Modeling Health Stage Development of Patients with Dynamic Attributed Graphs in Online Health Communities

Yuyang Gao*, Tanmoy Chowdhury*, Lingfei Wu, and Liang Zhao

Abstract—In this paper, we propose a novel DynAttGraph2Seq framework to model complex dynamic transitions of an individual user's activities and the textual information of the posts over time in online health forums and learning how these correspond to the health stage development. To achieve this, we first formulate the transition of user activities as a dynamic attributed graph with multi-attributed nodes that evolves over time, then formalize the health stage inference task as a dynamic attributed graph to sequence learning problem. Our proposed model consists of a novel dynamic graph encoder along with a two-level sequential encoder to capture the semantic features from user posts and an interpretable sequence decoder that learn the mapping between a sequence of time-evolving user activity graphs as well as user posts to a sequence of target health stages. We go on to propose new dynamic graph regularization and dynamic graph hierarchical attention mechanisms to facilitate the necessary multi-level interpretability. A comprehensive experimental analysis on health stage prediction tasks demonstrates both the effectiveness and the interpretability of the proposed models.

Index Terms—Deep Learning, Dynamic Graph, Sequence Prediction, Health Stage Development.

1 INTRODUCTION

NLINE healthcare forums and communities such as the Breast Cancer Community [1], American Bladder Cancer Society [2], and eHealth Forum [3] have become an important channel for patients to share and acquire health-related information and have greatly changes patients' daily lives. Unlike traditional approaches, where patients only receive information about their disease from their care providers, these online forums and communities have enabled millions of patients to ask questions related to their diseases, interact with other patients with similar prognoses, and provide support to each other across the world. The communications and interactions between patients in online forums can provide valuable information about a patient's emotional well-being and behaviors related to the management of their health that conventional clinical data collected from hospital information systems and electronic health records (EHR) is unable to capture. Moreover, beyond conventional online communities and social media, online health communities provide a unique way to analyze and infer patients' health stages and disease history. Figure 1 shows an example of a patient's health stage information extracted from a patient signature that contains the cancer diagnosis and treatment history, along with the relevant dates. In all, the synergies between the information on patients' online communication and health status make possible a unique and wide range of research topics on health informatics, such as patient behavior statistical analyses [4], [5], longitudinal communication network analyses [6], and patient participation prediction [7], all of which rely on both patients' online participation and their health stage records.

However, the health stage information in the online

health community has some unique challenges and characteristics. First, though some patients share their disease history, as shown in Figure 1, such information is not provided or is simply missing for many others. For instance, over 36% active users that registered within recent 2 years have not yet shared their disease history in the Breast Cancer Community [1]. Important information about patients' health stages can significantly facilitate comprehensive investigations about patients' health conditions and thus it is highly desirable to be able to infer or predict these patients' health stage information. Second, different subforums under specific topics are often correlated to specific disease stages. For example, in the online breast cancer forum, the patients who are active in the "Chemotherapy - Before, During, and After" subforum typically look for information related to their Chemotherapy treatment. Similarly, user's communication can convey schematic and informative information about the user's current and past health status. For example, the user post shown in Figure 1, especially the words marked in blue, contains rich information about the patient's health status. Thus, users' activities within these subforums could serve as a strong indicator of an individual user's current health stage. Third, as the patients' health conditions progress over time, they often move from one set of subforums to others that are more related to their new health stages. Therefore, for each patient, these transitions among subforums can lead to an inter-connected subforum activity network that evolves over time, which could be highly entangled with the progress of patient's health status or disease stage, as shown in Figure 2.

The ability to accurately infer users' missing health stage information is crucial, as it can benefit patients, health care organizations, and clinical research. Specifically, the benefits for accurately inferring users' missing health stage informa-

[•] Authors contributed equally

Hey all! I was diagnosed with IDC triple positive (er+, pr+, HER2+) Chemo worked well, ... After chemo the same 2 lymph nodes were still enlarged. Since they don't know, they have to assume they are positive with cancer and for that reason I will be doing 6 1/2 weeks of radiation. Does anyone else have the same thing going on? ...

Dx 9/24/2014, IDC, 2cm, Grade 2, ER+/PR+, HER2+ Targeted Therapy 10/27/2014 Herceptin (trastuzumab) Chemotherapy 10/27/2014 Carboplatin (Paraplatin), Taxotere (docetaxel) Targeted Therapy 10/27/2014 Perjeta (pertuzumab)

Fig. 1: An example of a patient post in the breast cancer forum along with the patient's signature that reveals the health status. The words that are highlighted in blue color such as 'IDC' and 'Chemo' in the post clearly show the correspondence to the user's health status in the signature.

tion via their forum activities are three-fold:

- 1) **For patients:** The predicted health stage information can help patients to establish potential connections to other patients with similar health status to provide support and offer help and information with each other.
- 2) For health care organizations: The health care organizations can be benefited from providing more personalized guidance and appropriate resources to each individual patient according to the particular health stage [8], [9].
- 3) For clinical research: Because of the large ratio of missing data on patient health stages information, our model can be used to provide label supports to benefit other clinical research such as health condition prediction [10] and health risk prediction [11], [12].

To infer the missing user health stage information, the correspondence between the users' forum activities (e.g. the posts and replies, and the movement from one subforum to another) and their health stage history needs to be accurately identified and modeled. Naturally, the networked and time-evolving forum activity data can be formulated as dynamic sequence of user posts and user activity transition graphs that change over time. In addition, the target user health stage history can be formulated as a sequence prediction problem that needs to be inferred from both the dynamic graph sequence and text sequence. Thus, without loss of generality, a new generic task is presented here where the goal is to learn the mapping from a sequence of graph-structured data and a sequence of text data to a target sequence. In this paper, we limit our scope to the domain of online health forums and focus on health stage sequence prediction based on online health forums data.

However, capturing the high-level mapping between the evolution of the user activity networks and the changes in the corresponding user's health stage cannot be easily handled by existing techniques due to the following four challenges: **1) Difficulty in modeling the forum data, which is dynamic, networked, and multi-attributed.** A user's activities in the various subforums can change dynamically over time and these activity transitions naturally bridge different subforums. In addition, each subforum contains both unique and shared content, and identifying how this content is shared among subforums is important. **2) Difficulty in learning the association between a sequence of user activity networks and the corresponding sequence of health stages**. The sequence of user activity networks contains complicated graph-structured information that dynamically evolves over time. Developing end-to-end learning between such dynamic complex data and a specific sequence is highly difficult. 3) Difficulty in jointly modeling the dynamic evolution of both the textual information in users' posts and their forum activities. The posts of the users also contain crucial information about their health status over time. For example in Figure 2, we can observe that both the user posts and the corresponding forum activity graphs are good indicators of the evolution of the user health status over time, i.e. from 'Chemotherapy' to 'Radiation Therapy'. However, capturing the complex correspondences and evolution of both user posts and activity networks as a whole is very challenging and haven't been explored before. 4) Lack of interpretability of the health stage sequence inference process. The sequence of user activity networks has a twolevel hierarchical structure, namely from the node (i.e., subforum) level to the network level, and from the network level to the health stage level. It is thus a major objective to incorporate this hierarchical structural information into the development of an interpretable health stage inference process.

In this paper, we formally define the generic learning problem of health stage sequence inference using online forum data and propose the first framework to address the aforementioned challenges effectively. The main contributions are as follows:

- Defining the novel problem of inferring user health stage information using dynamic user behavior graph and text information extracted from online health forum data. We define the health stage inference problem in online health forums and formulate the user activities as transition graphs and time evolving text sequence that are capable of modeling user dynamic transitions between subforums and their complex relationships.
- 2) Proposing a generic DynAttGraph2Seq framework for inferring target sequence from a sequence of attributed graphs and texts. We propose a novel deep neural encoder-decoder framework for learning the mapping between complex dynamic graph and text sequence inputs and the target output sequence.
- 3) Proposing dynamic graph regularization and a dynamic graph hierarchical attention mechanism for enhancing model effectiveness and interpretability. We propose a dynamic graph regularization that enforces the smooth learning of consecutive graphs while preserving the heterogeneity across the graph sequence. In addition, we propose a new dynamic graph hierarchical attention mechanism that captures both the time-level and node-level attention, thus providing model transparency throughout the whole inference process.
- 4) Proposing sequential encoder to capture user post level semantic features and bidirectional temporal dependencies from users' communication. We propose a two level bidirectional sequential encoder which used bi-directional long short-term memory neural network to capture both the semantic information from user posts as well as their dynamic evolution over time. This network enabled the feature learning of both semantic





Fig. 2: An example of user forum activities and the corresponding health stage evolution. In the first two time windows, the user is mainly active in Subforum #13 while going through chemotherapy treatment. In the third time window, the user starts to be active in Subforum #22 at about the time when she undergoes IDC treatment. Finally, in the last three time windows, the user becomes active in Subforum #14 when she enters the "Radiation Therapy" health treatment stage.

and dynamic characteristics from the user posts.

