Mining Time-aware Actor-level Evolution Similarity for Link Prediction in Dynamic Network

Faculty of Engineering and Information technology

Nazim Ahmed Choudhury

Submitted in fulfilment of the requirements of the degree Doctor of Philosophy August 2018

Statement of Authentication

This thesis is submitted to the University of Sydney in fulfilment of the requirement for the Doctor of Philosophy.

The work presented in this thesis is, to the best of my knowledge and belief, original except as acknowledged in the text. I hereby declare that ethical clearance was gained for this work and I have not submitted this material, either in full or in part, for a degree at this or any other institution..

Signed: .....Nazim..... On: 1st August 2018

Nazim Ahmed Choudhury

Abstract

Background

Topological evolution over time in a dynamic network triggers both the addition and deletion of actors and the links among them. A dynamic network can be represented as a time series of network snapshots where each snapshot represents the state of the network over an interval of time (for example, a minute, hour or day). The duration of each snapshot denotes the temporal scale/sliding window of the dynamic network and all the links within the duration of the window are aggregated together irrespective of their order in time. The inherent trade-off in selecting the timescale in analysing dynamic networks is that choosing a short temporal window may lead to chaotic changes in network topology and measures (for example, the actors' centrality measures and the average path length); however, choosing a long window may compromise the study and the investigation of network dynamics. Therefore, to facilitate the analysis and understand different patterns of actor-oriented evolutionary aspects, it is necessary to define an optimal window length (temporal duration) with which to sample a dynamic network.

In addition to determining the optical temporal duration, another key task for understanding the dynamics of evolving networks is being able to predict the likelihood of future links among pairs of actors given the existing states of link structure at present time. This phenomenon is known as the link prediction problem in network science. Instead of considering a static state of a network where the associated topology does not change, dynamic link prediction attempts to predict emerging links by considering different types of historical/temporal information, for example the different types of temporal evolutions experienced by the actors in a dynamic network due to the topological evolution over time,

known as actor dynamicities. Although there has been some success in developing various methodologies and metrics for the purpose of dynamic link prediction, mining actor-oriented evolutions to address this problem has received little attention from the research community. In addition to this, the existing methodologies were developed without considering the sampling window size of the dynamic network, even though the sampling duration has a large impact on mining the network dynamics of an evolutionary network. Therefore, although the principal focus of this thesis is link prediction in dynamic networks, the optimal sampling window determination was also considered.

Method

Considering the trade-offs in selecting the time scale with which to sample a dynamic network, as described above, this thesis developed a novel approach to determine an optimal sliding window by considering a variance analysis of network positional evolutions experienced by the actors in the dynamic network. The determination of an optimal time-scale was followed by calculations of three different actor-level dynamicities (structural, neighbourhood and community) in an optimally sampled dynamic network. Computing the similarity between a pair of actors is an intuitive and dominant solution to the problem of link prediction. Therefore, similarities between the actor-level evolutions experienced by a pair of actors were computed to measure the likelihood of future link formation between them. Three methods were used to compute evolutionary similarity: dynamic time warping, cross-correlation and the Bray-Curtis ecological similarity). Another dynamic feature was developed by considering evolutionary community-aware network structural information in dynamic networks. In a supervised dynamic link prediction setup, a total of nine dynamic similarity metrics/dynamic features were used for

the purpose of dynamic prediction to determine if evolutionary similarity between actor-pairs can measure the likelihood of future link formation between those pairs.

Result

By exploiting actor-level evolutionary network-positional information, this study developed a novel algorithm to discretise dynamic networks. The rationale behind using actor-level measures was that choosing an actor-level measure would create an equilibrium distribution of actor-level network activities over time. The algorithm developed could work in the absence of any actor-level attributes, was applicable to any size of network regardless of size and actor count and was also computationally inexpensive. Different validation methods were proposed to test the optimality of the identified window. The algorithm was found to be effective in all types of networks with any kind of candidate temporal window sizes.

The dynamic features constructed by computing the similarities of micro- and meso-level evolutionary aspects of actors were also found persuasive in the dynamic link prediction task. Considering a list of evolutionary similarity-based features, it was observed that they perform better than the existing prediction methodologies used in static networks and time series-based dynamic link prediction methods. Further, it was found community-aware evolutionary information is advantageous in the task of predicting dynamic links. Furthermore, although both similar and dissimilar actors participate in future links in regard to their evolution similarity, actors with a positive correlation between their evolutionary aspects have better chances of forming emerging links. In relation to the different performance metrics used in this thesis, it was found that these features are not only suitable for the dynamic link prediction task (for example,

predicting the purchasing patterns of online customers, the growth of terrorist networks, etc.) but that they can also be used to understand the underlying network growth effectively.

Conclusion

In network science, it is intuitively presumed that similar actors form links among themselves. Considering the impact of similarity on link formation, this study computed the evolutionary similarities between different types of actor-level dynamicity measures. Since the rate of evolutions depends on optimal sampling of the corresponding dynamic network irrespective of network structure, neighbourhood and community, it is imperative to define the optimal sampling duration for the dynamic network. It was also observed that dynamic similarity metrics/dynamic features constructed in optimally sampled network snapshots perform well in prediction tasks when used in a supervised link prediction model for dynamic networks.

Acknowledgement

With this lodgment of my Ph.D. thesis, my doctoral research journey is about to reach its endpoint after years of relentless work, hustling days and sleepless night. I start by thanking my almighty Allah, the most merciful and most gracious. I am thankful to HIM for not only giving me the opportunity to pursue a research degree but also, giving me the strength and patience to enduring and finally reaching a destination. In this inexorable journey, I remember to miss my daughter, Tazmia's birthdays, my son, Hassan's birthdays, failed to accompany them during their overseas visit and support their mother in all related issues. I am grateful to them for their patience and perseverance during these years and hope one day I can repay my debt and make them happy.

Although, it was not easy for me, and at times, I was about to leave it being demoralized; however, few people encouraged me a lot. Among them, I should first express my sincere gratitude towards Dr Pierre Rognon and Dr Shahadat for their support, motivation, and cooperation that they have extended. Along with them, I am also indebted to Dr. Li Liu and Dr. Mahendra Piraveenan, head and panel member of my annual Ph.D. performance review panel, for their valuable support and constructive feedback. My gratitude to A/Professor Javid Atai, the associate dean of research education, knows no bounds for his support during the review process. I am simultaneously grateful not only to those who inspired me but also to those who dissuade me on this journey. I wish to acknowledge the editing service provided by Dr. Nicole Smits, a research scientist at the Norris Cotton Cancer Center, Geisel school of medicine at Dartmouth and Ms. Belinda Glynn (MA Grad Dip Editing & Publishing). Their quick and prompt proof-editing services and important feedback are highly appreciable.

I want to thank all the office staffs at the University of Sydney - Maria, Daniela, Lorraine, and Jinping for their tremendous support. I also met few fellow researchers and cheerful friends during my doctoral journey. Nazmul, Navid, Abdullah, Nur and Fattha were to name a few in my regular contacts and it was a pleasant experience for all of us to share the positive and negative sides of our research life together. In the end, I want to thank my family – my mother, father, sister, and brother for their constant support. I am also thankful to my extended family members, my uncles, aunties, cousins, and in-laws, especially my sister-in-law Shanu who provided excellent assistance during the time of this thesis writing. This Ph.D. thesis would not be possible without all of your support and inspiration. To my wife, daughter, and son, words will fail to express my gratitude; however, big thank you from the core of my heart for all you have gone through in the last six months. To all my friends and families, wish you many happy returns for your love and support!

Nazim Ahmed Choudhury

1st August, 2018

Articles published arising from this thesis

1. **Choudhury. N,** Uddin. S, (2017) Evolution Similarity for Dynamic Link Prediction in Longitudinal Networks. In: Gonçalves B., Menezes R., Sinatra R., Zlatic V. (eds) Complex Networks VIII. CompleNet 2017. Springer Proceedings in Complexity. Springer, Cham, pp 109-118**.**

2. **Choudhury, N.,** & Uddin, S. (2017). Mining Actor-level Structural and Neighborhood Evolution for Link Prediction in Dynamic Networks. Paper presented at the Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017 ***ASONAM***, Sydney, Australia.

3. **Choudhury N.,** Uddin S. (2018) Evolutionary Community Mining for Link Prediction in Dynamic Networks. In: Cherifi C., Cherifi H., Karsai M., Musolesi M. (eds) Complex Networks & Their Applications VI. COMPLEX NETWORKS 2017. Studies in Computational Intelligence, vol 689. Springer, Cham

4. Uddin, S., **Choudhury, N.,** Farhad, S. M., & Rahman, M. T. (2017). The optimal window size for analysing longitudinal networks. Scientific Reports, 7(1).

5. **Choudhury, N.,** Uddin, M. (2016). Time-aware link prediction to explore network effects on temporal knowledge evolution. *Scientometrics*, 108(2), 745-776

Book Chapter published related to this thesis

1. Uddin, S., **Choudhury, N.,** Piraveenan, M., & Chung, K. S. K. (2017). Exploring Actor-level Dynamics in Longitudinal Network: The State of the Art. In R. Alhajj & J. Rokne (Eds.), Encyclopedia of Social Network Analysis and Mining (pp. 1-17). New York: Springer

Posters presented related to this thesis

1. ***Time-aware Network Structural Similarity Measured for Link Prediction in Longitudinal Networks***, 1st Australian Social Network Analysis Conference' 2016, Swinburne University of Technology, Victoria, Australia.

2. ***Evolution Similarity for Dynamic Link Prediction in Longitudinal Networks***. Paper presented at the 8th Workshop on Complex Networks ***CompleNet,*** Duvrovnik, Croatia

3. ***Complex Knowledge Networks for Scientific Foresight***, Research Conversazione' 2015, Faculty of Engineering and Information Technology, The University of Sydney, Sydney, Australia

Other articles and abstract not related to this thesis

1. Khan, M., Uddin, M., **Choudhury, N.,** (2015). *Fear, Criticism and Awareness – Understanding Sentiment Propagation during the 2014 Ebola Outbreak from Social Media Data*, in the 6th International Conference on Social Media and Society (**Abstract Only**)

2. Khan, M., **Choudhury, N.,** Uddin, M., Hossain, L., Baur, L. (2016). Longitudinal trends in global obesity research and collaboration: a review using bibliometric metadata. *Obesity Reviews*, 17(4), 377-385

3. Ahmed, M., **Choudhury, N.,** & Uddin, S. (2017). *Anomaly Detection on Big Data in Financial Markets*. Paper presented at the Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017 ***ASONAM***, Sydney, Australia

4. Shahbazi, M., **Choudhury, N.,** Shahbazi, M., & Bunker, D. (2018). *Predicting Opinion Leaders in Large Scale Enterprise Online Social Networks.* Paper presented at the IADIS International Conference on Information Systems, Lisbon, Portugal

PUBLICATION STATEMENT (FOR THESIS CHAPTER 3 and 7)

Statement from co-authors confirming the authorship contribution of the PhD candidate:

As co-authors of the paper 'The Optimal Window Size for Analysing Longitudinal Networks', we confirm that Nazim Choudury's contribution to the paper is consistent with him being named as the second author. In particular the candidate's contribution to the following items should be noted:

- Longitudinal datasets collection, data preparation and actor dynamicity calculation
- Proposed algorithm implementation and optimal temporal window calculation
- Conception and complete implementation of the algorithm evaluation section
- Literature review, manuscript preparation and critical appraisal of the content

Author Name	Signed	Date
Shahadat Uddin	Shahadat	11/07/2018
Sardar M. Farhad		12/7/2018
Md. Towfiqur Rahman		15/7/2018

Dedication

To my mum, Sharifunnessa Begum, for everything she did for us, her blessings, and what she went through in her life. I wish all your happiness, Ma.

Table of Contents

List of Figures

List of Tables

Chapter 1

Research Motivation, Objectives & Thesis Organisation

1.1 Introduction

Link prediction is a fundamental task in a complex networked system, such as a social network, where the principal task is to predict the future associations or interactions between networked entities, individuals or actors. These associations are driven by mutual interests inherent to a group of actors [1]. In network science, the principal goal of link prediction is to estimate the likelihood of new link formation [2]. However, most real-life systems are described as evolving networks, where entities (actors) and links (edges) may appear and disappear or attributes of entities and links may vary over time [3]. These evolving networks are called dynamic networks and they can be represented as a time series of network snapshots. In each snapshot, a specific temporal duration is considered to aggregate links regardless of their order of appearances. The aim of link prediction in dynamic networks is to predict future information based on historical data and this information is considered valuable in applications including national security, online recommendations, and organizational studies. In addition to this, link prediction has important practical significance. For example, it can support modelling information diffusion in online social networks, recommender systems for product recommendation and friend or co-author recommendations in a social (collaboration) network, and predict future interactions among biological entities that are expensive to discern through laboratory experiments [4]. Most link prediction strategies consider a static version of the corresponding network where the actor and link structures do not change. This means that the prediction methodologies are insufficient for the task of link prediction in dynamic networks. Although researchers have used time series information for this purpose, they have only considered temporal relational changes (for example, when friends of friends become friends) or the characteristics of pairs of actors (dyadic covariates) instead of the temporal network characteristics of the actors (actor-level evolutionary covariates). Further, most link prediction methodologies in dynamic networks overlook the

problem of defining an optimal sampling resolution to discretise the corresponding dynamic networks. Considering these two issues, this research developed a supervised link prediction strategy in dynamic networks by using some novel features. These novel features denote the similarity between actors in regard to the different evolutionary aspects demonstrated by them in dynamic networks. Further, these evolutionary aspects were quantified in an optimally sampled dynamic network.

This chapter introduces the thesis. First, background information on link prediction is provided, followed by related definitions, a formulation of the research problem and a description of the research issues explored in this thesis. This background discussion is followed by a description of the motivation behind this research in regard to link prediction in dynamic networks, also known as dynamic link prediction, from the perspectives of dynamic network analysis and link prediction methodologies. It also summarizes the rationale behind the research objectives of this thesis. The chapter concludes with an outline of the thesis, explaining the different topics discussed in the subsequent chapters.

1.2 Background

1.2.1 Networks

A network is a graph structure that consists of a set of nodes, also known as vertices or actors, and a set of ties among these actors, known as links. Alternatively, a network is a pattern of interconnections, also known as link structures, among a set of network components known as nodes, actors or vertices. Mathematically, a network can be formally defined as a graph $G = (V, E)$ that consists of the set V of nodes and the set E of edges, which are unordered pairs of elements of V. In this thesis, the words, ‘‘graph’ and ‘network’ are used interchangeably.

Nowadays, a network is a prevalent abstract structure that is used to understand and represent complex systems. Examples of their use include a society that requires billions of individuals to cooperate in order to run smoothly, communication infrastructure that integrates numerous mobile phones with computers and satellites, cognitive systems that require the coherent activity of billions of neurons in our brain and the biological existence of humans, which is dependent on seamless interactions between thousands of genes and metabolites within our cells. Referring again to these examples of systems that exist all around us, nodes, vertices or actors in networks can be individuals, mobile phones, transport vehicles, communication devices, cells, proteins, neurons, animals or any other entity. Links or edges can be any type of connection, relationship or interaction between these nodes, including societal, metabolic, infrastructural or even co-appearances.

1.2.2 Network Topology

In general, topology means the way in which the constituents of a system are interrelated or organized. Network topology is defined by a complete description of the way the components of a network (i.e., the nodes/actors) are connected to each other. There are three fundamental attributes of a network topology: degree, clustering and path length [5]. The topology of networks has been the subject of intensive attention, since it plays an extremely important role in many systems and processes including the flow of data in computer networks [6], the energy flow in food webs [7] and the diffusion of information in social networks [8].

1.2.3 Network Communities

Most real-world networks demonstrate inhomogeneity and reveal a high level of order and organization instead of randomness [9]. Actors in these networks demonstrate a community structure where some groups of actors have a higher density of links among them and other groups have a lower density of links. These densely connected groups of actors organized in networks are commonly referred to as network communities, clusters or modules [10]. J.

Yang and Leskovec identify various reasons why actors form groups in networks [11]. These include individuals forming families, villages, groups and associations to organize a society, topically related webpages on the internet densely linking among themselves and, finally, groups of actors in metabolic networks that are related to functional units, such as pathways and cycles.

1.2.4 Static and Dynamic Networks

In network analyses, a network can be static or dynamic. In a static network, the actors, links, corresponding network topology and communities of actor never change. No new actors or links are added and no existing actors and links get deleted. In contrast, in a dynamic network, new actors are added, new relationships are established between actors, existing actors disappear, and old relationships dissolve over time. These simultaneous appearances and disappearances of actors and links trigger alterations of corresponding network topologies and communities of actors in a dynamic network. Lu, Savas, Tang and Dhillon identified four different factors that contribute towards these dynamics: relational changes (for example, friends of friends become friends), characteristics and/or attributes of actors (i.e., actor covariates), characteristics or properties of pairs of actors (i.e., dyadic covariates) and random unexplained influences [12].

A dynamic network is comprised of different static network snapshots observed at different points in time. These observed networks are called short interval networks (SINs). Figure (1.1) shows an abstract representation of dynamic network consisting of five SINs at five different timestamps (i.e., $t = 1, 2, 3, 4, 5$). A dynamic network $G_T = (V, E_T)$ consists of a set of uniquely labeled actors $V = [v_1, v_2, v_3, \dots v_n]$ and $E_T = [e_t(v_i, v_j, t) | v_i, v_j \in V; t \in T]$ where t represents the timestamp of link e between actor-pair $e(v_i, v_j)$. In addition to this, both the static and dynamic networks can be undirected where $e = (v_i, v_j)$ and $e =$

(v_j, v_i) denotes identical links or is directed where two links are not same. Thus, a dynamic network is composed of an evolutionary sequence of network snapshots $G_T = [G_{t_1}, G_{t_1+\tau}, G_{t_1+2\tau} \dots G_{t_1+n\tau} \dots G_{t'-\tau}, G_{t'}]$ where G_{t_i} denotes an individual SIN or a static network at time t_i. In this thesis, the words, 'dynamic', 'temporal' and 'longitudinal' are used interchangeably.

In temporal (dynamic) networks, there are three aspects of change over time. The first aspect is the temporal changes of associated attributes of actors and links. The second aspect is when the number of actors remain unchanged but the links change over time, represented by $G_T = (V, E_T)$. The third aspect is when both the actors and links experience temporal changes $G_T = (V_T, E_T)$. This study considers the second aspect of the dynamic network where the number of actors remains unchanged but the links change over time.

The temporal arrival and departure of links lead a dynamic network to grow or shrink over time. The process of link formation is considered as a tenet behind the growth and evolution of a dynamic network [13,14]. This process of link formation considers the question of which actors will form associations with each other. In network science, this question is addressed by the link prediction problem.

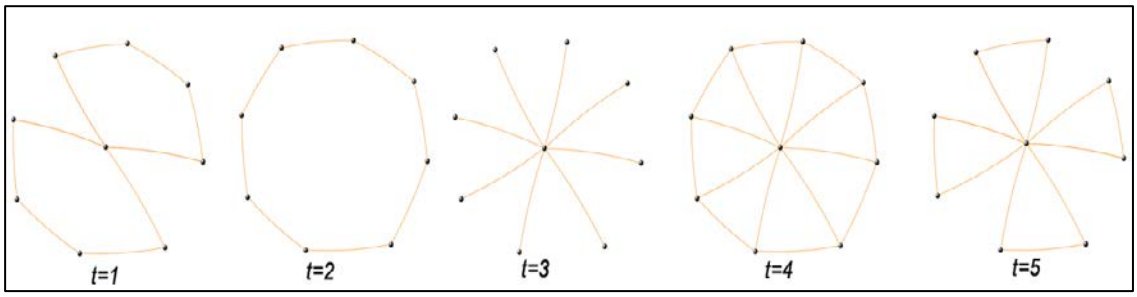


Figure 1.1: An abstraction of a dynamic network in which the state of the network changes over time. Each network snapshot at each individual timestamp is known as a short interval network (SIN).

1.2.5 Link prediction

The task of link prediction is to predict the occurrence of a future link between two actors based on actors' observed links and attributes. Mathematically, given a structure of a network at time t, link prediction models predict which new links are formed in structure at time $t + 1$. The link prediction problem is also considered important for mining and analysing the evolution of social networks [15]. Link prediction methods use the properties of the network, such as link existence, link weights, common neighbours, node degree and clustering coefficients, to predict the link between a pair of actors [16]. There are three categories of algorithms that are predominantly used in network link prediction: similarity-based algorithms, maximum likelihood models and probabilistic models [17]; however, most methods that are situated in these categories consider the underlying network as static. In addition to their inherent limitations (described later in this chapter and in Chapter 2), these methods are found to be unsuitable for the link prediction task in dynamic networks, also known as dynamic link prediction.

1.2.6 Dynamic Link Prediction

In the context of the link prediction problem, two different time intervals (t_1, t') , (t', t_1') where $t_1 < t' < t_1'$, are considered. As discussed above, the primary objective of a link prediction mechanism is to analyse the network structure and actors' attributes in the training phase $[\, t_1, t' \,]$ in order to predict the possibility of future links in the test phase $[t', t_2]$. Therefore, considering two different time intervals, (t_1, t') and (t', t_1') , the network $G_T[t_1, t']$ is used as the network in the training phase and $G_{T+1}[t', t_1']$ is the network in the test phase. In dynamic link prediction task, a finite set of discrete time points within the range $T = [\, t_1, t' \,]$ are considered as $T = [t_1,\ (t_1 + \tau),\ (t_1 + 2\tau) \ldots\ (t_1 + n\tau) \ldots\ (t' - \tau),\ t']$, where τ denotes the temporal sampling interval mentioned above. Fluctuations of the total number of actors are taken into consideration across the time series of network

snapshots. Any link may appear in multiple network snapshots at different timestamp(s). Considering this temporal sequence of network snapshots $[G_{t_1},\ G_{t_1+\tau},\ G_{t_1+2\tau}\ \dots\ G_{t_1+n\tau} \dots G_{t'-\tau},\ G_{t'}]$ for a given pair of actors $(v_i\ , v_j)$, dynamic link prediction attempts to predict the likelihood of link formation between them during the interval (t', t_1') in G_{T+1} by analysing the link formation and temporal information in $[G_{t_1},\ G_{t_1+\tau},\ G_{t_1+2\tau}\ \dots\ G_{t_1+n\tau} \dots G_{t'-\tau},\ G_{t'}]$ at timestamps $[t_1,\ (t_1+\tau),\ (t_1+2\tau) \dots\ (t_1+n\tau) \dots\ (t'-\tau),\ t']$.

Dynamic link prediction has practical significances. Sett, Basu, Nandi and Singh note that it has been studied for prediction tasks in various areas including information retrieval, user and product relationships in recommendation systems, determining the structure of terrorist networks, surveillance systems in communication networks, and the describe the relationships between individuals in a friendship network [18]. In the next section, some important applications of dynamic link prediction are discussed, exploring its wider applicability, thus also denoting the rationality behind this research.

1.3 Applications of Dynamic Link Prediction

Apart from its theoretical value in supporting the study of underlying network evolution mechanisms, dynamic link prediction has a wide range of practical values. Researchers have applied link prediction techniques in different types of networks including social, transportation, disease, communication, and biological networks. A list of interesting real-life problems can be modelled as link prediction problems. ranging from the outbreak of disease, spam email detection, route recommendation to collective classification [19] as well as specialists' predictions of receiving future referrals in healthcare systems [20] and predicting irregular links in disease-gene networks to find genes responsible for diseases [21].

Applications where dynamic link prediction can be exploited are discussed in the following subsections.

1.3.1 Recommender Systems

Recommender systems utilize various sources of information and data to infer users' interests. The basic underlying principle of recommendation algorithms exploits the dependencies between users and item-oriented activities. These dependencies can be better learnt through analysing the historical information of user-item relationships. Many forms of recommendation activities can be performed by using dynamic link prediction strategy. For example, personalized movie recommendations from Netflix [22], job recommendations [23], potential friend recommendations in online social networks [24], potential business or scholarly (for example, patent) collaborator recommendations [25,26], international trade recommendations [27], item/commodities recommendations [28] and predicting users' online ad-clicking patterns from the historical information about their actions and their friends [29].

1.3.2 Security Systems

Link prediction is already used in anomalous mail detection to single out spam emails [30]. It is also applied to discover the missing and/or incomplete information inherent to criminal networks [31], anomalous link discovery [32], and fraudulent call detection in mobile networks [33]. Link prediction also supports privacy control in social networks. For example, Al-Oufi, Kim et al. propose a model that identifies trustworthy people for a given user based on weighted relationships and hence protects the corresponding user's privacy and security from unreliable users [34].

1.3.3 Biological Systems

In biological networks like protein interactions [35] or metabolic networks [36] where discovering potential interactions through laboratory experiment is expensive, link prediction

can provide support and reduce overheads. An important problem in computational biology is predicting gene–disease associations in order to identify the causal disease genes. Some representative researches in this category include studies in [37-39]. Similar foundational applications of link prediction in biological networks include discovering and/or developing new drugs [40], predicting drug sensitivity and/or drug responses [41,42] and symptoms of abnormal parameters of disease [43].

1.3.4 Scholarly Systems

A vast majority of link prediction literature deals with scientific collaboration or citation networks where the objective is to predict future collaboration between scholars [44] or citations of a scholarly contribution [45]. In addition to collaboration and citation networks, link prediction mechanisms were also applied to predict the type and experts of academic research [46,47], identifying missing references to avoid plagiarism [48] and scientist-article cooperation analysis [49].

1.3.5 Communication Systems

Identifying optimal routes is a conventional problem in communication networks (for example, wireless technology). To avoid frequent breaks in routes in mobile ad-hoc networks and to improve the quality of routing in mobile wireless networks, different link prediction mechanisms have been used by researchers [50,51]. In addition to wireless network applications, link prediction has also supported improving transportation efficiency by identifying efficient routing strategies [52] including ensuring information transfer secrecy [53] and optimal routing [54] in sensor networks.

1.3.6 Social Systems

Link prediction mechanisms support the study of social network evolution. The principal application domain of dynamic link prediction is in social networks, including social media.

In online social networks such as Facebook, it can provide potential friend suggestions [24,55-57], special interpersonal links can be advised to users by analysing different social relations [58,59] , social influence detection [60] and information diffusion prediction [61]. Further, link prediction supports complete network inference from partially observed ones to better understand social network evolution [62,63].

1.4 Statement of the Problem

In recent years, inherently network-oriented and ubiquitous web and social media applications have resulted in a strong focus on network and relational data. Network data structure or graph models have become a common framework used to represent and analyse a large number of complex, integrated and real-world interacting systems from nature, society and technology, ranging from the billions of neurons in the human brain and the enormous collection of connected autonomous systems in the internet to the billions of users of social media. These networked systems are massive in size and contain tremendous amounts of content. They are also dynamic in nature and inherently evolutionary. Examples include romantic partners from online dating sites [64], protein interactions, nervous systems, power grids, ecosystems and physical and electronic communication infrastructure [65-67]. One of the inherent underlying structures of these networked systems is their evolution over time in experiencing temporal changes in the overall network dynamics. In these evolutionary networks, temporal patterns emerge through the simultaneous arrivals and/or departures of actors as well as the creation and/or deletion of links among these actors. Characterizing network structures in a time-dependent way or incorporating temporal information to model the dynamics of networks is often complex due to the intermittent existence of actors and links among them [68]. It has also been also found that temporality impacts on most of the dynamic processes taking place in networks [69-71]. Further, according to X. Li et al., high

dimensionality, the quantity of observations, complexity in selecting explanatory variables, sensitivity to noise due to sparsity and computational costs due to non-linear transformations pose major challenges in dynamic network analysis [72]. However, although the mining and analysis of evolutionary networks is a complex non-trivial task, it has drawn considerable research interest [73,74]. Other difficulties in analysing dynamic networks includes associated dynamicity, incomplete data due to topological approximation and limitations in time and space or experimental conditions [75] [15]. Although the mechanism by which evolution takes place in dynamic networks is yet to be congruously standardized or fully understood, network science proposes various methods supporting the study and modelling of the network evolutionary process that governs their dynamics [76]. One of these methods is link prediction. Link prediction is the basic and fundamental computational problem that models the underlying growth mechanism of an evolving network [77]. The emergence of new links, affecting the growth of underlying networks (as mentioned above), is paid the most attention in the analysis of network evolution. Therefore, link prediction mechanisms have attracted extensive research focus as they allow for the extraction of missing information and the evaluation of network dynamics [78]. Researchers consider link prediction the fundamental problem of network science, since it unfolds the mechanism governing the micro-dynamics of a network [79].

1.4.1 Research Motivation

As a time-evolving model, the problem of link prediction in network science has both theoretical and practical significance. Link prediction aims to uncover the underlying relationships among actors in a network to either help find missing links or to infer the future interactions among them by evaluating the likelihood of a link between two actors yet to be connected [80,81]. Link prediction models consume different types of information for these two purposes, including existing historical information either in regards to the network

structure and topology [82] or actor-oriented attributes [17]. The list of network structural information includes the number of common neighbours [83], clustering co-efficient [84] or actor attributes (for example, the degree of connections) [85]. Due to its wide range of applicability, a number of methodological improvements have been proposed to support this partial link analysis in networks. Most of these methods estimate the possibility of the emergence of new links among non-connected network actors by leveraging topological properties, actor/link attributes, local or global network structure [86] or probabilistic models [87]. Two of the major issues with these methods are dependency on static topological feature engineering [88] and failure to acknowledge the temporal changes in networks [89]. Although the link prediction problem is believed to be a time-evolving network analysis model, traditional similarity metrics-based methods generally fail to take the evolutionary aspects of the network into account. Sarkar, Chakrabarti and Jordan identify three weaknesses of these strategies: a dependency on heuristics (for example, counting the number of common neighbours); the use of heuristic measurements in static snapshots of the network (for example, counting the number of common neighbours between two actors in one network snapshot to predict their future association in the next snapshot; and, finally, the non-integration of the temporal components in heuristic measurements (for example, disregarding the temporal neighbourhood changes between two actors) [90].

Temporal patterns emerge in evolutionary networks through the simultaneous arrival and/or departure of actors as well as the creation and/or deletion of links among these actors. This has led scholars to reconsider the evolutionary information in link prediction tasks, resulting in the concept of dynamic link prediction. Dynamic link prediction, also referred to as link prediction in dynamic networks, is the process of inferring the possibility of future links among dynamic entities or network actors through exploring historical or temporal information [91]. Different dynamic link prediction methods explore a wide range of

techniques (described in Chapter 2), including topological evolution in conjunction with different forecasting methods, sub-graph evolution, the dynamic latent space representation of actors and random walk-in temporal networks, the correlation between different types of links along with temporal features (for example, 'recency', temporal activeness), temporal probabilistic measures, non-parametric link prediction methods based on both the features of individual actors and their neighbourhood, machine learning models, statistical models, and matrix or tensor analysis [92].

Despite their improved performance in predicting emerging or hidden links, some of these methods are subject to inherent limitations. For example, probabilistic models require the prior definition of the distribution of link occurrences, which is difficult to define before the actual prediction task and especially in temporal networks [93]. Most existing approaches (discussed in detail in Chapter 2) perform the task of dynamic link prediction by considering the temporal sequences of topological or structural features incident to actor-pairs instead of measuring their similarity and/or proximity by mining actor-level evolutionary aspects including the temporal patterns of neighbourhood changes or evolutionary community-aware information. Further, using the time series forecasting method to predict the future values of topological changes and then using these values for classifier training in supervised link prediction can be incoherent, since the prediction is performed using unrealistic values. To address these issues, it is imperative to consider the evolutionary similarity between actor-pairs when developing features for supervised dynamic link prediction. Further, since the rate of evolution experienced by each actor in a network depends on the sampling duration of network snapshots that make up the corresponding dynamic network, it is also crucial to determine the optimal sampling interval in order to discretise the dynamic network and generate network snapshots.

1.4.2 Research Objectives

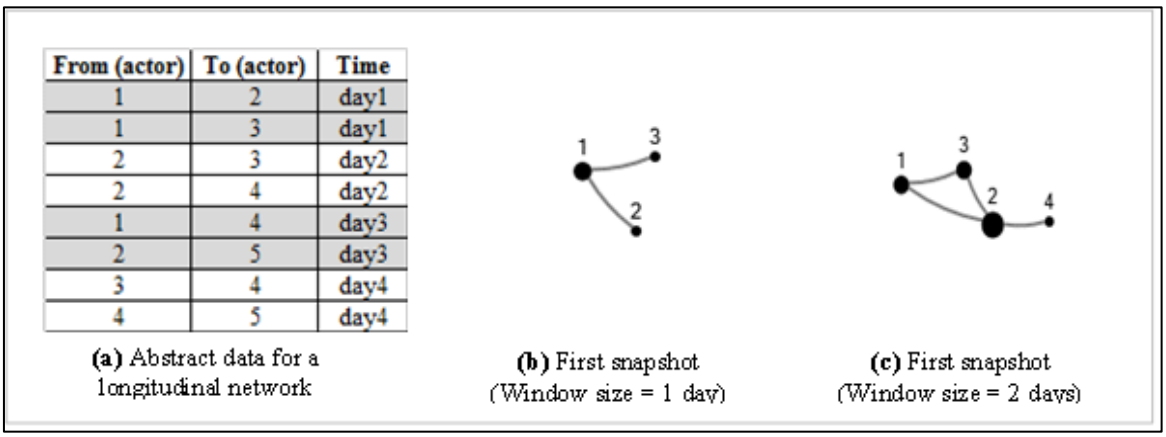
As mentioned earlier, a dynamic network is a time series of network snapshots where each snapshot represents the state of the network over the temporal interval of different granularity (for example, minute, hour, day, month) [94]. The duration of the interval denotes the temporal scale of the dynamic networks, since all links within this duration are aggregated together irrespective of their temporal order. Links within a dynamic network are represented either as streaming interactions in time or a collection of finer aggregated snapshots. In order to achieve meaningful knowledge in dynamic network analysis, it is essential that in the transition from the streaming temporal interactions or a series of aggregated snapshots to a dynamic network abstraction, the extent of discarded information should be insignificant. While researchers usually pay scrupulous attention to the design of their longitudinal studies, they typically pay less attention to the temporal design of their studies. This temporal design refers to the timing and spacing of occasions of measurements [95] or simply the time scale/sliding window mentioned above. Although the selection of this time scale to sample dynamic networks is often done opportunistically [96], the complex temporal structure of a dynamic network is very sensitive to the appropriate selection of this temporal sliding window. This is because a too fine or too coarse window size will either conceal or unravel the important temporal dynamics of the network and the underlying interaction structures of actors [97].

Further, actors in dynamic networks are subject to varying temporal changes (i.e., dynamicity) within the temporal network snapshots due to alterations of different network activities (for example, link formation and link deletion) over time. This triggers temporal changes in the actors' positions and neighbourhood in dynamic networks and, subsequently, this actor-level dynamicity instigates both micro (for example, neighbourhood) and mesoscopic (for example, community participation) changes in dynamic networks. By mining

the similarity or correlation between these diverse actor-level temporal fluctuations (i.e., structural position, neighbourhood and community), it is possible to generate dynamic features for the purpose of dynamic link prediction.

Considering the aforementioned two research issues, this study has two principal research objectives. First, this study develops an algorithm that effectively defines an optimal temporal window to sample/discretise a dynamic network, including some validation methods to evaluate the optimality of the sampling window size. Second, this study develops some dynamic similarity metrics (also called as dynamic features) by measuring the evolutionary similarity between actor-pairs for the purpose of dynamic link prediction. The dynamic similarity metrics, which are similar to the topological similarity metrics computed in traditional link prediction for static networks, are constructed by mining the temporal evolutionary similarity of actor-level evolution (the terms ‘evolution’ and ‘dynamicity’ are used interchangeably in this thesis) between actor-pairs in dynamic networks. In regards to the dynamic features developed in this study, the optimality of the sliding window selection is crucial since an individual actor’s link structure and its structural evolution and network position will vary in each network snapshot depending on the link aggregation [98].

In Figure (1.2), this phenomenon is described pictorially. In Figure (1.2a), a list of time-stamped (i.e., daily) links are collected where the source (i.e., From Actor) and target (i.e., To Actor) actors form links and the temporal granularity of each link occurrence is a day. Figures (1.2b) and (1.2c) show the pattern of network snapshots where the sampling duration for the dynamic network is one and two days respectively. The sizes of the actors denote their number of connections (i.e., degree). The figures show that both the pattern and degree of connections for each actor are dependent on the interval duration considered for



From (actor)	To (actor)	Time
1	2	day1
1	3	day1
2	3	day2
2	4	day2
1	4	day3
2	5	day3
3	4	day4
4	5	day4

Figure 1.2: Differences in network analysis results of an abstract dynamic network that evolved in four days with the consideration of different window sizes. The sizes of actors are proportionate to their degree centrality values. (a) a list of date-stamped links, (b) first network snapshot considering one day time scale, (c) second network snapshot considering a time scale of two days.

link aggregation in a network snapshot. Therefore, before developing dynamic features, an algorithm to detect the optimal time scale (i.e., sampling duration) to sample a dynamic network is proposed in this study.

Considering the aforementioned rationales behind the research objectives, the outcome of this research can be of great importance. By defining an optimal time window/scale to sample a dynamic network and generate a time series of network snapshots, researchers can now map their sampling resolution to the inherent temporal resolution of the

underlying processes of the system considered. Further, instead of the arbitrary sampling of dynamic networks, optimal sampling will demonstrate the actual temporal dynamics of the corresponding network. Furthermore, it will allow researchers to consider temporal information related to the actual prediction task. Similarly, the dynamic features constructed in this study will introduce the notion of similarity-based algorithms in dynamic link prediction tasks. Earlier in this chapter, it was noted that similarity-based algorithms are one of the principal methods used in the prediction task. In a static network, these algorithms generally compute different graph-based topological similarity or actor attribute-based similarity. In the case of a dynamic network, the outcome of the dynamic features from this study will not only allow researchers to model different types of actor-level dynamicity but also to compute their evolutionary similarity, a concept yet to be explored in the complex network research. Further, as explained in Section 2, in conjunction with relational changes and dyadic covariates, actor covariates also contribute towards the network dynamics. Thus, by mining different actor dynamicities, this study will also benefit the future studies on network evolution.

1.4.3 Problem Formulation

Although, link prediction in dynamic networks is complex and challenging, it is important for analysing the associated network evolution and is applicable to a wide variety of applications. Due to its inherent evolutionary nature, link prediction is vital for exploring interesting trends on evolutionary aspects of non-connected actors in dynamic networks. For example, mining micro-scale (for example, network structure and neighbourhood) or meso-scale (for example, community participation) changes that are incidental to actors in dynamic networks can be helpful for predicting the possibility of their future associations. Figure (1.3) shows some of the changes experienced by actors in a dynamic network that is sampled into two individual network snapshots. In this figure, in order to predict a link between actors a_1 and a_2 at

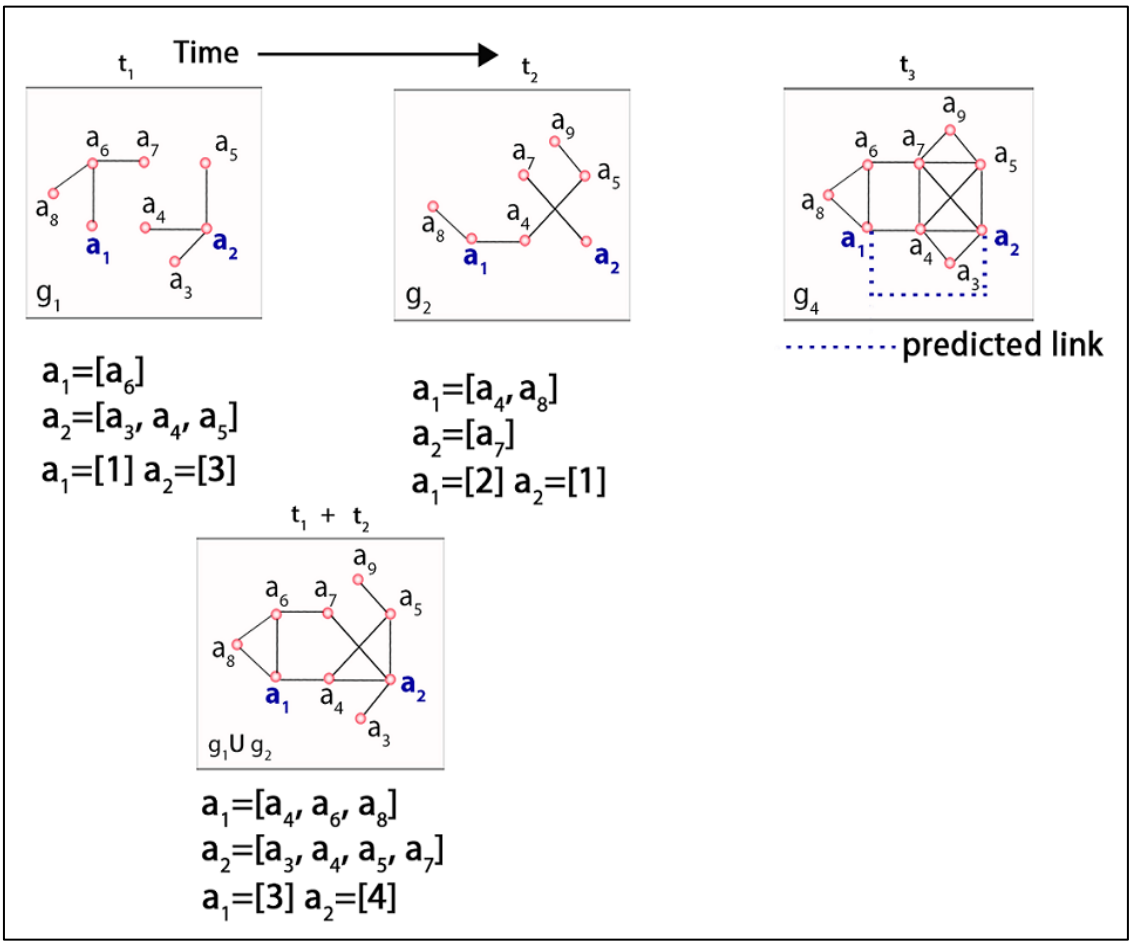


Figure 1.3: Visual representation of addressing the dynamic link prediction problem by considering actor-level evolutionary similarity explored in this thesis. The top row represents a dynamic network which is sampled into two network snapshots g_1 and g_2 at timestamp $t_1$and t_2. To predict the future link between actors a_1 and a_2 at timestamp t_3, this thesis compared similarity between different types of evolutions experienced by both actors over time. For example, at timestamp t_1, actor a_1 has one connection whereas the same actor has two connections at timestamp t_2. Similarly, the same actor has different neighbours at timestamp t_2 (i.e., a_4, a_8) than those at timestamp t_1 (i.e., a_6). Thus, actor a_1 has lower neighbourhood retention rate but higher gaining rate. In the bottom row, the pattern and topology of the network snapshot is represented if the sampling interval was different (i.e., $t_1 + t_2$). Thus, the bottom network snapshot consists of an aggregated network of $g_1 \cup g_2$.

timestamp t_3, the pattern and degree of connections experienced by these two actors are analysed in two SINs at timestamps t_1 and t_2. At these two timestamps, both these actors

have different degrees of connection. For example, at timestamp t_1, actor a_1 has one neighbour and actor a_2 has three neighbours. However, at timestamp t_2, the former has gained one more neighbour whereas the later has lost two. Further, at timestamp t_2, actor a_1 had lost its only neighbour from the previous timestamp t_1 despite gaining a new neighbour. This denotes that actor a_1 has a low neighbour retention rate but a high gaining rate. Similarly, actor a_2 lost three of its old neighbours but gained a new neighbour at timestamp t_2.

