

Detecting Path Anomalies in Sequential Data on Networks

Tim LaRock

larock.t@husky.neu.edu

In collaboration with



Vahan Nanumyan



Ingo Scholtes



Giona Casiraghi



Tina Eliassi-Rad



Frank Schweitzer

Detecting Path Anomalies in Time Series Data on Networks

Timothy LaRock
Network Science Institute
Northeastern University
Boston, MA, USA

Ingo Scholtes
Data Analytics Group
University of Zurich
Zürich, Switzerland

Vahan Nanumyan
Chair of Systems Design
ETH Zürich
Zürich, Switzerland

Giona Casiraghi
Chair of Systems Design
ETH Zürich
Zürich, Switzerland

Tina Eliassi-Rad
Network Science Institute
Northeastern University
Boston, MA, USA

Frank Schweitzer
Chair of Systems Design
ETH Zürich
Zürich, Switzerland

<https://arxiv.org/abs/1905.10580>



This Talk

Motivation: Understanding mechanisms behind sequential data on networks

Today:

Motivate the study of **path anomalies**

Introduce **de Bruijn graph** representation of sequential data

Develop tractable **null model** to measure deviation of path data from expectation

Validate null model in synthetic data + compare with naïve baseline method

Application of methodology to a real system

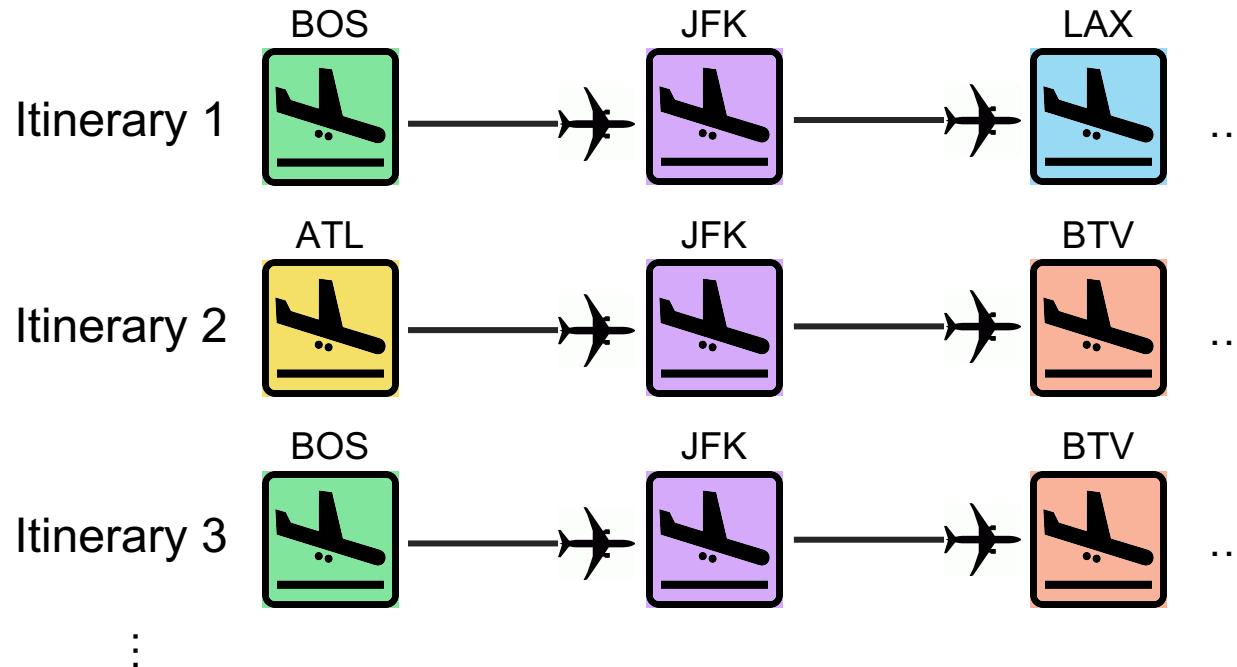


Intuitive Example: Passenger Flight Data

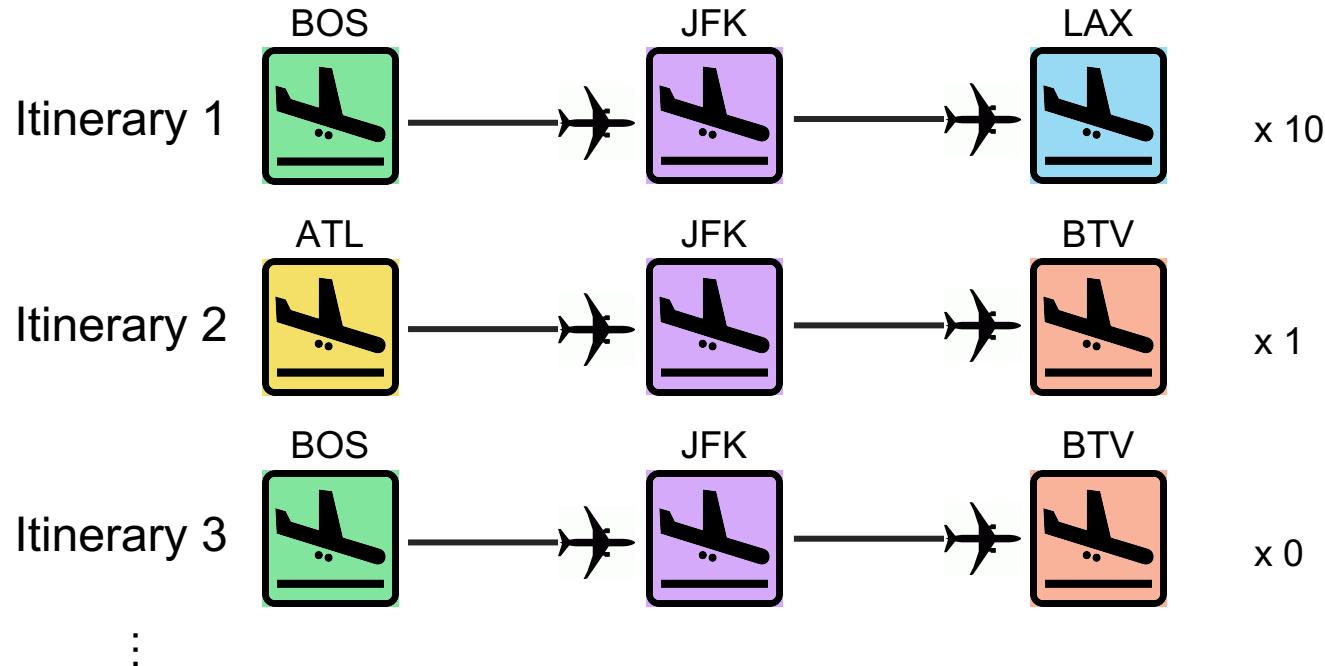


Northeastern University
Network Science Institute

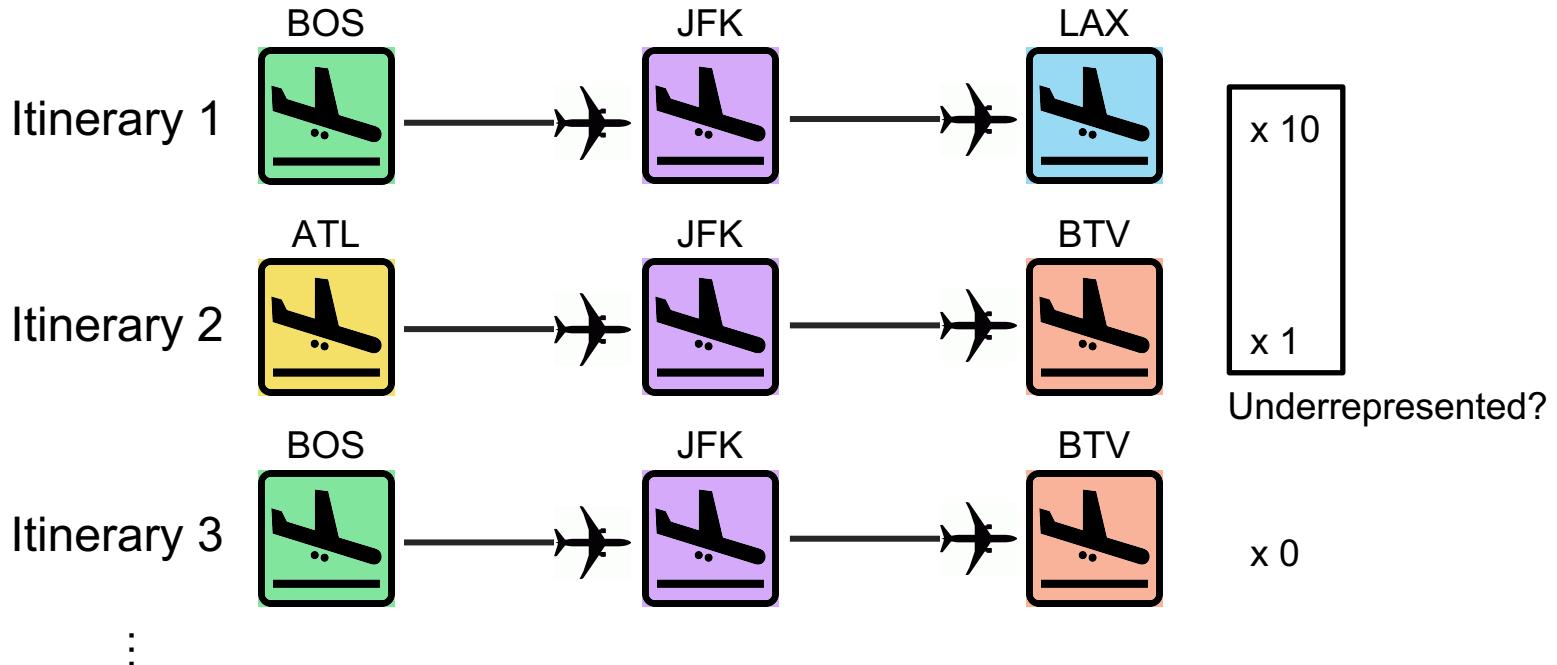
Intuitive Example: Passenger Flight Data



