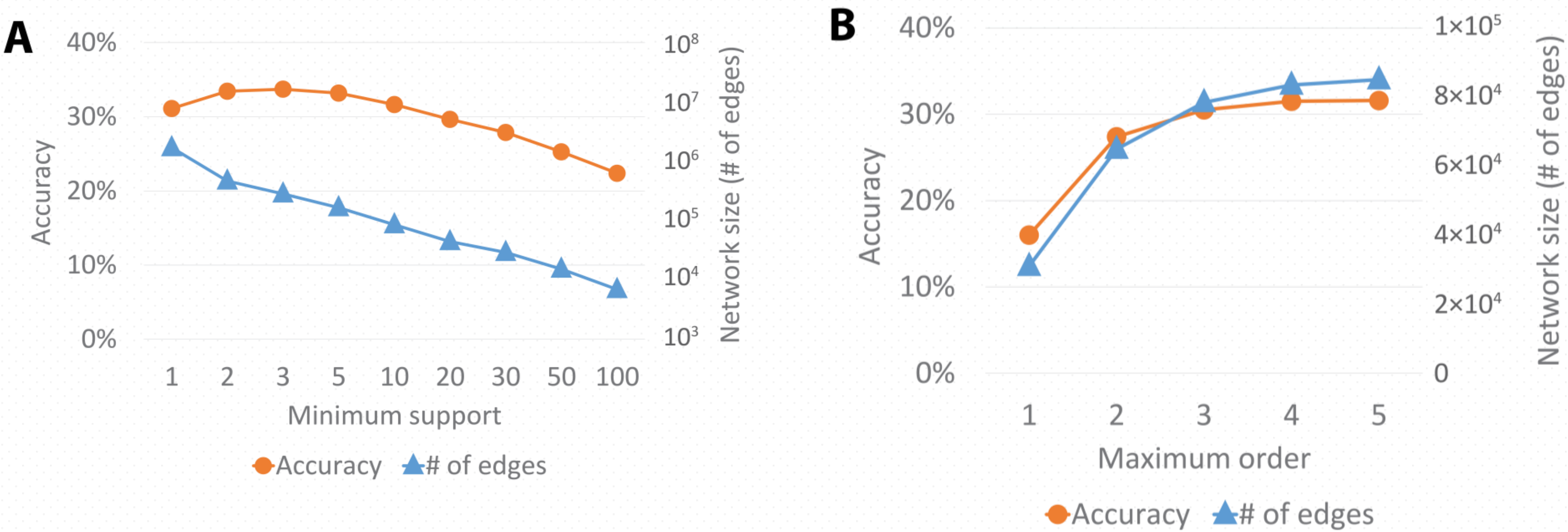
Influence on dynamics





Network representation	Number of edges	Number of nodes	Network density	Prob. of returning after two steps	Prob. of returning after three steps	Entropy rate (bits)	Clustering time (mins)	Ranking time (s)
Conventional first-order	31,028	2,675	4.3×10^{-3}	10.7%	1.5%	3.44	4	1.3
Fixed second-order	116,611	19,182	3.2×10^{-4}	42.8%	8.0%	1.45	73	7.7
HON, max order two	64,914	17,235	2.2×10^{-4}	41.7%	7.3%	1.46	45	4.8
HON, max order three	78,415	26,577	1.1×10^{-4}	45.9%	16.4%	0.90	63	6.2
HON, max order four	83,480	30,631	8.9×10^{-5}	48.9%	18.5%	0.68	67	7.0
HON, max order five	85,025	31,854	8.4×10^{-5}	49.3%	19.2%	0.63	68	7.6

Parameter sensitivity



Comparison with VOM

	HON	VOM	In HON but not in VOM	In VOM but not in HON
0th order	0	3,029	0	3029
1st order	31,028	31,028	0	0
2nd order	32,960	35,288	427	2,755
3rd order	15,642	21,536	550	6,444
4th order	4,632	8,973	302	4,643
5th order	763	2,084	23	1,344
Total	85,025	101,938	1,302	18,215

- **Global shipping data.** This data made available by Lloyd's Maritime Intelligence Unit (LMIU) contains ship movement information such as vessel_id, port_id, sail_date and arrival_date. Our experiments are based on a recent LMIU data set that spans one year from May 1st, 2012 to April 30th, 2013, totaling 3,415,577 individual voyages corresponding to 65,591 ships that move among 4,108 ports and regions globally. A minimum support of 10 is used to filter out noise in the data.

- **Clickstream data.** This data made available by a media company contains logs of users clicking through web pages that belong to 50 news web sites owned by the company. Fields of interest include user_ip, pagename and time. Our experiments are based on the clickstream records that span two months from December 4th, 2012 to February 3rd, 2013, totaling 3,047,697 page views made by 179,178 distinct IP addresses on 45,257 web pages. A minimum support of 5 is used to filter out noise in the data. Clickstreams that are likely to be created by crawlers (abnormally long clickstreams / clickstreams that frequently hit the error page) are omitted.

- **Retweet data.** This data *(50)* records retweet history on Weibo (a Chinese microblogging website), with information about who retweets whose messages at what time. The data was crawled in 2012 and there are 23,755,810 retweets recorded, involving 1,776,950 users.