5) Conducting comprehensive experiments and case studies to validate the effectiveness and interpretability of the proposed model. Experiments on online health forum dataset demonstrate that our proposed models outperform conventional sequence inference methods. In addition, our qualitative analyses and case studies provide interpretable insights into the learning results of the proposed model and its variations.

The rest of the paper is organized as follows. Section 2 outlines related research work in online health communities analysis, dynamic graph representation learning, hierarchical attention mechanism, and neural encoder-decoder models. Section 3 represents the problem formulation of dynamic graph, user post and sequential output of user health stage. Section 4 describes in details of each component of the proposed DynAttGraph2Seq framework. Dynamic graph encoder was introduced in Section 4.2 along with regularization that enforces the smooth learning of consecutive graphs while preserving the heterogeneity across the graph sequence. In Section 4.3, a two-level post encoder was proposed which enabled the feature learning of both semantic and dynamic characteristics from the user posts. Finally, we did ample amount of experimental analysis and case studies in Section 5 and demonstrated the effectiveness of proposed framework.

2 **RELATED WORK**

2.1 **Online Health Communities Analysis**

A number of studies have focused on the analysis and utilization of online health communities data. Popular social media such as Twitter and Facebook are good for aggregate level pattern mining tasks such as discovering epidemic outbreaks [13] and other type of events [14], [15]. However, compared to specialized health forums such as the Breast Cancer Forum, their power is limited for discovering individual-level health stages and health network patterns due to the privacy issues involved and data scarcity. There have been several analyses of breast cancer forum data [4], [5] and, more recently, machine learning models have been used for longitudinal analysis [6] as well as for some binary

classification tasks such as patient participation prediction [7] and cancer type classification [9]. However, we are the first to propose a general framework that can achieve health stage sequence inference using online forum data.

2.2 Dynamic Graph Representation Learning

As an emerging topic in the graph representation learning domain, dynamic graph learning has attracted a great deal of attention from researchers in recent years [16]. Most of the current research falls into the dynamic graph embedding domain. Some of the proposed methods intuitively extend the idea from static graph embedding approaches by adding regularization [17], [18], while others propose specific models for capturing dynamic characteristics of the graph [19], [20]. There has also been some work on streaming graph learning [21], [22], where network representations are learned dynamically as the network evolves. Furthermore, dynamic graph is also demonstrated to be helpful for modeling the inter-dependency relationships among nodes in time series data [23], [24]. However, these graph embedding techniques typically focus on learning representations of the graphs, such as node embedding, but in many real-world applications the aim is to learn some high-level knowledge from the graph data, such as graph classification tasks [25], [26] and graph to sequence tasks [27]. An end-to-end learning model is thus needed to learn the mapping between the whole sequence of graph data and the target output sequence, instead of merely focusing on learning node representations.

2.3 **Hierarchical Attention Mechanism**

The attention mechanism first proposed by [28], [29] was used for machine translation tasks. Here, the attention mechanism was used to select reference words in the original language that matched specific words in the foreign language before translation. Later on, Hierarchical Attention Networks [30] was proposed to model the natural hierarchical structure of word-to-sentence and sentence-to-document level attention in document classification tasks. Besides being widely used for machine translation, the attention mechanism has also been introduced in the graph representation learning domain. Graph Attention Networks [31] introduced node-to-node attention mechanism for graph

embedding, and many others followed and extended this idea [32], [33]. However, there is little to no work on studying the unique hierarchical structure that is naturally present in dynamic graphs.

2.4 Neural Encoder-Decoder Models

The neural encoder-decoder models originally designed to solve neural machine translation problems [28] have been widely extended to model the mapping of general object inputs to their corresponding sequences. A major focus has been on addressing the limitations of Seq2Seq when dealing with more complex objects, including Tree2Seq [34], Set2Seq [35], and TreeStructured LSTM [36]. Recent advances in graph deep learning and graph convolutional networks have enabled researchers to utilize various graph deep learning models to handle challenges in the domains of machine translation and graph generation [37], [38]. Most recently, the graph2seq models [39], [40] were proposed as a general-purpose encoder-decoder model for graphto-sequence learning. However, to the best of our knowledge, as yet there is no work that explores dynamic graph to sequence learning, where the natural sequential order contained in a dynamic graph and its sequences might be advantageous for neural encoder-decoder models.

3 PROBLEM FORMULATION

The online health forum data typical consists of complex user activities and textual information among different subforums as shown in Figure 2, which is hard to model directly. In this section, we formally define the learning problem of health stage sequence prediction with the complex online forum data. Specifically, in Section 3.1 we formulate the user activity transition among sub-forums as dynamic attributed graphs. We then move on to formulate the user posts as text sequence over time in Section 3.2. Finally, we formally define the learning problem of health stage sequence prediction as a dynamic attributed graph to sequence learning problem in Section 3.3.

3.1 User Forum Activities as a Dynamic Graph

The online forum data records the path of each user's transition from one subforum to another, as well as their activities within each subforum. In order to capture these complex transitions and model the relationships between subforums, we propose a novel method to formulate the raw user subforum activities into activity transition networks that preserve these characteristics.

An activity transition network is formulated naturally as follows. User activities are first partitioned into a series of time windows. We then begin by formulating a node for each subforum, with a transition from one forum to the other deemed to occur if the most active forum (based on visiting time or number of postings) switches from the former to the latter, creating a directed 'edge' between them. Each node (i.e., subforum) also records the user activity in the forum to build the activity transition network. For example in Figure 2, the subforum transition sequence is $\{30 \rightarrow 13 \rightarrow 6 \rightarrow 29\}$, where 30, 13, 6, and 29 are the IDs of the subforums visited. Thus, the transition edges for the first snapshot graph will be (30,13), (13,6), and (6,29). The graph in each time window records all the transitions in and previous to it.

Naturally, such time-ordered activity transition networks can be formally defined as dynamic graphs, also known as temporal networks in the network science literature [41], that capture the complex dynamic characteristics and time-evolving features of graphs, as defined in the following.

Definition 1. (dynamic attributed graph). A dynamic attributed graph $\mathcal{G} = \{G_1, G_2, \dots, G_T\}$ is an ordered sequence of $t = 1, \dots, T$ separate graphs on the same set of |V| = N nodes, with each snapshot graph $G_t(V, E_t)$ characterized by a weighted adjacency matrix $A_t \in \mathbb{R}^{N \times N}$ and a set of node features $F_t \in \mathbb{R}^{N \times D}$ for a given time window, where D represents the total number of node features.

We can now formulate the activity transition networks as a dynamic attributed graph, illustrated in Figure 2. Here, the dynamic attributed graph contains a sequence of snapshot graphs G_1, G_2, \cdots, G_6 that characterize user activities in the online forum for a given time period, where G_t represents the snapshot graph $G_t(V, E_t)$ for simplicity. In this example, the time windows are shown as blue boxes. Each node $v \in V$ represents a subforum devoted to a specific topic and the edges E_t capture the user's movement between different subforums at a given time window. In addition, each node v contains a set of features $F_{t,v}$ that represents the topics covered by the specific subforum. By formulating user online forum activities as dynamic graphs, the mapping between the evolution of the user activity graphs and the changes of the corresponding user's health stages will be preserved.

3.2 User Posts as Encoded Sequence

In the forum, users' post convey good amount of information about the evolution of their health status. For example in Figure 3, P_2 shows an example of several posts from a user in timestep 2. We can see from this time period, the patient is going through the diagnosis testing, and few days later the patient also finished chemotherapy treatment. Thus, the posts can be crucial for the health status prediction task as it not only contains rich semantic information that depicts patient's current health stages, but also conveys the evolution of the health status of the user over time.

