A Snapshot-Based Knowledge Graph Model for Temporal Link Prediction

Philipp Plamper[1,*], Oliver J. Lechtenfeld[2,*], Wolf von Tümpling[3] and Anika Groß[1,*]

[1]Anhalt University of Applied Sciences, Department Computer Science and Languages, Köthen (Anhalt), 06366, Germany
[2]Helmholtz Centre for Environmental Research – UFZ, Department of Analytical Chemistry, Leipzig, 04318, Germany
[3]Helmholtz Centre for Environmental Research – UFZ, Central Laboratory for Water Analytics and Chemometrics, Magdeburg, 39114, Germany

Abstract

Many systems can be intuitively modeled as knowledge graphs using entities and their relationships. However, we often have only partial or little knowledge of the inherent processes of complex, changing systems such as biomedical, economic or ecological systems. As a result, the construction of knowledge graphs often suffers from incompleteness which can lead to inaccurate analysis results and incorrect conclusions. A widely used approach is to monitor and analyse complex changing systems using time series of measurements. To understand a complex temporal network of processes, it is crucial to identify inherent temporal relationships and interactions. A complete temporal knowledge graph model could provide a better foundation for applications in complex systems and increase its potential to add context and connections that allow uncovering hidden or unknown relationships in data. We propose a snapshot-based knowledge graph model and temporal link prediction algorithm to find relationships between examined objects in successive time points of multivariate time series. We evaluate and demonstrate the functionality in an environmental chemistry use case and predict the transformations of molecules for two datasets. Our approach is able to discover previously unknown relationships in a snapshot-based knowledge graph helping to better understand the dynamics of the examined system.

Keywords

Knowledge representation and reasoning, Graph-based database models, Temporal data, Network algorithms

1. Introduction

Understanding mechanics and dynamics in complex networks has become an active field of research in recent years [1]. Since the introduction of knowledge graphs, the topic has become even more prominent [2, 3]. Despite the different naming conventions, the idea behind complex networks and knowledge graphs have many similarities. Both use a graph-based data model to represent a system and consist of nodes and edges describing entities and their relationships. Additionally, the nodes and edges are enriched with descriptive properties [4, 5, 6]. The representation of network-like data in the form of graphs is often used to model real-world

2nd International Workshop on Knowledge Graph Reasoning for Explainable Artificial Intelligence, December 09, 2023, Tokyo, Japan

*Corresponding author.

philipp.plamper@hs-anhalt.de (P. Plamper); oliver.lechtenfeld@ufz.de (O. J. Lechtenfeld); wolf.vontuempling@ufz.de (W. v. Tümpling); anika.gross@hs-anhalt.de (A. Groß)

applications to capture knowledge from various domains and to integrate, manage and add value to linked data sources [7, 8].

Most real-world systems are subject to constant change and therefore should not be treated as a static network [9]. In a graph, this means that nodes and edges can appear or disappear over time and that properties can change, leading to a completely new topology and dynamics. In many complex networks the transitions between the individual time points, i.e. the relationships, are completely unknown and can hardly be determined by experts. Nevertheless, it is common to record various values experimentally with the help of time series in order to better understand the development of e.g. processes over time [10, 11].

In most cases, the actual temporal relationships between individual time series cannot be derived directly from the collected multivariate time series [12, 13]. In order to track temporal developments in the data and to increase the accuracy of graph analyses, the underlying system should store data over a longer period of time and assign temporal properties such as timestamps [9].

We here present a new snapshot-based model for temporal knowledge graphs to find so far unknown relationships and interactions in complex systems. The prediction is based on measured quantities for large numbers of objects from multivariate time series such as molecule intensities in mass spectrometry experiments or vehicle counters for traffic monitoring. Our graph structure inherently integrates temporal properties to model and predict directed edges between entities in multivariate time series. The snapshot-based knowledge graph model and temporal link prediction are particularly useful when no temporal edges are known in advance and real-time monitoring is not or insufficiently possible. Our approach is generic and could be adopted to different complex changing systems, e.g. biological, economic or traffic networks. The temporal graph approach allows us to better represent the dynamic nature of temporal networks and to gain new insights into temporal dynamics based on graph analysis. We evaluate our approach to show its ability to uncover unknown information about the analysed system. Our main contributions are

1. A novel snapshot-based knowledge graph model to represent multivariate time series as a temporal graph structure.
2. A temporal link prediction algorithm to identify directed temporal edges between nodes across successive time points.
3. An evaluation of the approach using two datasets representing complex environmental chemistry systems.

The methods have been implemented in a graph database management system and can be analysed visually and statistically. In [14] we first introduced a temporal graph model specifically designed for molecules and chemical transformations. Compared to the previous publication, we here generalize the data model to objects and temporal edges without specific domain focus. Moreover, we here present the generalized methodology and actual link prediction algorithm that were not the focus of the use case in [14]. We further add a new dataset to the evaluation.

The paper is further organized as follows. We first discuss related work in Section 2. We present the snapshot-based temporal knowledge graph model in Section 3 and the temporal link prediction algorithm in Section 4. We then evaluate the model and algorithm in an environmental chemistry use case in Section 5 and conclude in Section 6.

2. Related work

The property graph model represents a directed heterogeneous multirelational network with properties at nodes and edges. It is one of the standard models for representing data with graph-like structure as a knowledge graph [15, 16, 2, 17, 3, 7]. Despite the lack of a standard formalization [18, 17, 19], several graph databases support the model because it is considered as comprehensive and easy to understand [18, 20, 15]. The original development of the property graph model was based on static graphs and therefore relies on static data, but many systems contain time-varying data. However, representing time-varying data in a static graph using static graph tools and algorithms can lead to inaccurate and false analysis results [21, 9, 22]. The need for temporal knowledge graphs and the lack of a formal definition is reflected in the different models proposed [23, 24, 25, 26, 27]. Temporal graph models can already be found for instance in social graphs [23, 25], logistic graphs [26] or network graphs [24]. The proposed temporal knowledge graph models can be roughly classified into duration-labeled, interval-labeled or snapshot-based models [23].

