

Time-variant Graph Learning and Classification



Haishuai Wang

Faculty of Engineering and Information Technology

University of Technology, Sydney

A thesis submitted for the degree of

Doctor of Philosophy

30 October 2016

This thesis is dedicated to my loving parents

CERTIFICATE OF ORIGINAL AUTHORSHIP

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Student: Haishuai Wang

Date: 30.10.2016

Acknowledgements

I benefited and learned a lot from my supervisors, my colleagues, and my friends during the PhD study at University of Technology, Sydney, Australia. I wish to take this opportunity to thank all of them.

Firstly, I would like to express my sincere gratitude to my advisors, Dr. Ling Chen, Dr. Peng Zhang and Prof. Xingquan Zhu, for the continuous support of my Ph.D study and related research, for their patience, motivation, and immense knowledge. Dr. Ling Chen has been supportive and has given me the freedom to pursue various projects without objection. She has also provided insightful discussions about the research. I am deeply indebted to Dr. Peng Zhang for his fundamental role in my doctoral work. Dr. Zhang provided me with every form of guidance, assistance, and expertise that I needed during my Ph.D study. In addition to our academic collaboration, I greatly value the close personal rapport that Dr. Zhang and I have forged over the years. I am also very grateful to Professor Xingquan Zhu for his scientific advice and knowledge and many insightful discussions and suggestions. Their friendship has also been important to me as they have often given me invaluable advice in a personal sense.

I would like to thank Professor Yixin Chen who has provided me with a great opportunity to visit Washington University in St. Louis and has given me a postdoctoral position. I am extremely grateful for his guidance and all the excellent discussions that I have had with him. His deep insights have helped me at various stages of my research. I also give thanks to Professor Chengqi Zhang, Professor Ivor W Tsang, Professor Huan Liu, Professor Xindong Wu for their very helpful comments and suggestions which were aimed at improving my research skills.

I would also like to take this opportunity to thank all my friends in the Quantum Computation & Intelligent Systems Centre at UTS for all the great times that we have shared, in particular, Jia Wu, Shirui Pan, Shaoli Huang, Bo Han, Qin Zhang, Sujuan Hou, Lianhua Chi, Chunyang Liu, Guodong Long, Bozhong Liu, Anjin Liu, Yu Bai, and Tongliang Liu. They are the ones who have given me support during both joyful and stressful times, and to whom I will always be thankful. I am also grateful to Jemima Moore for proof reading my submission drafts.

Finally, I am deeply thankful to my parents and sisters for their endless love, encouragement, support, and various sacrifices. Without them, this thesis would never have been written. I dedicate this thesis to them.

Abstract

Graph classification is an important tool for analyzing data with structure dependency. In traditional graph classification, graphs are assumed to be independent where each graph represents an object. In a dynamic world, it is very often the case that the underlying object continuously evolves over time. The change of node content and/or network structure, with respect to the temporal order, presents a new time-variant graph representation, where an object corresponds to a set of time-variant graphs (TVG). A time-variant graph can be used to characterize the changing nature of the structured object, including the node attribute and graph topological changing over time. Therefore, the evolution of time-variant graphs could be either network structure or node content over time. In this dissertation, we formulate a new time-variant graph learning and classification (TVGLC) task.

To learn and classify time-variant graphs, the vital steps are feature extraction, modeling and algorithm design. However, for time-variant graph classification, frequent subgraph features are very difficult to obtain. Because one has to consider the graph structure space and the temporal correlations to find subgraph candidates for validation, the search space for finding frequent subgraph features is infinite and unlikely to obtain stable structures. Secondly, graph structures that imply subgraph features may irregularly change over time. Thus, to extract effective and efficient features is a great challenge for TVGLC. In addition, carrying out applicable models and algorithms to cater for the extracted features for TVGLC is also a challenge.