To capture both the semantic and time-evolving information from the user post, we propose a novel two-level bidirectional encoder. As shown in Figure 3, we denote $\mathcal{P} = \{P_1, P_2, \dots, P_T\}$ as an ordered sequence of the collection of the user posts over time using the same time partition as in \mathcal{G} . Furthermore, each word tokens q in the accumulated posts t is further projected into the corresponding word vectors by a pre-trained word2vec model [42], denoted as $P_t = \{P_{t,1}, P_{t,2}, \dots, P_{t,Q}\}$, where Q is the total number of word tokens within the accumulated posts P_t . The word vectors, $P_{t,q}$ will be used as the input to the model to learn the semantic and contextual information encoded in user posts within a specific time period t.

3.3 Learning Sequence from Dynamic Graph and Post

As we can see from Figure 2, there is a clear mapping between the evolution of the user activity dynamic graph



Fig. 3: The proposed end-to-end Dynamic Attributed Graph to Sequence learning (DynAttGraph2Seq) framework. It includes a novel dynamic graph encoder, a two-level user post encoder, and a sequence decoder. The proposed framework can not only encode the complicated interactions of user's activities in the dynamic graphs, but also capture the timeevolving semantic features from user posts. Moreover, it provides both time-level and subforum level interpretability of the correlations between a user's online forum activities and their current health stages through the proposed two-level attention mechanism.

and changes in the corresponding user's health stage. In addition, from the user posts example shown in Figure 3, we can clearly see the user posts also contain critical semantic information about the user's current health status. Motivated by these observations, we propose to formulate such learning problems as follows:

Given a dynamic graph $\mathcal{G} = \{G_1, G_2, \cdots, G_T\}$ and the dynamic user posts $\mathcal{P} = \{P_1, P_2, \cdots, P_T\}$ as input data for each user, the goal is to predict the target sequence $S = \{s_1, s_2, \cdots, s_M\}$, where M is the output sequence length and $s_m \in \mathbb{V}$ is the *m*th token of the output sequence in vocabulary \mathbb{V} that consists of health stages of patients including 'Dx'¹, 'Chemotherapy', 'Targeted therapy', 'Hormonal therapy', 'Radiation therapy', and 'Surgery'. An example of S could be $\{'Dx', 'Targeted Therapy', 'Chemotherapy', 'Targeted Therapy', 'Is shown in Figure 1. Formally, this problem is equivalent to learning a translation mapping from input dynamic graph <math>\mathcal{G}$ and the dynamic user posts \mathcal{P} to a sequence S as $\{G_1, G_2, \cdots, G_T\} \cup \{P_1, P_2, \cdots, P_T\} \rightarrow \{s_1, s_2, \cdots, s_M\}$.

The translation mapping problem between some source objects and target sequences is widely studied, including both graph-to-sequence [39] and sequence-to-sequence [43] formulations. However, learning the translation from the user activity in the online forum to the user stage sequence is a more complex problem as it involves learning of both the dynamic behavior graphs that describe the user's activity transition between subforums as well as the texts that capture the semantic features from user post. Specifically, the challenges are four-fold: 1) Difficulty in modeling the forum data, which is dynamic, networked, and multi-attributed; 2) Difficulty in learning the association between a sequence of user activity networks and the corresponding sequence of health stages; 3) Difficulty in jointly modeling the dynamic evolution of both the textual information in users' posts and their forum activities. 4) Lack of interpretability of the health

stage sequence inference process.

4 MODEL

4.1 The DynAttGraph2Seq Framework

In this section, we introduce our proposed model framework that includes three main components: the dynamic graph encoder, the dynamic user post encoder, and the sequence decoder, as shown in Figure 3. To the best of our knowledge, this is the first end-to-end dynamic graph-tosequence learning framework. Our new DynAttGraph2Seq framework enables the learning paradigm of generating target sequence outputs by capturing both the complicated dynamic interactions between a user's activities over time from the dynamic attributed graph, and bi-directional semantic features evolution from user posts. In addition, the dynamic graph encoder also provides both time-level and subforum level interpretability of the correlations between a user's online forum activities and that user's current health stages through our two-level attention mechanisms.

For capturing the complex relationships represented in the graph input and the dynamic changes represented by the whole sequence of the dynamic graph, we propose a dynamic graph encoder that consists of three main components as follows: the first component contains a sequence of graph convolutional networks that learns the node level embeddings h_t for each graph snapshot G_t ; the learned node level embeddings are then aggregated into a graph level embedding g_t by an aggregation function; finally, a sequence encoder is used to take the learned graph level embedding sequence $g = \{g_1, g_2, \cdots, g_T\}$ and generate a sequence of patient health stages that capture the entire dynamic graph characteristics. In addition, we propose a novel dynamic graph regularization for sparse feature selection of the graph convolutional networks that enforces smooth feature selection for consecutive snapshot graphs locally, while at the same time preserving the heterogeneity of the features selected across the entire dynamic graph globally.

To capture both the semantic and temporal information from user posts, we propose a novel two-level sequential

^{1.} Short for Oncotype DX test, an initial diagnosis that analyzes how a cancer is likely to behave and respond to treatment.

encoder. As shown in Figure 3, the first level encoder, i.e. post level encoder, consists of a set of biLSTM networks that learn the semantic information from a set of user posts c_t within each timestep t; then the learned post level embedding sequence $c = \{c_1, c_2, \cdots, c_t\}$ is further feed into the second level encoder, i.e. user level encoder, to generate a single user level embedding $C_T^{(u)}$ that captures the overall dynamic and semantic features of all the user's posts.

Since our dynamic graph encoder together with our post encoder are capable of learning the representation of the entire dynamic graph and time evolving semantic information from the user posts as a single global vector C_T , we will be able to use a conventional sequence decoder as the decoder for our framework to generate the desired target sequence. Moreover, we propose a novel dynamic graph hierarchical attention mechanism that incorporates both **node-to-graph attention** and **graph-to-sequence attention** in order to promote better interpretability between graph sequences and output sequences and provide more effective aggregation function from node embeddings to graph embeddings [44]. A detailed introduction to the proposed encoder and decoder will be described in the next two subsections.

4.2 Dynamic Graph Encoder

The base model of our graph convolutional network for each snapshot graph is inspired by graph2seq [39], which was originally proposed for addressing static graph-tosequence learning problems. The Graph2Seq model employs an inductive node embedding algorithm that generates bi-directional node embeddings by aggregating information from a node local forward and backward neighborhood within k hops for a static graph. We extend this idea for dynamic graphs by applying such graph convolution on each snapshot graph within dynamic graph inputs. Specifically, suppose the total number of hops is k, then the hidden representation of n-th node in the snapshot graph G_t after applying the first graph convolutional layer will be computed as follows:

$$h_{t,n}^{\vdash} = mean(\{\sigma(W_t^{\vdash(1)}F_{t,u} + b_t^{\vdash(1)}), u \in \mathcal{N}_{\vdash}(v)\})$$
(1)

$$h_{t,n}^{\dashv} = mean(\{\sigma(W_t^{\dashv(1)}F_{t,u} + b_t^{\dashv(1)}), u \in \mathcal{N}_{\dashv}(v)\})$$
(2)

$$h_{t,n}^{(1)} = concat[h_{t,n}^{-}, h_{t,n}^{-}]$$
(3)

where $\mathcal{N}_{\vdash}(v)$ represents the set of forward neighbor nodes of node v, whereas $\mathcal{N}_{\dashv}(v)$ represents the set of backward neighbor nodes; $W_t^{\dashv(1)}, b_t^{\dashv(1)}$ and $W_t^{\vdash(1)}, b_t^{\vdash(1)}$ are learnable parameters for the first convolution layer. $F_{t,u}$ is the feature vector of node u in a snapshot graph at time step t; $\sigma(\cdot)$ represents the activation function of the network (e.g. ReLU); the $mean(\cdot)$ function takes the element-wise mean of the set of vectors in the equation; and concat[vec1, vec2]concatenates the two row vectors into a single row vector.