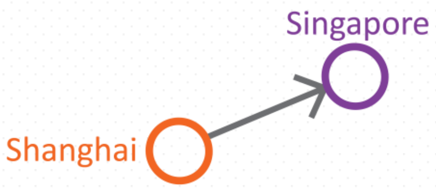
Synthetic data. We created a trajectory data set (data and code are available at https://github.com/xyjprc/hon) with known higher-order dependencies to verify the effectiveness of the rule extraction algorithm. In the context of shipping, we connect 100 ports as a 10×10 grid, then generate trajectories of 100,000 ships moving among these ports. Each ship moves 100 steps, yielding 10,000,000 movements in total. Normally each ship has equal probabilities of going up/down/left/right on the grid in each step (with wrapping, e.g., going down at the bottom row will end up in the top row); we use additional higher-order rules to control the generation of ship movements. For example, a second-order rule can be defined as whenever a ship comes from Shanghai to Singapore, instead of randomly picking a neighboring port of Singapore for the next step, the ship has 70% chance of going to Los Angeles and 30% chance of going to Seattle. We predefine 10 second-order rules like this, and similarly 10 third-order rules, 10 fourth-order rules, and no other higher-order rules, so that movements that have variable orders of dependencies are generated. To test the rule extraction algorithm, we set the maximum order as five to see if the algorithm will incorrectly extract false rules beyond the fourth order which we did not define; we set minimum support as five for patterns to be considered as rules.

Algorithm

How can we tell if this network representation more accurately captures the pattern in raw data?

A

- Convert all first-order rules into edges



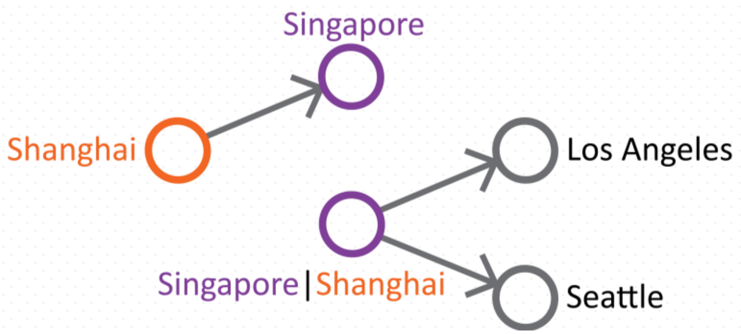
Raw data | Rule extraction | **Network wiring** | HON

A

- Convert all first-order rules into edges

B

- Convert higher-order rules
- Add higher-order nodes when necessary



A

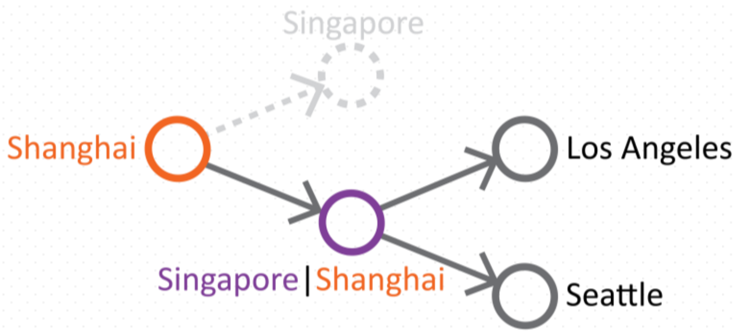
- Convert all first-order rules into edges

B

- Convert higher-order rules
- Add higher-order nodes when necessary

C

- Rewire edges
- The edge weights are preserved



A

- Convert all first-order rules into edges

B

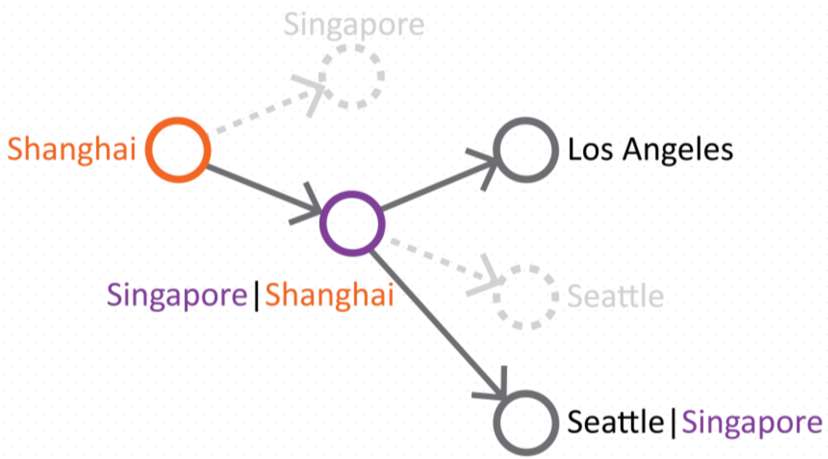
- Convert higher-order rules
- Add higher-order nodes when necessary

C

- Rewire edges
- The edge weights are preserved

D

- Rewire remaining edges



Effectiveness

Higher-order dependencies revealed by HON

Data	**# Records**	**Dependencies revealed**	**Similar observations**
Ship movement	3,415,577	Up to 5th order	N/A
Clickstream	3,047,697	Up to 3rd order	*"… appear to saturate at k = 3 for Yahoo… browsing behavior across websites is definitely not Markovian but can be captured reasonably well by a not-too-high order Markov chain."* --- Chierichetti et al. (2012)
Retweet	23,755,810	N/A	Assuming the second order has *"marginal consequences for disease spreading"* --- Rosvall et al. (2014)

Higher-order dependencies revealed by HON

Data	# Records	Dependencies revealed	Similar observations
Ship movement	3,415,577	Up to 5^{th} order	N/A
Clickstream	3,047,697	Up to 3^{rd} order	*"… appear to saturate at k = 3 for Yahoo… browsing behavior across websites is definitely not Markovian but can be captured reasonably well by a not-too-high order Markov chain."* --- Chierichetti et al. (2012)
Retweet	23,755,810	N/A	Assuming the second order has *"marginal consequences for disease spreading"* --- Rosvall et al. (2014)

Higher-order dependencies revealed by HON

Data	# Records	Dependencies revealed	Similar observations
Ship movement	3,415,577	Up to 5th order	N/A
Clickstream	3,047,697	Up to 3rd order	*"… appear to saturate at k = 3 for Yahoo… browsing behavior across websites is definitely not Markovian but can be captured reasonably well by a not-too-high order Markov chain."* --- Chierichetti et al. (2012)
Retweet	23,755,810	N/A	Assuming the second order has *"marginal consequences for disease spreading"* --- Rosvall et al. (2014)