A central topic in graph analysis is the search for missing structures in an incomplete network, this is called *link mining*, more precisely *link prediction* in the domain of complex networks [28] and knowledge graph *completion* in the domain of knowledge graphs [29, 21]. Link prediction algorithms use the existing structure given by the nodes and edges to try to infer missing relationships based on properties and already observed relationships. The various approaches to predict missing relationships in a temporal knowledge graph use, for example, graph embedding [30], machine learning methods [31], similarity-based methods, or probabilistic methods [32]. The adapted algorithms have the goal of predicting missing links in an existing graph or predicting links that will occur in the future [28, 33]. Existing approaches usually try to predict edges based on already known structures in the graph, but lack the ability to find edges when relationships are mostly unknown in the first place.

In many domains, multivariate time series are used to understand the underlying structure of the data [10, 11]. Many of the systems under consideration have a graph-like structure [13], but it is often difficult or impossible to capture the relationships between the individual time series. Proposed approaches for analysing the relationships between time series include transforming the time series as a multilayer visibility graph [34, 13], as recurrence networks [35, 36] or mapping it as a graph based on causality patterns [12]. These methods focus on comprehensive analysis to find general correlations and patterns in a complex system, rather than inferring actual transformations or interactions between thousands to millions of objects undergoing unknown changes. Previous network-like approaches on multivariate time series do not preserve snapshots while enriching it with temporal edges. We are interested in the relationships that take place between successive snapshots and lead to changes in the measured values of the individual time series. Our approach introduces a new knowledge graph model that represents multivariate time series with a graph-like structure to allow the tracking of single time series and understanding influences on their development.

3. Snapshot-Based Knowledge Graph model

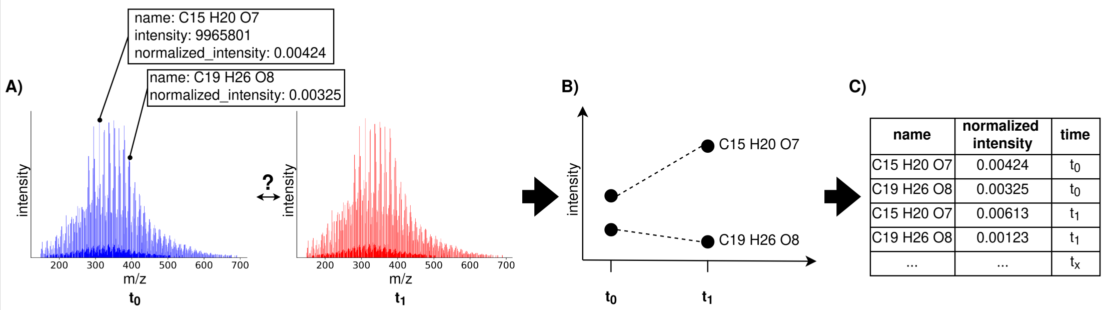
3.1. Scenario

Our snapshot-based knowledge graph model and temporal link prediction is based on multivariate time series and is particularly useful when actual relationships representing transformations, interactions or flows between nodes in succeeding snapshots are unknown, although there is evidence that they exist.

We here consider multivariate time series data, i.e. there is a time series for each considered object of interest in a larger set of objects. We assume each discrete time point in each time series contains some kind of measurable *quantity* per *object*. This quantity changes over time and we expect the objects to be related to each other. Depending on the domain, the repeatedly measured quantity of objects could be manifold, e.g. number of transferred goods or vehicles at monitoring stations, measured intensities of molecules, concentrations of chemical compounds or nutrients in a water treatment plant. Potential connections may be known in advance, but the actual relationships are unknown. For instance,

- there are chemical rules about how molecules can be transformed into other molecules, but we do not know what transformations have actually taken place,
- we know a network of roads, but we do not know which routes have actually been used by vehicles.

Figure 1 shows a chemical example of two mass spectra that can be enriched with temporal links based on our snapshot-based knowledge graph model. A mass spectrum provides measured m/z-values from which chemical formulas of molecules can be derived, e.g. ”C15 H20 O7” and ”C19 H26 O8”. The value on the y-axis is the intensity of the molecule, e.g. 9965801 at ”C15 H20 O7”. The intensity might also be normalized as proportion of the intensity of a molecule to the sum of all intensities of the mass spectrum. For instance, molecule ”C15 H20 O7” has a normalized intensity of 0.00424 or 0.424%.



name	normalized intensity	time
C15 H20 O7	0.00424	t_0
C19 H26 O8	0.00325	t_0
C15 H20 O7	0.00613	t_1
C19 H26 O8	0.00123	t_1
...	...	t_x

Figure 1: Environmental chemistry scenario. A) Two consecutive mass spectra for complex mixtures of environmental samples measured at time t_0 and t_1, B) Changing intensity of two molecules at t_0 and t_1, C) Tabular representation of multivariate time series for mass spectra data.

Applying our approach to this scenario, the quantity value is based on the measured intensities from the mass spectra. Identified molecules will be object nodes in the graph with their quantities as property. The molecules of the first mass spectrum (t_0) form the first snapshot, the molecules of the second mass spectrum (t_1) form the second snapshot of the knowledge graph, and so on. Between the two successive mass spectra, the quantities and proportions of the molecules change, indicating that chemical transformations have taken place between the molecules [37]. We will focus on this chemical scenario for examples in the remainder of this paper, but the approach is applicable to other time series data with repeated measurements on large object sets.

3.2. Graph Model

We propose a snapshot-based knowledge graph model to represent multivariate time series as a temporal graph structure. Our graph model is a temporal extension of the widely used standard ”Labeled Property Graph” model [15] and belongs to the group of the snapshot-based temporal graphs, e.g. [24, 27]. It includes one graph (or snapshot) for each discrete time point. Snapshots consist of object nodes that will be connected by temporal edges (see Section 4) and together form the snapshot-based knowledge graph.