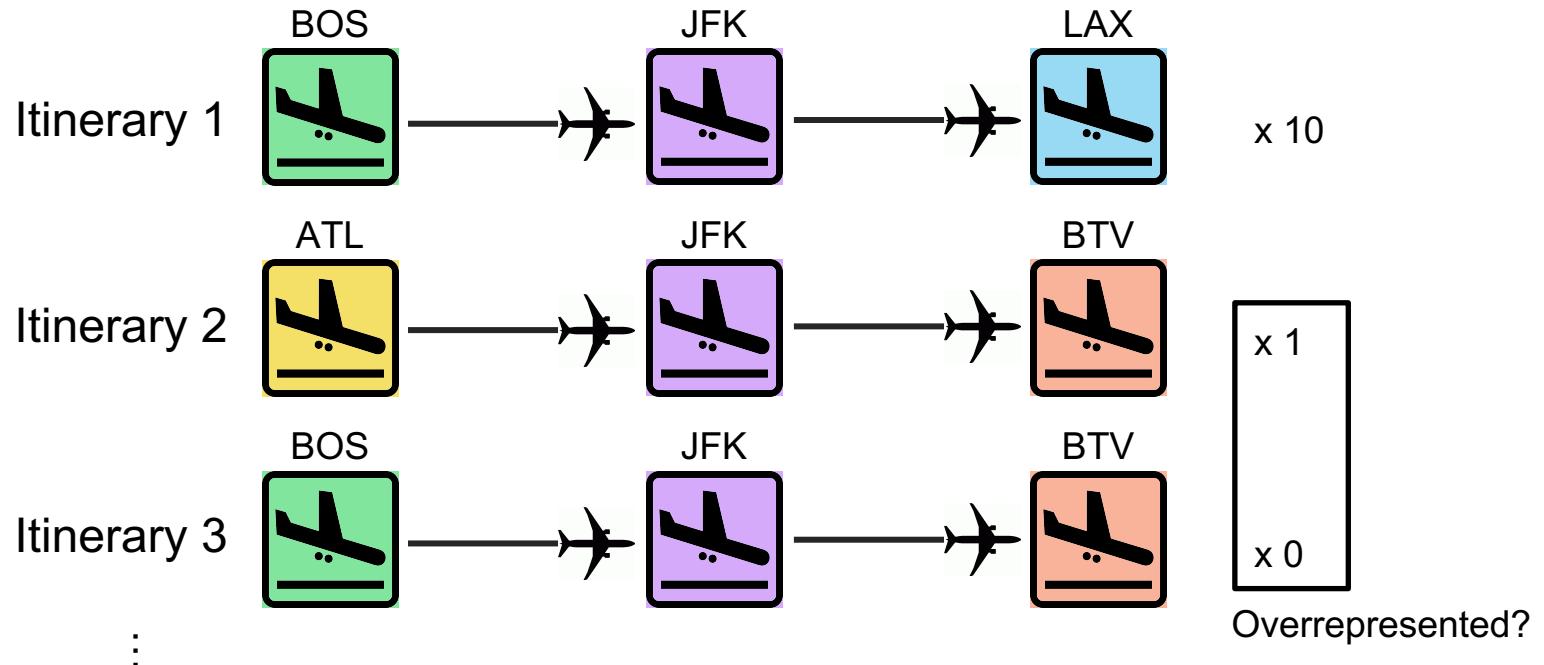
Intuitive Example: Passenger Flight Data



Intuitive Example: Passenger Flight Data

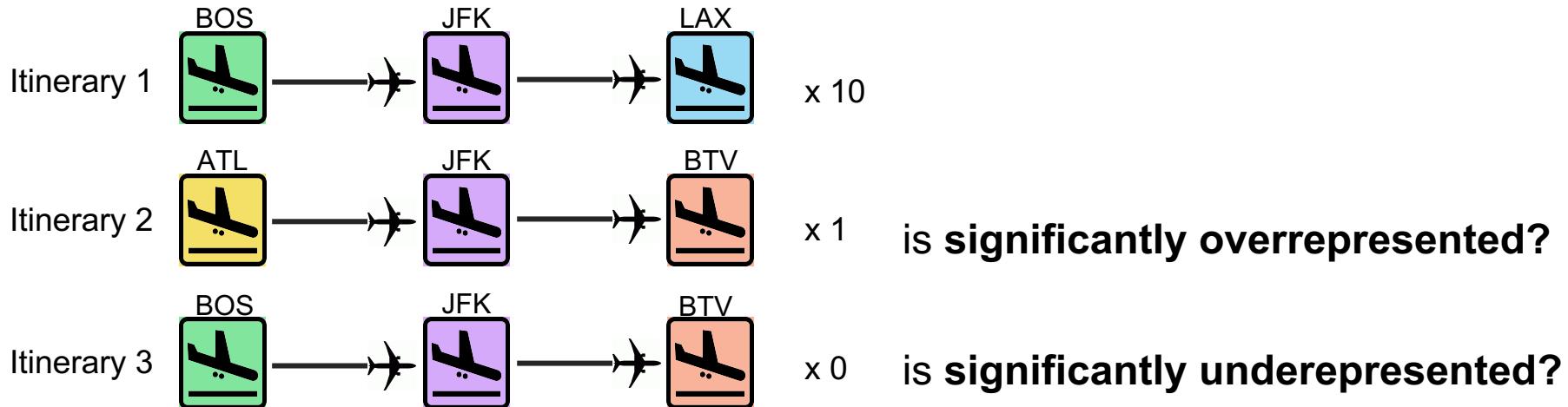


Intuitive Example: Passenger Flight Data



Research Question

Given this pathway dataset, can we determine whether...



Problem: Path anomaly detection

For a given pathway dataset S , graph G , and integer k , identify paths of length k through G whose observed frequencies in S deviate significantly from random expectation in a $(k-1)$ -order model of paths through G .

Problem: Path anomaly detection

For a given pathway dataset S , graph G , and integer k , identify paths of length k through G whose observed frequencies in S deviate significantly from random expectation in a $(k-1)$ -order model of paths through G .

When $k=2$, this corresponds to comparing a random walk with a single step of memory to a memoryless (Markovian) random walk on G .

Toy Example

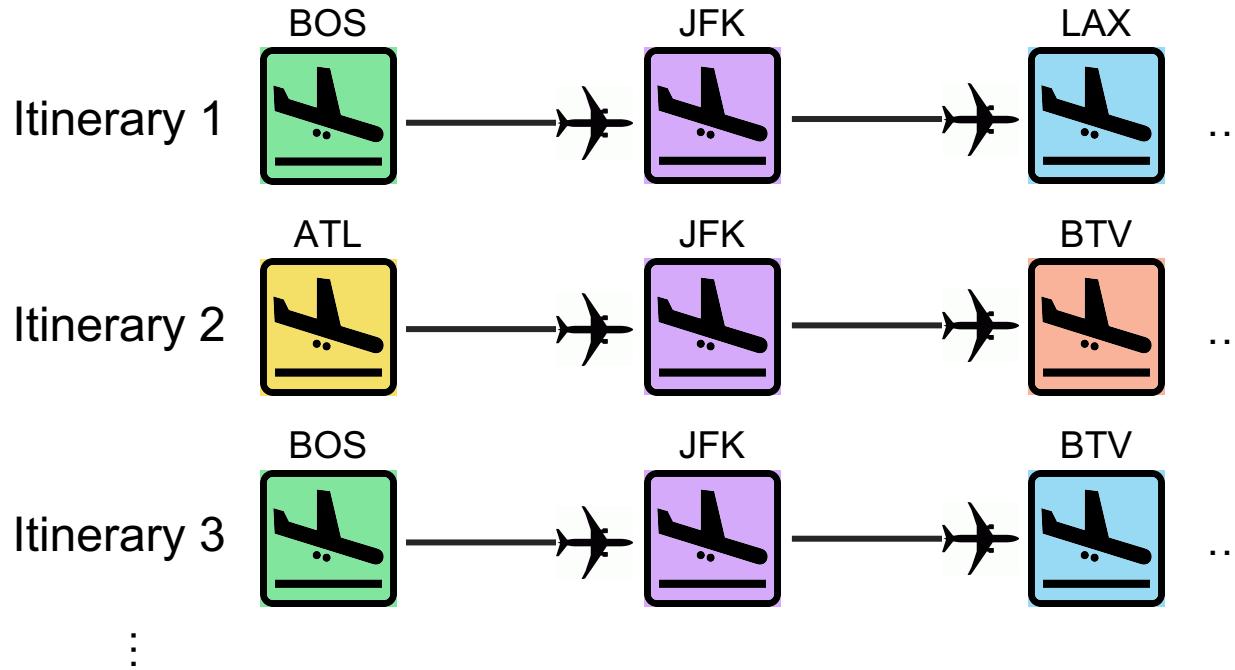
Three Goals:

1. Introduce de Bruijn graphs as representations of sequential data
2. Show how path anomalies emerge in a simple setting
3. Show how path anomalies can be detected through a random walk simulation approach

