Representing higher-order dependencies in networks

and how it improves the result of a variety of tasks such as random walking, clustering, and ranking

Jian Xu (jxu5@nd.edu) Thanuka Wickramarathne (twickram@nd.edu) Nitesh Chawla (nchawla@nd.edu, corresponding author)

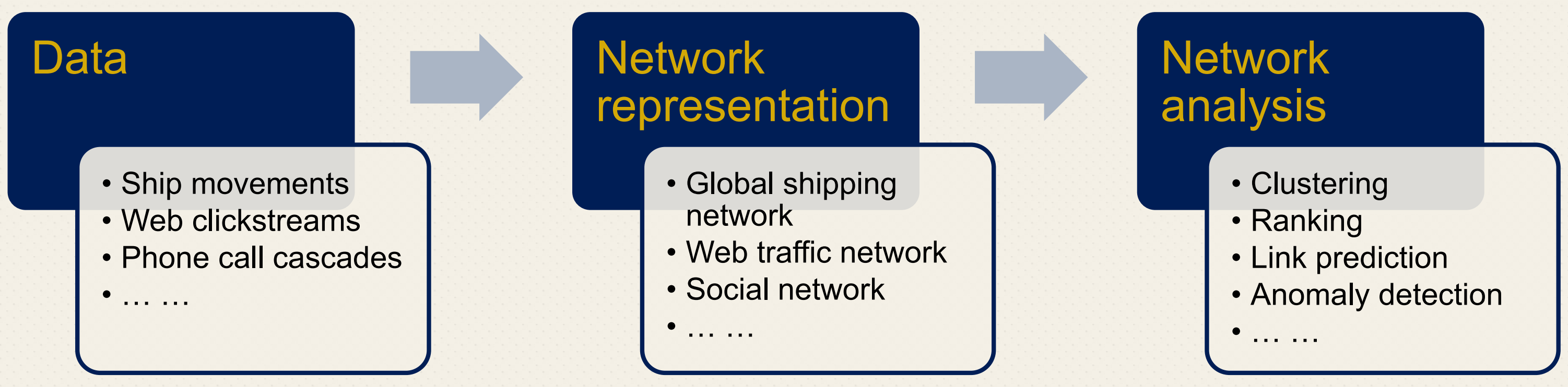
Overview

Prerequisite: To ensure the correctness of network analysis methods, the network as the input has to be a precise representation of the underlying data.

Question: How to effectively represent data as networks, without losing crucial information about higher-order dependencies?

Our answer: We propose the ***Higher-order Network (HON)*** that can embed higher-orders of dependencies to represent data more accurately, can use variable orders for concise representation, and is directly compatible with the existing suite of network analysis methods.

Representing data as network



What people want: A network representing the complex system.

