2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

Predicting Influential Higher-Order Patterns in Temporal Network Data

Christoph Gote[*, †]
Data Analytics Group
University of Zurich
Zurich, Switzerland
Email: cgote@ethz.ch

Vincenzo Perri[*]
Data Analytics Group
University of Zurich
Zurich, Switzerland
Email: perri@ifi.uzh.ch

Ingo Scholtes[‡]
Chair of Machine Learning for Complex Networks
Julius-Maximilians-Universität Würzburg
Würzburg, Germany
Email: ingo.scholtes@uni-wuerzburg.de

***Abstract*—Networks are frequently used to model complex systems comprised of interacting elements. While edges capture the topology of *direct* interactions, the true complexity of many systems originates from higher-order patterns in paths by which nodes can *indirectly* influence each other. Path data, representing ordered sequences of consecutive direct interactions, can be used to model these patterns. On the one hand, to avoid overfitting, such models should only consider those higher-order patterns for which the data provide sufficient statistical evidence. On the other hand, we hypothesise that network models, which capture only direct interactions, underfit higher-order patterns present in data. Consequently, both approaches are likely to misidentify influential nodes in complex networks. We contribute to this issue by proposing five centrality measures based on MOGen, a multi-order generative model that accounts for all indirect influences up to a maximum distance but disregards influences at higher distances. We compare MOGen-based centralities to equivalent measures for network models and path data in a prediction experiment where we aim to identify influential nodes in out-of-sample data. Our results show strong evidence supporting our hypothesis. MOGen consistently outperforms both the network model and path-based prediction. We further show that the performance difference between MOGen and the path-based approach disappears if we have sufficient observations, confirming that the error is due to overfitting.**

I. Introduction

Network models have become an important foundation for the analysis of complex systems across various disciplines, including physics, computer science, biology, economics, and the social sciences [29]. To this end, we commonly utilise network models in which *nodes* represent the interacting elements, and *edges* represent dyadic interactions between those elements. A significant contribution of this perspective on complex systems is that it provides a unified mathematical language to study how the topology of the interactions between individual elements influences the macroscopic structure of a system or the evolution of dynamical processes [3].

In a network, edges capture the *direct* influence between adjacent nodes. However, for most networked systems with sparse interaction topologies, the true complexity originates from higher-order patterns capturing *indirect* influence mediated via *paths*, i.e., via sequences of incident edges traversed by dynamical processes. The general importance of paths for analysing complex systems is expressed in many standard techniques in social network analysis and graph theory. Examples include measures for the importance of nodes based on shortest paths [2, 9], methods for the detection of community structures that are based on paths generated by random walkers [21], but also algebraic and spectral methods that are based on powers of adjacency matrices or the eigenvalues of graph Laplacians [6], which can be thought as implicitly expanding edges into paths.

Standard network methods typically analyse systems based on paths that are generated by some model or algorithm operating on the network topology, e.g., shortest paths calculated by an algorithm, random paths generated by a stochastic model, or all paths transitively expanded based on the network topology. The choice of a suitable model or process generating those paths is a crucial step in network analysis, e.g., for the assessment of node importance [4]. On the other hand, rather than using paths generated by models, we often have access to time-series data that captures real paths in networked systems. Examples include human behavioural data such as time-stamped social interactions, clickstreams on websites, or travel itineraries in transportation networks.

Recent works have shown that, for many complex systems, the patterns in time series data on such paths cannot be explained by the network topology alone. They instead contain higher-order patterns that influence the causal topology of a system, i.e., who can indirectly influence whom over time. To capture these patterns, higher-order generalisations of network models have been proposed [1, 13, 31]. While the specific assumptions about the type of higher-order structures included in those models differ, they have in common that they generalise network models towards representations that go beyond pairwise, dyadic interactions. Recent works in this area have used higher-order models for non-Markovian patterns in paths on networks to study random walks and diffusion processes [14, 22, 27], detect communities and assess node centralities [7, 19, 22, 26, 34], analyse memory effects in clinical time series data [12, 17, 18], generate node embeddings and network visualisations based on temporal network data

[*]Contributed equally.
[†]Also at Chair of Systems Design, ETH Zurich, Switzerland.
[‡]Also at Data Analytics Group, University of Zurich, Switzerland.

IEEE/ACM ASONAM 2022, November 10-13, 2022
978-1-6654-5661-6/22/$31.00 © 2022 IEEE

[20, 23, 30], detect anomalies in time series data on networks [15, 24], or assess the controllability of networked systems [35]. Moreover, recent works have shown the benefit of *multi-order models* that combine multiple higher-order models, e.g., for the generalisation of PageRank to time series data [25] or the prediction of paths [11].

This work extends this view by making the following contributions:

- We consider five centrality measures for nodes in complex networks and generalise them to MOGen, a multi-order generative model for paths in complex networks [11]. Those measures can be considered proxies for the influence of specific node sequences on dynamical processes like, e.g., epidemic spreading and information propagation.
- We show that the direct use of observed paths to calculate those centralities yields better predictions of influential nodes in time series data than a simpler network-based model if there is sufficient training data. At the same time, this approach introduces a substantial generalisation error for small data sets. This motivates the need for a modelling approach that balances between under- and overfitting.
- We develop a prediction technique based on a probabilistic graphical model that integrates Markov chain models of multiple higher orders. Unlike previous works that used multi-order models to model paths in networks, our framework explicitly models the start and end nodes of paths. We show that this explicit modelling of start/end probabilities is crucial to predict influential node sequences.
- Using five empirical data sets on variable-length paths in human clickstreams on the Web, passenger trajectories in transportation systems, and interaction sequences in time-stamped contact networks, we show that our approach provides superior prediction performance.