Figure 2 shows our temporal knowledge graph model. The model contains nodes of one type (‘Object’) and connects them by three different types of temporal edges (‘SAME_AS’, ‘POTENTIAL’, ‘PREDICTED’). All types of temporal edges in the model connect a node in one snapshot to a node in a subsequent snapshot and do not exist between nodes within one snapshot as illustrated in Figure 3 C. An ‘Object’ node describes an entity such as a molecule in a chemical network. The node can be identified by its name and snapshot properties, e.g. a molecule’s formula and snapshot 0, which describes the object at the first time point in the data. In addition, each node has a quantity property that describes a value that can change over time, such as the measured intensity of a molecule in a sample.

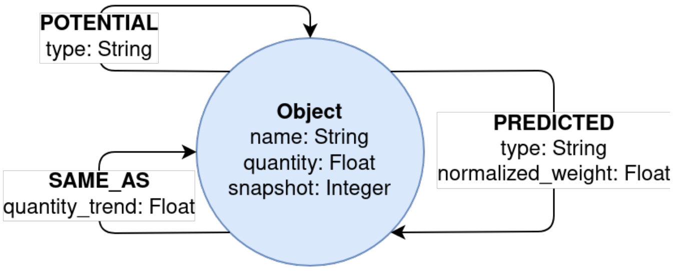


Figure 2: Snapshot-based knowledge graph model with ‘Object’ node, properties and temporal edges.

Initially, each discrete time point in the time series forms a snapshot, which means that the same objects can occur more than once and be identified by the name property. An ‘Object’ can be uniquely identified by the combination of name property and snapshot. Snapshots can be distinguished and ordered by the snapshot property at the ‘Object’ nodes. ‘Object’ nodes with the same name property are connected across successive snapshots via the ‘SAME_AS’ edges. Every ‘SAME_AS’ edge contains the relative change of the quantities of the successive nodes as property, which is termed quantity trend. The calculation is shown in Formula 1:

$$quantity\_trend = \frac{quantity_{succ} - quantity_{prev}}{quantity_{prev}} \tag{1}$$

It subtracts the quantity of the previous node ($quantity_{prev}$) from the quantity of the successive node ($quantity_{succ}$) and divides the result by the quantity of the previous node ($quantity_{prev}$). The possible interpretations of the quantity trend are increasing, decreasing and consistent, which is important for our temporal link prediction algorithm (see Section 4). A quantity trend of exactly 0 is considered *consistent*, above it is considered *increasing* $(0, \infty)$ and below it is considered *decreasing* $(-\infty, 0)$. Additionally, an error margin e around zero $(+/-\frac{1}{2}e)$ can be applied to expand the range considered as *consistent*. The error margin takes into consideration the immanent analytical uncertainties caused by instrumental variability (i.e., noise).

The second edge type describes the potential transformations from an ‘Object’ node to another node in a successive snapshot and is called a ‘POTENTIAL’ edge. ‘POTENTIAL’ edges must be determined based on existing knowledge beforehand. For instance, experts in a domain know basic rules on possible interactions or transformations or for transport networks maps contain all known roads. However, transformations or interactions that actually took place or actual routes of transportation are unknown. In the chemical scenario, the ‘POTENTIAL’ transformations might be computed based on all chemically possible transformations of the molecules in the considered use case (e.g. single reactions or chains of reactions).

The third type of edges describes the predicted transformations that are likely to occur and are called ‘PREDICTED’ edges. They are a subset of the ‘POTENTIAL’ edges and indicate the transformations that likely occurred out of all potential transformations. For instance, the ‘PREDICTED’ edges could describe the various chemical transformations that have actually taken place in a molecule after an conditional experiment measured as multivariate time series. The ‘PREDICTED’ edges are computed using the ‘Transformation Prediction Algorithm’ described in Section 4.

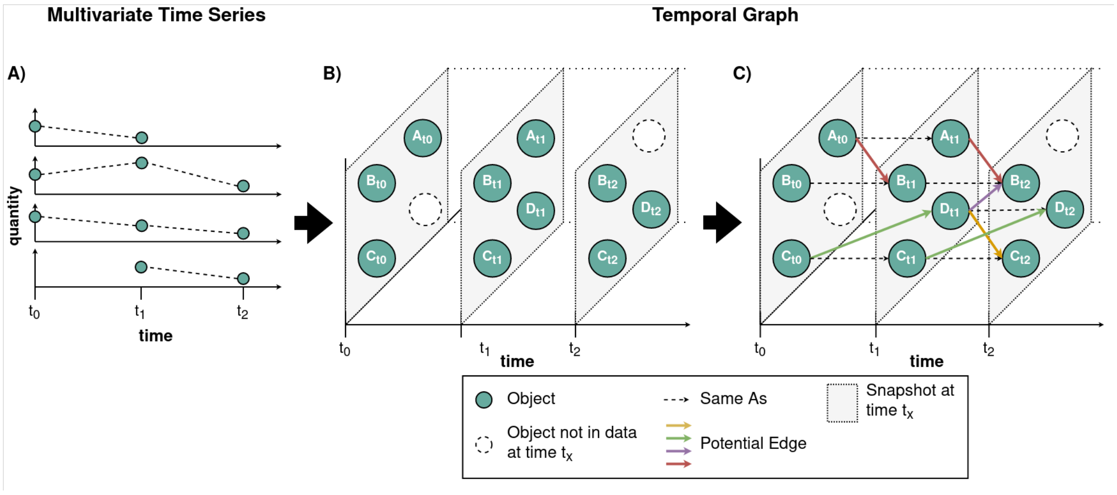


Figure 3: Construction of the snapshot-based knowledge graph from multivariate time series data.