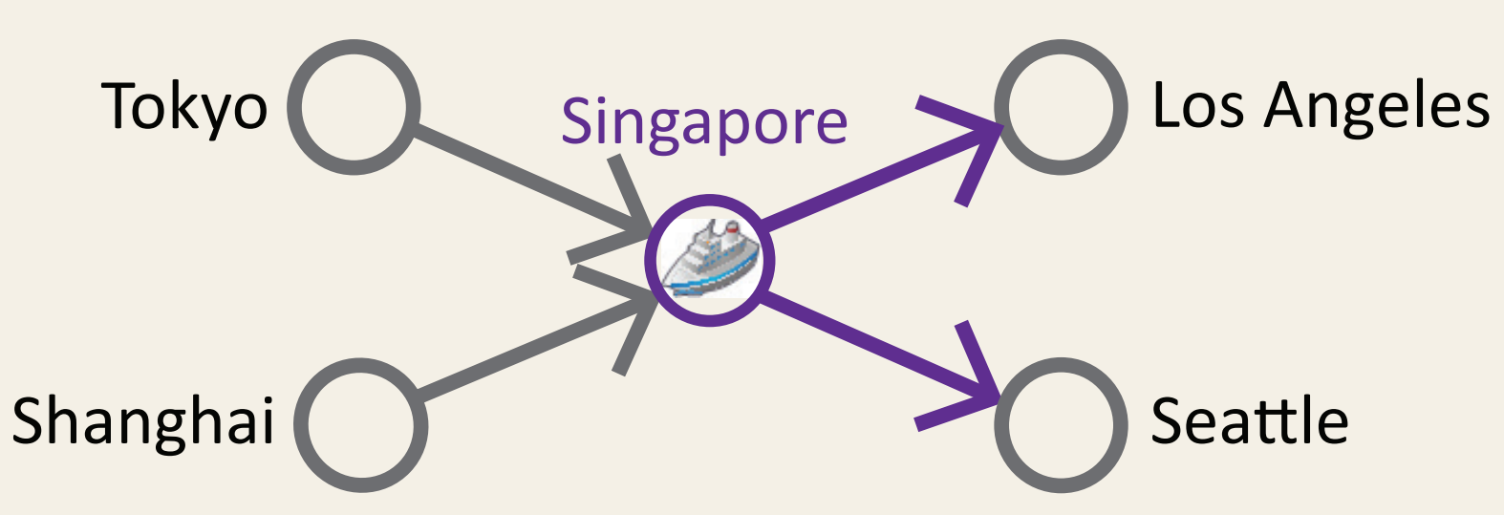
In reality: The network is usually not directly available.

What people have: Recorded sequences of events, such as trajectories of vehicles, retweets of messages, streams of web clicks, etc.

Example: Global shipping data, containing ship movements.

Vessel	Depart	Sailing_date	Arrive	Arrival_date
