Frequency of Significant Sequential Motifs Reveal Patterns in Pathway Data

Presented by

Timothy LaRock

larock.t@northeastern.edu

PhD Candidate in Network Science

Network Science Institute, Northeastern University

In collaboration with

Vahan Nanumyan

Ingo Scholtes

Tina Eliassi-Rad

Northeastern University
Network Science Institute

Motivation

- **Motifs** are meso-scale structures that characterize complex networks. They have long been studied for **static** networks.

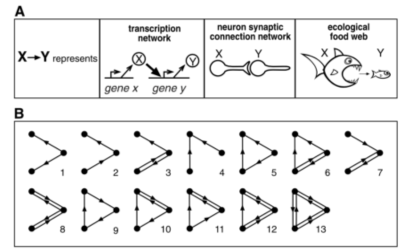


Fig. 1. (**A**) Examples of interactions represented by directed edges between nodes in some of the networks used for the present study. These networks go from the scale of biomolecules (transcription factor protein X binds regulatory DNA regions of a gene to regulate the production rate of protein Y), through cells (neuron X is synaptically connected to neuron Y), to organisms (X feeds on Y). (**B**) All 13 types of three-node connected subgraphs.

25 OCTOBER 2002 VOL 298 SCIENCE www.sciencemag.org

Milo et al. Network Motifs: Building Blocks of Complex Networks. Science, 2002.

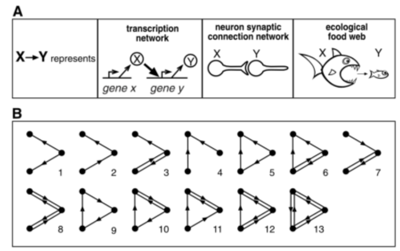
Northeastern University
Network Science Institute

Motivation

- **Motifs** are meso-scale structures that characterize complex networks. They have long been studied for **static** networks.

- More recently, motifs in **temporal** and **dynamic** networks have been analyzed, as well as their role in **dynamics on networks**

Fig. 1. (A) Examples of interactions represented by directed edges between nodes in some of the networks used for the present study. These networks go from the scale of biomolecules (transcription factor protein X binds regulatory DNA regions of a gene to regulate the production rate of protein Y), through cells (neuron X is synaptically connected to neuron Y), to organisms (X feeds on Y). (B) All 13 types of three-node connected subgraphs.



25 OCTOBER 2002 VOL 298 SCIENCE www.sciencemag.org

Milo et al. Network Motifs: Building Blocks of Complex Networks. Science, 2002.

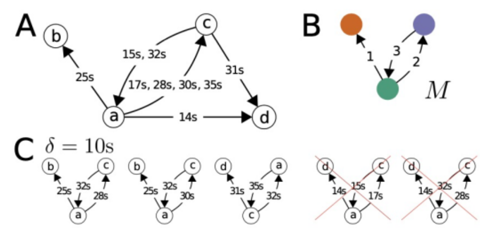


Figure 1: Temporal graphs and δ-temporal motifs. A: A temporal graph with nine temporal edges. Each edge has a timestamp (listed here in seconds). B: Example 3-node, 3-edge δ-temporal motif M. The edge labels correspond to the ordering of the edges. C: Instances of the δ-temporal motif M in the graph for δ = 10 seconds. The crossed-out patterns are not instances of M because either the edge sequence is out of order or the edges do not all occur within the time window δ.

Paranjape et al. Motifs in Temporal Networks. ICWSM, 2017.

Northeastern University
Network Science Institute

Motivation

- **Motifs** are meso-scale structures that characterize complex networks. They have long been studied for **static** networks.

- More recently, motifs in **temporal** and **dynamic** networks have been analyzed, as well as their role in **dynamics on networks**

- **Today:** Network data often comes in the form of **pathways** or **sequences**. This data has different characteristics that require specific methodology.

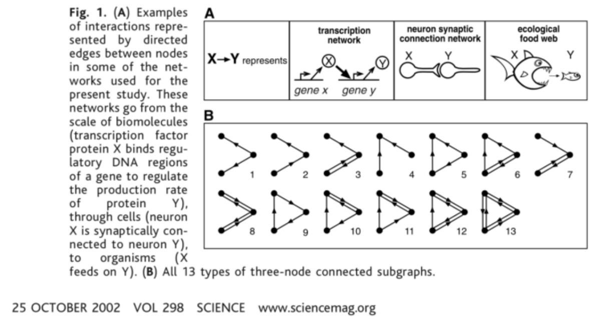


Fig. 1. (**A**) Examples of interactions represented by directed edges between nodes in some of the networks used for the present study. These networks go from the scale of biomolecules (transcription factor protein X binds regulatory DNA regions of a gene to regulate the production rate of protein Y), through cells (neuron X is synaptically connected to neuron Y), to organisms (X feeds on Y). (**B**) All 13 types of three-node connected subgraphs.

25 OCTOBER 2002 VOL 298 SCIENCE www.sciencemag.org

Milo et al. Network Motifs: Building Blocks of Complex Networks. Science, 2002.

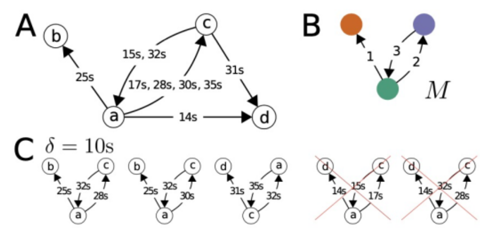
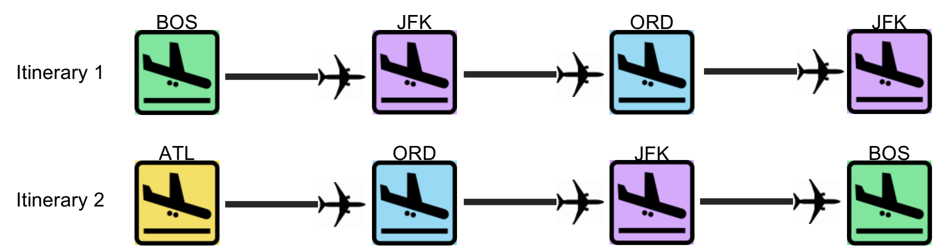


Figure 1: Temporal graphs and δ-temporal motifs. A: A temporal graph with nine temporal edges. Each edge has a timestamp (listed here in seconds). B: Example 3-node, 3-edge δ-temporal motif M. The edge labels correspond to the ordering of the edges. C: Instances of the δ-temporal motif M in the graph for δ = 10 seconds. The crossed-out patterns are not instances of M because either the edge sequence is out of order or the edges do not all occur within the time window δ.

Paranjape et al. Motifs in Temporal Networks. ICWSM, 2017.

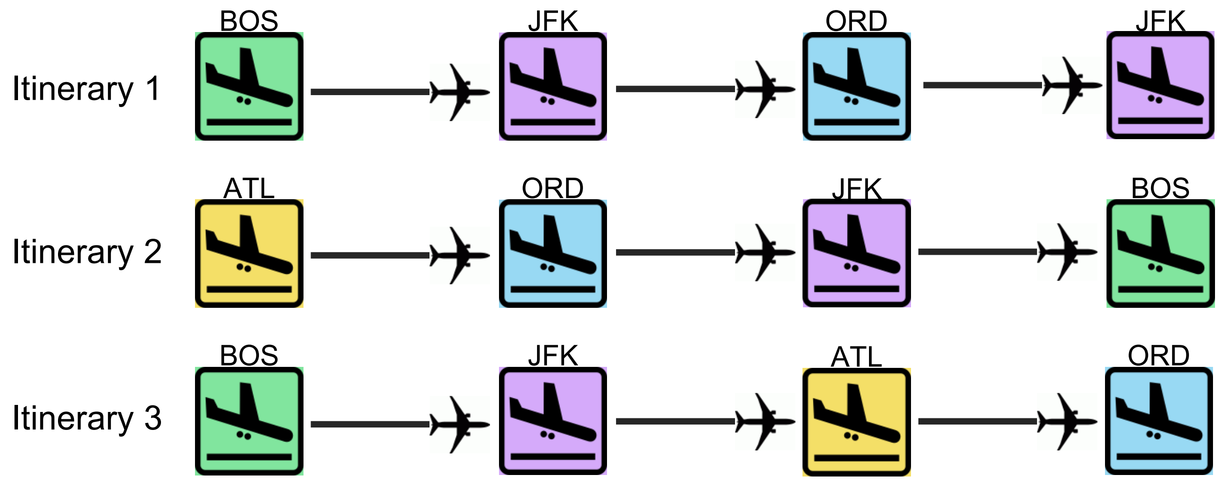


Northeastern University
Network Science Institute
netsi

Motivating Example: Flight Data

Given: Flight itinerary data, where each itinerary is a sequence of airports visited by a passenger on a trip

High-level Goal: Characterize the airport network based on the *frequency* and *significance* of substructures found in the itinerary data

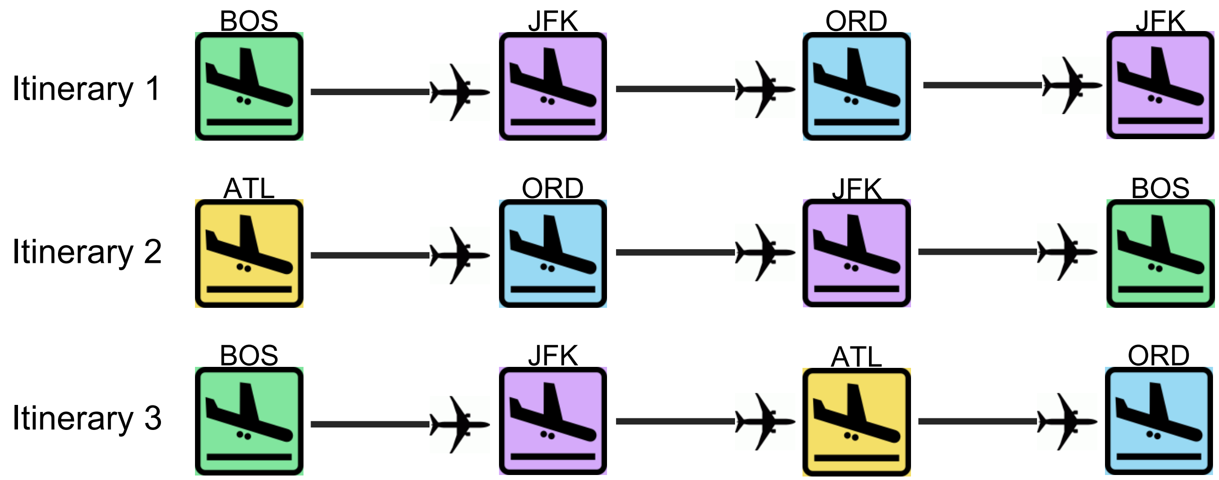


Northeastern University
Network Science Institute

Motivating Example: Flight Data

Given: Flight itinerary data, where each itinerary is a sequence of airports visited by a passenger on a trip