Likewise, for hop k, the hidden representation of the n-th node in the snapshot graph G_t can be computed via the hidden representations computed from layer k - 1, as follows:

$$h_{t,n}^{\vdash} = mean(\{\sigma(W_t^{\vdash(k)} h_{t,u}^{(k-1)} + b_t^{\vdash(k)}), u \in \mathcal{N}_{\vdash}(v)\}) \quad (4)$$

$$h_{t,n}^{\dashv} = mean(\{\sigma(W_t^{\dashv(k)}h_{t,u}^{(k-1)} + b_t^{\dashv(k)}), u \in \mathcal{N}_{\dashv}(v)\})$$
(5)

$$h_{t,n}^{(k)} = concat[h_{t,n}^{\vdash}, h_{t,n}^{\dashv}]$$

$$\tag{6}$$

In order to capture the high-level representation of graphs for end-to-end graph learning, aggregating node level embeddings to graph level embedding that conveys the entire graph information is essential. To achieve this, we adopt the max pooling operation proposed by [25], [39] as the base aggregation function, feed the node embeddings $h_{t,n}$ to a fully-connected layer and then applies the max pooling method element-wise for each snapshot graph G_t to yield a sequence of graph-level representations g_t .

To model the graph dynamic changes and long-term dependencies throughout the M steps, we utilize Long Short Term Memory (LSTM) networks [45] as a graph embedding sequence encoder to learn the entire dynamic graph-level embedding. The computation of the LSTM network at time step t is defined as:

$$f_t = \sigma(W_f \cdot [o_{t-1}, g_t] + b_f) \tag{7}$$

$$in_t = \sigma(W_{in} \cdot [o_{t-1}, g_t] + b_{in}) \tag{8}$$

$$C_t^{(g)} = tanh(W_C \cdot [o_{t-1}, g_t] + b_C)$$
(9)

$$C_t^{(g)} = f * C_{t-1}^{(g)} + in_t * \widetilde{C}_t^{(g)}$$
(10)

$$out_t = \sigma(W_{out} \cdot [o_{t-1}, g_t] + b_{out}) \tag{11}$$

$$o_t = out_t * tanh(C_t^{(g)}) \tag{12}$$

where o_t is the output of the LSTM network at time step t, $C_t^{(g)}$ is the new cell state for the next time step computation, and the initial cell state for the encoder is set to all-zeros.

In the above encoder formulation, each graph convolutional network needs to learn a set of parameters for each snapshot graph in order to capture their unique characteristics. However, this will lead to several problems for the entire model during training: 1) the node embeddings learned from adjacent snapshot graphs G_t , G_{t+1} may yield inconsistent node embeddings even when the graph characteristics are similar, since there is no constraint on the parameter set; and 2) the resulting model tends to suffer from severe over-fitting issue since too many parameters need to be learned, especially when the total number of time steps T is large for a given dynamic graph.

4.2.1 Dynamic Regularization for Sparse Feature Selection To cope with the aforementioned challenges, we propose a novel temporal feature selection regularization that characterizes feature sparsity, local feature selection consistency, and global feature selection flexibility across the evolving graphs over time. Inspired by group sparsity $\ell_{2,1}$ regularization from group lasso [46] and overlapping group lasso [47], we propose the following dynamic graph regularization for the first layer of graph convolutional networks:

$$\mathcal{L}_{reg} = \beta \sum_{t=1}^{T-w+1} \|\hat{W}_{[t:t+w]}^{(1)}\|_{2,1}$$
(13)

where w is the sliding window size; and $\hat{W}_{[t:t+w]}^{(1)}$ is the concatenated weight matrix from the weight parameters of a group of consecutive graph convolutional networks between time step t and t + w. Each row i of the weight matrix represents the ith feature weights across the w time steps; β controls the relative strength of the regularization.

Dynamic graph typically enjoys temporally consistent characteristics, since user transition activity graphs such as the example shown in Figure 2 change smoothly over time. Thus, the model can achieve temporal local consistency feature selection by adding a sliding window to force the local model to select similar features, which retains the flexibility of the feature selection process while still evolving gradually with time.

The proposed regularization brings several advantages. First, it promotes the interpretability of the model in term of node attributes, enabling us to visualize important features at any given time step and providing useful insights into how the importance of features evolves through time. Second, it also serves as a good regularizer to restrict the large number of model parameter sets, thus preventing possible model over-fitting. Lastly, it enhances the generalization power of the model. These results and analysis will be discussed in detail in the subsequent experimental section.

4.3 Two-level Post Encoder

Our dynamic graph was efficient to provide the spatialtemporal information of the node attributes over the time. Sequential information evaluation over time from the post conveyed useful information in both direction. Posts' content can carry the dependencies of the user's experience. To illustrate, at the initial stage ('Dx') an user may not have the experience, questions or answers of questions related to later stages like Chemotherapy, Radiation, etc. But in the evolution of time an user gains more knowledge and his/her contribution in question-answering or in post content also evolved. Our motivation for introducing post encoder along with graph encoder inspired by the context of the user's post.

We proposed to use a two-level biLSTM post encoder that can capture both semantic information from user posts and their evolution over time. The proposed post encoder would give us upper hand to train our model for understanding users' participation in a better way. There are two angles of information growth in the context of users' posts. First advancement is happened inside a time step. For example, in a forum one user can post on different date in a month. Graph encoder can not capture the information of sub-step or day by day evaluation in a month within a forum. So our first level encoder will be useful to encode the sub-steps evaluation within a time step. Second advancement follows along time steps. As we see in the graph encoder, the participation in different forums varies month to month. Same way the content of the users' posts vary month to month. Our second encoder will serve the purpose to capture this time-to-time evaluation.

The overall information flow of our proposed two level Bi-directional Long Short Term Memory (biLSTM) has been shown at the lower portion in Figure 3. The initial stage starts by combining all the posts of a specific user in a single time step, P_t . So, the whole context, P_t is endued with sub-step evaluation information within a time step. At the next stage, a domain specific 'Word2vec' model plays a vital role to place the word token at the right place in the embedding space as P_t would contain lot of domain specific terminologies.Thereafter, the 'Post Level Encoders' encodes the P_t as c_t which is fed to the 'User Level Encoder' as a part of sequence with other encoded time steps. Afterwards the encoded output from the 'User Level Encoder', $C_T^{(u)}$ ameliorates the decoder initial state with more contextual information along with graph encoder. Concretely, the proposed two level Bi-directional Long Short Term Memory (biLSTM) network is stated in Equation (14) to (22).

4.3.1 Post-level Encoder

The first level biLSTM encoder (i.e. post level) focus on encoding all the posts of a single user in that specific timestep *t*. We preferred biLSTM to LSTM, as patterns lying ahead in the earlier posts of a timestep is also relevant as well as the latest post of the timestep. In these accumulated posts ($P_t = \{P_{t,1}, P_{t,2}, \cdots, P_{t,q}\}$); every word token, $P_{t,q}$ was initialized by using a pre-trained vector embedding and act as the input of the encoder. The final state of the encoder c_t will encode the semantic information of the user within timestep t, and also served as the input of the second level encoder.

$$\vec{f}_{t,q} = \sigma(\vec{W}_f \cdot [o_{q,n-1}, P_{t,q}] + \vec{b}_f)$$
(14)

$$i\tilde{n}_{t,q} = \sigma(\tilde{W}_{in} \cdot [o_{q-1}, P_{t,q}] + \tilde{b}_{in})$$
 (15)

$$\vec{\tilde{c}}_{t,q} = tanh(\vec{W}_c \cdot [o_{q-1}, P_{t,q}] + \vec{b}_c) \tag{16}$$

$$\vec{c}_{p,q} = f * c_{q-1} + in_{t,q} * \tilde{c}_{t,q}$$
 (17)

$$f_{t,q} = \sigma(W_f \cdot [o_{q+1}, P_{t,q}] + b_f)$$
(18)

$$in_{t,q} = \sigma(W_{in} \cdot [o_{q+1}, P_{t,q}] + b_{in}) \tag{19}$$

$$c_{t,q} = tanh(W_c \cdot [o_{q+1}, P_{t,q}] + b_c)$$
(20)

$$\overline{c}_{p,q} = \overline{f} * c_{q+1} + i\overline{n}_{t,q} * \overline{\widetilde{c}}_{t,q}$$
(21)

$$c_t = concat[\vec{c}_{p,q}, \vec{c}_{p,q}] \tag{22}$$

4.3.2 User-level Encoder

The user level encoder is another biLSTM network that takes the post level embedding sequence $c = \{c_1, c_2, \dots, c_T\}$ of over all timesteps and focus on generating an overall user level embedding considering both the semantic and the temporal information in all the user posts. Following the same architecture like post level encoder, we concatenated two directional encoded state ($\vec{C}_{u,T}$ and $\vec{C}_{u,T}$) to get the final state of the user level encoder, $C_T^{(u)}$. Finally, the final state of the encoder $C_T^{(u)}$ will be used along with the final state of graph encoder $C_T^{(g)}$ as the initial state of the decoder.

$$C_T^{(u)} = concat[\vec{C}_{u,T}, \vec{C}_{u,T}]$$
⁽²³⁾

$$C_T = \sigma(concat[C_T^{(g)}, C_T^{(u)}]) \tag{24}$$

The concatenation in Equation (24) helped our model to adopt two pathways to capture the user level activities from dynamic graph encoder and user level context from post encoder.