V-001	Shanghai	2013-01-01	Singapore	2013-01-15
V-001	Singapore	2013-01-16	Los Angeles	2013-02-05
V-002	Singapore	2013-02-01	Los Angeles	2013-03-08
… …	… …	… …	… …	… …

Problem of first-order representation



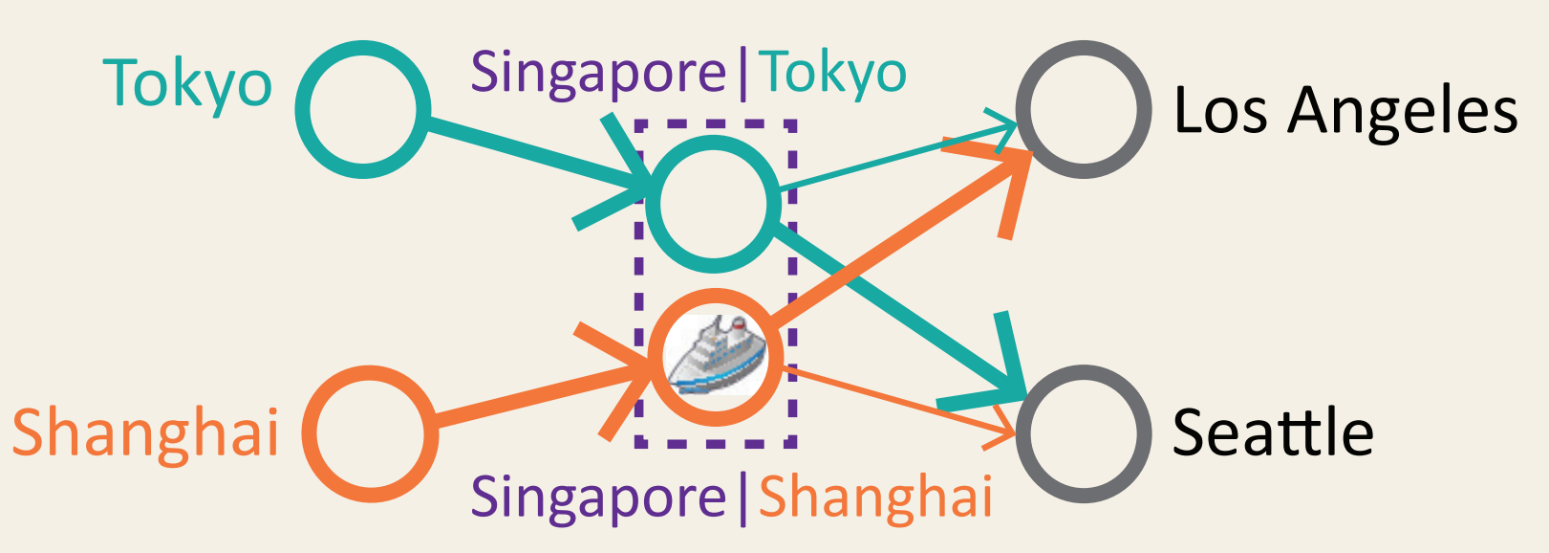
What people usually do: Direct conversion, the sum of pairwise connections in the raw data --> the edge weights in the network.