II. Methods

In the following, we introduce our approach to predict influential nodes and higher-order patterns based on MOGen, a multi-order generative model for path data [11].

A. Paths on Network Topologies

We mathematically define a *network* as tuple $G = (V, E)$, where V is a set of nodes and E is a set of edges. In the example of a public transport system, the individual stations are the nodes, and an edge exists between two nodes if there is a direct connection between the two stations. Users of the system move from start to destinations following *paths* that are restricted by the network topology. A *path* is defined as an ordered sequence $s = v_1 \to v_2 \to \cdots \to v_{l_s}$ of nodes $v_i \in V$, where l_s is the length of the path and nodes can appear more than once. We refer to a set of paths constrained by the same network topology as path data set P.

While empirical paths can come from various sources, we can differentiate between two main types: (i) data directly recorded in the form of paths; (ii) paths extracted from data on temporal interactions, i.e., a temporal network. Examples for the first case include clickstreams of users on the Web or data capturing passenger itineraries from public transportation systems. The primary example of temporal data are records on human interactions, which are a common source for studying knowledge transfer or disease transmission.

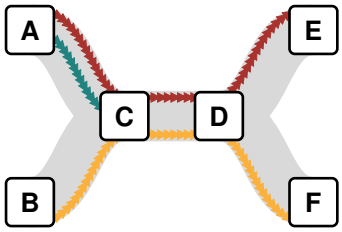

Fig. 1. Exemplary set of paths on a network topology. We observe three colour coded paths from A to B (■), from A to E (■), and from B to F (■). The underlying network topology is shown in grey (■)

A *temporal network* is a tuple $G^{(t)} = (V, E^{(t)})$, where V is a set of vertices and $E^{(t)}$ is a set of edges with a time stamp $E^{(t)} \subseteq V \times V \times \mathbb{N}$. We can extract paths from a temporal network by setting two conditions. First, for two time edges $e_i = (v_1, v_2; t_1)$ and $e_j = (v_2, v_3; t_2)$ to be considered consecutive in a path—i.e., $s = \cdots \to v_1 \to v_2 \to v_3 \to \cdots$— they have to respect the arrow of time, i.e., $t_1 < t_2$. Second, consecutive interactions belong to the same path only if they occur within a time window δ, i.e., $t_2 - t_1 \leq \delta$. Using these conditions, we can derive a set of paths P from any temporal network.

In summary, the network topology constrains the paths that are possible in real-world systems, such as transport or communication systems. However, empirical path data contain additional information on the start and endpoints of paths and the sequences in which nodes are traversed that the network topology does not capture.

B. Modelling Higher-Order Patterns in Path Data

In the previous section, we showed that empirical paths capture information not contained in the network topology. Based on our arguments, one might assume that paths are always better to capture the dynamics on a networked system compared to the topology alone. However, the validity of this argument strongly depends on the number of paths that we have observed.

Let us consider the example shown in Figure 1. As we can infer from the colour coded paths, a path in D will always continue to E if it started in A. In contrast, if the path started in B, it will continue to F. But does this mean that paths from A to F do not exist, despite being possible according to the underlying network topology? To address this question, we need to consider how often we observed the paths from A to E and B to F. If, e.g., we observed both paths only once each, we would have little evidence suggesting that a path from A to F would not be possible. Hence, in this case, using the observed paths as indicators for all possible paths would overfit

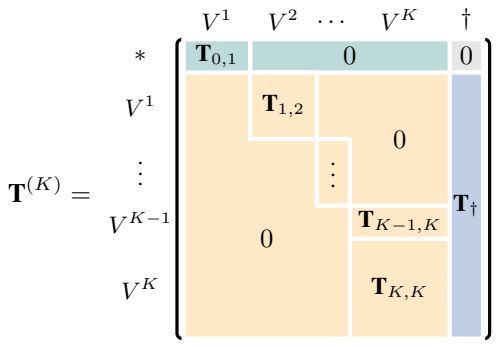

Fig. 2. Multi-order transition matrix $\mathbf{T}^{(K)}$ of a MOGen model with maximum-order K. We split $\mathbf{T}^{(K)}$ into transient part $\mathbf{Q}$ ( ) and absorbing part $\mathbf{R}$ ( ). $\mathbf{S}$ ( ) represents the starting distribution of paths.

the data, and a network model would be more appropriate. In contrast, observing both paths many times without ever observing paths from A to F would indicate that paths from A to F do not exist or are at least significantly less likely than the observed paths. In this case, a network model would underfit the data by not adequately accounting for the patterns present in the empirical path data.

These examples underline that to capture the influence of nodes in real-world networked systems, neither a network model nor a limited set of observed paths is sufficient. Instead, we require a model that can both represent the non-Markovian patterns in the path data, and allow transitions that are consistent with the network topology and cannot be ruled out because path data have not provided enough evidence.

C. MOGen

Our work is based on MOGen, a multi-order generative model for paths [11] that combines information from multiple higher-order models. In addition, MOGen explicitly considers the start- and end-points of paths using the special initial and terminal states $*$ and †. MOGen represents a path $v_1 \to v_2 \to \cdots \to v_l$ as

$$* \to v_1 \to (v_1, v_2) \to \cdots \to (v_{l-K+1}, \ldots, v_l) \to \dagger, \quad (1)$$

where K denotes the maximum memory the model accounts for. Combining the representations of all paths in a set P, the resulting MOGen model is fully described by a multi-order transition matrix $\mathbf{T}^{(k)}$ shown in Figure 2. The entries $\mathbf{T}^{(k)}_{ij}$ of $\mathbf{T}^{(k)}$ capture the probability of a transition between two higher-order nodes.