Considering the above challenges, this research aims to extract efficient features and design new algorithms to enable the learning of the

time-variant graph. Because time variant graphs may involve changes in the network structures and changes in the node content, which complicate the algorithm designs and solutions, our research employs a divide and conquer principle to first solve a simplified case where (1) network topology is fixed whereas the node content continuously evolves (i.e., networked time series classification). After that, we advance to the setting to (2) evolving network structure and propose solutions to TVGLC with incremental subgraph features. To enhance the subgraph feature exploration for time variant graph classification, we propose (3) graph-shapelet features for TVGLC. Last, but not the least, we study (4) an application of online diffusion provenance detection.

Temporal Feature Selection on Networked Time Series: As the time-variant graph can be graph node content and/or graph structure evolution, we first study a simple case where the structure is fixed but the node content continuously evolves. The problem forms time series data when the node content changes over time, and we combine time series data with a static graph to form a new problem called networked time series. We formulate the problem of learning discriminative features (i.e., segments) from networked time series data considering the linked information among time series (e.g., social users are taken as social sensors that continuously generate social signals (tweets) represented as time series). The discriminative segments are often referred to as *shapelets* of time series. Extracting shapelets for time series classification has been widely studied. However, existing works on shapelet selection assumes that time series are independent and identically distributed (i.i.d.). This assumption restricts their applications to social networked time series analysis. This thesis proposes a new Network Regularized Least Squares (NetRLS) feature selection model, which combines typical time series data and user network graph data for analysis.

Incremental Subgraph based TVGLC: To learn and classify the

time-variant graph with network structure evolve, the key challenges are to extract features and build models. To date, subgraphs are often used as features for graph learning. In reality, the dimension of the subgraphs has a crucial dependency on the threshold setting of the frequency support parameter, and the number may become extremely large. As a result, subgraphs may be incrementally discovered to form a feature stream and require the underlying graph classifier to effectively discover representative subgraph features from the subgraph feature stream. Moreover, we propose a *primal-dual incremental subgraph feature selection* algorithm (*ISF*) based on a max-margin graph classifier. The ISF algorithm constructs a sequence of solutions that are both primal and dual feasible. Each primal-dual pair shrinks the dual gap and renders a better solution for the optimal subgraph feature set. To avoid the bias of the ISF algorithm on short-pattern subgraph features, we present a new *incremental subgraph join feature selection* algorithm (*ISJF*) by forcing graph classifiers to join short-pattern subgraphs and generate long-pattern subgraph features.

Graph-shapelet based TVGLC: As graph structure continuously evolves over time, the search space for finding frequent subgraph features is infinite and unlikely to obtain stable structures. To tackle this challenge, we formulate a new time-variant graph classification task, and propose a new graph feature, *graph-shapelets*, for learning and classifying time-variant graphs. Graph-shapelet is compact and discriminative *graph transformation subsequences*. A graph-shapelet can be regarded as a graphical extension of *shapelets* – a class of discriminative features designed for vectorial temporal data classification. In order to discover graph-shapelets, we propose to convert a time-variant graph sequence as time-series data, and use shapelets discovered from the time-series data to find *graph transformation subsequences* as graph-shapelets. By converting each graph-shapelet as a unique tokenized graph transformation sequence, we can use the editing distance to calculate the distance between two graph-shapelets for time-variant graph classification.

Application of Online Diffusion Provenance Detection: In social network analysis, the information propagation graph (i.e., cascade) is a kind of time-variant graph because the information diffusion forms a graph at a certain time and the graph evolves over time. An important application of information diffusion networks (i.e., time-variant graph) is provenances detection. Existing work on network diffusion provenance identification focuses on offline learning where data collected from network detectors are static and a snapshot of the network is available before learning. However, an offline learning model does not meet the needs of early warning, real-time awareness and real-time response to malicious information spreading in networks. In this part, we study a new problem of online discovering diffusion provenances in large networks. To this end, we propose an online regression model for real-time diffusion provenance identification. Specifically, we first use offline collected network cascades to infer the edge transmission weights, and then use an online l_1 non-convex regression model as the identification model. The proposed methods are empirically evaluated on both synthetic and real-world networks.

Experiments on synthetic and real-world data validate and demonstrate the effectiveness of the proposed methods for time-variant graph learning and classification.