Figure 3 illustrates the step-wise construction of the snapshot-based knowledge graph from multivariate time series data for several objects. Each discrete time point in the time series represents a snapshot of the graph and each 'Object' represents a node (see Figure 3 (A and B)). The exemplary graph consists of 3 consecutive snapshots (t_0, t_1, t_2) and four 'Object' nodes (green A-D). Each snapshot has a different composition, i.e. nodes can be present in all (e.g., B_{t0}, B_{t1}, B_{t2}) or some of the snapshots (e.g., D_{t1}, D_{t2}). Figure 3 (C) illustrates the creation of new edges between succeeding points in time. This includes the 'SAME_AS' edges between same 'Object' nodes and the 'POTENTIAL' edges, which describe possible transformations. The edges are directed and always end at a node in the successive snapshot, i.e. no edges exist within a snapshot. Note, that the snapshots do not necessarily represent equidistant points in time . The same edge color describes the same transformation (e.g., from C_{t0} to D_{t1} and C_{t1} to D_{t2}).

4. Temporal Link Prediction

At this point, the snapshot-based knowledge graph contains all the necessary information to describe temporal data as a complex network. However, snapshots are not interconnected via temporal edges describing the transformations in the considered system. We developed the 'Transformation Prediction Algorithm' to identify the likely occurred transformations based on the potential transformations and the known quantity trends (e.g. from measurements). Knowing the direction of the transformations and assuming that a decrease in quantity in one 'Object' node is reflected by an increase in quantity in another 'Object' node and vice versa, it is possible to predict the likely transformations responsible for the dynamics in the snapshot-based knowledge graph. For instance in a network of molecules, potential chemical transformations can be specified by experts and the intensity of the molecules in samples is measured by mass spectrometry. When the quantity of one molecule decreases, the quantity of another increases. The prediction is based on the following three considerations and illustrated in Figure 4:

1. Consideration 1: If the quantity between 'Object' node A_{tx} and A_{tx+1} increases from snapshot t_x to t_{x+1}, the transformations of B_{tx} and C_{tx} from snapshot t_x to A_{tx+1} are assumed to be relevant and the transformation from A_{tx} to other nodes in snapshot t_{x+1} are assumed to be non-relevant. Therefore, outgoing potential transformations from the node A_{tx} are excluded and incoming edges to the node A_{tx+1} are included in the consideration of the 'PREDICTED' edges.
2. Consideration 2: If the quantity between 'Object' node A_{tx} and A_{tx+1} decreases from snapshot t_x to t_{x+1}, the transformations of B_{tx} and C_{tx} from snapshot t_x to A_{tx+1} are assumed to be non-relevant and the transformation from A_{tx} to other nodes in snapshot t_{x+1} are assumed to be relevant. Therefore outgoing potential transformations from the node A_{tx} are included and incoming edges to the node A_{tx+1} are excluded in the consideration of the 'PREDICTED' edges.
3. Consideration 3: If the quantity between 'Object' node A_{tx} and A_{tx+1} is consistent, the 'Object' node has a balancing number of transformations. Therefore, all incoming and outgoing potential transformations from A_{tx} and to A_{tx+1} are excluded.

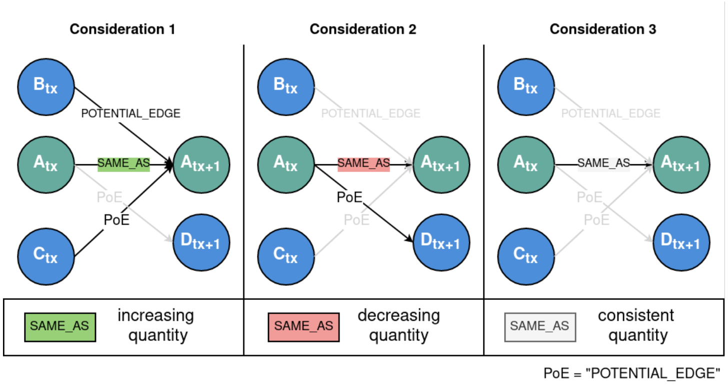


Figure 4: The three considerations of the 'Transformation Prediction Algorithm' are based on the quantity trend at the 'SAME_AS' edges and can be interpreted as increasing, decreasing or consistent.

In order to cover all nodes in all snapshots, the 'Transformation Prediction Algorithm' is implemented as an iterative algorithm, starting with the different interpretations and considerations of the quantity trends at the 'SAME_AS' edges. In short, the 'PREDICTED' edges are a subset of all 'POTENTIAL' edges and start always on a node from snapshot t_x with a decreasing quantity trend to snapshot t_{x+1} at the 'SAME_AS' edges and end on a node at snapshot t_{x+1} with an increasing quantity trend at the respective 'SAME_AS' edge from snapshot t_x to t_{x+1}.

Algorithm 1 'Transformation Prediction Algorithm'

```
Input: Graph G with nodes N and edges E, error margin e
Output: Graph G with edges of type 'PREDICTED' PrE

$typePoE \leftarrow$ 'POTENTIAL'
$typeSaS \leftarrow$ 'SAME_AS'
$e \in [0, 1]$
for $n \in N$ do
    ▷ get quantity trend from preceding snapshot of n ◁
    $qtyTrend \leftarrow$
        $getQtyTrend(n.getIncomingEdges(typeSaS))$
    if $qtyTrend > +\frac{1}{2}e$ then
        ▷ collect all nodes with a 'POTENTIAL' edge to n ◁
        $N' \leftarrow getStartNode(n.getIncomingEdges(typePoE))$
        for $n' \in N'$ do
            ▷ get quantity trend of n' to succeeding snapshot ◁
            $qtyTrend' \leftarrow$
                $getQtyTrend(n'.getOutgoingEdges(typeSaS))$
            if $qtyTrend' < -\frac{1}{2}e$ then
                $G \leftarrow addEdgetypePrE(n', n)$
return G
```

This means that the node in snapshot t_x and the node in snapshot t_{x+1} can be connected by a 'PREDICTED' edge only if both nodes have an opposite quantity trend.