What is assumed: The Markov property (first-order dependency).

What does it mean: When movements are simulated on the network, where the flow moves next dependes only on its current node.

Example: In this first-order network weighted by shipping frequency, a ship at Singapore has similar probabilities of going to Los Angeles and Seattle, no matter where the ship came to Singapore from; but in fact ships from Tokyo are more likely to go to Seattle, and ships from Shanghai are more likely to go to Los Angeles.

Proposed higher-order network (HON)



What we do: Break down nodes into higher-order nodes that carry different dependency relationships.

Example: In this higher-order network, by breaking down the node Singapore, the ship's next step can depend on previous steps when following different paths to Singapore, thus more accurately simulate movement patterns in data.

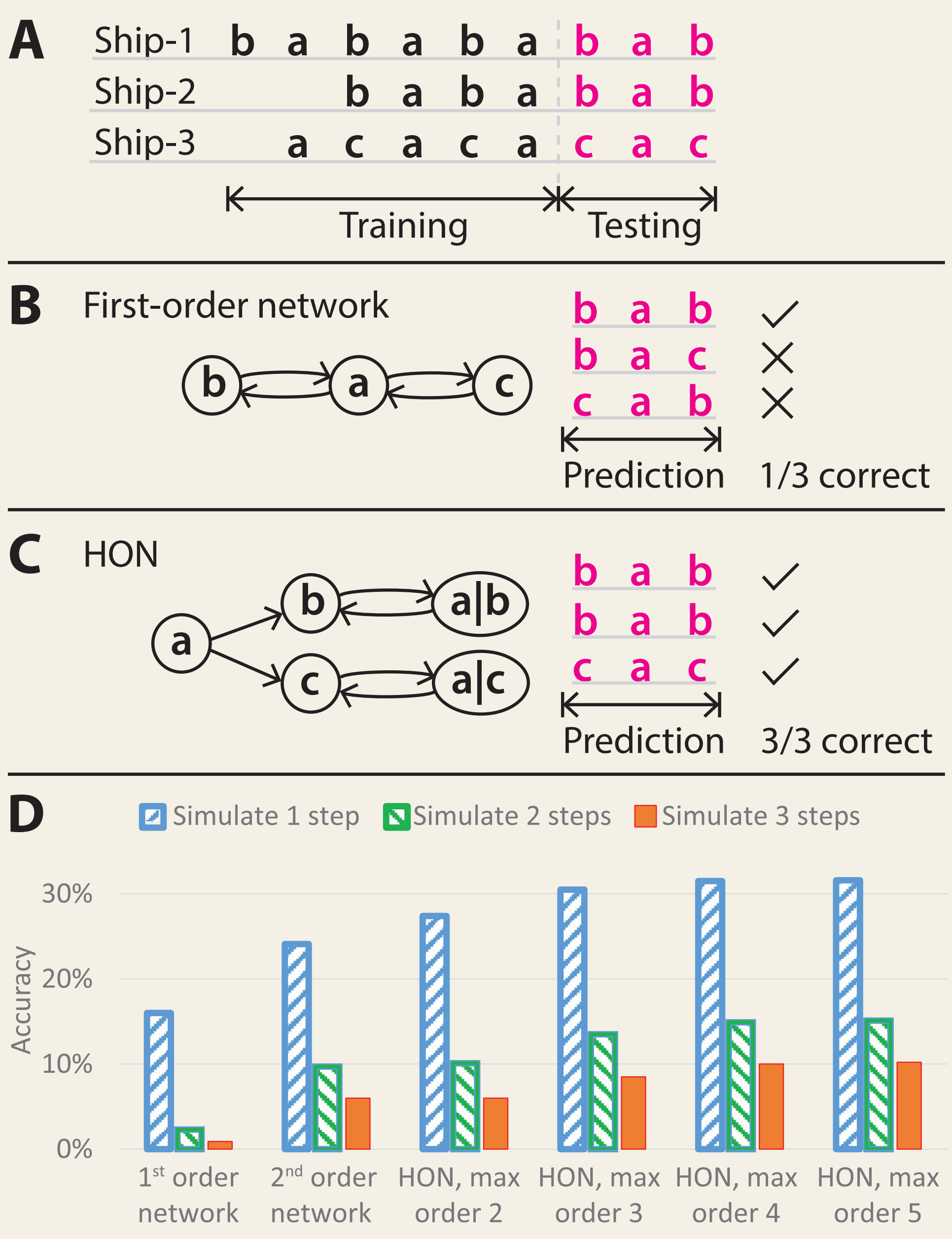
Compatibility: The data structure is consistent with the conventional network representation, allowing for a variety of network analysis methods and algorithms to run on HON without modification.