5 EXPERIMENTS

For this study, we evaluated the performance of our proposed model utilizing two real-world online health forum, namely the *breast cancer community* and *American bladder cancer society*. We conducted comprehensive experiments with both quantitative evaluation and qualitative analyses of the learning results.

5.1 Experimental Settings

Online Breast Cancer Community Dataset: The Breast Cancer Community [1] is one of the largest online forums

designed for patients to share information related to breast cancer. So far, the forum has enrolled 215,671 registered members since the forum launch and the site contains a total of 81 subforums discussing 153,338 topics. The forum data collected for this study covers an 8 year period from the beginning of 2010 to the end of 2017. To create user subforum activity transition graph sequences, we defined user activities as being when they posted new topics or replied to existing topics and the time window was set as one month. After removing common words and stop words, we extracted the 100 top frequency keywords from the forum content to construct the feature vectors for the subforums. For model training and evaluation, we used a total of 2882 users who provided their health stage information. More specifically, we randomly selected 70% of users who provided their health stage history for training, another 10% for validation, and the remaining 20% for testing. The predicted health stage sequences in the test data were validated against the real health stage history extracted from the corresponding users' signatures, as exemplified in Figure 1. The vocabulary of the health stages used in breast cancer consists of {'Dx', 'Chemotherapy', 'Targeted', 'Hormonal', 'Radiation', 'Surgery'}.

Online Bladder Cancer Dataset: The American Bladder Cancer Society [2] offers a online forums for users to discuss topics related to bladder cancer, covering topics like treatment, unusual symptoms, coping, caregiving, recent research, and etc. So far, the forum has enrolled 6,862 registered members since the forum launch and the site contains a total of 11 subforums discussing 7,233 subjects. The forum data collected for this study covers an 13 year period from the beginning of 2007 to the end of 2020. We followed the same ways to process user subforum activity transition graph sequences, and data prepossessing steps for constructing the feature vectors for the subforums. For model training and evaluation, we randomly selected 70% of users who provided their health stage history for training, another 10% for validation, and the remaining 20% for testing. The predicted health stage sequences were validated against the real health stage history extracted from the users' signatures. The vocabulary of the health stages used in breast cancer consists of {'Dx', 'TURBT'², 'BCG'³, 'Chemotherapy', 'Surgery' }.

5.1.1 Evaluation Metrics

We used BLEU scores [48] as the primary evaluation metric for determining the closeness of the model predicted health stage history and the ground truth. In addition, we also tested the model with ROUGE-1 score [49], which is commonly used for evaluating machine summarization and translation tasks.

5.1.2 Comparison Methods

NMT(seq2seq) The Neural Machine Translation model implemented by Luong et al. [50] is a widely used state-of-the-art sequence-to-sequence model for machine transition

tasks. Since the NMT model can only handle simple sequence inputs, we simplified the input data by concatenating the transition sequences of user activity for each month together in time order. The subforum features are omitted in such formulations. We tested the model settings both with and without the attention mechanism.

Graph2seq The Graph2seq model [39] was proposed as a general-purpose encoder-decoder model for static graph to sequence learning. Since the model cannot handle dynamic graphs as input, we simplified the input by aggregating all the edges that appeared in the dynamic graph together into a single static graph. We tested the model settings both with and without the attention mechanism.

5.1.3 Hyper-parameter Settings

For the models tested in this experiment, the Adam optimizer [51] was used with a learning rate of 0.001 and a batch size of 50 for model training; greedy search was used for all the sequence decoders, selecting the highest scoring token at each stage. Hyper-parameters were searched and chosen based on the highest scores achieved on the validation set. For the graph encoders used in Graph2Seq, DynGraph2Seq, and DynAttGraph2Seq models, the hop number k was set to 4. For the proposed dynamic graph regularization, the window size was set to 12 and β was set to 0.0003.

5.2 Performance

5.2.1 Breast Cancer Dataset

Table 1 shows the model performance of the baseline and proposed models on Breast cancer dataset. The scores were obtained from 20 individual runs and presented in a mean \pm standard deviation (SD) format. In general, our proposed DynAttGraph2Seq framework significantly outperformed both the Seq2Seq and Graph2Seq baselines for the various model settings and evaluation metrics. Specifically, the full DynAttGraph2Seq model with the proposed post encoder, dynamic graph regularization, and the dynamic graph hierarchical attention achieved the best score on all the metrics, outperforming the static Graph2Seq and Seq2Seq baseline models by 12% - 25% on the BLEU scores and 7% - 14% on the ROUGE scores. This is because neither the static Graph2Seq nor the Seq2Seq model is able to capture the dynamic evolution of the user forum activities over time, and thus is less effective on this prediction task.

Moreover, comparing to the best results from Dyn-Graph2Seq model [44] with dynamic graph encoder and hierarchical attention, our new model (i.e. with post encoder and regularization) further improved the performance by 3.33% to 4.67% on the BLEU score and 1.27% on the ROUGE score. This is because, different from the formulation in [44], here we further extended the node attributes to be user specific and dynamically changes over time, which gave the model more enriched information about the users in terms of their health stages evolution. Moreover, the newly added post encoder could capture the dynamic evolution of the user specific textual information over time, thus facilitated the effectiveness of the new model. In addition, we further studied the effect of the proposed PostEncoder attached to the baseline methods. Our experimental results in Table 1, depicts that, Graph2Seq can be enhanced to get a competitive performance with DynGraph2Seq (w/att) with the help

^{2.} Short for Trans Urethral Resection of Bladder Tumour, is usually the first treatment for early bladder cancer.

^{3.} Short for Bacillus Calmette-Guerin, it is the most common intravesical immunotherapy for treating early-stage bladder cancer.

TABLE 1: Performance Evaluation for Health Stage Prediction on Breast Cancer Dataset. The scores were obtained from 20 individual runs and presented in a mean \pm standard deviation (SD) format.

1		< ,			
Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE
NMT(seq2seq) (w/o att) NMT(seq2seq) (w/ att) NMT(seq2seq) (w/ att & post)	$ \begin{array}{c c} 55.5{\pm}2.38\\ 57.8{\pm}1.86\\ 58.05{\pm}8.78 \end{array} $	38.4 ± 0.91 40.4 ± 1.21 38.34 ± 7.72	27.1 ± 0.90 29.0 ± 1.28 26.47 ± 5.24	$\begin{array}{c} 19.2{\pm}0.87\\ 20.1{\pm}1.06\\ 18.88{\pm}4.01\end{array}$	71.6 ± 1.04 72.9 ± 0.86 79.67 ± 2.05
Graph2Seq (w/o att) Graph2Seq (w/ att) Graph2Seq (w/ att & post)	$ \begin{array}{c c} 57.5 \pm 1.72 \\ 58.2 \pm 2.19 \\ 62.52 \pm 3.39 \end{array} $	41.5 ± 0.94 41.1 ± 1.38 44.63 ± 3.01	29.8 ± 0.72 30.1 ± 0.83 32.09 ± 2.30	20.3 ± 0.85 21.0 ± 0.51 22.63 ± 1.90	75.8 ± 1.20 76.2 ± 0.96 79.73 ± 2.58
DynGraph2Seq (w/o att) DynGraph2Seq (w/ att)	$ \begin{vmatrix} 60.9 \pm 1.53 \\ 62.3 \pm 1.46 \end{vmatrix} $	43.7±1.00 44.7±1.29	31.5 ± 0.63 32.0 ± 0.94	22.1±0.48 22.5±1.13	$79.3{\pm}0.80\\80.8{\pm}0.36$
DynAttGraph2Seq (w/ reg) DynAttGraph2Seq (w/ reg & att) DynAttGraph2Seq (w/ reg & att & BERT+UL) DynAttGraph2Seq (w/ reg, att & post)	$ \begin{vmatrix} 61.5 \pm 2.42 \\ 64.1 \pm 0.84 \\ 63.14 \pm 1.83 \\ \textbf{65.2 \pm 2.00} \end{vmatrix} $	$\begin{array}{c} 45.1{\pm}1.86\\ 45.4{\pm}0.31\\ 44.04{\pm}2.96\\ \textbf{47.1{\pm}1.63}\end{array}$	32.3±1.31 33.1±0.41 30.73±3.80 33.6±1.24	$\begin{array}{c} 23.1{\pm}1.05\\ \textbf{24.1{\pm}0.70}\\ 21.03{\pm}3.82\\ 23.3{\pm}\ 0.98 \end{array}$	$78.5 \pm 0.86 \\ 81.0 \pm 0.69 \\ 78.09 \pm 1.06 \\ \mathbf{81.8 \pm 0.58}$

