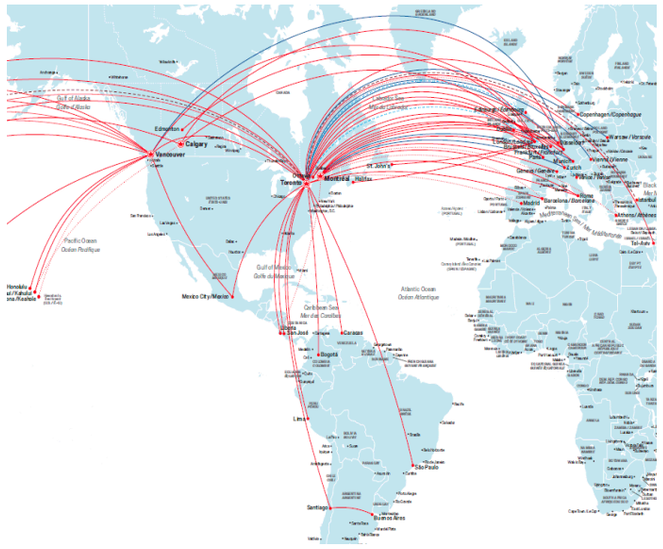
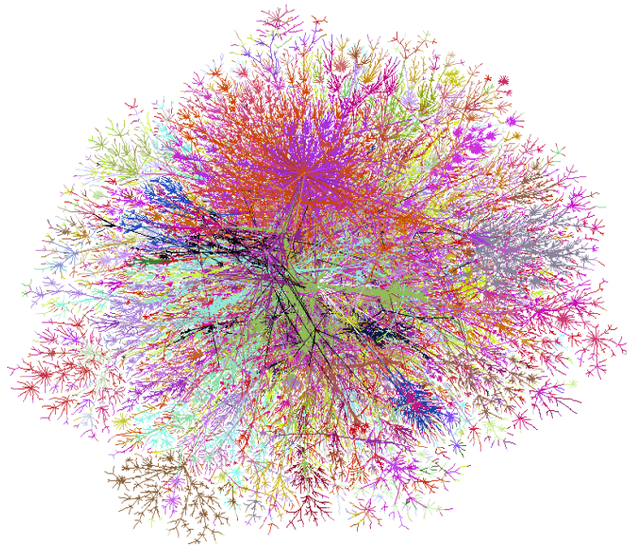
Scalable Hashing-Based Network Discovery

Tara Safavi, Chandra Sripada, Danai Koutra
University of Michigan, Ann Arbor

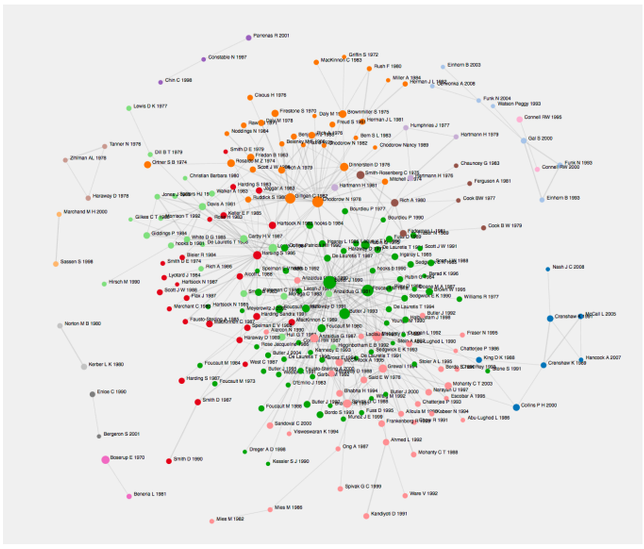
Networks are everywhere….



Airport connections

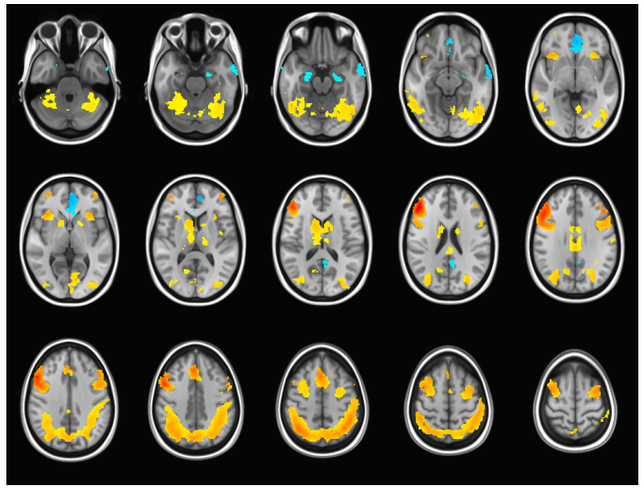


Internet routing

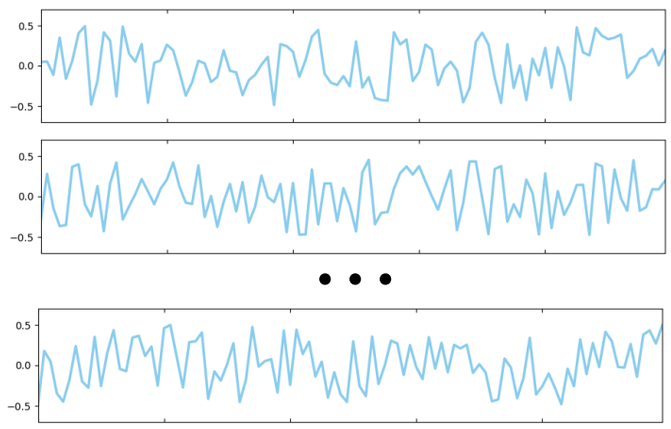


Paper citations

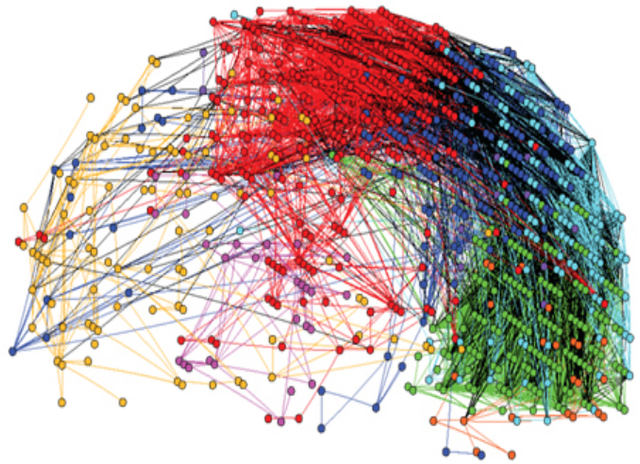
…but are not always *directly* observed



1. fMRI scans



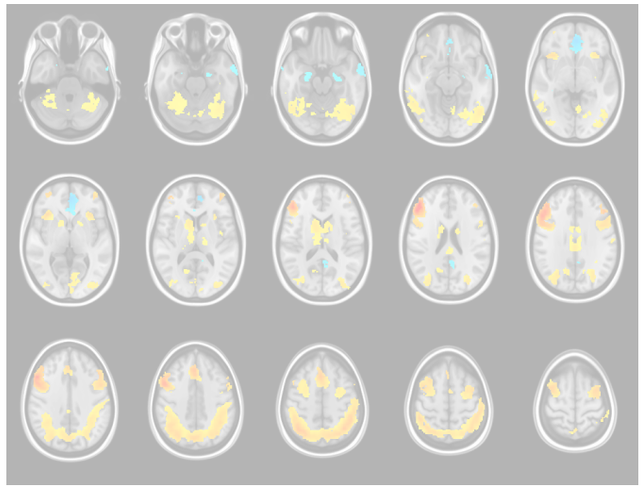
2. Time series



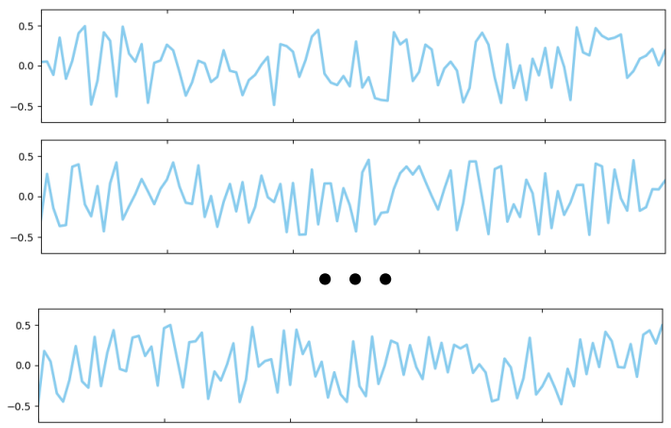
3. Brain network

See *Network Structure Inference, A Survey,* Brugere, Gallagher, Berger-Wolf

How to build this network?



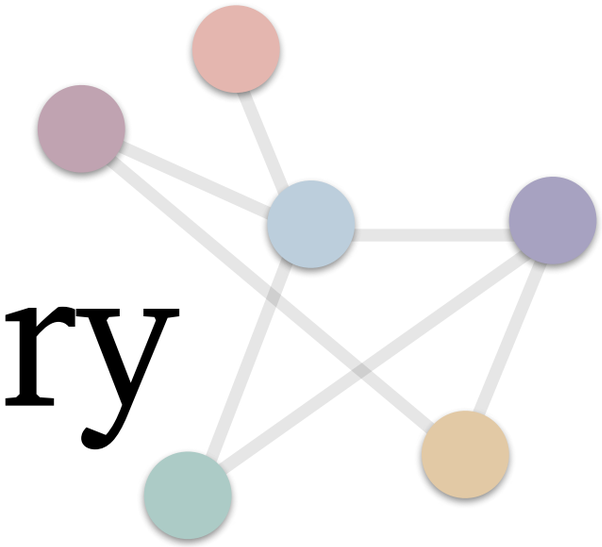
1. fMRI scans



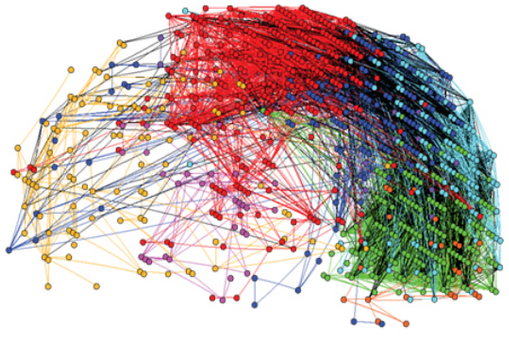
2. Time series

3. Brain network

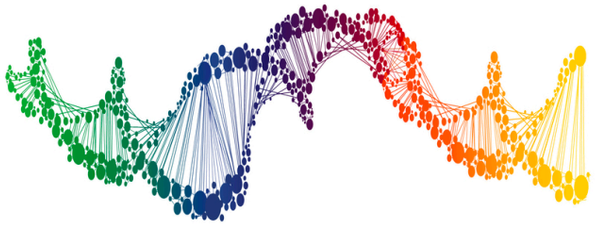
Network discovery



Reconstructing networks from **indirect, possibly noisy measurements** with **unobserved interactions**