Algorithm 1 describes the 'Transformation Prediction Algorithm' for creating the 'PREDICTED' edges in the snapshot-based knowledge graph. The input is graph G consisting of nodes N and edges E as well as the defined error margin e (line 1). As output, we generate the graph G enriched with the added predicted likely occurring transformations PrE (line 2). We define two variables for the edge types $typePoE$ and $typeSaS$ to increase readability of the pseudo code (line 4-5). The error margin e represents the uncertainty in the measured quantity in the data and is set to a value between 0 and 1 (line 6) (e.g., 0.05 for 5 % or 0 to neglect error margin). A quantity trend $qtyTrend$ above $+\frac{1}{2}e$ is considered as increasing while a trend below $-\frac{1}{2}e$ is considered as decreasing. For each node n in graph G (line 7), we check the quantity trend $qtyTrend$ of the current node n from the preceding snapshot to the current snapshot along 'SAME_AS' edge $typeSaS$ (line 8-9) and continue only if the quantity trend $qtyTrend$ is considered as increasing (line 10) (see also step 1 in Figure 5). We then collect all nodes n' with a 'POTENTIAL' edge $typePoE$ to the current node n (line 11-12). For each collected node n' (line 13), we check the quantity trend $qtyTrend$ from their current snapshot to the succeeding snapshot along their 'SAME_AS' edge $typeSaS$ (line 14-15) (see also step 2 in Figure 5). If the quantity trend $qtyTrend$ is considered as decreasing (line 16), we add a 'PREDICTED' edge $typePrE$ from the remaining collected nodes n' to the currently observed node n (line 17) (see also step 3 in Figure 5). The process is repeated until all nodes n have been viewed (line 7). As a result, we return the graph G with all added predicted likely occurring transformations (edge $typePrE$) (line 18).

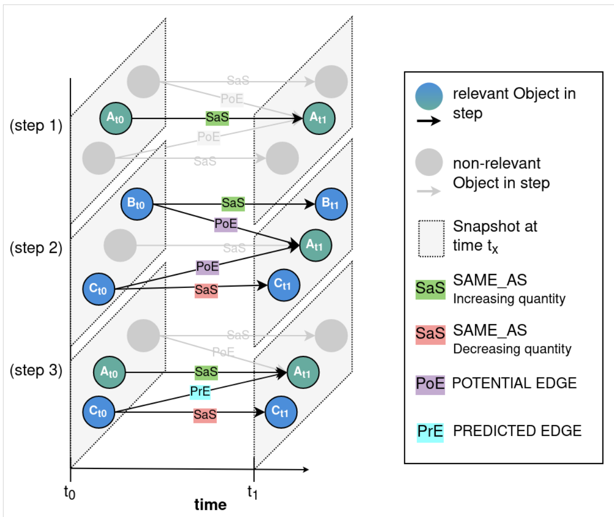


Figure 5: The 'Transformation Prediction Algorithm' follows iterative steps to predict the likely occurring transformations: (step 1) Select a node and continue if quantity trend at the incoming 'SAME_AS' edge increases. (step 2) Collect all nodes at the start of the incoming 'POTENTIAL' edges. (step 3) Keep all nodes with a decreasing quantity trend and add the new 'PREDICTED' edges next to the remaining 'POTENTIAL' edges.

We further assign each ‘PREDICTED’ edge a weight as edge property. The weight describes the most influential ‘PREDICTED’ edges that lead to an increase in the quantity of a node. For instance, if a node has two incoming ‘PREDICTED’ edges describing the increase in quantity, then the higher weight describes the edge with the greater influence on the increase. The weight is calculated during link prediction according to Formula 2:

$$weight = qty_{decr} * |qtyTrend_{decr}| \tag{2}$$

The formula for calculating the weighting requires two factors. The first factor is the quantity of the starting node at the ‘PREDICTED’ edge (i.e. the node with a decreasing quantity trend) (qty_{decr}). The second factor is the absolute value of the quantity trend at the ‘SAME_AS’ edge of the same node ($|qtyTrend_{decr}|$). Note that the weight of the edge increases with a high quantity trend and a high quantity. Each weight is further normalized based on the weights of all incoming edges of an end node.

5. Evaluation

We evaluate our graph model and algorithm based on two data sets from environmental chemistry. One of the data sets has been previously published and discussed in more detail [37, 14]. A qualitative evaluation of our approach within the considered domain is difficult because the considered system is complex and the processes that take place are largely unknown. Moreover, there is currently no comparable approach to track the transformations that are likely to occur. However, a test with a simplified systems, i.e. a chemical model with known reactions proved the general suitability of the model [14]. The suggestions generated by our approach cannot be easily verified by straight forward labeling. The predicted transformations rather serve the planning of new experiments by providing a novel perspective on the complex system and filling gaps between previously unconnected snapshot experiments.

5.1. Datasets

So called dissolved organic matter (DOM) plays an important role in freshwater, e.g., it alters available light for photosynthesis and must be removed during drinking water production [38]. DOM is a complex system with widely unknown processes and chemical relationships. Therefore, the investigation of DOM is a suitable use case to benefit from our approach. The extreme chemical diversity and reaction pathways make it difficult to identify the actual transformations. Ultra-high resolution mass spectrometry (UHRMS) can be used to qualitatively and quantitatively measure the composition of DOM. The molecules and their quantities measured over multiple time points form the data basis of the snapshot-based knowledge graph. Figure 1 showed an example of the structure of data generated by ultra-high resolution mass spectrometry (UHRMS) on DOM over multiple time points. The first dataset considered was generated from samples of a drinking water reservoir inflow. Within the samples, molecules were determined over 13 time points using natural sunlight to induce photochemical transformations. The second dataset originates from samples of a wastewater treatment plant outflow, in which the molecules were determined over 8 time points. Here, photochemical transformations were induced by an artificial light source.

5.2. Setup

We use Neo4j[1] as the graph database management system together with its query language Cypher to create the snapshot-based knowledge graph. Besides manipulating the graph structure (insert, update, delete), users of the graph database can describe patterns to interactively select matching parts in the graph. Both datasets were transformed into nodes and properties according to the snapshot-based knowledge graph model and loaded into the graph database. Programming and preprocessing was done in Python along with the neo4j library, an official Neo4j driver that serves as an interface between the Neo4j graph database and Python. Both datasets are configured the same and processed with the same hardware and an error margin of 5 %. The open source project is publicly available on GitHub[2].