Higher-order dependencies revealed by HON

Data	# Records	Dependencies revealed	Similar observations
Ship movement	3,415,577	Up to 5th order	N/A
Clickstream	3,047,697	Up to 3rd order	*"… appear to saturate at k = 3 for Yahoo… browsing behavior across websites is definitely not Markovian but can be captured reasonably well by a not-too-high order Markov chain."* --- Chierichetti et al. (2012)
Retweet	23,755,810	N/A	Assuming the second order has *"marginal consequences for disease spreading"* --- Rosvall et al. (2014)

Existing approaches

Ignore higher-orders

Inaccurate

Modify existing algorithms

Cannot generalize

Existing approaches

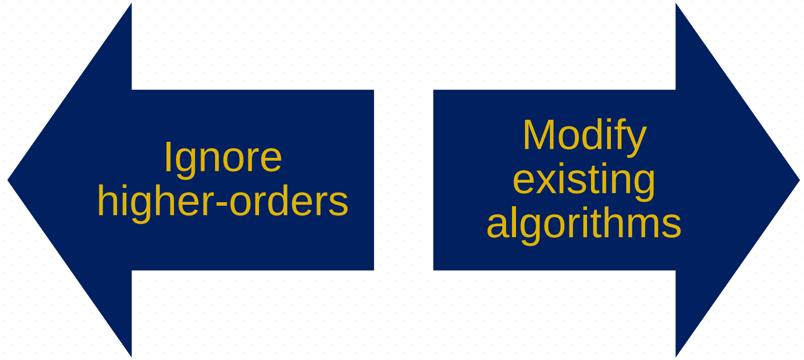
Ignore higher-orders

Inaccurate

Modify existing algorithms

Cannot generalize

Existing approaches



Higher-order network

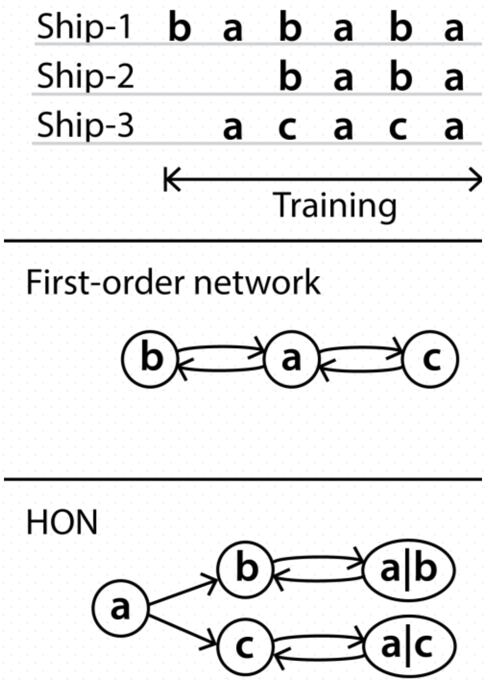
- ✓ Accurate representation
- ✓ Generalizes to existing algorithms

Influence on dynamics

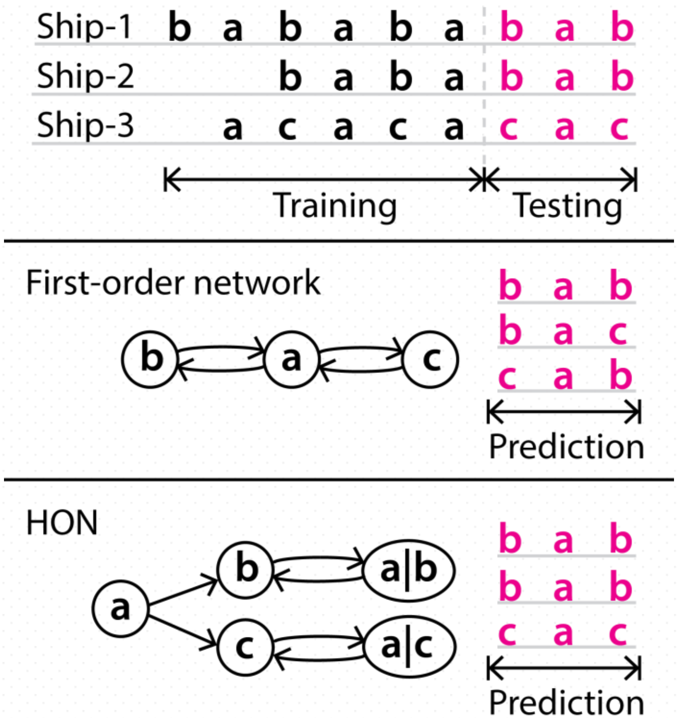
Ship-1	b	a	b	a	b	a
Ship-2			b	a	b	a
Ship-3		a	c	a	c	a