Brain scans

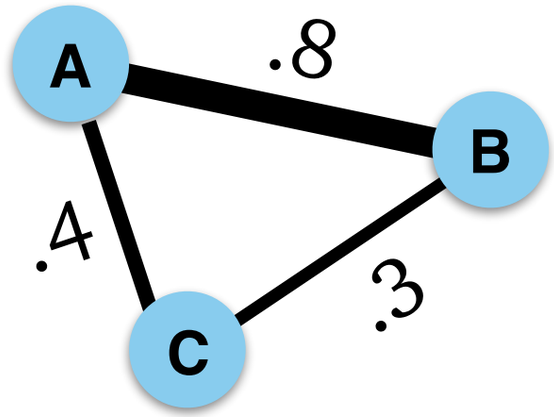


Gene sequences

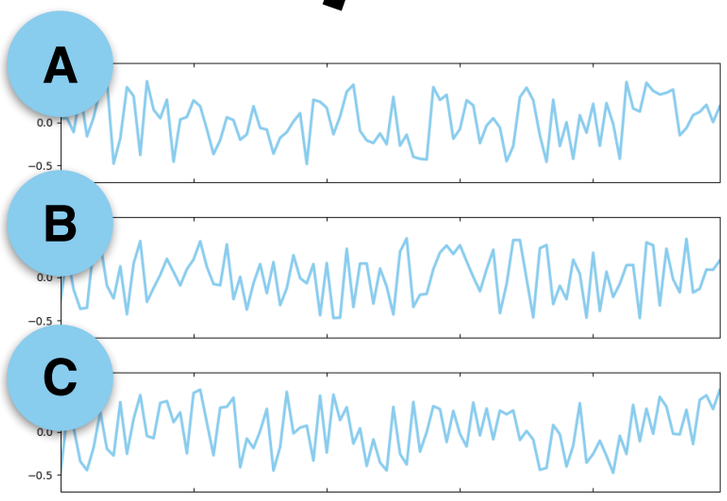
Stock patterns

Traditional method

2. Fully-connected weighted network



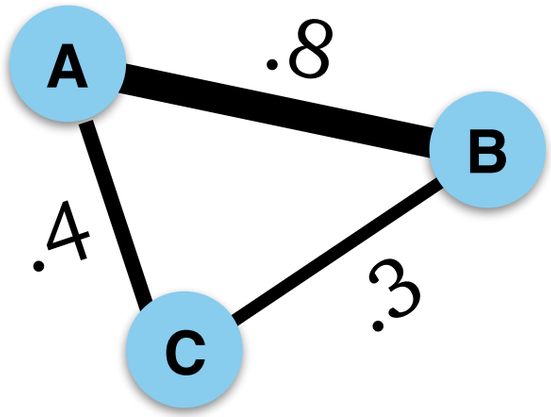
All-pairs correlation



1. *N* time series

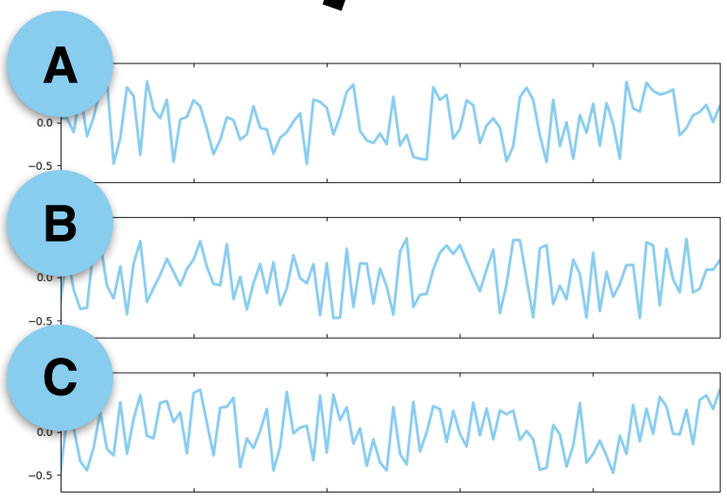
Traditional method

2. Fully-connected weighted network

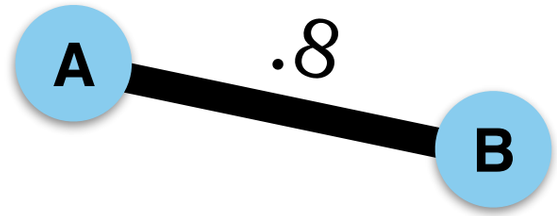


All-pairs correlation

Drop edges below threshold θ



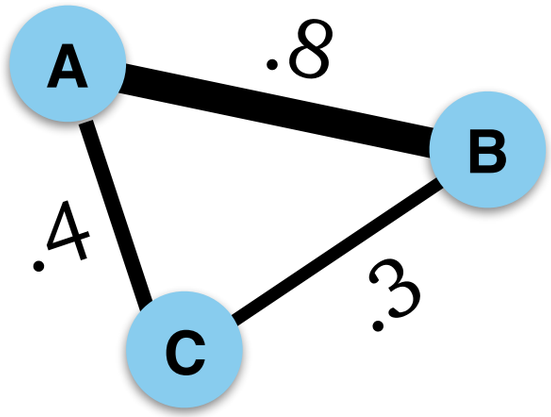
1. *N* time series



3. Sparse graph

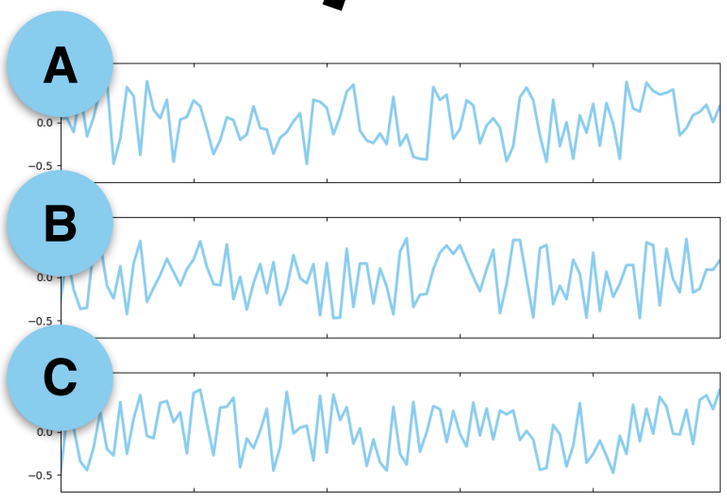
Traditional method

2. Fully-connected weighted network



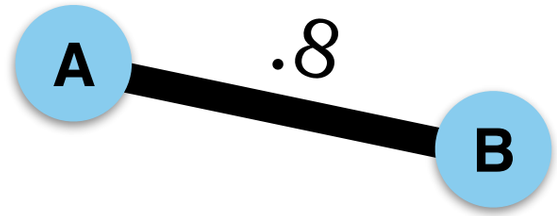
All-pairs correlation

Drop edges below threshold θ



1. *N* time series

Widely used in many domains, **interpretable**, but…



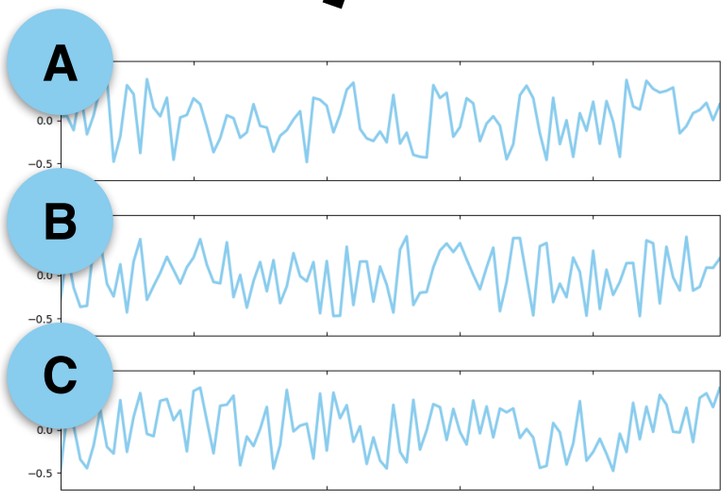
3. Sparse graph

Traditional method

2. Fully-connected weighted network

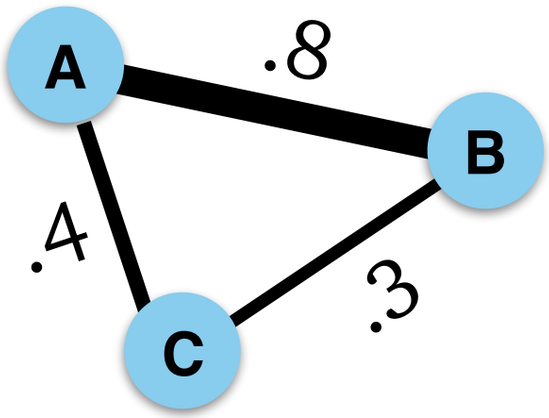
All-pairs correlation

Drop edges below threshold θ

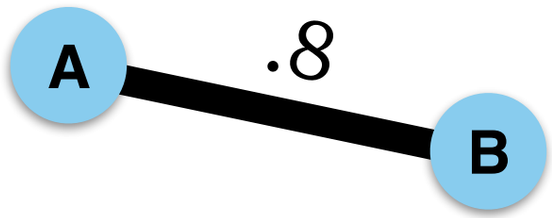


1. *N* time series

O(N^2) comparisons



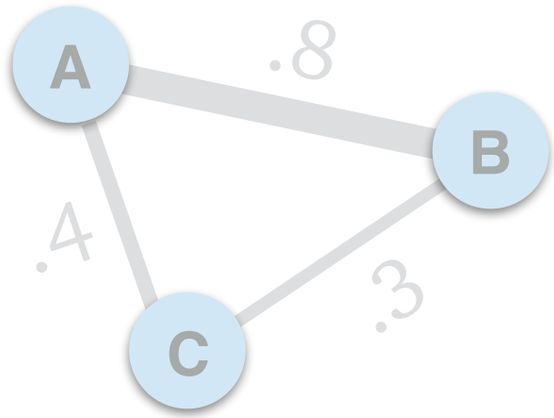
How to set?



3. Sparse graph

New hashing-based

2. Fully-connected weighted network

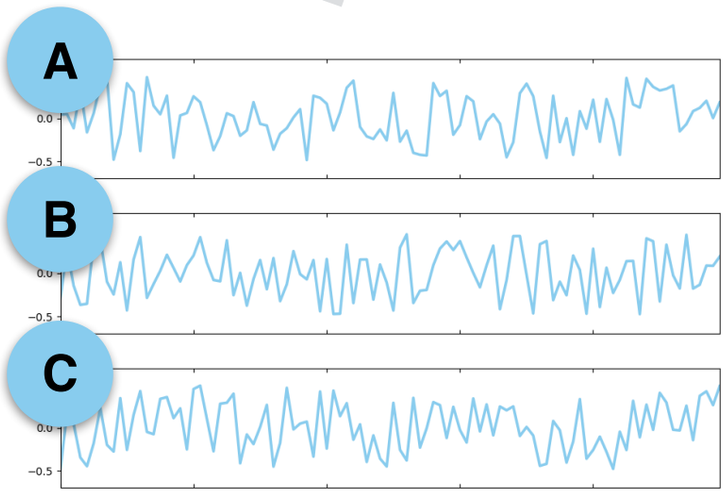


All-pairs correlation

$O(N^2)$ comparisons

Drop edges below threshold θ

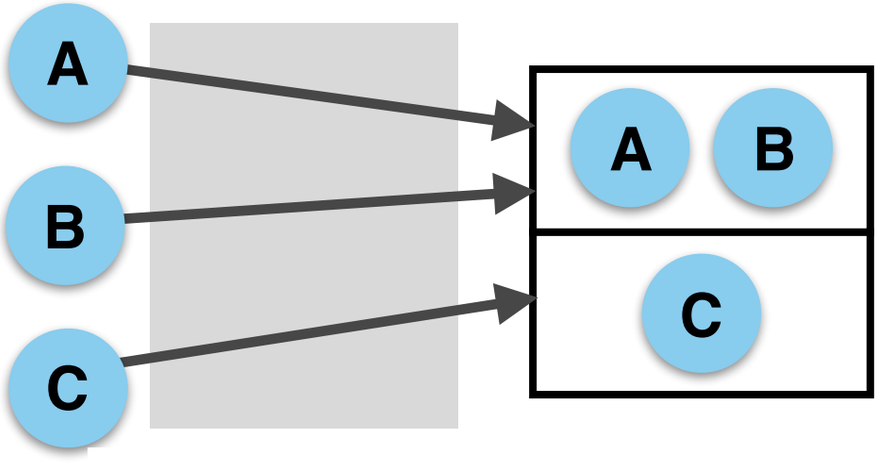
How to set?