Considering the different types of evolutions that are experienced by actor-pairs, in this study, I develop dynamic features by considering the similarity of evolutions incident to pairs of actors. Detecting the optimal and meaningful sliding window (i.e., time scale resolution) for dynamic networks is a fundamental prerequisite for developing these features, given that temporal sampling duration affects the temporal changes of the underlying network structures in dynamic networks. In order to detect the optimal and meaningful sliding window, in this study I propose a novel method to detect an optimal time scale that is applicable to any dynamic network. Looking at optimally sampled dynamic networks, ranges of evolutionary information (for example, network topology, link structure, neighbourhood and community participation) were used to develop dynamic features to use in a supervised link prediction model. The performances of these features were compared against a static predictor (for example, ResourceAllocation [1]) and an existing time series-based link prediction strategy in dynamic networks. Improved performances in dynamic link prediction, as investigated in this study, represent the dynamic features as the prospective candidates not only for dynamic link prediction tasks but also to further understand the underlying evolutionary mechanisms involved in the dynamic networks.

[1] Appendix A.

Choosing the appropriate feature set to describe instances in the classification dataset and classifier's training is one of the most important tasks in supervised link prediction. Traditional link prediction in static networks generally emphasizes the presence or absence of links and simultaneously considers topological information or actor attributes to construe the similarity between actors without considering the temporal information or the evolutionary aspects associated with actors. A key aspect in dynamic link prediction is to generate dynamic similarity metrics (i.e., dynamic features) that consider the evolutionary changes incident to actors. Therefore, in this thesis, I develop dynamic similarity metrics where i^{th} metric will assign a score $sim_i(v_i, v_j)$ to non-connected actor pairs (v_i, v_j) by considering the similarity/proximity of their evolutionary information in $[G_{t_1}, G_{t_1+\tau}, G_{t_1+2\tau} \ldots G_{t_1+n\tau} \ldots G_{t'-\tau}, G_{t'}]$. These scores will measure the likelihood of future links that emerge in G_{T+1}. As mentioned earlier, in dynamic link prediction, the network in the training phase $G_T[t_1, t]$ is sampled using an aggregation granularity (i.e., sliding window/temporal scale) to generate evolutionary network snapshots (i.e., SIN).

1.5 Research Questions

Considering the aforementioned research objectives (i.e., optimal temporal sampling resolution determination and evolutionary similarity-based dynamic link prediction) and the problem formulation in dynamic link prediction, various research questions arose concerning both research objectives. This section describes the associated research issues within both research objectives and the different questions pertaining to these issues. This section also points out the methods this study considered to answer these questions and outlines the research contributions of this research.

1.5.1 Optimal Sampling of Dynamic Network

1.5.1.1 Research Issue

Due to the incorporation of the time component into dynamic network analysis, the choice of the time scale to discretise and aggregate a given dynamic network has considerable implications on the observed of network structures, performing network analysis and the inference procedure adopted on the nature of the network including its processes. However, researchers generally select arbitrary temporal resolutions for the purpose of discretisation. Further, there is a lack of any simple algorithm considering the principal constituents of a network (i.e., the actors) that can work efficiently in networks of any size in absence of any actor attributes. Although researchers have attempted different methodologies to define the optimal sampling resolution, there exists a lack of appropriate validation methods to assess the optimality of window size.

1.5.1.2 Research Questions

- How can actor-level measures be used to determine the optimal sampling interval to discretise a dynamic network?
- How can the optimality of the sampling resolution be validated?

1.5.1.3 Methods

To define the optimal window resolution in discretising a dynamic network, this research considers the variance analysis of actor-level network positional evolutionary measures. To validate the optimality of the identified temporal window, this study utilizes time series-based methods, anomaly-based methods and, finally, an unsupervised clustering approach that is widely practiced in data science.

1.5.2 Actor-level Dynamicity

1.5.2.1 Research Issue

In dynamic networks, actors experience different types and rates of micro-level (i.e., link/neighbours) and meso-level (i.e., community/group) changes. To develop dynamic similarity metrics based on the evolutionary similarity between actor-pairs, it is necessary to determine the types of evolution actors experience in dynamic networks due to those micro and meso-level changes. Therefore, the different types of actor-level dynamicity measures must be defined and quantified to quantify actor-level evolution in dynamic networks.

1.5.2.2 Research Questions

- What kinds of evolutions or dynamicities are demonstrated by actors in dynamic networks?
- How can the actor-level dynamicities be quantified?

1.5.2.3 Methods

The mathematical quantification of actor-level dynamicities is measured by different network metrics (centralities, clustering, neighbourhood, etc.). This study also performs empirical analyses to determine the optimal (i.e., best) and the near-optimal (i.e., second best) sampling resolution for six real-world dynamic networks including an empirical assessment of the optimality of the identified window resolutions.

1.5.3 Dynamic Similarity Metrics

1.5.3.1 Research Issue

Similarity-based algorithms are the most intuitive and dominant methods used in the link prediction models. In case of link prediction in static networks, similarity-based algorithms generally compute a similarity metric by considering the network topological similarity (for example, the number of common neighbours) between actors. However, in an evolutionary

network where both actors and links are dynamic, it is believed that actor covariates (for example, actor-level evolution) have an impact on emerging link formation. Therefore, in the case of dynamic link prediction, it is necessary to define the evolutionary similarity between actor-pairs. Further, it is also an important requirement to assess the impact of the optimal sampling resolution on dynamic link prediction.

1.5.3.2 Research Questions

- How can the evolutionary similarity between actor-pairs be calculated by considering different actor-level evolutions?
- What impact do the evolutionary similarities between actor-pairs have on dynamic link prediction?
- What is the impact of an optimal sampling window interval on dynamic link prediction?
- What kind of actors participate in emerging links of a dynamic network in regard to their evolutionary similarity (i.e., similar/closer or dissimilar/distant)?
- What are the performance enhancements of evolutionary similarity-based features over traditional neighbourhood-based prediction or time series-based link prediction in dynamic networks?

1.5.3.3 Methods

Temporal similarity measures (for example, dynamic time warping), cross-correlation, ecological similarity measures and community-aware structural similarity are used to compute evolutionary similarity-based features. This study also performs empirical analyses to apply the evolutionary similarity-based features in supervised link prediction in six real-world dynamic networks to assess the performance of these features. This performance is then compared to traditional prediction methods.

1.6 Thesis Organization

Before discussing the organization of the thesis, it is worth remembering that this research pertains to two research objectives: optimal time scale determination and evolutionary similarity-based supervised link prediction in dynamic networks. Considering these two objectives and the research questions mentioned in the previous section, an outline of the structure of the thesis is provided in Figure 1.4 that summarises the chapters' contents and the research outcome. A brief description of the contents of each chapter (not including the present chapter) is provided in the following subsections.

1.6.1 Chapter 2

Considering both of the research objectives mentioned above, Chapter 2 reviews the literature related to both research objectives. All related methodologies that define optimal sliding windows and sampling dynamic networks into a series of network snapshots are also discussed. This chapter also categorizes existing dynamic link prediction methodologies and presents a list of the systems where dynamic link prediction mechanisms are utilized.

1.6.2 Chapter 3

This chapter describes the conception of the algorithm proposed to address the research issue related to the first research objective. Further, it also identifies different evaluation methods that can be used to validate the optimality of the identified sampling window.

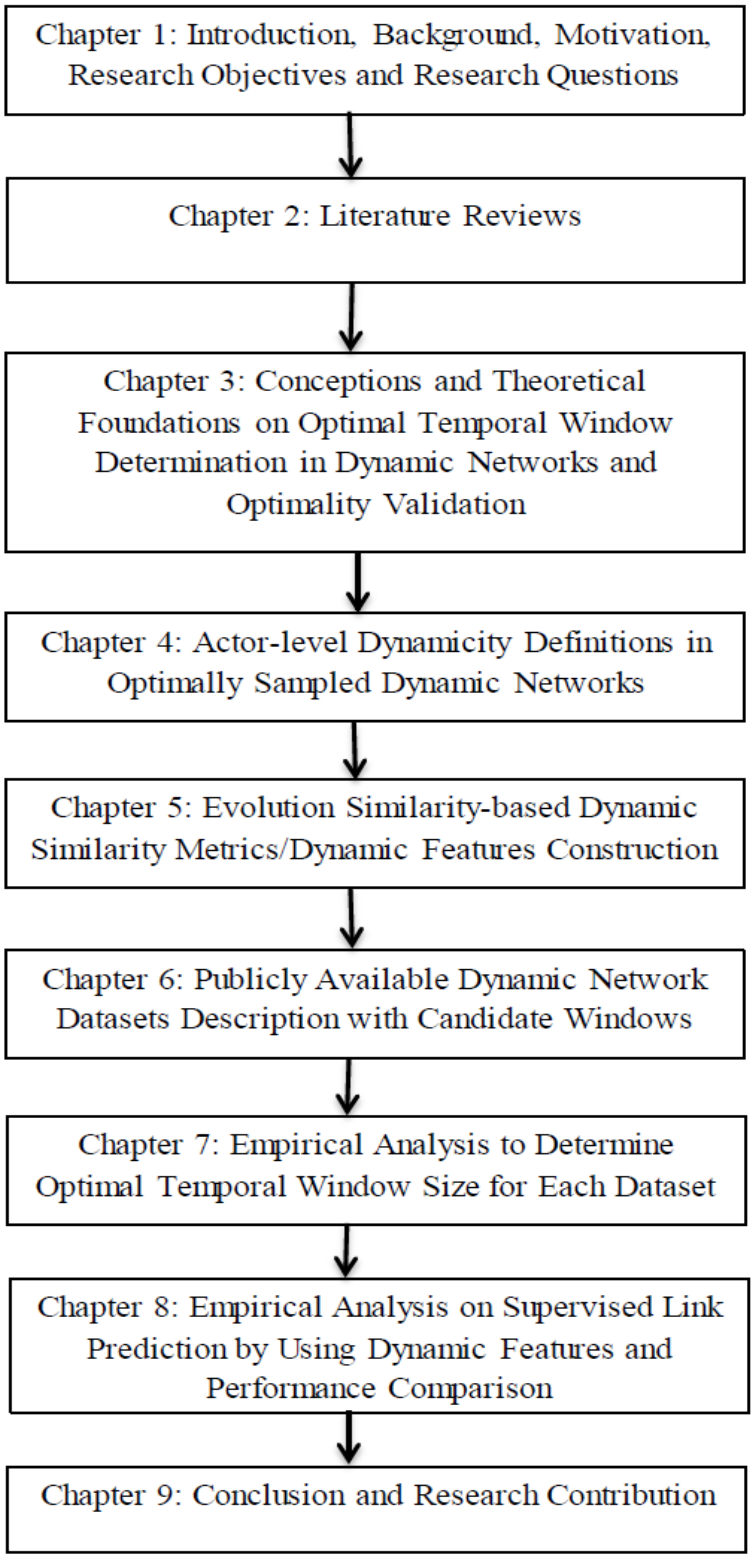


Figure 1.4: Diagram outlining the structure of the thesis

1.6.3 Chapter 4

With a view to address the research issues described in Section 4.2, this chapter develops the quantification of different types of actor-level evolutions. The actor-level evolutions are classified according to their link structure, neighbourhood and the community-aware changes experienced by individual actors in dynamic networks. These different types of actor-level evolutions (i.e., dynamicities) are used in the next chapter, Chapter 5, to develop evolutionary similarity-based dynamic features.

1.6.4 Chapter 5

This chapter discusses the theoretical background and conception of the different methods used in this research. The evolutionary similarity between actor-pairs is then computed. This chapter also discusses the framework of developing dynamic similarity metrics/dynamic features by considering different types of evolutionary similarity between actors.

1.6.5 Chapter 6

This chapter describes the dynamic network datasets, the supervised dynamic link prediction experimental setup, and the performance metrics used for empirical analyses.

1.6.6 Chapter 7

This chapter describes the results of the empirical analysis that was used to determine the optimal sampling window in the dynamic network datasets used in this study. The evaluation methods described in Chapter 3 were applied to the identified optimal window sizes to validate their optimality.

1.6.7 Chapter 8

This chapter describes the empirical results of the supervised link prediction experiment using the dynamic features developed and described in Chapter 5. It also extensively details

the outcomes of the application of dynamic features in link prediction tasks, including feature importance, feature distribution and feature comparisons against existing prediction methodologies. This chapter addresses the research questions outlined in this chapter by using the developed features in the dynamic link prediction tasks in publicly available real-life networks where links are time-stamped.

1.6.8 Chapter 9

This chapter concludes this thesis with a detailed reflection on the research performed, including answers to the research questions of the thesis and avenues of future research.

Chapter 2

Literature Review

2.1 Introduction

A large number of systems (e.g., cell structure in human body constructed by micro-molecules, various social relationships among human beings, and the present day's large technological infrastructure based on the internet) can be modelled as dynamic networks since both network and time aspects exist in these systems. Incorporating temporal information with big data mining on social and other types of networked systems may result I comprehensive big data analyses [99]. The principal advantage of modelling systems as temporal networks, as identified by Holme & Saramäki, is that the behaviour of the dynamical systems can now be expressed clearly without even studying the actual dynamics at all [100]. The list of such behaviours include: (i) network influence measurement among its different parts, (ii) optimization of networks in regards to the dynamic systems, or (iii) similarity of roles played by different actors within the network. Further, temporal pattern of link appearances and disappearances in dynamic networks can affect the dynamics of different systems interacting through networks such as disease contagion or information diffusion [101].

In dynamic networks, since both actors and links among them continuously arrive and leave over time, consequently the corresponding network may grow or shrink temporally. A central scientific challenge in this area is to model the network dynamics with precise description and explanation of the networks growths and shrinkages. Link prediction model is one of the prominent network growth models that capture the localized network dynamics (i.e., which actor will interact with which actor) instead of the global dynamics [102]. Unlike static network, link prediction models in dynamic network attempt to predict future actor interactions based on historical information that would be the most valuable aspects in applications like national security, online recommendations and organizational studies [91].

Since dynamic networks datasets carry additional temporal information (e.g., creation time of links), therefore, this type of network datasets are viewed either as a sequence of network snapshots or as a continuous time process [103]. In this thesis, a dynamic network is considered as a sequence of network snapshots wherein each snapshot links are aggregated for a specific duration irrespective of their appearances within that duration.

In the introduction chapter, this thesis (i.e., figure 1.3) demonstrated that different selections of temporal sampling to discretise a dynamic network effectively impact on the linkage patterns of its actors. Based on the temporal duration of each temporal network snapshot (i.e., short interval network), different types of actor-level evolutionary aspects (e.g., neighbourhood, communities) vary over time. The primary objective of this thesis is to developed different features by mining actor-level evolutions for the purpose of dynamic link prediction. However, the measurement of actor-level evolutions, by using social network metrics, greatly varies depending on the sampling interval of dynamic networks. Therefore, it is imperative to find out the optimality of this temporal sampling interval. Considering this, before delving into existing methods of dynamic link prediction, this thesis will illustrate different approaches, practiced by the research community, to address the optimality issue in sampling dynamic networks first. This chapter will also the motivation behind this issue including the limitations and challenges of existing methodologies to define the optimal temporal scale. This literary background on the first research objective will be followed by a comprehensive literature review on dynamic link prediction methodologies in different categories including their associated limitations.

2.2 Temporal Scale in Dynamic Networks

This section discusses the background information and existing methods on the first research objective of this thesis which is determining the optimal time scale/window to split a dynamic network dataset into a sequence of network snapshots.

2.2.1 Motivation and Background

The evolution in dynamic networks occurs temporally among a set of actors and their links. In recent years, the study of longitudinal social networks has attracted enormous research interest across a wide range of disciplines [104-106] as researchers seek to determine the underlying mechanism(s) of networks' formation, development and evolution overtime. Temporal networks, evolving over time, can be derived from a collection of network interactions by considering any temporal granularity (e.g., minutes, hours, day, month, year etc.). For each of these collections, a static network snapshot is created and the network evolution is measure in regards to the collective rate of changes demonstrated by all actors in these snapshots together.

Temporal streams of interactions in dynamic systems often occur over a range of time scales and are generally aggregated into dynamic networks for temporal analysis [94]. The temporal window or resolution, at which these interactions are aggregated, greatly impact on the results of this temporal analyses [107]. The disparity between the inherent temporal resolution of the underlying process, and the resolution at which the corresponding analyses are performed, can either obscure important insights or reveal inappropriate information. This phenomenon is widely observable in longitudinal studies (e.g., seasonality in mobility patterns of animals, medical cohort study or animal population behaviour [108]). Researchers usually pay more attention to the design of their longitudinal studies and typically ignore the temporal design of their studies. As a result, temporal data are typically collected

opportunistically. It is noteworthy here that by temporal design, it refers to the timing and spacing of occasions of measurements [95]. One of the most important components of temporal design is the size of the time scale under consideration (i.e., the amount of time that elapses between occasions of consecutive measurements). This window size also denotes the bin size and as identified by Fish and Caceres, researchers have named this problem differently, such as, change point detection, time scale detection, oversampling correction, temporal resolution inference, aggregation granularity detection or windowing selection [109]. All time-stamped network activities within each window are generally aggregated for conducting a longitudinal network data analysis. Therefore, the choice of the time scale is a central component in the design of any longitudinal research study which is often overlooked. To stress on the importance of this fact, Moody et al. pointed out that more fine-grained time scale will unfold a great deal of temporal details of the network but conceal some interesting and meaningful patterns (e.g., communities) as the temporal granularity was too short to form them [110]. The authors also pointed out that too coarse temporal scale will not only cause loosing critical temporal information but also fail to engender meaningful observations in regards to the temporal changes of the system and its processes.

2.2.2 Related Work

The problem of identifying optimal time scale for streaming data analysis span over multiple research areas including information theory [111], signal processing [112], time series analysis in econometrics [113], time series segmentation [114] and model granularity [115]. Considering the trade-off between information loss and noise reduction, although the list of literature offers a discretization process of longitudinal/dynamic systems; however, they don’t address the context of dynamic networks. Often the decision of temporal sampling a dynamic network is performed opportunistically [94] depending on a wide range of factors, as identified by Timmons and Preacher in [95]. These include types of social networks,

competing objectives, processes and measurements, a planned time horizon of the respective study, availability of funds, logistic and organizational constraints, availability and expected behaviours of participants of the study and desired accuracy and precision of the outcome of a study. Most theoretical and methodological approaches to define optimal time scale of dynamic networks focus on the aggregation of links in time-window graphs [116]. This impacts the observation bias and affect the accuracy and significance of analysis since dynamic network processes (e.g., formation and dissolution of ties) may begin or end during inter-event times [117]. Other approaches use the rule of thumb in analysing dynamic network which refers that higher number of sampling generates better results [118,119], or , in case of randomized clinical trials, a time scale was chosen that maximizes the efficiency in estimating the treatment effects [120]. Framework of statistical analysis including separable temporal exponential random graph model (STERGM) [121] can also be used by relating the timing of network snapshots to the precision of parameter estimates.

Recent studies focused on empirical analysis by comparing the network statistics of temporal aggregations or some graph metrics over time against some threshold values to determine 'appropriate' or 'meaningful' temporal window size. Few examples include the Temporal Window in Network (TWIN) algorithm by Sulo, Berger-Wolf, & Grossman where the algorithm analyses the compression ratio and variance of time series of graph metrics, computed over a series of graphs comprised of temporal links, as functions of sampling window size [97]. A time-scale or the length of a temporal window, for which the variance and compression ratio are close to each other, defined the optimal time scale. A study by Soundarajan et al. defined another algorithm that identified the variable-length aggregation intervals by considering 'structurally mature graph' that represents the stability of network with respect to network statistics as well [122]. A detailed study in [94] illustrated this time-windowing problem with different types of formalizations and approaches to identify the

optimal resolution of link aggregation. The study considered the time scale of dynamic networks including their corresponding advantages and limitations. Darst et al. followed a parameter free approach by using Jaccard similarity metric [123] to measure similarity between SINs and thus finding the optimal temporal resolution [124]. As mentioned earlier, Fish and Caceres used the quality of the performance of link prediction mechanism to determine the appropriate temporal resolution [125].

Considering the task of determining the optimal time scale of temporal networks as task-dependent, Rajmonda S Caceres and Fish setup a supervised machine learning approach [99]. The authors leveraged ground truth on training data to find best window size in test data. For this purpose, the best window was selected that maximized the performance of the 'task algorithm' in performing a given task effectively. The list of three task algorithms included: (i) link prediction, (ii) attribute prediction, and (iii) change point detection. A list of dynamic network analyses [126-128] incorporated the fact in their assumption that topology (e.g., degree distribution, clustering coefficient) is effectively static and any fluctuation in the network structure only contributes a small amount of unbiased noise to any network measurements. Therefore, in these studies, the authors considered quick sampling of dynamic networks relative to the speed of fluctuations and considered the fluctuations into dynamic network analysis by measuring topological features over a sequence of SINs where the duration of each SIN was very small. As opposed to these study, Eagle considered dynamic topology to analyse network structure and attempted to characterize the effect of the window size in defining SINs on three different measured topological parameters [129]. These included (i) degree statistics, the correlation coefficient and a topological similarity measure. The author also demonstrated that spectral methods can support revealing the known periodicity in the network dynamics and select an appropriate window size. With the help of a new topological similarity measure called 'adjacency correlation coefficient' for comparing

the topology of networks at different timestamps, the author empirically determined the appropriate temporal window size that represented the inferred topological changes over time. With an objective to understand communities, and the discontinuous time points in streaming graphs, Sun et al. presented the GraphScore algorithm [130]. The algorithm monitored communities and their changes in stream of network snapshots efficiently where consecutive snapshots with similar descriptions, in regards to the communities in them, were group together into a time segment. When a new snapshot could not fit well into the current segment, the algorithm introduced a change point and started a new segment at that time stamp. In this way, SINs were aggregated (compressed) to generate optimal length for each temporal graph.

2.2.3 Challenges and Limitation of Temporal Sampling Methods

In general, as mentioned in the previous chapter, researchers are reluctant to pay more attention to the optimal selection of the inherent time scale or the duration of temporal window to discretise dynamic networks in comparison to the consideration paid to the design of associated longitudinal studies. Despite the complex temporal structure of dynamic networks is sensitive to the appropriate selection of this temporal sliding window, however, often, the selection of this time scale to sample dynamic networks is performed opportunistically [96]. The aforementioned approaches to determine the appropriate or optimal timing scale to analyse dynamic networks suffer from their inherent drawbacks. For example, Timmons and Preacher found deteriorating outcome from study using more network snapshots (i.e., SINs) and suggested researchers to consider the trade-offs between precision and sampling time. On the other hand, statistical frameworks are parametric dependent and will only work in small networks with few hundred actors. Therefore, few studies focused on either heuristics-based methods or attempted to optimize for a specific metric over networks. The downside of these methods is the selection of appropriate/perfect

metric for this purpose. Further, the parameter free methods are developed on the assumption of a 'ground truth' time scale which is determined through generative models.

Depending on the amount of intervals for link aggregation in network snapshots of dynamic networks, the link structure of an individual actor will vary accordingly, and so will its structural evolution and network position [98]. To measure the rate of temporal network changes demonstrated by individual actor, it is crucial to consider the optimality of temporal choices in sampling dynamic networks by considering the rate of actor-level evolution. The rationale behind considering actor-oriented evolutionary aspects is that they are one of the central components of the dynamic network and contribute to the localized network dynamics. Furthermore, and most interestingly, despite their consideration of temporal network snapshots, almost all dynamic link prediction methodologies failed to shed light on this important issue. Therefore, prior to developing dynamic features by mining actor-level dynamicities, it is imperative to develop an algorithm to determine the optimal time scale to sample dynamic networks.

In the next section, the background information on dynamic link prediction and a detail description of different methods addressing this issue are elaborated.

2.3 Dynamic Link Prediction

Understanding and characterizing different processes driving interactions in networked systems is one of the fundamental research issues that have drawn considerable research attention from the network science community. Link prediction problem, dealing with prognosticating different types of interactions, collaborations, associations, or influence between actors in a network, encompasses an unprecedented amount of literary contents alongside its application domains in sociology, biology, anthropology, and information systems. To deal with the challenging task of inferring emerging and/or missing links,

researchers exploited a wide range of network structural, relational and temporal features, and compounded different sources of both network and actor-related information in their models to improve prediction performances.

Initially, link prediction problem was formulated as a generic data mining problem within the field of relational learning that included both the link structure and rich sets of descriptive attributes of the linked data objects. Probabilistic relational models, developed by Getoor & Sahami, was one such earliest relational learning model that represents the statistical correlation between one entity to the related others in regards their properties [131]. With the help of rich knowledge structure, encoded by relational archetypes, these models supported the reasoning of behind entity relationships. Lisa Getoor et al. extended the probabilistic realtional model with the help of two mechanisms to represent a probabilistic distribution over link structures of different relational entities [132]. Among others, relational Markov networks [133], structural logistic regression with a process to systematically generate features from relational data [134] and stochastic relational models [135] are few examples of relational link prediction algorithms. These models are only applicable to abstract graphs (i.e., networks without any actor or link attributes) and static graphs (temporal changes of links structures are not incorporated) where the link structures are the solitary source of predictive patterns, and incompetent in comprehending the complex graph-patterns (e.g., cliques, cluster).

Subsequently, with the advent of network science [8], and social network analysis [136,137], to comprehend the underlying relational structure among different entities, network representation became the dominant data structure to model interactions among actors of different complex systems. Simultaneously, these have supported studies of ego-centred networks [138] and outlined measures to describe and understand actor-oriented

complex network structures [139]. Consequently, different link prediction strategies in networks were proposed as these studies identified the basic models of axioms governing network formation and its structural features.

By analysing social networks of co-authorship networks, Liben-Nowell and Kleinberg studied first the link prediction problem in social networks to provide some apposite insights of the problem including special reference to some classical topological measures [77]. The prediction paradigm in this study typically extracts the similarity or proximity between a pair of actors in the network by exploiting various graph-based similarity scores and ranked them for the prediction of emerging link among those actors. Another notion of characterizing link prediction problem was rendered by Al Hasan et al. where the authors, in one hand, demonstrated that using extrinsic attributes other than graph topology can significantly increase prediction performance, and on the other hand, they used these features in a supervised learning setup by considering the link prediction problem as a binary classification task [140]. To date, these two studies are considered as the most pioneering and influential study in link prediction and reminiscent to the subsequent methods suggested by other scholars addressing the link prediction problem.

Several commendable survey studies attempted to enumerate different strategies addressing link prediction problem. Lü & Zhou summarized popular link prediction algorithms for complex networks emphasizing on actor-similarity indices, maximum likelihood and probabilistic methods [17]. Al Hasan & Zaki categorized some representative link prediction methods for social networks into four different categories: (i) feature-based, (ii) Bayesian probabilistic, (iii) probabilistic relational, and (iv) linear algebraic models [86]. P. Wang et al. presented the most comprehensive and systematic survey study on link prediction in social networks which is suitable for beginners to understand the underlying problem definition, scopes, concepts, and different aspects focusing on this problem [15]. In

conjunction with a further comprehensive link prediction definition, general solution framework and evaluation metrics, the authors proposed a state-of-the-art categorization of link prediction strategies in two perspectives: (i) link prediction technique, (ii) link prediction problem. The former includes four different aspects: (i) actor, (ii) topology, (iii) social theory, and (iv) learning features. The later included six categories: (i) temporal link prediction, (ii) active/inactive link prediction, (iii) link prediction in bipartite networks, (iv) link prediction in heterogeneous networks, (v) unfollow or disappearing link prediction and (vi) link prediction scalability. Haghani & Keyvanpour classified the link prediction problem into two categories, independent of the procedures followed by different methods: (i) missing link prediction and (ii) future link prediction [141]. The latter (i.e., future link prediction) was further subcategorized into two categories, namely (i) periodic and (ii) non-periodic, where periodic link prediction that includes a series of evolutionary networks resembles the dynamic link prediction strategies. Similarly, Srinivas and Mitra categorised the link prediction literature into six different categories, namely: (i) static link prediction using local and global similarity metrics, (ii) link prediction in heterogeneous networks, (iii) link prediction in signed networks, (iv) unsupervised and supervised learning based algorithms, (v) semi-supervised learning algorithms and, finally, (vi) dynamic link prediction [142]. Among all these categorizations, temporal/dynamic/periodic link prediction deals with the challenge of predicting dynamic interactions among actors in the network over time which is different from the traditional link prediction problem that has no temporal aspects associated. Before delving into formal definition of dynamic link prediction, we need to define the concept of dynamic networks:

2.3.1 Dynamic Link Prediction in Homogeneous Network

A homogenous network is composed of similar type of actors and links. For example, Facebook Friendship is a homogeneous network where actors are Facebook users and a link

denotes a friendship relation between two users. In the next sections, different dynamic link prediction methodologies, applied over homogeneous networks, are discussed

2.3.1.1 Matrix Factorisation

In dynamic networks, nothing is stable. The latent positions of the actors in network evolve with the temporal evolution that takes place in the network. Considering the latent space modelling of static networks, a naïve approach for dynamic networks is to extend this method to model each actor using a single latent representation and subsequently update its position as the corresponding network evolves. However, this modelling tends to suffer, as identified by L. Zhu et al., from poor incorporation of historical information and abrupt transitions due to its overfitting tendency on the current time step [143]. Therefore, the authors attempted to infer the temporal latent positions for all actors in dynamic networks using two variants of block-coordinate gradient descent (BGCD) algorithm [144] which is widely used to infer low-rank latent space in networks through matrix-factorisation. A local BGCD algorithm was introduced that sequentially infers the latent space at each timestamp with a single SIN and previous temporal latent spaces instead of jointly inferring temporal latent space in all timestamps. It also supports reducing the computational cost. Dunlavy et al. and Acar et al. exploited singular value dicomposition (SVD), and Eigen-decomposition (ED), with the help of low-rank approximation, to predict links in dynamic networks [145,146]. The authors used both weighted and unweighted methods to collapse temporal data inot matrix format, followed by an extended version of Katz method [147] known as 'truncated Katz score' by using truncated matrix single value decomposition and CANDECOMP/PARAFAC (CP) [148] tensor decomposition method for multi-periods temporal link prediction. Both matrix and tensor factorisation were used by the authors where tensor factorisation is a higher-order extension of matrix factorisation and dramatically improves the prediction accuracy. The CP algorithm produces a highly interpretable factorisation in regards to the time dimension. W.

Yu et al. developed a dynamic link prediciton model that supported both spatial and temporal consistency [149]. They leveraged the time-dependent matrix factorisation to decompose the network adjacency matrices into time-dependent matrices, and capture the features of actors in a dynamic network. The authors also introduced the network propagation constraint by adopting label propagation principle under the practical assumption that two actors are similar if they have the similar feature and having the same label. The network propagation constraint ensures that actors remain in the close proximity of their neighbours in the hidden feature space, learnt by the time-dependent matrix factorisation.

The aforementioned matrix and tensor factorisation based methods predict temporal links from a collapsed temporal network by ignoring the connection between temporal network snapshots. As reported by Ma et al. in their recent study that these methods are incompetent to incorporate the evolving information into feature extraction which is critical to dynamic network analysis and results in undesirable prediction performance [150]. According to the authors, regularization method that integrates the intrinsic geometrical structure of the data space is considered as a novel approach to address this issue. Consequently, by considering Non-negative Matrix Factorisation (NMF), they proposed a new algorithm, named Graph Regularized Non-negative Matric Factorisation (GrNMF) for dynamic link prediction with a view to improve the prediction performance by using a graph regularization strategy. Unlike others, instead of collapsing a dynamic network, it factorizes the SIN at time t by setting SINs from timestamp 1 to $t-1$ as a regularizer and each SIN is weighted to be incorporated into the objective function of GrNMF. The principal advantage of this algorithm is that in one hand, it can leverage the power of Non-negative Matrix Factorisation (NMF) and graph regularization, and on the other, the framework can be extended to incorporate other information about dynamic networks (e.g., community membership).

By considering the equivalence between the Eigen-decomposition (ED) and Non-negative Matrix Factorisation algorithms (NMF) and graph communicability, Ma et al. proposed two NMF-based frameworks for temporal link predictions [151]. The first framework collapses temporal features and the second one collapses the temporal networks. On the basis of matrix Factorisation formalism that combines both content and link information in conjunction with the aforementioned graph regularization method, Gao et al. proposed a unified model of dynamic link prediction that integrated three types of information: (i) global network structure, (ii) actor's content, and (iii) network proximity information [152]. The model used both latent matric Factorisation and graph regularization methods including efficient optimization procedure that supported learning from latent factors.

2.3.1.2 Statistical Model

In this section, different statistical models, dependent of probabilistic distribution, are described:

2.3.1.2.1 Probabilistic Generative Models

In dynamic link prediction, probabilistic generative models were fitted to a sequence of observed networks and in these models; a dynamic network is represented by a set of unobserved parameters. The values of the parameters, estimated from a sequence of t SINs, provide the probability score of a link between actors at time $t + 1$. In these models, the prediction accuracy scores are used to measure the goodness-of-fit. Junuthula et al. identified two different types of generative models, namely: (i) latent feature model and (ii) dynamic stochastic block model [153].

In latent feature models, each actor has an unobserved feature vector, and a link between an actor pair is defined conditionally independent of all other actors, given their

feature vectors. Although, related studies [154-156] demonstrated them as tremendously flexible; however, they exploited Markov Chain Monte Carlo (MCMC) method which is incompetent to scale up for large number of actors. On the other hand, in stochastic block models, actors are divided into classes where actors in the same class have identical statistical features. Although the probability of a link formation between two actors is independent of statistical properties of all the other actors; however, it depends on the classes of the corresponding actors. Temporal changes of link probabilities and class membership are associated with stochastic block models to support dynamic link prediction. The representative studies [157-159] of stochastic block models are advantageous over latent space models in regards to their scaling capability in regards to the actor quantity (i.e., few thousands than few hundreds in the latent feature model).

2.3.1.2.2 Other Probabilistic Models

With a view to propose a user recommendation model in social networks, Barbieri et al. proposed a stochastic generative model that jointly factorized both social connections and feature assoications [160]. The model, called WTFW (Who to Follow and Why), is a type of stochastic topic model that can not only predict links in directed and actor-attributed networks but also provide explanations on each predicted links whether it is ‘topical’ or ‘social’ in the context of online social networks. WTWF depends on latent factors (e.g., communities containing actors with similar behaviour) and explicit modelling of the underlying latent nature of the corresponding observed links. Besides, Hanneke and Xing proposed an extension of the Exponential Random Graph Models (ERGM) [161,162], where apart from adopting many methods and theorems from ERGM, Markov Chain Monte Calro (MCMC) maximum likelihood estimatmation algorithms was also applied to model the evolution of social network over multiple sequential observations [163].

By considering valuable temporal trends, emergent in dynamic networks, Potgieter et al. developed different temporal metrics and used Bayesian networks to model the interrelationships between local and emergent behaviours of actors [164]. The temporal metrics measured the local evolutionay behaviour of actors with an assumption that the formation of future associations denotes their emergent behaviours (e.g., the change percentage of the number of common neighbours) in a social network. The list of temporal metrics included degree, betweenness of actors, topological metrics (e.g., preferential attacment, Katz and AdamicAdar) and 'recency' to denote the time elapsed since an actor formed the last link. The authors also considered dyamic Bayesian network which is a directed acyclic graph (DAG) and the concept of social resource combinations [165]. The actual realtionships between the temporal metrics and link formations in each SIN at timestamp t was determind by mining the trained components of dynamic Bayesian network.

Markov chains, the favourite framework to model website navigational behaviours of users, were chosen by J. Zhu et al. to develop a Markov model based dynamic link prediciton strategy [166]. Networks were constructed from the web log files including an algorithm for transition probability matrix compression to cluster web pages with similar transition behaviours. The authors also used a mechanism called 'maximal forward path' to improve the prediciotn performance that denoted a sequence of maximally connected pages by a user in the probaility calculation [167].

2.3.1.2.3 Statistical Relational Models

As mentioned in the introduction section, in temporal networks, three aspects changes over time. Firstly, the temporal changes of associated attributes of actors and links, secondly, the number of actors remain unchanged, however, the links change over time, and finally, both the actors and links experience temporal changes. Considering the first case (i.e., changes of attributes of actors and links over time), Sanghai et al. extended the probabilistic relational

models (PRM) [168] to develop a separate PRM for each SIN in given dynamic network [169]. The authors modelled the dependencies of attribute values in a SIN at an individual timestamp to the next. Milch and Russell considered the final case where both actors and links experienced temporal evolution and developed a dynamic link prediction model by introducing Bayesian Logic [170]. It is considered as a first-order probabilistic modelling language that specifies probability distributions with varying sets of objects. Among other statistical models, Sharan and Neville attempted dynamic link prediction by incorporating time-varying dependencies into relational models [171]. The authors believed that temporal interaction dynamics contain valuable information that can improve prediction accuracy. The authors represented dynamic networks by aggregating the sequence of links between any pair of actors into one link with a weight, calculated by an exponential weighting scheme. As a result, a sequence of temporal network snapshots (i.e., SIN) would become a static weighted graph. The authors then incorporated the link weights in a relational Bayes classifier for the prediction purpose with a view to moderate the influence of attributes throughout the SINs.

2.3.1.2.4 Probabilistic and Matrix Factorisations

In these models, probabilistic dynamic modelling was proposed based on matrix Factorisation to deal with the time-varying relational data. One such model was proposed by Hayashi, Hirayama,and Ishii based on dynamic extension of matrix Factorisation [172]. In this model, the dynamic evolution of a sequence of relational matrices was modelled using low-rank matrices sampled from an exponential family distribution. The authors applied Laplace approximation to derive the sequential Bayesian estimation and capture the temporal variations of latent low-rank relationships effectively. By introducing original generalized linear models (GLM) [173] in the context of matrix Factorisation, known as Exponential family Matrix Factorisation (EMF), the authors also demonstrated that their model performed well in both real-world and synthetic networks. Sarkar & Moore proposed a similar method

of probabilistic dynamic model based on matrix Factorisation, where the authors introduced a latent space model for temporal networks [174]. Their model estimated the distance between actors in the network by using Bernoulli natural parameter space and extended the latent space model for static link prediction by considering temporal correlation of actors' position in latent spaces. However, the authors relied on Markov assumption in identifying latent locations of actors where the latent position of actor at time $t + 1$ is independent of all previous locations given its latent location at time t.

2.3.1.3 Machine Learning Model

Machine learning strategies have also been exploited in dynamic link prediction. Based on supervised rank aggregation, Pujari and Kanawati developed a model that aggregates unique information, provided by each attributes of actors in the network, and introduced weighting scheme to the rank aggregation method to predict future association between them [175]. The authors expressed the link prediction problem as a political election process where the voters are different topological measures and candidates are the non-connected pairs of actors. Vu, Hunter, Smyth, and Asuncion proposed a regression based modeling framework for dynamic link prediction that incorporates both time-dependent network statistics and time varying regression coefficients [103]. Leveraging the concept of survival and event history analysis [176], the authors employed a multivariate counting process including both multiplicative and additive intensity functions. These functions incorporated random network statistics and time varying regression coefficients. The additive approach, supported by this model, provided an efficient inference scheme to estimate the time varying coefficients and allowed scaling up for large networks. Zeng et al. developed a dynamic link prediction method called Self-trining based Link Prediction using Temporal features (SLiPT) by using semi-supervised learning [177]. The authors were motivated by the fact that the potential information from large number of non-connected actor pairs can improve the prediction performance. Their

dynamic link prediction strategy used two temporal features, borrowed from the study by Potgieter et al. [164] (e.g., 'recency', the degree of actors in each time stamp), in conjunction with six topological features.