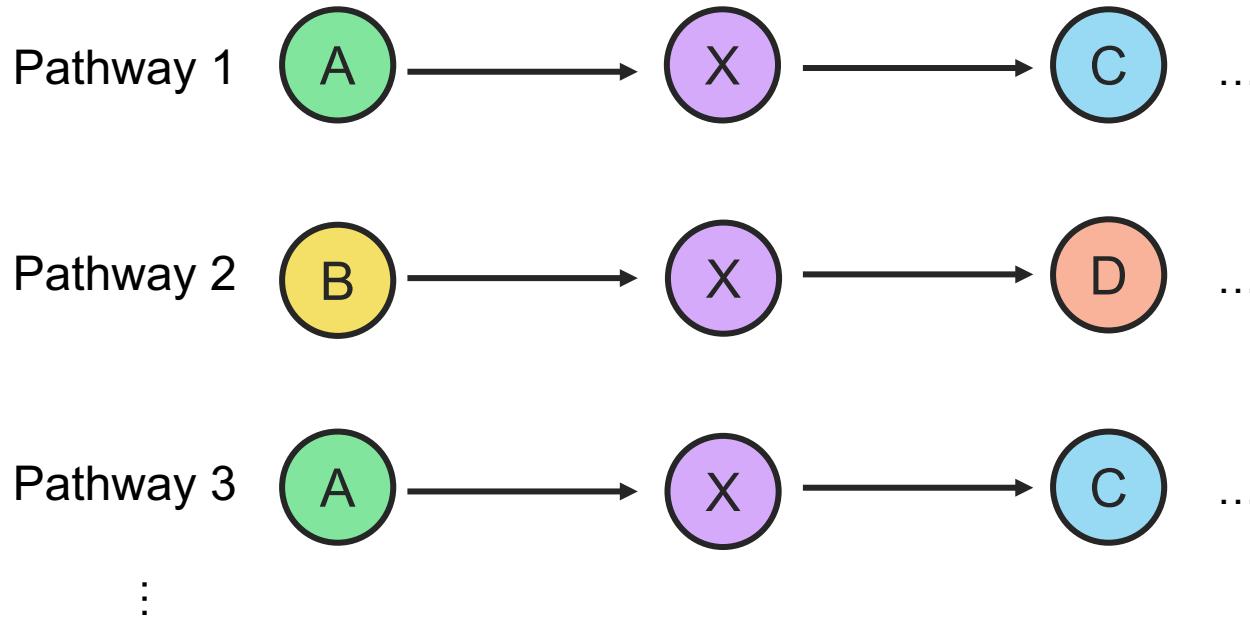
(Spoiler: Simulation approach is infeasible for real world datasets!)



Toy Example

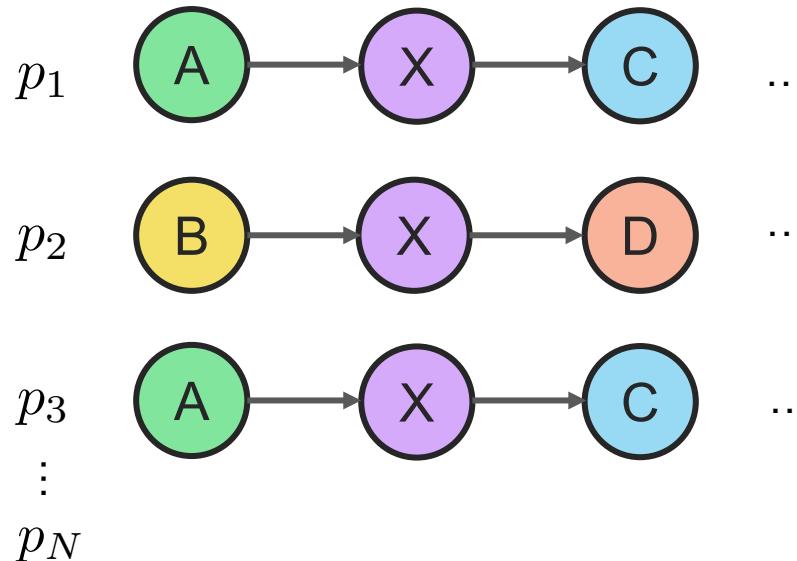


Toy Example

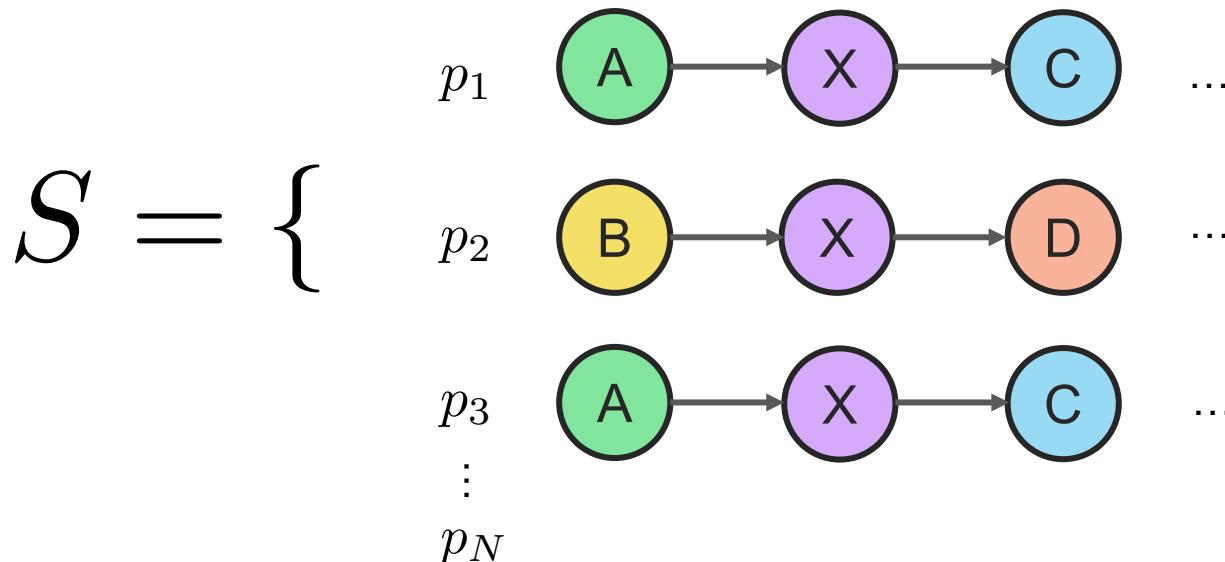


Toy Example: Data

$$S = \{$$

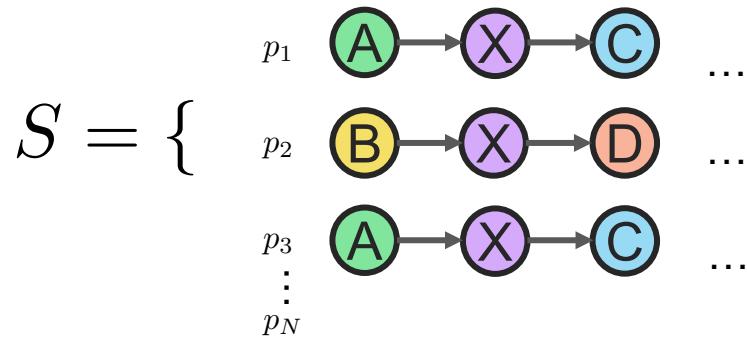


Toy Example: Data

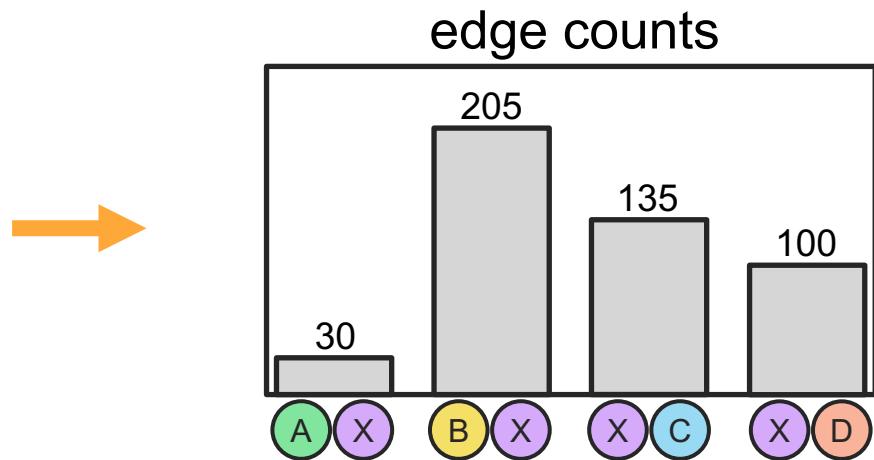


Total number of paths $N = |S| = 235$

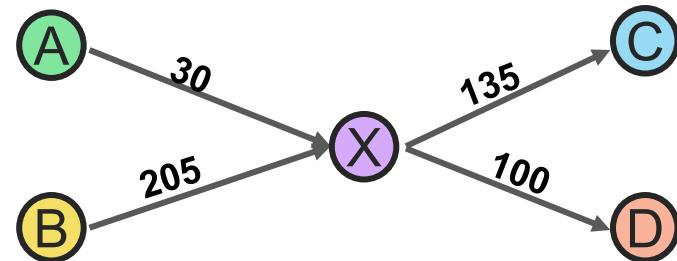
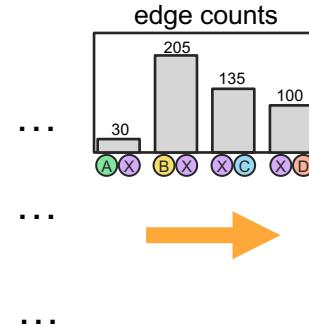
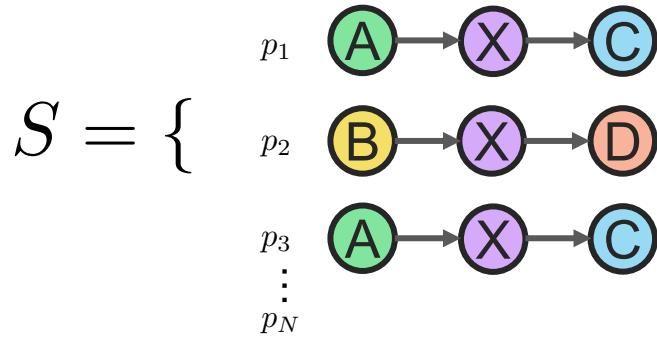
Toy Example: Data to (first-order) graph