Contents

Contents	xi
List of Figures	xv
List of Tables	xxiii
Nomenclature	xxiii
1 Introduction	1
1.1 Background	1
1.2 Motivation	3
1.3 Research Problems	7
1.3.1 Time-variant Graph Feature Extraction	7
1.3.2 Time-variant Graph Classification Algorithms	8
1.3.3 Evaluation of Proposed Features and Algorithms	8
1.3.4 Applications of Time-Variant Graph Learning	8
1.4 Thesis Contributions and Road Map	9
1.5 Publications	11
2 Literature Review	15
2.1 Dynamic Graph	15
2.2 Temporal Features for Networked Data	16
2.2.1 Discriminative Features for Temporal Data	17
2.2.2 Feature Selection in Networked Data	17
2.3 Incremental Subgraph base TVGLC	18
2.3.1 Graph Classification	18

CONTENTS

2.3.2	Incremental Feature	19
2.3.3	Cascade Outbreak Prediction	19
2.3.4	High Dimensional Data Learning and Data Stream Mining	20
2.4	Graph-shapelet based TVGLC	21
2.4.1	Graph Features	21
2.4.2	Graph Stream Mining	22
2.5	Online Diffusion Provenances Detection	22
3	Temporal Shapelet Feature for Networked Time Series	25
3.1	Introduction	25
3.2	Preliminaries and Problem Definition	26
3.2.1	Time Series Segments	26
3.2.2	Shapelets	27
3.2.3	Goal	27
3.3	Network Regularized Least Squares Shapelets Learning	28
3.3.1	Shapelets Selection	28
3.3.2	Challenges and Convexity	29
3.3.3	Networked Time Series Classification Algorithm	31
3.4	Experiments	34
3.4.1	Data Sets	34
3.4.1.1	Twitter	34
3.4.1.2	DBLP	35
3.4.2	Experimental Measures	36
3.4.3	Experimental Comparisons	36
3.4.4	Industry case study on clinical data	41
3.5	Conclusion	44
4	Incremental Subgraph based TVGLC	45
4.1	Introduction	45
4.2	Preliminaries	49
4.3	Graph Classification	51
4.3.1	Max-margin Graph Classifier	51
4.3.2	Incremental Subgraph Features	54

4.3.3	Long-pattern Subgraph Features	57
4.4	Analysis	60
4.5	Experiments	61
4.5.1	Data Sets	61
4.5.1.1	Real-world Data	62
4.5.1.2	Synthetic Data	63
4.5.2	Parameter study	66
4.5.3	Experimental Results	69
4.5.4	Case Study on Cascading Outbreak Early Prediction . . .	75
4.6	Conclusions	77
5	Graph-shapelets Feature based TVGLC	79
5.1	Introduction	79
5.2	Problem Formulation and Preliminaries	82
5.2.1	Problem Definition	82
5.2.2	Preliminaries	83
5.3	Overall Framework of Graph-shapelet based TVG Learning	88
5.4	Graph-shapelet Feature Exploration	89
5.4.1	Graph Sequences to Time Series	89
5.4.2	Graph-Shapelet Pattern Candidates	90
5.4.3	Finding Graph-Shapelet Patterns	90
5.5	Graph-shapelet based TVG Classification Algorithm	92
5.5.1	Classification with Graph-Shapelet Patterns	92
5.5.2	Time Complexity Analysis	93
5.6	Experiments	93
5.6.1	Data sets	93
5.6.1.1	Synthetic Time-Variant Graph Data	93
5.6.1.2	Real-World Time-Variant Graph Data	94
5.6.2	Experimental Settings	95
5.6.2.1	Baseline Approaches	95
5.6.2.2	Evaluation Measures	96
5.6.3	Experimental Results	96
5.6.3.1	Effectiveness Results	96