Improving the result of a variety of network analysis methods

Random walking

Experiment: See how well random walkers on different network representations of the same global shipping data can simulate the true ship movements in the raw data.

Result: The accuracy of simulating one step on HON doubles that of the conventional first-order network, and is higher by one magnitude when simulating three steps (algorithms based on random walking such as PageRank usually need to simulate multiple steps of movement).



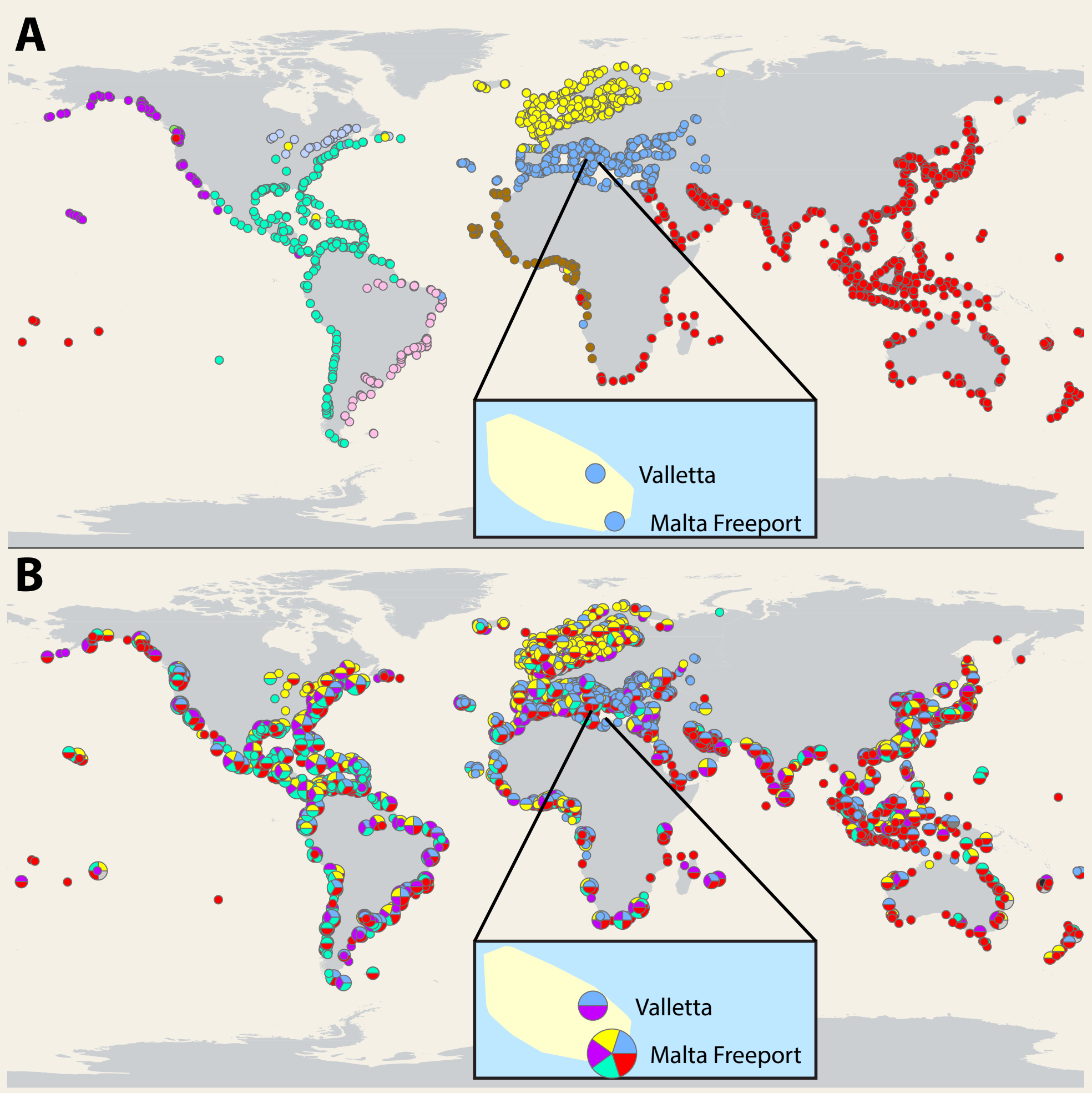
Clustering

Experiment: Given the global ship movements which drives the diffusion of invasion species, compare the clustering of ports (ports tightly coupled by species flows) on different network representations.

Result on first order network: (Fig. A) non-overlapping clusters are generated. Although Valetta and Malta Freeport are local and international ports respectively, the clustering result does not distinguish the two.

Result on HON: (Fig. B) clusters may overlap, and international ports (such as Malta Freeport) are effectively identified by belonging to multiple clusters and potentially suffering from multiple sources of invasions. Fourty-four ports belong to five clusters, including New York, Shanghai, Gibraltar, and so on.

What does it mean: Species may be introduced from multiple previous ports a ship has stopped at; these indirect species introduction pathways are already captured by HON and factored in the clustering.