Total number of paths $N = |S| = 235$

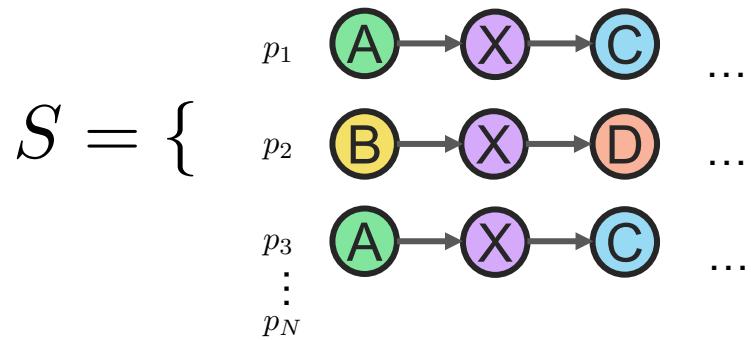


Toy Example: Data to (first-order) graph

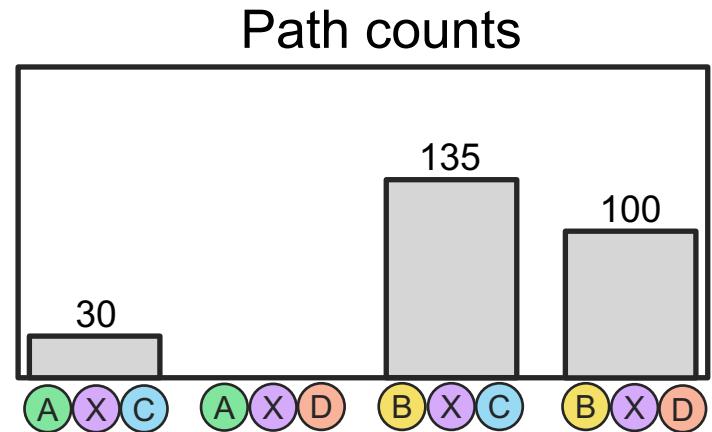


Total number of paths $N = |S| = 235$

Toy Example: Data to 2nd order de Bruijn graph



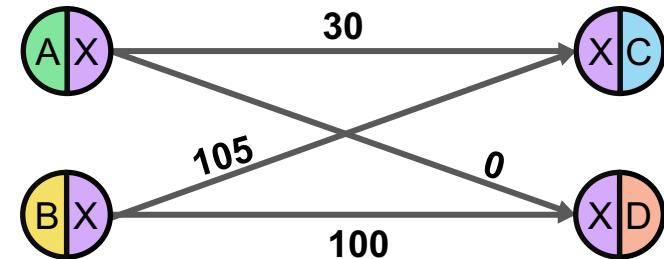
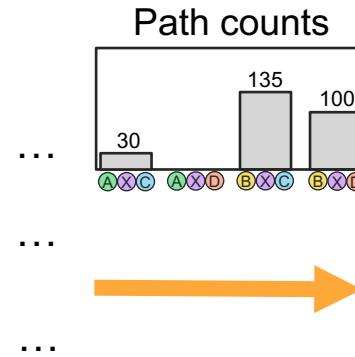
Total number of paths $N = |S| = 235$



Toy Example: Data to 2nd order de Bruijn graph

$S = \{$

p_1 
 p_2 
 p_3 
...
 p_N



Total number of paths $N = |S| = 235$

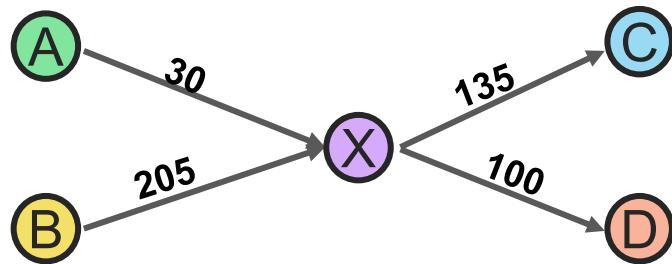
Toy Example: Path Anomalies via Simulations

For a given pathway dataset S , graph G , and integer k , identify paths of length k through G whose observed frequencies in S deviate significantly from random expectation in a $(k-1)$ -order model of paths through G .

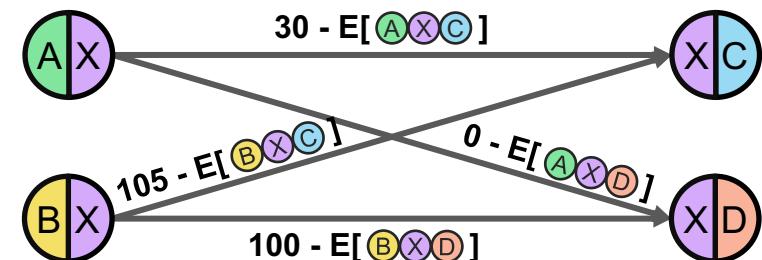
Toy Example: Path Anomalies via Simulations

For a given pathway dataset S , graph G , and integer k , identify paths of length k through G whose observed frequencies in S deviate significantly from random expectation in a $(k-1)$ -order model of paths through G .

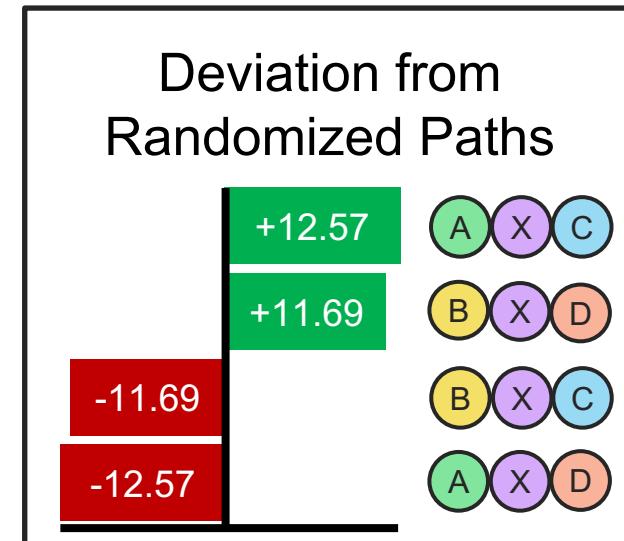
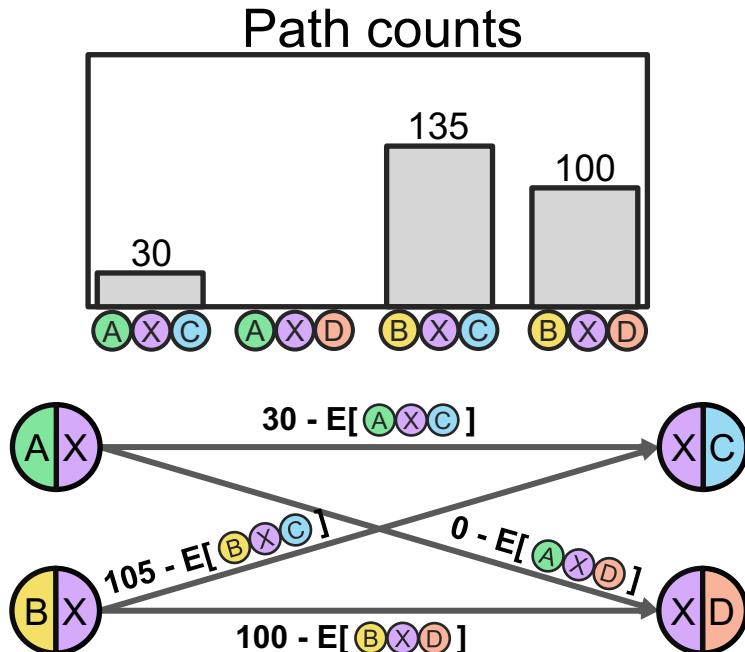
Simulate many random walk datasets



Compute expected frequency of each pathway and subtract from observed value



Toy Example: Path Anomalies via Simulations



Challenges



Path Anomaly Detection: Challenges

Detecting path anomalies via simulations → computationally intensive

Result is expected value, no concrete notion of significance

Alternative: detect path anomalies analytically by developing a tractable null model

Null Model: Challenges

Traditional null models (e.g. configuration model) cannot be applied directly



Null Model: Challenges

Traditional null models (e.g. configuration model) cannot be applied directly

Edges between higher-order nodes can not be randomized by stub-matching



Null Model: Challenges

Traditional null models (e.g. configuration model) cannot be applied directly

Edges between higher-order nodes can not be randomized by stub-matching



Need to randomize *edge weight distribution* in de Bruijn graph models, since connectivity structure is fixed by 1st-order topology

HYPA: Efficient Detection of Path Anomalies



Northeastern University
Network Science Institute

Generalized Hypergeometric Ensemble

Generalized Hypergeometric Ensembles: Statistical Hypothesis Testing in Complex Networks

Giona Casiraghi,^{1,*} Vahan Nanumyan,^{1,†} Ingo Scholtes,^{1,2,‡} and Frank Schweitzer^{1,§}

¹ETH Zürich, Chair of System Design, Weinbergstrasse 56/58, 8092 Zürich, Switzerland

²AIFB, Karlsruhe Institute of Technology, Karlsruhe, Germany

(Dated: 5th August 2016)

Statistical ensembles of networks, i.e., probability spaces of all networks that are consistent with given aggregate statistics, have become instrumental in the analysis of complex networks. Their numerical and analytical study provides the foundation for the inference of topological patterns, the definition of network-analytic measures, as well as for model selection and statistical hypothesis testing. Contributing to the foundation of these data analysis techniques, in this Letter we introduce *generalized hypergeometric ensembles*, a broad class of analytically tractable statistical ensembles of finite, directed and weighted networks. This framework can be interpreted as a generalization of the classical configuration model, which is commonly used to randomly generate networks with a given degree sequence or distribution. Our generalization rests on the introduction of *dyadic link propensities*, which capture the *degree-corrected* tendencies of pairs of nodes to form edges between each other. Studying empirical and synthetic data, we show that our approach provides broad perspectives for model selection and statistical hypothesis testing in data on complex networks.