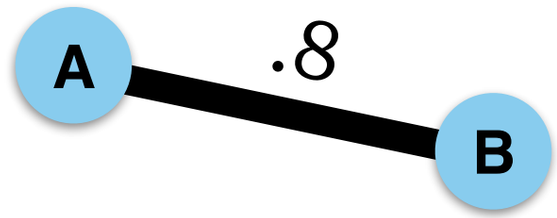


1. *N* time series

Binarize

2. Hash series



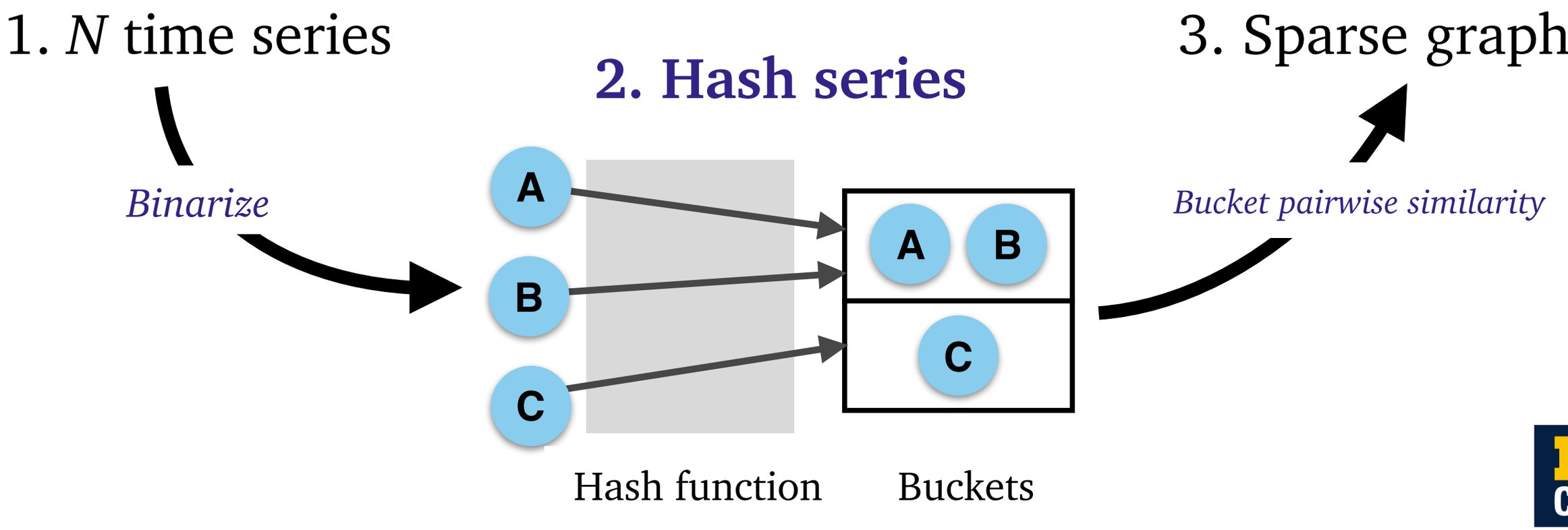


3. Sparse graph

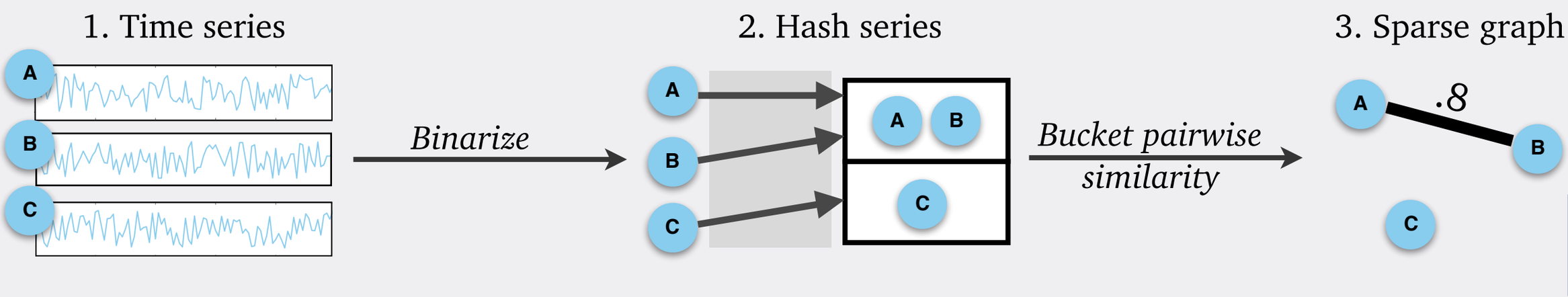
Bucket pairwise similarity

Contributions

- Network discovery via new **locality-sensitive hashing** (LSH) family
  - Quickly find similar pairs — circumvent wasteful extra computation
- Novel **similarity measure** on sequences for LSH
  - Quantify *time-consecutive similarity*
  - Complementary distance measure is a **metric**: suitable for LSH!
- **Evaluation** on real data in the neuroscience domain

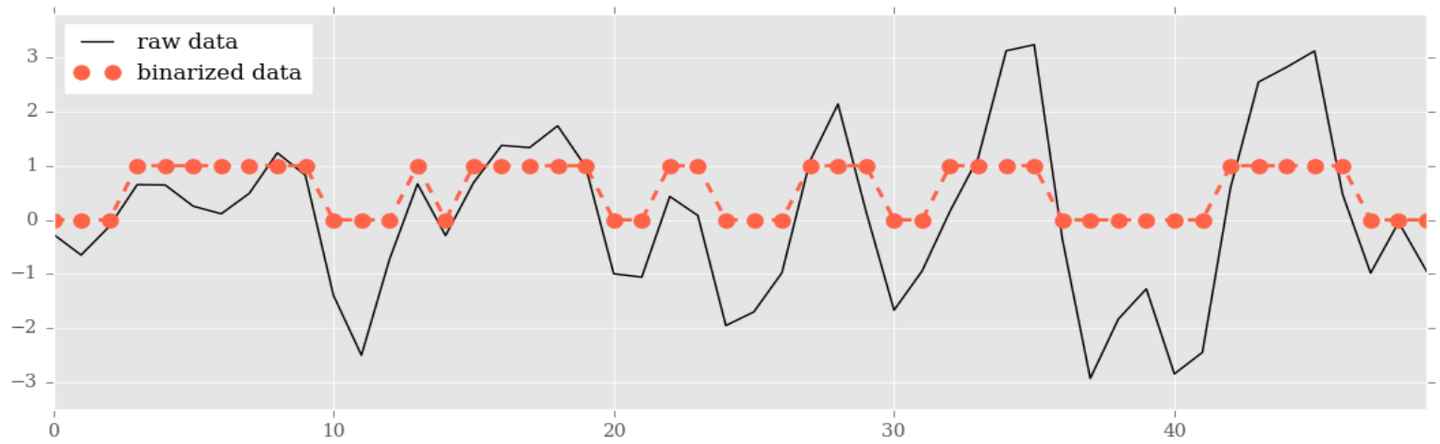


Method


1. Time series
2. Hash series
3. Sparse graph
A
B
C
Binarize
A
B
C
A
B
C
Bucket pairwise
similarity
A
.8
B
C

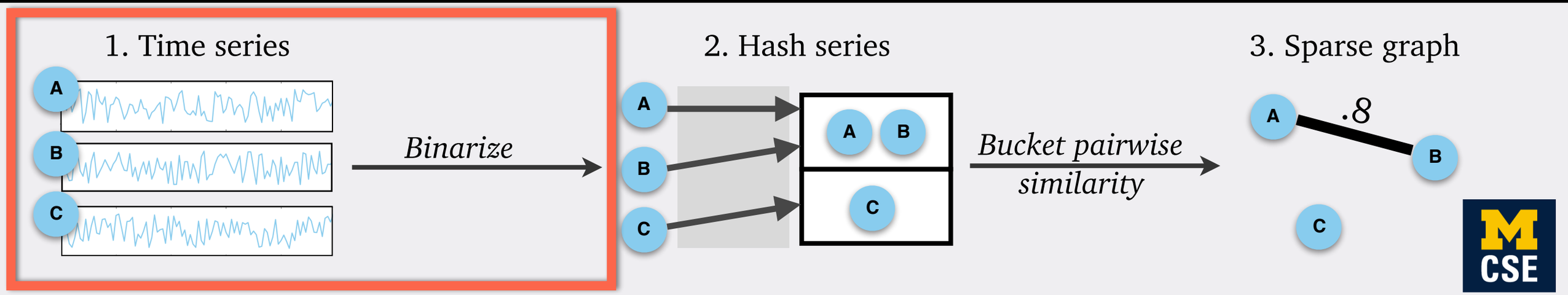
Approximate time series representation

- Binarize w.r.t series mean ("clipped" representation[1])
- Why?
  - Capture approximate *fluctuation trend*
  - Preprocess for hashing



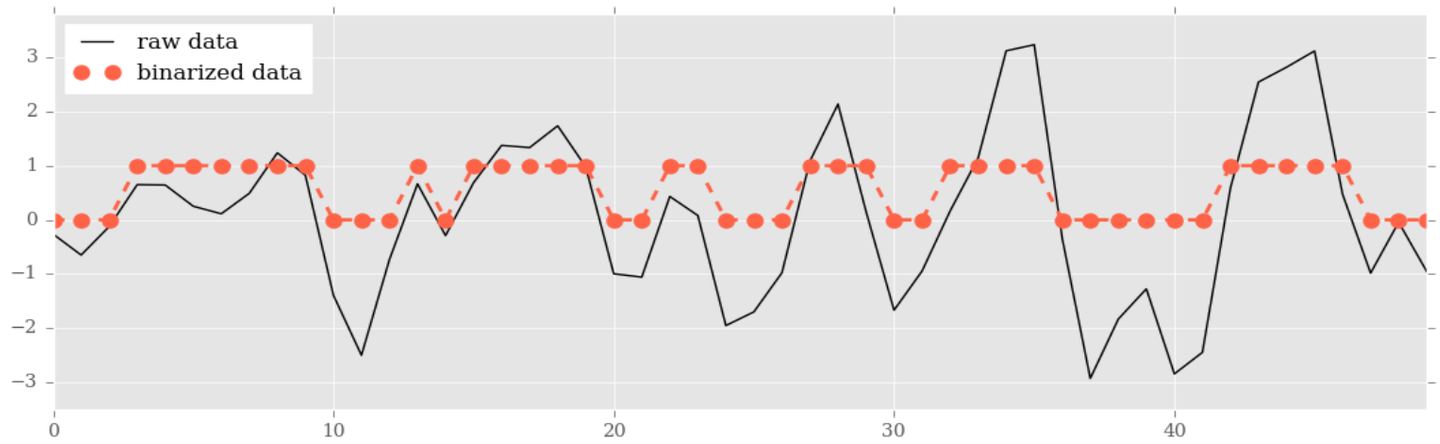
[1]Ratanamahatana et al, 2005

Pipeline



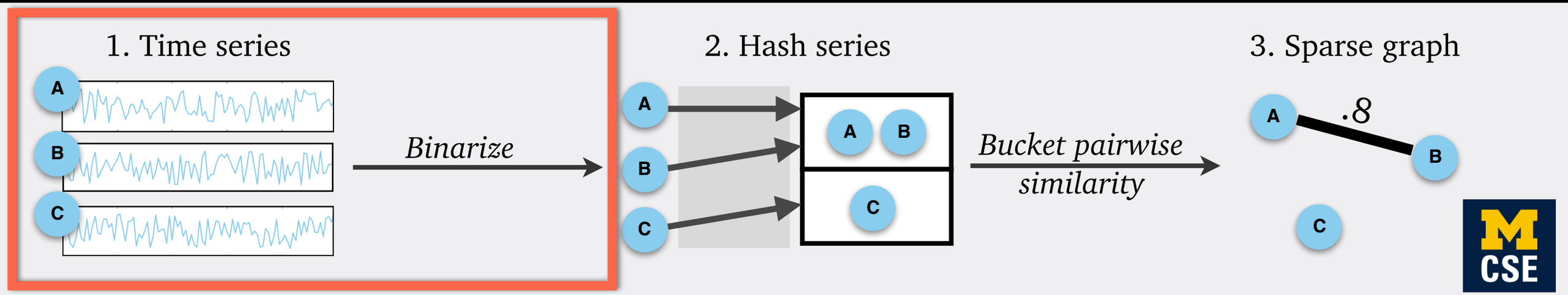
Approximate time series representation

- Binarize w.r.t series mean ("clipped" representation[1])
- Why?
  - Capture approximate *fluctuation trend*
  - Preprocess for hashing
- But — binary sequences only have two possible values
  - Emphasize **consecutive similarity** between sequences over pointwise comparison



Pipeline

[1]Ratanamahatana et al, 2005



ABC: Approximate Binary Correlation

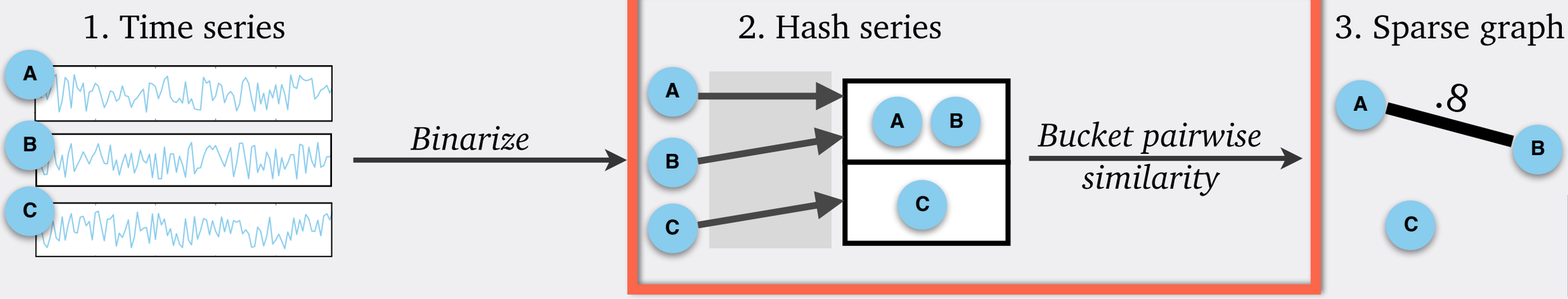
- Capture variable-length **consecutive runs** between binary sequences

x: 1 1 **0** 1 0 0 **0**

y: 1 1 **1** 1 0 0 **1**

$(1+\alpha)^0 + (1+\alpha)^1$ + $(1+\alpha)^0 + (1+\alpha)^1 + (1+\alpha)^2$

Pipeline



ABC: Approximate Binary Correlation

- Capture variable-length **consecutive runs** between binary sequences
- Similarity score s a sum of p geometric series, each of length k_i
  - Common ratio $(1+\alpha)$: $0 < \alpha \ll 1$ is a consecutiveness weighting factor