Training

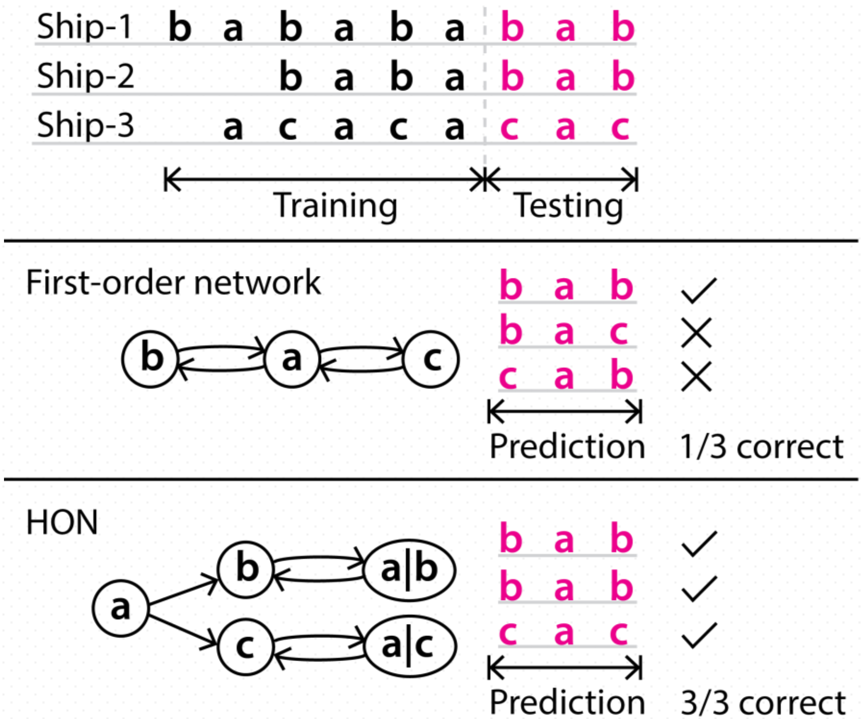
Influence on dynamics



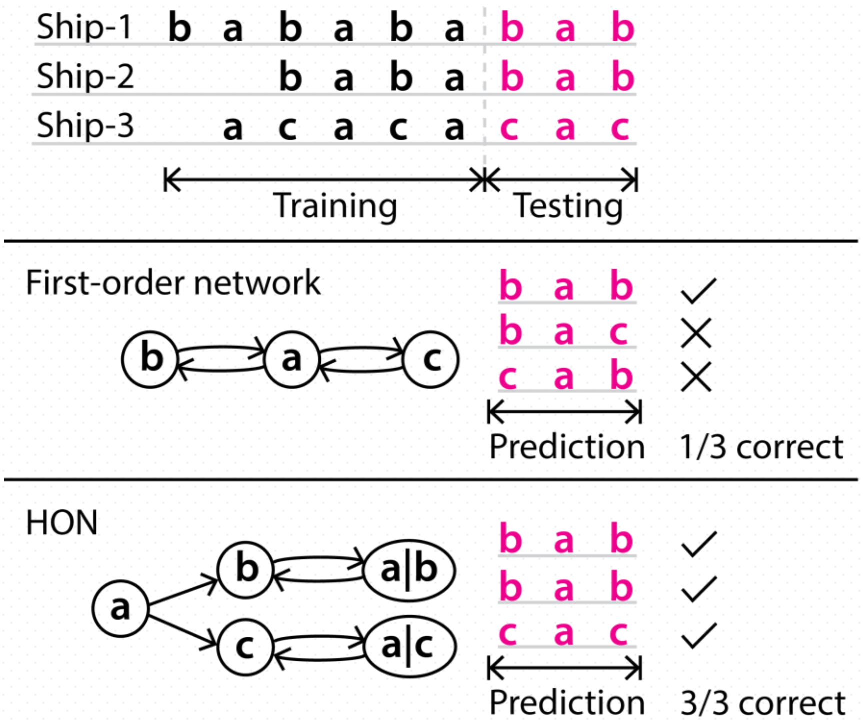
Influence on dynamics



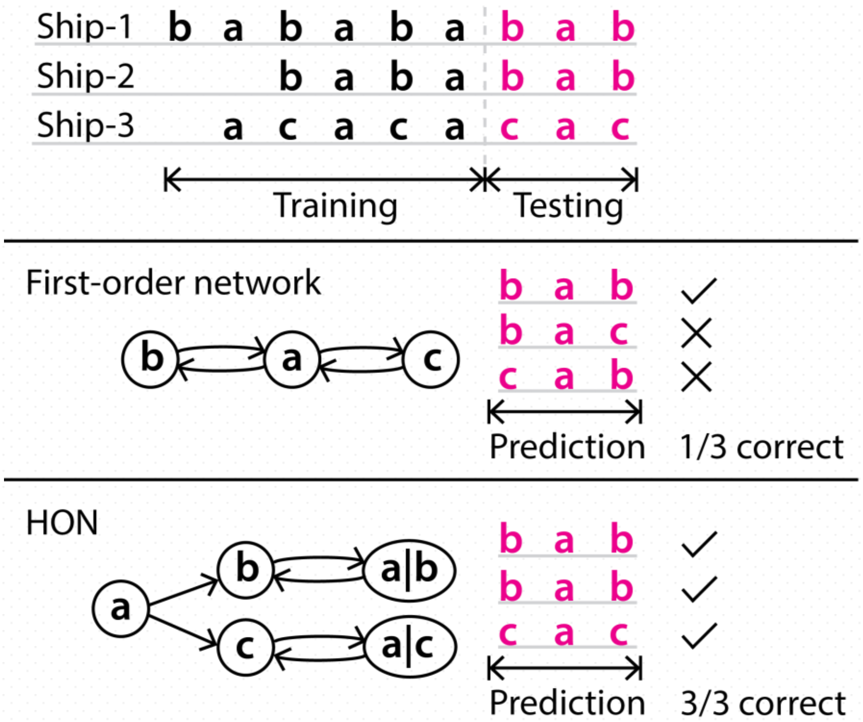
Influence on dynamics



Influence on dynamics



Influence on dynamics



Higher accuracy
in simulating real movements

Application: ranking

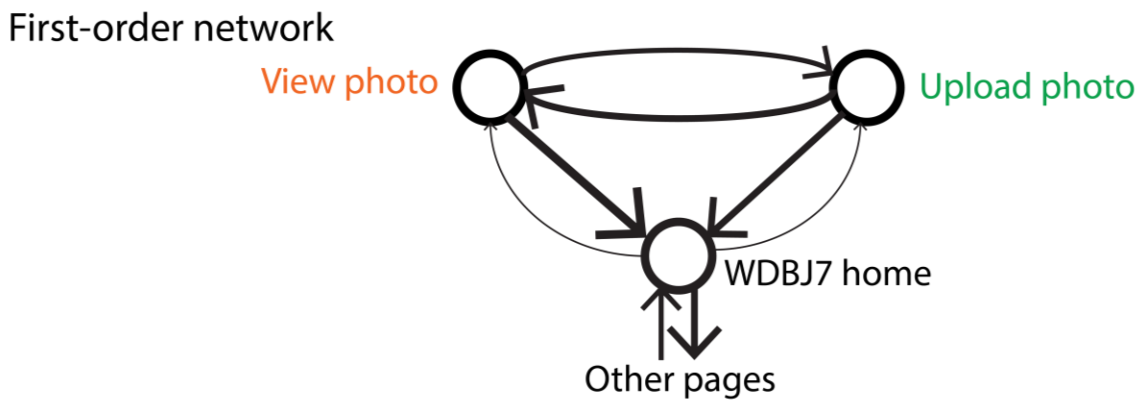
Web page access behaviors for server optimization and advertising