Considering no memory, a MOGen model with $K = 1$ is equivalent to a network model but for nodes $*$ and † that additionally consider starts and ends of paths. In turn, a MOGen model with K matching the maximum path length observed in P is a lossless representation of the set of paths. Thus, MOGen allows us to find a balance between the network model—allowing all observed transitions in any order—and the observed set of paths—only allowing for transitions in the order in which they were observed.

a) MOGen: Fundamental matrix: Building on the original model [11], we interpret the multi-order transition matrix $\mathbf{T}^{(K)}$ of MOGen as an absorbing Markov chain where the states $(v_1, \ldots, v_{n-1}, v_n)$ represent a path in node v_n having previously traversed nodes $v_1, \ldots, v_{n-1}$. Using this interpretation allows us to split $\mathbf{T}^{(K)}$ into a transient part $\mathbf{Q}$ representing the transitions to different nodes on the paths and an absorbing part $\mathbf{R}$ describing the transitions to the end state †. We can further extract the starting distribution $\mathbf{S}$. All properties are represented in Figure 2.

This representation allows us to compute the Fundamental matrix $\mathbf{F}$ of the corresponding Markov chain.

$$\mathbf{F} = \left(\mathbf{I}^{(n\times n)} - \mathbf{Q}\right)^{-1} \quad (2)$$

Here, $\mathbf{I}^{(n\times n)}$ is the $n \times n$ identity matrix, where n is the number of nodes in the multi-order model without counting the special states $*$ and †. Entries (i, j) of this Fundamental matrix $\mathbf{F}$ represent the expected number of times a path in node i will visit node j before ending. The Fundamental matrix $\mathbf{F}$ is essential as it allows us to compute path centrality measures for the MOGen model *analytically*.

D. Centrality measures

We now introduce five MOGen-based centrality measures that we use in our comparison. For all MOGen-based centrality measures, we also introduce the corresponding measures for the network and a set of paths.

1) Betweenness Centrality: Betweenness centrality considers nodes as highly important if they frequently occur on paths connecting pairs of other nodes. In a network, the betweenness centrality of a node v is given by the ratio of shortest paths $\sigma_{st}(v)$ from s to t through v to all shortest paths from s to t σ_{st} for all pairs of nodes s and t:

$$b_v^{(N)} = \sum \frac{\sigma_{st}(v)}{\sigma_{st}}. \quad (3)$$

Standard betweenness centrality calculated in a network model relies on the assumption that only shortest paths are used to connect two nodes. Using actual path data, we can drop this assumption and consider paths that are *actually* used. Therefore, we can obtain the betweenness of a node in a given set of paths P by simply counting how many times a node appears between the first and last node of all paths.

For MOGen, we can utilise the properties of the Fundamental matrix $\mathbf{F}$. Entries (v, w) of $\mathbf{F}$ represent the number of times we expect to observe a node w on a path continuing from v before the path ends. Hence, by multiplying $\mathbf{F}$ with the starting distribution $\mathbf{S}$, we obtain a vector containing the expected number of visits to a node on any path. To match the notions of betweenness for networks and paths, we subtract the start and end probabilities of all nodes yielding

$$b_v^{(M)} = (\mathbf{S} \cdot \mathbf{F})_v - s_v - e_v^{(M)}. \quad (4)$$

Equation (4) allows us to compute the betweenness centrality for all nodes in the MOGen model—i.e. higher-order nodes.

The betweenness centrality of a first-order node v can be obtained as the sum of the higher-order nodes ending in v.

2) Closeness Centrality (Harmonic): When considering the closeness centrality of a node v, we aim to capture how easily node v can be reached by other nodes in the network. For networks, we are therefore interested in a function of the distance of all nodes to the target node v. The distance matrix $\mathbf{D}$ capturing the shortest distances between all pairs of nodes can be obtained, e.g., by taking powers of the binary adjacency matrix of the network where the entries at the power l represent the existence of at least one path of length l between two nodes. This computation can be significantly sped up by using graph search algorithms such as the Floyd-Warshall algorithm [8] used in our implementation. As our networks are based on path data, the resulting network topologies are directed and not necessarily connected. We, therefore, adopt the definition of closeness centrality for unconnected graphs, also referred to as harmonic centrality [16]. This allows us to compute the closeness centrality of a node v as

$$c_v^{(M)} = \sum_{d \in \mathbf{D}_v} \frac{1}{d}, \tag{5}$$

where $\mathbf{D}_v$ is the v-th row of $\mathbf{D}$.

As MOGen models contain different higher-order nodes, $\mathbf{D}$ captures the distances between higher-order nodes based on the multi-order network topology considering correlations up to length K. While we aim to maintain the network constraints set by the multi-order topology, we are interested in computing the closeness centralities for first-order nodes. We can achieve this by projecting the distance matrix to its first-order form, containing the distances between any pair of first-order nodes but constrained by the multi-order topology. For example, for the distances $d\{(A,B),(C,A)\} = 3$ and $d\{(B,B),(C,A)\} = 2$, the distance between the first-order nodes B and A is 2. Hence, while for the network, the distances are computed based on the shortest path assumption, multi-order models with increasing maximum order K allow us to capture the tendency of actual paths to deviate from this shortest path. Based on the resulting distance matrix $\mathbf{D}$, closeness centrality can be computed following Equation (5).