PACS numbers: 89.75.Hc, 02.50.Sk, 89.75.Kd

Generalised hypergeometric ensembles of random graphs: the configuration model as an urn problem

Giona Casiraghi* Vahan Nanumyan†

Chair of Systems Design,

ETH Zurich, Weinbergstrasse 56/58, 8092 Zurich, Switzerland

*gcasiraghi@ethz.ch

†vnanumyan@ethz.ch

Abstract

We introduce a broad class of random graph models: the generalised hypergeometric ensemble (GHypEG). This class enables to solve some long standing problems in random graph theory. First, GHypEG provides an elegant and compact formulation of the well-known configuration model in terms of an urn problem. Second, GHypEG allows to incorporate arbitrary tendencies to connect different vertex pairs. Third, we present the closed-form expressions of the associated probability distribution ensures the analytical tractability of our formulation. This is in stark contrast with the previous state-of-the-art, which is to implement the configuration model by means of computationally expensive procedures.



Generalized Hypergeometric Ensemble

Generalization of the configuration model to weighted, directed networks.



Generalized Hypergeometric Ensemble

Generalization of the configuration model to weighted, directed networks.

Fixes the *expected* weight of every node, rather than the *exact* degree sequence.



Generalized Hypergeometric Ensemble

Generalization of the configuration model to weighted, directed networks.

Fixes the *expected* weight of every node, rather than the *exact* degree sequence.

Intuition: Urn problem where each pair of nodes i, j that can possibly connect is assigned a color, and K_{ij} balls are added, where K_{ij} is the total possible multi-edges between the nodes (e.g. $K_{ij} = k_i^{\text{out}} k_j^{\text{in}}$, as in the config model).

Generalized Hypergeometric Ensemble

Generalization of the configuration model to weighted, directed networks.

Fixes the *expected* weight of every node, rather than the *exact* degree sequence.

Intuition: Urn problem where each pair of nodes i, j that can possibly connect is assigned a color, and K_{ij} balls are added, where K_{ij} is the total possible multi-edges between the nodes (e.g. $K_{ij} = k_i^{\text{out}} k_j^{\text{in}}$, as in the config model).

To sample from the ensemble, draw $m = \sum_{ij} W_{ij}$ (total observed edges) multi-edges randomly from the urn.

$$m = \sum_{ij} W_{ij}$$
$$K_{ij} = k_i^{\text{out}} k_j^{\text{in}}$$

Hypergeometric Ensemble

$$\Pr(X_{vw} = f(v, w)) \propto \binom{K_{vw}}{f(v, w)} \binom{\sum_{ij} K_{ij} - K_{vw}}{m - f(v, w)}$$

$$m = \sum_{ij} W_{ij}$$

Hypergeometric Ensemble

$$K_{ij} = k_i^{\text{out}} k_j^{\text{in}}$$

$$Pr(X_{vw} = f(v,w)) \propto \underbrace{\binom{K_{vw}}{f(v,w)}}_{\text{Hypergeometric}} \binom{\sum_{ij} K_{ij} - K_{vw}}{m - f(v,w)}$$

Probability of observing frequency $f(v,w)$ given the entire weighted network structure.

$$m = \sum_{ij} W_{ij}$$

$$K_{ij} = k_i^{\text{out}} k_j^{\text{in}}$$

Hypergeometric Ensemble

$$\Pr(X_{vw} = f(v,w)) \propto \binom{K_{vw}}{f(v,w)} \binom{\sum_{ij} K_{ij} - K_{vw}}{m - f(v,w)}$$

Number of ways to pick $f(v,w)$ multiedges from K_{vw} possible.

$$m = \sum_{ij} W_{ij}$$

Hypergeometric Ensemble

$$K_{ij} = k_i^{\text{out}} k_j^{\text{in}}$$

$$Pr(X_{vw} = f(v, w)) \propto \binom{K_{vw}}{f(v, w)} \binom{\sum_{ij} K_{ij} - K_{vw}}{m - f(v, w)}$$

Number of ways to pick everything else.

Putting it all together: HYPA scores

$$\text{HYPA}^{(k)}(\vec{v}, \vec{w}) := \Pr(X_{\vec{v}\vec{w}} \leq f(\vec{v}, \vec{w}))$$

Putting it all together: HYPA scores

$$\text{HYPA}^{(k)}(\vec{v}, \vec{w}) := \Pr(X_{\vec{v}\vec{w}} \leq f(\vec{v}, \vec{w}))$$

If “close” to **0**, then the pathway is **underrepresented**.

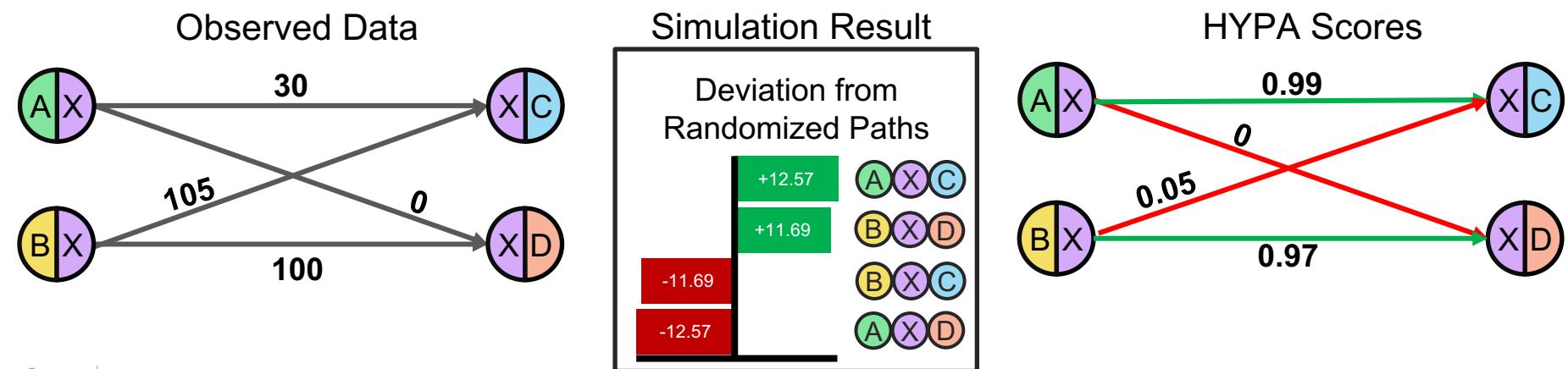
If “close” to **1**, then pathway is **overrepresented**.

Putting it all together: HYPA scores

$$\text{HYPA}^{(k)}(\vec{v}, \vec{w}) := \Pr(X_{\vec{v}\vec{w}} \leq f(\vec{v}, \vec{w}))$$

If “close” to 0, then the pathway is **underrepresented**.

If “close” to 1, then pathway is **overrepresented**.