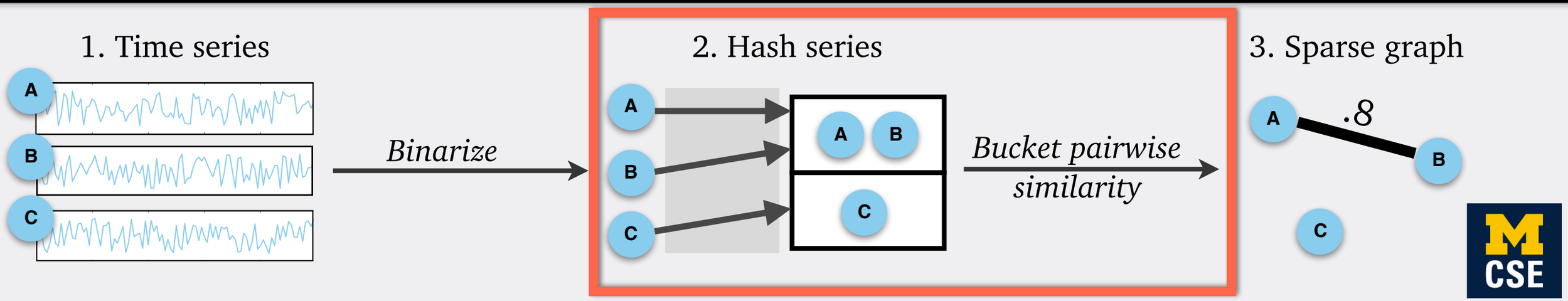
$$s(\boldsymbol{x},\boldsymbol{y}) = \sum_{i=1}^{p}\sum_{b=0}^{k_i}(1+a)^b = \frac{\sum_{i=1}^{p}(1+\alpha)^{k_i} - p}{\alpha}$$

x: 1 1 **0** 1 0 0 **0**

y: 1 1 **1** 1 0 0 **1**

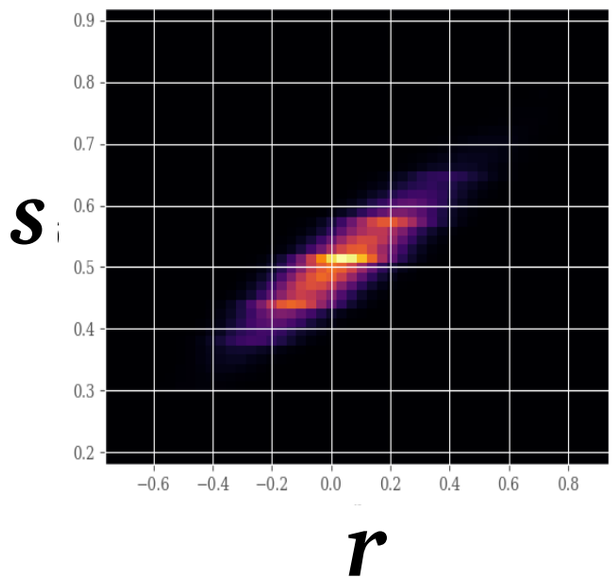
$(1+\alpha)^0 + (1+\alpha)^1$ + $(1+\alpha)^0 + (1+\alpha)^1 + (1+\alpha)^2$

Pipeline

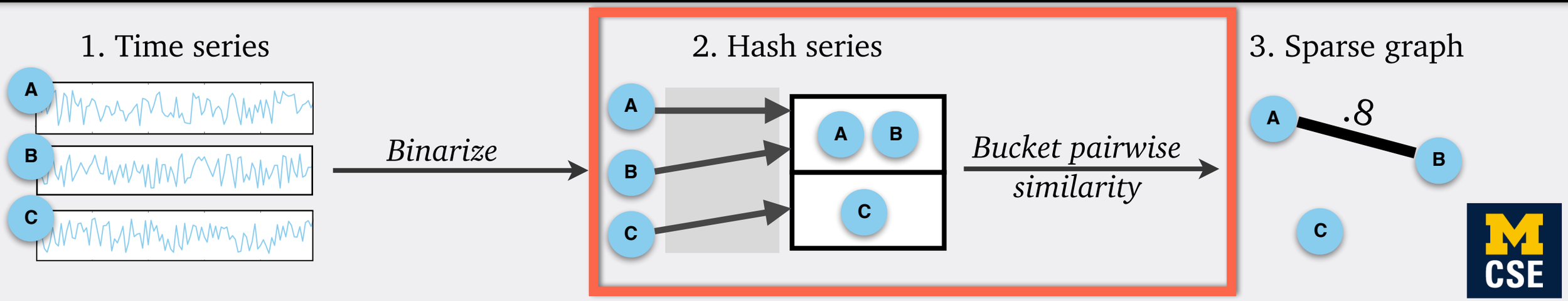


ABC: Approximate Binary Correlation

- Empirically, a good estimator of correlation coefficient ***r***
  - Similarity scores ***s*** correlate well with ***r***

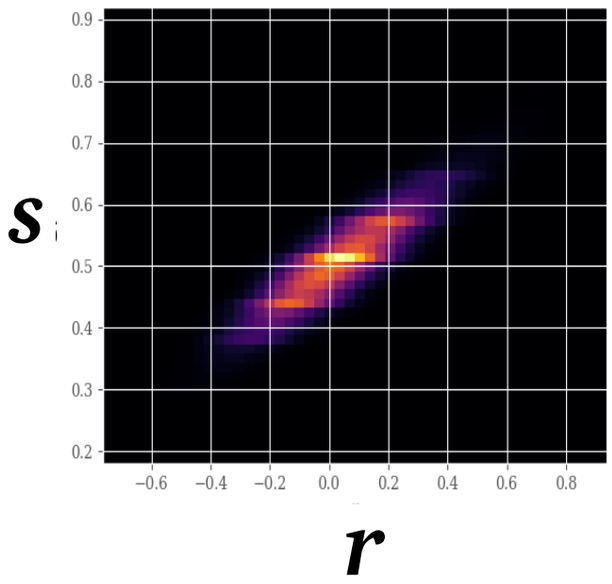


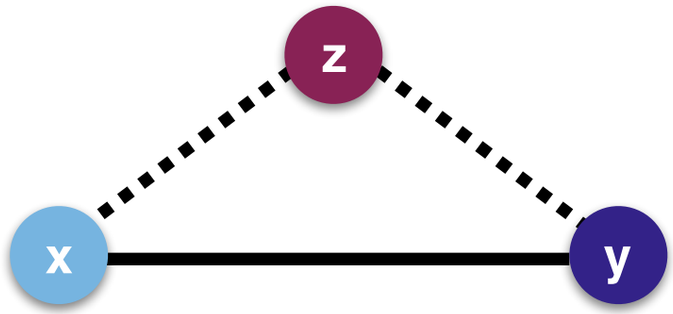
Pipeline



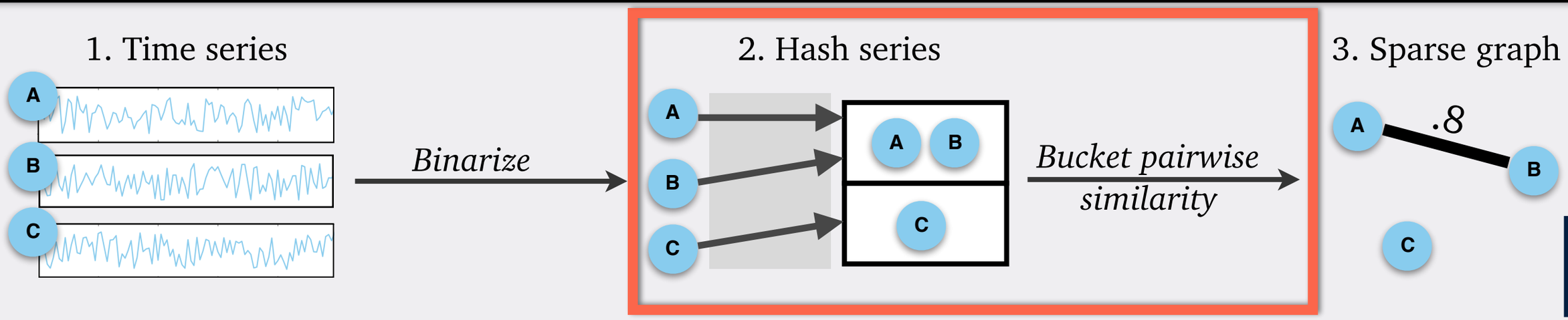
ABC: Approximate Binary Correlation

- Empirically, a good estimator of correlation coefficient ***r***
  - Similarity scores ***s*** correlate well with ***r***
- Added benefit of time-aware hashing
  - LSH requires a *metric*: satisfies triangle inequality
  - We can show ABC distance **is a metric (critical)**



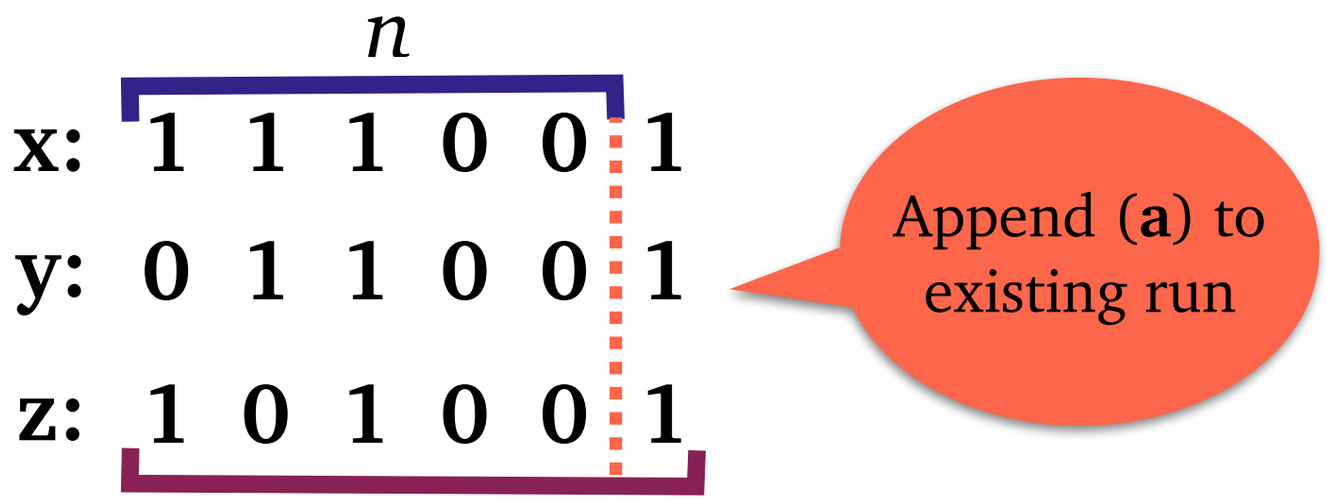


Pipeline



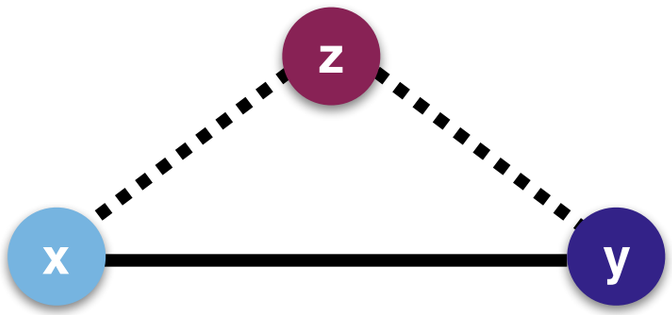
ABC distance triangle inequality: sketch of proof

- Induction on n, sequence length
- Induction step: identify feasible cases between sequence pairs
  - Disagreement (**d**)
  - New run (**n**)
  - Append to existing run (**a**)
- Compute all deltas
  - Triangle inequality holds!

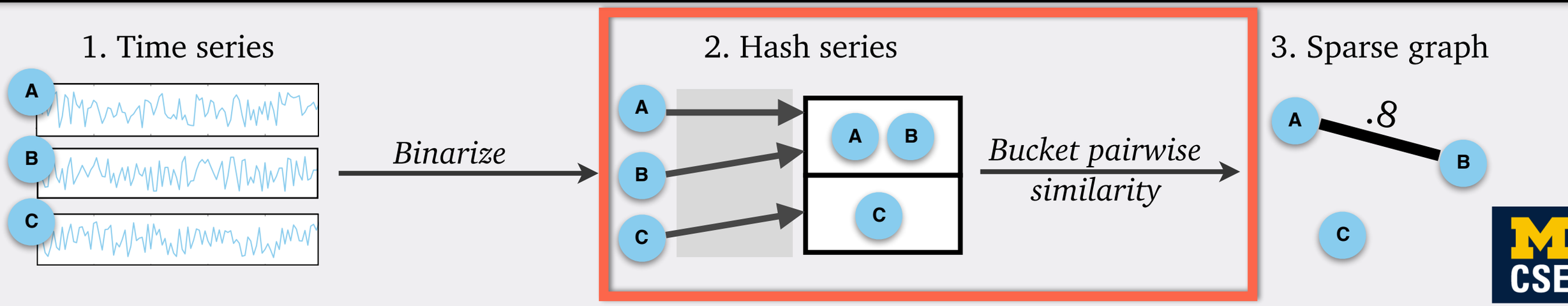


$$n + 1$$

Theorem 1 (ABC is a metric). *The ABC distance in Eq. (III.2) is a metric. It satisfies all the axioms of a metric, including the triangle inequality.*