Finally, for paths, the distance between two nodes v and w can be obtained from the length of the shortest sub-path starting in v and ending in w among all given paths. Again, the closeness centrality is then computed using Equation (5). Therefore, while for all representations, we compute the closeness centrality of a node using the same formula, the differences in the results originate from the constraints in the topologies considered when obtaining the distance matrix $\mathbf{D}$.

3) Path End Probability: The path end probability e_v of a node v describes the probability of a path to end in node v. For paths, $e_v^{(E)}$ is computed correspondingly by counting the fraction of paths ending in node v. For MOGen, all paths end with the state †. Therefore, $e_v^{(M)}$ is obtained from the transition probabilities to † of a single path starting in $*$. This last transition can—and is likely to—be made from a higher-order node. We can obtain the end probability for a first-order node by summing the end probabilities of all corresponding higher-order nodes. The path end probability cannot be computed for a network model as the information on the start and end of paths is dropped for this representation.

4) Path Continuation Probability: When following the transitions on a path, at each point, the path can either continue or end. With the path continuation probability f_v, we capture the likelihood of the path to continue from node v. Similarly to the path start and end probabilities, we obtain the path continuation probability from a set of paths P by counting the fraction of times v does not appear as the last node on a path compared to all occurrences of v.

For MOGen, the path continuation probability is given directly by summing the probabilities of all transitions in the row of $\mathbf{T}^{(K)}$ corresponding to node v leading to the terminal state †. As for other measures, for MOGen, the continuation probabilities are computed for higher-order nodes. We can obtain continuation probabilities for a first-order node v as the weighted average of the continuation probabilities of the corresponding higher-order nodes, where weights are assigned based on the relative visitation probabilities of the higher-order nodes. As path information is required, no comparable measure exists for networks.

5) Path Reach: Finally, we consider path reach. With path reach, we capture how many more transitions we expect to observe on a path currently in node v before it ends. To compute path reach for a set of paths P, we count the average number of nodes on all paths before the path ends for all nodes, in a procedure very similar to the one used to compute path closeness. For MOGen, we can again use the properties of the Fundamental matrix $\mathbf{F}$ and obtain the expected number as the row sum

$$\rho_v^{(M)} = \sum \mathbf{F}_v - 1 \tag{6}$$

We subtract 1 to discount for the occurrence of node v at the start of the remaining path. Analogous to the continuation probability, we obtain the path reach of a first-order node v by weighting the path reach of all corresponding higher-order nodes according to their respective relative visitation probabilities. Again, the path reach requires information on path ends. Therefore, it cannot be computed using the network model.

III. Analysis approach

In Section II, we argued that network models are likely to *underfit* patterns in observed paths that are due to some paths occurring less often (or not at all) while others appear more often than we would expect based on the network topology alone. Similarly, we expect the centralities computed directly on the paths to *overfit* these patterns. We, therefore, expect that when computing centralities based on the network or the paths directly, we misidentify the nodes that are actually influential. We further conjecture that the errors caused by overfitting are particularly severe if the number of observed paths is low, i.e., if we have insufficient data to capture the real indirect influences present in the complex system.

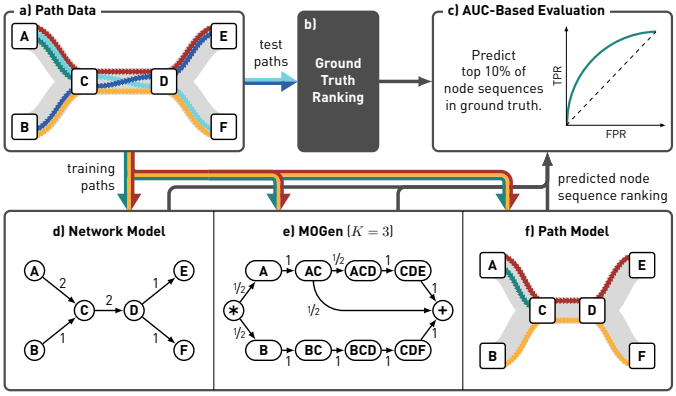

Fig. 3. Overview of our approach to predict influential nodes and node sequences based on path data. We start from path data which we split into training and test sets. We learn three different models bases on the training data: (i) a network model containing all transitions from the training data, (ii) a multi-order generative model containing observed higher-order transitions up to a maximum order of K, which is determined by model selection, and (iii) a path model containing the full paths in the training set. Based on these models, we predict the influence of node or node sequences according to a broad range of centrality measures. We compare the ranking of node sequences to the ground truth rankings obtained from the test paths using AUC-based evaluation.

We now test our MOGen-based centrality against network- and path-based measures in five empirical path data sets. To this end, we compare three types of models for a set of observed paths. First, a network model containing all nodes and edges observed in the set of paths. Second, a path model which precisely captures the observed paths, i.e., the model is identical to the set of paths. Third, MOGen models with different maximum orders K that capture all higher-order patterns up to a distance of K.