The prediction performances of different topological measures vary intensively. Considering their performance variance and instability, researchers attempted to organize them together to develop ensemble algorithm to reduce the likelihood of selecting the worse preforming one and hence obtaining the stable link predictor. This motivated Y. He et al. to developed an ensemble algorithm for dynamic link prediction by consiering three different ordered weighted averaging (OWA) operators [178]. The operators included were maximum entropy, minimum variance and chi-square methods. These operators assign weights to nine common neighbourhoods-based topological measures and then aggregate their results to obtain the final prediction score. Temporal evolution in dynamic networks may engender new dimension of attributes (e.g., community memberships) of actors and links including multiple features for each dimension. Considering this fact, Bao et al. developed a prediction strategy based on principal component analysis (PCA) [179]. By using principal component regression (PCR), the authors attempted to devise a robust link prediction mechanism with optimal time complexity by using automatically identified features. These features were contextually important and mostly not derived from link topology.

As dynamic social networks evolve over time, the volume of network datasets become larger and contains a large quantity of network events (e.g., publications, communications). O'Madadhain et al. developed a novel prediciton strategy based on the co-participation likelihood of actors in different network events in conjunction with temporal changes of their ranks (e.g., influence, level of participation) in regards to the participations in series of events. predicting potential cooperation between social entities in social events [180]. The authors explicitly incorporated two important aspects of event data (i.e., time and

sequence) in conjunction with scalable and robust machine learning techniques. Instead of temporal events, Bringmann et al. considered the typical patterns of structural chnages in temporal networks including association rule mining and frequent pattern mining to develop a paradigm of learning and predicting social network evolution [181]. Motivated by their preceding study on graph evolution rules [182], the authors developed Graph Evolution Rule Miner (GERM) software that extracts graph evolution rules and support the prediction of emerging links in dynamic network. Considering three challenges in dynamic network analyses (i.e., high dimensionality of responses, large number of observations, and complexity associated with explanatory variable selection), X. Li et al. proposed a deep learning framework (i.e., Conditional Temporal Restricted Boltzmann Machine) for dynamic link prediction [91]. It is a generative model in exponential family that integrates neighbour influence as adaptive bias into the energy function and employed the exponential capability to capture nonlinear variance in dynamic link prediction. According to the authors, this machine learning based approach is robust to noise and tackles the computational cost of learning and inference with the support of efficient Neighbour Influence Clustering algorithm. By employing low-dimensional latent space, Z. Zhang et al. proposed a machine learning based incremental dynamic link prediction algorithm [183]. In this algorithm, the authors employed the non-negative symmetric matrix decomposition in conjunction with block-cordinate gradient descent (BCGD) [184] algorithm to optimize the learn process. The assumption behind this algorithm is that actors with shorter distance in latent space have higher likelihood of forming links in future.

Optimization of machine learning algorithms is computationally intensive. To minimize the cost incurred by optimization process, Bliss et al. developed a linear model that combined both neighbourhood-based topolgical similarity metrics and actor attributes in an evolutionary algorithm in an inquest for the coefficients supportive to the optimization

objective [102]. The authors used Covariance Matrix Adaptation Evolution Strategy (CMA-ES) to optimize weights of a linear combination of sixteen topological metrics and actor attributes. A further optimization technique, ant colony optimization [185] was used by Sherkat et al. to develop a subgraph evolution based dynamic link prediction technique [186]. The unsupervised structural link prediction algorithm, based on the foraging behaviour of ants, studied the evolution of specially constructed subgraphs to predict emerging links.

2.3.1.4 Temporal Measures

Link topologies, in association with temporal information, play a critical role in dynamic link prediction. For example, the rate and length of communications provide indications on the type of relationship involved (e.g., family, commercial) [187]. It is also understood that the time of interactions between actors in a network is a dominant feature for ranking neighbours in regards to the likelihood of future association with a particular actor [89]. Therefore, different temporal methods have also gained attention from the researchers:

2.3.1.4.1 Univariate Temporal Sequence

This is the most generic and highly used method of dynamic link prediction. In a most straight approach, a time series of frequency of link occurrences is built by considering temporal sequence of network snapshots (i.e., SINs) and Auto Regressive Integrated Moving Average (ARIMA) [188] models are used to predict the future links [189]. Another similar approach was developed by Ibrahim and Chen where the authors developed a reduced static graph approach by incorporating both frequency and temporal information [93]. In this method, highly frequent links have higher probability of future appearances and a damping factor was used to denote the importance of a link based on the time of its occurrences. This type of time series model is capable of predicting the future occurrences of only repetitive links through interlink dependencies over time. Instead of considering link frequency, some other studies incorporated different time-varying structures. These included density and,

diameter [155], sub-graph and cycle structures [128] and clustering patterns [190] in a series of SINs. Recently, learning automata-based time series link prediction (LA-TSLP) was proposed by Moradabadi and Meybodi that used a set of Learning Automaton (LA) to predict future links [191]. As an adaptive decision-making tool, LA attempts to learn the optimal action from a set of allowable actions based on a probability distribution over the action set by interacting with the random environment [192]. Both a learning algorithm and reinforcement signal govern the updates of probability distribution of its action(s).

2.3.1.4.2 Network Structural and Topological Metrics

Some dynamic link prediction methods exploited the temporal network structural variations in dynamic networks. For example, Sarkar et al. developed a non-parametric algorithm for link prediciton in dynamic networks by considering both topological features and local neighbourhood of actors in different partitions of time domain [193]. The rationale behind their method is that the sociality of actors in social networks affected by its neighbours. By investigating the relationship between graph structure and link occurrences, Murata and Moriyasu developed a weighted graph proximity score for dynamic link prediction [194]. The authors assumed that association of link weights with graph proximity score can boost the performance of the dynamic link prediction

Topological similarity metrics, widely used in link prediction in static networks, are incorporated in time series to devise the most eminent dynamic link prediction method. In this method, a time series of a chosen topological metric, denoting the proximity and/or similarity between a pair of actors, is constructed to acquire historical information of their topological changes. Similar to the univariate temporal sequence based methods, ARIMA models were used to estimate the probability of future link occurrences based on the forecasted topological similarity scores from the constructed time series. Representative studies in this category were performed by different scholars in [195,196,189].

Using neighbourhood based topological structure in conjunction with temporal information is a common process of generating time-aware topological metrics for dynamic link prediction. For example, Zhang et al. used an improved version of 'ResourceAllocation' algorithm [183]. Despite, ResourceAllocation [1] has a similar form like AdamicAdar[1] algorithm; however, the former suppresses the contribution of the high-degree common neighbours more than the latter. The authors had the ResourceAllocation algorithm modified by considering the degree of the common neighbours including the degree of the concerned actors of a link. Subsequently, their modified algorithm was capable of avoiding re-computation of the whole network in case of any temporal structural changes.

In case of dynamic link prediction in temporal directed, Bütün et al. proposed a neighbour and graph pattern based topological measure that not only considered the link direction but also link weights and associated temporal information to imporove the prediciton performance [197]. Besides this, few other studies explored the dynamic link prediction mechanism in directed networks. For example, Schall proposed a new metric called Traidic Closeness (TC) that calculated the ratio of the number of closed traids in comparison to the number of potential triads [198]. Similarly, Romero and Kleinberg demonstrated the importance of directed triadic closure on link formation in online social networks like Twitter [199].

2.3.1.4.3 Temporal Communities/Cluster

In dynamic networks, structure of the network evolves over time where new links may arrive among new actors, new and existing actors, or between two existing ones. Link inference decision in temporal networks can be made using the combination of types of actors and links, including the heterogeneous contents associated with them within a particular structural locality (e.g., communities) of the network [200,201]. Therefore, dynamic community

[1] Appendix A

detection/clustering approach was exploited for link prediction in both homogeneous and heterogeneous networks. Aggarwal et al. used a dynamic graph-clustering approach in the content-rich networks where clusters were created based on structural similarity [202]. In this method, fine-grained clusters were created and maintained in evolving networks to generate local regions. Then content-specific linkage behaviours of different attributes within these local regions were exploited to predict future links. The authors demonstrated that clustered subnetworks in conjunction with relational attributes of different actors can further support the prediction of likelihood of future links.

In homogeneous networks, Rossetti et al. formalized the link prediction problem as a 'interaction prediction' paradigm [203]. The authors combined dynamic social network analysis, time series forecasting, feature selection and network community structure to predict future interactions. In their method, the modularity structure of dynamic networks was considered as an important topology since it represents the boundary of sociality of social actors. The evolution of such boundaries denotes changes of actors' social behaviours. The authors also divided the original problem into two disjoint tasks, namely (i) intra-community and (ii) inter-community interaction prediction. To extract community-aware features, the authors took advantage of three community detection algorithms, namely: (i) Louvain [204], (ii) Infohiermap [205], and (iii) DEMON [206]. Likewise, considering clustering information, Yuan et al. proposed a new dynamic link prediciton model [207] in association with MapReduce technology, a core component of Hadoop distributed system [208]. A similar appraoch of link prediciton in temporal networks was proposed by Ibrahim and Chen, known as Integrated Time series Model (ITM) where the authors used a combination of information from actors' communities and centrality measures in conjunction of time series information [93]. In this model, the authors used modularity Louvain method [204,209] for community

detection purpose and eigenvector centrality measures to compute the importance (i.e., centrality) of actors.

To study the link prediction in followee-follower networks of Twitter, Castaneda constructed communities of interest over time by considering organizational principles (e.g., hierarchy, user interest) [210]. To improve the protocol of Twitter in creating and maintaining a list of followees by the followers and let the followers avoid getting information overloaded, the author applied social network analysis and leveraged semantically enriched content (e.g., embedded links, URLs) to construct a triadic (user-object-user) networks. In these networks, followers and followees were implicitly connected through tweet contents. With the help of evolutionary communities of interest, the author proposed a model to identify connections between Twitter users and henceforth predict alternative information sources for followers.

2.3.1.4.4 Time-aware Features

Munasinghe and Ichise developed a new time index called 'Time-Score' for dynamic link prediction by incorporating temporality with the topological features and link strengths [211]. This time-aware measure is an extension of common neighbours with integrated time components. The assumption behind this method, as reported by the authors, is that the likelihoods of future links depends on the topological information (e.g., number of common neighbours, frequency of co-occurrences) and associated temporal duration (e.g., how long both actors have those common friendships). Munasinghe also developed another time-aware feature called 'T_Flow' which is an extension of the algorithm by Lichtenwalter et al. called 'PropFlow' [2]. The algorithm used link weights as transition probabilities and temporal random walk to compute the rate of information flow between two actors [212]. The rationale behind PropFlow algorithm is that if a pair of actor has higher transition probability, more information flow between them denoted the likelihood of their future association.

Munasinghe associated the duration of link activeness, as defined in the Time-Score measure, to compute the temporal aspects of the information flow between a pair of actors. Likewise, the authors in another study [213] developed a Time Path Index called 'Link-Score' by integrating temporal information with the path between actors and demonstrated that it performed better than the neighbourhood based Time-Score, despite its detrimental effect in regards to the execution time. To infer friendships in an online multiplayer game setting, Merritt et al. developed temporal statistical feature to capture the interaction patterns among players [214]. The authors considered periodicities, interaction volume, and the similarity in players' action within the online system. Different types of temporal features included in this study were: (i) pair autocorrelation (i.e., interaction continuity over a span of time), (ii) individual entropy calculated by considering individual's schedule of interaction and the context of the interaction (i.e., game type) and finally, (iii) pair frequency to denote the quantity of interactions between friends over non-friends. The authors demonstrated that interactions periodicity in combination with prosocial behaviours across these interactions are good indicators of inferring future friendships. Yao et al. developed three temporal metrics based on common neighbourhood within two hops by considering the size of the common neighbours as the representation of network's transitivity property [1]. These metrics included (i) time-varying link weight to reflect the topological variations over time and where the recent weight was preferred over the past and old link weights, (ii) change degree of common neighbours to reflect the stability of common neighbours where smaller-degree neighbours were considered more stable and finally, (iii) intimacy between common neighbours to determine the similarity between common neighbours. The final metric denoted that if two actors have common neighbours who are similar to each other, than these actors have higher likelihood of forming links in future.

Considering the periodic appearances and disappearances of links between actor in different SINs of a dynamic network as temporal network events, Soares and PrudêNcio developed new proximity measures for dynamic link predicton task [215]. The authors defined three different types of temporal events based on a specific activity between two actors from one SIN to its subsequent SINs. These are (i) conservative, an event that occurs when a link between two actors is preserved from one SIN to the next one, (ii) innovative, an event that occurs when a new link is created in a SIN which was not present in the previous one and finally, (iii) regressive, an event which is opposite to the innovative event (i.e., link disappears in a SIN which was present in the previous one). The authors proposed an event-based scoring mechanism considering these three events over time. The proximity score between a non-connected actor pair was computed by a rewarding scheme that was updated along time depending on these three events occurring between the actors including their neighbourhood.

Among other methods, Salem Narasimhan proposed three state-of-the-art supervised link prediction methods those are mutually exclusive [216].The author attempted to model the pattern of relationship formation between any two agents in a multi-agent dynamic network. These methods were called FELP (Feature Evolution based Link Prediction), HELP (History-based Eccentric Link Prediction) and MCLP (Minority Credit-based Link Prediction). In FELP, novel meta-feature vectors were constructed by considering a combination of time-augmented domain and topological attributes (e.g., AdamicAdar, degree mixing probability) of the network. In HELP, a complementary class of links between two connected components in dynamic networks was predicted using intuitive temporal network topological features (e.g., eccentric probability, group size). These links were considered complementary because the actors at both ends of these links were not reachable at the point of prediction (inter-group links). Finally, to address the inherent class skewness of the

supervised link prediction (i.e., positively and negatively labeled links), MCLP used single class learning on minority class examples only by extracting additional information from the network evolution. It was also capable of predicting inter-group links by considering temporal features like average seconds, average weekdays etc.

Time-aware features were also used for dynamic link prediction in bipartite networks. By considering an User-Object network, Liu and Deng developed a time-weighted network model for this purpose [217]. The authors considered both time attenuation (i.e., time scale to denote the recent vs old network events) to put mode weight on recent events and diversion delay (i.e., delay duration between two different network events) to weigh the link weights in their time-weighted network model.

2.3.1.4.5 Temporal Probabilistic

Type-awareness of relationships can provide additional information for different data mining tasks (e.g., expert recommendation). One such application is to model Advisor-advisee relationship from a co-authorship network which provides additional semantic information than simple coauthor relationship. For example, it may support identifying different research communities, how research topics are emerging and who are the influential figures in different research communities. These facts motivated Chi Wang et al. to adopt a probabilistic ranking method and propose a time-constrained probabilistic factor graph (TPFG) model that integrated intuitive features to predict dynamic links in collaboration networks [218]. On the other hand, Lakshmi and Bhavani developed a measure called Temporal Co-Occurrence Probability (TCOP) to predict links in homogeneous networks [219]. Their algorithm was considered as an extension of the Co-occurrence probability algorithm developed by Chao Wang et al. which is a probabilistic graphical model using higher order topological information [220]. However, Lakshmi and Bhavani incorporated temporal information with such graphical models to compute clique potentials. The authors

also extended the 'Time-Score' measure, mentioned above, to include the cliques where the corresponding actors belong. Further, they normalized the interaction quantity to assign high probability to the recent cliques and low to the old ones. Ahmed and Chen designed a new method of dynamic link prediction based on a random walk in temporal networks [221]. By considering a probabilistic random walk, the similarity scores between an actor and its neighbours were computed within a SIN around that actor to reduce the computational time. The proposed method Time Series Random Walk (TSRW), as called by the authors, exploited a global topological similarity metric called SimRank [222]. In conjunction with temporal information and SimRank algorithm, the authors computed a sequence of probabilistic random walk transition matrices for each SIN in a dynamic network and combined them together to generate the final one. With this final transition matrix, a damping factor was used to give more importance to the recent link information.

To predict the time (i.e., 'When') of the future link occurrences, given its features from the current network snapshot, Sajadmanesh et al. developed a probabilistic non-parametric approach named Non-Parametric Generalized Linear Model (NP-GLM) [223]. NP-GLM modeled the distribution of link creation time given the feature vectors and was capable of learning the underlying distribution of the data including the amount of contribution of each extracted feature.

2.3.1.5 Actor-oriented Measure

Most of the aforementioned dynamic link prediction methods measure the likelihood of future links connecting pairs of actors. These methods attempted to define the probability of future link between two actors, by considering topological information, network structure, and attributes shared by both actors, irrespective of any individual actor-level measure. Neighbourhood based methods are incompetent in differentiating two pairs of actors with similar common neighbourhood but having different likelihoods of link formation [219].

Further, according to Tylenda et al., link-based prediction strategies are based on the assumption that actors are interested in links irrespective of the interests of the actors themselves [89]. Therefore, the authors proposed actor-centric time-agonistic link prediction method where they demonstrated that the time-aware interaction information can support as an important feature to rank neighbouring actors, even beyond immediate direct neighbourhood, based on the likelihood of their future associations with the central actor. In addition, they also contributed with a novel actor-oriented approach to address the evaluation of link prediction. Other researchers also attempted to develop actor-oriented measures for link prediction in dynamic networks. By considering the temporal trends of actor popularity, T. Wang et al. developed a hypothesis where the likelihood of emerging links depends both on the structural importance and popularity (activeness) of actors in the links [81]. The authors proposed Popularity Based Structural Perturbation Method (PBSPM) that integrated both actors' popularity and observed network topology. Considering the community participation of actors over time, Adrian et al. extended the sociability index of actors in temporal networks to define weighted sociability index [224]. Intuitively, sociability index (SoI) [225] measures the number of times an actor changes its cluster along time. Among other studies, Tabourier et al. considered link prediciton in ego-networks and defined a list of features to capture different temporal information in regards to the time of interactions between an ego (i.e., actor) and its neighbours [79].

Semi-supervised label propagation-based learning algorithm, in conjunction with actor-oriented information, was also used as a dynamic link prediction strategy and hence acquired the name as link propagation [226]. To overcome the computational complexity, in regards to time and space requirements for network with thousands of actors, like many other actor-oriented strategies, the authors developed a fast and scalable method for link propagation. They used matrix Factorisation technique instead of widely used conjugate

graduate method widely used in semi-supervised link propagation. The authors also demonstrated a compact representation of the solution to the associated linear equation, and used a non-trivial combination of linear algebraic methods to solve the link prediction problem in dynamic networks.

2.3.1.6 Other Methods

Among other methods, Ahmed, Chen et al. presented a sampling based method of dynamic link prediciton where similarity scores, between a given actor and each of its neighbours, were computed by constructing a subgraph around that actor [227]. After combining the temporal network snapshots (i.e., SINs) into a weighted network, the actor-cantered subgraph was constructed by considering a chosen actor as the central one and then following a random walk in that weighted network from that actor. The underlying objective was to reduce the computational complexity by processing a smaller subnetwork rather than the whole network itself. The subgraph contained an optimal number of sampled paths to restrict the error of the estimated similarities within a given threshold.

To discover all the interaction patterns occurring at regular intervals in dynamic networks, Lahiri and Berger-Wolf proposed periodic subgraph mining algorithm [228], a concept borrowed from periodic pattern mining [229] and its other variants [230]. The authors also proposed a novel measure to rank periodically mined subgraph to determine its closeness of being perfectly periodic. Rahman and Hasan used a collection of induced subgraphs, known as Graphlet, in large scale graph analysis to predict links in dynamic networks [231]. By using the graphlet transition events (GTEs) in temporal network snapshots (SINs), the authors proposed GraTFEL (Graphlet Transition and Feature Extraction for Link Prediction) algorithm. It was defined as a novel learning method to obtain feature representation of actor-pairs for the purpose of predicting the likelihood of future links among them. Based on the

assumption that subgraphs distribution in complex networks is statistically stable and typical even during significant structural changes [232], Juszczyszyn, Musial,and Budka attempted to characterize the network structural changes by statistical data, extracted from the evolutionary subgraphs [233]. Subsequently, the authors also proposed a dynamic link prediction method considering triads discovery and their transition measurement during network evolution. Thus, the link prediction strategy TTM-predictor (Triad Transition Metrix predictor) used the information related to temporal transition of triads found in dynamic networks.

2.3.2 Dynamic Link Prediction in Heterogeneous Networks

Most of the aforementioned link prediction strategies are designed for homogenous networks; however, many important real-world networks are inherently heterogeneous. These include bibliographic network (e.g., author-keyword), biological networks (e.g., gene-disease) or recommendation network (user-item). Due to the structural complexity and actor/link heterogeneity, link prediction in heterogeneous dynamic networks is challenging. Aggarwal et al. proposed a dynamic graph-clustering-based approach where both macro and micro decisions supported the link inference process in dynamic content-rich heterogeneous networks [234]. In this approach fine-grained clusters were generated based on structural similarity and constantly maintained in SINs. Structural behaviour of this dynamic summarization approach that divided the network into densely-connected regions supported the macro decision process whereas both structural and attribute information were used in the micro-decision process. Conversely, instead of considering structural similarity, meta-path based similarity was introduced by Sun et al. to define similar objects in hetergeneous network to aiding the dynamic link prediction [235]. The meta-path was defined as a sequence of relations existed between different types of objects or alternatively, the structural paths at the meta-level. Considering this meta-path framework, Li et al. also presented a

novel similarity measure called 'PathSim' heterogeneous military/combat networks [236]. This measure was capable of finding out the peer objects in the network (e.g., finding the viewers having similar movie choice or similar ratings).

Probabilistic models were also exploited for link prediction purpose in heterogeneous networks. These models attempted to optimize a target function to develop a model composed of different parameters that best-fit the corresponding network. By modeling the influence (probability) propagation among heterogeneous relationships, Y. Yang et al. developed a novel probabilistic method, known as Multi Relational Influence Propagation (MRIP), for dynamic link prediciton task in heterogeneous networks [237]. MRIP was also capable of capturing the correlation between different types of links in heterogeneous networks. The authors also introduced temporal link predictors in heterogeneous networks by considering time-augmented variants of classical link predictors (e.g., CommonNeighbours[1], AdamicAdar). Existing literatures further include several other probabilistic methods [238,239].

2.3.3 Challenges and Limitations in Dynamic Link Prediction Strategy

In case of link prediction in static networks, local similarity indices are constructed using neighbourhood-related topological information whereas global similarity indices use the whole network to extract topological information and compute similarity/proximity between actor pairs. Fundamentally, many real-world networks are dynamic in nature. With the temporal evolution of networks, actors simultaneously experience both micro and mesoscopic network structural changes forcing the topological properties to change over time. This phenomenon made the traditional topological similarity indices incompetent in dynamic link prediction.

[1] Appendix A

Sajadmanesh et al. pointed out three important challenges in dynamic link prediciton that cannot be solve trivially: (i) due to the indispensable integration of time component in network evolution, the formulation of dynamic link prediction is quite complex. Dynamic link prediction includes a 'when' query in addition to 'which' query in traditional binary link prediction task (ii) information deficiency in regards to the creation time of non-existent links whereas only the creation time of the existent links can be known and finally (iii) inferring a temporal probability distribution for each actor-pair in the network by considering their available features, and answer time-related queries about the link creation time, can be inefficient as the underlying distribution of the links' creation time is unknown and a priori distribution considered may become unrepresentative or limited representation of the reality [223].

From the aforementioned different categories of dynamic link prediction strategies, it is evident that the temporal pattern of dynamic networks imposes the first constraints in analysing them. Many dynamic link prediction methodologies exploited random walk methods over temporal networks or some sorts of aggregation techniques to transform the temporal networks into a weighted static network and compute similarity between actors from that static version. These methods appeared to ignore the temporal components inherent to the dynamic networks and temporal information was not used in the principled manner of weighted static graph construction. It is also evident that some of the above-mentioned methods are subject to their inherent limitations. For example, many methodologies used time series to model the temporal pattern of dynamic networks and then exploited link occurrence frequencies, topological, network structural or heuristics based measures in each network snapshot (i.e., SIN) to compute the likelihood of future links between actors. Time series-based methods, considering only link occurrence frequency, can only allow us to predict repetitive links or links observed in any SIN. These methods are incompetent of predicting

unobserved links. On the other hand, time series-based methods, considering topological or structural metrics, can predict the unobserved links but not the repetitive links. Further, some of these methods included the time series forecasting method (e.g., ARIMA) to predict the future values of topological changes. This exercise can be counterproductive since prediction is performed using predicted and unrealistic values.

Most of the dynamic link prediction methods predominantly calculate or estimate likelihood score for a non-existent link which quantifies emergence possibility of that link in future. This likelihood score is calculated by different topological similarity-based algorithms or probabilistic methods. The problem with the similarity-based algorithms is that different similarity measures denote different likelihood scores for the same link. For example, common neighbourhood based topological information may denote different likelihood score for two links despite both having actors with same number of common neighbours. On the other hand, probabilistic models involve the prior definition of distributions of link occurrences which is problematic for temporal networks. Further, the probabilistic model is only suitable for small networks with few hundred actors (e.g., ERGM). The underlying reason behind this is the inclusion of maximum likelihood methods. According to Y. He et al., the maximum liklihood calculation in probabilistic methods, by using surmised information, stringent rules and numerous parameters, is time consuming [178]. Similarly, matrix or tensor-based methods are not feasible for real-time link predictions in large networks due to the computational complexity and time requirements [93]. Further, in most machine learning based dynamic link prediction strategies, the corresponding algorithm need to be optimized. It is also evident that most of the methods, described above, considered only dyadic evolutionary information instead of actor-level evolution experienced by individual actors as a result of temporal changes in dynamic networks. Despite some existing methods defined actor-oriented features in each temporal network snapshots and combined them

together to denote similarity between them, however, these methods are incompatible of considering temporal evolution similarity between actors.

Dynamic link prediction task needs to have both spatial and temporal consistencies [149]. The first property signifies the neighbourhood based topological information and the second enforces the smooth temporal network evolution [73]. In temporal link prediction, it requires to unify both these spatial and temporal factors [149]. To address this issue, firstly, it is imperative to identify the optimal time scale to discretise or sample streaming interactions among dynamic actors to generate temporal short interval networks (SIN). This time scale determines the length of the window or the duration of the interval for each SIN. Too coarse or fine-grained/smooth temporal window size will either generate topological error that lead to erroneous results or sampling error that causes to obscure important information [240].

2.4 Conclusion

This chapter provided extensive literature reviews on both research objectives (i.e., optimal window length detection and dynamic link prediction) including the limitations prevalent in existing methodologies addressing both research issues. The research gaps and challenges, those provide the rationales behind this research work, were presented after each literature review section. For example, in most cases, the existing methodologies of determining optimal window length to sample dynamic networks are dependent on either global network metrics or actor attributes. These facts left them ignoring the local-level network metrics, attributed to individual actors - the principal constituents of a network. These methodologies are also considered as incompetent to work with large networks. On the other hand, the extant link prediction methodologies in dynamic networks ignored not only the optimal sampling of the corresponding networks but also actor-level evolutionary measures, in case of absence of any actor attributes, in predicting future links among them. In addition, some of these

methods either consider probability distribution of future links which is hard to presume prior to the actual prediction task (e.g., Exponential Random Graph Model) and some methods are computationally intensive in nature (e.g., Non-negative Matrix Factorisation). Considering these, this research not only develops a novel algorithm to optimally sample a dynamic network but also proposes some validation criteria to validate the optimality of the window size. Once optimal sampling is achieved, this study then develops different features by considering actor-level evolution similarity, the concept yet to be explored by the research community. These features also consider different types of actor-level evolutions (i.e., dynamicity) and can work in any size of the network in case any actor-level attributes are not presented. The next chapter provides the theoretical conception of an algorithm to detect the optimal sampling window size to discretise dynamic network including some evaluative measures to validate the optimality.

Chapter 3

Optimal Time Scale (Window) in Dynamic Networks

<u>The following article was published based on the theoretical contents of this chapter</u>

1. Uddin, S., **Choudhury, N.**, Farhad, S. M., & Rahman, M. T. (2017). The optimal window size for analysing longitudinal networks. *Scientific reports, 7*(1), 13389. doi:10.1038/s41598-017-13640-5. (**Impact Factor 4.259**)

3.1 Introduction

The design of any longitudinal study requires the selection of a time scale or window size (i.e., the amount of time that elapses between occasions of consecutive measurements) either before or after the data collection. This window size also denotes the bin size and alternatively known as ‘resolution’, ‘aggregation granularity’, ‘temporal sampling’, ‘sliding window’, or ‘temporal window’, used in dynamic data collection process. Although, in most longitudinal studies, the system under consideration naturally suggests the size of such a temporal resolution [241]; however, the routine practice, in regards to the choice of time scale or temporal window size, is arbitrary and followed according to the convenience of the data representation and analysis. All time-stamped network activities within each window are generally aggregated for conducting a longitudinal network data analysis. This practice is common in clinical cohort, epidemiological and psychological studies where researchers frequently use a chosen time scale without any rational or deterministic explanation [242]. Simultaneously, to date, designing longitudinal network study overlooked actor-oriented approach that can be adopted to precisely determine the appropriate time scale [243]. Fish and Caceres used link prediction technique, the underlying growth mechanism of networks, to demonstrate that link prediction performs well on aggregation sequences close to the ‘ground truth’ [125]. The term ground truth here denotes maximizing the number of network snapshots for a given dynamic network. Further, in a longitudinal study on smoking behaviour, Collins and Graham [118] found a positive correlation between smoking and peer smoking; however, the strength of the relationship decreased as the window size increased.

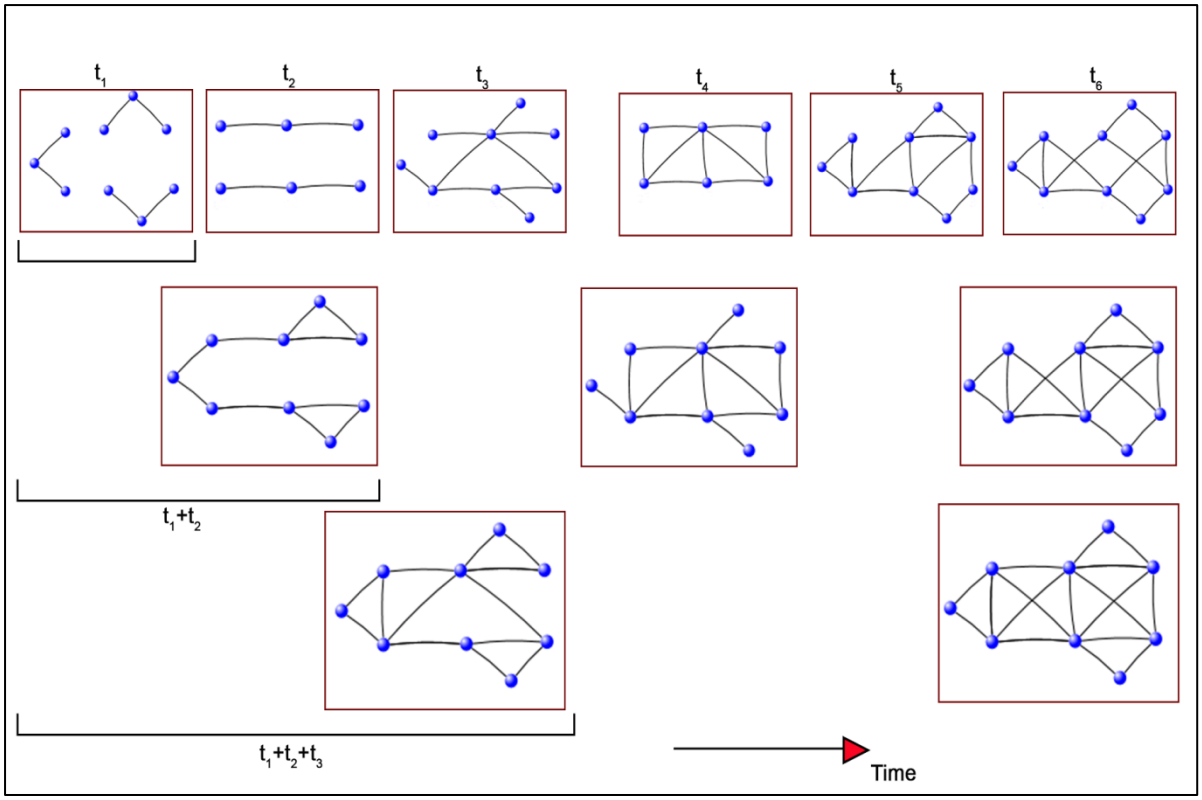


Figure 3.1: An abstract visual representation of a dynamic network that evolved over six timestamps $(t_1, t_2, t_3, t_4, t_5, and\ t_6)$ to demonstrate how a given dynamic network can be described as a collection of multiple network snapshots (i.e., Short Interval Networks, SINs). An individual SIN denotes the length of the duration for which streaming links/interactions were aggregated into the corresponding snapshot. For example, in the top row, six individual SINs donated a collection of links formed at the corresponding timestamps and each timestamp can be of any duration (e.g., an hour, day, or month). In the second row, each aggregation window was increased twice (i.e., $t_1 + t_2$) leading the dynamic network to have three SINs instead of six. Simultaneously, the network structure changed among the actors. Finally, in the bottom row, the aggregation window represents three timestamp units from the top row combined (i.e., $t_1 + t_2 + t_3$) denoting a larger duration for each SIN. Similarly, due to the altered duration, the network structure in each SIN was changed.

Conversely, in case of analysing human interaction patterns over time, using calendric time scale is considered convenient and supports the underlying social interaction system; however, practicing similar temporal scale choice in case of animal interaction pattern would be unrealistic since periodic animal behaviour does not rely on weekdays or weekend. Therefore, determining the optimal or appropriate sampling time scale is imperative to effectively analyse dynamic networks whereas a poorly chosen window size can cause researchers to make inaccurate conclusions about the variables or hypotheses being studied [100].

In dynamic network abstraction, where a dynamic network is represented by a time series of network snapshots or short interval networks (SIN), the duration of each SIN represents the time scale or temporal window by which the dynamic network was sampled. Alternatively, it is called temporal network snapshot. In transition from streaming interactions to dynamic network abstraction, data may already come as a series of aggregated snapshots in some cases, or in other instances, from a given stream of interactions in time; we may have to aggregate into different SINs. Irrespective of these two design considerations, the level of aggregation of the temporal stream has a great impact on the observed patterns in the corresponding dynamic network, and the inferential process made on the network and its processes [244-246,97]. This phenomenon is visually presented in (Figure 3.1). In this figure, a dynamic network is presented that evolves over six individual timestamps. At each time window the network structure (randomly generated) varies from the periodic patterns observable in others. In the second row, the duration of link aggregation was doubled (i.e., consists of two individual time windows together) resulting the total number of SINs to three from six in the upper row. In the final or third row, the aggregation window was tripled to generate only two SINs. In each row, as the aggregation window or time scale to accumulate links over time varies, so do the network structure in each window including the number of

SINs. Therefore, it is understood that the selection of time scale of duration of a window can greatly impact the study and analysis of dynamic networks. When assigning a time span to a window of a dynamic network, the underlying principle is that actors of the network must have sufficient time to initiate network processes such as the formation and dissolution of ties.

Actors in a dynamic network usually exhibit different rates participation in different network activities (e.g., for the formation of new ties or the dissolution of existing ties). In any given SIN with specified time length, some actors may show higher network activities than others while network activities by some actors may be under-represented. Further, an actor may demonstrate a high volume of network activity at the very beginning time of a particular window or another actor may create all its new ties at the end stage of the immediate window. Furthermore, in relation to a given window size, an actor might reveal all its network activities (e.g., create 15 ties and delete 10 ties) within a longitudinal network in only one window while the other might engage in the same activities in five windows. This would significantly affect the involvement of the actors contributing towards the evolution of the underlying dynamic network. Considering these phenomenon, the analysis of a given network could produce different results for the actor-level social network measures (e.g., network centrality) with the consideration of different time scale sizes. An appropriate or optimal time scale should reduce the differences in network activities demonstrated by the group of dynamic actors. This is because these dynamic actors show somehow similar or different level of network activities during the evolution of the underlying network. Therefore, in this thesis, a novel approach is proposed to determine the optimal or appropriate sampling size of the temporal window or time scale to analyse dynamic network.

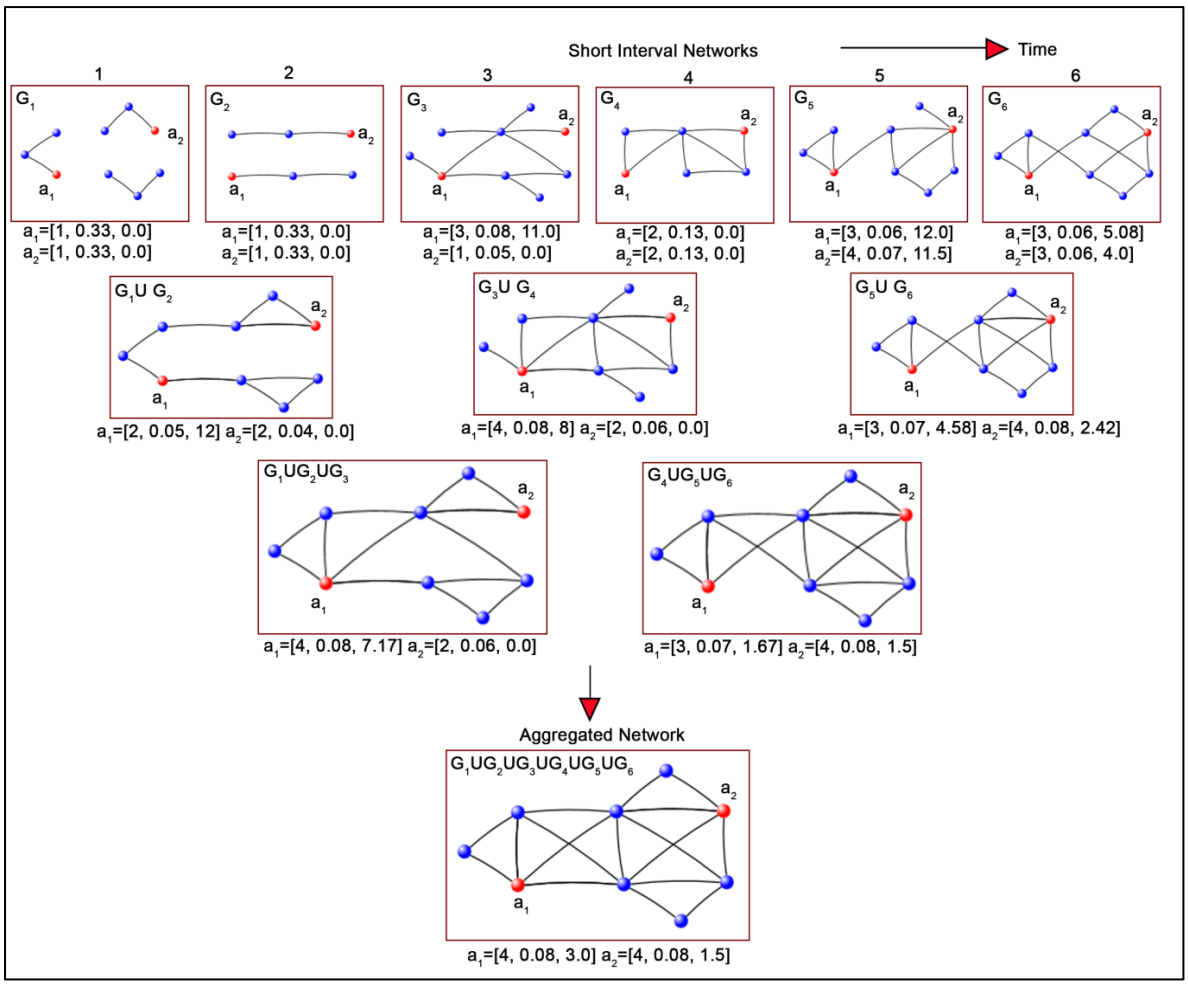


Figure 3.2: An abstract visualization of dynamic networks to demonstrate how actor-oriented network structure changes over time. The dynamic network consists of a series of evolutionary network snapshots at different discrete timestamps ($t = 1, 2, 3, 4, 5, 6$) (e.g., days). The top row represents a time series of six network snapshots $G_1, G_2 \ldots, G_6$, known as short-interval networks (SINs). The second row represents the aggregated networks at timestamps $t = 2, 4, 6$ where an individual SIN denotes ($G_{t-1} \cup G_t$). The third row represents the aggregated networks at timestamps $t = 3, 6$ where an individual SIN denotes ($G_{t-2} \cup G_{t-1} \cup G_t$). The bottom row represents an aggregated network (i.e., union of all SINs). Each network snapshot is accompanied by three normalized centrality measures (i.e., degree, closeness, and betweenness), incident to actor a_1 and a_2 at different timestamps, both in individual SINs and in aggregated networks.