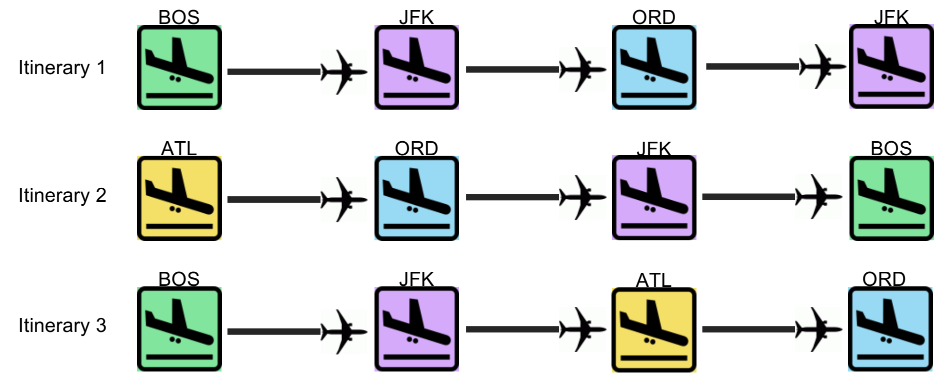
High-level Goal: Characterize the airport network based on the *frequency* and *significance* of substructures found in the itinerary data



Note: We use people traveling through an airport network as motivation, but any network that individual items move through can be studied.

Northeastern University
Network Science Institute

Potential Solution 1

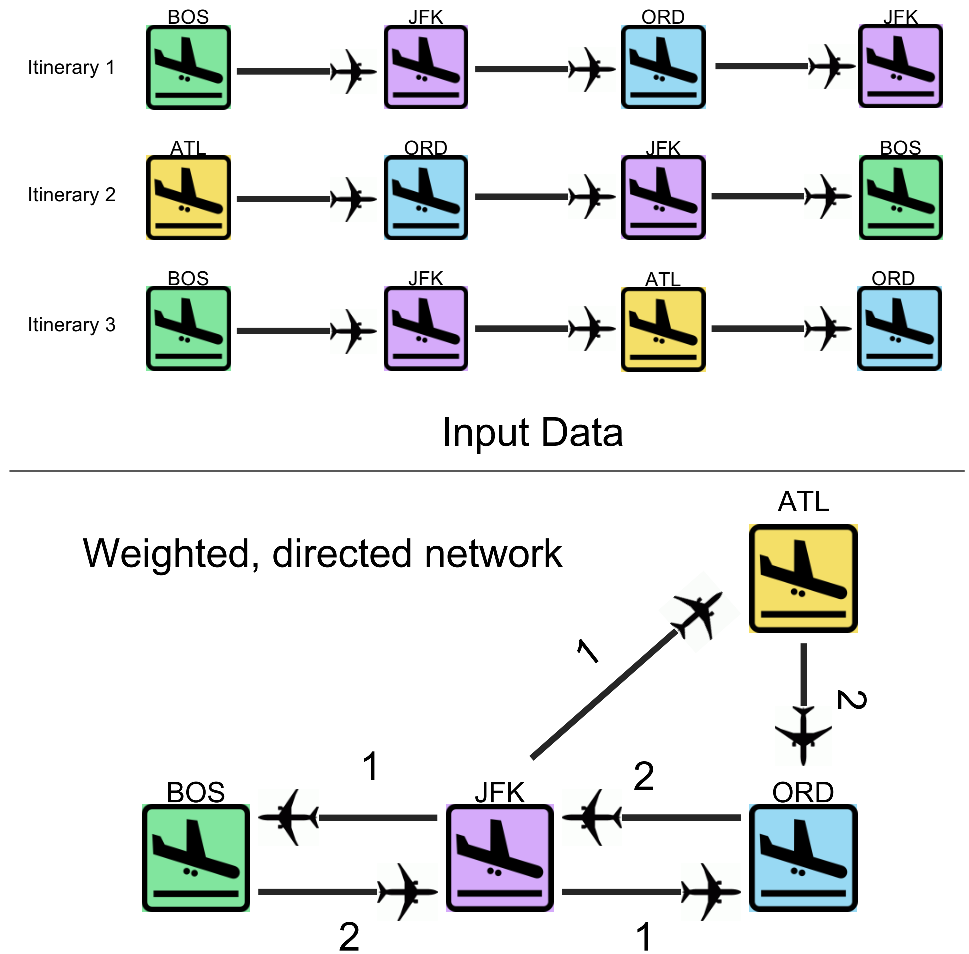


Input Data

Northeastern University
Network Science Institute

Potential Solution 1

Why not construct a weighted and directed network, then count motifs?

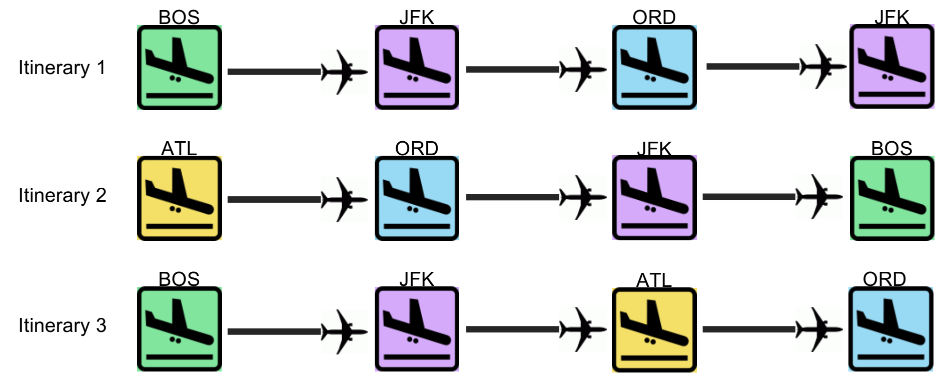


Northeastern University
Network Science Institute

Potential Solution 1

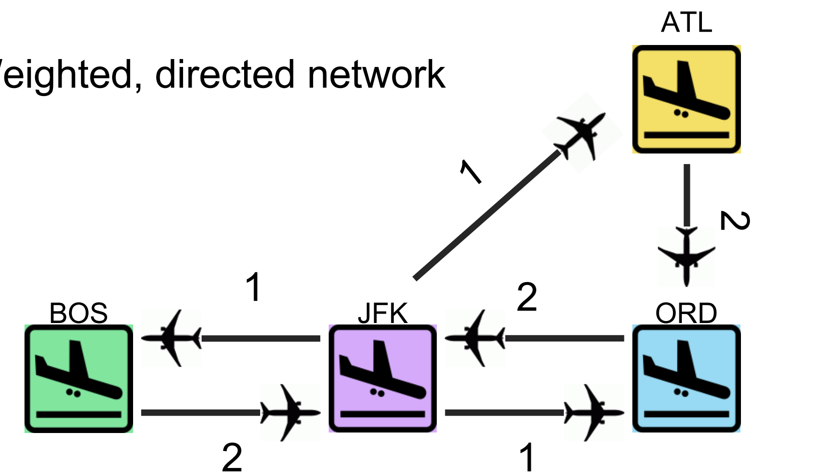
Why not construct a weighted and directed network, then count motifs?

- We have methods to do this, e.g.
  - Milo et al. Network Motifs: Building Blocks of Complex Networks. Science, 2002.
  - Saramäki et al. Characterizing Motifs in Weighted Complex Networks. AIP Conference Proceedings, 2005.
  - Alon. Network Motifs: Theory and Experimental Approaches. Nature Reviews Genetics. 2007.
  - Underwood et al. Motif-based spectral clustering of weighted and directed networks. Applied NetSci, 2020.



Input Data

Weighted, directed network

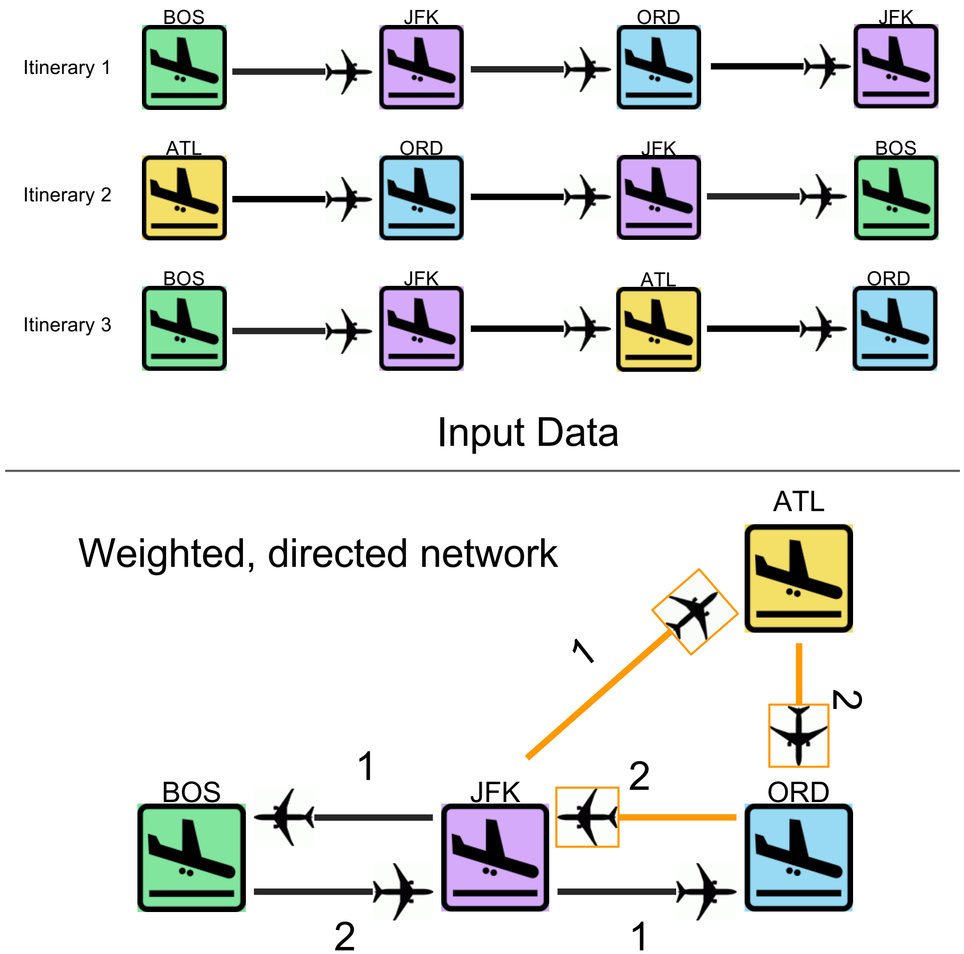


Northeastern University
Network Science Institute

Potential Solution 1

Why not construct a weighted and directed network, then count motifs?