We operationalise our comparison in a prediction experiment in which we aim to predict influential nodes and higher-order patterns in a set of test data based on training data. Figure 3 provides an overview of our evaluation approach.

a) Train-test split: For our prediction experiment, we first split a given set of N paths into a training and test set, while treating all observed paths as independent. We denote the relative sizes of the training and test sets as n_{tr}/N and n_{te}/N, respectively.

b) Ground truth ranking: As introduced in Section II, our path-based centrality measures exclusively capture the importance of nodes in a set of observed paths. While we expect this to lead to overfitting when making predictions based on training data, they yield precise ground truth influences when applied to the test data directly. To obtain a ground truth ranking (see Figure 3b), we sort the nodes and node sequences according to their influence in descending order.

c) Prediction of Influential Nodes and Node Sequences: The network model is the least restrictive model for a set of paths. In contrast, the path model always considers the entire history. With $K = 1$, a MOGen model resembles a network model with added states capturing the start- and endpoints of paths. By setting $K = l_{max}$, where l_{max} is the maximum path length in a given set of paths, we obtain a lossless representation of the path data. By varying K between 1 and l_{max}, we can adjust the model's restrictiveness between the levels of the network and the path model. We hypothesise that network and path models under- and overfit the higher-order patterns in the data, respectively, leading them to misidentify influential nodes and node sequences in out-of-sample data. Consequently, by computing node centralities based on the MOGen model, we can reduce this error.

To test this, we train a network model, a path model, and MOGen models with $1 \leq K \leq 5$ to our set of training paths. We then apply the centrality measures introduced in Section II-D to compute a ranking of nodes and node sequences according to each of the models. In a final step, we compare the computed rankings to the ground truth ranking that we computed for our test paths.

d) Comparison to ground truth: While our models are all based on the same set of training paths, they make predictions for node sequences up to different lengths. We allow the comparison of different models' predictions through an upwards projection of lower-order nodes to their matching node sequences. To this end, we match the prediction of the closest matching lower-order node $v_l \in \mathcal{L}$ as the prediction of the higher-order node $v_h \in \mathcal{H}$. Here, $\mathcal{L}$ is the set of lower-order nodes, e.g., from the network model, whereas $\mathcal{H}$ is the set of higher-order nodes from the ground truth. We define the closest matching lower-order node v_l as the node with highest order in $\mathcal{L}$ such that v_l is a suffix of v_h.

We evaluate how well the predictions match the ground truth using an AUC-based evaluation approach. Our approach is built on a scenario in which we aim to predict the top 10% most influential nodes and node sequences in the ground truth data. By considering this scenario, we transform the comparison of rankings into a binary classification problem, where for each node or node sequence, we predict if it belongs into the top 10% of the ground truth or not. All results reported throughout this manuscript refer to averages over at least five validation experiments.

e) Datasets: We test our hypothesis in five empirical path data sets containing observations from three different categories of systems: (i) user clickstreams on the Web (BMS1 [5]), (ii) travel itineraries of passengers in a transportation network (TUBE [32]), and (iii) time-stamped data on social interactions (HOSPITAL [33], WORKPLACE [10], SCHOOL [28]). BMS1 and TUBE are directly collected in the form of paths. For SCHOOL, HOSPITAL, and WORKPLACE

TABLE I
SUMMARY STATISTICS FOR OUR FIVE EMPIRICAL DATA SETS.

	paths		nodes on path		network topology	
	total	unique	mean	median	nodes	links
BMS1	59,601	18,473	2.51	1	497	15,387
TUBE	4,295,731	32,313	7.9	7	276.0	663
SCHOOL	103,260	25,831	2.5	2	242	8,297
HOSPITAL	62,676	13,578	4.8	5	75	1,137
WORK	7,832	1,170	2.5	2	92	753

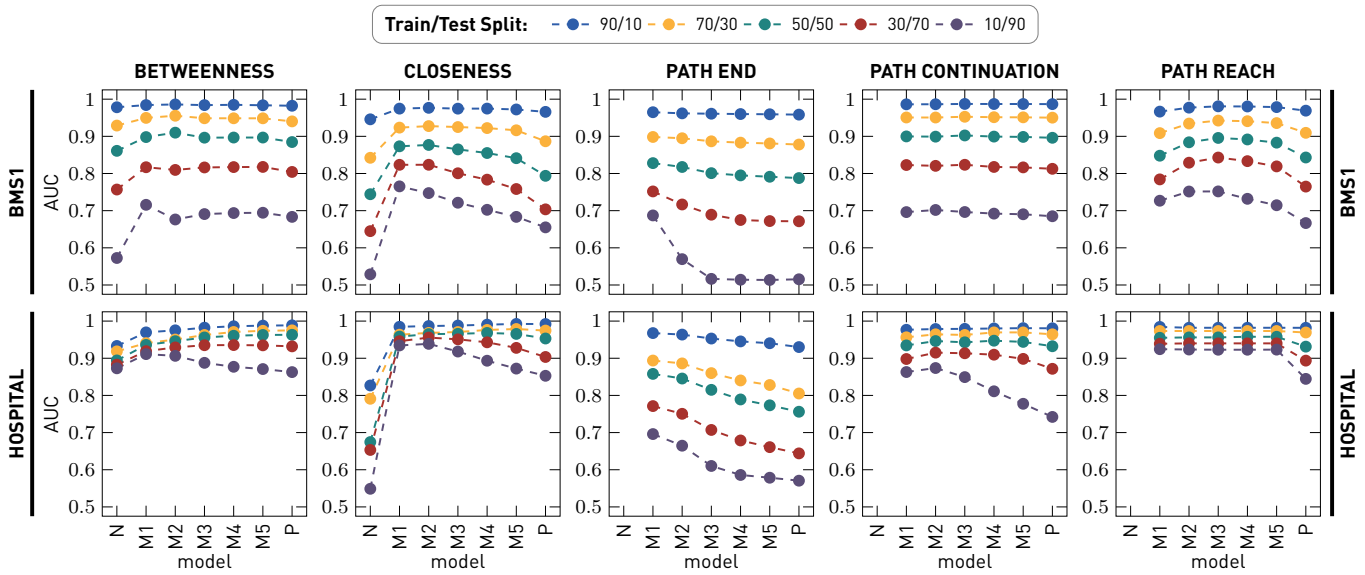

Fig. 4. Prediction results for five centrality measures for the BMS1 and SCHOOL data sets and different train/test splits. N and P indicate the network and path model, respectively. M1 through M5 are MOGen models with maximum orders between 1 and 5.

we extracted paths following Section II-A, using δ as 800s, 1,200s, and 3,600s, respectively. The raw data for all data sets are freely available online. We provide summary statistics for all data sets in Table I.