The simple and novel approach to determine the sampling duration of dynamic networks is hypothesized on the minimum variance analysis of dynamicities demonstrated by the network components, the actors, participating in a given dynamic network. The minimum variance will ensure that the suggested time scale will neither be too coarse for some actors that reveal high rates of network activities nor be too small for some other actors that reveal slow rates of network activities. This approach will be applied to different dynamic network datasets to empirically determine their appropriate time scale or duration of network activity aggregation from different actor-level perspectives. In this thesis, the optimality of the identified temporal window sizes/time scales is further evaluated by different approaches including time series analysis.

3.2 Actor-oriented Positional Evolution

As mentioned earlier, a dynamic network comprises different network snapshots observed at different points in time known as short interval networks (SINs). An accumulation of these SINs into a bigger network is known as an aggregated network. In two consecutive SINs, the network involvement of an actor can be changed in two different ways. First, an actor may change its neighbourhood connectivity within the two SINs and thus change the network position. It can be captured by different network measures such as degree, closeness and betweenness centrality. Second, an actor may change its presence; for example, an actor may be present in the tth SIN, leave the network at the $(t+1)^{th}$ interval and re-join the network at $(t+2)^{th}$ interval. In (Figure 3.2), these phenomena are demonstrated visually. Similar to the (Figure 3.1), this figure represents a given dynamic network by using six SINs (i.e., $G_1, G_2, G_3, G_4, G_5, G_6$) in six individual timestamps (e.g., days) in the top row, three SINs in the middle where the duration of time scale was doubled (i.e., two days) and two SINs in the third row where time scale duration was tripled (i.e., three days). In addition, this figure

also present an aggregated network that combined all SINs together (i.e., $G_1 \cup G_2 \cup G_3 \cup G_4 \cup G_5 \cup G_6$) to represent the static version of this dynamic network. It is observable from this figure that due to the consideration of different time scale to aggregate links in dynamic networks, network positions and participations of actors and links vary. These positional evolutions are observable through the alterations of network centrality measures. In this figure, three different network centrality measures (i.e., degree, closeness and betweenness) of two actors (i.e., a_1 and a_2) are presented in all SINs of different temporal window sizes. It is evident that due to network evolutionary changes in different snapshots, in regards to the presence/absence of actors and formation/dissolution of links, the centrality measures of these two actors change accordingly. Thus, the evolution of an actor in a given longitudinal social network has two components: (i) positional; and (ii) participation [247]. Positional evolution denotes changes of network positions of actors in different network snapshots of a dynamic network relative to their positions in the aggregated network. On the other hand, participation changes exemplify the changing network participation of actors in any two consecutive temporal network snapshots. For example, an actor may participate in the $(t-1)^{th}$ network snapshot but disappear in the tth snapshot, or alternatively, it may appear to participate in the tth network snapshot and disappear in the subsequent $(t+1)^{th}$ snapshot. These types of actor participatory transitions in consecutive network snapshots contribute to the participation dynamicity in the dynamic network

According to social network analysis, a given dynamic network needs to be analysed in regards to the temporal aggregation of links among its actors [248,249]. Quantification of different aspects of the dynamicity in this network depends on both static and dynamic topology of social network analysis [250]. Due to the temporal nature of SINs, dynamic topology is exercised on temporal network snapshots, whereas; the static topology is applied to the aggregated network. SINs may have different durations that principally depend on the

underlying longitudinal network data. The accumulation of all SINs creates an aggregated network which is a big cross-sectional network (Figure 3.2).The rationale behind the concept of positional dynamicity and the reasoning behind using this concept to identify optimal sampling window size can further be perceived better from (Figure 3.2). In this figure, the network snapshots in the top three rows represent SINs of different durations (i.e., one, two and three days) and the bigger network in the bottom row denotes the aggregated cross sectional networks. It is evident from the varying sized SINs that changes in network positions or importance of actors depend both temporally and sampling resolution of SINs. Depending on temporal granularity to define SINs, an actor may either participate in network activities by forming links or may get disappeared by severing all links with its neighbours. Concurrently, the rate and volume of neighbourhood changes by actors fluctuate considering time and window size, which can be observed for two actors (i.e., a_1 and a_2) in this figure. The concept of *positional dynamicity* was developed to quantify these temporal positional variations considering both dynamic and static social network topology. Consequently, our objective in this study is to define the optimal time scale duration by considering the uniformity of positional dynamicities demonstrated by actors over time.

The positional dynamicity of a longitudinal network represents changes in network positions of its member actors across different SINs compared to their network positions in an aggregated network [251]. SINs may have different durations that mainly depend on the underlying longitudinal network data. Given a dynamic network G_T that can be observed at $T = t_1, t_2, \dots\ t_m$ different equal-time intervals where $t_m > t_{m-1} > \dots\dots\ t_2 > t_1$, the list of SINs can be defined as $G_{t_1}, G_{t_2}, \dots\dots\ G_{t_{m-1}}, G_{t_m}$. Considering the aggregated network has N actors and m SINs have $n_1, n_2, \dots n_m$ actors where, $|n_1 \cup n_2 \cup n_3 \cup \dots\dots \cup n_{m-1} \cup n_m| = N$, respectively, an equation to quantify the positional dynamicity of a member actor in a dynamic network was proposed by Uddin et al. [251]:

$$PoD_i = \frac{\sum_{t=1}^{m}\left[\frac{|NP_{AN}^i - NP_{SIN(t)}^i|}{|NP_{AN}^i + NP_{SIN(t)}^i|} * M(i,t)\right]}{m} \times 100\% \quad \dots\dots\dots\dots \ (3.1)$$

In Equation (1), PoD_i denotes the positional dynamicity demonstrated by the ith actor, NP_{AN}^i indicates the network position measure, calculated by using any actor-level social network measure individually (e.g., closeness centrality) or as a combinations of multiple measures (i.e., $degree + closeness + betweenness$) for the ith actor in the aggregated network, $NP_{SIN(t)}^i$ denotes the network position measure based on the same social network measure as described above in the tth SIN for the ith actor, $M(i,t)$ represents the participation details (i.e., whether an actor is present or absent in the corresponding SIN) of all actors in all SINs and m indicates the number of SINs in the longitudinal social network. The value of $M(i,t)$ is one if the ith actor is present in t^{th} SIN or zero (0) otherwise. Subsequently, an equation to determine the positional dynamicity of a particular SIN was also proposed by the authors as follows:

$$PoD_{SIN(t)} = \frac{\sum_{i=1}^{N}\left[\frac{|NP_{AN}^i - NP_{SIN(t)}^i|}{|NP_{AN}^i + NP_{SIN(t)}^i|} * M(i,t)\right]}{n_t} \times 100\% \quad \dots\dots\dots\dots (3.2)$$

where $PoD_{SIN(t)}$ denotes the positional dynamicity of the tth SIN, n_t denotes the number of actors in the tth short-interval network and N denotes the total number of actors in the aggregated network. In Equation (3.1), the value of NP_{AN}^i for an actor depends on the underlying social network analysis (SNA) measure used to capture the network position of that actor. The value of NP_{AN}^i will differ depending on the SNA measure considered (e.g., degree centrality or closeness centrality). Similarly, the value of $NP_{SIN(t)}^i$ depends on the SNA measure is used to quantify the network position of the actors. This, in turn, implies that the

value of PoD_i (i.e., in the right hand side of Equation 3.1), depends on the SNA measure used to capture the network position of actors. The SNA measure, developed for this study, is a function of degree, closeness and betweenness centrality of an actor in an individual SIN. The rationale behind using a composite measure consists of these three centrality measures are as follows: firstly, they are the simplest, well-defined, and can successfully quantify an actor's connectivity, position, communication dynamics, influence and broadcasting capabilities, and importance in a network. Secondly, these measures are correlated. For example, an actor with high betweenness and low closeness centrality can monopolise links from a small number of actors to many others. Likewise, high degree with low closeness centrality denotes that the actor is embedded in cluster far from the rest of the network [252]. The measure is defined as follows:

$$a_i(g_t) = C_a^{Deg}(g_t) + C_a^{Cls}(g_t) + C_a^{Betwn}(g_t) \quad \dots\dots (3.3)$$

where $a_i(g_t)$ denotes SNA measure of actor i in a particular short-interval network (SIN) g at time t, and expressed through a composite centrality measure. $C_i^{Deg}(g_t)$ denotes degree centrality, C_i^{Cls} denotes closeness centrality and C_i^{Betwn} betweenness centrality of actor i in a SIN g at time t. The measure $a_i(g_t)$ quantifies both NP_{AN}^i and $NP_{SIN(t)}^i$. Thus, aggregation of degree, closeness and betweenness centrality measures of actors are considered to quantify the positions of actors both in SIN and aggregated cross-sectional networks. Definitions of these three centrality measures are described below:

3.2.1 Degree Centrality

Degree is the simplest centrality measure among all that counts the number of direct neighbours. It refers to the number of ties an actor has to other actors or the number of connections an actor has. In directed network, the direction of the connections is considered.

The in-bound connections are considered as in-degree and out-bound connections are considered as out-degree. Degree centrality assigns importance score based purely on the number of links held by an actor. Degree centrality is useful for identifying the most connected and popular actor who can quickly connect with wider network with a possibility of holding most information. The degree centrality is normalized by dividing the number of direct neighbours by the maximum number of actors $(n-1)$. The degree centrality C_a^{Deg} of actor a is defined as:

$$C_a^{Deg} = \frac{\sum_{b:b\neq a} p_{ab}}{n-1} \quad \dots\dots (3.4)$$

where $p_{ab} = 1$ if there is a link between actor a and b and n is the total number of actors.

3.2.2 Closeness Centrality

The closeness centrality measures the degree to which an individual actor is near all other actors in a network. It is defined by the inverse of the length of the geodesic distance (i.e., shortest path) to/from all other actors. Closeness centrality measures the momentum of influence by an actor or finds the best-placed actors to influence the entire network most quickly. Sometimes in a networked system, it is better to stay between others or in the middle rather than staying far from the rest. In this case, closeness centrality is an important measure to denote the 'broadcaster' actors. The normalized closeness centrality of an actor is represented by:

$$C_a^{Cls} = \frac{n-1}{\sum_{b=1}^{n-1} d(a,b)} \quad \dots\dots (3.5)$$

where $d(a, b)$ denotes the geodesic path between actor a and b and n denotes the total number of actor.

3.2.3 Betweenness Centrality

Betweenness is a measure of the extent to which an actor is connected to other actors those are not connected to each other. It is a measurement of actor in the network to what extent it serves as a bridge. By definition, an actor's betweenness centrality measures the number of times it lies on the shortest path between other actors. This measure is useful to analyse communication dynamics within a network as high betweenness denotes authoritative or controlling power between disparate actors or clusters within the network. It can also denote peripheral actors among clusters. The normalized betweenness centrality of an actor a is defined as:

$$C_a^B = \frac{\sum_{x \neq y} d_{xy}(a)}{\sum_{x \neq y} d_{xy}}$$

$$C_a^{Betwn} = \frac{C_a^B}{[\frac{(n-1)(n-2)}{2}]} \dots \dots (3.6)$$

where C_a^B denotes the un-normalized betweenness centrality of an actor a, $d_{xy}(a)$ denotes the number of shortest path going through actor a, d_{xy} denotes total number of shortest path and the denominator in the normalized version C_a^{Betwn} denotes the total number or pairs of actors excluding the actor itself.

3.3 Proposed Algorithm

This section describes the study's approach to determine the window size of the underlying longitudinal network.

A two-step procedure was followed to determine the window size of a longitudinal network. In Figure (3.3), a visual demonstration of the algorithm is provided. In this figure, for the sake of clarity, a dynamic network was selected that was sampled using different time scales of days (e.g., 1, 2, 3 …, 7 days).

3.3.1 Step One

Step One used Equation (3.1) to quantify actors' positional dynamicity values for a longitudinal network dataset and considered different lengths (e.g., one day to seven days in Figure 3.3) to define the SINs. The centrality measures defined in equation (3.3) were used to calculate actors' network positions in each SIN and in the aggregated network.

3.3.2 Step Two

Step two compared different sets of actors' dynamicity values. A consideration of different lengths for SINs led these different sets of actors' dynamicity values. According to the dynamic network presented in Figure (3.3), there were seven different sets of actors' dynamicity values since seven different values (i.e., one day, two days, three days … seven days) were considered to define the SINs. A variance comparison approach was used to compare different sets of actors' dynamicity values. For a group of numbers (e.g., a set of actors' dynamicity values), the variance measured how far all these numbers spread out from their mean or average [253]. A higher variance among a set of numbers indicated a wide spread of numbers around the mean.

The corresponding length (used to define the SIN) was considered to be the right window size, if it produces the lowest variance for the dynamicity values of all actors. Thus:

Window size = S; for which **Variance** $\left(AD_i^S\right)$ is the minimum ……(3.7)

Actor \ Day	1	2	3	4	5	6	7
1	0.0137	0.0167	0.0191	0.0238	0.0361	0.0264	0.0227
2	0.0246	0.0458	0.0581	0.0665	0.0655	0.0893	0.0662
3	0.1060	0.1598	0.1836	0.2204	0.2330	0.2143	0.3049
4	0.0063	0.0086	0.0125	0.0182	0.0253	0.0252	0.0217
5	0.0100	0.0195	0.0256	0.0349	0.0293	0.0245	0.0240
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...
1893	0.0220	0.0296	0.0260	0.0288	0.0406	0.0319	0.0308
1894	0.0035	0.0041	0.0058	0.0073	0.0094	0.0110	0.0122
1895	0.1680	0.2246	0.2495	0.2766	0.3211	0.3220	0.3157
1896	0.0038	0.0069	0.0096	0.0121	0.0143	0.0145	0.0168
1897	0.0520	0.0665	0.0841	0.0833	0.0920	0.0887	0.1102
1898	0.0061	0.0136	0.0210	0.0294	0.0325	0.0323	0.0562
1899	0.0057	0.0097	0.0119	0.0159	0.0165	0.0118	0.0244
---	---	---	---	---	---	---	---
Variance	0.0026	0.0049	0.0061	0.0077	0.0097	0.0094	0.0120

Actors

Different Window-Size

Dynamicity of actor 1897 for the 7 day window-size

Figure 3.3: An illustration of the changes in variances of positional dynamicity values of actors where three centrality measures (degree, betweenness, and closeness) were considered to quantify an actors' position in Short Interval Networks (SINs). The time scale duration of each SIN may vary from one day to seven days. The blue shaded top row denotes a different duration/window size/time scale (in days) that is utilized to split the entire network dataset for generating SINs. The light-yellow shaded left most column denotes actors in the network and the values in the table denotes actor dynamicity values.

Where, AD_i^S is the *Actor Dynamicity* (i.e., positional dynamicity) for the ith actor (where, $i = 1, 2, ...n$) given that a length of S has been considered in defining SINs and there are n actors in the longitudinal network.

Algorithm:
Input: List of time labeled edges
$(e_1,t_1),(e_2,t_2),(e_3,t_3),...(e_i,t_i)$
List of actors A
Maximum window length $\omega_{\max}$
Output: Optimal window length $\omega_{optimal}$

declare $N \leftarrow |A|$ //total number of actors
$\omega_{optimal} \leftarrow 1$
for $i \leftarrow 1$ to $\omega_{\max}$ **do**
  Declare $POD_i[N]$//positional dynamicity array
  Compute the time-series of short-interval networks $S_i(\omega)$
  for $n \leftarrow 1$ to N **do**
    $POD_i[N] \leftarrow$ Compute Actors' Positional Dynamicities
  end for
  $V_i \leftarrow$ Compute Variance of Actor Dynamicities
  $Variance\ (POD_i[N])$
  if $V_i < V_{i-1}$ **then**
    $\omega_{optimal} \leftarrow i$
  end if
end for
Output $\omega_{optimal}$

Figure 3.4: Algorithm to compute optimal temporal sliding window size to analyse a given longitudinal network with time labelled edges.

In (Figure 3.4), the algorithm to determine the optimal window length is presented. The lowest variance for the dynamicity values indicates that actors revealed the least difference in their positional dynamicity values over the time in the underlying longitudinal network, which is the underlying principle of this research as discussed in the first section of this article. If actors' positions in different SINs and in the aggregated network have been quantified by three centrality measure together, for example, then the lowest variance exemplifies that actors showed a minimum difference among themselves in terms of the variability in their degree centrality values over the time in the underlying dynamic network. Therefore, for a suggested window size the lowest variance will confirm that – (i) an active actor will not exhibit a large number of network activities; and (ii) an actor showing low rate of network activity will display a minimum volume of network activities.

3.4 Evaluation

This section describes some evaluation approaches, based on time series analysis and supervised learning methods that were used to validate the optimality of the identified window sizes in dynamic networks. Three different methods were used to validate the effectiveness of the proposed approach of this study in identifying the optimal window length to sample a given longitudinal network: (i) Auto Regressive Integrated Moving Average (ARIMA) model; (ii) Time series anomaly detection method; and (iii) Unsupervised clustering method known as K-means clustering.

3.4.1 ARIMA Model

Understanding the dynamics of time-dependent complex social networks using time series of network variables has drawn attention of network science researchers. Time series analysis has broadly adopted in network analysis methods such as link predictions that model the underlying growth pattern of social networks. In time series analysis, past observations of a

time variable can be analysed to develop a model predicts the future values of the variable. This study considers the ARIMA univariate time series method to model the temporal dynamicity of longitudinal networks. Under the ARIMA model, the future values of a variable are determined using a linear combination of past values and past errors. The model can be expressed as follows:

$$y_t = \theta_0 + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \dots\dots + \varphi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \theta_q \varepsilon_{t-q} \quad \dots\dots (3.8)$$

where, y_t = actual value, ε_t = random error at time t, φ_i and θ_j are the coefficients, and p and q are the integers for the *auto regressive (AR)* and *moving average (MA)* polynomials. *ARIMA (p, d, q)* represents an ARIMA model where p equals the number of autoregressive terms, q equals the number of lagged forecast errors in the prediction equation and d equals the number of non-seasonal differences needed for stationarity.

Stationarity is important in developing time series model since the properties of stationary time series depend only on white noises which is Gaussian (random) in nature. Trends and seasonality, affecting the value of time series at different times, are two of the most important contributors towards time series stationarity. One way to make a time series stationary is known as differencing that computes the differences between consecutive observations. It helps stabilising the mean of a time series by eliminating trend and seasonality and thus the changes in the level of a time series. In ARIMA model, the parameter d represents the level of differencing required to make a time series stationary. The higher the order, higher trends or seasonality is present in the time series. Similarly, parameters p and q represent the lag order for both autoregressive and moving average process, where these lag orders determine the level of auto-correlation between time series values and associated error measures. For example, lag order p denotes that the time series

value v_t at time t is correlated with the value v_{t-p} at time $(t-p)$. Interested readers are directed to the work by Hyndman [254] for detail description in ARIMA. Considering this, $ARIMA\ (0,0,0)$ is a white noise time series with no predictive pattern and trend components, requires no differencing, and demonstrate zero correlation among time series values and associated error terms.

Considering the temporal SINs in a dynamic network as random networks, the total actor-level positional dynamicity represented by an individual SIN has to be random in nature without any correlation to the positional dynamicity generated by any previous SIN(s). Therefore, the time series of positional dynamicity demonstrated by the temporal sequence of SINs should be independent of any trend or seasonal components and random in nature. In order to prove this fact, I generated a univariate time series, with the help of equation (3.2), which denotes the positional dynamicity of individual SIN in a dynamic network. For different window lengths considered in this research (i.e., one-six days, weekly, fortnightly and monthly), different time series of SIN's positional dynamicity were constructed. For all these series, the best fit ARIMA model was determined. Since the underlying concept is that the dynamicity distribution across SINs of any length will be free from trends and patterns (i.e., stationary series), the series which is be close to $ARIMA\ (0,0,0)$ will denote the best candidate for optimal SIN length.

3.4.2 Time Series Anomaly Detection

A time series is defined as a collection of observations of data items collected through repeated measurements temporally. It can be decomposed into three components: (i) long term variations or trend, (ii) systematic or calendar related movements or seasonal and (iii) out-of-control, irregular, and short term fluctuations known as residuals. Seasonality generally consists of regular, periodic, repetitive and predictable pattern whereas the trend

component is known as secular variation that denotes long-term non-periodic variations. In time series anomaly detection, it is imperative to identify the trend component(s) in time series which may introduce artificial anomalies within the time series [255]. In time series analysis, anomalies are denoted by point-in-time irregular data points. These anomalous data points can be global or local, and positive or negative. Global anomalies extend above or below expected seasonality; and local anomalies appear inside seasonal patterns and are always masked making them hard to detect. The anomalies can further be categorized as positive and negative anomalies. Positive anomalies represent point-in-time increase of observed values (e.g., number of tweets during a famous gaming tournament) and negative anomalies represent point-in-time decrease of observed values (e.g., number of service request to a server during server malfunctioning). In this step, besides global and local anomalies, this study adopted a novel anomaly detection technique [256], as used in cloud data to identify positive and negative anomalies within the time series of positional dynamicity values. It employs statistical learning approach, time series decomposition and robust statistical metrics (e.g., median together ESD). This approach is known as Seasonal Hybrid ESD builds upon the Generalised ESD (Extreme Studentized Deviate) test for detecting in time series.

In this approach, similar to that followed for ARIMA evaluation approach, positional dynamicity values of SINs were calculated using equation (3.2) to define different time series of positional dynamicity values, incident to SINs, and considering different temporal window lengths. Hybrid ESD was then applied to detect the percentage of anomalies in each time series and select the time series with minimum number of anomalies. The window length of the time series of SINs considering their positional dynamicity with minimum anomalies will denote the optimal window length.

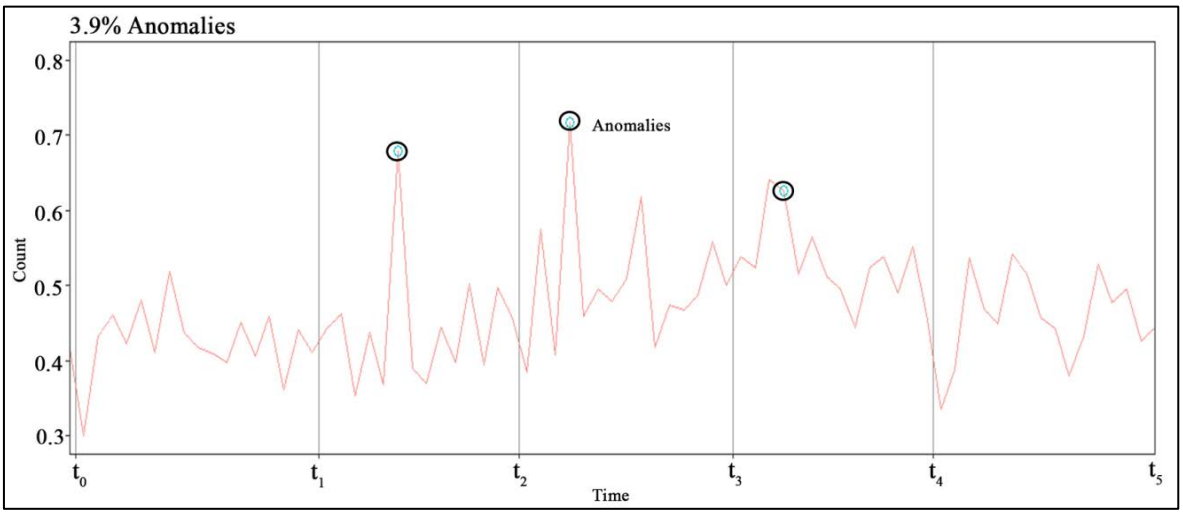


Figure 3.5: Percentage of anomalies present in a time series as determined by the Seasonal Hybrid Extreme Studentized Deviate algorithm.

In Figure (3.5), a visualization of time series anomalies as calculated by the Hybrid ESD algorithm is presented.

3.4.3 *K*-means Clustering

In this approach, an unsupervised learning approach was applied to cluster actors considering their actor-level dynamicity values for different window sizes, computed using equation (3.1). We employed a popular unsupervised learning approach, known as *K*-means clustering. This research clustered actors into *K* groups based on the similarity of their positional dynamicity values and considering the distribution of actor-level positional dynamicity as Gaussian distribution. *K*-means clustering algorithm starts with initial estimates for randomly selected *K* centroids where each centroid defines one cluster and then each data point is assigned to its nearest centroid based on the distance function described below. The objective of K-means clustering is to minimize total intra-cluster variance, known as squared error function. The objective function is defined as:

$$J = \sum_{m=1}^{M} \sum_{n=1}^{N} \left\| x_n^{(m)} - c_m \right\|^2 \quad \dots\dots (3.9)$$

where, M denotes the number of clusters, N represents number of samples, c represents centroid for cluster m and $\left\| x_n^m - c_j \right\|$ denotes the distance function. If the distribution of calculated actor dynamicity values is sparse (i.e., the variance is high or too many extreme values) and the range of the actor dynamicity values is high, then there will be more clusters in comparison to the distribution with low variance. Therefore, in this step, the optimal number of clusters in *K*-means clustering was determined by considering the actor dynamicity values for different window lengths to identify the window size, for which window size the distribution of actors' positional dynamicity has minimum centroids or lower number of clusters. In regards to the optimality of the number of clusters, the ultimate objective is to minimize the error measure, denoted by the total within-cluster sum of squares around the cluster means, as denoted in equation (3.9). Then this study attempted to find out the total within-cluster variance or the total within-cluster sum of square (i.e., square of the distance function in equation 3.9) of these clusters. The window size, for which the total value of within-cluster sum of square or the value of the distance function in equation (3.9) in *K*-means clustering over actor dynamicity values is lowest, will be the potential candidate for the optimal window length.

In clustering approach, identification of the optimal number of clusters is somehow subjective and depends on the methods used for measuring similarities among data points and the parameters used for partitioning. Generally, clustering algorithms are designed for multivariate environment where dataset is a collection of features describing each data point. However, the popular heuristic *K*-means algorithm is unable to guarantee the optimal number of clusters for univariate data. Since, in this study, I have one-dimensional data (i.e., actors' positional dynamicity), I used '*Ckmeans.1d.dp*' algorithm, developed by Wang & Song, which performs optimal one-dimensional k-means clustering using dynamic programming [257]. In this evaluation method, the first attempt is to find out the optimal number of clusters

considering univariate actors' positional dynamicity distribution using different window sizes and, secondly, for each window size, calculate the within-cluster sum of squared distances. In every network datasets, the window size that gives lowest values for both quantities are considered as the optimal window size.

3.5 Conclusion

In dynamic or longitudinal networks, one important task is to identify the correct, appropriate or optimal choice of aggregation granularity in order to perform binning any stream of time stamped links to discern meaningful information and understand the rate of dynamics demonstrated by these networks. This identification of correct window length strongly impacts the structural analyses, efficacy of network mining and dynamics demonstrated by networks [258,246,107]. Having too coarse or too fine temporal granularity may conceal or fail to unravel critical information about network dynamics and impair the understanding of the structure of underlying interactions. Further, appropriate temporal binning decision in dynamic networks will enable to distinguish between noisy, local and critical temporal orderings.

In the literature, there is a lack of actor-oriented measurement/method on the selection of optimal window size to analyse longitudinal networks and often the task is left on arbitrary choices of scholars depending on the experimental contexts or the requirements of the corresponding study. Sometimes it is also left up to the data collection process which is impractical. Researchers also attempted to exploit network-level structural properties across temporal network snapshots to identify the appropriate window length as discussed in section one. Therefore, this thesis proposed an approach that can be used to determine the appropriate window size for the analysis of any longitudinal network in relation to different actor-level perspectives. The approach was based on the concept of an actor-level dynamicity that

quantifies changes in actors' network involvements (in terms of network position) during the evolution of the underlying longitudinal network. In detecting optimal window length, firstly, positional dynamicity was defined that quantifies the change associated to actor's structural positions in networks over time. A combination of three well defined centrality measures was used to measure the positional dynamicity values of actors. These centrality measures have long been exploited in network science not only to quantify actor's network activities but also to define actor's prominence, communicability and reachability. To determine the optimal window length, the variances of actor dynamicity values will be compared by considering different time scale durations. The window length with minimum variance in actor dynamicity distributions define the appropriate sampling window to analyse dynamic network because the minimum variance will ensure that the suggested window size will neither be too large for some actors that reveal high rates of network activities and consequently exhibit the maximum network dynamicity nor be too small for some other actors that reveal slow rates of network activities and consequently exhibit the minimum dynamicity. It is noteworthy that the proposed method determines the optimal sampling time scale/window from a list of candidate time windows whereas these time windows are network dependent and can be of any durations (e.g., second, minute, hour, day, month). For example, if in a dynamic network, streaming links are collected or aggregated in every second then choices of candidate windows in multiple of day(s) would be inappropriate. Similarly, if links are aggregated in a dynamic network by considering temporal unit of single day, then choices of candidate windows in seconds or minutes would produce inaccurate results. Further, three validation tests were also proposed using time series analysis and unsupervised learning methods to support justification of the resultant optimal window length from the approach suggested in this research study.

Chapter 4

Actor-oriented Evolution

<u>The contents of this chapter were published in the following articles</u>

1. **Choudhury, N.,** & Uddin, S. (2017). Mining Actor-level Structural and Neighbourhood Evolution for Link Prediction in Dynamic Networks. *Paper presented at the Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017 ASONAM*, Sydney, Australia.

2. **Choudhury, N.,** & Uddin, S. (2018). Evolutionary Community Mining for Link Prediction in Dynamic Networks. In C. Cherifi, H. Cherifi, M. Karsai, & M. Musolesi (Eds.), *Complex Networks & Their Applications VI: Proceedings of Complex Networks 2017* (pp. 127-138). Springer International Publishing, Lyon, France.

4.1 Introduction

Increasing size and complexity of dynamic networks raised the necessity of splitting a large network into small-scale manageable components and thereby facilitating the visualisation and inference procedure. Such splitting not only simplifies the exploration of different aspects of network but also describe the network without computational difficulties. Therefore, in the previous chapter, a sampling/discretization strategy was suggested to split the stream of links in dynamic networks into small-scale temporal network snapshots. An individual snapshot is called short interval network (SIN)[1]. In these temporal network snapshots, actors experience varying dynamicity in regards to their network positions, and neighbourhood formed over time.

Temporal variations of different network activities (e.g., forming or severing links) result in micro-scale temporal changes of actors' network structural positions and neighbourhood. Further, these actor-oriented microscopic network changes may result in mesoscopic alterations of network structure (e.g., communities of actors). Communities in social networks implicitly denote groups of actors with similar features or attributes or actors closely tied according to their roles, social interests, or collective behaviour. As attributes, social patterns, roles and interest of actors change over time, so do their network activities and association patterns. Consequently, these result in fluctuations of both local and global network structures. Further, due to the evolutionary patterns of link structures of actors in dynamic networks, across the different time intervals, they may either retain existing community memberships or gain new membership to different communities. Consequently, communities of actors may shrink or increase in size or completely disappear, erode, or new communities may form over time. Therefore, it is believed that in evolving social network, temporal microscale actor-level changes trigger mesoscopic or collective changes.

[1] Please see appendix A

Considering three different types of actor-oriented evolutionary aspects (i.e., network structure, neighbourhood and community-aware), experienced by each actor in dynamic networks, in this chapter, I define three different types of actor dynamicities. The similarity between a pair of actors, in regards to these three types of evolution, representing their evolutionary proximity, will be used in the later chapter to develop features for dynamic link prediction.

4.2 Actor Dynamicity

In the previous chapter, the term actor dynamicity was explained in detail including two classes of actor dynamicities: (i) positional and (ii) participation[1]. By considering the degree of actor-level temporal fluctuations, Uddin et al. proposed these two types of actor-level dynamicities in their recent study [98]. The term '*actor dynamicity*' refers to the variable involvement of individual actors in dynamic social networks. Focusing on this concept, I

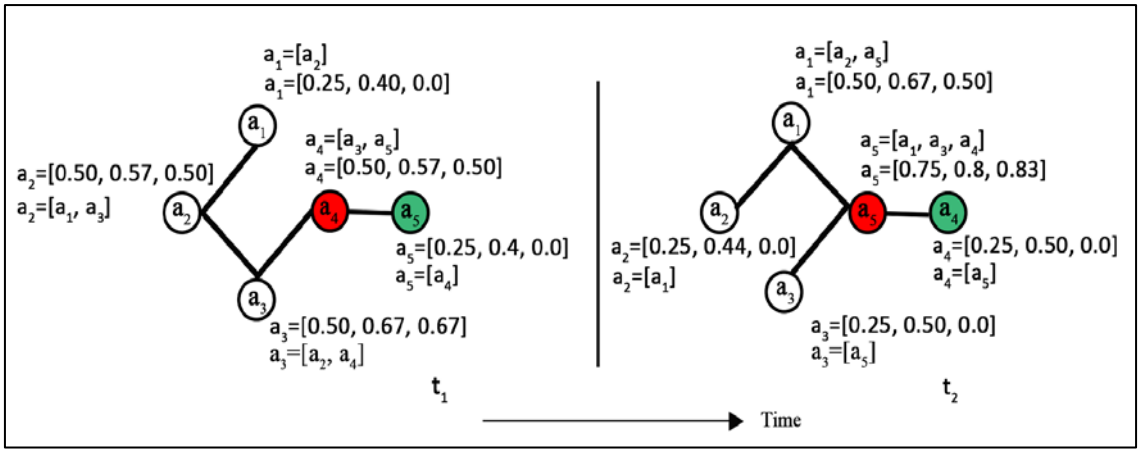


Figure 4.1: Visualization of an actors' positional and neighbourhood changes in a dynamic network consists of two Short Interval Networks (SINs) at two different timestamps t_1 and t_2. All actors are accompanied with their [degree, closeness, and betweenness] centrality measures in the corresponding SIN including their direct neighbourhoods. Actor a_4 and a_5 are coloured red and green to represent how their centrality measures are changed due to their positional changes in the SINs over time.

explored the temporal changes, incident to actors' in the dynamic networks, in regards to their link structures, neighbourhoods, and community participations in every SIN, to construct my dynamic similarity metrics/dynamic features. Modifications of actors' network positions in SINs over time due to their varying nature of network activities (i.e., link formation, link deletion) and changing neighbourhoods is visualized in (Figure 4.1). In this figure, link structures and neighbourhoods of all actors including their normalized centrality measures (i.e., For example, degree centrality of an actor a is calculated as $\frac{degree\ of\ a}{n-1}$, see detail in chapter 3) are presented in two different SINs at two different timestamps (i.e., t_1 and t_2 where $t_2 > t_1$). It is observable that the varying network positions of actors in temporal networks can effectively be mapped by the centrality measures. For example, actor a_4 experienced higher degree, closeness, and betweenness centrality in comparison to a_5 in the SIN at t_1 whereas actor a_5 achieved higher measure in the SIN at t_2, in comparison to a_4, due to their corresponding network positions changes. Likewise, actor a_4 lost one of its neighbours in the SIN at t_2 where in the SIN at t_1, it had two direct neighbours. Simultaneously, in addition to its retention of a previously gained neighbours (i.e., a_4) in the SIN at t_1, a_5 gained two new neighbours in the SIN at t_2 (i.e., a_3, a_1). These temporal changes in SINs occur over time due temporal micro-scale network activities performed by actors.

On the other hand, in most social networks, there are parts where the actors are more densely connected to each other than the actors in the rest of the network. These condensed regions, known as clusters or communities, consist of actors with common structural properties, objectives or goals. With the wide adoption of networks to understand the social interaction pattern, the term 'community' started representing closely-connected actors demonstrating certain common characteristic structural properties [259]. According to Santo Fortunato, both global and local heterogeneous distributions of links within networked

systems result in spawning community structure within networks [9]. In evolving social network, interactions among its actors evolve over time and so do their community patterns. The underlying reasons are divergent, for example, actors may change their roles, acquire new links, severe old links with others, or new actors and links emerged. Simultaneously, with network evolution, owing to various network events, actors may join or leave a

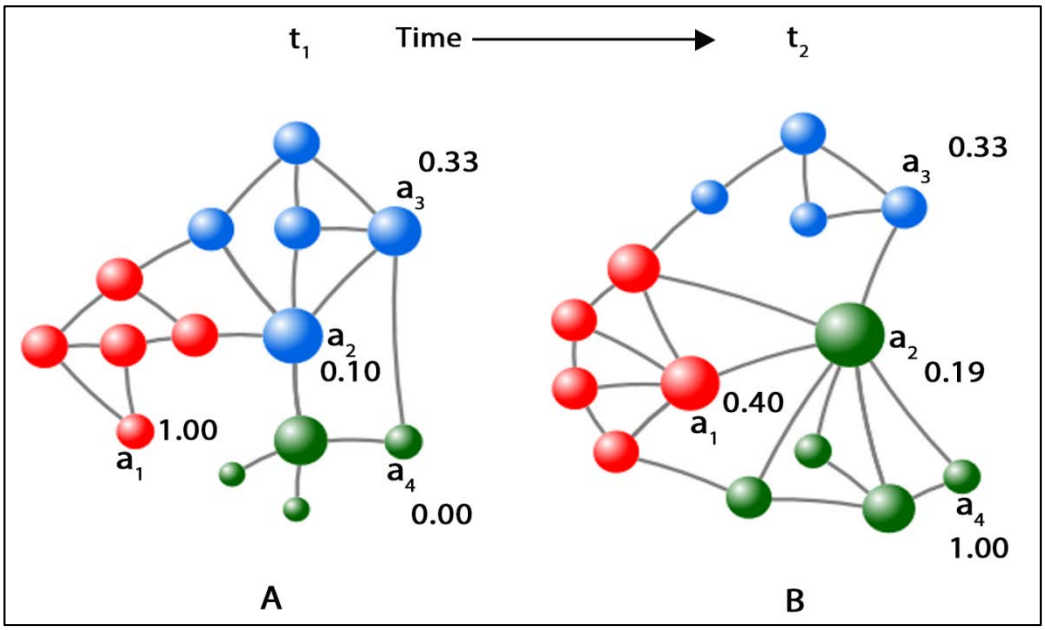


Figure 4.2: Visualization of an actors' clustering tendency changes in a dynamic network consists of two Short Interval Networks (SINs) at two different timestamps t_1 and t_2. Four actors (i.e., $a_1, a_2, a_3,$ and a_4) are accompanied with their clustering coefficient values in the corresponding SIN. The sizes and colours of the actors represent their respective degree centrality and communities they belong to.

community that results shrinking or expanding the size of communities, merging, splitting or diminishing the existing communities or even engendering new communities. In (Figure 4.2), this phenomenon is visualized with the help of two abstract SINs at two different timestamps (i.e., t_1, t_2) in a dynamic network metaphor. The sizes of actors in the network snapshots are proportionate to their degree of connections and the colour codes represent different communities where different actors participated. Four actors (i.e., a_1, a_2, a_3, a_4) are

accompanied with their clustering coefficient values at two different timestamps. This figure demonstrates various aspects of actor-level temporal microscale changes resulting in community-aware mesoscale network alterations. For example, in this figure, at time t_2, actor a_4 changes its community as a result of its neighbourhood changes. Likewise, although, the clustering tendency of actors changes as a result of altering link structures among actor's neighbourhood; however, acquiring more neighbourhoods does not implicitly extend actor's cliquishness. It is also evident that varying neighbourhood and actors' network positional changes simultaneously affect their clustering disposition.

Considering these facts, in conjunction to the aforementioned observations, evident from Figures (4.1 & 4.2), I defined three types of actor-oriented dynamicities. Firstly, motivated by the concept of positional dynamicity, as described in the previous chapter, and considering the link structural changes (i.e., link formation and dissolution by an actor), I defined actor-level *structural dynamicity*. Secondly, by considering the alterations of neighbourhoods over time in a series of network snapshots, I defined *neighbourhood dynamicity*. Finally, the *community dynamicity* is defined by the degree of evolutionary changes, in regards to actor's participation in communities or its clustering tendency, in SINs over time. In the following sections this study describes these two dynamicity measures:

4.2.1 Structural dynamicity

Motivated by the concept of the positional dynamicity, as defined by Uddin et al. in [98], the change in link structures and network positions, experienced by actors in every SIN over time, can be measured using different network measures used in social network analysis [251]. Therefore, I used the average of the composite measure, described in chapter 3 to measures actor's positional dynamicity, to quantify an actor's structural properties in each network snapshot (SIN) :

$$a_i(g_t) = \begin{cases} \left[C_i^{Deg}(g_t) + C_i^{Cls}(g_t) + C_i^{Betwn}(g_t)\right]/3 & i \in v_t \\ 0 & i \notin v_t \end{cases} \quad \dots\dots (4.1)$$

where $a_i(g_t)$ denotes SNA measure of actor i in a SIN g at time t. $C_i^{Deg}(g_t)$, C_i^{Cls} and C_i^{Betwn} denotes degree, closeness, and betweenness centrality measures of actor i in the SIN g at time t. The term v_t denotes the set of actors in the SIN g at time t. The underlying reasons for using such a composite measure of three centrality measures were described in the previous chapter. Those are, firstly, these measures are well-defined and can successfully quantify an actor's connectivity, position, communication dynamics, influence and broadcasting capabilities, and importance in a network, and, secondly, these measures are correlated [260]. For example, an actor with high betweenness and low closeness centrality can monopolize links from a small number of actors to many others. Likewise, high degree with low closeness centrality denotes that the actor is embedded in cluster far from the rest of the network. Using the aggregating function (i.e., average of the three centrality measures) will normalize the score so that it will be within the range ($0 \leq a_i(g_t) \leq 1$).