Consider the directed cycle on the 3 edges between JFK, ATL, and ORD.



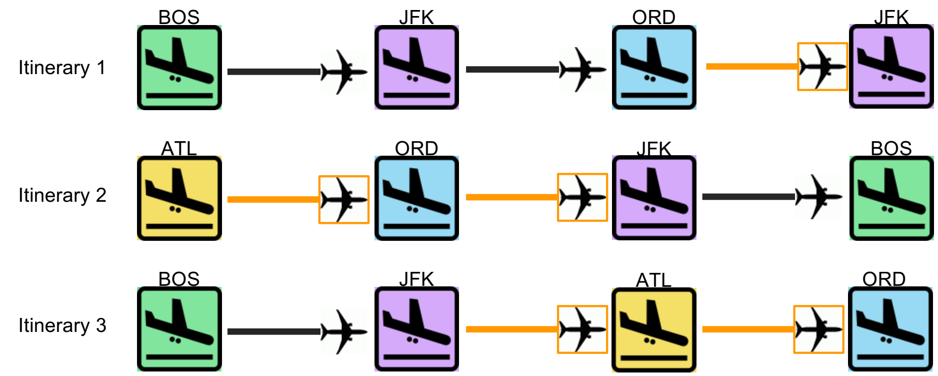
Northeastern University
Network Science Institute

Potential Solution 1

Why not construct a weighted and directed network, then count motifs?

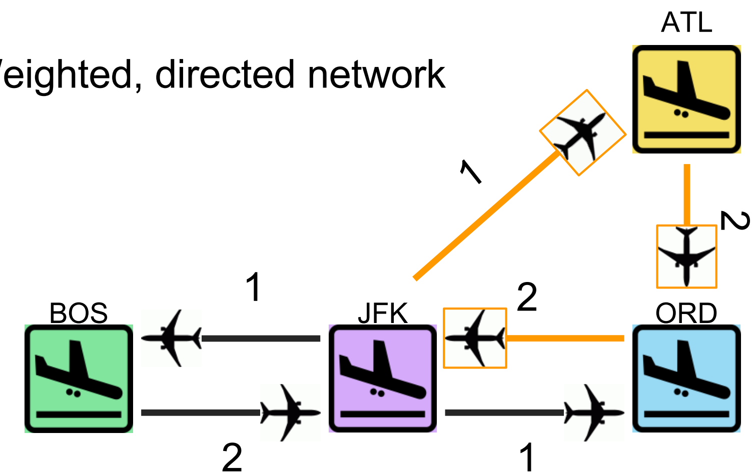
Consider the directed cycle on the 3 edges between JFK, ATL, and ORD.

1. Does this cycle actually appear in the data?



Input Data

Weighted, directed network



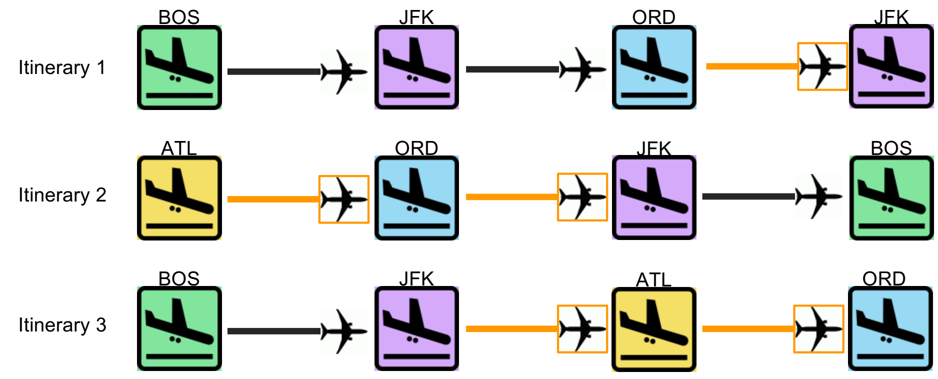
Northeastern University
Network Science Institute

Potential Solution 1

Why not construct a weighted and directed network, then count motifs?

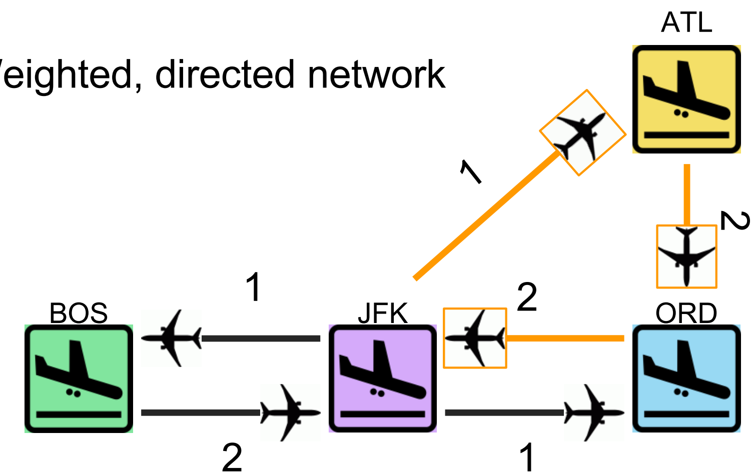
Consider the directed cycle on the 3 edges between JFK, ATL, and ORD.

1. Does this cycle actually appear in the data?
   - No! And we already know this from the original data.



Input Data

Weighted, directed network



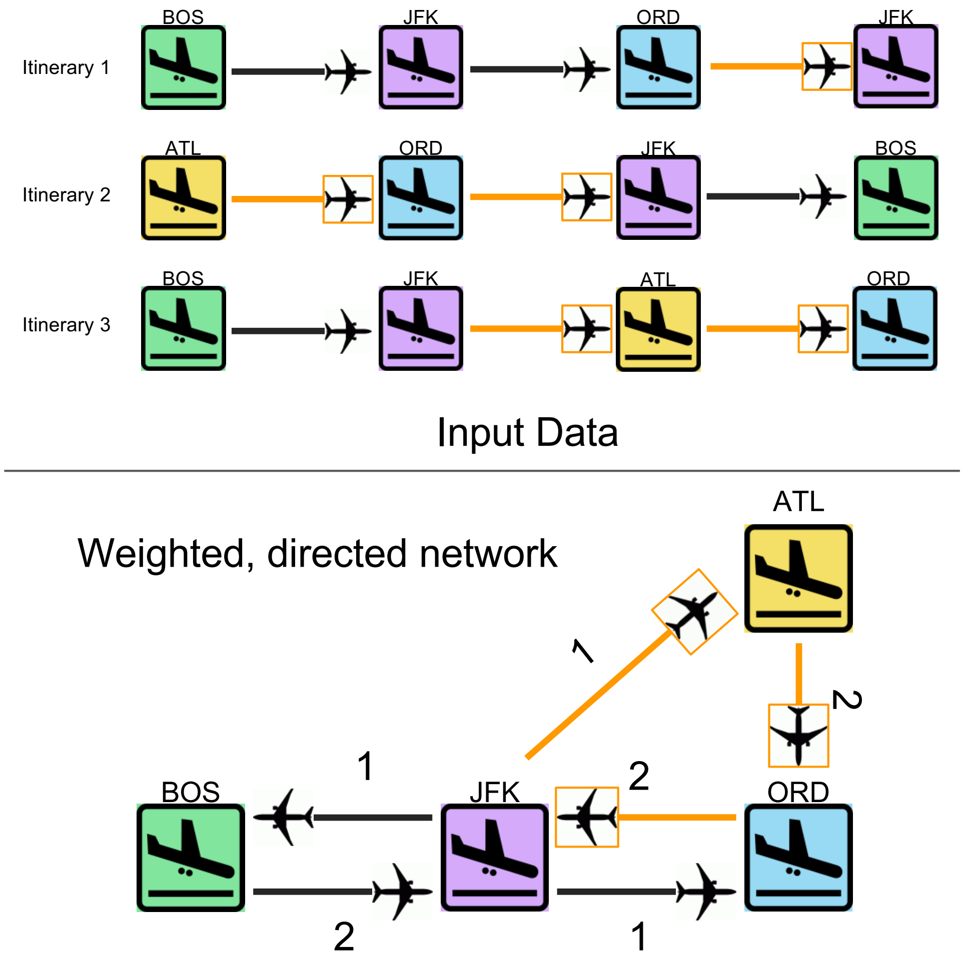
netsi
Northeastern University
Network Science Institute

Potential Solution 1

Why not construct a weighted and directed network, then count motifs?