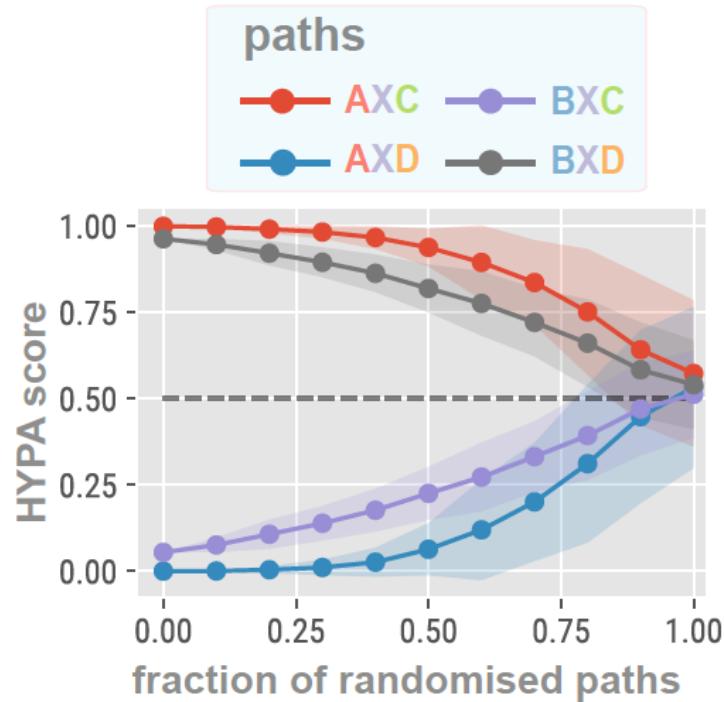


Validation



Northeastern University
Network Science Institute

Noise via Path Randomization



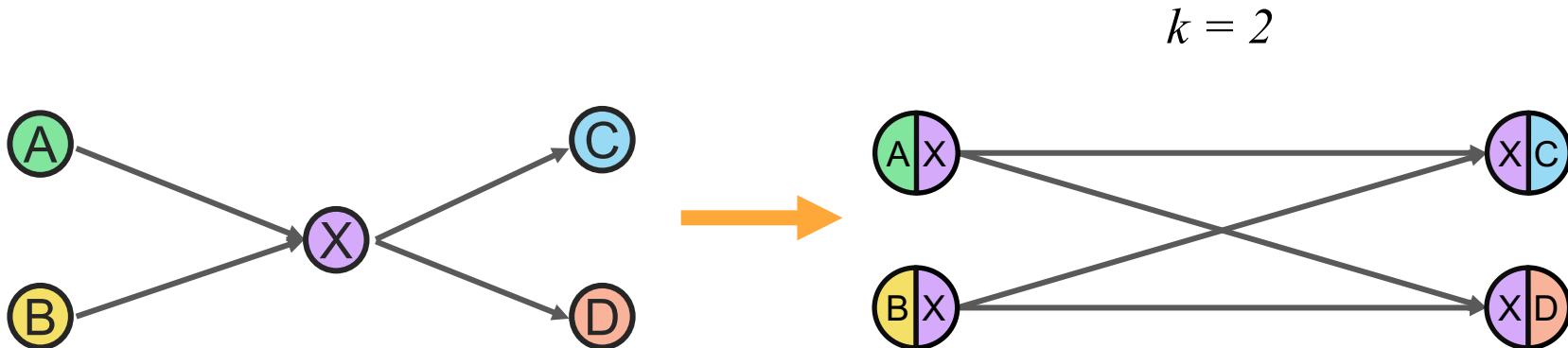
Synthetic Anomalies: Setup



Northeastern University
Network Science Institute

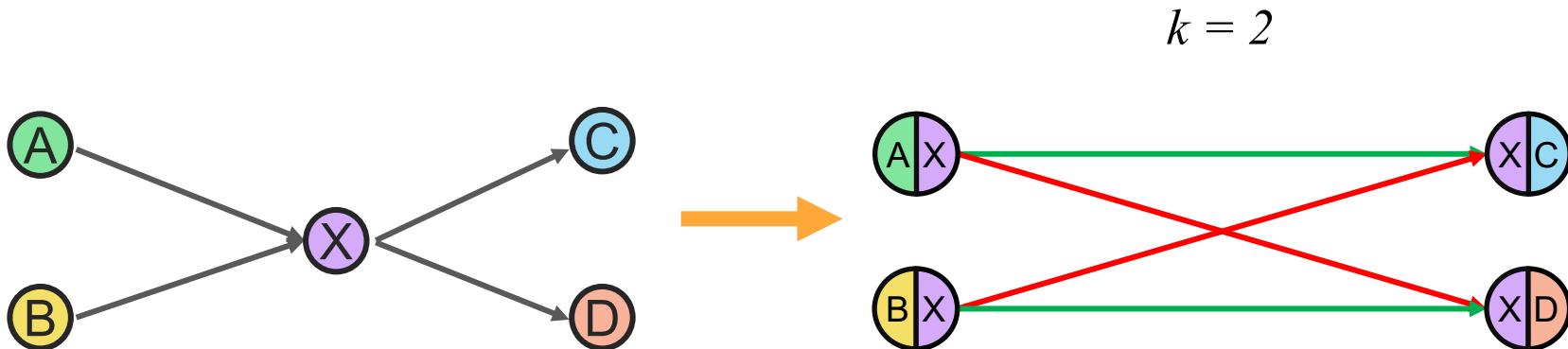
Synthetic Anomalies: Setup

Start with an arbitrary first order topology, then construct the k -order de Bruijn graph



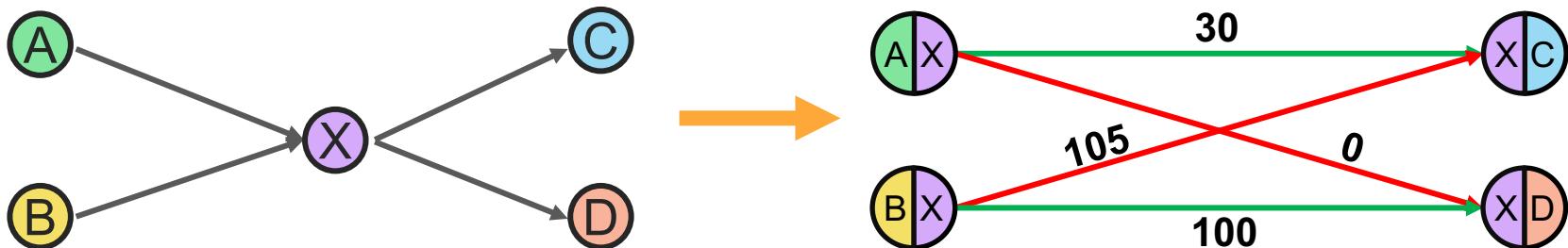
Synthetic Anomalies: Setup

Randomly choose some edges to label over-represented



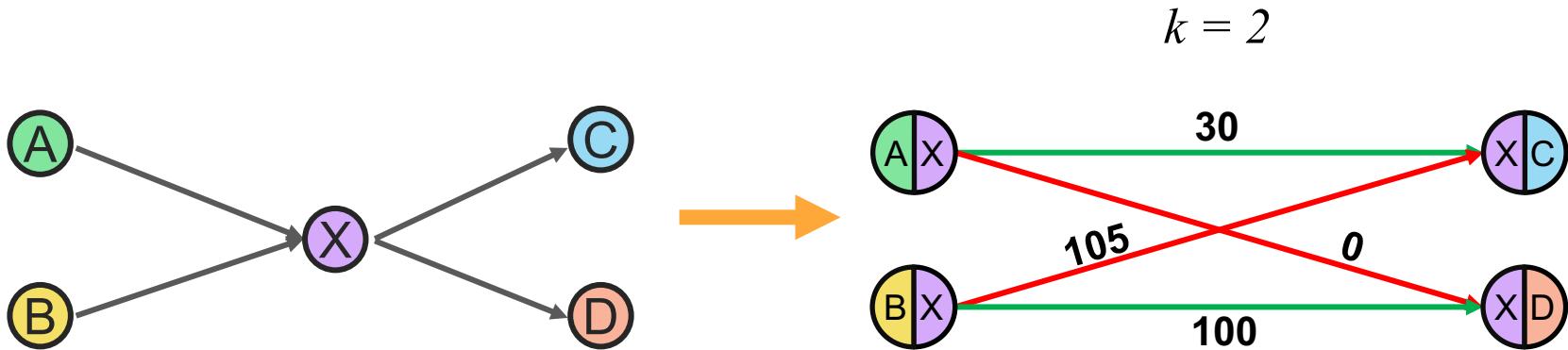
Synthetic Anomalies: Setup

Assign heterogeneous weights based on label



Synthetic Anomalies: Setup

Generate paths via random walks on this model, then evaluate ability of HYPA to detect injected anomalies (binary classifier).

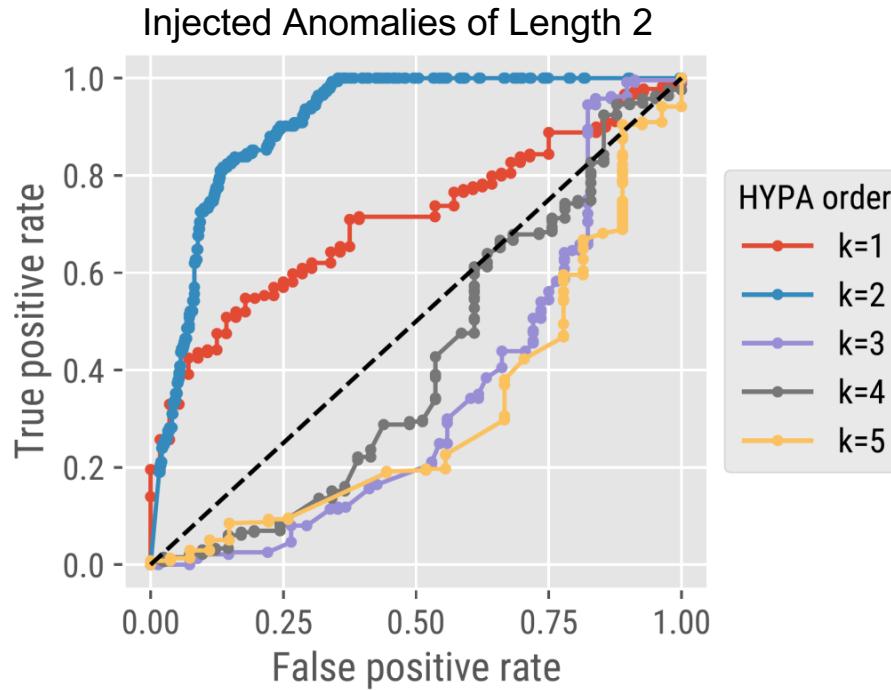


Synthetic Anomalies: ROC Example

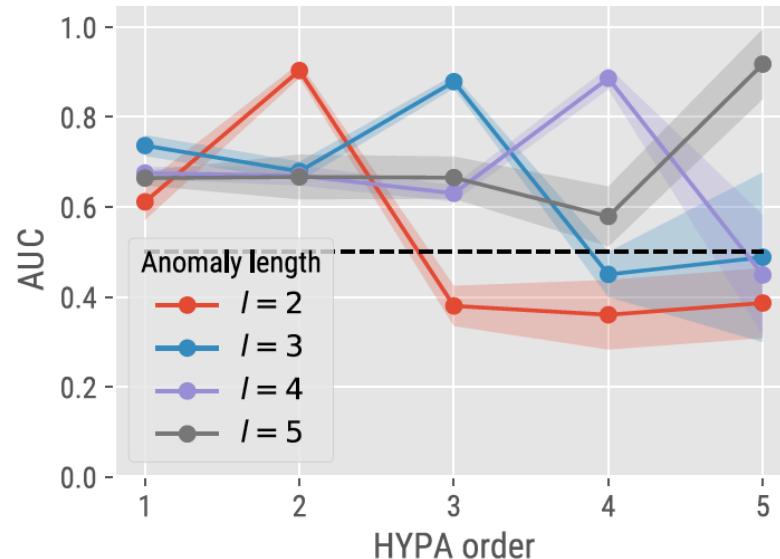


Northeastern University
Network Science Institute

Synthetic Anomalies: ROC Example



Synthetic Anomalies: AUC Results



Application to Flight Data



Airlines

5% sample of all US domestic flights in 2018

	Topology		Sequences				
Data	Nodes	Edges	Total	Unique	l^{\max}	$\langle l \rangle$	
Flights	382	6933	185871	88539	10	2.48	

Airlines

Hypotheses:

1. Return flights should be over-represented, since people most often travel round trip.



Airlines: Return trips are over-represented

α	Return	Non-return
0.05	0.915	0.340
0.01	0.851	0.130
0.001	0.760	0.023
0.0001	0.688	0.004
0.00001	0.628	0.001

Fraction of over-represented return/non-return flights for various discrimination thresholds.