IV. Results

We now present the results of our prediction experiments comparing the performance of network, path, and MOGen models to predict the influence of nodes and node sequences in out-of-sample data. For ease of discussion, we start our analysis focusing on the two data sets BMS1 and HOSPITAL. Figure 4 shows the results for our five centrality measures. For betweenness and closeness, we do not require information on the start- and endpoint of paths. Therefore, equivalent measures for the network model exist. In contrast, no equivalent measures for the network model can be computed for path end, path continuation, and path reach.

We show the AUC values for the different models and for different relative sizes for our training and test sets. The models shown on the x-axis are sorted according to the maximum distance at which they can capture indirect influences. Thus, starting from the network model (N), via the MOGen models (MK) with increasing K, the models become more restrictive until ending with the path model (P).

Overall, the MOGen models outperform both the network model and the path models. With less training data, the AUC scores of all models decrease. However, as expected, these decreases are larger for the network and path models. For the betweenness and closeness measures, this results in AUC curves that resemble "inverted U-shapes". For the remaining measures, for which no equivalent network measures are available, we generally find that MOGen models with K between 1 and 3 perform best and the prediction performance decreases for more restrictive models, such as the path model. Our results highlight the risk of underfitting for network models and overfitting for path models. We further show that this risk increases when less training data is available.

In Table II, we show the results for all data sets and centrality measures for a 30/70 train/test split. In general, we find similar patterns to those discussed with Figure 4. However, for WORK and TUBE, the difference in prediction quality between the MOGen and path models decreases and for some measures, the path model even yields better performance. WORK and TUBE are those data sets for which we have the highest fraction of total observed paths compared to the number of unique paths in the data sets. As shown in Table I BMS1 contains 59,601 total paths of which 18,473 are unique. This means that, on average, each unique path is observed 3.2 times. These counts increase to 4 for SCHOOL, 4.6 for HOSPITAL, 6.7 for WORK, and 132.9 for TUBE. The good performance of the path model for these data sets shows that the error we found with fewer observations is indeed due to overfitting. In other words, if we have a sufficient number of observations, we can compute the centralities on the path data directly. However, if the number of observations is insufficient, the path model overfits the patterns in the training data and consequently performs worse on out-of-sample data. How many observations are required to justify using the path model depends on the number of unique paths contained in the data set.

In conclusion, our results support our hypothesis. By not capturing the higher-order patterns present in path data and not considering the start- and endpoints of paths, the network model consistently underfits the patterns present in path data. Similarly, the path model overfits these patterns. Consequently, when using either model to rank the influence of nodes and node sequences in path data, we obtain rankings that

TABLE II

AUC VALUES FOR ALL MODELS AND MEASURES ON FIVE DATA SETS FOR A 30/70 TRAIN-TEST SPLIT. N AND P INDICATE THE NETWORK AND PATH MODEL, RESPECTIVELY. M1 THROUGH M8 ARE MOGEN MODELS WITH MAXIMUM ORDERS BETWEEN 1 AND 8 (SHOWN IN). THE BEST PERFORMING RESULT FOR EACH DATA SET AND MEASURE IS HIGHLIGHTED IN BOLD.

		N	M1	M2	M3	M4	M5	M6	M7	M8	P
BMS1	betweenness	0.7569	0.8169	0.8096	0.8163	0.8173	**0.8177**	—	—	—	0.8042
	closeness	0.6449	0.8234	**0.8235**	0.8006	0.7834	0.7582	—	—	—	0.7035
	path end	—	**0.7517**	0.7166	0.6891	0.6749	0.6720	—	—	—	0.6714
	path continuation	—	0.8228	0.8206	**0.8234**	0.8176	0.8165	—	—	—	0.8126
	path reach	—	0.7841	0.8291	**0.8429**	0.8332	0.8191	—	—	—	0.7648
SCHOOL	betweenness	0.7963	0.8331	**0.8407**	0.8357	0.8335	0.8326	—	—	—	0.8270
	closeness	0.6198	0.8069	**0.8221**	0.7806	0.7628	0.7584	—	—	—	0.7521
	path end	—	**0.6521**	0.6270	0.5641	0.5677	0.5703	—	—	—	0.5719
	path continuation	—	**0.8100**	0.7968	0.7767	0.7619	0.7573	—	—	—	0.7552
	path reach	—	**0.8547**	0.8547	0.8547	0.8547	0.8547	—	—	—	0.7462
HOSPITAL	betweenness	0.8828	0.9191	0.9291	0.9351	**0.9355**	0.9347	—	—	—	0.9320
	closeness	0.6533	0.9459	**0.9556**	0.9509	0.9429	0.9279	—	—	—	0.9034
	path end	—	**0.7713**	0.7505	0.7071	0.6788	0.6608	—	—	—	0.6440
	path continuation	—	0.8979	**0.9151**	0.9134	0.9096	0.8983	—	—	—	0.8716
	path reach	—	0.9390	**0.9401**	0.9401	0.9401	0.9401	—	—	—	0.8936
WORK	betweenness	0.7973	0.8542	0.8290	0.8406	0.8416	0.8418	—	—	—	**0.8829**
	closeness	0.5886	0.8495	0.8445	0.8349	0.8342	0.8345	—	—	—	**0.8819**
	path end	—	**0.6955**	0.6844	0.6842	0.6863	0.6877	—	—	—	0.6438
	path continuation	—	0.7431	0.7751	0.7651	0.7648	0.7633	—	—	—	**0.7894**
	path reach	—	**0.8862**	0.8847	0.8828	0.8831	0.8831	—	—	—	0.8419
TUBE	betweenness	0.7634	0.8223	0.9008	0.9241	0.9393	0.9474	0.9453	0.9500	0.9542	**0.9700**
	closeness	0.5497	0.7415	0.8679	0.9046	0.9329	0.9598	0.9707	0.9742	0.9749	**0.9786**
	path end	—	**0.7995**	0.7974	0.7721	0.7378	0.6965	0.6023	0.5614	0.5277	0.5719
	path continuation	—	0.6920	0.7179	**0.7269**	0.7196	0.7196	0.6809	0.6757	0.6683	0.6704
	path reach	—	0.7093	0.8787	0.8996	**0.9131**	0.9101	0.9005	0.8933	0.8845	0.8430