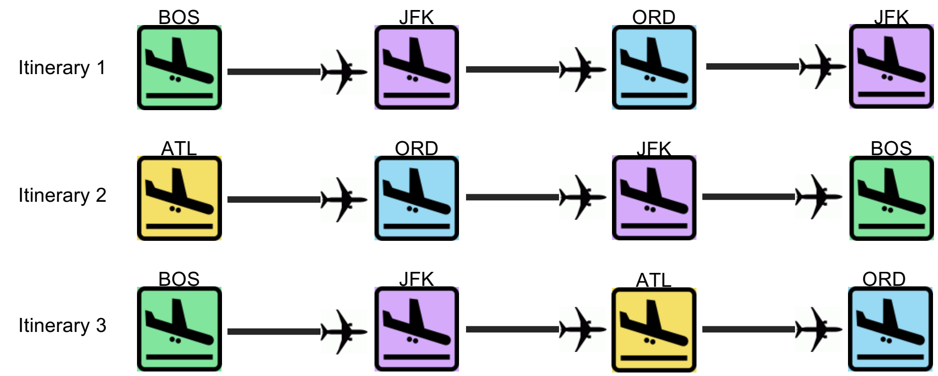
Consider the directed cycle on the 3 edges between JFK, ATL, and ORD.

1. Does this cycle actually appear in the data?
   - No! And we already know this from the original data.
2. How should we include edge weight information?
   - Combining edge weights does not correspond to what we want to count!



Northeastern University
Network Science Institute

Potential Solution 2

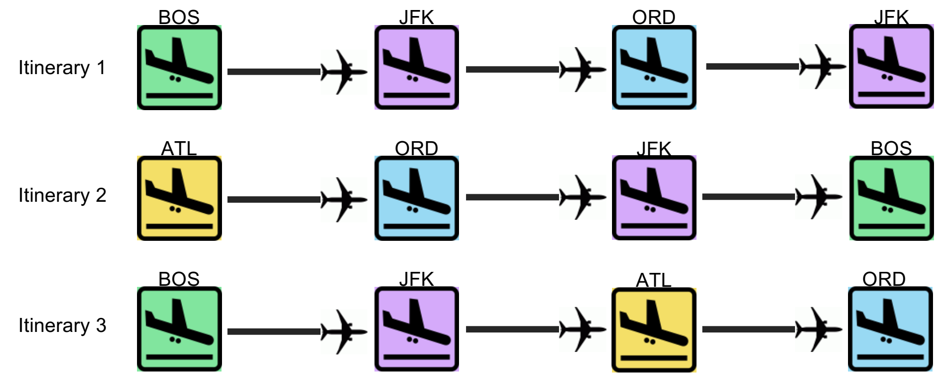


Input Data

Northeastern University
Network Science Institute

Potential Solution 2

Why not adopt methods for temporal and/or dynamic network data?



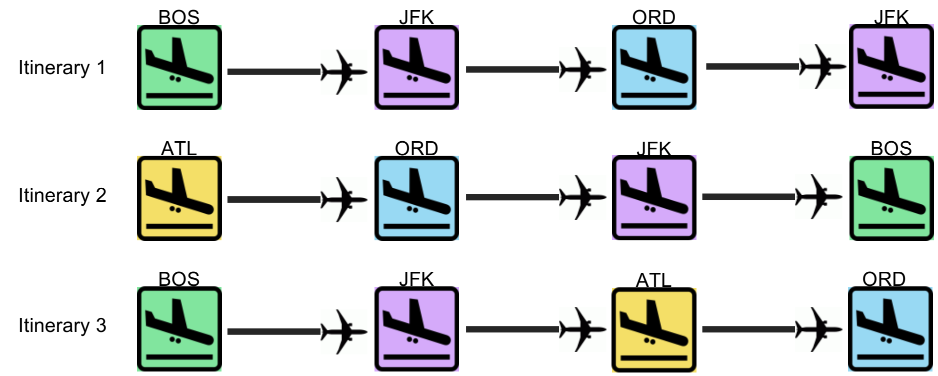
Input Data

Northeastern University
Network Science Institute

Potential Solution 2

Why not adopt methods for temporal and/or dynamic network data?

- Again, we have methods to do this, e.g.
  - Kovanen et al. Temporal Motifs in time-dependent networks. Journal of Stat Mech, 2011.
  - Jurgens et al. Temporal Motifs Reveal the Dynamics of Editor Interactions in Wikipedia. ICWSM, 2012.
  - Kovanen et al. Temporal motifs reveal homophily, gender-specific patterns, and group talk in call sequences. PNAS 2015.
  - Xuan et al. Temporal motifs reveal collaboration patterns in online task-oriented networks. PRE, 2015.
  - Paranjape et al. Motifs in Temporal Networks. ICWSM, 2017.
  - Liu et al. Temporal Network Motifs: Models, Limitations, Evaluation. ArXiv Preprint, 2020.
- But...

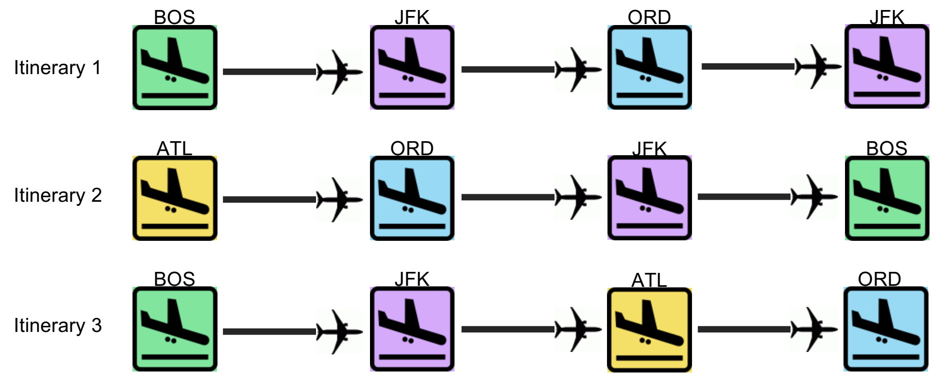


Input Data

Potential Solution 2

Why not adopt methods for temporal and/or dynamic network data?

These types of input data are fundamentally different!



Input Data

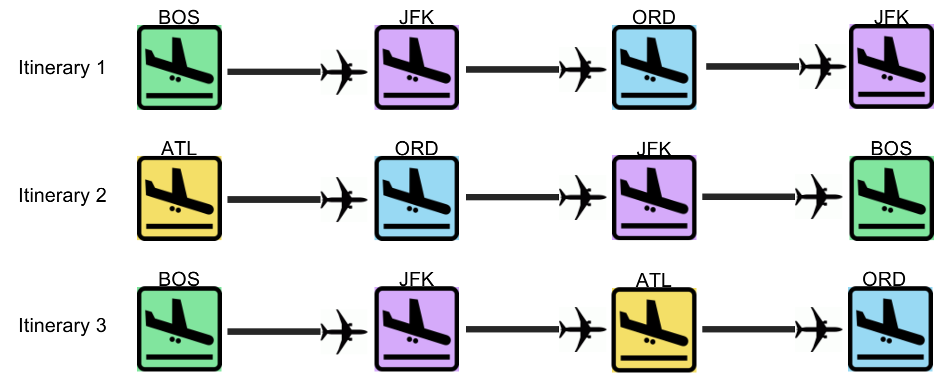
Northeastern University
Network Science Institute

Potential Solution 2

Why not adopt methods for temporal and/or dynamic network data?

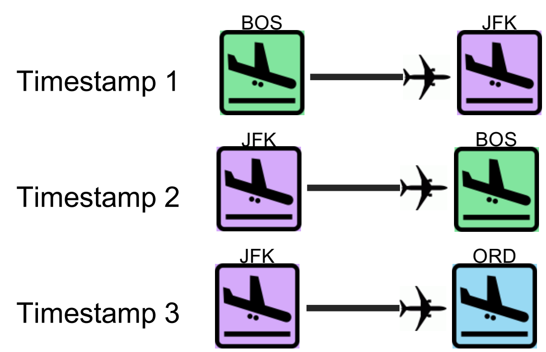
These types of input data are fundamentally different!

- Timestamped edge data is independent observations of individual edge events
    - Path data is **full path observations over multiple edges, typically without timestamps**



Input Data

Temporal Network Data

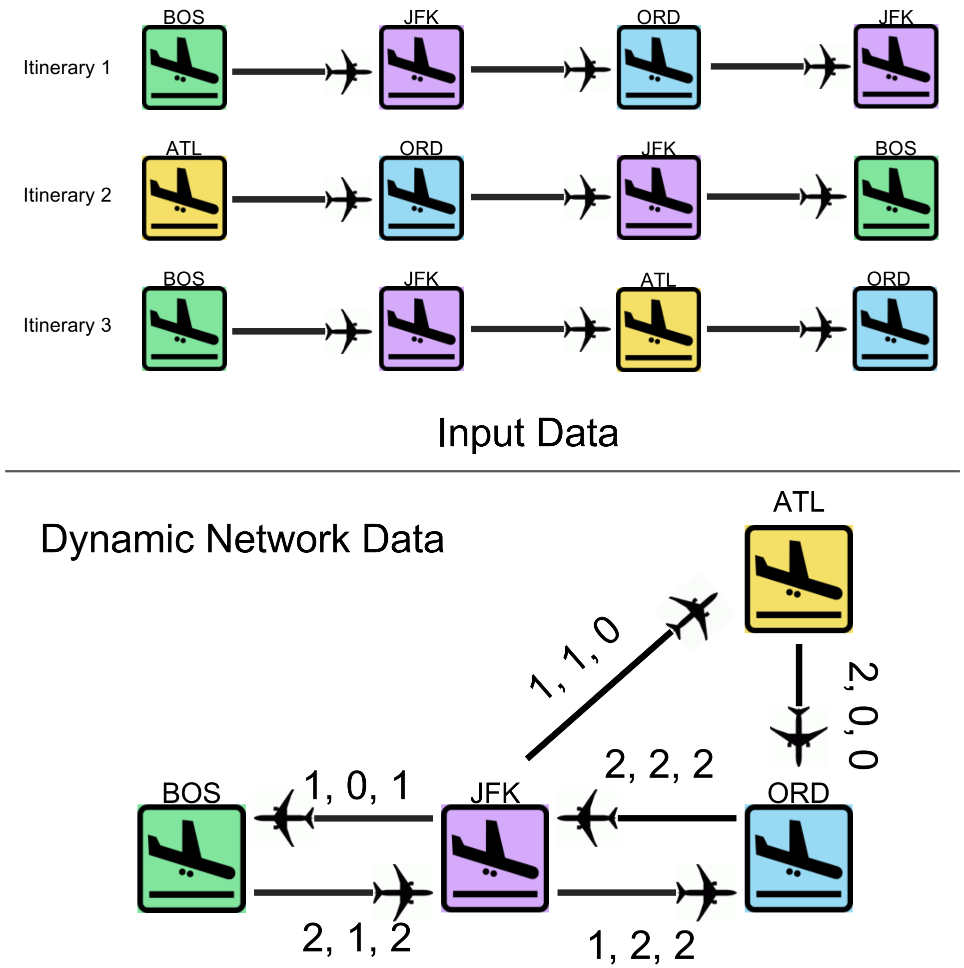


Potential Solution 2

Why not adopt methods for temporal and/or dynamic network data?