Airlines: Return trips are over-represented

α	Return	Non-return
0.05	0.915	0.340
0.01	0.851	0.130
0.001	0.760	0.023
0.0001	0.688	0.004
0.00001	0.628	0.001

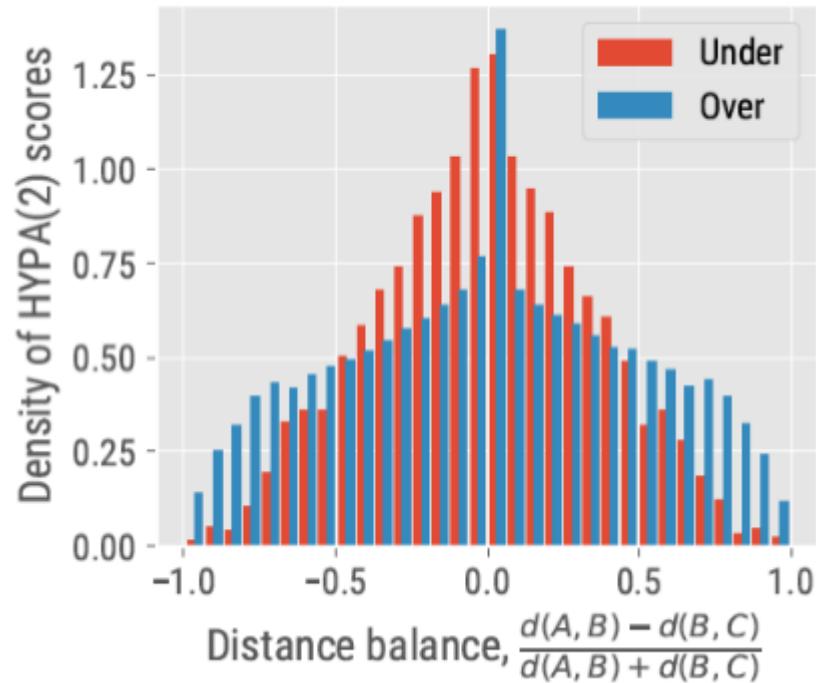
Fraction of over-represented return/non-return flights for various discrimination thresholds.

Airlines

Hypotheses:

1. Return flights should be over-represented, since people most often travel round trip.
2. Over-represented non-return flights are due to regional/national hubs, since people need to fly from small airports → regional hub → large airport.

Airlines: Trip Balance

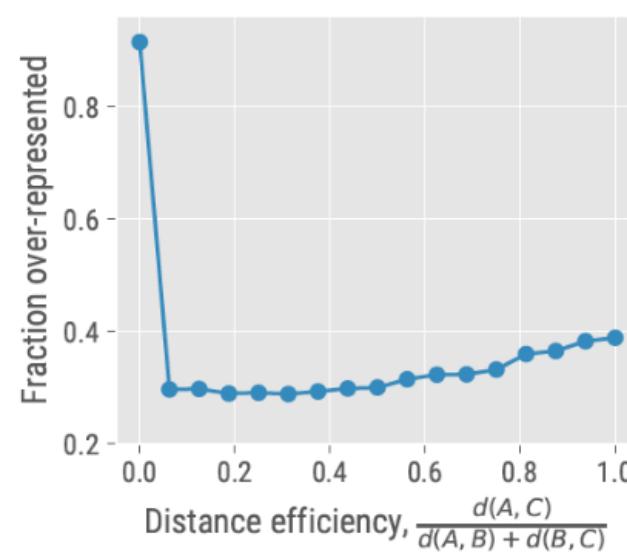
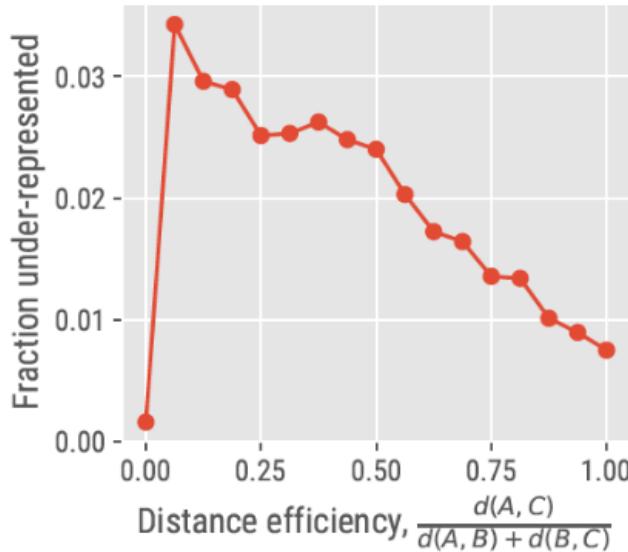


Airlines

Hypotheses:

1. Return flights should be over-represented, since people most often travel round trip.
2. Over-represented non-return flights are due to regional/national hubs, since people need to fly from small airports → regional hub → large airport.
3. “Efficient” paths are more likely to be over-represented.

Airlines: Efficiency



Thanks!

Tim LaRock

larock.t@husky.neu.edu

tlarock.github.io

<https://arxiv.org/abs/1905.10580>

References

Scholtes, Ingo. "When is a network a network?: Multi-order graphical model selection in pathways and temporal networks." *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2017.

Casiraghi et al. "Generalized hypergeometric ensembles: Statistical hypothesis testing in complex networks." *arXiv:1607.02441* (2016).

Casiraghi & Nanumyan. "Generalised hypergeometric ensembles of random graphs: the configuration model as an urn problem." *arXiv:1810.06495* (2018)

R. TransStat. Origin and destination survey database. http://www.transtats.bts.gov/Tables.asp?DB_ID=125, 2018.

Definition: kth-order de Bruijn Graph

For a given graph $G = (V, E)$ and positive integer k we define a k -th order De Bruijn graph of paths in G as a graph $G^k = (V^k, E^k)$, where (i) each node $\vec{v} := v_0v_1 \dots v_{k-1} \in V^k$ is a path of length $k - 1$ in G , and (ii) $(\vec{v}, \vec{w}) \in E^k$ iff $v_{i+1} = w_i$ for $i = 0, \dots, k - 2$.

Pseudocode

Algorithm 1 ComputeHYPA(S, k): *Compute kth order HYPA scores for sequence dataset S.*

Input: S (sequences), k (desired order)

Output: HYPA $^{(k)}$ score for all k -th order paths

- 1: $G^k \leftarrow \text{DeBruijnGraph}(S, k)$ # Construct k th order graph
 - 2: $\Xi \leftarrow \text{fitXi}(G^k, \text{tolerance})$ # Optimization (Algorithm 2 in Appendix A.1)
 - 3: **for** $(\vec{v}, \vec{w}) \in G^k$ **do**
 - 4: $\text{HYPA}^{(k)}(\vec{v}, \vec{w}) \leftarrow \Pr(x_{vw} \leq (\vec{v}, \vec{w}) \mid m, \Xi)$
 # Compute CDF
 - 5: **return** HYPA $^{(k)}$
-



Naïve Baseline Comparison

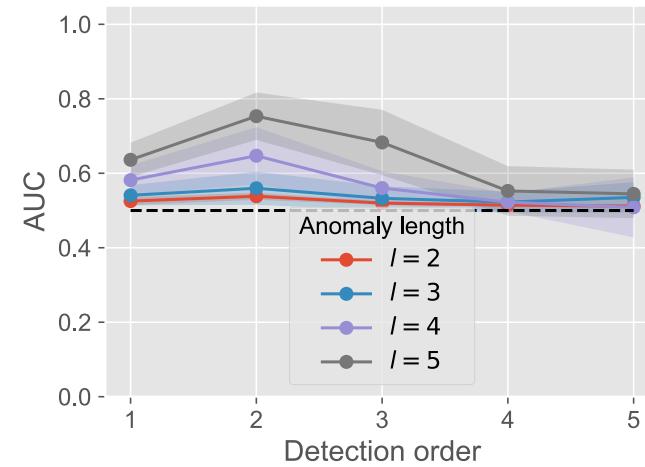
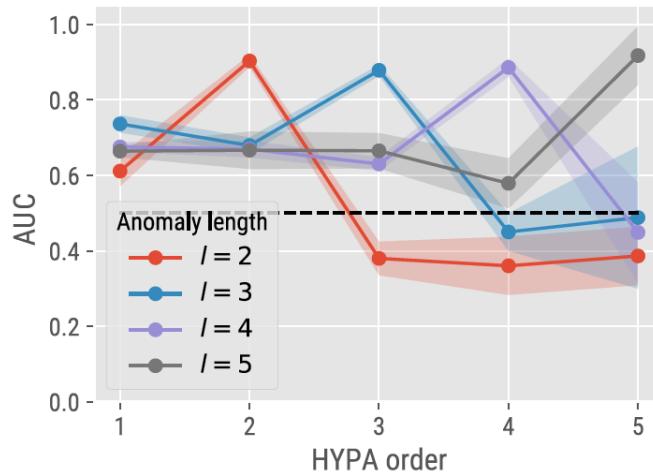
Frequency-Based Anomaly Detection (FBAD)

Compute mean, μ , and standard deviation, σ , of kth order edge weights

Given scaling factor α , label edges as

- Overrepresented if frequency is larger than $\mu + \sigma\alpha$
- Underrepresented if frequency is smaller than $\mu - \sigma\alpha$

