Code available on GitHub

Community Detection on Higher-Order Networks: Identifying Patterns in US Air Traffic

Bradley Dice[1] **Daniel McCusker**[2] **Shannon Moran**[3]

[1]*Department of Physics, Scientific Computing Program*
[2]*Applied Physics Program*
[3]*Department of Chemical Engineering, Scientific Computing Program*

EECS 598-008, W19
Advanced Data Mining

Problem Definition

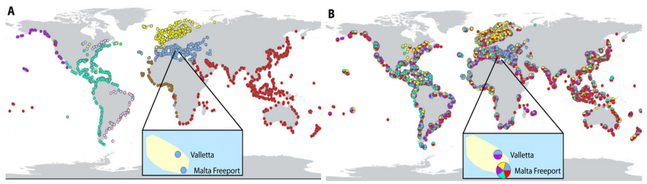
Network analyses usually assume *first-order* dynamic processes. *Higher-order* networks account for path histories, but this can be computationally complex.

INPUT: Quarterly flight ticket data from 1993-2018.

OUTPUT: Airport "communities" from a higher-order network (HON) representation of the underlying airline traffic.

Motivation

Community detection on a transportation network HON (e.g. shipping routes) reveals structure beyond geographic proximity.



Xu, J., Wickramarathne, T. L., & Chawla, N. V. (2016). Representing higher-order dependencies in networks. *Science Advances*, *2*(5), e1600028.

Data sets and management

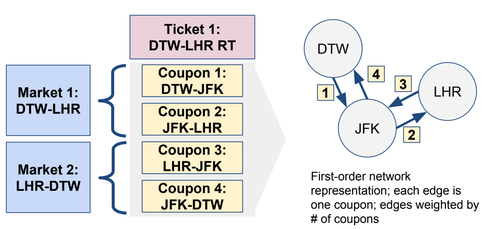
signac.io

jupyter pandas

APACHE Spark

Data from Bureau of Transportation Statistics

- DB1B flight tickets quarterly, (1993-2018)
  - 837 million coupon observations
  - 507 million source/origin observations
  - 286 million ticket observations



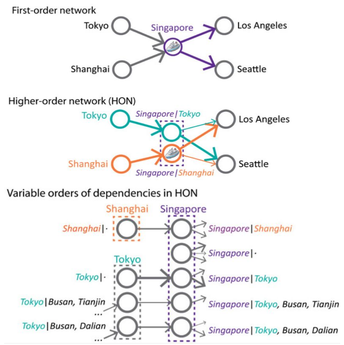
First-order network representation; each edge is one coupon; edges weighted by # of coupons

Data management & workflow automation via *signac*

- The *signac* data management framework is designed to accelerate file-based workflows like this project
- A series of pre- and post-conditions are used to define a workflow of operations: cleaning, data extraction, analysis, and visualization are all automated

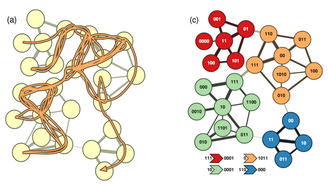
Our Approach

Step 1: Build higher order networks (HONs) from coupon data



- Flight trajectories between airports ("physical nodes") are extracted from CSVs of coupon data using PySpark
- HON algorithm produces a network of variable-order "state nodes" by proposing transition rules

Step 2: Detect communities on the HON using InfoMap

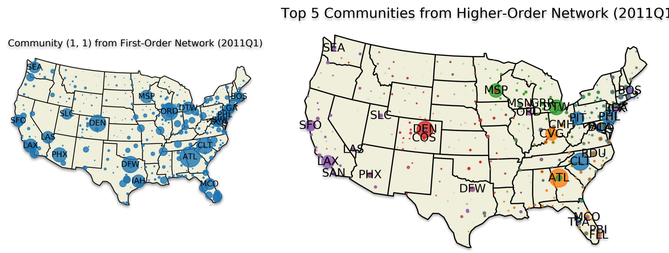


- InfoMap models a random walker process on the HON's state nodes (a "memory network")
- A random walker tends to stay within a community
- Merge state nodes back to physical nodes

Experimental Results

Result 1: HONs accurately account for regional airline communities

Node colors indicate communities. Sizes correspond to flow volume. Labels indicate three letter IATA airport codes.