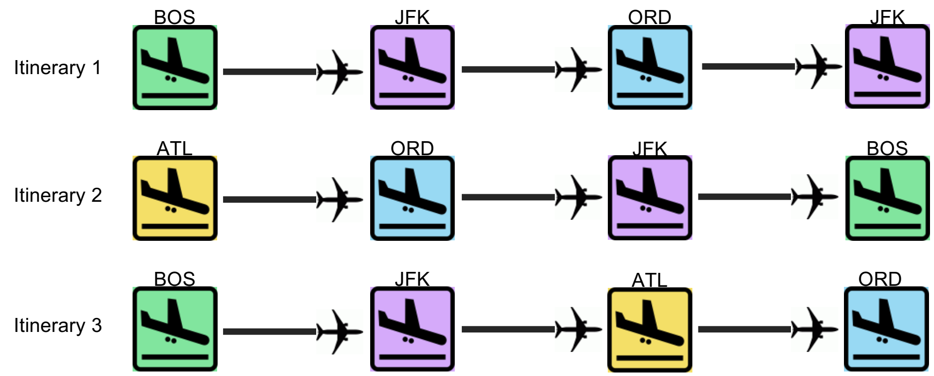
These types of input data are fundamentally different!

- Timestamped edge data is independent observations of individual edge events
  - Path data is **full path observations over multiple edges, typically without timestamps**
- Dynamic network data allows edge deletion between snapshots and assumes some synchronization of dynamics
  - Path data has **no edge deletion and a (typically) unknown and unsynchronized timescale**



Northeastern University
Network Science Institute

Potential Solutions



Input Data

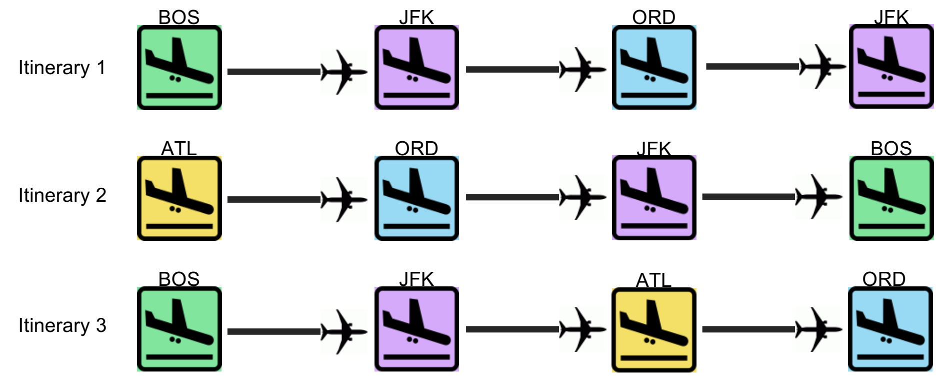
1. **Construct a weighted, directed network and apply existing methods**
   - Existing methods may be misleading

2. **Adopt methods for temporal and/or dynamic data**
   - Pathway data is fundamentally different and requires its own methodology

Potential Solutions

1. **Construct a weighted, directed network and apply existing methods**
   - Existing methods may be misleading

2. **Adopt methods for temporal and/or dynamic data**
   - Pathway data is fundamentally different and requires its own methodology



Input Data

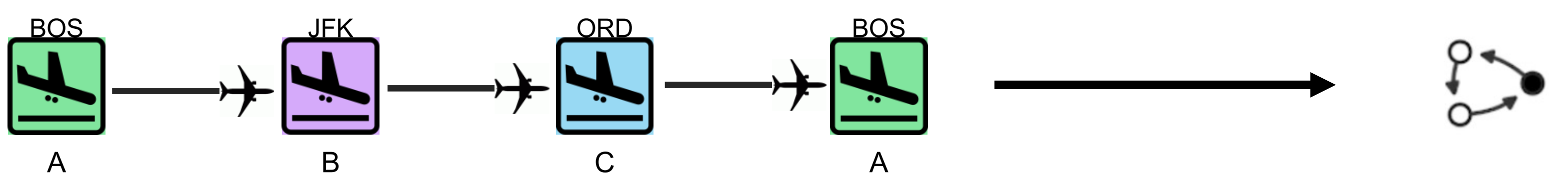
<u>The bottom line:</u> Existing solutions are of limited use in this data, so new methodology is warranted!

Northeastern University
Network Science Institute

Sequential Motifs

Observation: Every *path* or *sequence* on *k* edges observed in a dataset corresponds to a *motif* of length *k*

We can move from specific paths to general motifs using a one-to-one mapping over an alphabet $\Sigma = (\sigma_1, \sigma_2, \ldots, \sigma_\ell)$, e.g. $\Sigma = (A, B, C)$



Northeastern University
Network Science Institute

Advantages of Sequential Motifs

Advantages of Sequential Motifs

- Results in counts of motifs that correspond directly to observed data
  - Does not count motifs based on paths that did not occur in the data

Advantages of Sequential Motifs

- Results in counts of motifs that correspond directly to observed data
  - Does not count motifs based on paths that did not occur in the data

- Weights can be incorporated intuitively
  - For each path that maps to a motif, count the frequency of that path as a (sub)path in the data

Advantages of Sequential Motifs

- Results in counts of motifs that correspond directly to observed data
  - Does not count motifs based on paths that did not occur in the data

- Weights can be incorporated intuitively
  - For each path that maps to a motif, count the frequency of that path as a (sub)path in the data

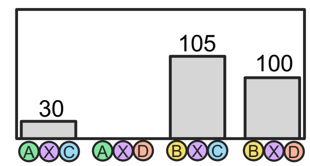
- **Next:** Corresponds to mapping the edges of a DeBruijn Graph to motifs

What is a DeBruijn Graph?

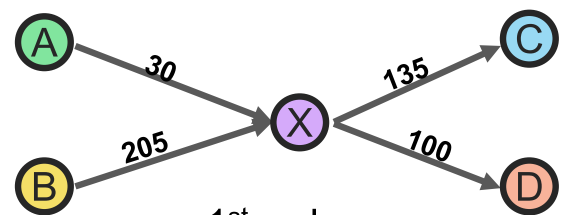
Natural graph representation for path and sequence data

A k^{th} order DeBruijn graph has nodes that represent paths of length $k-1$

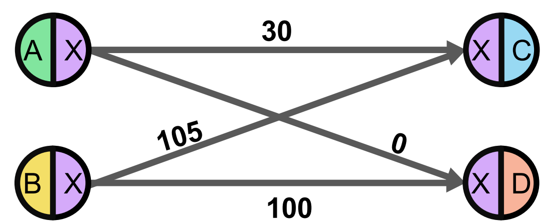
Higher-order nodes can connect if they overlap in $k-1$ first-order nodes, and each edge is a path of length k



Input Paths



1st-order
"DeBruijn" graph



2nd-order
DeBruijn graph

Northeastern University
Network Science Institute

HYPA = Higher Order Hypergeometric Path Anomaly

Why use a DeBruijn graph?

Leverage a null model for weighted DeBruijn graphs, HYPA

HYPA gives a probability of observing a smaller edge (path!) frequency according to a *k-1st* order model of the data

- Need to select threshold value α on this probability
- We use *1/M*, where *M* is the sum of all weights in the DeBruijn graph

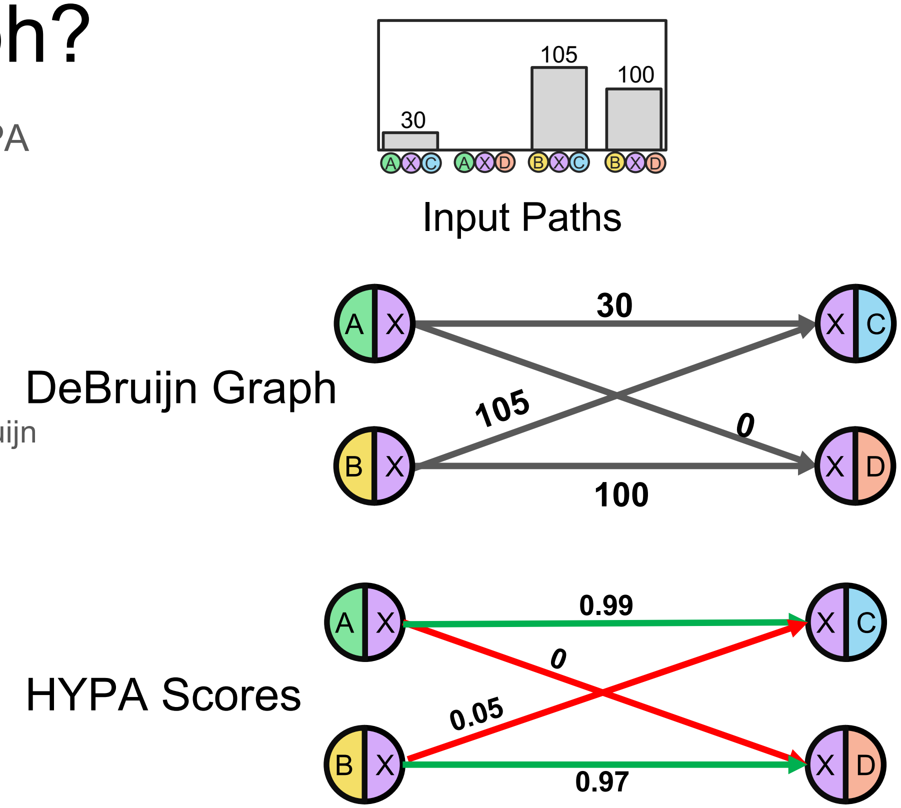
$$\mathrm{HYPA}^{(k)}(\vec{v}, \vec{w}) := \Pr\left(X_{\vec{v}\,\vec{w}} \leq f(\vec{v}, \vec{w})\right)$$

If $< \boldsymbol{\alpha}$ then the pathway is **underrepresented**.