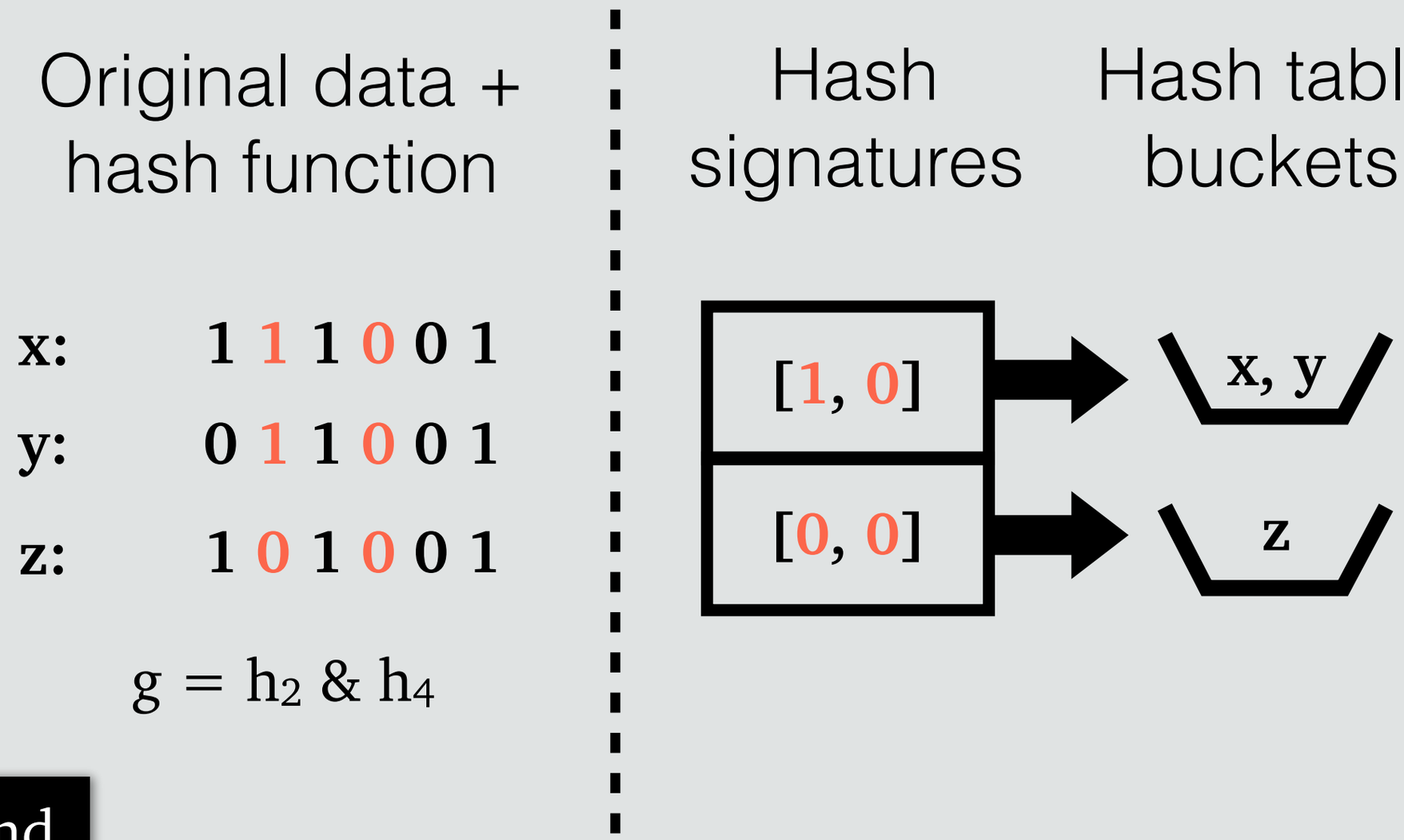


Pipeline



Locality-sensitive hashing

- Hash data s.t. **similar items likely to collide**
- Family of hash fns **F**: (d_1, d_2, p_1, p_2)-sensitive
  - Control false **negative**/positive rates
- Parameters
  - b: number of hash tables, increases p_1
  - r: number of hash functions to concatenate, lowers p_2



Proposed: ABC-LSH window sampling

Original data + hash function

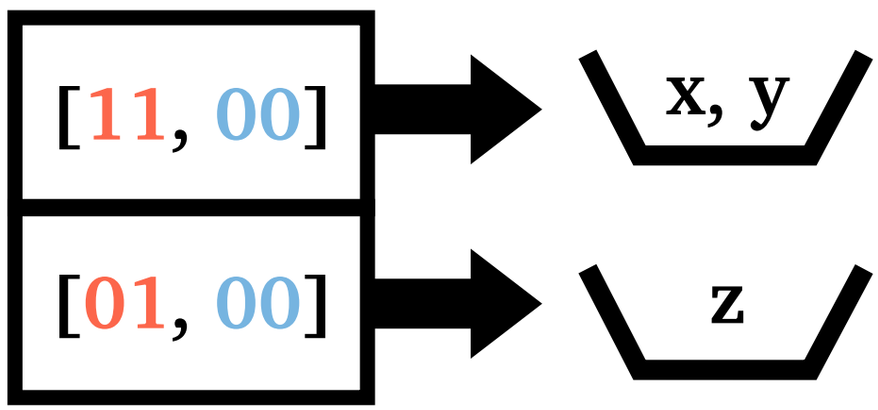
Hash signatures

Hash table buckets

x: 1 1 1 0 0 1

y: 0 1 1 0 0 1

z: 1 0 1 0 0 1

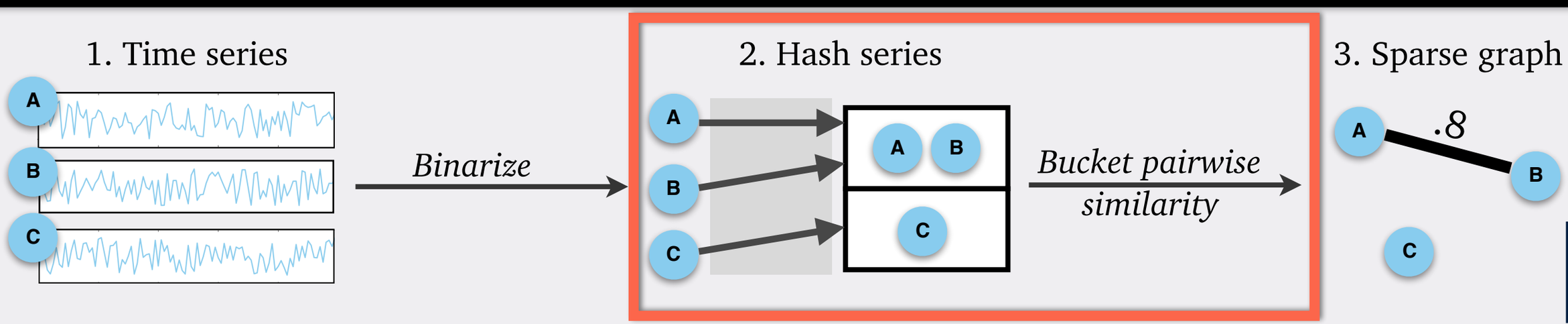


$k = 2$

$g = h_2 \,\&\, h_4$

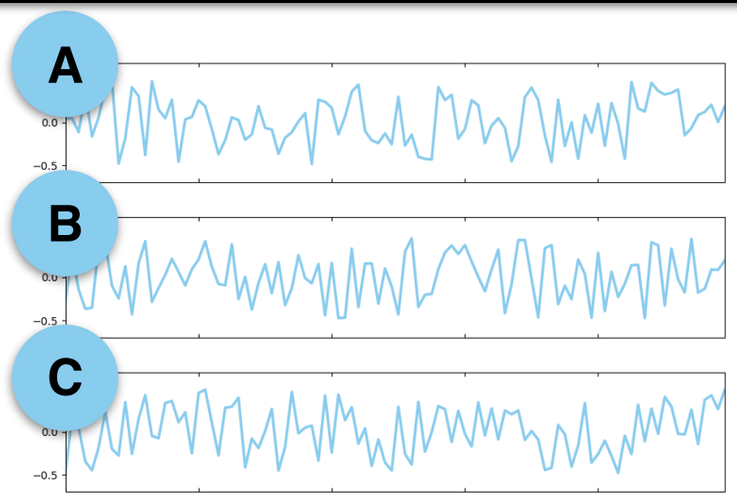
$(d_1, d_2, 1 - \alpha\frac{d_1}{(1+\alpha)^n - 1}, 1 - \alpha\frac{d_2}{(1+\alpha)^n - 1})$ - sensitive

Pipeline



Summary

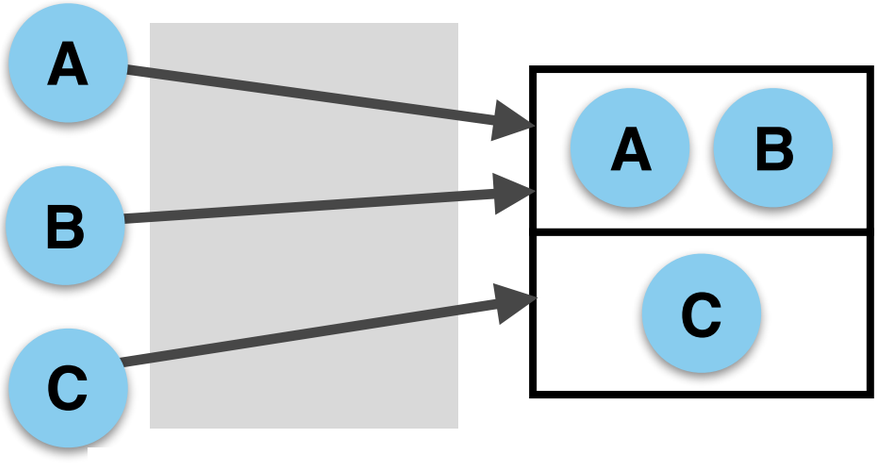
- Time-consecutive **locality-sensitive hashing** (LSH) family
- Novel **similarity measure** + distance metric on sequences



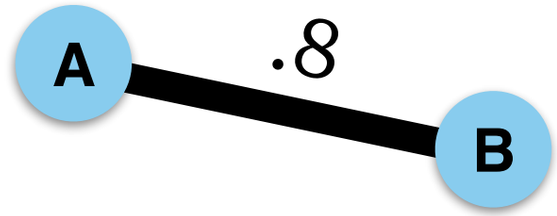
1. *N* time series

Binarize

2. Hash series

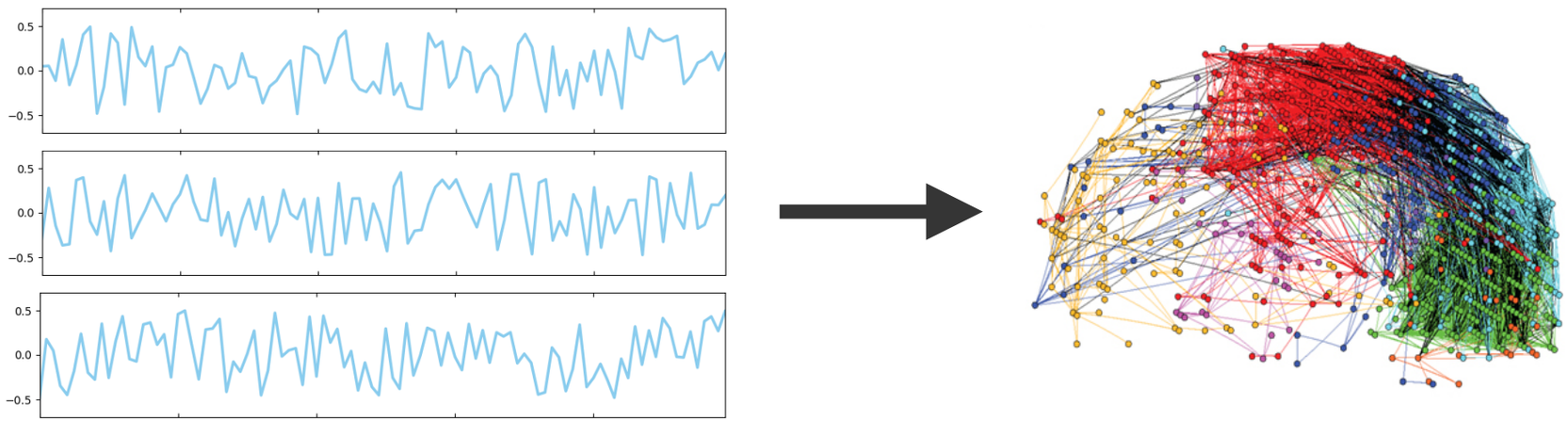


Bucket pairwise similarity



3. Sparse graph

Evaluation



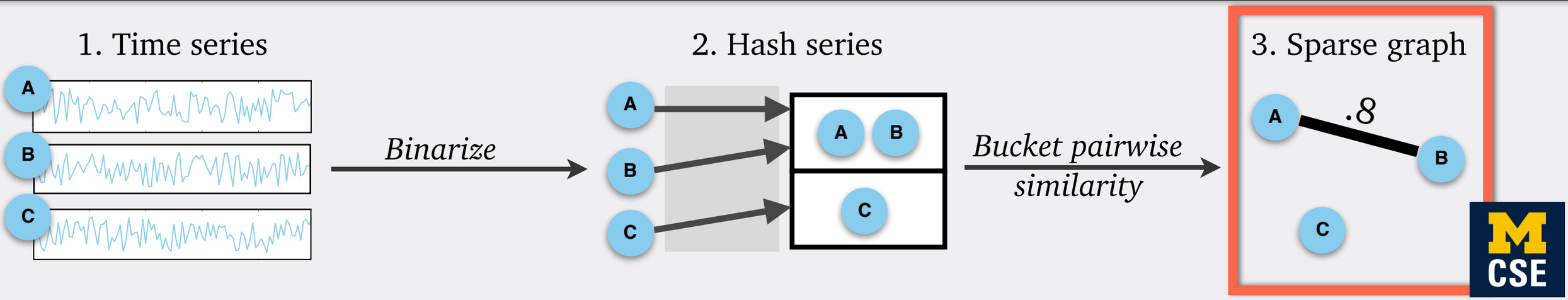
Evaluation questions

1. How **efficient** is our approach compared to baselines?
   - Baseline: pairwise correlation
   - Proposed: pairwise ABC, ABC-LSH
2. How **predictive** are the output graphs in real applications?
   - Can we predict brain health using graphs discovered with ABC-LSH?
3. How **robust** is our method to parameter choices?