CONTENTS

5.6.3.2	Efficiency Results	98
5.6.3.3	Analysis of gShapelet Algorithm	100
5.7	Discussions	102
5.8	Conclusion	103
6	Application of Online Diffusion Provenance Detection from TVG	105
6.1	Introduction	105
6.2	Preliminaries	109
6.3	Regression Model	112
6.4	Online Algorithm	115
6.4.1	Relative Time Difference	115
6.4.2	Convex Approximation	116
6.4.3	Online Sub-gradient	116
6.4.4	The Online Stochastic Sub-gradient (OSS) Algorithm	119
6.5	Experiments	123
6.5.1	Experimental Data	123
6.5.2	Experimental Setup	125
6.5.3	Experimental Results	125
6.6	Conclusions	133
7	Conclusions and Future Work	135
7.1	Summary of This Thesis	135
7.2	Future Work	138
	Appendix A	141

List of Figures

1.1	An example of a time-variant graph representation for large-scale information propagation. The information propagation can be regarded as a series of graphs. At different time periods, both the node volume and the graph structure are diverse, <i>i.e.</i> , the information propagation is a process that takes place on the graph. The information propagation outbreak prediction aims to build a time-variant graph classification solution to accurately identify an outbreak information diffusion graph (above) from a non-outbreak information diffusion graph (below).	2
1.2	An example of an information diffusion network. The information propagation cascade can be regarded as a graph. The cascade on the left (with bold-faced edges and green nodes) quickly grows and propagates to an increasing number of nodes (<i>i.e.</i> outbreak), whereas the cascade on the right (with bold-faced edges and yellow nodes) remains steady and is therefore a non-outbreak cascade. Cascade outbreak prediction aims to build a graph classification model to accurately identify outbreak cascades from non-outbreak cascades.	4
1.3	An illustration of social robots identification. We aim to identify social robots (Left) from real social users (Right). Each social node can be taken as a social sensor [104] that generates continuous social signals (tweets). Each social signal is independent and identically distributed because nodes are mutually connected in social networks.	5

LIST OF FIGURES

1.4	An example of time-variant graph provenance detection. In social networks, information propagation forms a cascade (time-variant graph). The cascade in the figure originates from a provenance (i.e., PKU_news) and propagates to a huge number of users over time (outbreak cascade). We aim to identify the provenance from the time-variant graph in a timely fashion to avoid malicious information break outs as early as possible.	6
3.1	The network (Left) has 200 nodes and 210 edges, social robots are densely connected while the remaining two groups of real users are sparsely connected. The generated time series (Mid.) of the three classes of nodes, social robots and active VIP users have discriminative patterns while ordinary users tend to have flat curves. The shapelets (Right) of a typical social robot is concave ([a, b, c]), while shapelets of an active VIP user is convex ([a,b,c]). Time series of ordinary users have heavy noise and it is hard to capture a shapelet.	33
3.2	Parameter test (a) and model comparison (b).	35
3.3	Accuracy comparison on Twitter data set <i>w.r.t.</i> various window length and parameter ρ	37
3.4	AUC comparison on Twitter data set <i>w.r.t.</i> various window length and parameter ρ	38
3.5	Accuracy comparison on DBLP data set <i>w.r.t.</i> various window length and parameter ρ	39
3.6	AUC comparison on DBLP data set <i>w.r.t.</i> various window length and parameter ρ	40
3.7	An illustration of the shapelets learned by NetRSL on the Twitter and DBLP datasets.	40

LIST OF FIGURES

3.8	A part of the heart rate data. The network (Left) is based on age of patients, and we link two patients if their age difference is within 3 years. The middle time series represents heart rate for one minute interval. The blue time series represents these kinds of patient were transferred to ICU, and the orange one represents the patient was not transferred to ICU. We use both the patients network and heart rate time series to classify if the patient will be transferred to ICU or not. Obviously, only using the time series data with inseparability can make the learning task difficult. Thus, together with the network data we can improve the final classification performance.	42
3.9	Performance on the clinical data set <i>w.r.t.</i> various window length and parameter ρ	43
3.10	AUC comparison on the clinical data set <i>w.r.t.</i> various window length and parameter ρ	44
4.1	The number of frequent subgraphs <i>w.r.t.</i> the support threshold in frequent pattern mining. The cascade data, containing about 2.76 million cascades and 3.3 million nodes, are obtained from the SNAP data set (http://snap.stanford.edu/infopath/data.html/). When the parameter <i>Supp</i> is 50, the number of discovered subgraph features is more than $8 * 10^5$!	46
4.2	Subgraph features. The graph (left) is converted into a binary feature vector (right) by examining the existence of subgraph features. The feature vector can be processed by traditional classifiers such as SVMs.	50
4.3	Joining correlated subgraph fragments.	51