If $\geq 1\text{-}\boldsymbol{\alpha}$, then pathway is **overrepresented**.

For each motif, we keep track of the

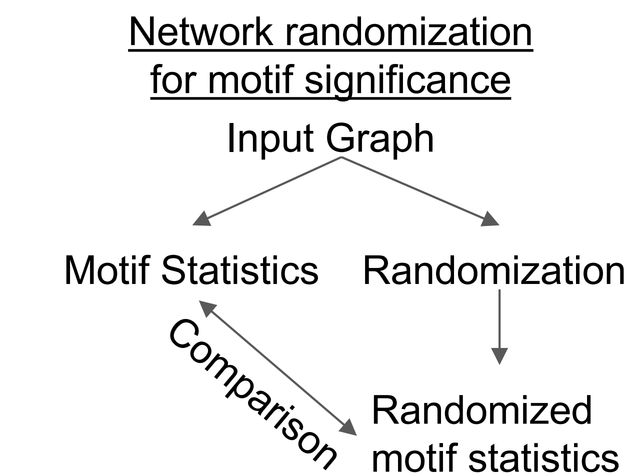
1. **total** number of edges that map to the motif
2. number of mapped edges that are **over** represented
3. Number of mapped edges that are **under** represented



LaRock et al. HYPA: Efficient Detection of Path Anomalies in Time Series Data on Networks. SDM, 2020. https://doi.org/10.1137/1.9781611976236.52

Some Subtlety in Significance

The method presented here **does not** measure the significance of motif frequency compared to a randomization of the first-order network structure (p-value per motif)

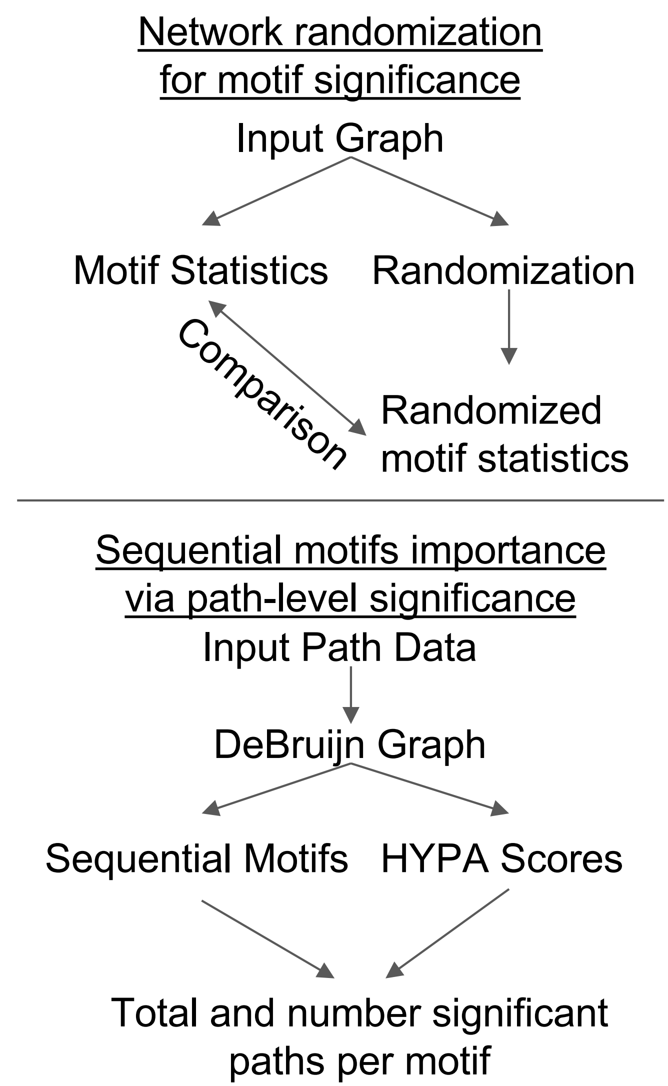


netsi Northeastern University
Network Science Institute

Some Subtlety in Significance

The method presented here **does not** measure the significance of motif frequency compared to a randomization of the first-order network structure (p-value per motif)

Instead, for each motif we count the **number of paths** mapping to that motif that are significantly over or under represented compared to a path-based null model (p-value per path, aggregated for each motif)