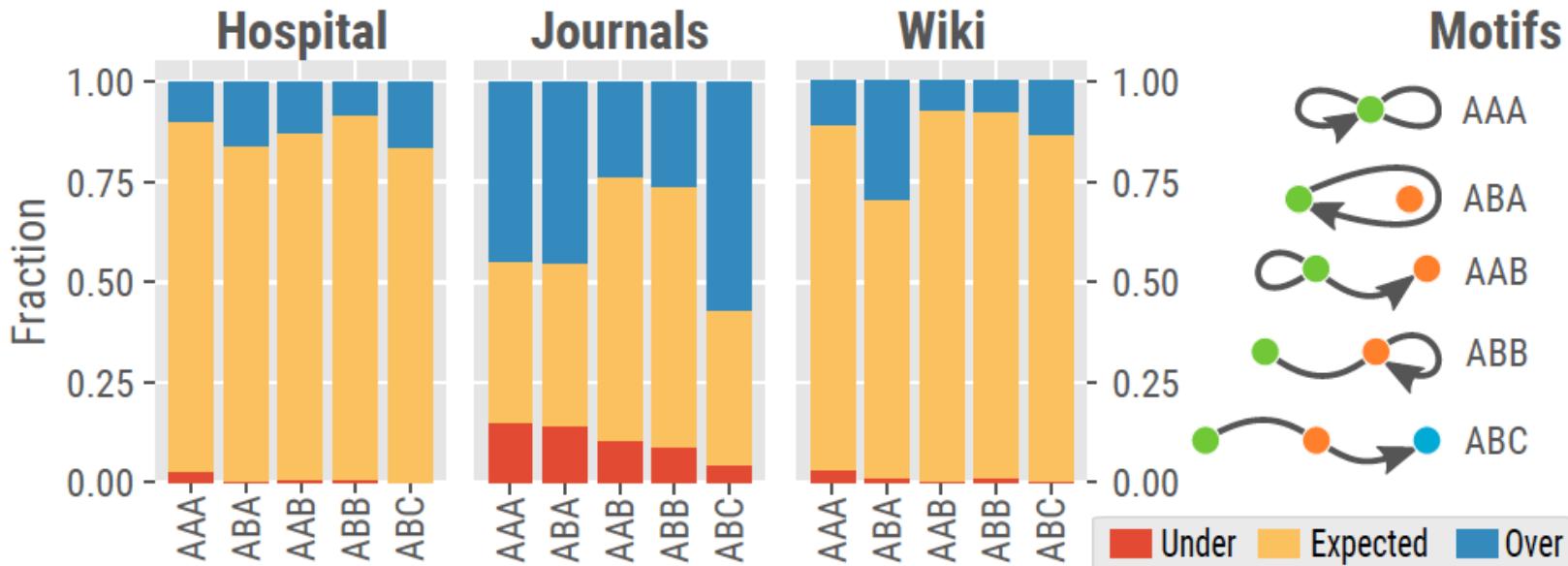
Synthetic Anomalies



Real Data

Data	Topology		Sequences			l^{\max}	$\langle l \rangle$
	Nodes	Edges	Total	Unique			
Tube	268	646	4295731	67015	35	6.75	
Flights	382	6933	185871	88539	10	2.48	
Journals	283	1743	480496	309565	35	14.8	
Hospital	75	1138	28422	2561	5	1.19	
Wiki	100	1598	29682	7431	21	1.64	

Exploring Motifs



Case Study: London Tube

Data:

- Origin → destination statistics between London Tube stations
 - (origin, destination, #observations)
- Shortest paths between stations
 - Assume people follow shortest paths



London Tube

Hypothesis:

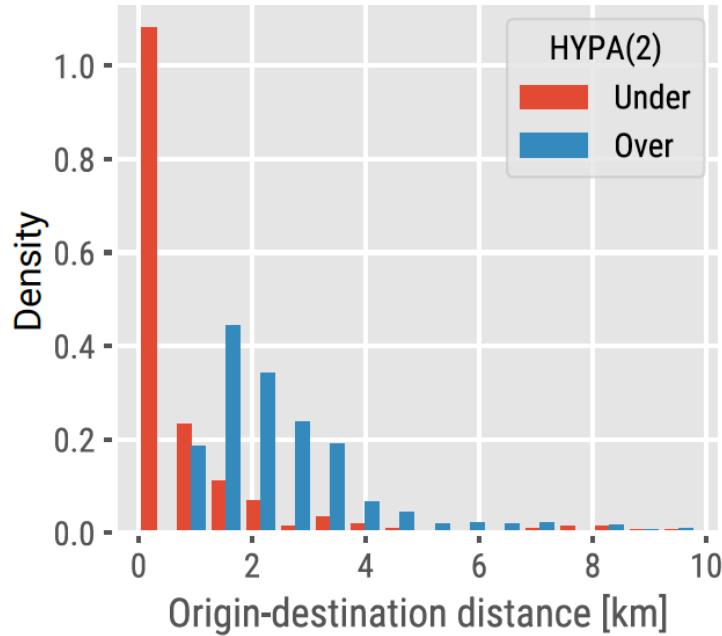
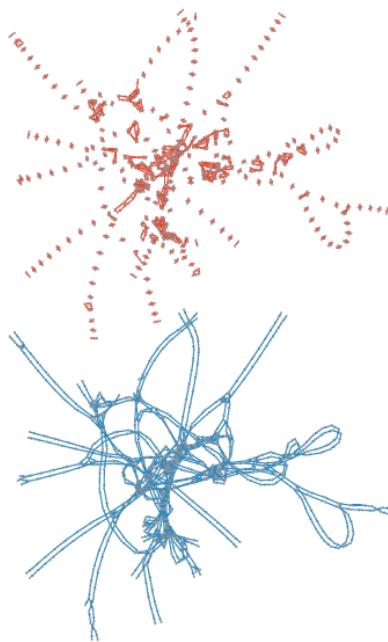
- People typically use public transportation to travel large geographic distances
- *Overrepresented pathways* should cover *larger* distances

Test:

- Measure distance between every station
- For 2nd order transitions A-B-C, compute distance between nodes A and C
- Analyze distributions of distance in over vs. under represented transitions
 - Expect to see distribution shifted towards higher values for over-represented transitions



London Tube



London Tube

HYP_A^(k)	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$
Under [km]	0.00	2.38	3.29	4.60	5.43
Over [km]	2.20	2.93	3.79	5.21	5.63
p-value	$< 10^{-170}$	$< 10^{-7}$	$< 10^{-4}$	0.006	0.08

Median distance between source and destination nodes in under/over represented transitions.

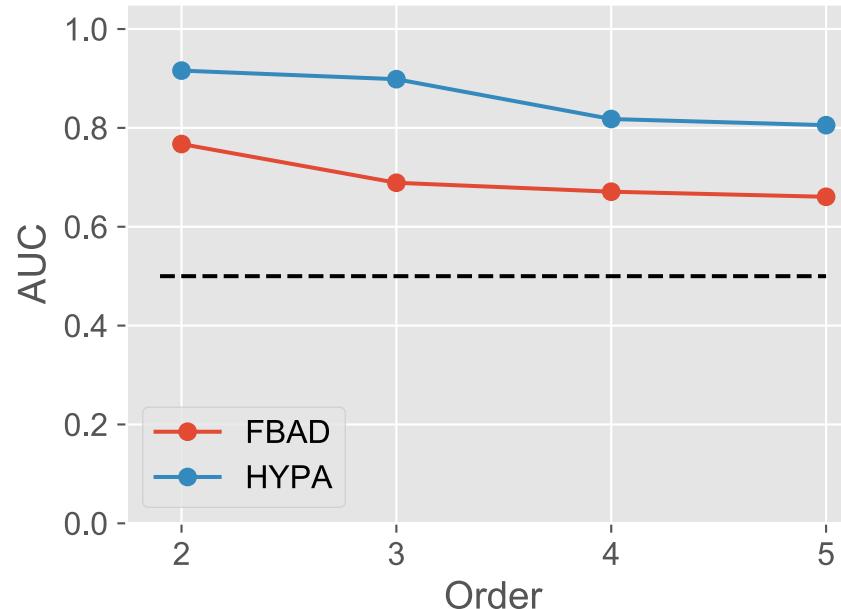
Constructing Ground Truth

Construct ground truth based on the method discussed earlier:

- Randomize path data using $k-1^{\text{st}}$ order random walks
- Compute k^{th} -order path statistics
- Repeat m times, noting the frequency of each path
- Estimate multinomial distribution and its CDF from these statistics
- If $\text{CDF}(\text{path}) > \text{threshold}$, label over-represented



Tube Data - Ground Truth



Computational complexity

$$O(N + |V|^2 \lambda_1^k)$$



Scalability

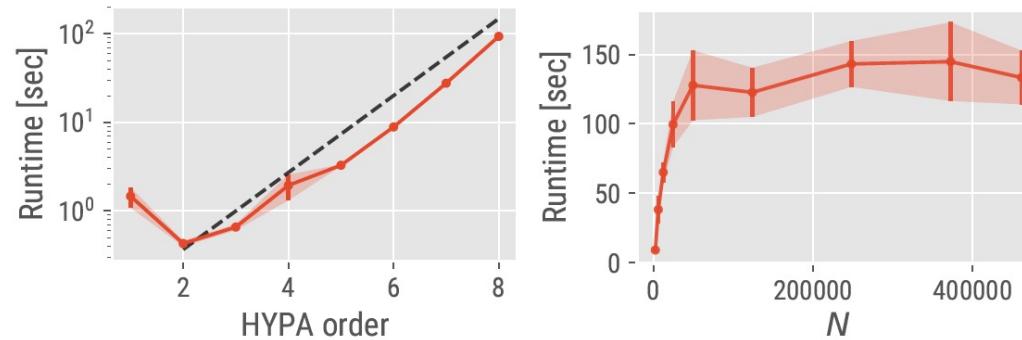


Figure 8: Empirical scalability of HYPA. Left: Required time to detect path anomalies of length k for the Tube data. Right: Runtime in Flights data for detection order $k = 1$ and varying data size N randomly sampled from the data. All data points correspond to the mean of ten repeated measurements, with the standard deviations shown as bars.