Ranking on clickstream network

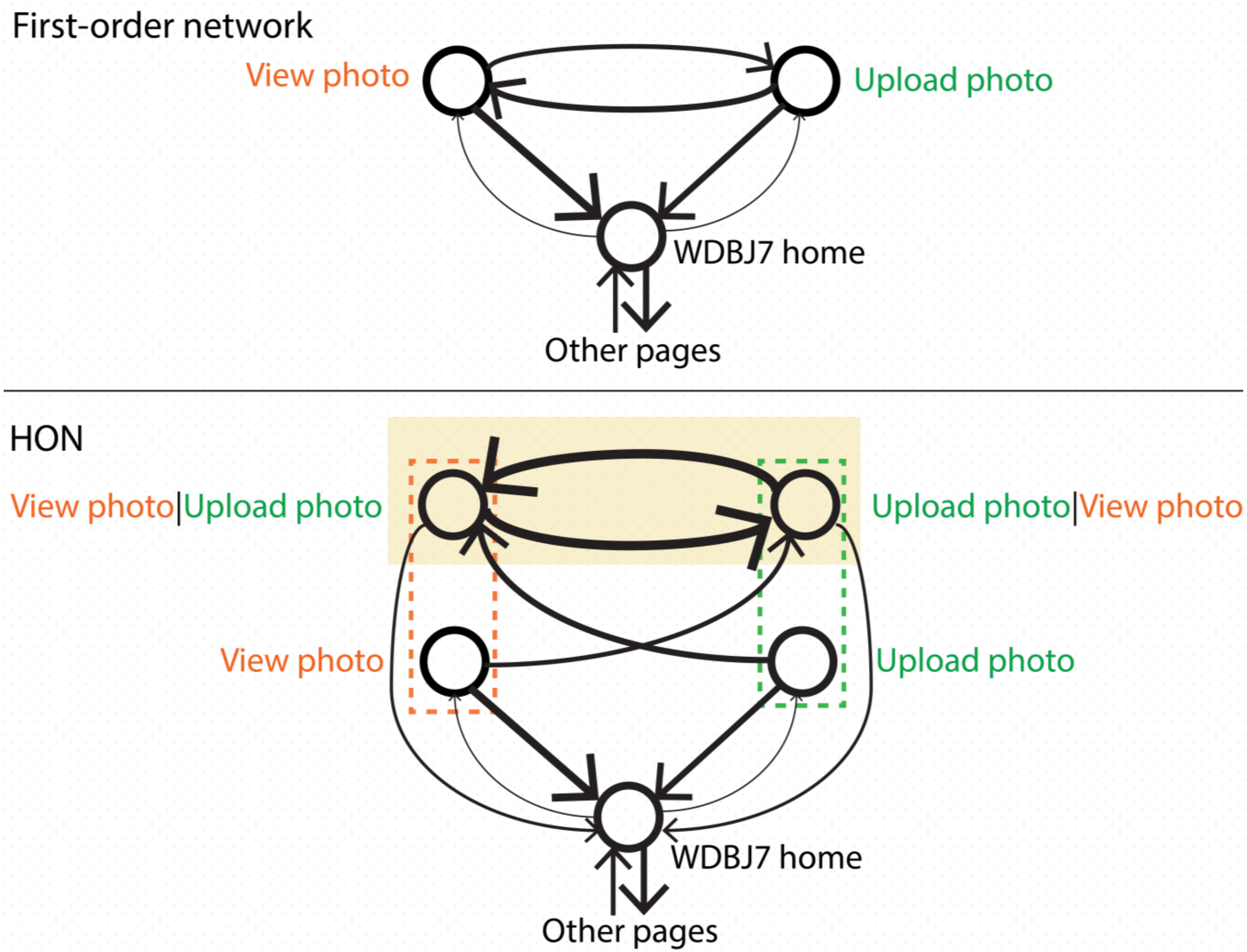
User 1 WDBJ7 home → View photo → WDBJ7 home → …
User 2 WDBJ7 home → View photo → Upload photo → …
User 3 View photo → Upload photo → View photo → …
User 4 WDBJ7 home → Upload photo → WDBJ7 home → …