Data

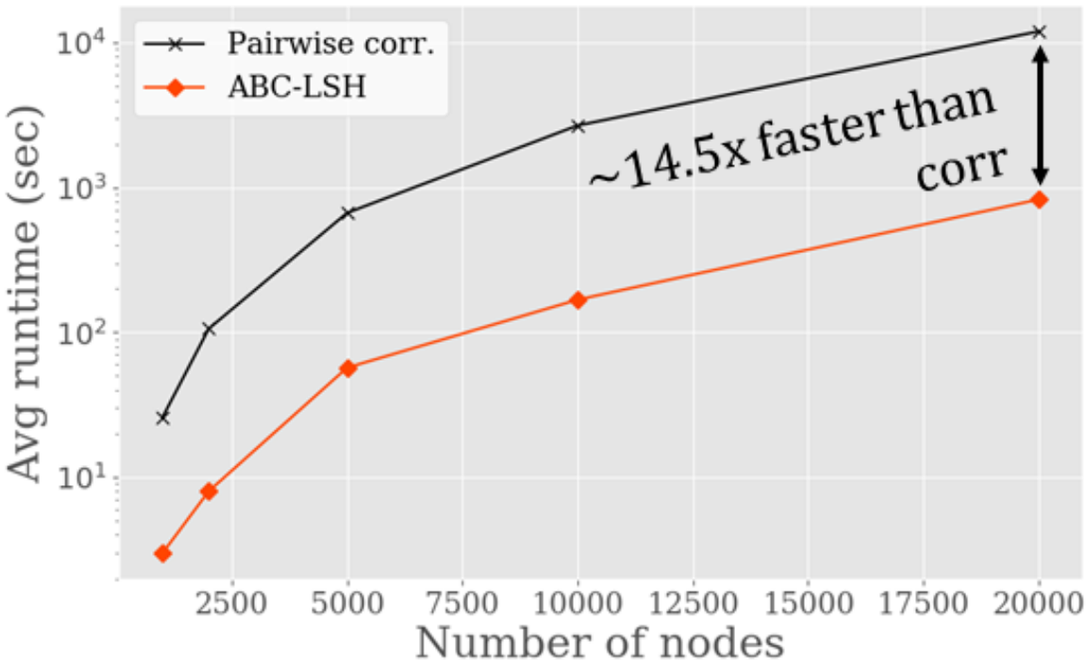
- Two publicly available fMRI datasets
- Synthetic data

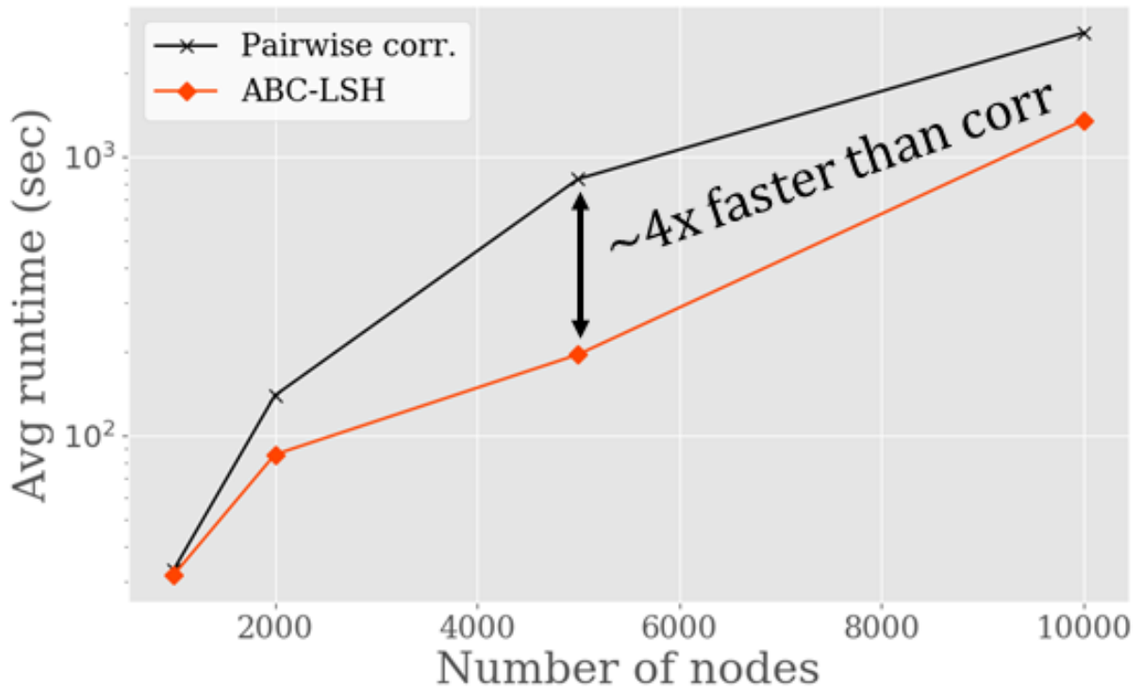
Pipeline



Question 1: scalability

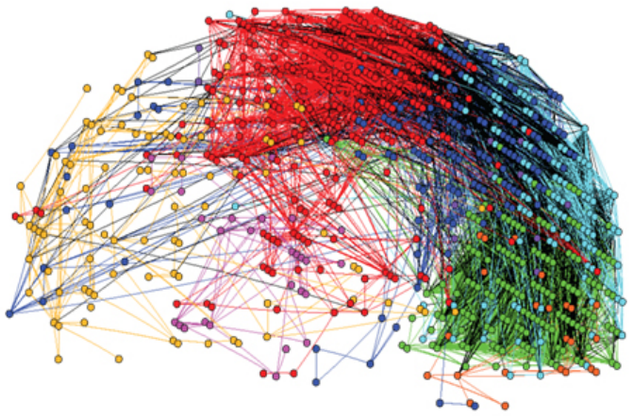
- 2 - 15*x* speedup with 2k - 20k nodes





Question 2: task-based evaluation

- Brain networks: identify **biomarkers** of mental disease
- Extract commonly used features from generated brain networks
  - Avg weighted degree
  - Avg clustering coefficient
  - Avg path length
  - Modularity
  - Density