To quantify actor-oriented dynamic changes, Uddin et al. also suggested to use both dynamic and static social network topology [251]. The underlying reason is that according to social network topology, dynamic network needs to analysed in regards to the temporal aggregation of links among its actors [248] and simultaneously, different aspects of dynamicity within dynamic networks can be quantified using both static and dynamic topology of social network analysis [250]. Social network analysis (SNA) supports the mapping and measuring of social relationships among actors in regards to links among them [137]. Therefore, in temporal network perspective, static SNA methods are applied to network data aggregated over the entire observation time and in contrast, dynamic SNA methods are applied to a temporal series of network snapshots or data that is collected in split intervals of the total period of observation [261]. For example, while analysing a

communication network from a list of mobile calls made by the users over the duration of a year, a dynamic SNA can be used over temporally sampled data, binned in hourly, daily, weekly, fortnightly or monthly. In contrast, the static SNA considers only one network constructed by aggregated all links and denoting all calls made over a year or more between the mobile network users.

Further, Chen et al. used local topological similarity indices (e.g., AdamicAdar, Jaccard Coefficient), and instead of building time series of these indices, they considered their variations between adjacent time steps [262]. This approach by Chen et al. is diffeent from the other supervised dynamic link prediciton methods [196,195],. Upon drawing the diffeences between temporal properties (e.g., 'return' that denotes the increase and decrease of an actor's degree from one time stamp to the next) and time-aware properties (e.g., change of the number of common neighbours of two actors over time), Chen et al. also developed a method that paid more attention to the evolutionary process of the network calculated by the variations of structural properties to train the classifier in a supervised link prediciotn setup. To find out the intrinsic relationship between the variations of topological propoerties and the formation of links between non-connected actor pairs, Chen et al. defined a measure to quntify the rate of change of the structural propoerties as

$$\Delta x_t(i,j) = \frac{x_{t+1}(i,j) - x_t(i,j)}{x_t(i,j)} \quad \dots\dots (4.2)$$

where $\Delta x_t(i,j)$ denotes the temporal rate of change of topological attribute x, $x_t(i,j)$ denotes the property values incident to actor i and j at timestamp t. Motivated by these two aforementioned concepts, I defined structural dynamicity as the rate or degree of actor-level structural changes computed at time t using the following equation:

$$\delta_i(t) = \frac{|a_i(g_t) - a_i(g_{t-1})|}{a_i(g_t \cup g_{t-1})} \qquad \dots (4.3)$$

where $\delta_a(t)$ denotes the degree of structural dynamicity demonstrated by an actor i at time t. $a_i(g_t)$ denotes the composite centrality measure defined in equation (4.1) incident to actor i in a SIN g_t at time t and finally $g_t \cup g_{t-1}$ in the denominator denotes the aggregation of two network snapshots at timestamp t and $t-1$. Although this structural dynamicity measure looks analogous to the positional dynamicity[1] measure defined in chapter 3 [98]; however, in equation (4.3), the aggregated network is computed at every time stamp differently. At every time stamp ($t > 1$), the structural difference of an actor between consecutive SINs is normalized by the structural measure of that actor computed in an aggregated network, constructed by combining g_t and g_{t-1}. Conversely, in positional dynamicity, Uddin et al. computed the aggregated network once and which is the union of all SINs together. The downside of this measure is that in streaming networks, assuming the total number SINs may not be feasible in advance.

[1]Appendix A

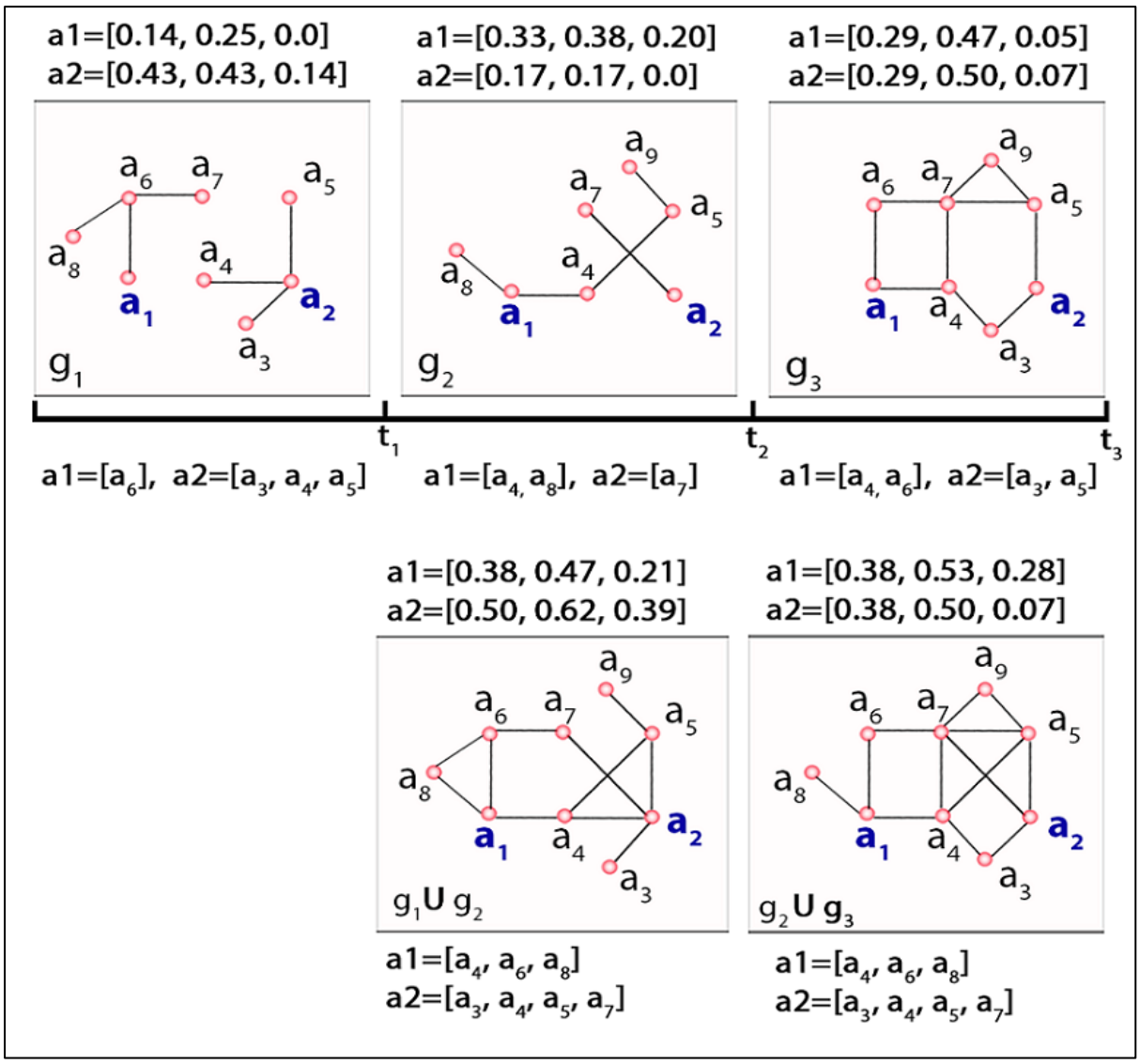


Figure 4.3: An abstract visualization of a dynamic network considering a series of evolutionary network snapshots at different discrete timestamps ($t = 1, 2,$ and 3) which is used to metaphorically demonstrate the computation of actor-level structural and neighbourhood dynamicity measures. The top row represents a time series of three Short Interval Networks (SINs) g_1, g_2, g and the bottom row represent the aggregated networks at timestamps $t = 2, 3$ where the first network denotes union of G_1 (i.e., $g_1 \cup g_2$) and the second denotes union of g_2, g_3 (i.e., $g_2 \cup g_3$). On top of each SIN, two actors a_1 and a_2 are accompanied by their degree, closeness and betweenness centrality measures computed in the corresponding SIN. At the bottom, their direct neighbourhoods are presented.

According to Figure (4.3), the composite measure of actor a_1 is 0.39 in g_1 whereas in g_2, it is 0.91 for the same actor. Similarly, from the bottom row of this figure we get the composite measure of this actor in the aggregated network ($i.e., g_1 \cup g_2$) at timestamp $t = 2$ as 1.06. Therefore, at time $t = 2$, the degree of structural evolution for actor a_1 can be measured as $\left[\frac{|0.91-0.39|}{1.06} = 0.491\right]$. Similarly, at timestamp $t = 3$, the degree of structural evolution, experienced by the same actor is measured as $\left[\frac{|0.81-0.91|}{1.19} = 0.084\right]$. Since link prediction algorithms predominantly focus on network growth [263], the denominator in equation (4.3) denotes the composite centrality measures that an actor ought to achieve in a growing network. To be more precise, I compared an actor's network structural difference in SINs at adjacent timestamps against the structural position in a static network consists of all links present at those two timestamps- an aggregated network without considering removal of any links.

4.2.2 Neighbourhood Dynamicity

In social network analysis, neighbourhood is defined as the local region around individual actors considering different path lengths [264]. The neighbourhood also includes all the links among all the actors having direct connection with egos. Neighbourhood based analysis within SINs can disclose different aspects of networks, including interesting features (e.g., local leadership changes, spurious/irregular activities) and structures not available from the aggregated global network [265]. Although, in this study, we considered the neighbourhood as an individual actor's immediate field of interactions (i.e., at distance one); however, a further study can explore neighbourhood at different distances to observe the prediction performance. Subsequently, the neighbourhood dynamicity of an actor is measured in a SIN at timestamp ($t > 1$) as the ratio of an actor's total neighbour count in G_t in comparison to the total neighbour count in an aggregated network at timestamp t. The aggregated network,

in this context, at time t is computed by aggregating all the links from all SINs, starting from the beginning till t. This ratio is further multiplied with the neighbourhood gaining rate at timestamp t for all actors in G_t. Thus the neighborhood dynamicity λ_a of actor i at time t is defined as:

$$\lambda_i(t) = \left[\frac{|\aleph_i(g_t)|}{|\aleph_i(\cup_{n=1}^{t} g_n)|}\right] * \frac{1}{\{V_t - \aleph_i(g_t)\}} \qquad \dots\dots (4.4)$$

where $\aleph_i(g_t)$ denotes the set of neighbours of actor i and V_t denotes the total number of actors in the SIN at timestamp t. The denominator in the first part of the equation denotes the neighbourhood of actor i in an aggregated network comprised of all SINs before and at timestamp t (i.e., $g_1 \cup g_2 \cup g_3 \cup \dots \cup g_t$). The reason behind considering the aggregated network differently from how it was considered in the case of structural dynamicity is that in this case we consider the neighbours of an actor by their identity. In case of structural dynamicity, acquiring one link in a SIN will increase the corresponding actor's degree centrality whereas losing a neighbour will simply decrease it irrespective of neighbour's identity. If an actor connects to another in the first SIN, severe this link in the second SIN, and then again form a link with the same actor in the third SIN, these facts will augment the corresponding centrality measures of that actor in aggregated network in the equation (4.3). In contrast, the aggregated network in the equation (4.4) that neighbour will be considered once in the SIN where the neighbourhood was established first irrespective of the removal of the link between these two actors. The rationale behind this is that in neighbourhood dynamicity, I analysed the rate of actor's neighbourhood retention and at the same time acquiring new neighbours at each time stamp. For example, in Figure 4.3, at timestamp t_3, actor a_2 gained two of its old neighbours (i.e., a_3 and a_5 from the SIN at t_1) and consequently, the denominator of the equation (4.4) would be three both in t_2 and t_3. However, in case of structural dynamicity calculation in the equation (4.3), considering

degree centrality, the value in SIN at t_3 (i.e., $a_2(g_1 \cup g_2) = 4$) would be higher than what a_2 achieved in the aggregated network at timestamp t_3 (i.e., $a_2(g_2 \cup g_3) = 3$).

From the equation (4.4), we can observe that an actor can have maximum score of one as neighbourhood dynamicity if it forms association with every other actor in SIN at timestamp $t = 1$, maintaining its neighbourhood in all the subsequent SINs in the dynamic network, and form association with every new actor appearing in subsequent SINs. On the other hand if an actor does not participate in any SIN (i.e., actor is not connected to any other actors or actor does not have any neighbours), its neighbourhood dynamicity score will be zero. From this equation, it is apparent that associations with more new actors in SINs and maintaining the acquired neighbourhood in subsequent SINs will ameliorate actor's neighbourhood dynamicity score. It is noteworthy that for the first SIN the aggregated network in the denominator of equation (4.4) consists of only one and the first network snapshot. Therefore, for the first SIN where an actor appears (i.e., an actor gains neighbourhood), the first part of this equation before the multiplication sign assigns a value of one for that actor. For example, in Figure (4.3), actor a_1, the neighbourhood dynamicity at $t = 1$ is computed as $\left[1 * \frac{1}{8-1} = 0.143\right]$. Similarly, for $t = 2$ and 3, the neighbourhood dynamicity values for a_1 is 0.133 and 0.095. For a_2, the time series of neighbourhood dynamicity is $[0.2, 0.042, 0.083]$. Conversely, considering actor a_9, the temporal sequence of neighbourhood dynamicity is $[0, 0.167, 0.167]$ where this actor's dynamicity is zero in g_1 due to its absence in that SIN.

4.2.3 Community Dynamicity

In conjunction with the actor dynamicity, varying roles and divergent network activities trigger changes in social communities within these network snapshots. Communities may appear, disappear, merge, split, shrink, expand or even sometimes remain unmodified without

incurring any changes. Understanding the evolutionary patterns of network communities, actors, and their community participations may support researchers understanding the underlying network evolution. Particularly, it can assist the social scientists to comprehend the underlying growth pattern of social networks. Different types of evolutionary changes, evident from the aforementioned figures; those are triggered by temporal variations of different network activities performed by network actors. Embracing this concept, I also computed the community-oriented actor dynamicity. The term '*community dynamicity*' in this study denotes the ratio of evolutionary changes of actor's participation in communities or its clustering tendency in SINs against its clustering tendency in temporally aggregated network over time. In conjunction, the community dynamicity also considers the corresponding temporal neighbourhood changes. The rationale behind using aggregated network, as mentioned earlier, is that link prediction mechanism of network science predominantly deals with network growth and in dynamic network analysis links are aggregated by considering an aggregation window size to accumulate links temporally.

In network theory, actor's *clustering coefficient*[1] measures the degree the actors in networks tend to cluster together. Since in social networks, actors tend to build friendship with friends of their friends, this coefficient measures the extent one actor's friends are also friends. For a complex social network, this measure characterises both global and local cliquishness of the actors and the network in regards to the triadic closure mechanism that characterizes the network evolution. Triadic closure emerges when friends to a common friend become friend as well and this is a general phenomenon in social networks. In this study, we consider the local clustering co-efficient with view to understand the actor-level evolution instead of the network itself. The clustering co-efficient of an actor i in a network snapshot G_t at timestamp t is defined as:

[1] Please see appendix A

$$CC_{G_t}^{i} = \frac{2\mathcal{L}_i}{D_{G_t}^{i}\left(D_{G_t}^{i} - 1\right)} \quad \ldots (4.5)$$

where, $D_{G_t}^{i}$ denotes the number of direct neighbours or degree of actor i in a network snapshot G_t at time t and $\mathcal{L}_i$ denotes the number of links between D^i neighbours of actor v. Subsequently, an actor's community dynamicity using its clustering coefficient is measured in a SIN at timestamp t as follows:

$$\partial_i(t) = \left[\frac{\left|CC_{G_t}^{i} - CC_{G_{t-1}}^{i}\right|}{CC_{G_T}^{i}}\right]^{e^{\left[\frac{2\left|\eta_{G_t}^{i} \cap \eta_{G_{t-1}}^{i}\right|}{D_{G_t}^{i} + D_{G_{t-1}}^{i}}\right]}} \quad \ldots (4.6)$$

where $C_{G_t}^{i}$ represents the clustering co-efficient and $\eta_{G_t}^{i}$ denotes the neighbourhood of actor i in SIN G_t at timestamp t and G_T denotes an aggregated network as the union of two SINs at two adjacent timestamps (i.e., $G_T = G_t \cup G_{t-1}$). The numerator in the base part of equation (4.6) represents the ratio of the rate of clustering coefficient changes of an actor in two adjacent SINs at timestamps t and $t - 1$. On the other hand, the denominator represents the clustering coefficient of that actor in an aggregated network consists of SINs at those timestamps. The denominator basically normalizes the difference in the numerator in regards to the cliquishness[1] of the actor what it ought to achieve in a static network without severing any links. This score is further amplified by an exponent that measures the neighbourhood achievement and retention score of that actor at two adjacent timestamps; and the power of the exponent considers the Sorensen index [266] of the actor's neighbourhood in G_{t-1} and G_t. An actor can achieve high community dynamicity score if more of its neighbours participate in triadic closure events and simultaneously, it retains its neighbourhood between two adjacent timestamps. However, if the actor changes its communities in consecutive timestamps then the minimum value the exponent in equation (4.6) can have is zero (0) and

[1] Please see appendix A

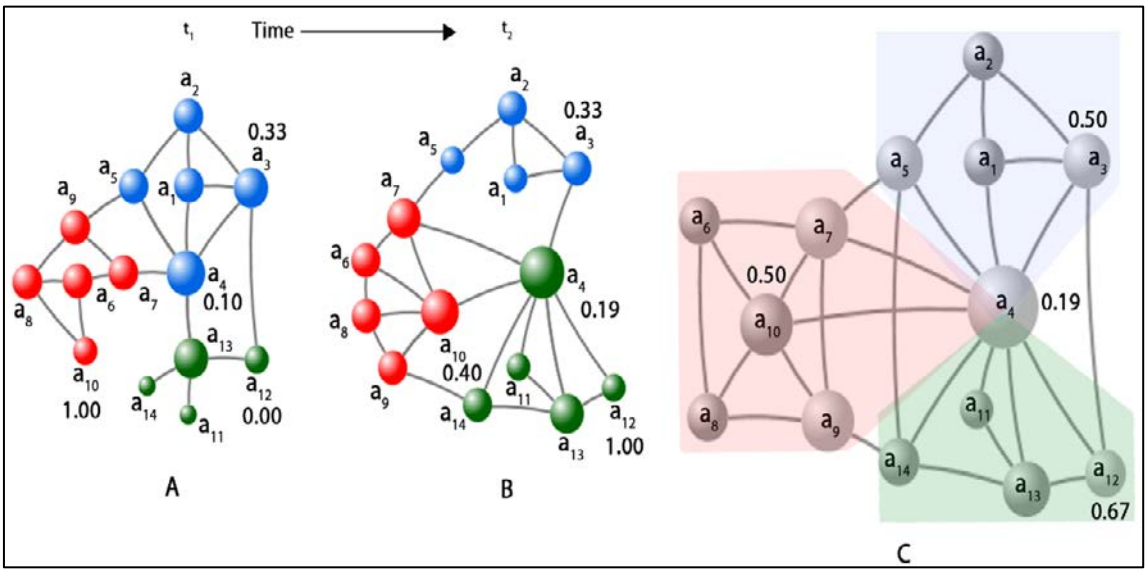


Figure 4.4: An abstract visualization of a dynamic network comprised of two Short-Interval Networks (SINs) (A) G_{t1} at time t_1 and (B) G_{t2} at time t_2 and (C) denotes an aggregation of G_{t1} and G_{t2} (i.e., $G_{t1} \cup G_{t2}$). Each SIN has three communities that are represented by three different colors and actors within these communities represent the color of the corresponding community. Actors a_3, a_4, a_{10}, and a_{12} are accompanied by their clustering coefficient values in G_{t1}, G_{t2} and the aggregated network on the right.

which will eventually penalize its dynamicity score. In this case, the community dynamicity score will be defined by the rate of clustering coefficient changes in consecutive SINs. Therefore, if the clustering tendency of an actor does not change in two adjacent SINs, then its lowest possible community dynamicity can plummet to zero. For example, from Figure (4.4), the community dynamicity of actor a_{10} at timestamp $t = 2$ is calculated as $\left[\frac{0.40-1.0}{0.50}\right]^{e^{\left[\frac{2*2}{2+5}\right]}} = 1.381$ in G_{t2}. Similarly, for actor a_4, at timestamp $t = 2$, the community dynamicity value is measured as $\left[\frac{0.19-0.10}{0.19}\right]^{e^{\left[\frac{2*4}{5+7}\right]}} = 0.233$. In this way, a time series of community dynamicity values can be formed by considering each actor in the SINs at each timestamp of a given dynamic network to develop dynamic features.

4.3 Conclusion

In a dynamic network, actors usually exhibit different rates of network activities such as the formation of new links or the dissolution of an existing links. Further, due to the evolutionary patterns of actors' link structures, they eventually endure existing community membership or gain new membership to different communities. Consequently, communities of actors may shrink or increase in size, erode, completely disappear or re-emerge over time. Thus, in evolving social network, temporal microscale actor-level changes trigger mesoscale or collective changes. On the other hand, a well-connected longitudinal network could reveal low dynamicity values for its member actors, if those actors had similar levels of dynamicities over time. Conversely, a sparse longitudinal network could reveal high dynamicity values for its member actors if those actors had different levels of dynamicity over time. Considering these, this study has defined three different types of actor-level dynamicities, demonstrated by actors in dynamic networks. Firstly, structural dynamicity quantifies an actor's network structural changes measured by three well-defined centrality measures. This also quantifies actor's popularity changes including changes in its broadcasting or brokerage capabilities in networks over time. Secondly, the neighbourhood dynamicity quantifies an actor's neighbourhood retention rate including its rate of gaining new neighbours or loosing exiting ones. Finally, the community dynamicity measures the temporal changes of actor's clustering tendency or its inclination of community participation. These three dynamicity measures can be used as units of changes, experienced by an individual actor in dynamic networks. In the next chapter, dynamic features will be developed by calculating similarity of these actor-level evolutions experienced by pairs of actors. The features will demonstrate actor evolutionary similarity/proximity in dynamic networks.

Chapter 5

Evolution Similarity and Feature Engineering for Link Prediction

<u>The contents of this chapter were published in the following articles</u>

1. **Choudhury, N.**, & Uddin, S. (2017b). Evolution Similarity for Dynamic Link Prediction in Longitudinal Networks. In B. Gonçalves, R. Menezes, R. Sinatra, & V. Zlatic (Eds.), *Complex Networks VIII: Proceedings of the 8th Conference on Complex Networks CompleNet 2017* (pp. 109-118). Cham: Springer International Publishing.

2. **Choudhury, N.,** & Uddin, S. (2017). Mining Actor-level Structural and Neighbourhood Evolution for Link Prediction in Dynamic Networks. *Paper presented at the Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017 ASONAM*, Sydney, Australia.

3. **Choudhury, N.,** & Uddin, S. (2018). Evolutionary Community Mining for Link Prediction in Dynamic Networks. In C. Cherifi, H. Cherifi, M. Karsai, & M. Musolesi (Eds.), *Complex Networks & Their Applications VI: Proceedings of Complex Networks 2017* (pp. 127-138). Springer International Publishing, Lyon, France.

5.1 Introduction

Among three different categories of link prediction strategies (i.e., similarity-based algorithms, maximum likelihood and probabilistic methods) [17], similarity-based algorithms are the most prevalent prediction methods. In social network analysis, it is widely believed that similar and/or closer actors tend to form link in future. These methods compute different network topological or actor-based similarity scores to denote the similarity and/or proximity between actor pairs and these scores are to denote the likelihood of link formations. For example, actors having large number of common neighbours or actors with similar smoking behaviour are presumed to be friend in future. In the earliest link prediction models, Liben-Nowell and Kleinberg concentrated on graph-based similarity metrics for prediction task [77]. Later, Hasan et al. used various similarity metrics as features in a supervised learning setup and showed that using external information outside the scope of graph topology can significantly improve the prediction result [140].

In previous chapters, we have observed that in temporal network snapshots, actors experience varying dynamicity in regards to their network positions, neighbourhood and communities formed over time. Temporal variations of different network activities (e.g., forming or severing links) result in microscale temporal changes of actors' network structural positions and neighbourhood. Further, these microscale network changes may result in mesoscopic alterations of network structure (e.g., communities of actors). Considering these different types of actor-level dynamicities, feature engineering will be applied in this chapter to generate some important features for the purpose of dynamic link prediction.

Feature engineering is a process of transforming data or raw information into features useful for predictive models. This chapter describes different similarity measurement methods used as part of the feature engineering process to compute the evolution similarity

score between actor-pairs. Different features engineered in this chapter represent similarity/proximity between actor-pairs in regards to different types of evolutionary aspects. Unlike static networks where similarity/proximity scores are calculated by mining different network (i.e., graph) topology and actor attributes, the similarity/proximity scores between actor-pairs are calculated by mining different types of actor-level evolutions. Therefore, the features constructed in this chapter are named as *dynamic similarity metrics* or simply, *dynamic features*. Therefore, this study sought to develop such dynamic features by computing similarity between actors by considering their changes in link structure, neighbourhood and community-specific information over time in dynamic networks.

5.2 Dynamic Similarity Metrics/Dynamic Features

In this section, I describe different methods to define dynamic features for the purpose of link prediction. These methods consider three evolutionary aspects of non-connected actor pairs defined in the previous chapter. These features will denote the similarity/proximity between actors in regards to their structural, neighbourhood or community evolution. To define the similarity/proximity between actor pairs, we compare the time series information constructed by using structural, neighbourhood, and community dynamicity, described in chapter 4, incident to actor pairs.

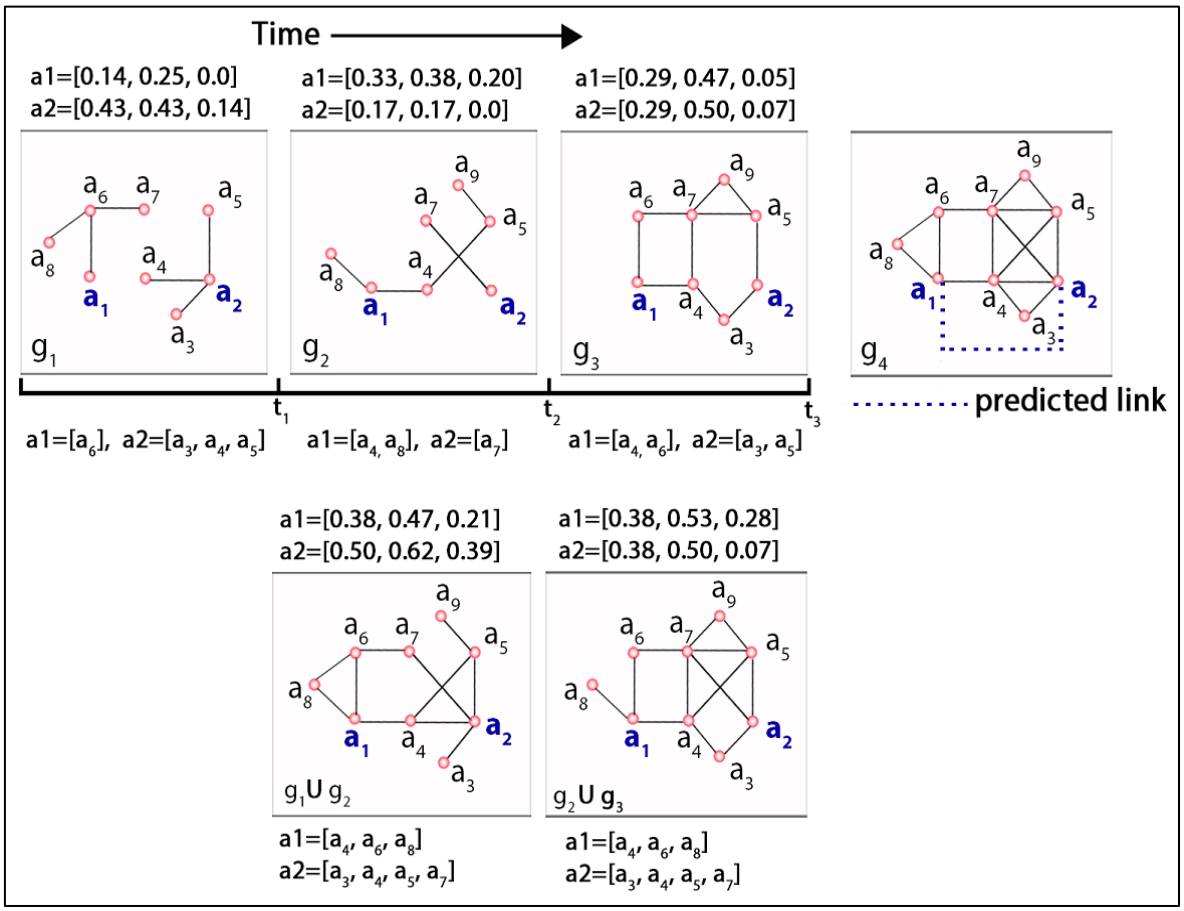


Figure 5.1: An abstract visualization of the dynamic link prediction framework considering a series of evolutionary network snapshots at different discrete timestamps ($t = 1, 2, 3$ and 4). The top row represents a time series of three Short-Interval Networks (SINs) g_1, g_2, g_3, where these evolutionary networks are analyzed to predict a future link between actor a_1 and a_2 at timestamp $t = 4$ in g_4. The bottom row represents the aggregated networks at timestamps $t = 2, 3$, in which the first network denotes union of g_1, g_2(i.e., $g_1 \cup g_2$), and the second denotes union of g_2, g_3(i.e., $g_2 \cup g_3$). The numbers on top of each SIN represent the degree, closeness, and betweenness centrality measures of actors a_1 and a_2 in each SIN. At the bottom of each SIN, the direct neighborhoods are presented, incident to these two actors

For example, according to Figure (5.1), to predict a link between actors a_1 and a_2 in G_4, this study builds three separate temporal sequences of $\delta_a(t)$, $\lambda_a(t)$ and $\partial_a(t)$ (i.e., actor-level structural, neighbourhood and community dynamicity) incident to actors a_1 and a_2. For these two actors, the temporal sequences of structural dynamicity are $a_1 = [0, 0.491, 0.091]$

and $a_2 = [0, 0.437, 0.436]$. It is noteworthy that for the first timestamp the structural dynamicity is assigned to zero since no variation can be computed using the first SIN. The similarity between a pair of actors is defined in regards to temporal similarity, correlation, and other similarity coefficients computed over temporal sequences encompassing their dynamicity values over time. In the following sections, different methods to compute the similarity/proximity between actor-level evolutionary information for non-connected actor, are described. Each method assigns a score $sim_i(a, b)$ to a pair of actors (a, b) where i^{th} method computes similarity or proximity between actors a and b.

5.2.1 Temporal Similarity

The time-series of different network structural and/or topological properties considering multiple snapshots of the network or temporal information of link occurrences in network snapshots are widely used as input for link prediction in dynamic networks. In chapter 2, we have observed that researchers used the time series approach for link prediction to emulate the dynamic behaviour of complex networks. They also used the link creation time to analyse the effect of the elapsed time since a link first appeared and/or to assess the effect of 'recentness' on new links around associated nodes. Although, few link prediction strategies have utilized time-series of actor level network attributes and time-aware techniques or forecasting methods to measure the probability of future links, however, they ignored the temporal similarity measures of these attributes.

Due to the pervasive nature of time series date in many scientific domains, comparing different time sequences and similarity based matching of time sequence data is common in scientific research including signal processing and speech recognition. In some studies, simple (i.e., Euclidean) distance measures suffice, however, there are instances where two time sequences have approximately the same overall component shapes, but they do not align

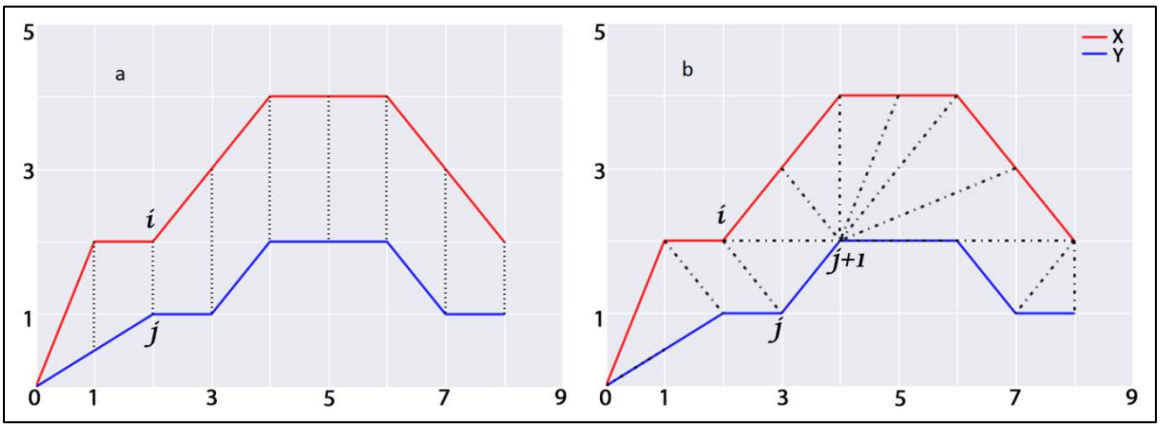


Figure 5.2: Visualizations of measuring similarity between two temporal sequences (a) traditional approach (b) Dynamic Time Warping (DTW) approach. Dashed lines represent the distance between corresponding points in both time series. The traditional approach aligns the ith point in one time series with the corresponding j^{th} point of the other. Moreover, DTW provides non-linear alignment to produce a more intuitive similarity measures and allows similar shapes to match ($i \rightarrow j, j+1$) even if it requires localized stretching along the time axis. In the DTW approach, the difference between the two time series is the warped path distance, which is measured by summing the distances between each pair of points connected by the dashed lines.

in X-axis. In (Figure 5.2), we represent a simple example of this phenomenon by using two time series represented by the red and blue lines in the plot. In this figure, we observe that two time-series having similar overall shape but not align in the time axis. Existing distance measures (i.e., Euclidean, Manhattan) produce unintuitive results and demonstrate incompetency to produce optimal alignment while measuring the similarity between such temporal sequences with varying speeds. For example, the Euclidian technique measures distance between two time series simply by summing the squared distances from each point in one time series to the corresponding point in the other Figure (5.2 a). If two time series are identical with one being shifted a little along the time axis, then Euclidean distance may consider them as totally different time sequences.

Therefore, in time series analysis, dynamic time warpping (DTW) technique [267] is widely used to overcome the aforementioned limitation of traditional distance measures to provide intuitive distance measurements between temporal sequences while ignoring both global and local deviations in the time dimension [268]. It measures the similarity between two time-series by shrinking or expanding, or simply 'wrapping' the time axis of one (or both) sequences to achieve better alignment. This wrapping technique is an example of dynamic programming and can measure the similarity between two time series. In Figure (5.2b), each black dotted line connects a point in time series marked 'i' with the corresponding similar point in the time series marked 'j'. If both time-series i and j were identical, all these dotted lines would be straight vertical lines and no-warpping would be necessary to align the temporal sequences. The difference between these two time series is the wrapped path distance which is measured by summing the distances between each pair of points connected by the dashed lines in the figure. Therefore, two time-series will have DTW distances as zero if they were identical except for localized stretching along the time axis. Due to its ability to determine the optimal alignment and similarity between temporal sequences, it is often used in time-series based classification in various domains including data mining, robotics, manufacturing and medicine.

Let $X^a = [x_1, x_2, x_3, \dots\ x_m]$ and $Y^b = [y_1, y_2, y_3, \dots\ y_n]$ be the time series of length $|m|$ and $|n|$ considering the chosen dynamicity measure, described in section 2.1 (i.e., structural and neighbourhood dynamicity), for actors a and b where $m, n \leq T$, and T is the total number of SINs. If $d(x_i, y_j)$ denotes local distance measure (e.g., Euclidean), defined to compare two different points in X^a and Y^b, then the goal of DTW technique is to find an optimal alignment between X^a and Y^b with minimum overall distance [269]. The notion of this alignment depends on the definition of a (m, n)-warping path. I then construct a wrap path which is a sequence $p = p_1, p_2, p_3, \dots, p_\ell$ with $p_\ell = (m_\ell, n_\ell) \in [1{:}m]x[1{:}n]$ for $\ell \in$

$[1{:}\mathcal{L}]$ where $max(|m|,|n|) \leq \mathcal{L} < |m|+|n|$, where $\mathcal{L}$ denotes the length of the warping path. The wrap path considers all points in both time series starting from $p_1 = (1,1)$ to $p_\ell = (|m|,|n|)$ such that $p_\ell = (i,j), p_{\ell+1} = (i',j')$ where $(i \leq i' \leq i+1)$ and $(j \leq j' \leq j+1)$ and i and j are indexes from time series X^a and Y^b respectively. The optimal warping path p^*, between X^a and Y^b , is defined as the minimum distance among all possible warping paths. To accomplish this it may encounter that a single point in one time series may be mapped to multiple points of the other. The optimal warping path is determined by following a dynamic programming method that recursively measures the following function at every step:

$$\gamma(i,j) = d(x_i, y_j) + min\big(\gamma(i-1,j-1), \gamma(i-1,j), \gamma(i,j-1)\big) \quad \text{... (5.1)}$$

The total cost $d_p(X^a, Y^b)$ of a warping path p between X^a and Y^b with respect to the local cost measure $d(x_i, y_j)$ is defined as $d_p(X^a, Y^b) = \sum_{\ell=0}^{\mathcal{L}} d(x_{m\ell}, y_{n\ell})$. Further, the optimal warping path between two temporal sequences is a warping path p^* with the minimal total cost among all possible warping paths:

$$d_{p^*}(X^a, Y^b) = min \sum_{\ell=0}^{\mathcal{L}} d(x_{m\ell}, y_{n\ell}) \mid p \text{ is an } (m,n)\text{-warping path} \quad \text{... (5.2)}$$

Considering temporal sequences of three dynamicity values (i.e., structural, neighbourhood and community), this study applied DTW technique to measure temporal similarity between them. Temporal similarity between time series of actors' dynamicity values will represent their evolutionary proximity/similarity. Therefore, the value of the first

dynamic similarity metrics, developed in this study for actor pair a and b considering structural dynamicity values, is computed as follows:

$$sim_1(a,b) = d_{p^*}(\delta_i^a,\ \delta_j^b) = min\left\{\sum_{\ell=1}^{\mathcal{L}} d(\delta_{ml}^a\ , \delta_{nl}^b)\right. \qquad \text{... (5.3)}$$

where δ_i^a and δ_j^b are the ith and j^{th} element of time series of structural dynamicity, m and n denotes the length of temporal sequences of structural dynamicity values incident to the actor pair a and b, respectively. Similarly, to compute the second and third dynamic similarity metric, the temporal similarity between neighbourhood and community dynamicity values over time, between a pair of actors a and b, can be computed as:

$$sim_2(a,b) = d_{p^*}(\lambda_i^a,\ \lambda_j^b) = min\left\{\sum_{\ell=1}^{\mathcal{L}} d(\lambda_{ml}^a\ , \lambda_{nl}^b)\right. \qquad \text{... (5.4)}$$

$$sim_3(a,b) = d_{p^*}(\partial_i^a,\ \partial_j^b) = min\left\{\sum_{\ell=1}^{\mathcal{L}} d(\partial_{ml}^a\ , \partial_{nl}^b)\right. \qquad \text{... (5.5)}$$

Where λ_i^a and ∂_i^a denote the neighbourhood and community dynamicity values of an actor a.

5.2.2 Correlation-based Similarity

Correlation analysis is a statistical evaluation method quantifies the extent two continuous variables tend to change together and captures the strength and direction of the linear association between these two variables. It is widely used in financial network analysis, asset allocation, portfolio optimization and risk management [270]. Correlation-based network analysis became a popular data-mining tool for visualizing and analyzing biological relationships within large data sets [271]. Actors and links in this type of network represent molecular elements (e.g., metabolites or genes) and their correlation coefficient (strength and sign), respectively [272,273] and links inferred by correlation analyses reflect a coordinated behaviour between actors across the data set (treatments, genotypes, conditions, and time)

[274]. Besides, correlation analysis is widely used in financial analyses where networks of shares can be constructed by using return correlations [275]. This study applied correlation analysis to measure the affinities or similarities between actor pairs in regards to the temporal sequences of dynamicity values in all SINs. The assumption here is that two actors are similar if they change in a similar fashion (i.e., dynamicity values of one actor increases or decreases with the other at the same time) considering three dynamicity measures proposed in this study. If $\delta_a(t)$ and $\delta_b(t)$ denote the structural dynamicity, $\lambda_a(t)$ and $\lambda_b(t)$ denote the temporal neighbourhood dynamicity and $\partial_a(t)$ and $\partial_b(t)$ of actor a and b at time t then the evolution similarity between them is computed in regards to the Pearson correlation coefficient. Considering $X = [x_1, x_2, x_3, \dots\ x_m]$ and $Y = [y_1, y_2, y_3, \dots\ y_n]$ as two continuous variables, the Pearson correlation between these two can be computed by the following:

$$r_{xy} = \frac{\sum_t \left[(x_i - \bar{x})(y_j - \bar{y})\right]}{\sqrt{\sum_t (x_i - \bar{x})^2}\sqrt{\sum_t (y_j - \bar{y})^2}} \qquad \dots (5.6)$$

Between two continuous variables X and Y, if X is a linear function of the other variable Y, then a positive value (i.e., $r > 0$) denotes the existence of a positive correlation between X and Y. Conversely, a negative value (i.e., $r < 0$) denotes the existence of a negative correlation between X and Y and a zero value (i.e., $r = 0$) indicates non-existence of any kind of association. Positive correlation indicates that if one variable increases than the other has a tendency to increase. In contrast, negative correlation denotes the opposite behaviour (i.e., if one variable increases then the other has a tendency to decrease). In case no correlation present, the variable does not demonstrate any tendency.