TABLE 2: Performance Evaluation for Health Stage Prediction on Bladder Cancer Dataset. The scores were obtained from 5 individual runs and presented in a mean \pm standard deviation (SD) format.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE
NMT(seq2seq) (w/o att)	42.52±5.05	28.1±3.57	$\begin{array}{c} 17.22{\pm}2.15\\ 20.08{\pm}1.94 \end{array}$	$0.0{\pm}0.0$	52.44±5.13
NMT(seq2seq) (w/ att)	46.88±1.32	33.42±2.0		$0.0{\pm}0.0$	56.74±3.78
Graph2Seq (w/o att)	46.52±4.48	31.83±5.19	21.17 ± 3.68	15.83±2.80	$59.33 {\pm} 4.47$
Graph2Seq (w/ att)	47.54±5.50	32.77±5.84	21.76 ± 4.27	16.57±3.83	$62.16 {\pm} 4.83$
DynAttGraph2Seq (w/ reg)	$ \begin{vmatrix} 53.27 \pm 1.10 \\ 48.62 \pm 4.43 \\ 53.23 \pm 0.54 \end{vmatrix} $	37.48 ± 1.10	24.32 ± 3.94	16.86 ± 9.42	68.99 ± 0.0
DynAttGraph2Seq (w/ reg & att)		33.51 ± 4.09	23.21 ± 3.38	18.15 ± 3.28	65.61 ± 6.35
DynAttGraph2Seq (w/ reg, att & post)		37.95 ± 0.88	26.75 ± 0.95	21.29 ± 1.78	67.08 ± 1.86

of the proposed post encoder. Specifically, the performance of Graph2Seq was increased by 6.61% to 8.60% for BLEU scores and 4.63% for the rouge score. Likewise, in case of NMT(seq2seq), the post encoder also greatly enhanced the the ROUGE score of the model by 9.29%. This demonstrated the importance of the textual information as well as the effectiveness of the proposed two-level post encoder.

Lastly, we also compared the proposed two-level post encoder with Bidirectional Encoder Representations from Transformers (BERT) [52], which is the state-of-the-art model for textual data. We replaced the proposed 'Post Level Encoder' by BERT and kept the 'User Level Encoder'. As shown in the second last row of Table 1, we see that the performance is decreased in general, especially for Bleu-4 score. One possible reasoning behind that can be the corpus BERT used for pre-training is too general, and thus the embedding is sub-optimal to domain specific dataset. In contrast, our 'Post Level Encoder' used the corpus from biomedical domain [42], which is more suitable for the medical domain applications.

5.2.2 Bladder Cancer Dataset

Table 2 shows the model performance of the baseline and proposed models on Bladder cancer dataset. The scores were obtained from 5 individual runs and presented in a mean \pm standard deviation (SD) format. Again, our proposed Dy-nAttGraph2Seq framework with the proposed post encoder, dynamic graph regularization, and the dynamic graph hierarchical attention significantly outperformed both the Seq2Seq and Graph2Seq baselines achieved the best score on all the metrics. Specifically, the proposed model outperformed the static Graph2Seq and Seq2Seq baseline models by 6.51% - 41.23% on the BLEU scores and 16.28% - 31.56% on the ROUGE scores. This further demonstrated the effectiveness and generalizability of the proposed framework

on accurately inferring the missing health stage information of the patients across different disease areas. Moreover, this model can work with lower sampled dataset. Whether NMT was not able to produce BLEU-4 score there our model was able to generate the result.

5.3 Interpretability Analysis

5.3.1 Dynamic Graph Hierarchical Attention

Figure 4 shows an example of the learned dynamic graph hierarchical attention by DynAttGraph2Seq for test data. The left part of the figure shows the graph-to-sequence attention learned by the model, where each column is a grayscale heatmap representing the amount of attention being paid to each snapshot graph when the model predicted a specific health stage. The darker the color, the greater the attention being paid. We can see much attention was paid to the graphs around the months being labeled in the figure. The graphs for each labeled months are shown on the right. Interestingly, the graphs in the first two months attracted more attention from the model because those were the months when the patient first became active in the breast cancer online forum. The last two labeled snapshot graphs relate approximately to the time when the user engaged in extensive activities in a wide variety of subforums.

However, it is still hard to understand why these particular snapshot graphs were important and of interest to the model when predicting the user health stage sequence. To explore this issue, we went one step deeper by examining the node-to-graph level attention of these graphs. The red spots on the nodes shown on the right side of Figure 4 represent the amount of attention being paid to each node (i.e. subforum) when the model aggregated the node level information into the graph level representation. Again the darker the red spot, the greater the attention being paid.



Fig. 4: An example of learned dynamic graph hierarchical attention by DynAttGraph2Seq on breast cancer dataset. The darker the color, the greater the attention being paid.

TABLE 3: Top 15 static subforum features (keywords) selected by dynamic graph regularization for the second half of the year 2017 of breast cancer dataset. The keywords in boldface are commonly selected during the year and have high correspondences with patient health stage evolution.

July	Aug	Sep	Oct	Nov	Dec
help	bone	bc	today	help	bone
sister	treatment	family	mets	support	chemo
treatment	therapy	sisters	scared	new	help
negative	scared	year	news	chemo	mets
today	mom	hair	diagnosed	results	please
bone	diagnosed	reconstruction	support	bone	newly
mom	family	care	negative	mets	bc
year	question	tumor	scan	please	survivor
please	radiation	positive	vs	ca	diagnosed
new	pain	lymph	results	pain	ladies
tumor	diagnosis	node	therapy	sisters	support
lymph	back	treatment	people	tamoxifen	anyone
share	looking	back	mom	diagnosed	today
lump	new	please	new	life	share
results	bc	research	newly	show	vs

Now the attention becomes even more interesting and interpretable. For example, when constructing the representation of the May-2012 snapshot graph, Subforums #14, #19, and #2 received attention, with Subforum #14 being assigned the most attention. The title of Subforum #14 is actually "Radiation Therapy - Before, During and After", which is strongly correlated to the health stage 'Radiation'. This explains why that particular graph received more graph-tosequence attention when the model predicted 'Radiation'. Likewise, we further discovered that Subforum #2, entitled "Not Diagnosed but Worried", has a strong correlation with 'Dx' and Subforum #19, entitled "DCIS (Ductal Carcinoma In Situ)", is a strong indicator for the model to predict 'Surgery'. These observed correspondences confirm that the proposed dynamic graph hierarchical attention mechanism greatly enhances the interpretability of the model.

5.3.2 Dynamic Graph Feature Selection

Table 3 shows an example of the top subforum features (keywords) selected by the proposed dynamic graph regularization for the second half of the year 2017. The keywords in boldface were commonly selected during the year and exhibited high correspondences with patient health stage

TABLE 4: Health Stage Sequence Prediction for Breast Cancer (Top two) and Bladder Cancer (Bottom one)

· · · ·	× /
Model	Health stage sequence
Ground Truth	Dx Surgery Surgery Hormonal Surgery Surgery
DynAttGraph2Seq	Dx Surgery Surgery Hormonal Surgery Surgery
Graph2Seq	Dx Surgery Chemotherapy Hormonal
Seq2Seq	Dx Surgery Chemotherapy Hormonal Surgery
Ground Truth	Dx Chemotherapy Chemotherapy Surgery Radiation
DynAttGraph2Seq	Dx Chemotherapy Chemotherapy Surgery Radiation
Graph2Seq	Dx Surgery Chemotherapy Radiation
Seq2Seq	Dx Chemotherapy Surgery Surgery
Ground Truth	Dx TURBT TURBT BCG
DynAttGraph2Seq	Dx TURBT TURBT BCG
Graph2Seq	Dx TURBT BCG
Seq2Seq	TURBT BCG

evolution. For instance, treatment-related keywords (e.g. 'diagnosed', 'treatment', 'therapy', and 'chemo') could be a strong indication of whether the patient was undergoing specific examinations or treatments. Moreover, the keywords containing temporal information, such as 'today', 'year', 'new', and 'newly', were also selected in many of the consecutive months. This is because these keywords could provide a temporal bridge to link dynamic graph sequences to the corresponding patient health stage sequences. Thus, the proposed dynamic graph regularization not only regularized the massive model parameters, but also brought significant benefits to enhance the model interpretability.