LIST OF FIGURES

4.4	An illustration of a long-pattern subgraph feature buried under two short-pattern subgraph features in the information cascade data. Consider four graphs g_1, \dots, g_4 . g_1 and g_2 from class “+1” while g_3 and g_4 from “-1”. Assume we have two short-pattern subgraphs $f_1 : U_1 \rightarrow U_2$ and $f_2 : U_2 \rightarrow U_3$, and a long-pattern subgraph $f_3 : U_1 \rightarrow U_2 \rightarrow U_3$ by joining f_1 and f_2 . If one feature is allowed to select for classification, then f_1 or f_2 is likely to be selected, instead of the more interesting f_3	57
4.5	Parameter study on min-batch size B at each iteration and value k in top- k	63
4.6	# of subgraph features <i>w.r.t.</i> support threshold on the MemeTracker data set.	65
4.7	Memory cost <i>w.r.t.</i> the support threshold $Supp$ under different propagation time stamps on the MemeTracker data set.	66
4.8	Running time <i>w.r.t.</i> the support threshold $Supp$ under different propagation time stamps on the MemeTracker data set.	67
4.9	Percentage of patterns <i>w.r.t.</i> the support threshold and pattern length.	68
4.10	Precision comparison under different $Supp$ on the MemeTracker data set.	68
4.11	Recall comparison under different $Supp$ on the MemeTracker data set.	69
4.12	F1 score comparison under different $Supp$ on the MemeTracker data set.	70
4.13	Accuracy and variance comparisons <i>w.r.t.</i> time stamp on the MemeTracker data set.	71
4.14	# of subgraph features <i>w.r.t.</i> support threshold on the DBLP data set.	72
4.15	Memory cost and running time <i>w.r.t.</i> the support threshold at different year on the DBLP data set.	73
4.16	Accuracy comparison under different $Supp$ on the DBLP data set.	74

4.17 The probability distribution of cascades. The dotted line is the linear fitting result to the red curve, showing that the distribution fits the power-law. The two dotted vertical lines indicate the threshold which discriminates outbreaks from non-outbreak cascades. The sizes in $[100, 300]$ are the gap cascades which are not used in our experiments. 75

4.18 Early prediction of information cascade outbreaks. We compare the subcascade-based method (red line) with the node-based method (blue line). The figure shows that the subcascade-based method provides better prediction accuracy than the node-based method. 76

5.1 An illustration of graph-shapelet patterns. Graph-shapelet patterns are compact and discriminative graph transformation subsequences that describe the graph transformation patterns shared in the same class of time-variant graphs, *e.g.* two time-variant graphs (the two entire rows above) with the outbreak labels. When we explore a univariate time series from a time-variant graph, we can see that the location of a *graph-shapelet pattern* is consistent with that of a *shapelet* in time series (detailed in Section 5.4.1). In this case, graph-shapelet patterns can be used for time-variant graph prediction such as dynamic graph outbreak prediction. 80

5.2 A toy example of graph-shapelet pattern mining to explain the definitions of related operations. We first explore univariate time series from a set of time-variant graphs (detailed in Section 5.4.1), where shapelets are discovered using shapelet pattern mining algorithms. The graph-shapelet patterns (discriminative graph transformation sequences) are then relocated by calculating the graph edit similarity (as in Definition 8) between shapelet patterns which match a set of graph subsequences. To make the above illustration clear, we use this example through *Examples 1 ~ 4*. 83