Therefore, considering three different dynamicity values, computed temporally at each time stamp, as series of continuous variables, the fourth, fifth and sixth dynamic

similarity metrics to measure similarity/proximity between the actor pair a and b are computed as follows:

$$sim_4(a,b) = \frac{\sum_t\left[(\delta_a(t) - \overline{\delta_a})(\delta_b(t) - \overline{\delta_b})\right]}{\sqrt{\sum_t(\delta_a(t) - \overline{\delta_a})^2}\sqrt{\sum_t(\delta_b(t) - \overline{\delta_b})^2}} \quad \text{... (5.7)}$$

$$sim_5(a,b) = \frac{\sum_t\left[(\lambda_a(t) - \overline{\lambda_a})(\lambda_b(t) - \overline{\lambda_b})\right]}{\sqrt{\sum_t(\lambda_a(t) - \overline{\lambda_a})^2}\sqrt{\sum_t(\lambda_b(t) - \overline{\lambda_b})^2}} \quad \text{... (5.8)}$$

$$sim_6(a,b) = \frac{\sum_t\left[(\partial_a(t) - \overline{\partial_a})(\partial_b(t) - \overline{\partial_b})\right]}{\sqrt{\sum_t(\partial_a(t) - \overline{\partial_a})^2}\sqrt{\sum_t(\partial_b(t) - \overline{\partial_b})^2}} \quad \text{... (5.9)}$$

where δ_a, λ_a and ∂_b denote the structural, neighbourhood and community dynamicity of an actor a respectively.

5.2.3 Dynamicity Abundance-base Similarity

Although a significant amount of dynamic link prediction studies have exploited the time series of topological similarity metrics (e.g., CommonNeighbours), this study used abundance-based similarity metric which is widely used in biology and ecology domain. Frequently used by marine ecologists to measure bio-diversity, the *Bray-Curtis* similarity measure was initially proposed by J. Roger Bray and John T. Curtis in 1957 [276] which is principally employed in multivariate analysis of biological assemblage data and signifies the 'relativisation' of species-wise differences in regards to the their total abundance in biological metaphor [277] . Despite the availability of traditional distance measures (e.g., Euclidean) that conform to the concept of distance, there are some more appropriate measures as distance metrics in an environment with multivariate samples. The Bray-Curtis dissimilarity measure is one such well-known measure to quantify the difference between samples when it

comes to ecological abundance data collected at different sampling locations. In ecological perspective, less abundant species including samples with lower total abundances have greater effect on Bray-Curtis index [278]. Using this measure, values for an individual species are standardized once by computing the abundances relative to maximum value attained by that species over all samples and standardized twice with respect to the sample total.

Using Bray-Curtis method, the distance between two entities X and Y in regards to n-dimensional feature space can be determined by the following formula as described by Legendre and Legendre [279]:

$$BC_{XY} = \frac{\sum_{i=1}^{n} |x_i - y_i|}{\sum_{i=1}^{n} |x_i + y_i|} \quad \ldots (5.10)$$

where x_i and y_i denotes the ith feature of species X and Y, respectively. The numerator signifies the differences between X and Y in regards to the abundance of feature i and the denominator normalizes these differences. Instead of considering traditional topological similarity metrics built upon commonality of neighbourhoods and network structure between actors, this study considered the evolutionary aspect (i.e., structural, neighbourhood and community dynamicity) of actors in T number of SINs to compute similarity between them. In the context of this study where each SIN will represent a sampling location, the Bray-Curtis distance between actors a and b using three dynamicity measures can be defined as:

$$BC_{ab} = \frac{\sum_{t=1}^{T} \sum_{i=1}^{n} |x_i - y_i|}{\sum_{t=1}^{T} \sum_{i=1}^{n} |x_i + y_i|} \quad \ldots (5.11)$$

where n denotes the total number of dynamicity values (i.e., $n = 3$). Since in this thesis, I considered three evolutionary features (i.e., structural, neighbourhood and community dynamicity), therefore:

$$BC_{ab} = \frac{\sum_{t=1}^{T}[|\delta_a(t) - \delta_b(t)| + |\lambda_a(t) - \lambda_b(t)| + |\partial_a(t) - \partial_b(t)|]}{\sum_{t=1}^{T}[|\delta_a(t) + \delta_b(t)| + |\lambda_a(t) + \lambda_b(t)| + |\partial_a(t) + \partial_b(t)|]} \quad \dots (5.12)$$

where, δ_a, λ_a and ∂_a are structural, neighbourhood, and community dynamicity values of actor a. Since the distance represents dissimilarity, therefore, $1 - BC_{ab}$ was used to represent similarity. Hence, the seventh dynamic similarity metric in this study is defined as follows:

$$sim_7(a, b) = 1 - BC_{ab} \quad \dots (5.13)$$

5.2.4 Temporal Community-aware Network Structure

To employ community-aware information in link prediction task, it is imperative to partition a network into communities. Most community-aware link prediction methods exploited an existing community detection algorithm to compute the similarity among actor-pairs considering the community-oriented structural information. For example, 'InfoMap' [280] algorithm minimizes the length of random walks and mostly used in information theory was used by Soundaranjan and Hopcroft [281] in their study. Likewise, Valverde-Rebaza and Lopes [282] used the 'Label Propagation' based community detection method [283] to develop a similarity measure for the purpose of link prediction in static networks. Following them, this study used Louvain algorithm [204] and greedy agglomerative hierarchical community detection algorithm proposed by Newman in [284] for the community detection purpose. The former method has been successfully and widely used for detecting communities in many different types of large networks with millions of actors and links. As a greedy optimization method, Louvain optimizes the modularity by firstly looking for smaller

communities locally with optimized modularity (i.e., numerical index to evaluate partitions in a network) and secondly, aggregating actors belonging to the same community to build a network where individual community act as an actor. The latter method of community detection follows a greedy method for the purpose of optimizing and maximizing the modularity and produces a tree-like dendogram as a presentation of hierarchical rendering of

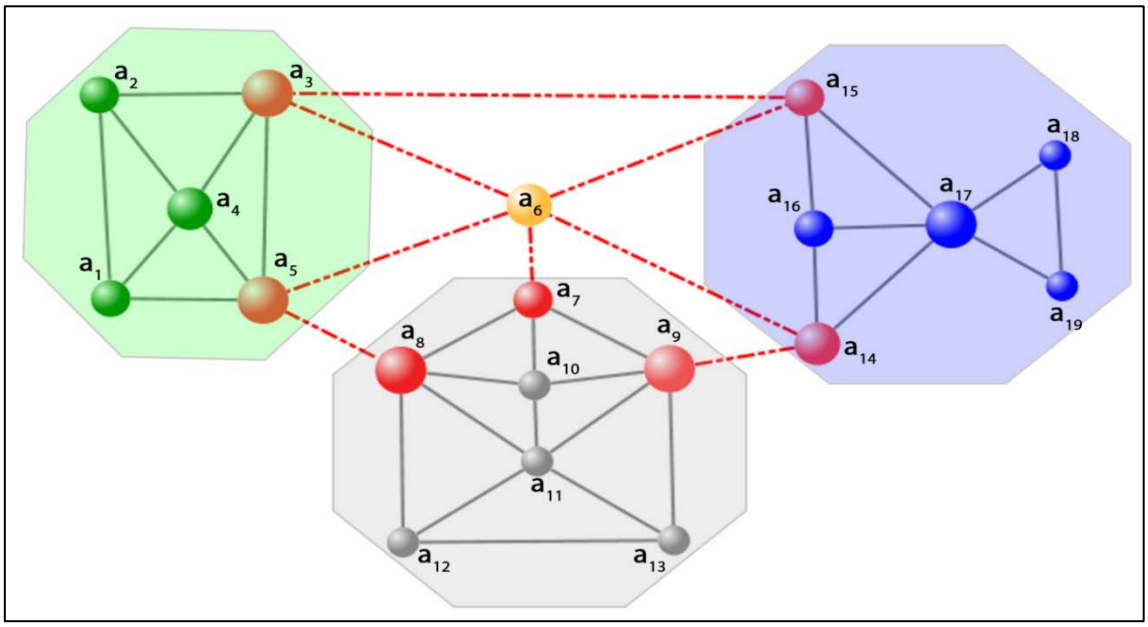


Figure 5.3: Community-aware network architecture supporting link prediction. The orange-coloured actor a_6 is an actor with multiple community memberships. The red-colored actors in each community represent the peripheral actors in each community. Red-colored dotted links denote the bilateral links bridging two communities. It is noteworthy that links connected to actor a_6 from individual communities are considered bilateral links.

the network communities. This algorithm can efficiently clusters large number of actors while generating the given number of communities and also known well for its scaling capability. The final dynamic similarity metric was computed in this research by using the community-aware information extracted from the communities detected in each SIN of a given dynamic network. For each non-connected actor pair, in each SIN using the identified

communities, similarity between a pair of actors was computed depending on their community participation including the structures of communities.

Before delving into the actual similarity/proximity score between actor-pairs, I first define few preliminary concepts and notation those will be used in the following sections with the help of Figure (5.3).

5.2.4.1 Peripheral Actors: If an actor simultaneously belong to more than one community, or resides in one community but belong to one end of a link where the other end belongs to another actor from a different community, is considerd as a peripheral actor. Similarly, if an actor is connected to another actor that has multiple community memberships is also considered as a peripheral actor. For example the green-coloured actor a_6 in Figure (5.3) is a peripheral actor that has multiple community memberships. Similarly, the red-coloured actors a_3 , a_5, a_7, a_8, a_9, a_{14} , and a_{15} are considered as peripheral actors for their respective communities since they are either part of links transcending more than one community or conneced to an actor having multiple community memberships. If C_i and C_j are two communities in a SIN g_t and V_i and V_j denote the set of actors belong to these two communities, then a peripheral actor is denoted by $v_{g_t}^{i.j}$. A set of peripheral actors between two communities C_i and C_j in a SIN g_t is denoted by $\left|V_{g_t}^{i.j}\right|$ where $i \neq j$.

5.2.4.2 Bilateral Links: The number of links connecting two different communities. Actor in both end of these links belong to different communities. Similarly, in the presence of actor with multiple community memberships, all links from a community connecting to that actor are also consideered as bilateral links. For example, in (Figure 5.3), the red-coloured dotted links ($e.g.$, (a_3, a_{15}), (a_5, a_8), (a_9, a_{14})) are bilateral links as they are connecting two communities. Likewise, links including (a_5, a_6), (a_6, a_{15}), (a_6, a_{14}), (a_3, a_6) are

considered as bilateral links since these contain an actor with multiple memberships at one end of them. If C_i and C_j are two communities in a SIN g_t, then a bilateral link between these two communities is denoted by $e_{g_t}^{i,j}$; $i \neq j$. A set of bilateral links between two communities C_i and C_j is denoted by $\left|E_{g_t}^{i,j}\right|$.

5.2.4.3 Actor Connectivity: The actor connectivity between two actors a and b in a SIN g_t at timestamp t is the total of minimum number of actors and links that must be removed to disconnect all paths from actor a to b. If $E_{g_t}^{c}(a,b)$ denotes the set links and $V_{g_t}^{c}(a,b)$ denotes the set of actors of minimum cardinality such that, when removed, would severe off the connectivity between actor a and b then the actor connectivity between actor a and b is defined as:

$$\lambda_{g_t}^{a,b} = \left|E_{g_t}^{c}(a,b)\right| + \left|V_{g_t}^{c}(a,b)\right| \qquad \ldots (5.14)$$

Larger value for $\lambda_{g_t}^{a,b}$ denotes that there are many different alternative paths in a SIN g_t are defined to maintain the connectivity between actor a and b.

To measure similarity/proximity between non-connected actor pairs using temporal community-aware network structural information in regards to the aforementioned concepts, three different contexts were taken into consideration. Firstly, if both actors belong to the same community within a SIN; then their similarity score for that SIN is strengthen by the rate of clustering tendency of their common neighbours within the same community. However, the score is weakened by a dividing factor that represents the clustering tendency of the common neighbours residing in other communities different from the community where the corresponding actor-pair belongs. The assumption here is that if more neighbours of the common neighbours, incident to a non-connected actor pairs, performs triadic closure then the possibility of that actor-pair to close the triangle between them is amplified and so as

the probability of forming link between them. Valverde-Rebaza and Lopes exploited a similar concept where common neighbours within the same community strenghten twice more in the similarity/proximity score [282]. Secondly, if both actors in a pair reside in different communities, then the similarity score between them is computed considering the number of peripheral actors, bilateral links, path length between actors and their actor connectivity score. Finally, if there is no path defined between a pair of actors residing in different community within any SIN G_t, then a score of zero is assigned to denote their proximity in that particular SIN.

Considering $C_i(g_t)$ denoting the ith community and $\eta_a^{C_i}(g_t)$ the neighbourhood of actor a in a SIN g_t at timestamp t, and the aforementioned three different contexts, the final similarity metric using community related and network structural information in every SIN is defined as follows:

$$sim_8(a,b) = \begin{cases} \sum_{t=1}^{T} \frac{\sum_{x \in \eta_a^{C_i}(g_t) \cap \eta_b^{C_i}(g_t)} CC_{g_t}^{x}}{\sum_{j=1, j \neq i}^{n} \sum_{y \in \eta_a^{C_i}(g_t) \cup \eta_b^{C_i}(g_t)} CC_{g_t}^{y}} & if\ a,b \in C_i(g_t) \\ \sum_{t=1}^{T} \left[\left|\mathrm{V}_{g_t}^{\mathrm{i.j}}\right| + \left|\mathrm{E}_{g_t}^{\mathrm{i,j}}\right| + \frac{\lambda_{g_t}^{a,b}}{\left|\mathrm{p}_{g_t}^{\mathrm{a,b}}\right|} \right] & if\ a \in C_i(g_t), b \in C_j(g_t), i \neq j \\ 0 & if\ a \in C_i(g_t), b \in C_j(g_t), i \neq j,\ \ p_{g_t}^{a,b} = \emptyset \end{cases}$$

... (5.15)

Considering equation (5.15), if two actors belong to the same community in a SIN g_t at the timestamp t; then the similarity between them is increased by the increasing rate of clustering tendency of the intra-community common neighbours incident to both actors but decreased by the clustering tendency of inter-community neighbours of them who belong to

other communities. The assumption here is, if neighbours of the common neighbours, incident to non-connected actor pair within the same community, tend to close traingles, then the possibility of forming links between them is enhanced. Conversely, if they belong to different communities, then similarity is calculated as the total of the number of peripheral actors, bilateral links and actor connectivity score for the actor-pair in conjunction with the inverse of the geodesic distance between both actors. The assumption here is, in regards to social network structure, the peripheral actors are considered as intercessor or negotiator between two distant actors, bilateral links signify the common attributes or properties between communities. Further, the higher the actor connectivity between non-connected actor pairs the higher the probability of emerging links between them since there are more possible ways actors can reach each other. On the other hand, the connectivity score is undermined by the length of the geodesic distance between the corresponding actors. The rationale behind this part of the equation is that despite higher connectivity score, if the corresponding actors reside in the furthest corner from each other, then the possibility of forming link between them is demeaned.

5.3 Conclusion

Since most networks inherently evolve over time, it is imperative to delve into temporal networks and network dynamics to resolve issues with link prediction problem in dynamic networks [285]. Since most future links emerge between similar actors, this research computed the similarity between actors in regards to their structural, neighbourhood and community-aware evolutions, measured by three different dynamicity values. To develop the dynamic features by considering evolution similarity, these three actor-level dynamicity measures were leveraged to quantify the similarity/proximity between actors in dynamic network. Dynamic programming-based temporal similarity measures (i.e., Dynamic Time

Warping), and Pearson correlation measures were applied to develop the first six dynamic similarity metrics. The seventh dynamic feature was constructed by considering a similarity measures widely used in ecology. In this measure, we quantify the normalized abundance of actor-level dynamicity in temporal networks. Finally, with the help of two different existing community detection algorithms, by integrating evolutionary community-aware topologies in conjunction with both inter and intra-community network structure, the last dynamic feature was developed. These dynamic features will describe the instances of the classification datasets in a supervised link prediction setup to train different classifiers. The traditional topological similarity measures (e.g., common neighbours) were avoided due to their incompetency and susceptibility to attenuation in link prediction task. For example, two non-connected pairs of actors can demonstrate different likelihood of future link formation having similar measurement in regards to neighbourhood based topological metrics. Similarly, although there is an abundance of topological metrics; however, selecting the right one that fits in the corresponding study's context is a challenging task. Despite researchers attempted to build ensemble of topological metrics to alleviate the critical issue, however, ensemble-based methods are always considered computationally intensive. Further, very few studies have considered actor-level attributes in dynamic link prediction task, let alone their evolutionary features. In contrast, this study developed dynamic features relying on actor-level evolutionary aspects where both temporal and time-aware features were considered. In conjunction, associated actor-level network structural features were also considered. These features will be beneficial in dynamic link prediction task where actor-level attributes (e.g., age, role, gender) are not available. Further, these features will support the quantification of dynamic behaviours demonstrated by actors in dynamic networks.

Chapter 6

Datasets and Experimental Settings

6.1 Introduction

In the previous two chapters, I described conception of two major objectives of this thesis: firstly, determination of optimal sampling interval/window size to discretise a dynamic network including some evaluative test measures to validate the optimality of time scale, and secondly, feature engineering to develop dynamic features by mining different actor-level temporal evolutions for dynamic link prediction purpose. These features denote the evolution similarity scores those will be used in supervised link prediction. In this chapter, I describe my empirical experimental settings including detail description of the dynamic network datasets, supervised link prediction setup, performance measurement metrics.

6.2 Network datasets

To Collect dynamic network datasets, this study used both 'KONECT Network Dataset'[286] (i.e., the Koblenz Network Collection) and 'Network Repository' [287]. KONECT project is run by Institute of Web Science and Technologies at the University of Koblenz as part of collecting large network datasets to facilitate research in network science and related fields and Network Repository is considered as the first and the largest interactive repository of network datasets. The first dynamic network dataset comes from a reality mining project at Massachusetts Institute of Technology (MIT) in 2004 where the actors were tracked with the help of their personal smart phones to study interpersonal interaction. In this undirected network an actor in the network represents a person and a link indicates a physical contact among two persons. The second dataset comes from internal email communications among employees of a mid-sized manufacturing company where actors represent employees and links represent individual emails between two employees. The next dataset contains an undirected network data from a Facebook-like social network originated from an online community for students at University

Table 6.1: Basic statistics of the dynamic network datasets used in this study. The actors and links denote the total unique number of actors and links found in the entire network. Temporal fluctuations of the quantity of actors and links occur in each temporal network snapshot of the network known as Short Interval Network (SIN). From the link prediction perspective, the total duration of the time-resolved network, data were split into two non-overlapping intervals (i.e., training and test). The start and end denote the beginning and end of each interval. Nine different sampling intervals (i.e., duration length/time scale of SINs) were used and the optimum was singled out from these time-scale durations.

Network	**Actors**	**Links**	**Training Duration yyyy/mm/dd (hh:mm)**		**Test Duration yyyy/mm/dd (hh:mm)**	**Temporal Granularity**	**Sampling Window Sizes**
G_{MIT}	96	2539	Start	2004/09/14	2005/02/01	day	1-7, 14, 30
			End	2005/01/31	2005/05/05		
G_{Email}	167	3250	Start	2010/01/02	2010/08/01	day	1-7, 14, 30
			End	2010/07/31	2010/09/30		
G_{UCI}	1899	13838	Start	2004/03/24	2004/06/01	day	1-7, 14, 30
			End	2004/05/31	2004/10/26		
G_{FF}	11715	34539	Start	2007/01/01	2007/04/01	day	1-7, 14, 30
			End	2007/03/31	2007/04/30		
G_{INF}	801	2631	Start	2009/04/28 10:03	2009/05/01 10:05	minute	30, 60, 90, 120, 180, 240, 360, 480, 720
			End	2009/04/30 17:54	2009/05/01 18:06		
G_{HT}	113	2196	Start	2009/06/29 06:00	2009/07/01 6:11	minute	30, 60, 90, 120, 180, 240, 360, 480, 720
			End	2009/06/30 11:51	2009/07/01 4:59		

of California, Irvine, where actors represent students within the community and a link represents that two students communicated via a message. The fourth undirected network dataset is a very small subset of total 'Facebook' friendship graph where an actor represents a Facebook user and a link represents a friendship between two users.

To collect more dynamic network datasets, I also attempt to capitalize on different efforts that were made to mine behavioural networks of direct interactions between individual actors, in two real-world events that included temporal settings [288]. The first event is the INFECTIOUS exhibition that was held at the Science Gallery in Dublin, Ireland. The exhibition held from 17/04/2009 to 17/07/2009. The second event was the ACM Hypertext conference, arranged by the Institute for Scientific Interchange (ISI) Foundation. It was held in Turin, Italy, from 29/06/2009 to 29/07/2009. These two events generated networks of proximity or interactions where in INFECTIOUS, the network of interactions was constructed among museum viewers. In the ACM conference, the network was generated based on the proximity of conference participants. Therefore, in these two networks actors are museum viewers and conference participants respectively and a link represents a contact or physical proximity between two actors. For the sake of brevity, we name these seven networks as G_{MIT}, G_{Email}, G_{UCI}, G_{FF}, G_{INF},and G_{HT} to denote the network originated from MIT reality project, small manufacturing company, University of California Irvine, real Facebook Friendship, INFECTIOUS exhibition, and ACM Hypertext conference respectively in the rest of the study. In the first network datasets, the links are date stamped with individual dates and the smallest temporal granularity of these networks is a day. On the other hand, in $G_{INF}$and G_{HT}, the links are time stamped where the smallest temporal granularity is minutes. Therefore, the first four dynamic networks are sampled using a single day and multiples of this. In contrast, the final two networks were discretised using temporal window size of 30 minutes and multiples of it. Table (6.1) sets out the basic statistics of these network

datasets. In this table, the numbers of unique actors and links in each network dataset are presented including their total temporal duration that was split into two non-overlapping intervals from link prediction perspective. In dynamic network, actors' participations and link formations vary temporally. So the same link may appear multiple time times in different temporally sampled short interval networks (SINs). For example, in the dynamic network dataset collected from MIT realty project, the total number of links within the temporal duration is 1086403 where as the number of unique links, as presented in the table, is 2539. This means that a lot of links appeared and disappeared over time. I have also presented the unit of temporal granularity (i.e., smallest temporal unit) for each network and nine different sampling window sizes from where the optimal one was determined by applying the method proposed in chapter three. For example, the smallest temporal granularity in G_{Email} is a day and I used nine different temporal window sizes to sample G_T(i.e., network in the training phase) and these are one, two, three, four, five, six, seven, fourteen and thirty days. The one, fourteen and thirty days were selected to emulate daily, fortnightly and monthly dynamic network. Similarly, in G_{INF}, the smallest temporal granularity is minute and thus I used half hour (i.e., 30 mins), hourly (i.e., 60 min), one and half hour (90 min), two hours (120 min) and so on. The final temporal duration or time scale was 12 hours (720 min). It is noteworthy here, that, although the duration of the INFECTIOUS exhibition was three months; however, for the temporal analysis's sake (since the temporal granularity considered is minute), I considered the specified durations as mentioned in the Table (5.1) for G_{INF}.

6.3 Supervised Link Prediction

Link prediction strategies, using network structural pattern, can be predominantly categorized into two categories: (i) unsupervised and (ii) supervised. In unsupervised approach, a non-connected actor-pair is chosen first and then get assigned a score based on the chosen metric

or feature. After assigning scores, all such actor-pairs are ranked according to the scoring scheme (e.g., number of common neighbours) and then top-*L* ranked pairs are considered as predicted links. In contrast, supervised methods for link prediction problems need to predict emerging links by successfully discriminating positive and negatively labelled links within a classification dataset where instances are described using a set of features. Different advantageous of supervised link prediction were reported over its unsupervised counterpart. For example, firstly, what should be the optimal value of top *L* or how many top ranked links should be considered as future probable links, and secondly, since the ranking is performed in decreasing order which denotes that only the actor-pairs with high score of the chosen metric will form the emerging links which is not always true in the real world [289]. Further, a high score in one chosen metric does not necessarily imply a high score in an alternative measure. Lichtenwalter et al. also pointed out supervised link prediction approach as expedient over unsupervised approach [2]. According to the authors, in supervised approach, learning algorithms are competent in capturing the interacting relationships among different structural properties of features. In addition, it is also considered as adaptive in comparison to unsupervised approach which is invariant in nature. Further, in supervised approach, a classifier, trained by using a single unsupervised method, can outperform the performances demonstrated by the ranking scheme of the corresponding method.

Supervised link prediction approach is also considered as a binary classification task. This approach learns and differentiates between positive and negative instances with the help of interesting features describing all instances. Considering the aforementioned dominance of, I considered to exploit it with the help of dynamic features. In a supervised link prediction setup, the total duration of the time-resolved network is partitioned into two non-overlapping sub-intervals. The first interval is considered as the training phase and the second one is known as the test phase. Network, link structure, actor attributes and topological features

during the training interval are analysed to predict links in the test phase. After selecting these sub-ranges of duration, the classification datasets are constructed to be used in different learning algorithms. A classification dataset consists of a set of actor-pairs that appeared in the training phase but did not form links between then during the test phase. Each pair in this set is either labelled as positive or negative sample depending on the existence of links between them in the test phase. If a pair has a link formed during the test interval then it is labelled as positive sample or negative otherwise. The classification model for supervised link prediction problem predicts future links by successfully distinguishing the positive samples from the negative ones in the classification dataset. Thus, it is considered as a binary classification task. Each actor-pair in the classification dataset is described by a set of features learnt by a supervised learning framework [290].

In this research, classification datasets were built for all network datasets. These classification datasets consist of positive and negative sample instances where each instance is a non-connected actor-pair found in the training phase but did not form links in the test phase. Since in the real-world evolutionary networks, the number of links that has actual physical existence is trivial in comparison to the all potential links, the supervised link prediction problem suffers from the class imbalance problem. To be more precise, in dynamic network, although there exists an enormous number of potential links between actors that can be formed during the test phase in supervised link prediction problem; however, only a trivial portion of those probable links physically occur in real life leaving the number of positively labelled sample instances in the classification dataset easily outnumbered by the total number of negatively labelled instances. This phenomenon is known as the class imbalance problem [291-293]. In pattern classification problem, class imbalance is a fundamental problem where number of training instances of a minority class is much smaller than the number of instances of other majority classes.

There are predominantly two categories of learning algorithms for class imbalance problem: (i) resampling and (ii) cost-sensitive based [294] Over-sampling and under-sampling are two resampling methods that attempt to obtain a more balanced number of instances for both minority and majority classes by modifying their prior probability. The under-sampling method is suitable for large–scale applications where it extracts a smaller set of majority class instances, while maintaining the minority instances. Reducing the number of

Predicted Link / Actual Link	TRUE	FALSE
TRUE	True Positive (**TP**)	False Negative (**FN**)
FALSE	False Positive (**FP**)	True Negative (**TN**)

$$Precision = \frac{TP}{TP+FP} \qquad Recall/True\ Positive\ Rate = \frac{TP}{TP+FN}$$

$$False\ Positive\ Rate = \frac{FP}{FP + TN}$$

Figure 6.1: Standard confusion matrix used in the evaluation of supervised link prediction performance (i.e., binary classification model)

training instances boost the training time and make the learning problem more controllable [295]. In contrast, the over-sampling technique increases the number of minority instances. Despite minority instances are over-represented, the principal advantage of this method is that no information is lost from the training samples [291]. On the other hand, cost-sensitive based techniques assign different costs to errors in different classes [296]. Although, by these methods, classification accuracy can be achieved for minority classes; however, subsequent

inclusion of cost function in the learning process will alter the initial probability distribution [297].

From the aforementioned description, to minimize the effect of class imbalance in the classification dataset, by following Choudhury and Uddin, I restricted the workload ratio of positive vs negative instances in each network as 1:5 [298]. In this regard, it is noteworthy here that although, Lichtenwalter et al. asymptotically imposed a higher limit on the ratio between the positive and negative samples in a classification dataset [2]; however, many studies [299,92,220,298] tend to restrict the ratio of positive and negative samples in supervised learning approach to a certain degree (e.g., 1:5, 1:10). Further, I used synthetic minority over sampling technique (SMOTE) [300] algorithm that over samples the minority class in classification problem by creating 'synthetic' samples rather than over-sampling with replacement. The technique is found to be efficient to boost classification performance. SMOTE generates synthetic examples by operating in the feature space rather than the data space. It oversamples the minority class by considering each instance from this class and introducing synthetic examples along the line segments that join any/all k-nearest neighbours of that instance. It provided a new approach to over-sampling that can improve the classifiers performances for minority class. Chawla et al. also demonstrated that a combination of SMOTE-based over-sampling the minority class and under-sampling the majority class can achieve better classifier performances.

6.4 Performance Evaluation

In supervised link prediction problem, performance evaluation metrics are broadly categorized into two classes. These are (i) fixed threshold metrics (e.g., accuracy, precision and recall) and (ii) k-equivalents, threshold curves (e.g., receiver operating characteristics (ROC) curve, precision-recall/P–R curve) and the area under the ROC curve (AUCROC) or

P-R curve (AUCPR) [301]. Generally, in data mining and machine learning research domain, *precision* and *recall* are the two widely used metric where precision denotes the fraction of predicted links those are actually true (i.e., physically appear in the test phase) and recall denotes the fraction of true links those are predicted. For example, if an supervised link prediction technique predict 10 instances in the classification dataset as true links appearing in future and out of the 10, only five links are actually found in the test phase then the precision is 50 % ($\frac{5}{10} * 100$). In case of recall, if there are actually 12 true links and supervised link prediction technique has predicted five of them then it is considered as 41% recall score is achieved ($\frac{5}{12} * 100$). A standard confusion matrix, presented in Figure (6.1), is generally used to calculate these measures. In this figure precision, recall (i.e., True Positive Rate) and False Positive Rate metrics are represented by the elements of the confusion matrix. An ROC curve is a two-dimensional plot where the true positive rate is plotted on the Y-axis and the false positive rate is plotted on the X-axis to show the relative trade-offs among the two class values (i.e., positive vs. negative).

As mentioned earlier, I used dynamic similarity metrics as dynamic features, as constructed in chapter 5, to describe both positively and negatively labelled instances (i.e., actor pairs) in the classification datasets. Dynamic feature values were normalized such that the distribution has zero mean and one standard deviation. To measure classification performances, accuracy score (i.e., a 10-fold cross-validation and the mean scores), AUCROC (Area Under Receiver Operating Characteristics Curve), and AUCPR (Area Under Precision-Recall Curve) were used. While the AUCROC measure is the de-facto standard for measuring supervised learning based classification, AUCPR is reported for a more differentiated view in regards to the learning task in imbalance dataset. Despite its criticism [301], AUCROC is a popular metric (after accuracy) used in binary classification. Accuracy

only classifies the class label right or wrong; however, AUCROC quantifies the uncertainty associated with classifiers by introducing a probability value. As an important traditional measure, AUCROC score is interpreted as the probability that a randomly chosen missing link (i.e., link to be predicted) in the test phase belonging to G_{T+1} is given higher probability score than a randomly chosen non-existent link absent both in the training G_T and test network G_{T+1}. The formula to calculate AUCROC is defined as $AUCROC = \frac{n' + 0.5n''}{n}$ where n denotes the number of independent comparisons, n' denotes the times where a missing link in the test network has been given a higher score and n'' denotes the times where a non-existent link has been given a higher score.

AUCROC curve demonstrates how the number of correctly classified positive examples varies with the number of incorrectly classified negative examples and shows an overly optimistic view of an algorithm's performance. In contrast, the area under precision recall (P-R) curve (i.e., AUCPR) often serves as a summary statistics while comparing the performances of several different algorithms. The minimum value of AUCPR can be determined as $AUCPR_{min} = 1 + \frac{(1-\pi)\ln(1-\pi)}{\pi}$ with skew $\pi = \frac{positive\ samples}{n}$ where n = total number of samples in the classification dataset [299]. According to this equation, considering the ratio of positive and negative samples as 1:5 (i.e., the ratio of positive and negative samples is 1:5 in this study) in the classification datasets of G_{MIT}, G_{Email}, G_{UCI}, G_{FF}, G_{INF} and G_{HT} with the value of the skew π = 0.167, the minimum value of AUCPR in these datasets should be 0.09.

For comparison sake, I also compared the performances of dynamic features with a well-known metric 'ResourceAllocation[1]' [302] which is widely used for link prediction purpose in static network and demonstrated improved performance. I also implemented the

[1] Appendix A

link prediction strategy in dynamic networks proposed by Soares & Prudêncio where the authors built time series of traditional topological metrics (e.g., Jaccard Coefficient) for non-connected actor pairs for each SIN in the training phase and used time series forecasting method (e.g., ARIMA) to predict the final score of the topological metrics and used those forecasted values to train the classifier [303]. Different variations of this method are also extensively followed by other authors to support link prediction in dynamic networks [1,196]. For the sake of brevity, in the rest of the study, we used $sim_{RA}(a, b)$ and $sim_{\text{Soares}}(a, b)$ to denote the values computed for the positively and negatively labelled actor-pairs considering 'ResourceAllocation' metric and dynamic link prediction strategy proposed by Soares & Prudêncio. It is noteworrthy that to compute sim_{Soares} we have considered the well-known 'Jaccard Coefficient[1]' measure as the topological similarity metric and used ARIMA forecasting method to predict the future values of the selected metric incident to actor pairs.

6.5 Conclusion

In this chapter, the experimental settings, dynamic network datasets, supervised learning setup and performance measurement metrics of dynamic link prediction task by using dynamic features (described in chapter 5) were discussed. In supervised link prediction setup, a classification dataset is constructed where each instance in the dataset is an actor pair with either a positive and negative label depending on their formation of a true link. Each instance is described by useful features where a learning algorithm is employed to classify each instance correctly and turning it into a binary classification problem. This research developed features by mining temporal evolutions experienced by each actor pair. Each feature score denotes the evolution similarity between a pair of actors. Temporal evolution rate varies in dynamic networks and the measurements of actor-level dynamicity greatly depend on the

[1] Appendix A

temporal sampling of the corresponding network. Therefore, it is imperative to define an optimal sampling scale to discretise the network. In the next chapter, I describe the results of experiments to determine the optimal and near optimal sampling resolution for each dataset described in this chapter. It is noteworthy that to determine the optimal window size, I used the candidate window sizes for each network described in the table of this chapter. Although, these sizes are defined for the sake of experimental purpose, however, candidate windows can be of any size irrespective of the duration. The candidate windows were chosen in such way so that each network snapshot has sufficient (at least one) link(s) without any loop or duplicity. Further, it is not optimal to select nonconforming temporal windows. For example, if a dynamic network consists of aggregating links over day then the selection of seconds, minutes or hours as candidate window sizes will be inappropriate. Similarly, if a dynamic network consists of the aggregation of links by microsecond then selection of a day, or its multiples as candidate windows will also be inappropriate. Once optimal temporal window is identified for each dataset then in the next chapter, I describe the results of dynamic link prediction over each network by using the dynamic features constructed in optimally sampled network.

Chapter 7

Optimal Temporal Scale in Dynamic Networks: Empirical Results

7.1 Introduction

Constructing time-aware dynamic features largely depends on the optimal length of temporal duration for each SIN since, as described in chapter 4, the structure of the network greatly varies by the total number of aggregated links within a particular timeframe. As the network structure varies due to the size variations of the link aggregation window in dynamic networks, so do the actor-oriented network measures and associated evolutionary information. As both these aspects are used in developing the dynamic features/dynamic similarity metrics, as described in the previous chapter (please see chapter 5 for details), it is imperative to define the optimal/appropriate time-scale duration for network snapshots in a given dynamic network.

To recap the problem formulation from the introduction chapter, in dynamic link prediction task, a finite set of discrete time points are considered as $T = [t_1,\ (t_1 + \tau),\ (t_1 + 2\tau) \ldots\ (t_1 + n\tau) \ldots\ (t' - \tau),\ t']$ where τ denotes the temporal sampling interval (i.e., time scale). A dynamic network $G_T = (V, E_T)$ consists of a set of uniquely labeled actors $V = [v_1, v_2, v_3, \ldots v_n]$ and $E_T = [e_t(v_i, v_j, t) | v_i, v_j \in V; t \in T]$ where t represents the timestamp of link e between actor-pair $e(v_i, v_j)$, is composed of an evolutionary sequence of network snapshots $G_T = [G_{t_1},\ G_{t_1+\tau},\ G_{t_1+2\tau}\ \ldots\ G_{t_1+n\tau} \ldots G_{t'-\tau},\ G_{t'}]$ where each G_{t_i} is known as short interval network (SIN). Fluctuations of the total number of actors are taken into consideration across the time series of network snapshots. Any link may appear in multiple network snapshots at different timestamp(s). Considering this temporal sequence of network snapshots $[G_{t_1},\ G_{t_1+\tau},\ G_{t_1+2\tau}\ \ldots\ G_{t_1+n\tau} \ldots G_{t'-\tau},\ G_{t'}]$, for a given pair of actors $(v_i\ , v_j)$, dynamic link prediction attempts to predict the likelihood of link formation between them during the interval (t', t_1') in G_{T+1} by analysing the link formation and temporal information in $[G_{t_1},\ G_{t_1+\tau},\ G_{t_1+2\tau}\ \ldots\ G_{t_1+n\tau} \ldots G_{t'-\tau},\ G_{t'}]$ at timestamps $[t_1,\ (t_1 + \tau),\ (t_1 +$

2τ) ... $(t_1 + n\tau)$ ... $(t' - \tau)$, $t']$. Further, as evident from the list of features constructed in chapter 6, each feature contains the time information t that generally denotes each timestamp $t_1 + n\tau$ (i.e., initial time + n^{th} temporal duration τ). Therefore, before exploring the results of dynamic link prediction task using the dynamic features, in this chapter, I describe the result of experiments performed to determine the optimal time-scale for each SIN in G_{MIT}, G_{Email}, G_{UCI}, G_{FF}, G_{INF},and G_{HT}. As mentioned earlier, to split each network G_T and generate time series of SINs, I used the method described in chapter 3 to determine the optimal temporal window size for each dynamic network considered in this research.

7.2 Determination of Optimal Time Scale

In this section, I describe the results of the optimal time scale determination method, described in chapter 3, applied over six different network datasets, which will help me to identify the most and second most appropriate (i.e., optimal and near optimal) temporal window size to sample or discretise the each dynamic network.

Table 7.1: Variances of actor-level positional dynamicity values in each dynamic network dataset sampled by considering nine different window sizes. The green-shaded cell represents the smallest value and according to the algorithm developed in chapter 3, denotes the best optimal window size in the respective dataset and the yellow-shaded cell represents the second best optimal window size for each dynamic network.