… … … …

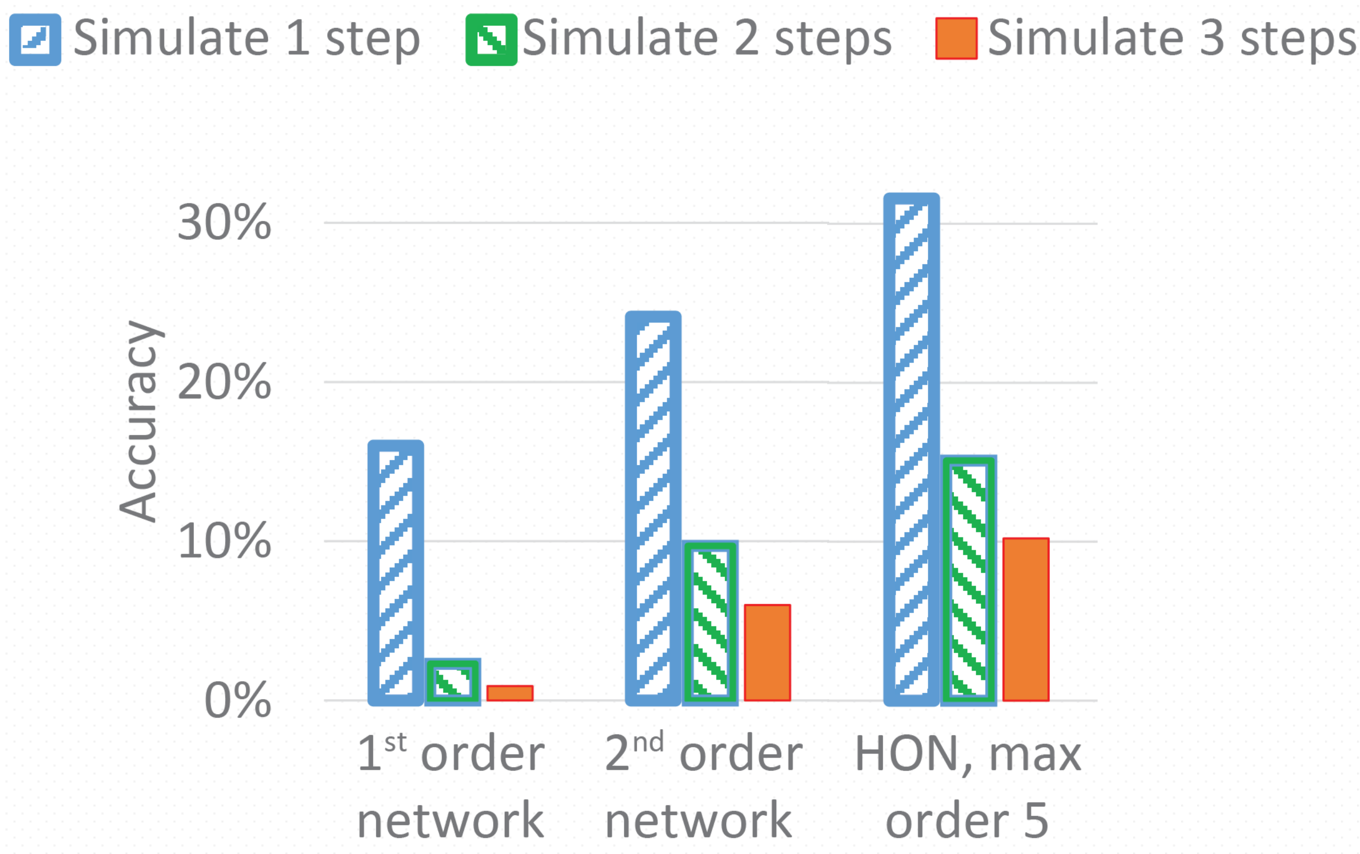
Ranking on clickstream network



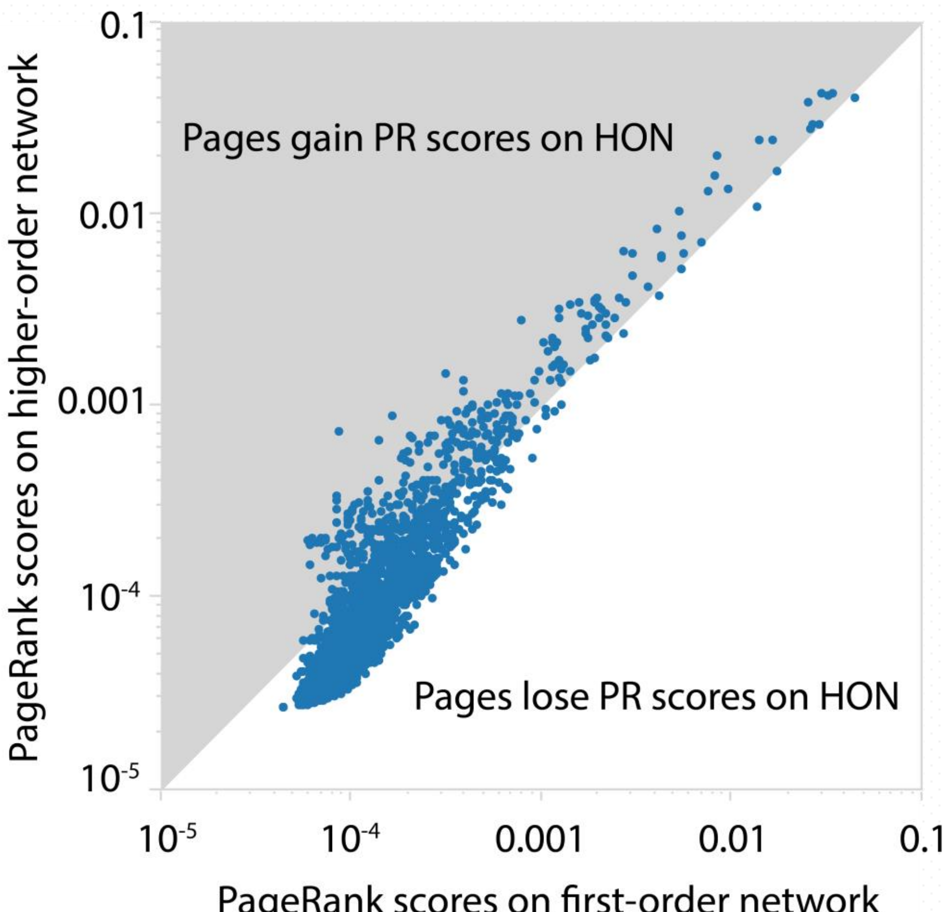
Ranking on clickstream network



Ranking on clickstream network



Ranking on clickstream network



- 26% pages show more than 10% changes in ranking
- More than 90% pages lose PageRank scores, while a few pages gain significant scores