5.4 Health Stage Sequence Analysis

5.4.1 Correct Health Stage Sequence Predictions

Table 4 shows examples of health stage sequences that DynAttGraph2Seq was able to infer correctly from breast cancer dataset and also bladder cancer dataset. In the first two examples from breast cancer, the first patient underwent four surgeries, while the second patient had two consecutive chemotherapy treatments. The baseline Graph2Seq failed miserably in terms of capturing such duplication due to the fact that a static transition network cannot preserve such information on the dynamic evolution of user forum activity. In the third example from bladder cancer, the patient went through two consecutive TURBT treatments and then underwent one BCG treatment after got diagnosed with bladder cancer. Although all 3 models are able to capture the pattern 'TURBT \rightarrow BCG', the seq2seq model missed the diagnosis stage ('Dx') due to a lack of overall understanding of the relationships between stages from sequential data alone, while the graph2seq model failed to capture the reoccurrence pattern of TURBT as it lacks temporal information from overall static behavior graph.

5.4.2 Interpretable Sequential Pathway Patterns in Stage Sequence Predictions

In general, there is no fixed sequential for each individual user's health stage sequence, as different patient can go through quite diverge direction of treatment based on their personal health status evolution over time. However, we indeed observed that there some pathway patterns in stage sequence predictions the proposed model is able to learn

from the data which are reasonable and interpretable in clinical perspective. For example, from the breast cancer dataset, we observed that the proposed model is likely to predict a pathway like 'Dx \rightarrow Surgery \rightarrow Chemotherapy \rightarrow Radiation \rightarrow Hormonal'. This order of treatments is well aligned with the typical pathway that doctors would recommend to the patients in common practice for breast cancer treatment. Similarly, for bladder cancer, the model learns a sequential pathway of 'Dx \rightarrow TURBT \rightarrow BCG \rightarrow Surgery'. Which is also will suited for the patients to go through in practice for bladder cancer treatment. Those observations suggest that the proposed model can effectively capture the common patterns of the health stage sequence, and therefore can be applied to infer the missing health stage of the patients and benefit the patients, the health care organizations, and the future clinical research.

6 CONCLUSION

In this paper, we formulated the task of health stage inference using online health forum data as a dynamic attributed graph as a dynamic attributed graph to sequence learning problem and propose a novel and generic DynAttGraph2Seq framework that can handle this new type of learning problem effectively. Our DynAttGraph2Seq consists of a novel dynamic graph encoder along with a twolevel sequential encoder to capture the semantic features from user posts and an interpretable sequence decoder that learn the mapping between a sequence of time-evolving user activity graphs as well as user posts to a sequence of target health stages. In addition, we propose a new dynamic graph regularization and dynamic graph hierarchical attention mechanisms to facilitate the necessary multi-level interpretability. Comprehensive experimental analysis on the health stage prediction task demonstrates the effectiveness and the interpretability of the proposed models.

Like any deep learning model, this model also faces the challenges of data-hungry. This model performed better for Online Breast Cancer Community Dataset [1] compared to Online Bladder Cancer Dataset [2] as the second one does not have abundant training samples. Moreover, it is important to have properly guided and well-documented health stage of the patients from the forum. Otherwise, as a supervised learning model, the model will face the challenge to train properly. We also observed that domainspecific 'Word2vec' is very important factor. This model performance can be improved further if we can use a pretrained 'Word2vec' model based on a cancer-specific corpus.

ACKNOWLEDGMENT

This work was supported by National Science Foundation (NSF) under grant number (1755850, 1841520, 1907805, 1942594, and 2007976), Defense Advanced Research Projects Agency (DARPA) under Bioelectronics for Tissue Regeneration program, The Design Knowledge Company (Award No. 10827.002.120.04), Jeffress Memorial Trust Award, and NVIDIA GPU Grant.

REFERENCES

- "Breast cancer community," https://community.breastcancer.org/. "American bladder cancer society," bladdercancersupport.org. [1]
- [2]
- [3] "ehealth forum," http://ehealthforum.com.

- [4] Y.-C. Wang, R. Kraut, and J. M. Levine, "To stay or leave?: the relationship of emotional and informational support to commitment in online health support groups," in Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work, 2012, pp. 833-842.
- [5] N. Elhadad, S. Zhang, P. Driscoll, and S. Brody, "Characterizing the sublanguage of online breast cancer forums for medications, symptoms, and emotions," in AMIA Annual Symposium Proceedings, vol. 2014. American Medical Informatics Association, 2014, p. 516.
- S. Zhang, E. Grave, E. Sklar, and N. Elhadad, "Longitudinal analy-[6] sis of discussion topics in an online breast cancer community using convolutional neural networks," Journal of biomedical informatics, vol. 69, pp. 1-9, 2017.
- F. Sadeque, T. Solorio, T. Pedersen, P. Shrestha, and S. Bethard, [7] "Predicting continued participation in online health forums," in Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis, 2015, pp. 12-20.
- N. V. Chawla and D. A. Davis, "Bringing big data to personal-ized healthcare: a patient-centered framework," *Journal of general* [8] internal medicine, vol. 28, no. 3, pp. 660-665, 2013.
- M. Jha and N. Elhadad, "Cancer stage prediction based on patient online discourse," in Proceedings of the 2010 Workshop on Biomedical Natural Language Processing, 2010, pp. 64–71. [10] P. K. Sahoo, S. K. Mohapatra, and S.-L. Wu, "Analyzing healthcare
- big data with prediction for future health condition," IEEE Access, vol. 4, pp. 9786–9799, 2016.
- [11] J. Gao, C. Xiao, Y. Wang, W. Tang, L. M. Glass, and J. Sun, "Stagenet: Stage-aware neural networks for health risk prediction," in Proceedings of The Web Conference 2020, 2020, pp. 530–540.
- [12] A. Perotte, R. Ranganath, J. S. Hirsch, D. Blei, and N. Elhadad, "Risk prediction for chronic kidney disease progression using heterogeneous electronic health record data and time series analysis," Journal of the American Medical Informatics Association, vol. 22, no. 4, pp. 872-880, 2015.
- [13] L. Zhao, J. Ye, F. Chen, C.-T. Lu, and N. Ramakrishnan, "Hierarchical incomplete multi-source feature learning for spatiotemporal event forecasting," in KDD 2016. ACM, 2016, pp. 2085–2094.
- [14] J. Wang and L. Zhao, "Multi-instance domain adaptation for vaccine adverse event detection," in WWW 2018, 2018, pp. 97–106.
- [15] L. Zhao, "Event prediction in the big data era: A systematic survey," ACM Computing Surveys (CSUR), vol. 54, no. 5, pp. 1-37, 2021.
- [16] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and S. Y. Philip, "A comprehensive survey on graph neural networks," IEEE transactions on neural networks and learning systems, 2020.
- [17] L. Zhu, D. Guo, J. Yin, G. Ver Steeg, and A. Galstyan, "Scalable temporal latent space inference for link prediction in dynamic social networks," TKDE, vol. 28, no. 10, pp. 2765–2777, 2016.
- Z. Zhang, P. Cui, J. Pei, X. Wang, and W. Zhu, "Timers: Error-bounded svd restart on dynamic networks," arXiv preprint [18] arXiv:1711.09541, 2017.
- [19] R. Trivedi, M. Farajtbar, P. Biswal, and H. Zha, "Representation learning over dynamic graphs," arXiv preprint arXiv:1803.04051, 2018.
- [20] P. Goyal, S. R. Chhetri, and A. Canedo, "dyngraph2vec: Capturing network dynamics using dynamic graph representation learning, arXiv preprint arXiv:1809.02657, 2018.
- [21] W. Yu, W. Cheng, C. C. Aggarwal, K. Zhang, H. Chen, and W. Wang, "Netwalk: A flexible deep embedding approach for anomaly detection in dynamic networks," in KDD 2018. ACM, 2018, pp. 2672-2681.
- [22] D. Eswaran, C. Faloutsos, S. Guha, and N. Mishra, "Spotlight: Detecting anomalies in streaming graphs," in KDD 2018, 2018, pp. 1378-1386.
- [23] Z. Wu, S. Pan, G. Long, J. Jiang, and C. Zhang, "Graph wavenet for deep spatial-temporal graph modeling," arXiv preprint arXiv:1906.00121, 2019.
- [24] Z. Wu, S. Pan, G. Long, J. Jiang, X. Chang, and C. Zhang, "Connecting the dots: Multivariate time series forecasting with graph neural networks," in Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020, pp. 753-763.
- [25] J. Wu, S. Pan, X. Zhu, C. Zhang, and S. Y. Philip, "Multiple structure-view learning for graph classification," IEEE Transactions on Neural Networks and Learning Systems, vol. 29, no. 7, pp. 3236-3251, 2018.