Dataset	**Window Size (days)**								
	1	**2**	**3**	**4**	**5**	**6**	**7**	**14**	**30**
G_{MIT}	0.0110	0.0104	0.0098	0.0090	0.0081	0.0076	0.0070	0.0055	0.0051
G_{Email}	0.0177	0.0170	0.0210	0.0196	0.0171	0.0156	0.0143	0.0099	0.0063
G_{UCI}	0.0042	0.0037	0.0028	0.0027	0.0027	0.0025	0.0021	0.0015	0.0023
G_{FF}	0.0019	0.0057	0.0104	0.0145	0.0186	0.0219	0.0236	0.0256	0.0273
	Window Size (minutes)								
	30	60	90	120	180	240	360	480	720
G_{INF}	0.0002	0.0004	0.0006	0.0010	0.0015	0.0017	0.0021	0.0024	0.0036
G_{HT}	0.0160	0.0167	0.0151	0.0133	0.1001	0.0078	0.0058	0.0055	0.0015

7.2.1 Optimal Window Size

I first performed the variance analysis of actors' positional dynamicity values considering nine window sizes for each datasets. In Table (7.1), I present the results of the algorithm developed in chapter 3 that determines the optimal window size to discretise dynamic network by analysing the variance analysis of actor-level positional dynamicity values. It is noteworthy that window size that represents the smallest variance in positional dynamicity values is considered as the optimal temporal scale. In this table, the green-shaded cell in each dataset represents the lowest variance in the distribution of actor-level positional dynamicity values and thus denotes the optimal sampling scale. On the other hand, the yellow-shaded cell

denotes the second best optimal scale having the second lowest variance. From this table, it is evident that in G_{MIT}, and G_{Email}, networks, the optimal window size is 30 days (i.e., monthly window). The second best window size for these two networks is 14 days. Therefore, according to the proposed algorithm, it concludes that monthly or fortnightly SINs will suffice to analyse these dynamic networks. On the other hand, in G_{UCI}, the proposed approach identified 14 days (fortnightly window) as the optimal time-scale duration to generate network snapshots. The second best optimal window size in this network is seven days (i.e., weekly) window. In case of G_{FF}, the daily window (i.e., one day) was identified as the optimal time-scale duration with the lowest variance of actor-oriented positional dynamicity values reported. The second best window size in G_{FF} is two days. In case of the other two dynamic network datasets (i.e., G_{INF} and G_{HY}), where the unit of temporal granularity is minute, two different temporal-scale durations were identified as their optimal window sizes. In G_{INF}, the optimal time-scale duration is half an hour (i.e., 30 minutes) and in G_{HY} it is 720 minutes (i.e., 12 hours). Consequently, hourly window (i.e., 60 minutes) and eight-hour window (i.e., 480 minutes) became the second best time-scale durations for them respectively. A general phenomenon observable from the table is that higher number of actors may leads towards lower-scale temporal duration. For example, in case of the dynamic networks with lowest link aggregation duration as day (G_{UCI}, G_{MIT}, G_{Email} and G_{FF}, the highest number of actors were found G_{FF} (i.e., 11715 in dataset description table from chapter 6) and the optimal temporal scale in this case was identified as the daily (i.e., one day) window. On the flip side, the lowest number of actors was found in G_{MIT} (i.e., 96) and in this case the optimal temporal scale was identified as the monthly window (i.e., 30 days). Similarly, in case of the networks where the lowest temporal granularity is minute (i.e., G_{INF} and G_{HT}), the optimal temporal granularity in G_{INF} is 30 minutes (i.e., half hour) which is lower than the temporal granularity of 720 minutes (i.e., 12 hours) identified in G_{HT}, although

the number of actor in the later (i.e., 113) is lower than the former (i.e., 801). This fact may not be necessarily true in all respects. This phenomenon solely depends on the rate of network activities (i.e., formation/deletion of links) demonstrated by the actors in dynamic networks over time. For example, as presented in the table of basic network statistics in chapter 6, the number of unique links in G_{MIT} was 2539 whereas the total number of links found within the temporal duration specified in the table was 1086403. This means that same link between a pair of actors was found in multiple times since in dynamic network links appear and disappear over time. On the other hand, in G_{FF}, the number of unique links found was 34539 and the total number of links within the temporal duration for this network dataset

Table 7.2: Evaluation results to justify the optimal time-scale duration out of nine sampling window choices as per the approach presented in chapter 3 in three dynamic networks (i.e., G_{MIT}, G_{Email}, G_{UCI}). Evaluation tests include the best-fit ARIMA model, percentage of time series anomalies present (Anomaly %) in the time series of positional dynamicity of Short Interval Networks (SINs) of nine different lengths and minimum total within-cluster variance (Minimum Variance) within optimal number of clusters (# Optimal Clusters). The univariate K-means clustering method was used for distribution of positional dynamicity values of actors. The green-shaded columns denote the optimal temporal window. The yellow-shaded columns are the contenders as the second-best window(s) in the respective dataset. The red-shaded column(s) represent the contender window to be the second best optimal window choice in the respective dataset.

G_{MIT}									
Window (days)	1	2	3	4	5	6	7	14	30
Best Fit ARIMA	(3,0,3)	(0,0,0)	(1,1,1)	(0,1,1)	(0,1,1)	(0,1,0)	(0,1,1)	(0,1,0)	(0,0,0)
Anomaly (%)	8.57	4.29	10.64	5.71	7.14	8.33	0	4.81	0
# Optimal Clusters	1	1	2	2	2	2	2	2	1
Minimum variance	1.0389	0.2226	0.2767	0.2525	0.2042	0.1893	0.1999	0.1688	0.4889
G_{Email}									
Best Fit ARIMA	(1,0,0)	(0,1,1)	(0,1,1)	(1,1,1)	(3,1,0)	(0,1,1)	(2,1,0)	(2,1,0)	(0,1,0)
Anomaly (%)	18.82	12.38	14.08	16.98	18.6	19.44	19.35	10.0	12.5
# Optimal Clusters	4	5	4	4	3	3	3	3	3
Minimum variance	0.0888	0.0887	0.2115	0.2021	0.3646	0.3311	0.3208	0.2416	0.1406
G_{UCI}									
Best Fit ARIMA	(0,1,0)	(0,1,0)	(2,0,1)	(0,0,0)	(1,0,0)	(0,1,0)	(1,1,0)	(0,0,0)	(0,0,0)
Anomaly (%)	13.33	8.7	12.5	15.38	20.05	11.11	14.29	0	0
# Optimal Clusters	8	9	9	9	8	9	9	8	8
Minimum variance	0.1340	0.0866	0.0814	0.0745	0.0948	0.0782	0.0667	0.0489	0.0631

Table 7.3: Evaluation results to justify the optimal time-scale duration out of nine sampling window choices in three dynamic networks (i.e., G_{FF}, G_{INF}, G_{HT}). Evaluation tests include the best-fit ARIMA model, percentage of time series anomalies present (Anomaly %) in the time series of positional dynamicity of Short Interval Networks (SINs) of nine different lengths and minimum total within-cluster variance (Minimum Variance) within optimal number of clusters (# Optimal Clusters). The univariate K-means clustering method was used for distribution of positional dynamicity values of actors. The green-shaded columns denote the optimal temporal window. The yellow-shaded columns are selected as the second-best window(s) in the respective dataset. The red-shaded column(s) represent the contender window to be the second best optimal window choice in the respective dataset.

G_{FF}									
Window (days)	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**14**	**30**
Best Fit ARIMA	**(0,0,1)**	**(0,0,0)**	**(0,0,1)**	**(0,0,1)**	**(0,0,0)**	**(0,0,1)**	**(0,0,1)**	(0,0,0)	(1,0,0)
Anomaly (%)	**3.33**	**4.44**	**4.44**	**0**	**16.67**	**0**	**0**	**14.29**	**0**
# Optimal Clusters	**9**	**9**	**9**	**9**	**9**	**9**	**9**	**9**	**9**
Minimum variance	**0.4882**	**1.3647**	**2.1300**	**2.4751**	**2.9288**	**3.0239**	**3.4416**	2.9288	3.5974
G_{INF}									
Window (minutes)	**30**	**60**	**90**	**120**	**180**	**240**	**360**	**480**	**720**
Best Fit ARIMA	**(0,0,1)**	**(0,0,0)**	**(0,0,1)**	**(0,0,0)**	**(1,0,0)**	**(2,0,1)**	**(0,1,0)**	**(1,1,0)**	**(0,0,0)**
Anomaly (%)	**0**	**15.38**	**15.79**	**14.29**	**0**	**12.5**	**14.29**	**0**	**0**
# Optimal Clusters	**4**	**4**	**5**	**3**	**7**	**4**	**4**	**8**	**8**
Minimum variance	**0.0096**	**0.0106**	**0.0092**	**0.0316**	**0.0089**	**0.0341**	**0.0411**	0.0011	0.0015
G_{HT}									
Best Fit ARIMA	(0,1,0)	(2,0,1)	(0,0,0)	(0,0,0)	(0,0,1)	(0,1,0)	(0,0,1)	(0,0,0)	(0,0,0)
Anomaly (%)	**0**	**4.17**	**0**	**0**	**7.81**	**5.75**	**10.64**	**0**	**0**
# Optimal Clusters	**1**	**1**	**1**	**2**	**2**	**2**	**3**	**4**	**2**
Minimum variance	**1.7954**	**1.8796**	**1.6965**	**0.2981**	**0.2858**	**0.1973**	**0.1217**	0.0578	0.0403

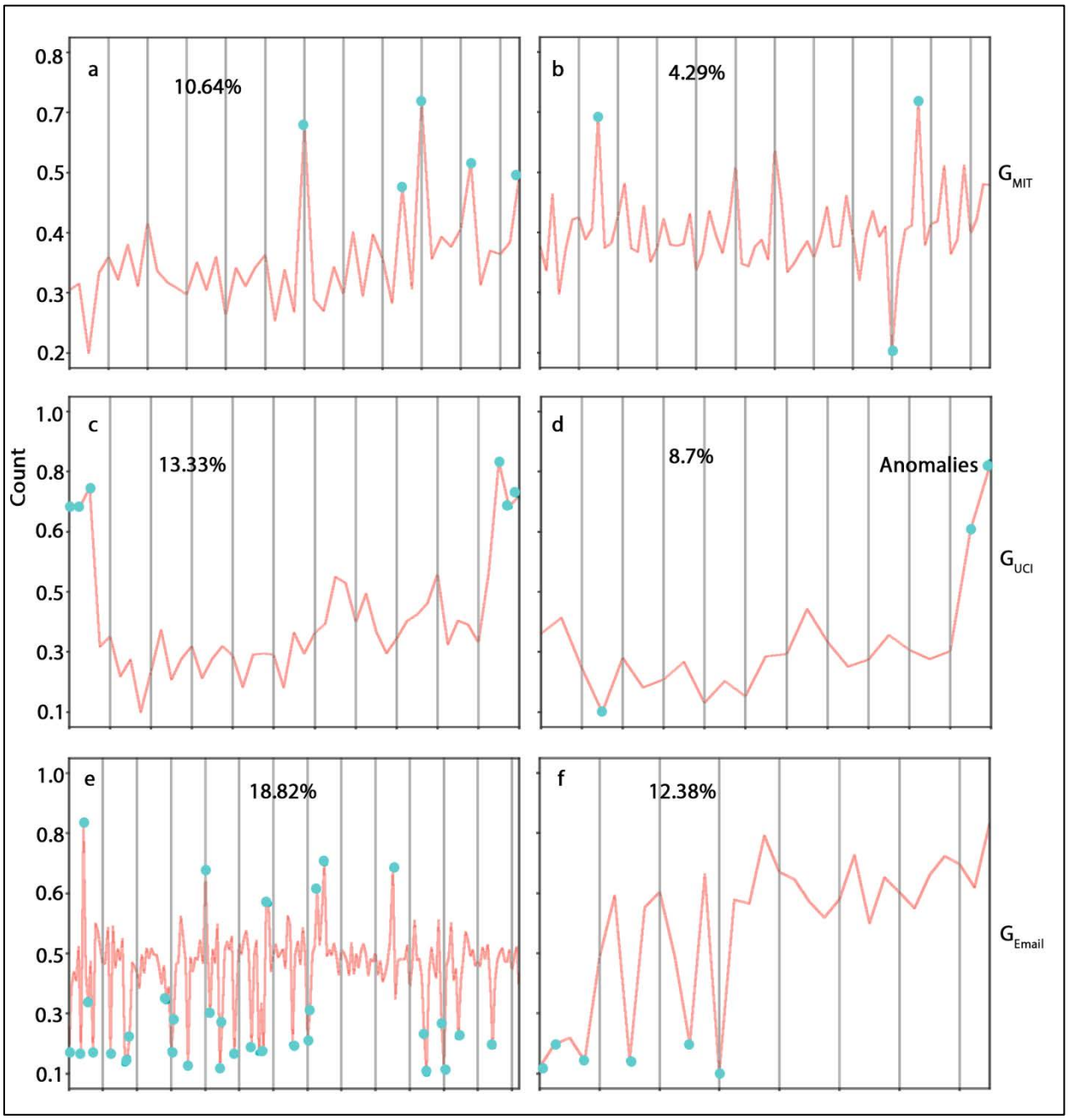


Figure 7.1: Visual presentations of the percentage of anomalies present in a time series of positional dynamicity values for every Short Interval Network (SIN). The time series were built for all SINs considering two different window sizes (i.e., time-scales) in $\boldsymbol{G}_{\mathbf{MIT}}$, $\boldsymbol{G}_{\mathbf{UCI}}$, and $\boldsymbol{G}_{\mathbf{Email}}$ networks. The respective time scales are (a) 3 days, (b) 14 days in $\boldsymbol{G}_{\mathbf{MIT}}$, (c) 1 day, (d) 2 days in $\boldsymbol{G}_{\boldsymbol{UCI}}$, and finally, (e) 1 day, (f) 2 days in $\boldsymbol{G}_{\boldsymbol{Email}}$. The blue dots represent the percentages of anomalies (numbers within each image) in the time series.

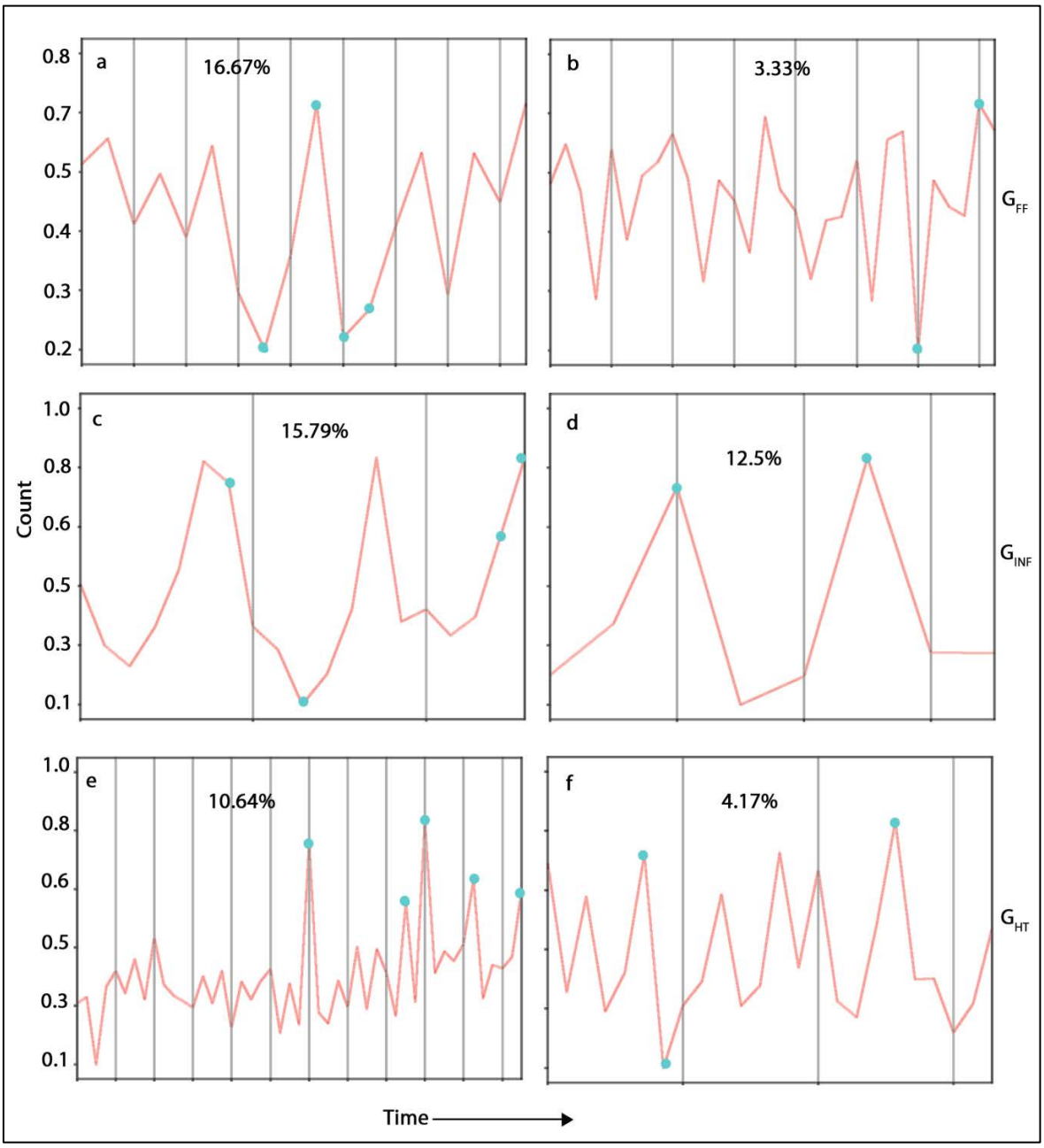


Figure 7.2: Visual presentations of the percentage of anomalies present in the time series of positional dynamicity values for every Short Interval Networks (SINs). The time series were built for all SINs considering two different window (i.e., time-scales) sizes in G_{FF}, G_{INF}, and G_{HT} networks. The respective time scales are (a) 5 day, (b) 1 days in G_{FF}, (c) 90 minutes, (d) 240 minutes in G_{INF} and finally, (e) 6 hour (360 min), (f) 1 hour (60 min) in G_{HT}. The blue dots represent the percentages of anomalies (numbers within each image) in the time series.

was 42698. It is observable that the number of network activities in G_{MIT} was exceedingly higher than the rate of network activities found in G_{FF}.

7.2.2 Optimal Window Size Validation

Once the optimal and near-optimal window lengths for each network were identified by the proposed algorithm in chapter 3, in this step, I applied the evaluation part, also described in chapter 3, to validate the identified optimality. This is performed by determining the best fit ARIMA model, percentage of anomalies in time series of positional dynamicity values of SINs and finally identifying minimum total within-cluster variance (i.e., sum of squared errors) within optimal number of clusters in K-means clustering considering positional dynamicity of actors in SINs. It is noteworthy that the standard threshold value of these validation tests were *ARIMA(0,0,0)* for the best-fit ARIMA model, lowest number of anomalies present in regards to time series anomalies and lowest intra-cluster variance in minimum number of clusters by considering K-means clustering. For details, an interested reader is referred to chapter 3). Tables (7.2 & 7.3) illustrate these validation results. In this validation phase, I used two R packages named as 'forecast' [304] for ARIMA validation and 'AnomalyDetection' [255] for the time series anomaly detection purpose. The later one is capable of detecting percentage of anomalies present in univariate time series including directionality (i.e., positive, negative or both) of anomalies. This package, with the help of Seasonal Hybrid Extreme Studentized Deviate (S-H-ESD) method, can detect maximum number of anomalies present as a percentage of the data. This percentage value was set to five. This means we consider time series with maximum five percent anomalies. Figures (7.1 & 7.2) present the visualisation of percentage of time series anomalies present in two different time series of positional dynamicity values for each SIN. Two time series of SINs in each dynamic network dataset was constructed by considering two different time scale or temporal window sizes. The blue dots represent the anomalies present in each time series. In

(Figure 7.1), two different time series for G_{MIT}, G_{UCI}, and G_{Email} were presented where one time series presents the window size containing comparably the higher amount of anomalies and the other with the lower amount of anomalies present. Similarly, (Figure 7.2), presents two different time series in G_{FF}, G_{INF}, and G_{HT} networks. On the other hand, for the evaluation task using k-means clustering, I used the associated R package [257].

From Table (7.2), it is apparent that the optimal window size of thirty days (i.e., monthly temporal window) in G_{MIT} network, as determined by the approach suggested in chapter 3, passed all the validation tests. In regards to the best-fit ARIMA model, the time series of positional dynamicity values, demonstrated by SINs where the SINs were generated by considering a temporal window size of 30 days, presents no auto-correlation between the dynamicities of SINs and associated error-values. Further, it demonstrated time series information without any trend and seasonality which was the presumption in regards to SINs being random networks. Considering the time series anomalies present in the time series of SINs' dynamicity values, it presented zero anomalies and considering univariate K-means clustering over actor-level positional dynamicity values, it demonstrated lowest optimal number of clusters with minimum total within-cluster variance. Similar to G_{MIT}, in G_{Email}, the optimal temporal window size of 30 days (i.e., monthly) was identified by the proposed algorithm. In regards to the validation test, the monthly optimal window in G_{Email}, demonstrated a weak fit in regards to the best-fit ARIMA model (i.e., *ARIMA(0,0,0)*) where the time series of SINs' dynamicity values revealed non-stationarity. This represent the presence of trends and/or seasonality (i.e., *ARIMA(0,1,0)*). Despite being poorly fit in regards to the ARIMA model, it was found to be the best one among all other temporal window choices (Table 7.2) in regards to the threshold values, considered for all the validation tests, mentioned earlier. The second best temporal scale of this dynamic network (i.e., fortnightly window) outperformed the best one in regards to the amount of anomalies present in the time

series of corresponding SINs' positional dynamicity values. Despite its poor performance, the optimal monthly window outplayed the second best window choice (i.e., fortnightly) in regards to number of optimal clusters in K-means clustering and considering the minimum total within-cluster variances. Although, both the optimal monthly and semi-optimal temporal window choices had similar optimal K-number of clusters in univariate K-means clustering; however, considering the other two tests (i.e., best-fit ARIMA, and minimum total within-cluster error rate/variances), I decided to consider it as the best window size for this dynamic network.

In G_{UCI}, the optimal window size of 14 days (i.e., fortnightly window) surpassed all the other temporal window choices. According to the evaluations presented in Table (7.2), the monthly window size demonstrated better performance than the original second best temporal window (i.e., weekly-7days) in Table (7.1). It showed better fit in all validation tests but the measurement of errors (total within-cluster variance) present in K-means clustering. Thus, considering its win in three out of four tests, in the following chapter(s), I will consider the monthly window as the second best choice to sample the dynamic network G_{UCI} rather than the weekly temporal window. A different scenario was found in G_{FF} in Table (7.3). In this dynamic network, considering the best-fit ARIMA test, the second best window outplayed the optimal one. The time series of positional dynamicity values, computed in SINs generated by the optimal daily (i.e., one day) window, demonstrated auto-correlation between the associated error terms (i.e., moving average) present in the time series information (please see chapter 3 section 4.2 for detail). In regards to the optimal number of K-clusters in K-means clustering, it also showed similar results like the second best one. Nonetheless, the daily window choice outperformed the second best choice (i.e., two days) in regards to the anomalies present in the time series information and the amount of errors present in K-means

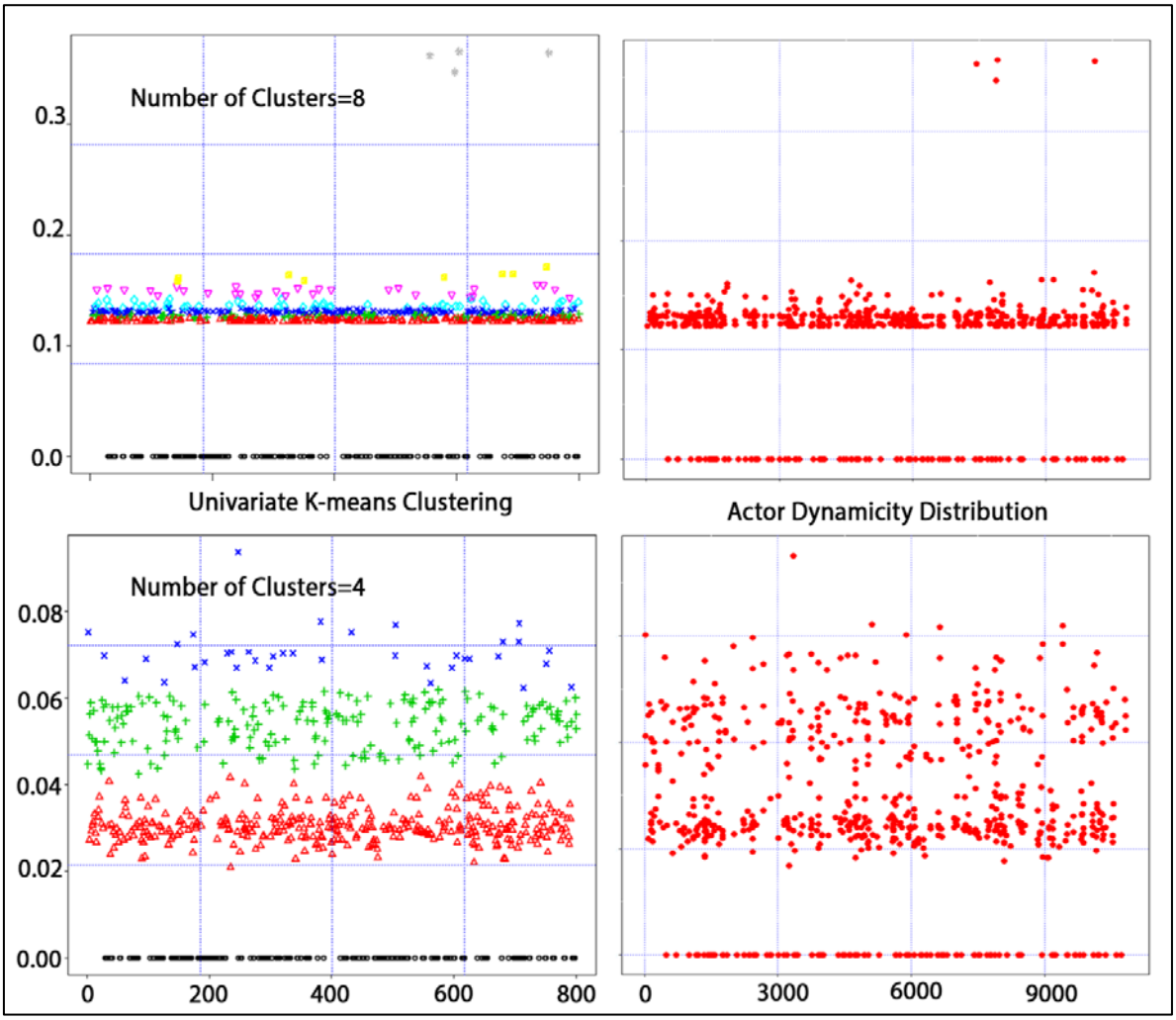


Figure 7.3: Distribution of the actors' positional dynamicity values and corresponding clusters of univariate K-means clustering in G_{INF} network considering a window size of 12 hours (720 minutes) (top row) and one hour (60 minutes) (bottom row). In each row, the left plot represents the optimal number of clusters in the univariate K-means clustering algorithm over the actors' positional dynamicity. The right plot represents the corresponding distribution of actor dynamicity values.

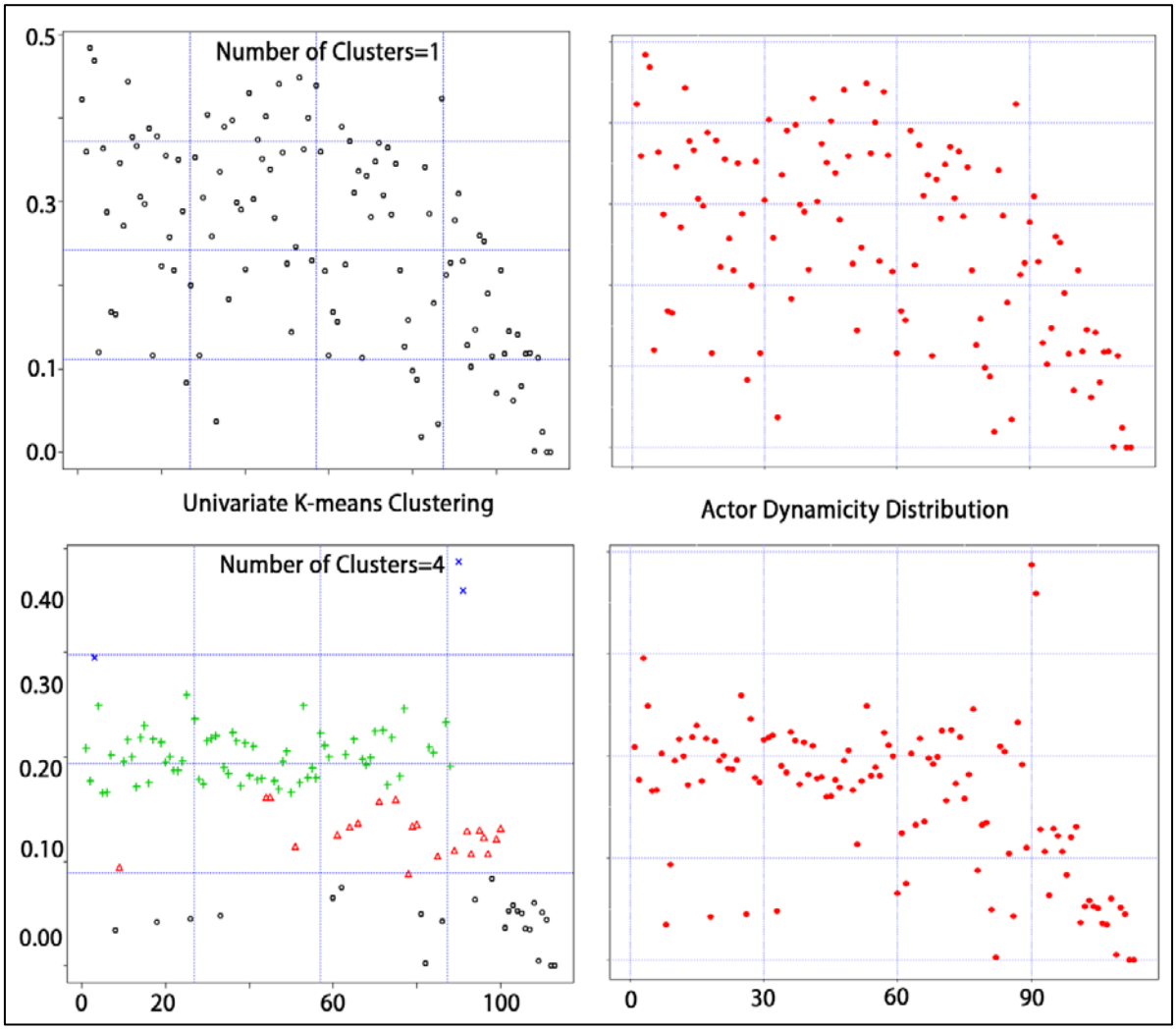


Figure 7.4: Distribution of the actors' positional dynamicity values and corresponding clusters of univariate K-means clustering in the G_{HT} network, considering a window size of 1.5 hours (90 minutes) (top row) and 8 hours (480 minutes) (bottom row). In each row, the left plot represents the optimal number of clusters in the univariate K-means clustering algorithm over the actors' positional dynamicity. The right plot represents the corresponding distribution of actor dynamicity values.

clustering (total minimum within-cluster variance). Therefore, I concluded with this window choice as the winner. In G_{INF} and G_{HT}, where the unit of temporal granularity was minute, I observed similar phenomena in both dynamic networks. In the former (i.e., G_{INF}) dynamic network, the best temporal window choice of 30 minutes (i.e., half hour) was outperformed by the second best choice (i.e., hourly window) in regards to the best-fit ARIMA model. However, in regards to the other three validation tests, I found it performing better than the others as evident in Table (7.3). Therefore, considering its win in three out of four validation tests, I considered the originally found optimal window choices as the winner. Similar to the case found in G_{UCI}, the original second best window size in this network (i.e., hourly window) was outperformed by the window size of 12 hours (i.e., 720 minutes) in all three tests except the optimal number of clusters in K-means clustering. Considering the three out four winner as mentioned above, in this case the better contender (i.e., 12 hours) will be considered as the second best temporal window choice in the following chapter(s). Similar observations were evident in G_{HT}. Although, the originally identified best/optimal window choice performed better in comparison to the others by considering all validation tests, however, the window size of 90 minutes was found to be better contender as the second best one. By the proposed algorithm, it was 480 minutes (i.e., 8 hours) was the originally chosen second-best window choice in this dynamic network. The 90 minutes window choice demonstrated better performance in regards to the optimal number of clusters in univariate K-means clustering despite having greater amount of total minimum within-cluster errors present. In Figures (7.3 & 7.4), the optimal number of clusters in univariate K-means clustering, by using '*Ckmeans.1d.dp*' algorithm[257], over actor-level positional dynamicity values and the corresponding distribution of these values are presented for dynamic networks G_{INF} and G_{HT} respectively.

Table 7.4: Number of Short Interval Networks (SINs) generated by different choices of temporal window sizes for each dynamic network used in this study. This also denotes the length of temporal network snapshots.

Dataset	Window Size (days)								
	1	**2**	**3**	**4**	**5**	**6**	**7**	**14**	**30**
G_{MIT}	140	70	47	35	28	24	20	10	5
G_{Email}	186	105	71	53	43	36	31	16	8
G_{UCI}	45	23	16	13	10	9	7	4	3
G_{FF}	90	45	30	23	18	15	13	7	3
	Window Size (minutes)								
	30	**60**	**90**	**120**	**180**	**240**	**360**	**480**	**720**
G_{INF}	50	26	19	14	10	8	7	5	4
G_{HT}	48	24	16	12	8	6	4	4	2

In Figure (7.3), two contender window sizes (i.e., 12 hours vs. one hour) were selected to compute actors' positional dynamicities. It is evident from the figure that considering 12 hours (720 minutes), the number of optimal clusters (i.e., eight) is higher than those of hourly window size (i.e., four). However, the minimum total within-cluster variance is higher in the latter. Similarly, in Figure (7.4), by considering 90 minutes, the optimal number of cluster was one whereas in case of 480minutes (i.e., 8 hours), it was four. Simultaneously, it was evident from the figure that in the latter case, the minimum total within-cluster variance is lower than the former. In G_{HT}, by considering the performances of these two window choices (90 minutes vs. 480 minutes), it is evident that both of them demonstrated equal performance in regards to best-fit ARIMA model and the amount of error present in the time series information. However, they both outperformed each other in one out of the rest two tests. Therefore, I will consider originally identified 480 minutes (i.e., 8 hours) as the second best choice in regards to the optimal temporal window selection. In Table (7.4), the final optimal and second-best optimal window choices are presented. In this

table, also the statistics in regards to the total number of SINs generated by considering all different temporal window choices are presented. In this table, the optimal time scale and the second best optimal time scale choices in each dataset are shaded as green and yellow respectively.

7.3 Conclusion

In dynamic networks, one important task is to identify the correct, appropriate or optimal choice of aggregation granularity in order to perform binning any stream of time stamped links to discern meaningful information and understand the rate of dynamics demonstrated by these networks. As identified by Fish and Caceres, researchers have named this problem differently, such as, change point detection, time scale detection, oversampling correction, temporal resolution inference, aggregation granularity detection or windowing selection [109]. This identification of correct window length strongly impacts the structural analyses, efficacy of network mining and dynamics demonstrated by networks [258,246,107]. Having too coarse or too fine temporal granularity may conceal or fail to disentangle critical information about network dynamics and impair the understanding of the structure of underlying interactions. Further, appropriate temporal binning decision in dynamic networks will enable to distinguish between noisy, local and critical temporal orderings.

The approach I followed to define the optimal time scale for dynamic networks, as described in chapter 3, is based on the concept of an actor-level dynamicity that quantifies changes in actors' network involvements (in terms of network position) during the evolution of the underlying longitudinal network. To determine the optimal window length from nine sampling resolutions in each dynamic network, I have compared the variances of nine sets of actor-level positional dynamicity values. The window length with minimum variance in actor dynamicity distributions define the appropriate sampling window to analyse the dynamic

network because the minimum variance will ensure that the suggested window size will neither be too large for some actors that reveal high rates of network activities to exhibit a large volume of network activities nor be too small for some other actors that reveal slow rates of network activities to exhibit a minimum number of network activities. In the first table of this chapter, I presented the optimal and second-best optimal sampling window sizes identified by the algorithm in each dynamic network dataset. Once identified, I have also evaluated their optimality with the help of four validation tests. The theoretical backgrounds of these tests are described in chapter 3. The threshold value for each test was described both in chapter 3 and this chapter. These values are best-fit ARIMA model closer to *ARIMA(0,0,0)* with stationary and lowest possible time series anomalies present in time series of positional dynamicity demonstrated by each SIN in a dynamic network, and lowest optimal number of clusters including minimum intra-cluster variance in K-means clustering. In two tables, I presented the validation test results for each optimal and second-best optimal window choices in each dataset. It was observed that in all cases the optimal window choices, identified by the proposed algorithm, passed maximum validations. However, in some cases, the second-best optimal window choices were subject to change due to other contenders was demonstrating better performance in regards to the validation tests. Finally, in the last table, I present the selected optimal window choices and second-best (near-optimal) window choices for each dataset including the number of SINs generated under each sampling window size.

Chapter 8

Supervised Dynamic Link Prediction: Empirical Results

8.1 Introduction

In this chapter, the results of an empirical analysis where the dynamic features (described in chapter six), applied in a supervised link prediction setup to predict links in six different dynamic networks, are presented. These dynamic features were then compared to one similarity metric, widely used in link prediction in static network, and one time series based link prediction approach to determine the superiority of the dynamic features developed in this study. Then this chapter also presents the distribution of feature values to determine whether similar or dissimilar actors in regards to their evolutionary aspects participate in emerging links. Since the dynamic features or dynamic similarity metrics denote the evolution similarity between actor-pairs, so a lower value of the feature would denote dissimilar actors and higher value denote higher evolution similarity between actor- pairs. In this chapter, I also describe the feature importance to identify which feature(s) performed better in which database in regards to the prediction task.

8.2 Preambles

Dynamic similarity metrics for link prediction task in this thesis was developed by considering the similarity and/or proximity between actors in dynamic networks regarding their evolutionary aspects. Three types of actor-level evolutionary information were defined including (i) structural, (ii) neighbourhood and (iii) community-aware dynamicity. The similarity between a pair of actors was defined by computing the temporal similarity and correlation between these evolutionary aspects. In addition, two other similarity measures were defined. The first was based on a measure used in ecology, known as Bray-Curtis Similarity measure, whereas the second considered evolutionary changes of actor-level community participation and network structure.

To use community-aware network-structural information, two different community detection techniques were used: (i) agglomerative hierarchical community detection and (ii) Louvain community detection algorithm. Table (8.1) sets out the Summary of dynamic features constructed in chapter 6. For the sake of brevity, the classification dataset for each network will be denoted as follows: $G^{\tau}_{Network}$ where τ denotes the length of the optimal time scale identified for each dataset in the previous chapter. Thus, six classification datasets, by considering the optimal time scale in each of the six dynamic networks, will be denoted by G^{30}_{MIT}, G^{30}_{Email}, G^{14}_{UCI}, G^{1}_{FF}, $G^{0.5}_{INF}$ and G^{12}_{HT} in which each network is accompanied with the identified optimal temporal window length in their respective dataset. For example, in G_{MIT}, it is the monthly window (i.e., 30 days), in G_{FF}, it is the daily window (i.e., 1 day), in G_{INF}, it is the half (0.5) an hour (i.e., 30 minutes), and in G_{HT}, it is 12 hours (i.e., 720 minutes).

The fundamental purpose of supervised link prediction in dynamic networks was to build a binary classification model and successfully differentiate between positive and negatively labelled actor-pairs. For this purpose, after developing the classification datasets, three different classifiers were considered. These included simple logistic regression, Random Forest and Bagging algorithms. In the latter two algorithms, ensemble-based methods were used. Ensemble is a machine learning concept in which the idea is to combine multiple models using the same learning algorithm, or alternatively, a set of weak learners are grouped to form a stronger learner to obtain better performance.

Table 8.1: A list of different dynamic features in which each feature computes $\boldsymbol{sim_i}$ $(\boldsymbol{a}, \boldsymbol{b})$, a similarity score between actor $\boldsymbol{a}$ and $\boldsymbol{b}$ by using different evolutionary aspects and actor-level network structures in dynamic networks.

Metrics	Equation	Description
$sim_1(a,b)$	$min\left\{\sum_{\ell=1}^{\mathcal{L}} d\left(\delta_{ml}^{a}, \delta_{nl}^{b}\right)\right.$	Temporal similarity of structural, neighbourhood and community dynamicity measured using the Dynamic Time Warping (DTW) Technique
$sim_2(a,b)$	$min\left\{\sum_{\ell=1}^{\mathcal{L}} d\left(\lambda_{ml}^{a}, \lambda_{nl}^{b}\right)\right.$	
$sim_3(a,b)$	$min\left\{\sum_{\ell=1}^{\mathcal{L}} d\left(\partial_{ml}^{a}, \partial_{nl}^{b}\right)\right.$	
$sim_4(a,b)$	$\frac{\sum_t\left[(\delta_a(t)-\overline{\delta_a})(\delta_b(t)-\overline{\delta_b})\right]}{\sqrt{\sum_t(\delta_a(t)-\overline{\delta_a})^2}\sqrt{\sum_t(\delta_b(t)-\overline{\delta_b})^2}}$	Correlation between structural, neighbourhood and community dynamicity of two non-connected actors computed using Pearson correlation
$sim_5(a,b)$	$\frac{\sum_t\left[(\lambda_a(t)-\overline{\lambda_a})(\lambda_b(t)-\overline{\lambda_b})\right]}{\sqrt{\sum_t(\lambda_a(t)-\overline{\lambda_a})^2}\sqrt{\sum_t(\lambda_b(t)-\overline{\lambda_b})^2}}$	
$sim_6(a,b)$	$\frac{\sum_t\left[(\partial_a(t)-\overline{\partial_a})(\partial_b(t)-\overline{\partial_b})\right]}{\sqrt{\sum_t(\partial_a(t)-\overline{\partial_a})^2}\sqrt{\sum_t(\partial_b(t)-\overline{\partial_b})^2}}$	
$sim_7(a,b)$	$1 - \frac{\sum_{t=1}^{T}[\lvert\delta_a(t)-\delta_b(t)\rvert + \lvert\lambda_a(t)-\lambda_b(t)\rvert + \lvert\partial_a(t)-\partial_b(t)\rvert]}{\sum_{t=1}^{T}[\lvert\delta_a(t)+\delta_b(t)\rvert + \lvert\lambda_a(t)+\lambda_b(t)\rvert + \lvert\partial_a(t)+\partial_b(t)\rvert]}$	Similarity by the abundance of structural, neighbourhood, and community dynamicity between two non-connected actors computed using Bray-Curtis dissimilarity measure
$sim_8(a,b)$	$\begin{cases} \sum_{t=1}^{T} \frac{\sum_{x\in\eta_a^{c_i}(g_t)\cap\eta_b^{c_i}(g_t)} CC_{g_t}^{x}}{\sum_{j=1,j\neq i}^{n}\sum_{y\in\eta_a^{c_i}(g_t)\cup\eta_b^{c_i}(g_t)} CC_{g_t}^{y}} \\ if\ a,b\in C_i(g_t) \\ \sum_{t=1}^{T}\left[\left\lvert V_{g_t}^{i,j}\right\rvert + \left\lvert E_{g_t}^{i,j}\right\rvert + \frac{\lambda_{g_t}^{a,b}}{\left\lvert p_{g_t}^{a,b}\right\rvert}\right] \\ if\ a\in C_i(g_t), b\in C_j(g_t), i\neq j \\ 0 \\ if\ a\in C_i(g_t), b\in C_j(g_t), i\neq j,\ \ p_{g_t}^{a,b}=\emptyset \end{cases}$	Actor similarity by using evolutionary community-aware network structural information

Ensemble-based learning models play a crucial role in alleviating the root causes of error in learning, which are due to noise, bias, and variance. On the other hand, Logistic Regression (i.e., binary logistic regression) is an example of a generalized linear model. This represents a special type of regression in which the binary response variable (i.e., label of the link, positive or negative) was related to a set of explanatory (i.e., predictor) variables (i.e., dynamic features). Before analysing the performance of the classifiers used in supervised link prediction in dynamic networks, a brief description of these three classifiers is presented in the following sections.