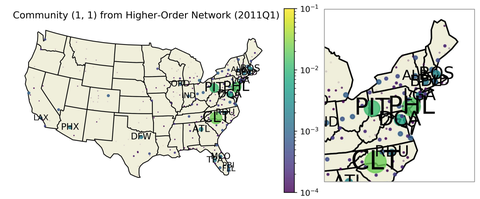


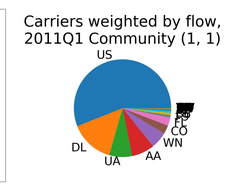
HON community detection identifies multiple clusters with meaningful geographic similarities.

- Running InfoMap on a first order network finds one large community (left) and two tiny communities in Alaska (not pictured)
- For HONs (right), fewer key nodes are identified, with communities built around them

Result 2: HON communities identified using community detection correspond to real-world networks

Node color and size indicates **PageRank** (see colorbar). Labels indicate three letter IATA airport codes.



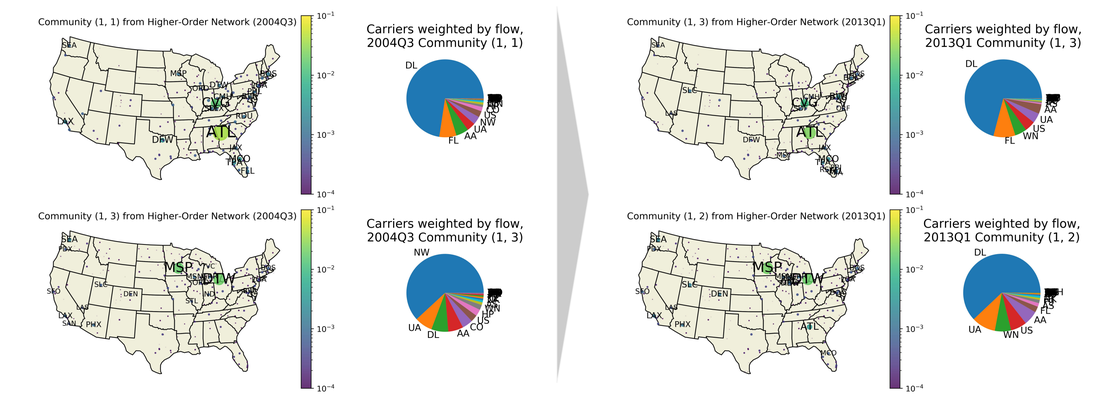


As an example, the largest community for 2011Q1 is highlighted, representing hubs for US Airways.

- At the time, US Airways was the largest airline in the US (until its 2015 merger with American)
- Its hubs included PIT, PHL, and CLT (inset)

Result 3: HON community detection highlights airline mergers and reveals preserved underlying route structure, even after integration is completed

Node color and size indicates **flow in the community** (see colorbar). Labels indicate three letter IATA airport codes.



Delta (DL) and Northwest (NW) officially merged in 2010Q1, when NW tickets moved to the DL call sign.

- Shown are data for 2004 (before) and 2013 (after) the merger
- Interestingly, we see that the underlying structure in the flight tickets is preserved
- This suggests that airline mergers do not significantly change the underlying transportation network structure

References

- Xu, J., Wickramarathne, T. L., & Chawla, N. V. (2016). Representing higher-order dependencies in networks. *Science Advances*, *2*(5), e1600028. https://doi.org/10.1126/sciadv.1600028
- Rosvall, M., Esquivel, A. V, Lancichinetti, A., West, J. D., & Lambiotte, R. (2014). Memory in network flows and its effects on spreading dynamics and community detection. *Nature Communications*, *5*, 4630. https://doi.org/10.1038/ncomms5630
- Edler, D., Bohlin, L., & Rosvall, M. (2017). Mapping Higher-Order Network Flows in Memory and Multilayer Networks with Infomap. Algorithms, 10(4), 112. https://doi.org/10.3390/a10040112
- Rosvall, M., Axelsson, D., & Bergstrom, C. T. (2009). The map equation. *Eur. Phys. J. Special Topics*, *178*, 13. https://doi.org/10.1140/epjst/e2010-01179-1

Conclusions

- **Higher-order networks** reveal network structure and dynamics not apparent in a first-order representation
- Using **community detection** on higher-order networks, we found **interpretable features** of the airline network in the continental United States