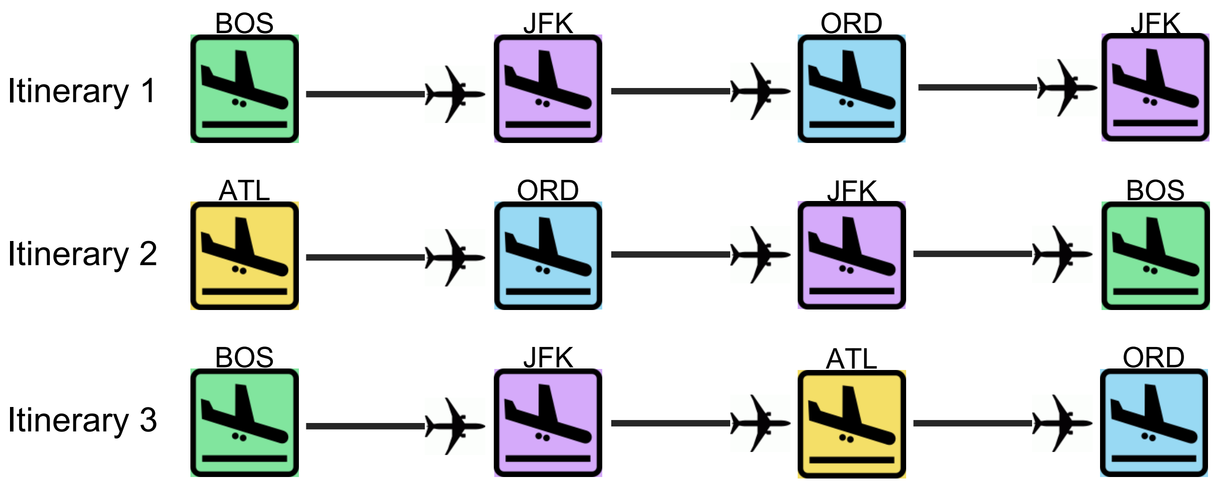


netsi Northeastern University
Network Science Institute

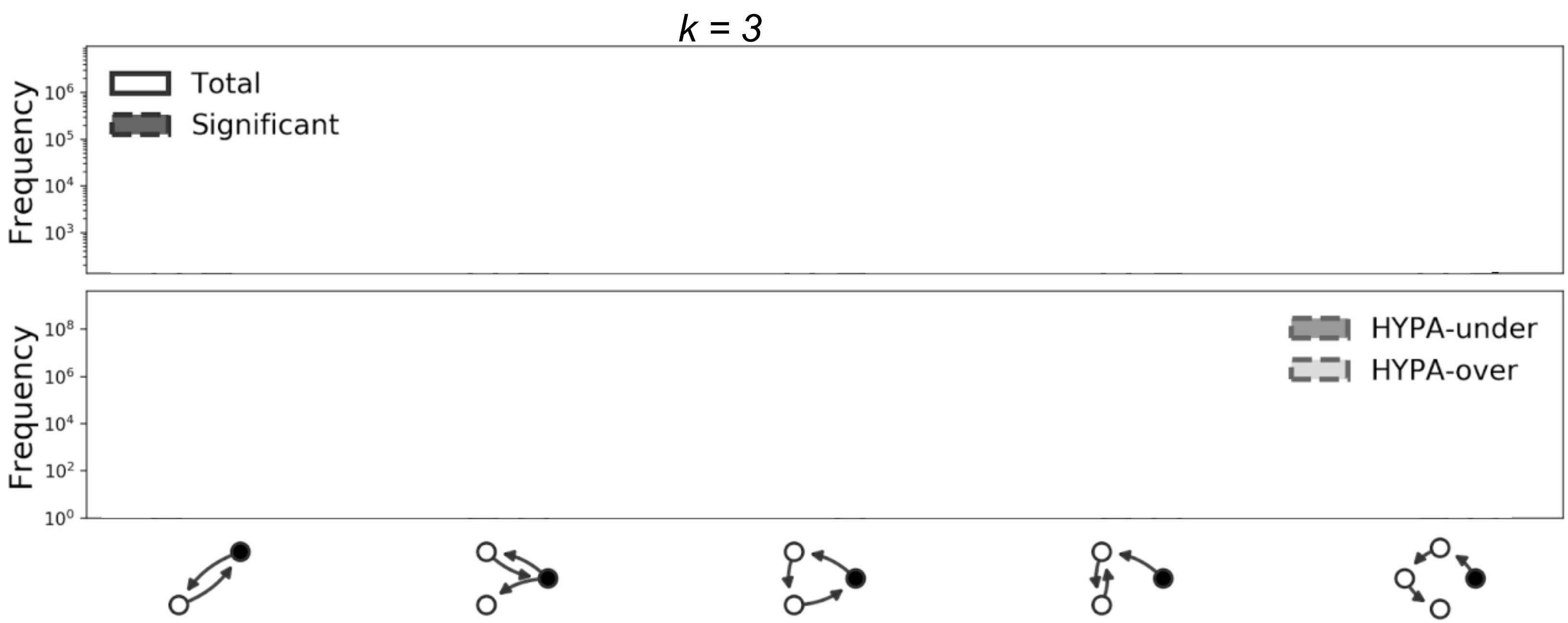
Dataset: Domestic Flights in the US

US passenger flight itineraries downloaded from https://transtats.bts.gov/

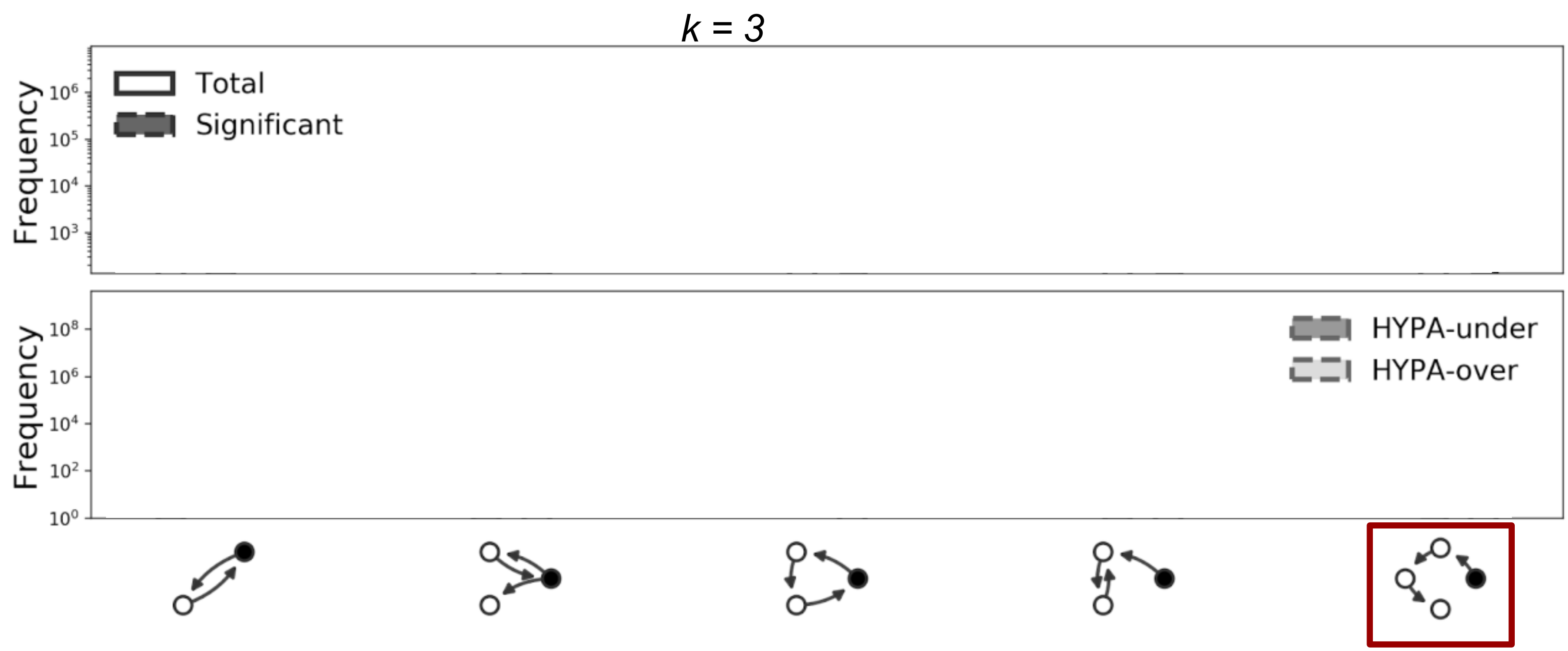
Data from quarter 1 of 2019 includes 3,981,589 itineraries with maximum length 12