5.3. Results and Discussion

A snapshot-based knowledge graph is created from each of the two datasets. The nodes in the graphs describe the molecules identified at a particular discrete time in the mass spectrum. Each node has several properties, e.g. the chemical formula, the atoms it consists of (i.e., C,H,O,N,S) and the quantity. Between each snapshot the molecules with the same chemical formula are connected by the 'SAME_AS' edges and are enriched with the quantity trend. The 'POTENTIAL' edges explain the potential chemical transformations between molecules of successive snapshots and are based on 22 different photochemical transformations considered [14]. The 'PREDICTED' edges describe the likely occurring chemical transformations calculated with the 'Transformation Prediction Algorithm' (see Algorithm 1). Note that in the current use-case of mass spectrometry data, the algorithm only provides relative trends of molecule abundances, not actual amounts changed. Figure 6 shows parts of the first three snapshots of the created graph based on dataset 1. Before using the 'Transformation Prediction Algorithm', it was not possible to determine the 'PREDICTED' edges drawn in yellow. Those novel edges enable new analyses and evaluations on the datasets.

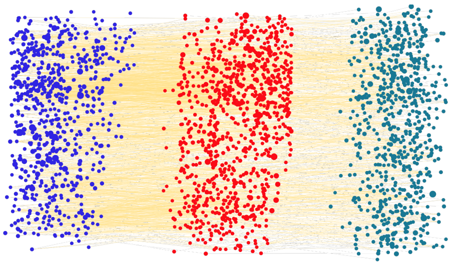


Figure 6: Three snapshots of the temporal knowledge graph. The yellow edges indicate the 'PREDICTED' edges between the nodes.

[1] https://neo4j.com/

[2] https://github.com/PhPlam/tegrom

Table 1
Summary of the snapshot-based knowledge graphs.

	dataset 1	dataset 2
snapshots	13	8
nodes (max)	4671	5935
nodes (min)	2973	2372
nodes (avg)	3578	4861
”SAME_AS” edges (max)	4161	5250
”SAME_AS” edges (min)	2685	2285
”SAME_AS” edges (avg)	3183	4336
qtyTrend increasing (avg)	1029	2010
qtyTrend consistent (avg)	675	587
qtyTrend decreasing (avg)	1479	1739
”PREDICTED” edges (max)	3685	8911
”PREDICTED” edges (min)	2073	2847
”PREDICTED” edges (avg)	2620	6853

Table 1 compares the nodes and edges of the generated snapshot-based knowledge graphs from the two datasets. Dataset 1 has 13 snapshots while dataset 2 has 8 snapshots matching the number of time points in the datasets. Dataset 2 has on average more 36 % more nodes per snapshot than dataset 1 and also a 27 % higher maximum number of nodes in a snapshot. In contrast the minimum number of nodes in a snapshot is 25 % lower in dataset 2 than in dataset 1 reflecting a more pronounced effect of the artificial irradiation. Dataset 2 has on average 36 % more ‘SAME_AS’ edges than dataset 1. The maximum (minimum) number of ‘SAME_AS’ edges between two snapshots is about 26 % higher (18 % lower) in dataset 2 than in dataset 1. We observe the three possible quantity trends along ‘SAME_AS’ edges. Averaged over all snapshots, dataset 1 (2) has 32 % (46 %) increasing, 21 % (14 %) consistent and 47 % (40 %) decreasing quantity trends. The differences in the fractions of increasing and decreasing quantity trends matches the different maximum and minimum node values and is consistent with the stronger photochemical transformations induced by the artificial light (dataset 2) as compared to the natural light (dataset 1).

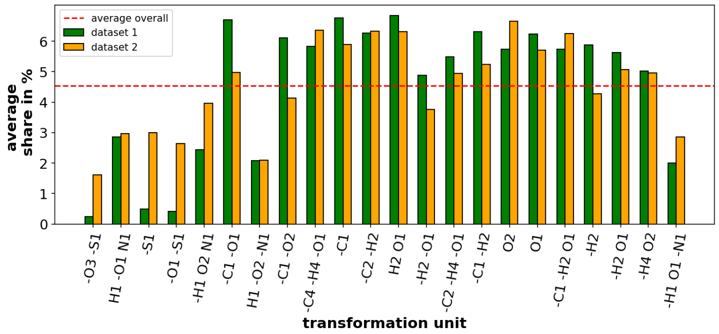


Figure 7: Comparison of the average shares of the transformation units reflected by the ‘PREDICTED’ edges of both datasets.

Using the ‘Transformation Prediction Algorithm’, we are able to identify novel, previously unknown ‘PREDICTED’ edges for both datasets (Table 1 bottom section). Dataset 2 has on average about factor 2.6 more ‘PREDICTED’ edges than dataset 1. The maximum (minimum) number of ‘PREDICTED’ edges between two snapshots is about 2.4 (1.4) times higher in dataset 2 than in dataset 1. Note, that we consider a closed system, i.e. no quantities enter from outside or leave the system, making the predicted edges particularly meaningful. The model results correspond to the expected behavior of molecules in the two datasets and will be analysed in more detail below.

Each of the previously unknown ‘PREDICTED’ edges reflects one of the 22 different photochemical transformations considered, which are identified by the ‘Transformation Prediction Algorithm’ as likely to have occurred. Figure 7 shows the average share of these transformations in both datasets. For instance, the figure depicts that in the sample of dataset 1, the transformations ”-C1 -O1”, ”-C1”, and ”H2 O1” occurred most frequently, while in the sample of dataset 2, the transformations ”O2”, ”-C4 -H4 -O1”, and ”-C2 -H2” occurred most frequently. The comparison of the two datasets reveals that the distribution of the transformations in the two samples and experiments is different. For instance, the share of ”-S1” is much higher (about 606 %) in dataset 2 as compared to dataset 1. This is expected since dataset 2 originates from a wastewater treatment plant outflow, in which a higher amount of sulfur-(S)-containing tensides is expected than in dataset 1, which originates from a pristine drinking water reservoir inflow. A higher share of transformations involving sulfur, i.e. ”-O3 -S1”, ”-S1” and ”-O1 -S1”, is thus expected reflecting pronounced tensides degradation over time. With this example the ‘Transformation Prediction Algorithm’ shows its basic ability to make reasonable predictions for the likely occurring transformations within the sample.