are not consistent with out-of-sample observations. Prediction performance can be significantly improved by using MOGen models that prevent underfitting by capturing higher-order patterns up to a distance of K while simultaneously preventing overfitting by ignoring patterns at larger distances.

V. CONCLUSION

Paths capture higher-order patterns, i.e., indirect influences, between elements of complex systems not captured by network topology. To accurately capture the influence of nodes and node sequences, we must accurately account for these higher-order patterns present in our data. However, not all higher-order patterns observed in a set of paths are representative of the actual dynamics of the underlying system. In other words, by computing centralities on the full paths, we are likely to overfit higher-order patterns and attribute centrality scores to nodes and node sequences different to the ones we obtain when further observing the system and collecting additional paths. Therefore, we require a model that captures only those higher-order patterns for which there is sufficient statistical evidence in the data. We argued that the multi-order generative model MOGen is an ideal model for this purpose as it captures higher-order patterns in paths up to a given length while simultaneously including representations for the start and end of paths.

Based on the MOGen representation, we proposed measures to quantify the influence of both nodes and node sequences in path data according to five different notions of centrality. Our centrality measures range from simple concepts like the betweenness to complex measures such as path reach. For all centrality measures, we also proposed equivalent measures computed directly on path data. While equivalent measures exist for the simple notions of centrality, networks cannot represent the start and end of paths and, hence, cannot represent the full information contained in a path. Consequently, for the more complex measures, no network equivalents exist.

In a prediction experiment with five empirical data sets, we showed that networks models underfit and path models overfit higher-order patterns in path data. Therefore, by computing the centralities of nodes or node sequences according to these models, we misidentify influential nodes. By using MOGen, we can avoid both under- and overfitting. Thus, when computing centralities for MOGen models, we obtain rankings that better represent influential nodes in out-of-sample data.

Our results highlight the potential consequences of applying networks—the most popular model for relational data—to sequential data. Similarly, MOGen-based centralities generally outperform those computed using the path model. The performance difference is greater if the ratio between the number of observed paths and the number of unique paths in a data set decreases. Thus, the larger the variance in the set of observed paths, the larger the potential for overfitting when using a path model to identify central nodes and node sequences in the data. Large variances in observed paths characterise many real-

world systems such as human interactions, where the range of possible interactions is extensive, and data is either costly to obtain or limited in availability. In these cases, our MOGen-based centrality measures provide significantly more accurate predictions on the true influential nodes and node sequences compared to both the network- and path-based measures.

Archival and Reproducibility

Sources for all data used in this paper are provided. A reproducibility package is available at https://doi.org/10.5281/zenodo.7139438. A parallel implementation of the MOGen model is available at https://github.com/pathpy/pathpy3.

Acknowledgements

All authors acknowledge support by the Swiss National Science Foundation, grant 176938.

References

[1] Battiston, F.; Cencetti, G.; Iacopini, I.; Latora, V.; Lucas, M.; Patania, A.; Young, J.-G.; Petri, G. (2020). Networks beyond pairwise interactions: structure and dynamics. *Physics Reports* .

[2] Bavelas, A. (1950). Communication patterns in task-oriented groups. *The journal of the acoustical society of America* **22(6)**, 725–730.

[3] Boccaletti, S.; Latora, V.; Moreno, Y.; Chavez, M.; Hwang, D.-U. (2006). Complex networks: Structure and dynamics. *Physics reports* **424(4-5)**, 175–308.

[4] Borgatti, S. P. (2005). Centrality and network flow. *Social networks* **27(1)**, 55–71.

[5] Brodley, C.; Kohavi, R. (2000). KDD-Cup 2000 homepage.

[6] Chung, F. R.; Graham, F. C. (1997). *Spectral graph theory*. No. 92, American Mathematical Soc.

[7] Edler, D.; Bohlin, L.; Rosvall, M. (2017). Mapping higher-order network flows in memory and multilayer networks with infomap. *Algorithms* **10(4)**, 112.

[8] Floyd, R. W. (1962). Algorithm 97: shortest path. *Communications of the ACM* **5(6)**, 345.

[9] Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry* , 35–41.

[10] Génois, M.; *et al.* (2015). Data on face-to-face contacts in an office building suggest a low-cost vaccination strategy based on community linkers. *Network Science* **3(3)**, 326–347.