Sequential Motifs: Flights

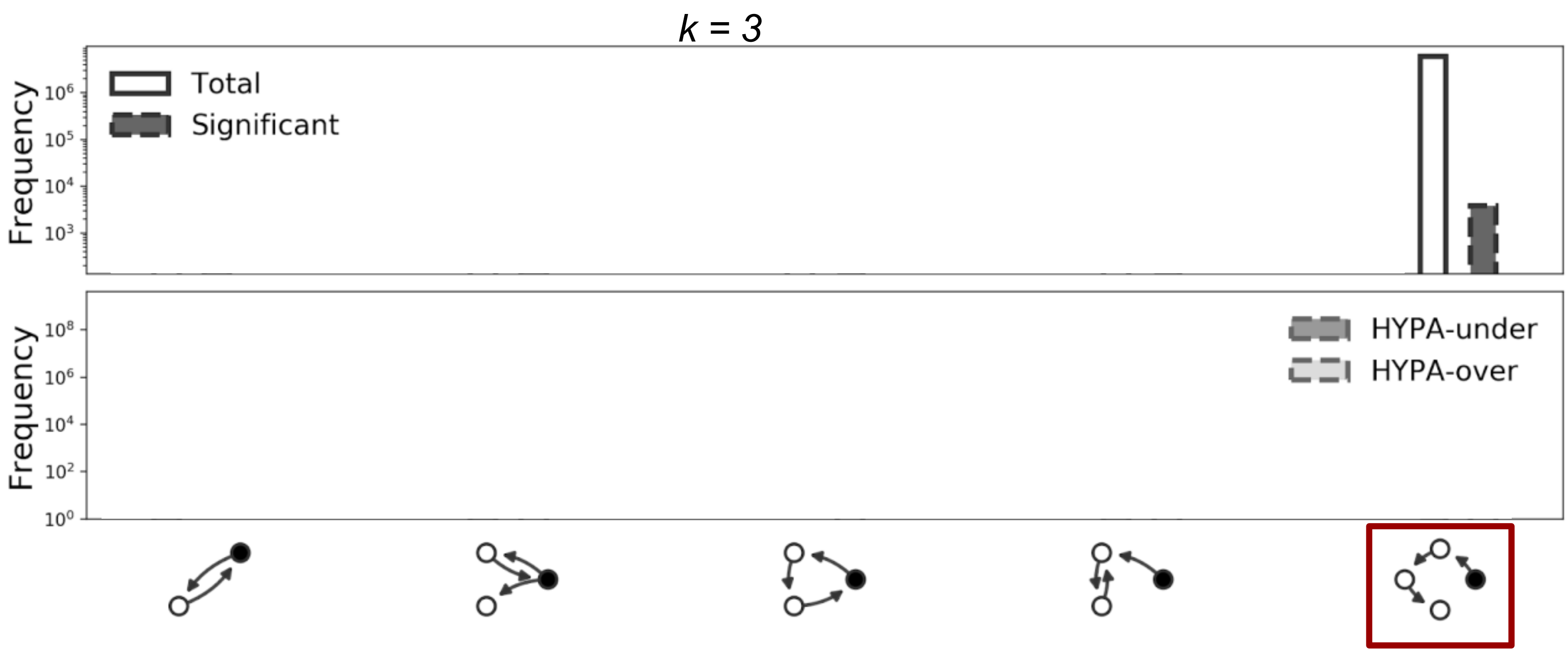


Sequential Motifs: Flights

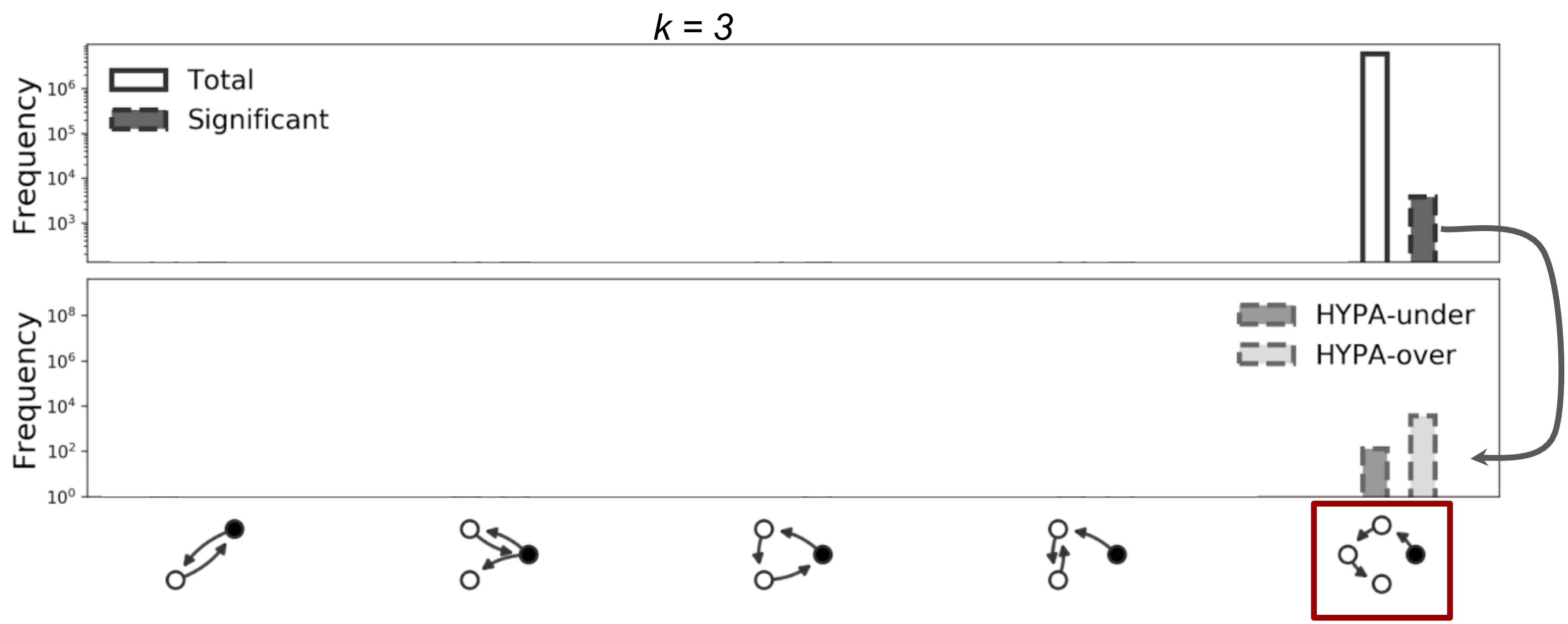


Northeastern University
Network Science Institute

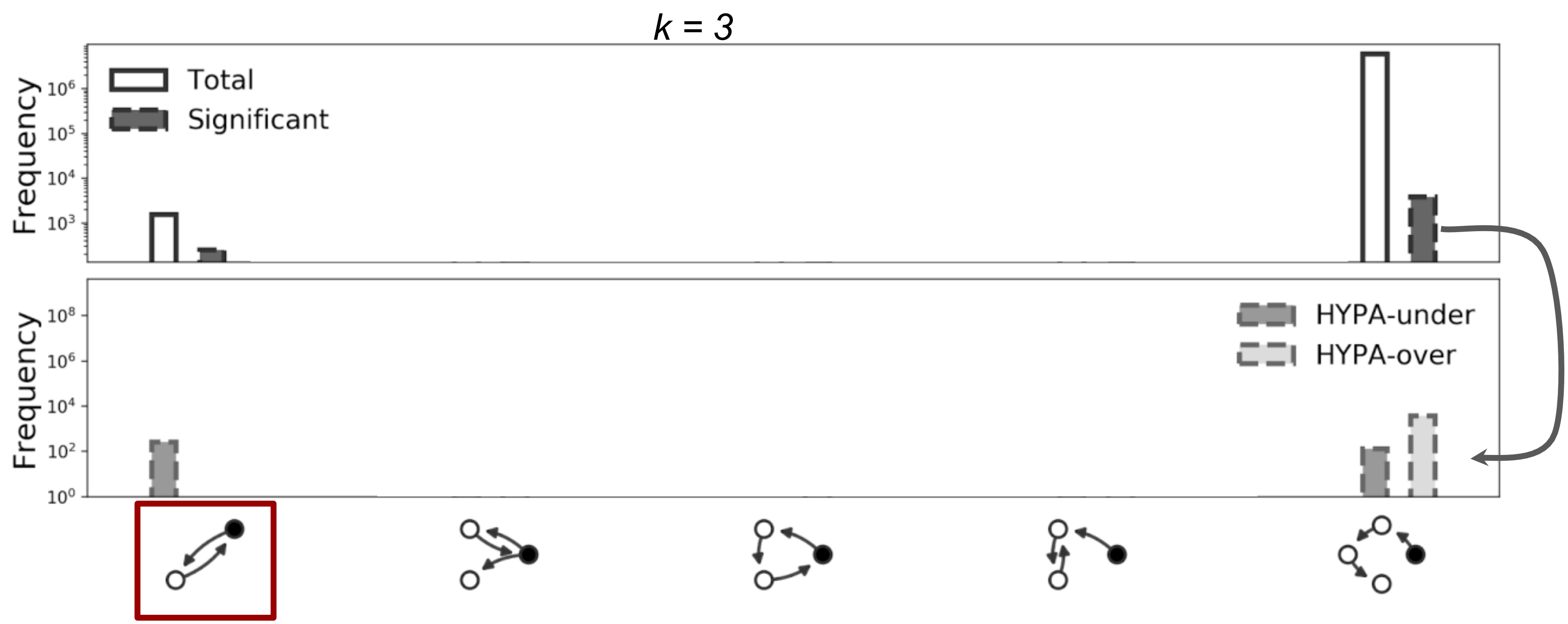
Sequential Motifs: Flights



Sequential Motifs: Flights

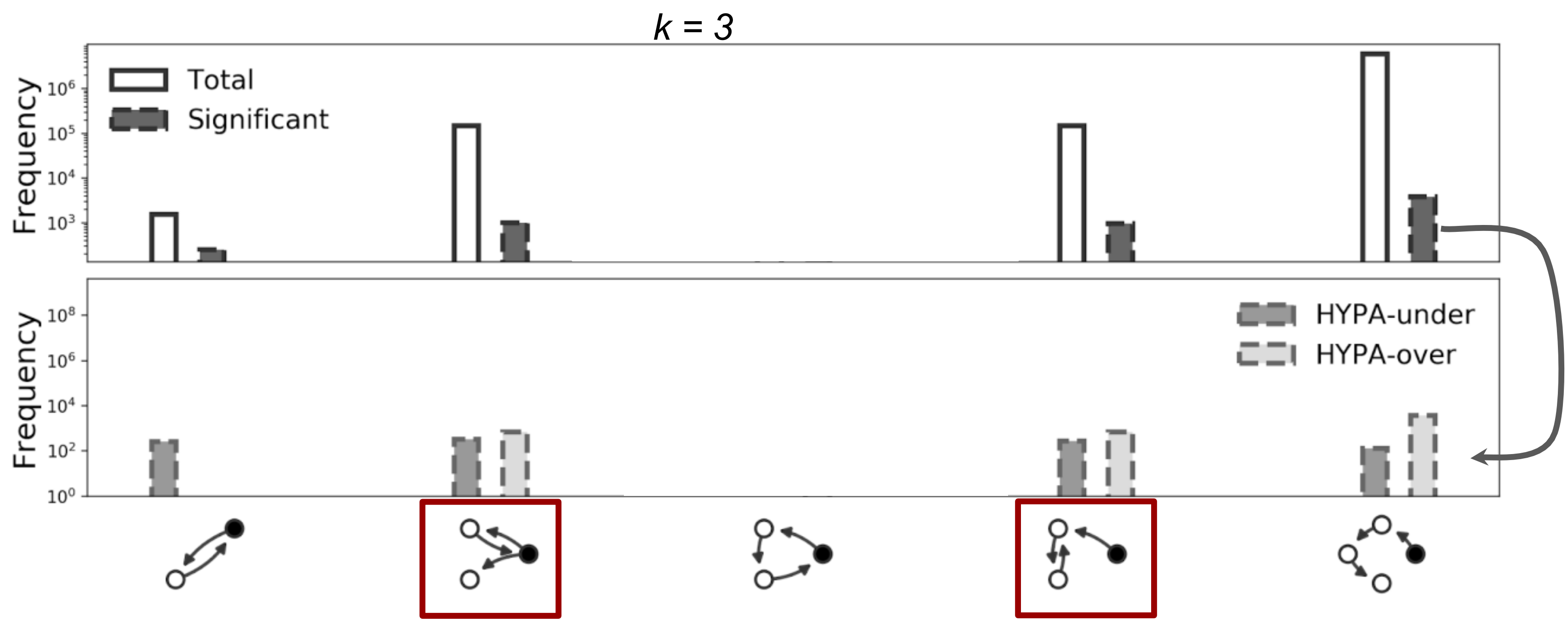


Sequential Motifs: Flights

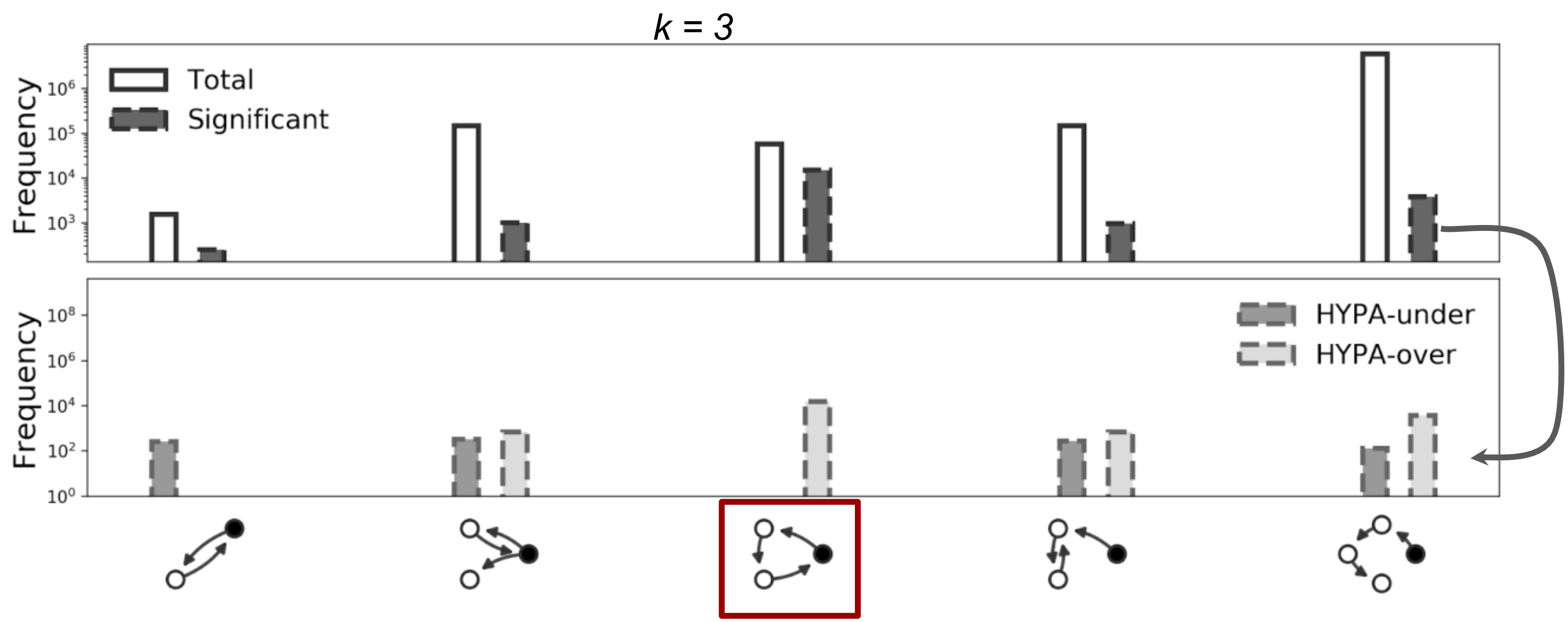


Northeastern University
Network Science Institute

Sequential Motifs: Flights



Sequential Motifs: Flights



Northeastern University
Network Science Institute

Conclusions

- Existing motif counting methodology does not apply directly to pathway data
- Sequential motifs via DeBruijn graphs provide motif counting and an indication of significance via HYPA
- Analysis of sequential motifs in flight data identifies patterns that correspond to expectations

---

Thank you!

Timothy LaRock
larock.t@northeastern.edu
tlarock.github.io

Preprint to come very soon!

Northeastern University
Network Science Institute