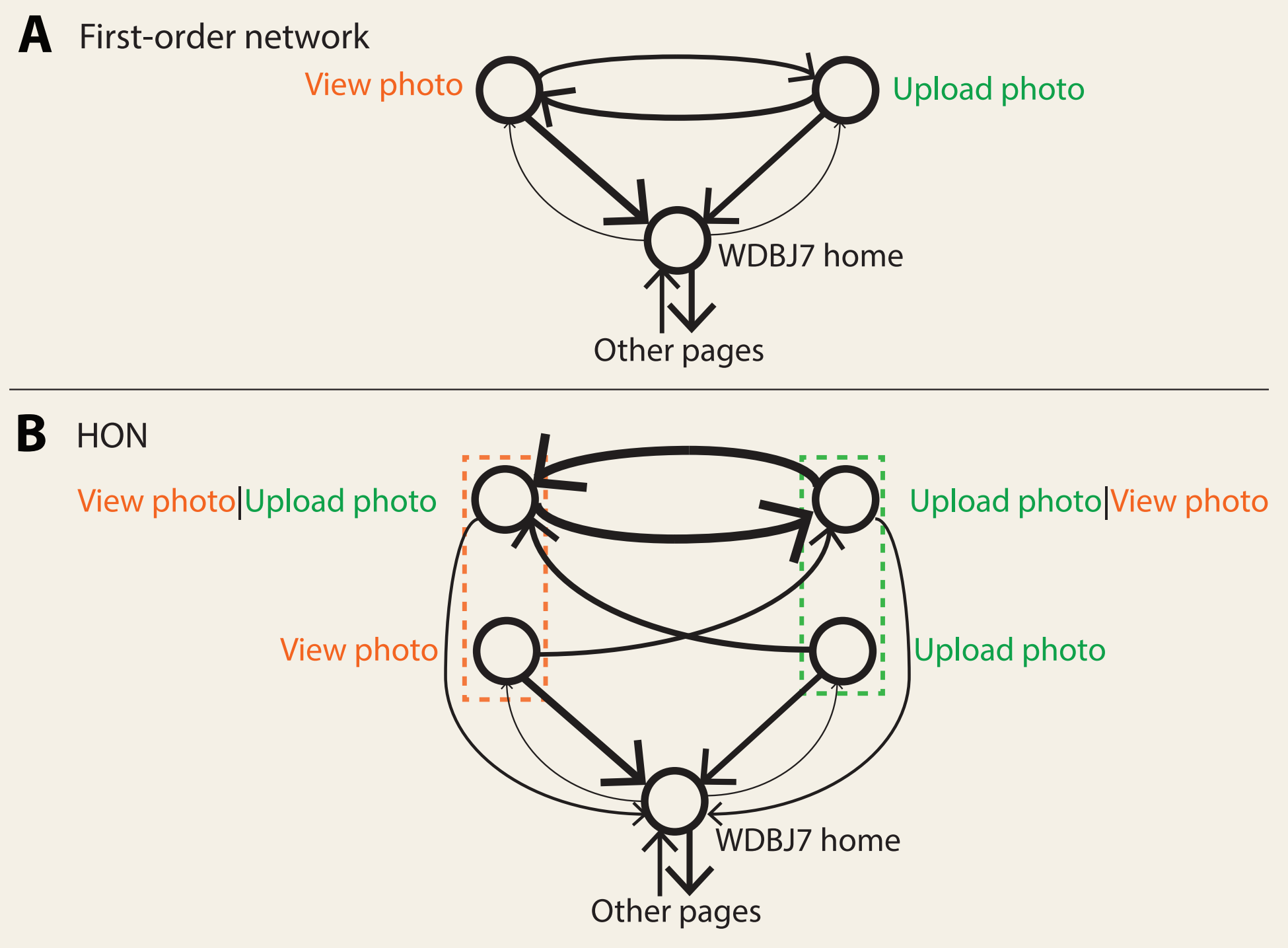
Ranking

Experiment: With a clickstream data set recording how users navigate through Web pages, compare Web PageRank scores by using HON instead of first-order network representation.

Result: More than 90% of the pages show a decrease of PageRank scores, such as pages of news personnels; while a few pages gain considerable PageRank scores, such as weather forecasts and obituaries.

Why such changes: A case study: (Fig. A) the first-order network representation indicates that a user is likely to go back to the homepage after viewing or uploading snow photos. (Fig. B) the HON representation uses additional higher-order nodes and edges to represent a natural scenario that once a user views ***and*** uploads a photo, the user is likely to repeat this process to upload more photos and is less likely to go back to the home page.

What does it mean: HON can better "guide" random walkers to simulate more complex movements such as user's non-Markovian web browsing behaviors, thus HON may be used to improve the ranking results of web ranking, citation ranking, keyphrase extraction, and more, without modification of the PageRank algorithm.



Scalability of HON

Problem: Previous work using fixed order for networks will result in exponential growth of network size when higher orders are incorporated due to combinatorial explosion.

Question: How to build a scalable representation for complex data with high orders of dependencies, and how much more space is needed to represent higher-order dependencies?