- [26] L. Wu, I. E.-H. Yen, Z. Zhang, K. Xu, L. Zhao, X. Peng, Y. Xia, and C. Aggarwal, "Scalable global alignment graph kernel using random features: From node embedding to graph embedding," in *KDD 2019*, 2019, pp. 1418–1428.
- [27] K. Xu, L. Wu, Z. Wang, M. Yu, L. Chen, and V. Sheinin, "Exploiting rich syntactic information for semantic parsing with graph-tosequence model," arXiv preprint arXiv:1808.07624, 2018.
- [28] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," arXiv preprint arXiv:1409.0473, 2014.
- [29] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," arXiv preprint arXiv:1508.04025, 2015.
- [30] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in NAACL, 2016, pp. 1480–1489.
- [31] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," arXiv preprint arXiv:1710.10903, vol. 1, no. 2, 2017.
- [32] J. Yang, J. Lu, S. Lee, D. Batra, and D. Parikh, "Graph r-cnn for scene graph generation," arXiv preprint arXiv:1808.00191, vol. 2, 2018.
- [33] J. Lee, I. Lee, and J. Kang, "Self-attention graph pooling," arXiv preprint arXiv:1904.08082, 2019.
- [34] A. Eriguchi, K. Hashimoto, and Y. Tsuruoka, "Tree-tosequence attentional neural machine translation," *arXiv preprint arXiv:1603.06075*, 2016.
- [35] O. Vinyals, S. Bengio, and M. Kudlur, "Order matters: Sequence to sequence for sets," arXiv preprint arXiv:1511.06391, 2015.
- [36] K. S. Tai, R. Socher, and C. D. Manning, "Improved semantic representations from tree-structured long short-term memory networks," arXiv preprint arXiv:1503.00075, 2015.
- [37] L. Zhang, L. Zhao, S. Qin, and D. Pfoser, "Tg-gan: Deep generative models for continuously-time temporal graph generation," arXiv preprint arXiv:2005.08323, 2020.
- [38] X. Guo, L. Zhao, Z. Qin, L. Wu, A. Shehu, and Y. Ye, "Interpretable deep graph generation with node-edge co-disentanglement," arXiv preprint arXiv:2006.05385, 2020.
- [39] K. Xu, L. Wu, Z. Wang, and V. Sheinin, "Graph2seq: Graph to sequence learning with attention-based neural networks," arXiv preprint arXiv:1804.00823, 2018.
- [40] Y. Chen, L. Wu, and M. J. Zaki, "Reinforcement learning based graph-to-sequence model for natural question generation," in *Proceedings of the 8th International Conference on Learning Representations*, Apr. 26-30, 2020.
- [41] A. Li, S. P. Cornelius, Y.-Y. Liu, L. Wang, and A.-L. Barabási, "The fundamental advantages of temporal networks," *Science*, vol. 358, no. 6366, pp. 1042–1046, 2017.
- [42] S. Moen and T. S. S. Ananiadou, "Distributional semantics resources for biomedical text processing," *Proceedings of LBM*, pp. 39–44, 2013.
- [43] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in NIPS 2014, 2014, pp. 3104–3112.
- [44] Y. Gao, L. Wu, H. Homayoun, and L. Zhao, "Dyngraph2seq: Dynamic-graph-to-sequence interpretable learning for health stage prediction in online health forums," in 2019 IEEE International Conference on Data Mining (ICDM). IEEE, 2019, pp. 1042– 1047.
- [45] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997.
- [46] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.
- [47] L. Jacob, G. Obozinski, and J.-P. Vert, "Group lasso with overlap and graph lasso," in *ICML 2009*. ACM, 2009, pp. 433–440.
- [48] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in ACL 2002, 2002, pp. 311–318.
- [49] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," Text Summarization Branches Out, 2004.
- [50] M. Luong, E. Brevdo, and R. Zhao, "Neural machine translation (seq2seq) tutorial," https://github.com/tensorflow/nmt, 2017.
- [51] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.
- [52] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pretraining of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.







Yuyang Gao is a Ph.D. student in Department of Computer Science at Emory University. He received the B.S. degree in Computer Science from Shandong University, China in 2014, and M.S. in Computer Science from George Mason University in 2018 with the Distinguished Academic Achievement. His research focuses on spatiotemporal data mining, graph deep learning, and bio-inspired machine learning. He has published multiple papers as the first author in top-tier conferences such as ICDM and AAAI.

Tanmoy Chowdhury is a Ph.D. student in Department of Information Sciences and Technology in George Mason University. He received B.Sc. degree in EEE from Chittagong University of Engineering and Technology in 2011, M.Sc. in Biomedical Engineering From Uiversity of Ulsan in 2016 and M.S. in Applied Information Technology from George Mason University in 2018. His research interests include data mining, computer aided design, natural language processing and computational model.

Lingfei Wu Dr. Lingfei Wu earned his Ph.D. degree in computer science from the College of William and Mary in 2016. He currently serves as a Principal Scientist at JD.COM Silicon Valley Research Center. Previously, he was a research staff member at IBM Research and led a research team (10+ RSMs) for developing novel Graph Neural Networks for various tasks, which leads to the #1 Al Challenge Project in IBM Research and multiple IBM Awards including Outstanding Technical Achievement Award. He

has published more than 80 top-ranked conference and journal papers and is a co-inventor of more than 40 filed US patents. Because of the high commercial value of his patents, he has received several invention achievement awards and has been appointed as IBM Master Inventors, class of 2020. He was the recipients of the Best Paper Award and Best Student Paper Award of several conferences such as IEEE ICC'19, AAAI workshop on DLGMA'20 and KDD workshop on DLG'19. His research has been featured in numerous media outlets, including NatureNews, YahooNews, Venturebeat, and TechTalks. He has co-organized 10+ conferences (KDD, AAAI, IEEE BigData) and is the founding co-chair for Workshops of Deep Learning on Graphs (with AAAI'21, AAAI'20, KDD'20, KDD'19, and IEEE BigData'19). He has currently served as Associate Editor for IEEE Transactions on Neural Networks and Learning Systems, ACM Transactions on Knowledge Discovery from Data and International Journal of Intelligent Systems, and regularly served as a SPC/PC member of the following major AI/ML/NLP conferences including KDD, IJCAI, AAAI, NIPS, ICML, ICLR, and ACL.



Liang Zhao is an assistant professor in Department of Computer Science at Emory University. He obtained his PhD degree in 2016 from Computer Science Department at Virginia Tech. His research interests include data mining, artificial intelligence, and machine learning, with special interests in spatiotemporal and network data mining, deep learning on graphs, nonconvex optimization, and interpretable machine learning. He has published over 70 peer-reviewed full research papers in top-tier conferences and jour-

nals such as KDD, ICDM, TKDE, Proceedings of the IEEE, TKDD, TSAS, IJCAI, AAAI, WWW, CIKM, SIGSPATIAL, and SDM. He won NSF CAREER Award in 2020. He has also won Jeffress Trust Award in 2019, Outstanding Doctoral Student in the Department of Computer Science at Virginia Tech in 2017, NSF CRII Award in 2018, and was ranked as one of the "Top 20 Rising Star in Data Mining" by Microsoft Search in 2016. He has won best papers in top-tier venues such as Best Paper Award in ICDM 2019. He has been organizing several prestigious venues such as KDD 2019, ACM SIGSPATIAL 2020, ICDM 2019, CIKM 2019, SecureCom 2020, and SSTD 2017. He is co-chairing workshops such as GeoAl 2019 co-located with SIGSPATIAL 2019 and DeepSpatial 2019 co-located with ICDM 2019.