[11] Gote, C.; Casiraghi, G.; Schweitzer, F.; Scholtes, I. (2020). Predicting Sequences of Traversed Nodes in Graphs using Network Models with Multiple Higher Orders.

[12] Krieg, S. J.; Robertson, D. H.; Pradhan, M. P.; Chawla, N. V. (2020). Higher-order Networks of Diabetes Comorbidities: Disease Trajectories that Matter. In: *2020 IEEE International Conference on Healthcare Informatics (ICHI)*. IEEE, pp. 1–11.

[13] Lambiotte, R.; Rosvall, M.; Scholtes, I. (2019). From networks to optimal higher-order models of complex systems. *Nature physics* **15(4)**, 313–320.

[14] Lambiotte, R.; Salnikov, V.; Rosvall, M. (2015). Effect of memory on the dynamics of random walks on networks. *Journal of Complex Networks* **3(2)**, 177–188.

[15] LaRock, T.; Nanumyan, V.; Scholtes, I.; Casiraghi, G.; Eliassi-Rad, T.; Schweitzer, F. (2020). Hypa: Efficient detection of path anomalies in time series data on networks. In: *Proceedings of the 2020 SIAM International Conference on Data Mining*. SIAM, pp. 460–468.

[16] Marchiori, M.; Latora, V. (2000). Harmony in the small-world. *Physica A: Statistical Mechanics and its Applications* **285(3-4)**, 539–546.

[17] Myall, A. C.; Peach, R. L.; Weiße, A. Y.; Mookerjee, S.; Davies, F.; Holmes, A.; Barahona, M. (2021). Network memory in the movement of hospital patients carrying antimicrobial-resistant bacteria. *Applied Network Science* **6(1)**, 1–23.

[18] Palla, G.; Páll, N.; Horváth, A.; Molnár, K.; Tóth, B.; Kováts, T.; Surján, G.; Vicsek, T.; Pollner, P. (2018). Complex clinical pathways of an autoimmune disease. *Journal of Complex Networks* **6(2)**, 206–214.

[19] Peixoto, T. P.; Rosvall, M. (2017). Modelling sequences and temporal networks with dynamic community structures. *Nature communications* **8(1)**, 1–12.

[20] Perri, V.; Scholtes, I. (2020). HOTVis: Higher-Order Time-Aware Visualisation of Dynamic Graphs. In: D. Auber; P. Valtr (eds.), *Graph Drawing and Network Visualization - 28th International Symposium, GD 2020, Vancouver, BC, Canada, September 16-18, 2020, Revised Selected Papers*. Springer, vol. 12590 of *Lecture Notes in Computer Science*, pp. 99–114.

[21] Rosvall, M.; Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences* **105(4)**, 1118–1123.

[22] 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(1)**, 1–13.

[23] Saebi, M.; Ciampaglia, G. L.; Kaplan, L. M.; Chawla, N. V. (2020). HONEM: learning embedding for higher order networks. *Big Data* **8(4)**, 255–269.

[24] Saebi, M.; Xu, J.; Kaplan, L. M.; Ribeiro, B.; Chawla, N. V. (2020). Efficient modeling of higher-order dependencies in networks: from algorithm to application for anomaly detection. *EPJ Data Science* **9(1)**, 15.

[25] Scholtes, I. (2017). When is a network a network? Multi-order graphical model selection in pathways and temporal networks. In: *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*. pp. 1037–1046.

[26] Scholtes, I.; Wider, N.; Garas, A. (2016). Higher-order aggregate networks in the analysis of temporal networks: path structures and centralities. *The European Physical Journal B* **89(3)**, 1–15.

[27] Scholtes, I.; Wider, N.; Pfitzner, R.; Garas, A.; Tessone, C. J.; Schweitzer, F. (2014). Causality-driven slow-down and speed-up of diffusion in non-Markovian temporal networks. *Nature communications* **5(1)**, 1–9.

[28] Stehlé, J.; Voirin, N.; Barrat, A.; Cattuto, C.; Isella, L.; Pinton, J.-F.; Quaggiotto, M.; Van den Broeck, W.; Régis, C.; Lina, B.; *et al.* (2011). High-resolution measurements of face-to-face contact patterns in a primary school. *PloS one* **6(8)**, e23176.

[29] Strogatz, S. H. (2001). Exploring complex networks. *nature* **410(6825)**, 268–276.

[30] Tao, J.; Xu, J.; Wang, C.; Chawla, N. V. (2017). HoNVis: Visualizing and exploring higher-order networks. In: *2017 IEEE Pacific Visualization Symposium (PacificVis)*. IEEE, pp. 1–10.

[31] Torres, L.; Blevins, A. S.; Bassett, D. S.; Eliassi-Rad, T. (2020). The why, how, and when of representations for complex systems. *arXiv preprint arXiv:2006.02870* .

[32] Transport for London (2014). Rolling Origin and Destination Survey (RODS) database.

[33] Vanhems, P.; Barrat, A.; Cattuto, C.; Pinton, J.-F.; Khanafer, N.; Regis, C.; Kim, B.-A.; Comte, B.; Voirin, N. (2013). Estimating Potential Infection Transmission Routes in Hospital Wards Using Wearable Proximity Sensors. *PLoS ONE* **8**.

[34] Xu, J.; Wickramarathne, T. L.; Chawla, N. V. (2016). Representing higher-order dependencies in networks. *Science advances* **2(5)**, e1600028.

[35] Zhang, Y.; Garas, A.; Scholtes, I. (2020). Higher-order models capture changes in controllability of temporal networks. *Journal of Physics: Complexity* .