Our solution: Use variable orders of dependency, and add higher-orders nodes and edges to a first-order network only where necessary.

Compact representation: HON has less nodes and half the number of edges compared with the fixed second-order network. Even when increasing the maximum order to five, HON still has less edges than the fixed second-order network, while all the useful dependencies up to the fifth order are incorporated in the network.

Speeding up analyses: Another important advantage of HON over a fixed-order network is that network analysis algorithms can run faster on HON, due to HON's compact representation. In addition, HON is sparser than the fixed-order representation, and many network toolkits are optimized for sparse networks. Compared with the fixed second-order network, these tasks run almost two times faster on HON with a maximum order of two, and about the same speed on HON with a maximum order of five (which embeds more higher-order dependencies and is more accurate).

Network representation	No. of edges	No. of nodes	Network density	Probability of returning after two steps	Probability of returning after three steps	Entropy rate (bits)	Clustering time (min)	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, maximum order of two	64,914	17,235	2.2×10^{-4}	41.7%	7.3%	1.46	45	4.8
HON, maximum order of three	78,415	26,577	1.1×10^{-4}	45.9%	16.4%	0.90	63	6.2
HON, maximum order of four	83,480	30,631	8.9×10^{-5}	48.9%	18.5%	0.68	67	7.0
HON, maximum order of five	85,025	31,854	8.4×10^{-5}	49.3%	19.2%	0.63	68	7.6

Interdisciplinary applications

Invasive species modeling & prediction: Species may be carried unintentionally by ships from port to port and cause invasions. Thus, ship movements connect ports in the world in an implicit species flow network. By identifying higher-order dependencies in ship movements, namely where a ship is more likely to go next given its previous steps, ship movements and species flow dynamics can be modeled more accurately.

Social interactions & information diffusion: More accurately representing flow of information on networks can give a more accurate representation of the complex social interactions and the flow of information, which can be of interest for telecom companies, social media and so on.

Healthcare: Representing and highlighting complex patterns in sequences of symptoms, and revealing their relationships with genes and patient features.

Traffic data: Representing complex taxi movements and human trajectories and highlighting emerging movement patterns, which the government can leverage for urban planning, and merchants can use for customer behavior analysis and prediction.

Anomaly detection: The ability to extract and represent higher-order navigation patterns can also be used to analyze web clickstreams and network access patterns, with potential applications from website optimization to intruder detection (based on anomalous access patterns) for security and defense.