LIST OF FIGURES

5.3	A concept view of the proposed time-variant graph classification framework. We first explore univariate time series from time-variant graphs via a sample kernel method in each graph (step ①, Section 5.4.1). Then, we find shapelet patterns from the time series by using shapelet pattern mining (step ②, Section 5.4.2). Next, we locate the sub-time-variant graphs that match the shapelet patterns from the original time-variant graphs (step ③, Section 5.4.3). Note that each sub-time-variant graph corresponds to a unique graph transformation subsequence by the proposed time-variant graph representation approach. At the last step, we calculate the graph edit similarity between graph transformation subsequences and find the most discriminative transformation subsequences as graph-shapelet patterns (step ④, 5.5.1). Step ⑤ shows the process of time-variant graph prediction.	85
5.4	Accuracy comparisons with respect to different time stages on both synthetic and real-world time-variant graph data sets.	97
5.5	The average CPU time with respect to different time stages on both synthetic and real-world time-variant graph data sets.	98
5.6	Comparisons with varying graph-shapelet length on both synthetic and real-world time-variant graph data sets.	99
5.7	Two graph transformation sequences extracted from the MIT phone call time-variant graph data. The symbol “N” represents normal person and “H” represents hub person. The first graph sequence shows that a weekly phone call time-variant graph data contain a hub person, while the second graph sequence shows that all the participants are normal persons.	100
5.8	An example of message propagated in a Sina weibo time-variant graph. At different propagation stage, the diffusion (including reached nodes and propagation edges) constitutes a graph. The propagation of the message in temporal order will form a set of temporally related graphs. Each weibo propagation is regraded as a time-variant graph.	101

LIST OF FIGURES

6.1	The rumor “ <i>Malaysian Flight 370 has been found</i> ” propagated on Twitter from March 22 to April 21, 2014. The x axis is the time and the y axis is the total number of tweets including the rumor.	106
6.2	An example of twitter diffusion path. At the unknown time $t = t^*$, the information provenance S_0 initiates the diffusion of a tweet. The propagation time delay between any two nodes is τ and the time window $T = [t^*, t^* + 3\tau]$	108
6.3	Gaussian time delay and the shortest-path propagation. The propagation path from the provenance s_0 to detector S_3 is approximated by the shortest path $\mathcal{P}(S_0, S_3) = \{S_0 \rightarrow S_2 \rightarrow S_3\}$. The propagation delay is an aggregate Gaussian distribution of paths p_1 and p_2	113
6.4	An illustration of the <i>OSS</i> algorithm. Detectors are split into two sets. The first set is observed (activated) within the time window, and the second set is unobserved (inactivated) outside the time window. Function 1 is called by <i>OSS</i> within the time window T , and Function 2 is called after the time window T	120
6.5	A network with one provenance S_0 , and three detectors S_1 , S_3 and S_5 . At the unknown time $t = t^*$, propagation starts from S_0 . Time delay along each edge is $\theta_i \sim N(1, 0.01)$. Monitoring time window $T = [t^*, t^* + T]$, where T is the size of the time window. Detector S_1 is activated at time point $t = t^* + \frac{1}{2}T$. Detectors S_3 and S_5 are inactivated during time window T . We only consider eight nodes $\{S_0, \dots, S_7\}$	122
6.6	Parameter study on synthetic and real-world data sets by using the proposed regression learning model and Online Stochastic Sub-gradient algorithm. The number of hops (error rate) <i>w.r.t</i> : (A) the diffusion provenances number k ; (B) the detector number m ; (C) the monitoring time window T ; (D) the parameter λ ; (E) the parameter ρ . (F) the online detection. The average distance on synthetic/real-world data sets <i>w.r.t</i> propagation time.	126
6.7	Parameter η_t in the proposed Online Stochastic Sub-gradient <i>OSS</i> algorithm.	127

LIST OF FIGURES

6.8	The online detection under the Linear Threshold propagation process.	128
6.9	Running time <i>w.r.t</i> the propagation time on the synthetic and real-world data sets.	129

List of Tables

3.1	The data sets summarization.	34
4.1	Analysis of the new constraint	60
4.2	List of the synthetic and real-world data sets.	62
4.3	F1 score under the parameter support=30 on the four synthetic data sets.	64
5.1	Operation Definitions	84
5.2	Symbols and notations	87
6.1	List of the four data sets.	124
6.2	Comparisons on the four data sets.	131