No changes
to the ranking algorithm

Algorithm 3 HONOR rule extraction algorithm. Given the raw sequential data T, extracts arbitrarily high orders of dependencies, and output the dependency rules R. Optional parameters include $MaxOrder$, $MinSupport$, and $ThresholdMultiplier$

```
define global C as nested counter
define global D,R as nested dictionary
define global SourceToExtSource, StartingPoints as dictionary

function ExtractRules(T, [MaxOrder, MinSupport, ThresholdMultiplier = 1])
    global MaxOrder, MinSupport, Aggresiveness
    BuildFirstOrderObservations(T)
    BuildFirstOrderDistributions(T)
    GenerateAllRules(MaxOrder, T)

function BuildFirstOrderObservations(T)
    for t in T do
        for (Source, Target) in t do
            C[Source][Target] += 1
            IC.add(Source)

function BuildFirstOrderDistributions(T)
    for Source in C do
        for Target in C[Source] do
            if C[Source][Target] < MinSupport then
                C[Source][Target] = 0
            for Target in C[Source] do
                if thenC[Source][Target] > 0
                    D[Source][Target] = C[Source][Target]/(∑C[Source][*])

function GenerateAllRules(MaxOrder, T)
    for Source in D do
        AddToRules(Source)
        ExtendRule(Source, Source, 1, T)

function KLDThreshold(NewOrder,ExtSource)
    return ThresholdMultiplier × NewOrder/log2(1 + ∑C[ExtSource][*])
```

Algorithm 3 *(continued)*

```
function ExtendRule(Valid, Curr, order, T)
    if Order ≤ MaxOrder then
        AddToRules(Source)
    else
        Distr = D[Valid]
        if −log2(min(Distr[*].vals)) < KLDThreshold(order + 1), Curr then
            AddToRules(Valid)
        else
            NewOrder = order + 1
            Extended = ExtendSource(Curr)
            if Extended = ∅ then
                AddToRules(Valid)
            else
                for ExtSource in Extended do
                    ExtDistr = D[ExtSource]
                    divergence = KLD(ExtDistr, Distr)
                    if divergence > KLDThreshold(NewOrder, ExtSource) then
                        ExtendRule(ExtSource, ExtSource, NewOrder, T)
                    else
                        ExtendRule(Valid, ExtSource, NewOrder, T)

function AddToRules(Source):
    for order in [1..len(Source) + 1] do
        s = Source[0 : order]
        if not s in D or len(D[s]) == 0 then
            ExtendSource(s[1:])
        for t in C[s] do
            if C[s][t] > 0 then
                R[s][t] = C[s][t]

function ExtendSource(Curr)
    if Curr in SourceToExtSource then
        return SourceToExtSource[Curr]
    else
        ExtendObservation(Curr)
        if Curr in SourceToExtSource then
            return SourceToExtsource[Curr]
        else
            return ∅
```

Algorithm 3 *(continued)*

```
function ExtendObservation(Source)
    if length(Source) > 1 then
        if not Source[1 :] in ExtC or ExtC[Source] = ∅ then
            ExtendObservation(Source[1 :])
    order = length(Source)
    define ExtC as nested counter
    for Tindex, index in StartingPoints[Source] do
        if index − 1 ≤ 0 and index + order < length(T[Tindex]) then
            ExtSource = T[Tindex][index − 1 : index + order]
            ExtC[ExtSource][Target]+ = 1
            StartingPoints[ExtSource].add((Tindex, index − 1))
    if ExtC = ∅ then
        return
    for S in ExtC do
        for t in ExtC[s] do
            if ExtC[s][t] < MinSupport then
                ExtC[s][t] = 0
            C[s][t]+ = ExtC[s][t]
        CsSupport = ∑ ExtC[s][*]
        for t in ExtC[s] do
            if ExtC[s][t] > 0 then
                D[s][t] = ExtC[s][t]/CsSupport
                SourceToExtSource[s[1 :]].add(s)

function BuildSourceToExtSource(order)
    for source in D do
        if len(source) = order then
            if len(source) > 1 then
                NewOrder = len(source)
                for startingin[1..len(source)] do
                    curr = source[starting :]
                    if not curr in SourceToExtSource then
                        SourceToExtSource[curr] = ∅
                    if not NewOrder in SourceToExtSource[curr] then
                        SourceToExtSource[curr][NewOrder] = {}
                    SourceToExtSource[curr][NewOrder].add(source)
```

Rank	Risk of single-step direct invasion	Risk of multi-step indirect invasion
1	Murmansk, RUS	Tromso, NOR
2	Tromso, NOR	Reykjavik, ISL
3	Dudinka, RUS	Murmansk, RUS
4	Glomfjord, NOR	Hammerfest, NOR
5	Hammerfest, NOR	Nuuk, GRL
6	Kirkenes, NOR	Kirkenes, NOR
7	Grundartangi, ISL	Harstad, NOR
8	Harstad, NOR	Dutch Harbor, USA
9	Hammerfall, NOR	Grundartangi, ISL
10	Bodo, NOR	Aasiaat, GRL

The method also adapts to

Transportation

Flow of information

Evolution of diseases

Sequential data

1

Ship-001: …, Tokyo, Singapore, Los Angeles, …
Ship-002: …, Shanghai, Singapore, Seattle, …
⋮

Sequential data

Ship-001: …, Tokyo, Singapore, Los Angeles, …
Ship-002: …, Shanghai, Singapore, Seattle, …
⋮

①

Count subsequences of various orders

Singapore → Los Angeles: 60
Singapore → Seattle: 65
Shanghai → Singapore → Los Angeles: 30
Shanghai → Singapore → Seattle: 5
⋮

Raw data → **Rule extraction** → Network wiring → HON

Sequential data

① ↓

Ship-001: …, Tokyo, Singapore, Los Angeles, …
Ship-002: …, Shanghai, Singapore, Seattle, …
⋮