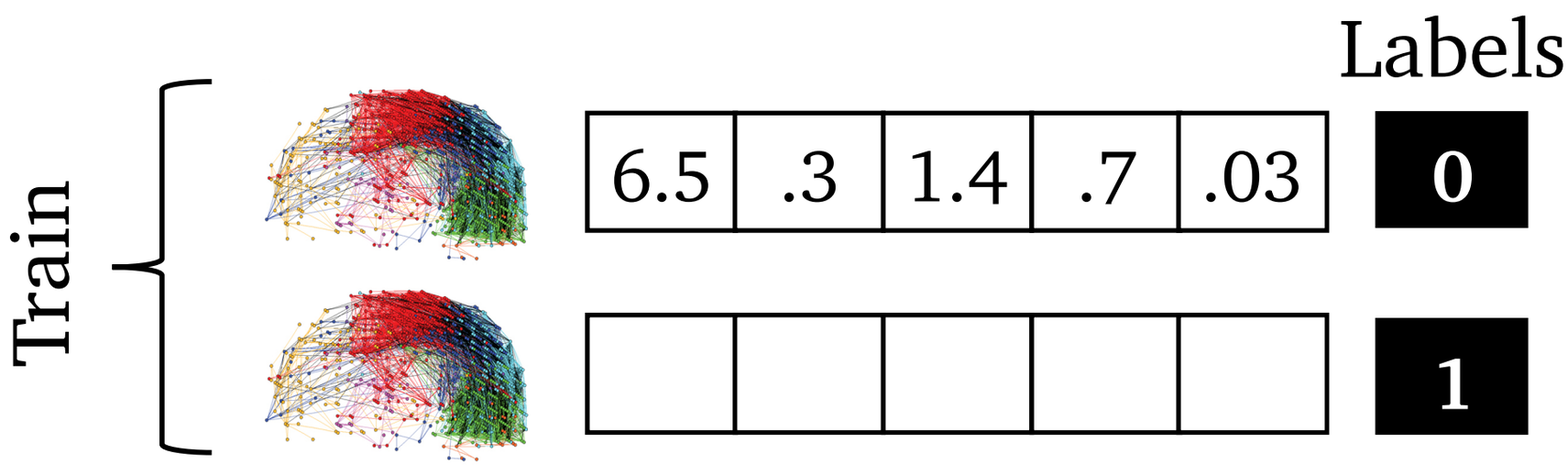


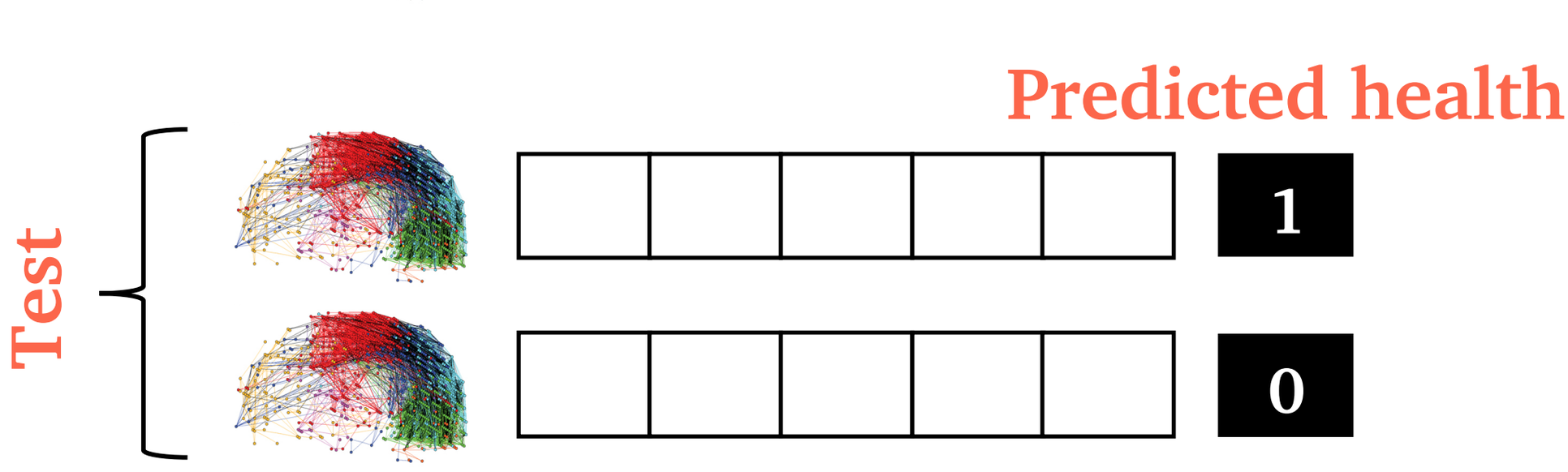
Feature selection

F_1	F_2	F_3	F_4	F_5	Healthy
6.5	.3	1.4	.7	.03	**0**

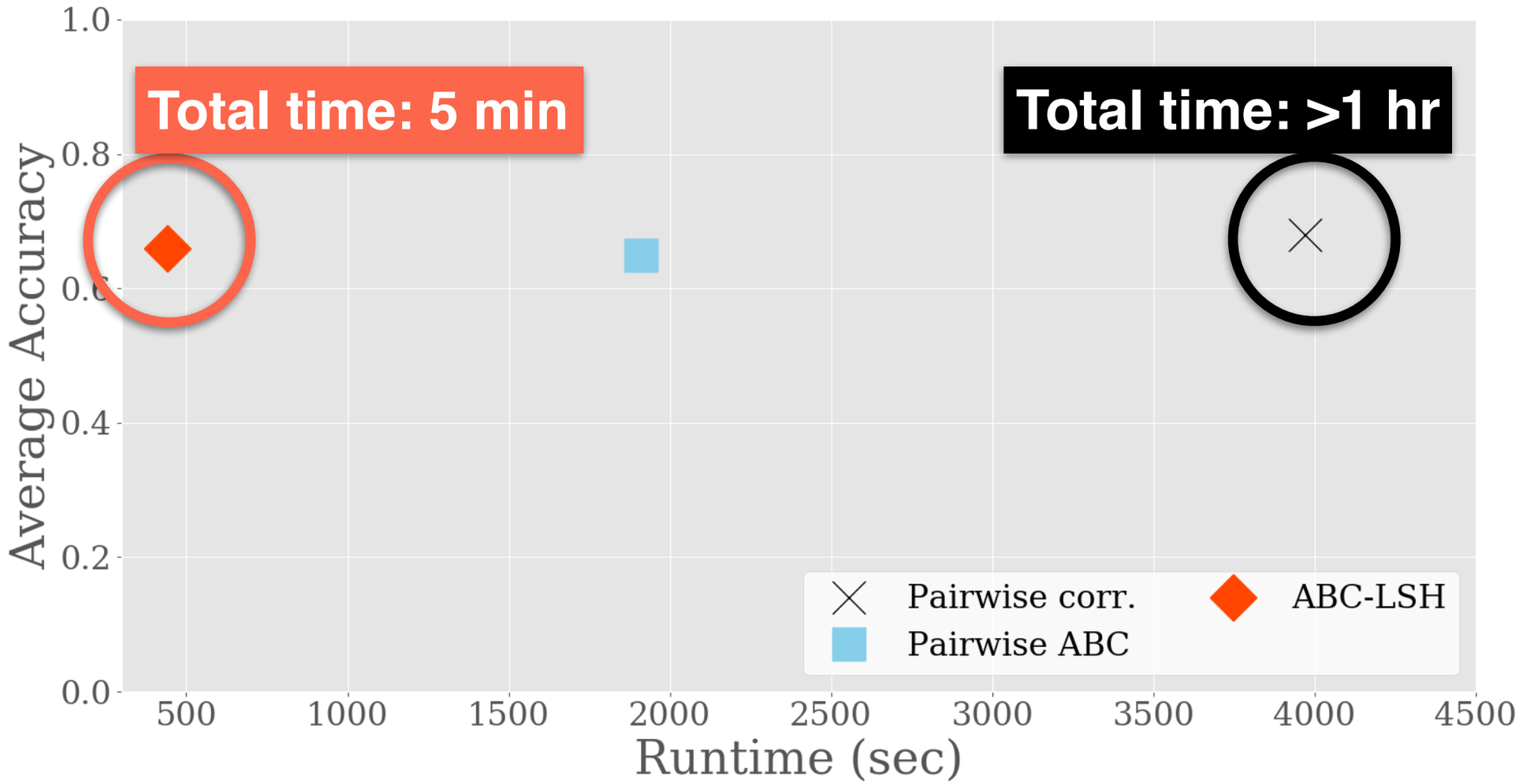
Question 2: task-based evaluation

- Logistic regression classifier, 10-fold stratified CV





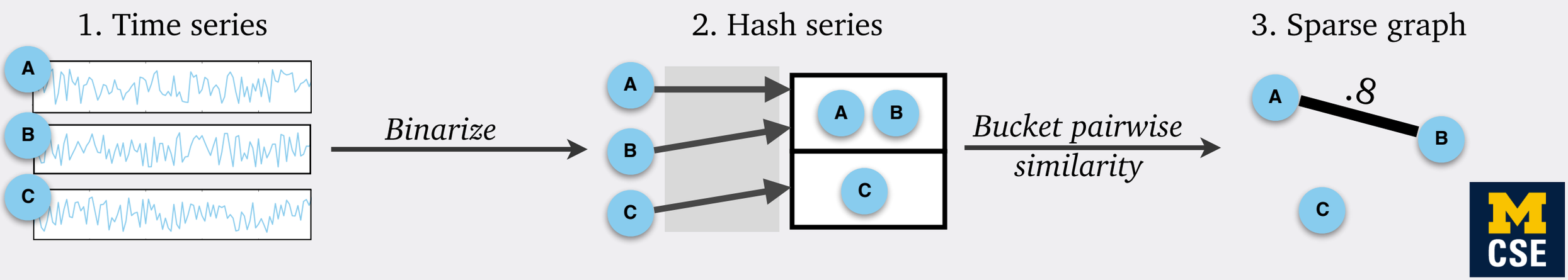
Question 2: task-based evaluation



Average accuracy same — runtime is not!

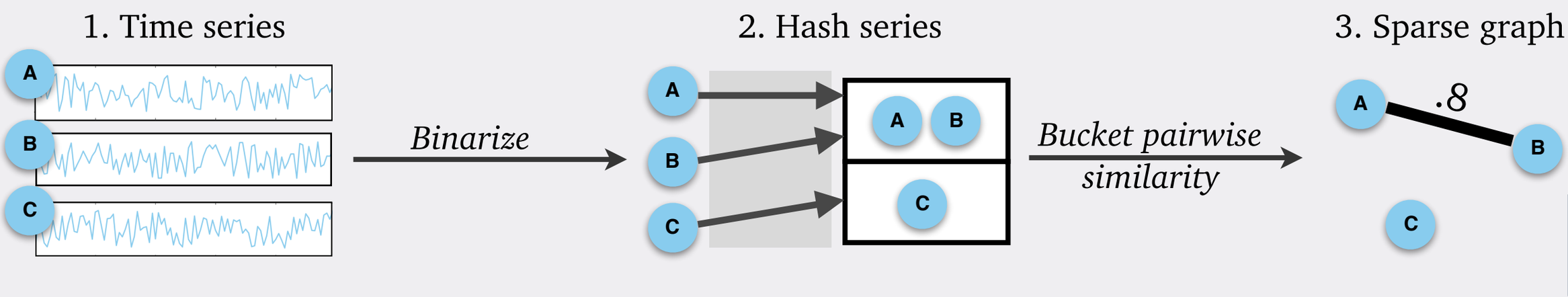
Conclusion

- Pipeline for network discovery on time series
  - ABC: time-consecutive **similarity measure** + distance metric on binary sequences
  - Associated **LSH family**
  - Modular + applicable in other settings



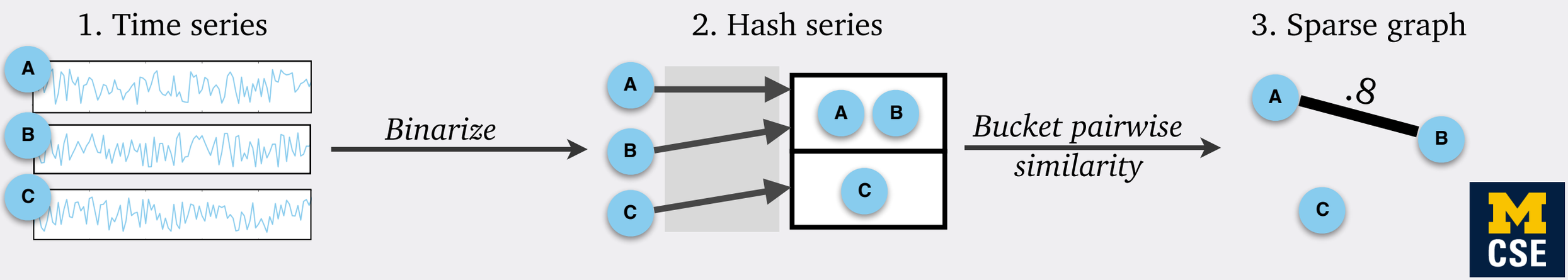
Conclusion

- Pipeline for network discovery on time series
  - ABC: time-consecutive **similarity measure** + distance metric on binary sequences
  - Associated **LSH family**
  - Modular + applicable in other settings
- Experiments: shown to be **fast** + **accurate**
  - Brain networks
  - More experiments on robustness, scalability, parameter sensitivity



Conclusion

- Pipeline for network discovery on time series
  - ABC: time-consecutive **similarity measure** + distance metric on binary sequences
  - Associated **LSH family**
  - Modular + applicable in other settings
- Experiments: shown to be **fast** + **accurate**
  - Brain networks
  - More experiments on robustness, scalability, parameter sensitivity
- Impact: integrated into production systems

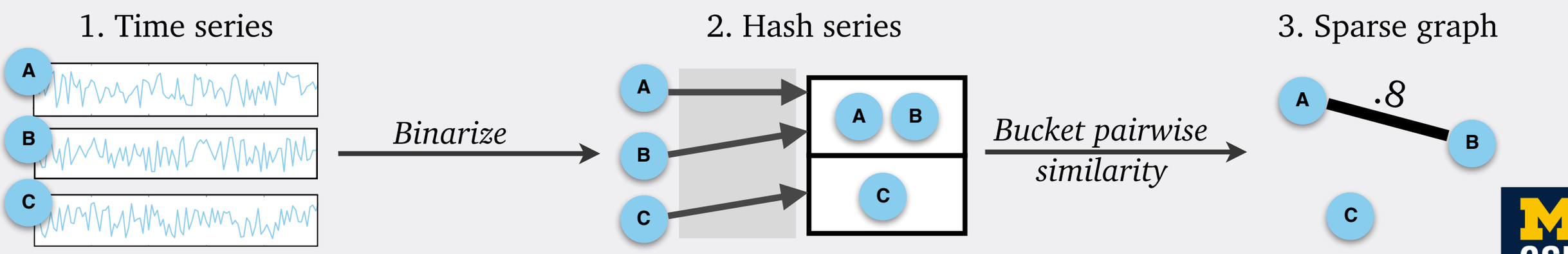


Conclusion

- Pipeline for network discovery on time series
  - ABC: time-consecutive **similarity measure** + distance metric on binary sequences
  - Associated **LSH family**
  - Modular + applicable in other settings
- Experiments: shown to be **fast** + **accurate**
  - Brain networks
  - More experiments on robustness, scalability, parameter sensitivity
- Impact: integrated into production systems

Thank you + questions

Supported by Google



Additional slides

Related work

Graphical model inference → Distributional assumptions

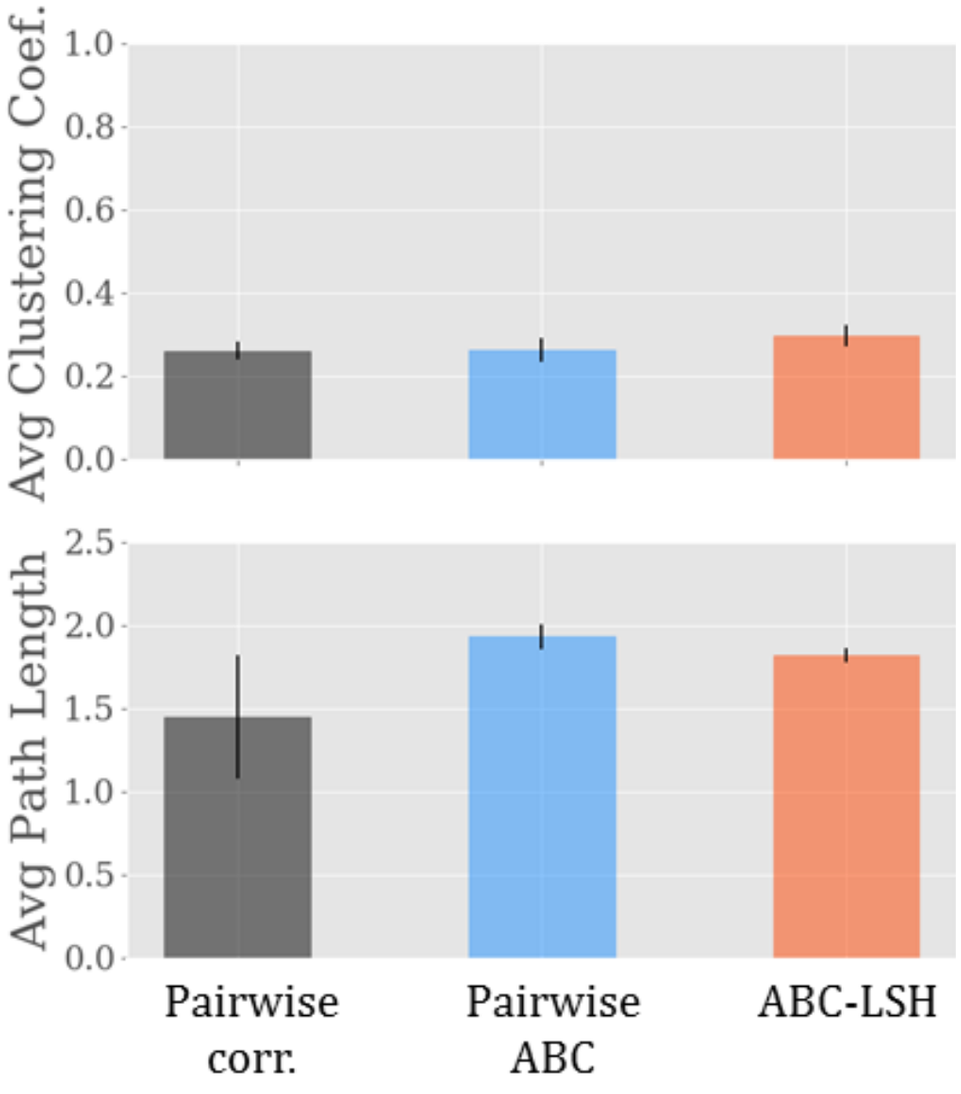
Time series hashing → Lacking consecutiveness and/or metrics