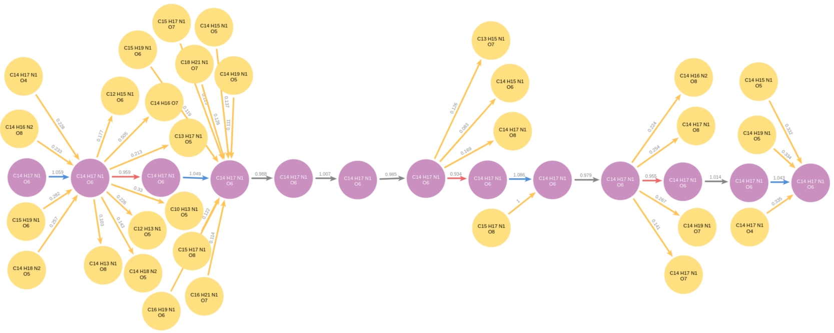


Figure 8: View of a molecule over time (purple nodes) and the predicted edges with weights.

The snapshot-based knowledge graph allows instant evaluation of the previously unknown chemical transformations and easy access to trends and time courses with interactive queries in the graph database. The result of the database query in Figure 8 shows the evolution of a molecule (purple nodes) over time. Blue arrows represent an increasing quantity, red arrows represent a decreasing quantity, and gray arrows represent a constant quantity. The incoming and outgoing weighted yellow arrows represent the predicted chemical transformations. The example shows a way to create more interactive queries that help to capture the structure and relationships within the data.

6. Conclusion

We presented a novel snapshot-based knowledge graph model that allows to transform multivariate time series data into a feature-rich complex and highly connected network. The nodes and edges store not only static but also time-dependent properties. Starting with a set of potential transformations as they are inherent in many domains such as complex chemical networks, our temporal link prediction algorithm calculates new edges between nodes representing the likely occurring predicted transformations in consecutive snapshots. We evaluated the snapshot-based knowledge graph model and temporal link prediction algorithm in an environmental chemistry use case and demonstrated its functionality to represent graph-like data into a complex network and identify previously unknown relationships.

The temporal graph model presented in this study is not limited to specific applications and can be readily extended to other systems and processes that benefit from the enrichment with temporal edges. To apply and comparatively evaluate the model and algorithms for other domains in the future, we will develop a benchmark including gold standard datasets for link prediction in different complex networks. Moreover, we will test different approaches such as answer set programming, logical reasoning and graph neural networks to be compared with the current approach. By modeling large temporal datasets from multivariate time series as temporal graphs, further graph-based algorithms can be used to perform comprehensive analyses. Future extensions of the approach could incorporate temporal clustering to analyse temporal patterns and dynamics of recognized communities.

Acknowledgments

We thank Alexander Sobolev and Christin Wilske for conducting the photodegradation experiments and collecting samples. We also thank Jan Kaesler for performing FT-ICR-MS measurements and Peter Herzsprung for supporting data processing.

References

[1] R. Albert, A.-L. Barabási, Statistical mechanics of complex networks, Reviews of Modern Physics 74 (2002) 47–97. doi:10.1103/revmodphys.74.47.

[2] A. Hogan, E. Blomqvist, M. Cochez, C. D'amato, G. D. Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier, A.-C. N. Ngomo, A. Polleres, S. M. Rashid, A. Rula, L. Schmelzeisen, J. Sequeda, S. Staab, A. Zimmermann, Knowledge graphs, ACM Computing Surveys 54 (2021) 1–37. doi:10.1145/3447772.

[3] M. Hofer, D. Obraczka, A. Saeedi, H. Köpcke, E. Rahm, Construction of knowledge graphs: State and challenges, arXiv preprint arXiv:2302.11509 (2023). doi:10.48550/ARXIV.2302.11509.

[4] N. Noy, Y. Gao, A. Jain, A. Narayanan, A. Patterson, J. Taylor, Industry-scale knowledge graphs: Lessons and challenges, Queue 17 (2019) 48–75. doi:10.1145/3329781.3332266.

[5] L. da F. Costa, F. A. Rodrigues, G. Travieso, P. R. V. Boas, Characterization of complex

networks: A survey of measurements, Advances in Physics 56 (2007) 167–242. doi:10.1080/00018730601170527.

[6] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, D.-U. Hwang, Complex networks: Structure and dynamics, Physics Reports 424 (2006) 175–308. doi:10.1016/j.physrep.2005.10.009.

[7] S. Sakr, A. Bonifati, H. Voigt, A. Iosup, K. Ammar, R. Angles, W. Aref, M. Arenas, M. Besta, P. A. Boncz, K. Daudjee, E. D. Valle, S. Dumbrava, O. Hartig, B. Haslhofer, T. Hegeman, J. Hidders, K. Hose, A. Iamnitchi, V. Kalavri, H. Kapp, W. Martens, M. T. Özsu, E. Peukert, S. Plantikow, M. Ragab, M. R. Ripeanu, S. Salihoglu, C. Schulz, P. Selmer, J. F. Sequeda, J. Shinavier, G. Szárnyas, R. Tommasini, A. Tumeo, A. Uta, A. L. Varbanescu, H.-Y. Wu, N. Yakovets, D. Yan, E. Yoneki, The future is big graphs, Communications of the ACM 64 (2021) 62–71. doi:10.1145/3434642.

[8] X. Zou, A survey on application of knowledge graph, Journal of Physics: Conference Series 1487 (2020) 012016. doi:10.1088/1742-6596/1487/1/012016.

[9] P. Holme, Modern temporal network theory: a colloquium, The European Physical Journal B 88 (2015). doi:10.1140/epjb/e2015-60657-4.

[10] R. H. Shumway, D. S. Stoffer, Time Series Analysis and Its Applications, Springer International Publishing, 2017. doi:10.1007/978-3-319-52452-8.