Count subsequences of various orders

② ↓

Singapore → Los Angeles: 60
Singapore → Seattle: 65
Shanghai → Singapore → Los Angeles: 30
Shanghai → Singapore → Seattle: 5
⋮

Probability distributions: 1st order

$$\underbrace{P(X_{t+1} \mid X_t = Singapore)}_{\text{Source node}} = \underbrace{\begin{cases} Los\ Angeles\text{: 4.8\%} \\ Seattle\text{: 5.2\%} \\ \vdots \end{cases}}_{\text{Target nodes}}$$

Sequential data

Ship-001: …, Tokyo, Singapore, Los Angeles, …
Ship-002: …, Shanghai, Singapore, Seattle, …
⋮

① ↓

Count subsequences of various orders

Singapore → Los Angeles: 60
Singapore → Seattle: 65
Shanghai → Singapore → Los Angeles: 30
Shanghai → Singapore → Seattle: 5
⋮

② ↓

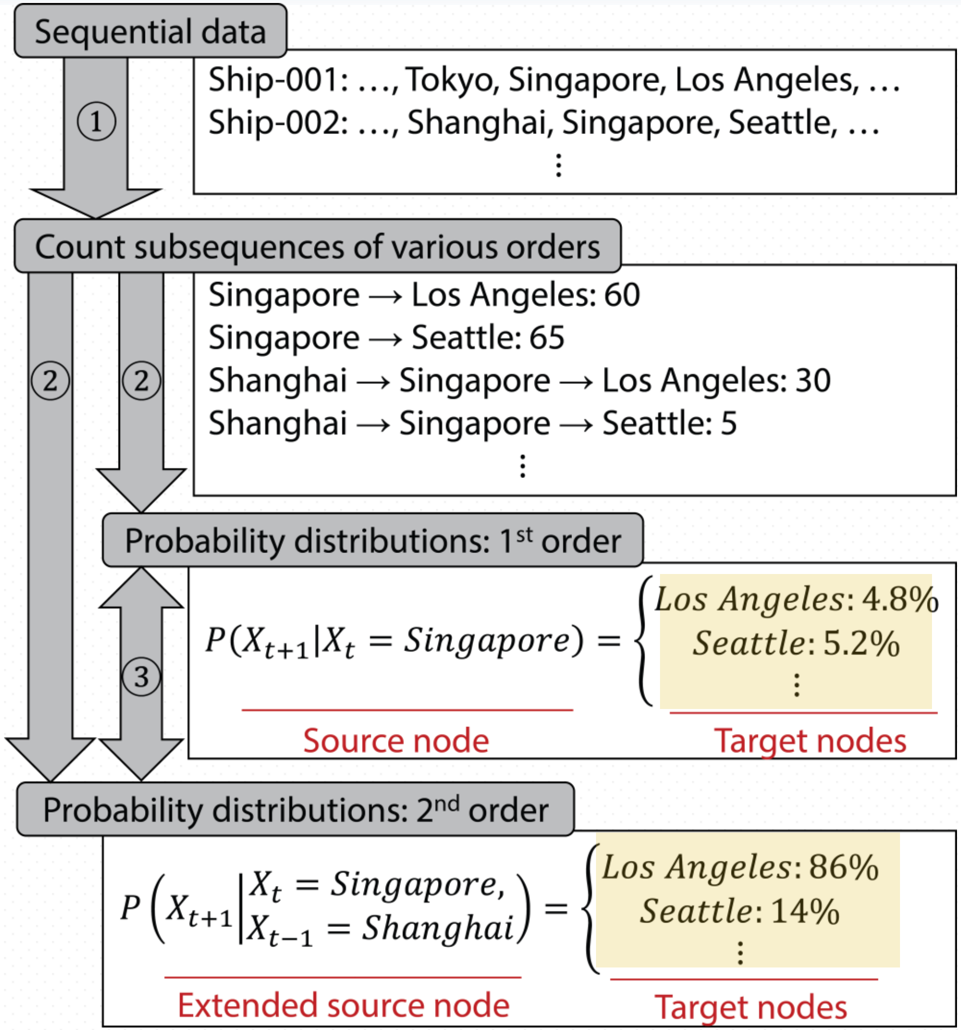
Probability distributions: 1st order

$$\underbrace{P(X_{t+1} | X_t = Singapore)}_{\text{Source node}} = \underbrace{\begin{cases} Los\ Angeles: 4.8\% \\ Seattle: 5.2\% \\ \vdots \end{cases}}_{\text{Target nodes}}$$

② ↓

Probability distributions: 2nd order

$$\underbrace{P\left(X_{t+1} \middle| \begin{matrix} X_t = Singapore, \\ X_{t-1} = Shanghai \end{matrix}\right)}_{\text{Extended source node}} = \underbrace{\begin{cases} Los\ Angeles: 86\% \\ Seattle: 14\% \\ \vdots \end{cases}}_{\text{Target nodes}}$$


Sequential data
Ship-001: …, Tokyo, Singapore, Los Angeles, …
Ship-002: …, Shanghai, Singapore, Seattle, …
⋮
1
Count subsequences of various orders
Singapore → Los Angeles: 60
Singapore → Seattle: 65
Shanghai → Singapore → Los Angeles: 30
Shanghai → Singapore → Seattle: 5
⋮
2
2
Probability distributions: 1st order
3
$P(X_{t+1}|X_t = Singapore) = \begin{cases} Los\ Angeles: 4.8\% \\ Seattle: 5.2\% \\ \vdots \end{cases}$
Source node
Target nodes
Probability distributions: 2nd order
$P\left(X_{t+1}\middle| \begin{matrix} X_t = Singapore, \\ X_{t-1} = Shanghai \end{matrix}\right) = \begin{cases} Los\ Angeles: 86\% \\ Seattle: 14\% \\ \vdots \end{cases}$
Extended source node
Target nodes