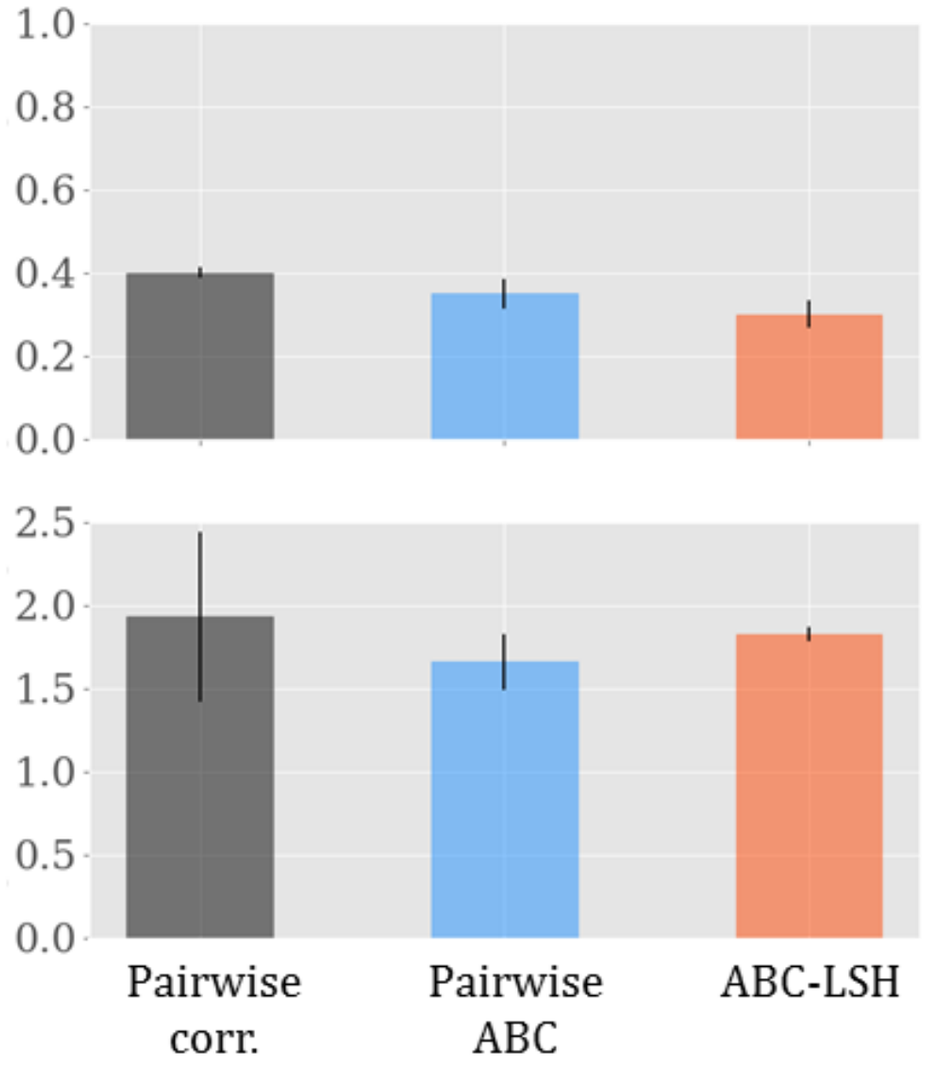
Similarity self-join

k-NN networks → User-set thresholds

Generated graph structure

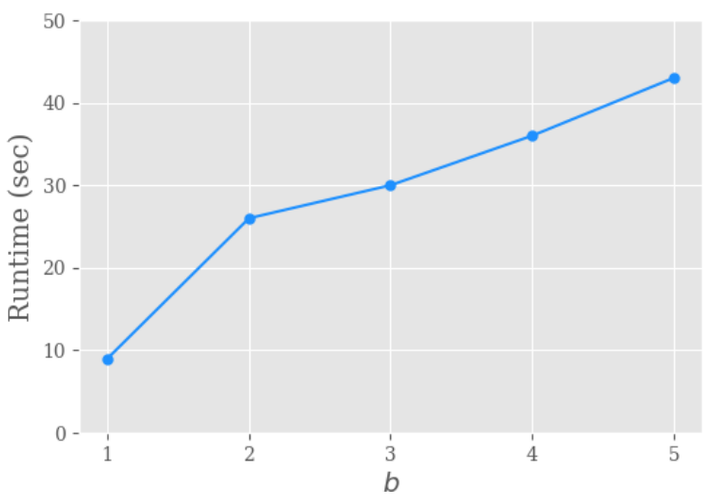
- Characteristics of generated graphs
  - Approximates correlation-based approach well

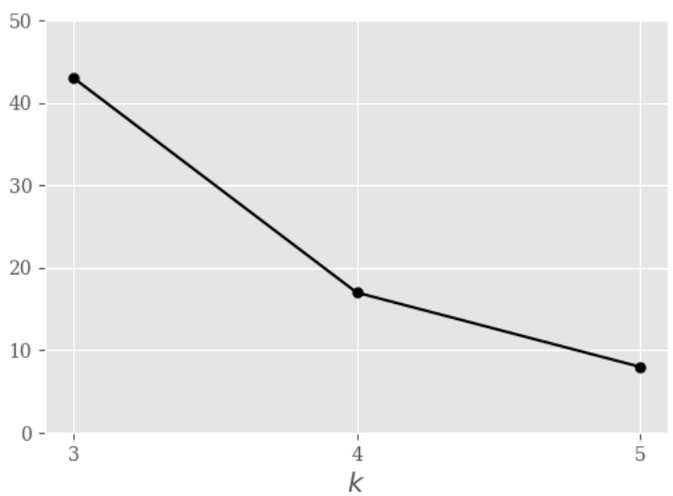


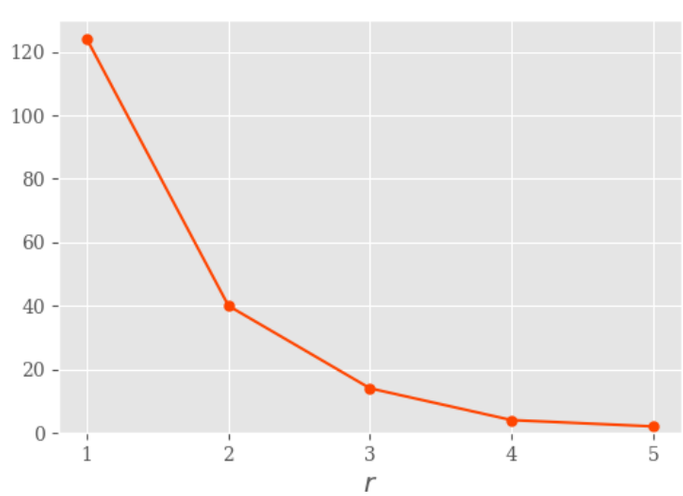


Parameter sensitivity

- How does scalability change varying k, b, and r?
- Results fairly intuitive
  - Increase b: more hash tables, slower
  - Increase k: longer windows, series less likely to collide, faster
  - Increase r: longer signatures, series less likely to collide, faster

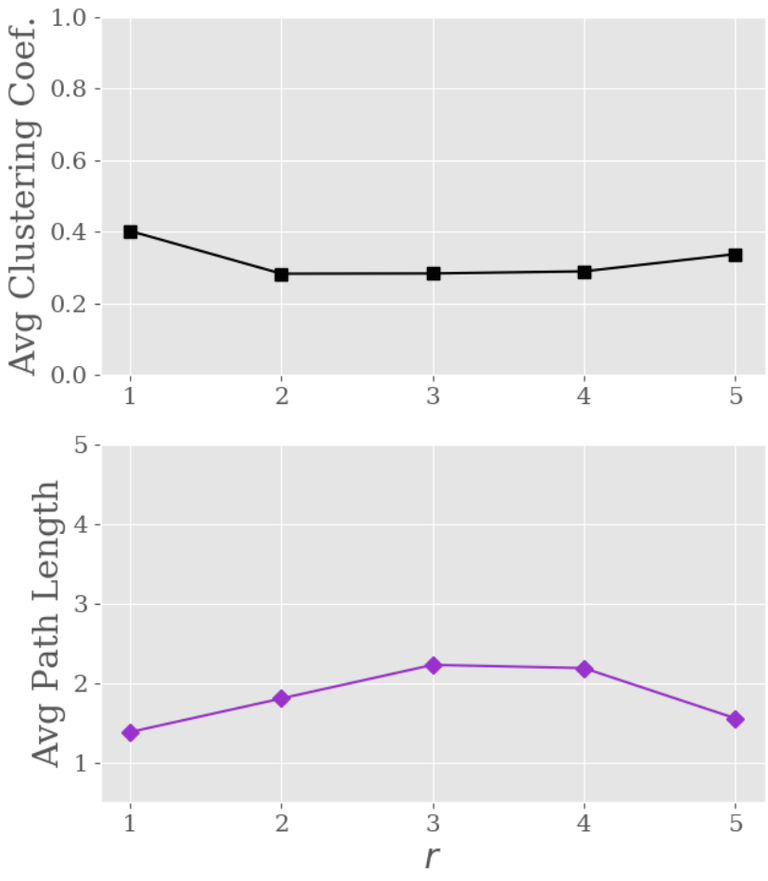


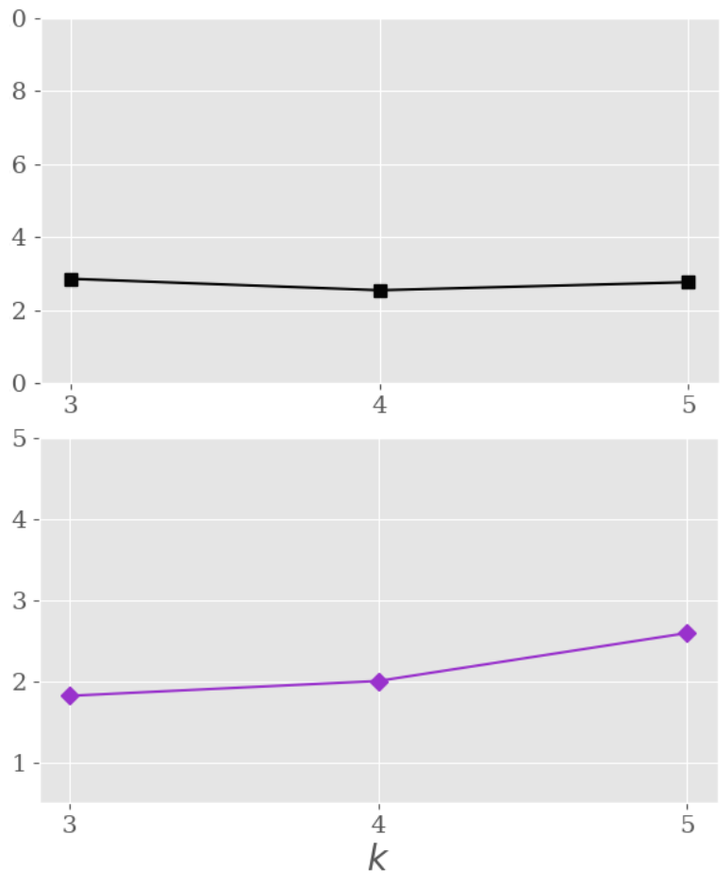




Parameter sensitivity

- How do graph properties change varying k, b, and r?
  - Not much





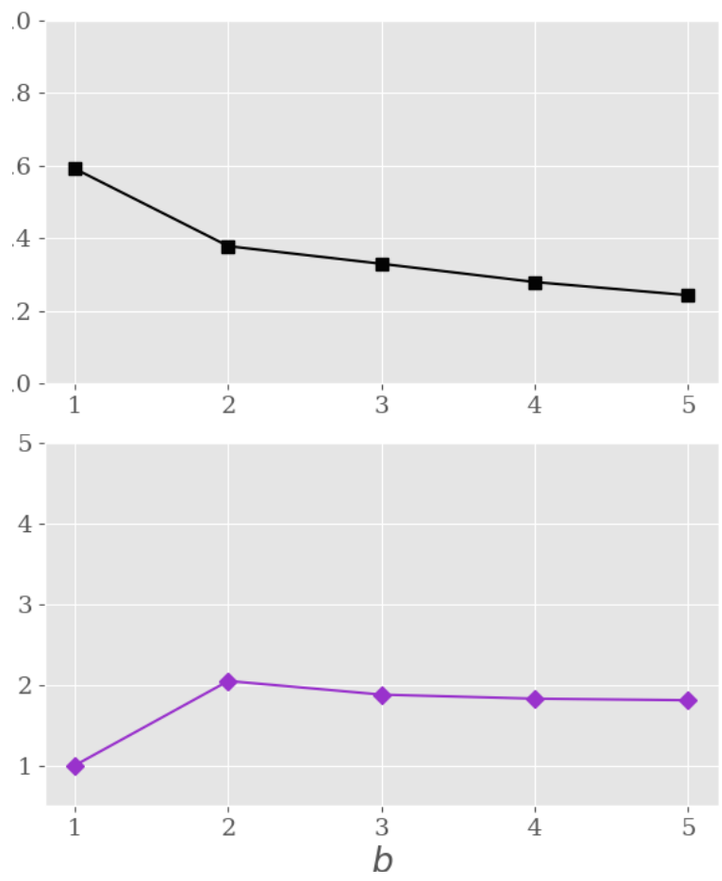


Image credits

- Slide 2, airline routes
- Slide 2, internet
- Slide 2, paper citations
- Slide 3, fMRI
- Slide 3, brain network
- Slide 5, gene sequences
- Slide 5, stocks