[11] Y. Zou, R. V. Donner, N. Marwan, J. F. Donges, J. Kurths, Complex network approaches to nonlinear time series analysis, Physics Reports 787 (2019) 1–97. doi:10.1016/j.physrep.2018.10.005.

[12] M. Jiang, X. Gao, H. An, H. Li, B. Sun, Reconstructing complex network for characterizing the time-varying causality evolution behavior of multivariate time series, Scientific Reports 7 (2017). doi:10.1038/s41598-017-10759-3.

[13] L. Lacasa, V. Nicosia, V. Latora, Network structure of multivariate time series, Scientific Reports 5 (2015). doi:10.1038/srep15508.

[14] P. Plamper, O. J. Lechtenfeld, P. Herzsprung, A. Groß, A temporal graph model to predict chemical transformations in complex dissolved organic matter, Environmental Science & Technology (2023). doi:10.1021/acs.est.3c00351.

[15] I. Robinson, Graph databases new opportunities for connected data, 2015.

[16] M. A. Rodriguez, P. Neubauer, Constructions from dots and lines (2010). doi:10.48550/ARXIV.1006.2361.

[17] O. Hartig, Reconciliation of rdf* and property graphs, arXiv preprint arXiv:1409.3288 (2014). doi:10.48550/ARXIV.1409.3288.

[18] R. Angles, The property graph database model., in: AMW, 2018.

[19] M. Ciglan, A. Averbuch, L. Hluchy, Benchmarking traversal operations over graph databases, in: 2012 IEEE 28th International Conference on Data Engineering Workshops, IEEE, 2012. doi:10.1109/icdew.2012.47.

[20] R. kumar Kaliyar, Graph databases: A survey, in: International Conference on Computing, Communication & Automation, IEEE, 2015. doi:10.1109/ccaa.2015.7148480.

[21] B. Cai, Y. Xiang, L. Gao, H. Zhang, Y. Li, J. Li, Temporal knowledge graph completion: A survey, 2022. doi:10.48550/ARXIV.2201.08236.

[22] P. Holme, J. Saramäki, Temporal networks, Physics Reports 519 (2012) 97–125. doi:10.1016/j.physrep.2012.03.001.

[23] A. Debrouvier, E. Parodi, M. Perazzo, V. Soliani, A. Vaisman, A model and query language for temporal graph databases, The VLDB Journal 30 (2021) 825–858. doi:10.1007/s00778-021-00675-4.

[24] A. Ferreira, Building a reference combinatorial model for MANETs, IEEE Network 18 (2004) 24–29. doi:10.1109/mnet.2004.1337732.

[25] C. Cattuto, M. Quaggiotto, A. Panisson, A. Averbuch, Time-varying social networks in a graph database, in: First International Workshop on Graph Data Management Experiences and Systems, ACM, 2013. doi:10.1145/2484425.2484442.

[26] C. Rost, K. Gomez, M. Täschner, P. Fritzsche, L. Schons, L. Christ, T. Adameit, M. Junghanns, E. Rahm, Distributed temporal graph analytics with GRADOOP, The VLDB Journal 31 (2021) 375–401. doi:10.1007/s00778-021-00667-4.

[27] P. Basu, A. Bar-Noy, R. Ramanathan, M. P. Johnson, Modeling and analysis of time-varying graphs, arXiv preprint arXiv:1012.0260 (2010). doi:10.48550/ARXIV.1012.0260.

[28] L. Getoor, C. P. Diehl, Link mining, ACM SIGKDD Explorations Newsletter 7 (2005) 3–12. doi:10.1145/1117454.1117456.

[29] Z. Chen, Y. Wang, B. Zhao, J. Cheng, X. Zhao, Z. Duan, Knowledge graph completion: A review, IEEE Access 8 (2020) 192435–192456. doi:10.1109/access.2020.3030076.

[30] Q. Wang, Z. Mao, B. Wang, L. Guo, Knowledge graph embedding: A survey of approaches and applications, IEEE Transactions on Knowledge and Data Engineering 29 (2017) 2724–2743. doi:10.1109/tkde.2017.2754499.

[31] M. Al Hasan, V. Chaoji, S. Salem, M. Zaki, Link prediction using supervised learning, in: SDM06: workshop on link analysis, counter-terrorism and security, volume 30, 2006, pp. 798–805.

[32] L. Lü, T. Zhou, Link prediction in complex networks: A survey, Physica A: Statistical Mechanics and its Applications 390 (2011) 1150–1170. doi:10.1016/j.physa.2010.11.027.

[33] A. Divakaran, A. Mohan, Temporal link prediction: A survey, New Generation Computing 38 (2019) 213–258. doi:10.1007/s00354-019-00065-z.

[34] L. Lacasa, B. Luque, F. Ballesteros, J. Luque, J. C. Nuño, From time series to complex networks: The visibility graph, Proceedings of the National Academy of Sciences 105 (2008) 4972–4975. doi:10.1073/pnas.0709247105.

[35] N. Marwan, J. F. Donges, Y. Zou, R. V. Donner, J. Kurths, Complex network approach for recurrence analysis of time series, Physics Letters A 373 (2009) 4246–4254. doi:10.1016/j.physleta.2009.09.042.

[36] F. Hasselman, A. M. T. Bosman, Studying complex adaptive systems with internal states: A recurrence network approach to the analysis of multivariate time-series data representing self-reports of human experience, Frontiers in Applied Mathematics and Statistics 6 (2020). doi:10.3389/fams.2020.00009.

[37] C. Wilske, P. Herzsprung, O. J. Lechtenfeld, N. Kamjunke, W. von Tümpling, Photochemically induced changes of dissolved organic matter in a humic-rich and forested stream, Water 12 (2020) 331. doi:10.3390/w12020331.

[38] J. A. Leenheer, J.-P. Croué, Peer reviewed: Characterizing aquatic dissolved organic matter, Environmental Science & Technology 37 (2003) 18A–26A. doi:10.1021/es032333c.