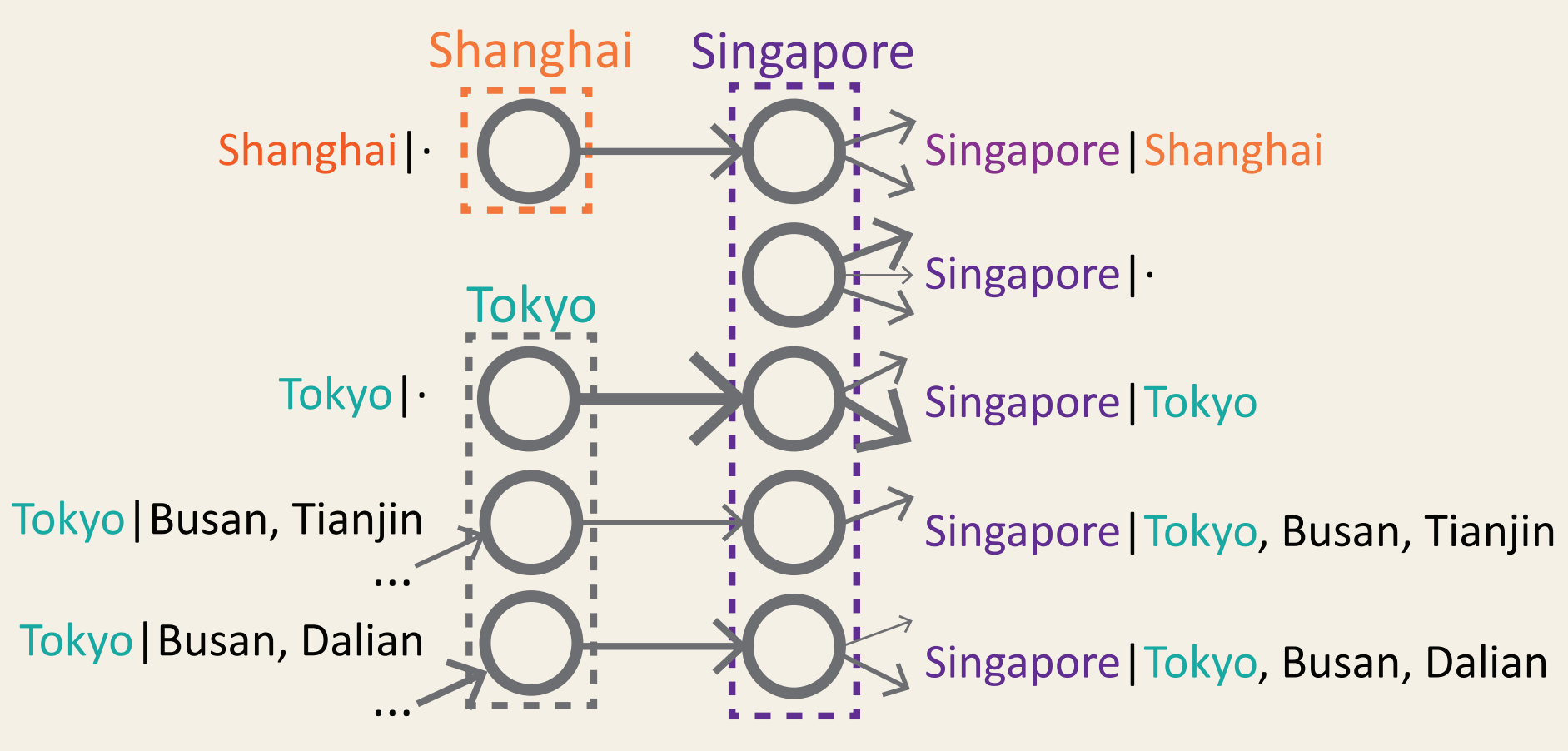
Contributions at a glance

Accurate: HON is a more accurate network representation that is reflective of the underlying real-world phenomena by embedding higher order dependencies.

Compact: Use higher orders only where necessary, scales well.

Compatible: HON is consistent with existing network representation by using simple nodes and edges only, so that existing network analysis tools can be applied on HON directly with no modifications.

Algorithms: Guarantee that HON can extract and represent arbitrary orders mixed in the same data set.



Related works and why we need HON

Hypergraph: Although its edges can connect to multiple nodes simultaneously, the nodes are unordered, thus cannot represent dependencies.

Fixed second order network: (Rosvall et al.) is effective but does not scale well. It is an overkill when first order is enough, and insufficient when higher orders are needed. HON scales better by using variable orders.

Variable Order Markov: For a network representation, VOM generates unnecessary probabilities while failing to capture some necessary probabilities, thus not guaranteeing a network representation.

More information

Full paper: DOI: 10.1126/sciadv.1600028 or scan QR code
Science Advances **2**, e1600028 (2016)
Source code is available at https://github.com/xyjprc/hon
Collaborations and discussions are welcomed.

Acknowledgements: This work is based on research supported by the University of Notre Dame Office of Research via Environmental Change Initiative (**ECI**) and the National Science Foundation (**NSF**) Awards EF-1427157, IIS-1447795 and BCS-1229450; the research was supported by the Army Research Laboratory (**ARL**), and was accomplished under Cooperative Agreement Number W911NF-09-2-0053 [the ARL Network Science Collaborative Technology Alliance (**NSCTA**). We thank David Lodge, Yuxiao Dong and Reid Johnson for their valuable comments.