8.3 Classifiers

This section describes three classifiers used in this research for supervised classification purpose.

8.3.1 Bagging

Bagging stands for 'Bootstrap AGGregatING' [305] that attempts to decrease the variance of the prediction by generating additional data for training purposes from the original dataset. It uses combinations with repetitions or random sampling with replacements from the original training dataset. Although the model's prediction performances are not necessarily always improved, increasing the cardinality or size of the training datasets supports reducing the variance. The basic idea behind this learning algorithm is to split the training instances from the classification dataset into multiple random subsets. A classifier (e.g., decision tree) is trained for each collection of data subsets and thereby generating ensembles of multiple models. The prediction average from different trees is used and is considered more robust compared to a single decision tree. Bagging uses bootstrap sampling to obtain these subsets of data to train the base learner (i.e., decision tree) in conjunction with voting for

classification (i.e., plurality voting) and average for regression to aggregate the outputs of the base learner(s).

8.3.2 Random Forest

The Random Forest is a notion of an ensemble technique in statistical learning that is utilized for classification and/or prediction purposes in both statistics and machine learning [306]. The ensemble method is a divide-and-conquer approach in which improved prediction performance is generated by considering a weighted average (vote) of multiple basic model such as a decision tree. A decision tree is a tree-like complex and deterministic data structure in which each branch node represents a choice condition between a number of alternatives and each leaf node represents a classification or decision. The learning process starts by building a multitude of decision trees for the sample data. To classify a new object from a vector of sample data, the input vector is put down each of the trees in the forest. Consequently, each tree provides a classification that is considered as its vote for a class. Subsequently, the forest chooses the classification voted by the maximum number of trees. At each tree, a fraction of samples is randomly chosen (with replacement) to make the tree grow. At each branch node, a random subset of features/attributes, describing the samples (e.g. dynamic features in this study), is chosen to achieve the best split of the samples. The best threshold value of these features, contributing towards the best split, is held constant while the forest grows.

Given ensemble methods, the difference between Bagging and Random Forest is that the latter is considered as an extension of the former. In addition to creating subsets of data, it a random selection of features is considered rather than considering all features to grow decision tress. Interestingly, bagging algorithms are used on features in which each decision

tree uses a random subset of features and ends up creating many random trees to signify its name as a Random Forest.

8.3.3 Logistic Regression

Logistic Regression is a statistical method that is considered a special type of regression in which the dependent variable is binary or dichotomous. The goal of Logistic Regression is to find the best fitting model to describe the relationship between the dichotomous characteristic of interest and a set of explanatory or independent variables (i.e., predictor). It considers a linear model, which is made up of a linear predictor and two functions: (i) link function, that describes how the expectation of the dependent variable depends on the linear predictor, and (ii) a variance function that describes how the variance of the response variable depends on the expected value. It differs from linear regression by modelling the probability of the response variable, thereby taking a particular value that is based on combinations of values taken by the predictors. In case of linear regression, instead of the probability, the expected value of the response variable is considered. The term 'logistic' is a synonym of the word 'Sigmoid' and thus uses a sigmoid activation function with an 'S' shaped curve. As the sigmoid function simplifies the mathematics involved during optimization, it is considered an ideal choice for a small-scale classification problem.

8.4 Results

In this section, the empirical results, obtained in supervised dynamic link prediction strategy using the dynamic features established in chapter 5, are described:

8.4.1 Classifier Performances

In this section, the classification performances of three classifiers are described regarding the dynamic features developed in the previous chapter that are summarized in the Table 8.2. A

Table 8.2: Classification performances of three classifiers (i.e., LR=Logistic Regression, RF=Random Forest, and B=Bagging) in classifying positive and negatively-labelled instances in the classification datasets of six different dynamic network datasets. The instances in the corresponding dataset were described by dynamic features constructed by considering temporal series of network snapshots. Two different time scales (optimal and second optimal) were considered to generate these network snapshots.

Optimal Time Scale					**Second Optimal Time Scale**			
	Classifier	**Accuracy %**	**AUC ROC**	**AUC PR**		**Accuracy %**	**AUC ROC**	**AUC PR**
G_{MIT}^{30}	LR	**83.89**	0.653	0.13	G_{MIT}^{14}	**83.88**	**0.647**	**0.13**
	RF	76.17	**0.660**	**0.18**		73.89	0.560	0.12
	B	75.16	0.560	0.11		73.88	0.580	0.11
G_{Email}^{30}	LR	**83.25**	0.634	0.17	G_{Email}^{14}	81.55	0.638	0.24
	RF	82.31	**0.682**	**0.22**		**82.78**	0.591	**0.25**
	B	81.34	0.573	0.16		81.34	**0.647**	0.18
G_{UCI}^{14}	LR	83.27	0.614	0.27	G_{UCI}^{30}	83.27	0.657	0.25
	RF	**83.65**	**0.637**	**0.39**		**83.45**	**0.712**	**0.38**
	B	83.10	0.616	0.36		83.05	0.641	0.29
G_{FF}^{1}	LR	83.47	**0.623**	0.24	G_{FF}^{2}	**84.01**	**0.605**	**0.29**
	RF	82.13	0.619	**0.29**		82.51	0.591	**0.29**
	B	**83.51**	0.618	**0.29**		83.84	0.602	0.27
$G_{INF}^{0.5}$	LR	94.49	0.987	0.85	G_{INF}^{12}	88.13	0.926	0.60
	RF	**95.47**	**0.989**	0.84		90.58	**0.961**	0.69
	B	93.27	**0.989**	**0.87**		**91.68**	0.957	**0.75**
G_{HT}^{12}	LR	80.07	0.590	0.26	$G_{HT}^{1.5}$	**80.07**	0.619	0.26
	RF	**81.05**	**0.623**	**0.29**		80.03	**0.646**	**0.28**
	B	77.64	0.557	0.26		78.22	0.547	0.24

comparable performance representation of a static topological similarity metric (i.e., ResourceAllocation RA) and a time series forecasting-based dynamic link prediction strategy, proposed by Soares and Prudêncio [195]. The well-known machine learning library WEKA [307] was used for classification purposes using default parameters. For example, in case of ensemble-based Random Forest classification algorithm, WEKA uses Random Tree as base classifier that construct a tree considering K randomly chosen attributes at each node. It also does not perform pruning and allows estimation of class probabilities based on a hold-out set. Further, Random Forest algorithm in WEKA also considers 10 trees by default including unlimited depth for each tree. On the other hand, in case of Bagging classification algorithm, WEKA uses Reduced Error Pruning Tree ("REPT") which is the fast decision tree learning algorithm that builds a decision tree based on the information gain or reducing the variance. It builds a decision/regression tree using information gain as the splitting criterion, and prunes it using reduced error pruning. The default parameters in REPT are: unrestricted depth of trees, the minimum total weight of the instances in a leaf is two, and the minimum proportion of the variance on all the data that needs to be present at a node in order for splitting to be performed in regression trees is set to 0.001). In conjunction with REPT, Bagging also uses a parameter known as the size of each bag which is set to 100 percent of the training set size.

Table (8.2) presents the classification performances that were demonstrated by three classifiers using dynamic features constructed in this study. In this table, the classifier's performances are described using three metrics described in the previous chapter. These include accuracy by 10-fold cross-validation, Area Under ROC Curve (AUCROC), and Area Under PR Curve (AUCPR). Considering the best optimal temporal window choices in all datasets, the classification performances demonstrated by all classifiers were significant. When considering the accuracy score, most of the classifiers achieved more than 80%

accuracy across all datasets, except for the Bagging algorithm in two network datasets (i.e., G_{HT} and G_{MIT}) and Random Forest classifier in G_{MIT} for the second-best temporal window choice. The best accuracy score was achieved by the Random Forest classifier in the INFECTIOUS dataset G_{INF}, considering both the best (i.e., half an hour) and the second-best (i.e., 12 hour) temporal window choices. The worst performance, for the accuracy percentage, was recorded in the G_{MIT} dataset by the Bagging classifier for both optimal and second optimal temporal scale choices. Moreover, when considering the AUCROC scores, the best performance was recorded in the same dataset (i.e., G_{INF}) by all classifiers. However, Random Forest exceeded others by nominal differences.

The worst performance, when considering the AUCROC score, was demonstrated by the Bagging algorithm in the Hypertext dataset (i.e., G_{HT}). Despite its trivial performances, Bagging classifier exceeded the others with regard to the AUCPR scores in G_{INF} when considering both the optimal and second-optimal temporal scale choices. In the G_{MIT} dataset, the lowest score in this performance metric was recorded by the same classifier. It is noteworthy that, the minimum AUCPR score of 0.09, was surpassed by all classifiers across six datasets. In addition, although in some instances, the AUCROC scores were much lower than the best score recorded in G_{INF}, however, it was observed that all classifiers performed better than a random classifier, which can achieve a maximum AUCROC score of 0.50.

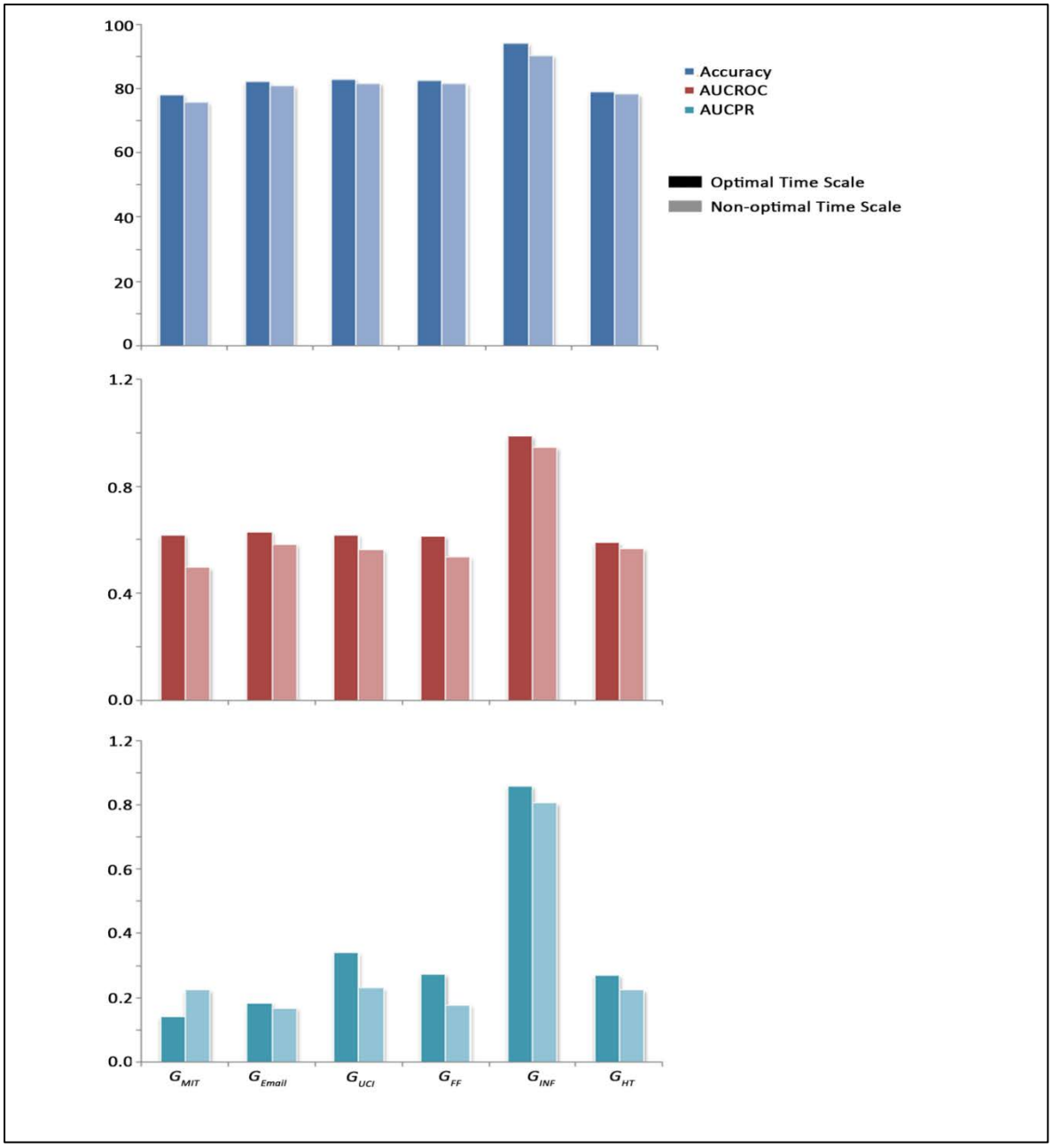


Figure 8.1: The average performances indicated by three classifiers (i.e., logistic regression, Random Forest, and Bagging) considering three performance metrics (Accuracy %, AUCROC, and AUCPR) in classification datasets. Each Performance metric denotes the average of aggregated performances demonstrated by the three classifiers together considering three performance metrics (Accuracy, AUCROC and AUCPR). Dynamic features were generated by considering Short Interval Networks (SINs) with both optimal (dark colour) and second-best optimal (light colour) time scale choices. The dark coloured bars represent the performance metrics when the optimal time scale was considered and the light coloured bars represent the performance metrics of the classifiers when the second best optimal time scale was considered in the corresponding dataset.

For the temporal window choices, it was observed that, other than few exceptions (e.g., G_{FF}), in most cases, three classifiers performed well in the classification dataset in which the dynamic features were constructed by considering the best and optimal temporal scale choice rather than the second-best choice. In the Facebook Friendship dataset, in AUCROC and AUCPR scores, three classifiers were observed with better performances when considering the best optimal time scale choice. However, in accuracy scores, three classifiers were outperformed by the second-best time scale. To further demonstrate that improved supervised dynamic link prediction performance depends on the optimal choice of the temporal window when actor-oriented evolutionary features were used, the average performances of three classifiers are demonstrated by considering both the optimal and non-optimal temporal window choices Figure (8.1). In this figure, the performances of each classifier in three different metrics were averaged and in the corresponding classification dataset, dynamic features were constructed, considering that SINs had both an optimal and a non-optimal temporal duration. For example, in G_{MIT}, considering the optimal temporal duration of 30 days, accuracy scores demonstrated by three classifiers, were aggregated and then divided by three to obtain the average accuracy score. Similarly, AUCROC and AUCPR scores were averaged for all classifiers. Next, the average performance scores of these three classifiers were computed in the similar fashion in a classification dataset in which each SIN had a temporal duration of one day (i.e., daily), a non-optimal temporal scale choice in this case. The figure demonstrates that optimal temporal duration affects the supervised link prediction performances in dynamic networks. In almost all cases, the optimal temporal window choice of the SINs supported the construction of actor-oriented evolutionary features that demonstrated better performance in predicting the future links. Thus, it is evident that an optimal time scale greatly impacted accurate link predictions in dynamic networks.

Further, it was evident that both linear and ensemble classifiers demonstrated notable performances in classification tasks. Regarding the accuracy scores in the optimal time scale choice, the Logistic Regression performed better in G_{MIT}, G_{Email} and G_{FF} datasets. In contrast, the ensemble classifier Random Forest performed well in G_{UCI}, G_{INF} and G_{HT}. In AUCROC scores, Logistic Regression displayed a better performance when compared to its counterpart Random Forest only in G_{FF}. An interesting observation was that Logistic Regression had either outperformed or performed closer to the other ensemble classifier (i.e., Bagging) across all datasets using the AUCROC scores. In addition, except for two datasets (i.e., G_{FF}, G_{UCI}), it downplayed Bagging in AUCPR scores. Nevertheless, in most cases, the Random Forest classifier performed notably better than others in AUCROC and AUCPR scores despite its setback in G_{FF} considering AUCROC scores, and in G_{INF}, considering AUCPR scores. it is noteworthy that between two ensemble-based classifiers, bagging, where a decision tree was used as a base classifier, was susceptible to overfitting and computationally expensive, as it considered all the available features to split a node in decision trees. Conversely, Random Forest, a special case of bagging, randomly considered only a subset of the best features of those available. Therefore, its performance was superior to that of bagging in several cases.

Table 8.3: Importance ranking of different dynamic features constructed in this study using different algorithms including Information Gain (IG), Chi-square statistical evaluation (Chi), attributes ranking in support vector machine classifier (SVM), and feature ranking in a Random Forest (RF) classifier. Ranks are in decreasing order in which number one (1) denotes the highest ranking. The 'Total' column represents the aggregation of all ranking score to generate the final ranking. sim_8^l denotes the 8th metric that used hierarchical agglomerative clustering approach and sim_8^h denotes the same metric using Louvaincommunity detection approach. The green-shaded cells represent the best performing features, whereas the yellow-shaded cells indicate the second-best features.

		IG	Chi	SVM	RF	Total		IG	Chi	SVM	RF	Total
G_{MIT}	sim_1	9	9	6	9	33	G_{FF}	6	6	7	3	22
	sim_2	4	4	2	6	16		4	4	4	7	19
	sim_3	5	5	4	1	15		9	9	9	1	28
	sim_4	3	3	5	3	14		5	5	5	2	17
	sim_5	2	2	8	8	20		8	8	8	8	32
	sim_6	7	7	7	5	26		7	7	3	9	26
	sim_7	8	8	1	7	24		3	3	6	4	16
	sim_8^l	6	6	3	2	17		2	2	1	5	10
	sim_8^h	1	1	9	4	15		1	1	2	6	10
G_{Email}	sim_1	9	9	6	7	31	G_{INF}	8	8	5	7	28
	sim_2	4	4	1	9	18		7	7	2	4	20
	sim_3	5	5	7	4	21		9	9	7	8	33
	sim_4	3	3	3	3	12		2	2	9	1	14
	sim_5	2	2	2	2	8		1	1	1	9	12
	sim_6	7	7	4	1	19		4	5	3	5	17
	sim_7	8	8	5	8	29		6	6	4	6	22
	sim_8^l	6	6	8	6	26		3	3	6	3	15
	sim_8^h	1	1	9	5	16		5	4	8	2	19
G_{UCI}	sim_1	9	9	9	3	30	G_{HT}	3	3	7	4	17
	sim_2	4	4	2	5	15		6	6	6	7	25
	sim_3	5	5	4	1	15		2	2	3	2	9
	sim_4	3	3	5	2	13		4	4	2	6	16
	sim_5	2	2	1	4	9		8	8	8	3	27
	sim_6	7	7	7	7	28		5	5	4	8	22
	sim_7	8	8	3	6	25		7	7	1	5	20
	sim_8^l	6	6	6	8	26		1	1	5	1	8
	sim_8^h	1	1	8	9	19		9	9	9	9	36

8.4.2 Feature Importance

After the performance measurement of classifiers when using nine different dynamic features, it was attempted to determine the relative importance of these dynamic features (i.e., $sim_1(a,b)$, $sim_2(a,b)$, ... $sim_8^h(a,b)$ and $sim_8^l(a,b)$), described in Table (6.1) to assess their relative competency in dynamic link prediction task in all six datasets. In the eighth feature, two community detection algorithms were considered (i.e., Louvain and hierarchical agglomerative clustering). Therefore, $sim_8(a,b)$ was further classified into ... $sim_8^h(a,b)$ and $sim_8^l(a,b)$ in which $sim_8^h(a,b)$denotes the metric that considered the hierarchical clustering approach, whereas $sim_8^l(a,b)$ denotes the feature constructed by considering Louvain community detection approach. For the ranking purpose, three different algorithms were considered from the WEKA machine learning software. These included information gain, chi-square evaluation, and Support Vector Machine (SVM) attribute evaluation. In Table (6.3), a comparable picture of these features was provided with regard to their rank of importance obtained by these algorithms. Information gain and chi-square evaluator algorithms evaluated the worthiness of a feature by calculating the information gain and chi-squared statistics with respect to the class variables. On the other hand, the SVM column denoted the rank of a feature with regard to the SVM. In this evaluation method, the worthiness of features was evaluated by using a SVM classifier in which the ranks of features were calculated by the square of the weight assigned by the SVM. Similarly, in this table, the column with Random Forest heading represents the importance of each feature according the feature evaluation mechanism employed in Random Forest (RF) classifier.

In Table (8.3), the ranks of the features were assigned in decreasing order with one denoting the highest ranking. Finally, all ranks for each four algorithms were aggregated to generate the final rank. The most important feature in each dataset was shaded green whereas

the second-best one was shaded yellow. This table indicates that $sim_4(a,b)$, constructed by considering the correlation between time series of actor-level structural dynamicity values, was the most prominent feature in G_{MIT}. Temporal similarity between community dynamicity values of non-connected actor pairs, measured by the Dynamic Time Warping method, denoted by $sim_3(a,b)$, was the second best dynamic feature in the same dataset accompanied by $sim_8^h(a,b)$. The latter represents similarity between actors using evolutionary community-aware network structural information by using a hierarchical agglomerative community detection algorithm. In contrast, $sim_5(a,b)$, the correlation between neighbourhood dynamicity values of actor-pairs, became the most valuable feature in G_{Email}, G_{UCI}, and G_{INF}. $sim_8^h(a,b)$ in G_{Email} and $sim_4(a,b)$ in G_{UCI} and G_{INF} were the second best important features in the respective datasets. In the rest of the two datasets (i.e., G_{FF} and G_{HT}), it is $sim_8^h(a,b)$ that was the most significant dynamic feature, although in G_{FF}, this feature was jointly accompanied by $sim_8^l(a,b)$ to win the first place. $sim_8^l(a,b)$ worked in a similar way such as $sim_8^h(a,b)$. However, it used the Louvain community detection method. Further, in the same dataset, $sim_7(a,b)$, which calculated similarity between actors by using Bray-Curtis similarity measure from ecology, was the second best emphatic feature, whereas in case of G_{HT}, it was $sim_3(a,b)$, temporal similarity between neighborhood dynamicity values of a pair of actors that was computed by the Dynamic Time Warping method. Thus, based on the aforementioned discussion, it was evident that although, most dynamic features perform well

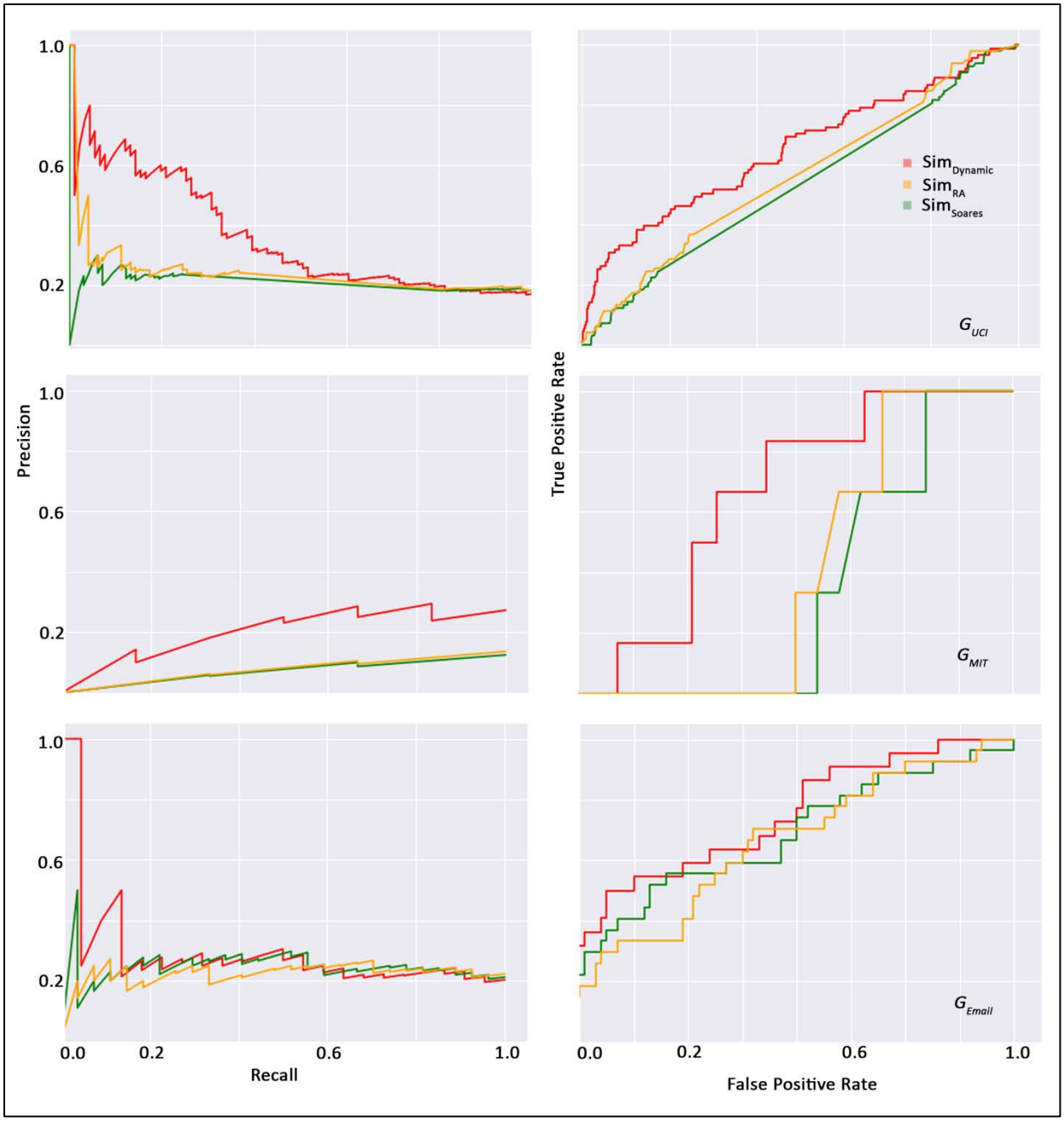


Figure 8.2: Visual representation of Precision-Recall (i.e., left column) and ROC curves (right column) of three network datasets G_{UCI} (top row), G_{MIT} (middle row) and G_{Email} (bottom row), considering the following features: (i) dynamic features $Sim_{Dynamic}$ (ii) topological similarity metric, Resource Allocation (RA) as a static link predictor Sim_{RA}, and (iii) Time series forecasting-based link prediction Sim_{Soares}.

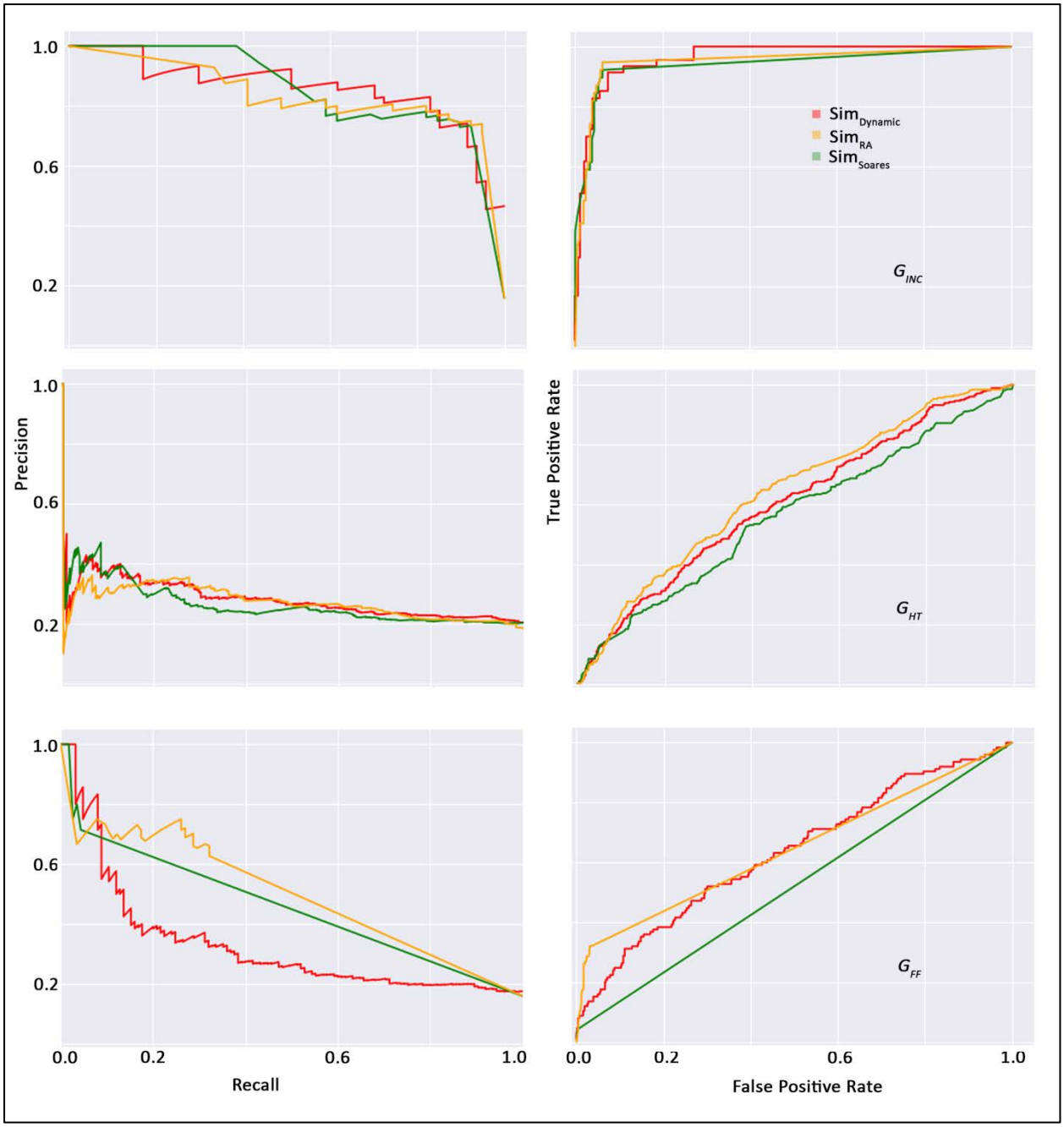


Figure 8.3: Visual representation of Precision-Recall (i.e., left column) and ROC curves (right column) of three network datasets G_{INF}(top row), G_{HT} (middle row) and G_{FF} (bottom row), considering the following features: (i) dynamic features $Sim_{Dynamic}$ (ii) topological similarity metric, Resource Allocation (RA) as a static link predictor Sim_{RA}, and (iii) Time series forecasting based link prediction Sim_{Soares}.

in supervised dynamic link prediction in dynamic networks, only few features performed decisive and applicable to datasets irrespective of their structure. The performance variations of three classifiers considering different metrics and variable importance factors, associated with different dynamic features constructed, is beyond the content and extent of this study. However, further studies can elucidate these research issues, which will help to determine the optimal temporal window size that is suitable for dynamic link prediction task in different contexts.

8.4.3 Comparison with Static Predictor

In this section, the performance of the best performing dynamic features was compared with a topological similarity metric, which is widely used in static network link prediction. The chosen topological similarity metric in this case was the 'ResourceAllocation[1]' (RA) metric [308]. This metric works in the similar fashion how the AdamicAdar (AA) index [309] works (i.e., $\frac{1}{\log d(z)}$ vs. $\frac{1}{d(z)}$). Both these metrics suppressed the contribution of high degree common neighbours. The principal difference between these two metrics lies within the degree (i.e., $d(z)$) of common neighbors. If the common neighbors of an actor-pair have more connections, than the differences between AA and RA are significant. To compute the RA index for each actor-pair in the classification dataset of individual network dataset, the temporal network snapshots $G_T = [G_{t_1},\ G_{t_1+\tau},\ G_{t_1+2\tau}\ \dots\ G_{t_1+n\tau} \dots G_{t'-\tau},\ G_{t'}]$ were aggregated into one cross-sectional network G_T, a static version of the corresponding dynamic network. Next, the RA index was computed for each positively and negatively-labelled actor-pair, which was then fed into three classifiers mentioned before, as a feature to describe the instances in the classification dataset. In Figure (8.2 & 8.3), comparable plots of dynamic features and RA topological similarity metric, in supervised link prediction context,

[1] ResourceAllocation is described in appendix A

by Precision-Recall (P-R) and ROC (Receiver Operating Characteristics) curves in six different network datasets are presented. It is noteworthy that in P-R plots, the goal of the curves is to appear in the bottom left corner of the graph to be optimal. The closer a curve is to the diagonal line, the higher the classifier's performance in classification. Conversely, in ROC plots, the goal of the curves is to be in the top-left region of the plots [310]. The higher the curve is away from the diagonal line, the better the predictor's performance. These figures present that apart from the ROC plot in G_{HT} and P-R plot in G_{FF}, in other 10 plots, the dynamic features established in this study, outperformed the static predictor, the Resource Allocation topological similarity metric. It is worth mentioning that in representing the performance comparison between dynamic and static topological feature by ROC and P-R plots, the corresponding winner classifier and the optimal time scale choices in each dataset were considered (described in section 2.1 and Table 6.2 of this chapter). These involve Random Forest in G_{MIT}^{30}, G_{Email}^{30}, G_{UCI}^{14}, $G_{INF}^{0.5}$, and G_{HT}^{12} and Logistic Regression in G_{FF}^{1} for ROC plots. Subsequently, for the P-R curve, a similar classifier was used across all datasets.

8.4.4 Comparison with Time Series Link Prediction

In this section, a performance comparison between the dynamic features and a time series forecasting-based link prediction strategy is presented in which the topological evolution was explored in dynamic networks by using temporal sequences of topology information. This strategy was developed by Soares and Prudêncio in which the authors built a time series of a chosen topological similarity metric (e.g., AdamicAdar[1]) for all non-connected actor pairs calculated in a temporal series of network snapshots or SINs in different past times $G_T = [G_{t_1},\ G_{t_1+\tau},\ G_{t_1+2\tau}\ \dots\ G_{t_1+n\tau} \dots G_{t'-\tau},\ G_{t'}]$. After the time series construction, a forecasting technique (e.g., ARIMA) was used to predict the next value of that metric in the next network

[1] Appendix A

G_{T+1}, which was used as input to supervised link prediction models [195]. In this section, their method was followed by using the same topological similarity metric RA, described above, to build the time series of the RA index for all instances of actor-pairs in the classification dataset. Then, the ARIMA procedure was followed to predict the future values of RA for each instance that was fed into the supervised learning setup as described in the previous chapter. The performance comparison was presented by using the same plots in Figure 6.2 and 6.3. These Figures demonstrate that in all cases, the dynamic features outplayed the time series forecasting-based link prediction in a supervised learning-based setting. Similar to the previous section, the best performing classifiers and the optimal time scale choice were used in the performance comparison.

8.5 Dynamic Feature Distribution

To determine whether similar or dissimilar actors, in regards to these evolving features in each network snapshots, participate in emerging links, the distributions of dynamic feature values were analysed. For this purpose, the best performing dynamic features were selected from Table (8.3). In Figure (8.4), distributions of the top two performing features from Table (8.3) in datasets G_{HT}, G_{Email} and G_{FF} are presented. In this figure, the top performing dynamic features found in these three datasets were $sim_8^h(a,b)$ and $sim_8^l(a,b)$, which denoted dynamic similarity metrics and computed the similarity between actor-pairs through using temporal community-aware network structural information. Although they performed in a similar fashion, their differences lie in the community detection method employed. The former employs hierarchical agglomerative clustering, whereas the latter used a Louvain community detection approach (see chapter 5 for details).

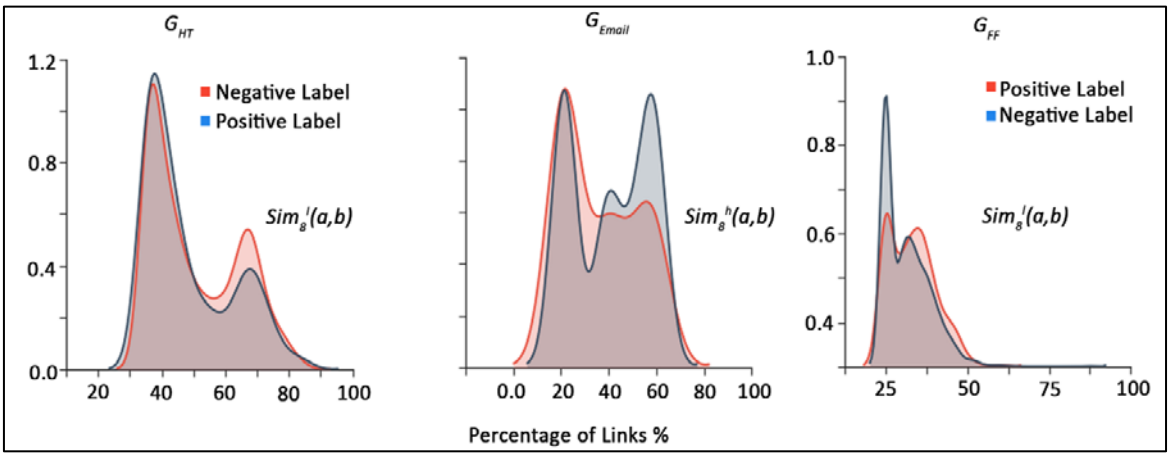


Figure 8.4: Distribution of three dynamic feature values in three network datasets $\boldsymbol{G_{HT}}$, G_{Email}, and G_{FF} for both positive and negatively-labeled actor-pairs in the corresponding classification datasets. The chosen features are the best performing features in the respective datasets. These are $sim_8^h\ (a,b)$ in G_{Email} and $sim_8^l(a,b)$ in G_{HT}, and G_{FF}. Both these metrics compute similarity between a pair of actors by considering evolutionary community-aware structural information. The first uses a hierarchical agglomerative, whereas the second uses the Louvain community detection method.

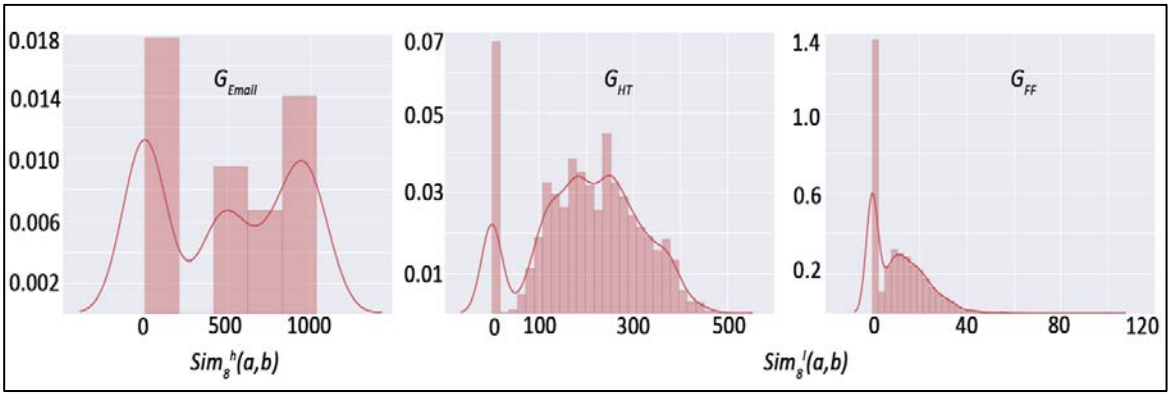


Figure 8.5: Binned distribution of three dynamic feature values in three network datasets G_{HT}, G_{Email}, and G_{FF} for positively-labeled actor-pairs in the corresponding classification datasets. The chosen features are the best performing features in the respective datasets. These include $sim_8^h(a,b)$ in G_{Email} and $sim_8^l(a,b)$ in G_{HT}, and G_{FF}.

From the classification performances that were demonstrated by different classifiers, it was observed that Logistic Regression classifier archived better accuracy in G_{Email} and G_{FF}. In contrast, Random Forest performed better in G_{HT}. The feature distribution from Figure (8.4) can help us to better understand the performance variations demonstrated by these two classifiers. For this purpose, the distributions of normalized feature values were plotted along the y-axis for the two features mentioned above, to investigate the spread of both positive and negatively labelled samples in the corresponding classification dataset. For the two features in the two right most plots (i.e., $sim_8^h(a,b)$ in G_{Email}, and $sim_8^l(a,b)$ in G_{FF}), the distributions of positive and negative classes exhibited differences with comparably reduced amount of overlapped regions. This facilitated the linear classification algorithm to pick patterns from the feature values and classify the samples correctly. In addition, the fraction of features values for both positive and negative class in the critical overlap region of the left most plot (i.e., $sim_8^l(a,b)$ in G_{HT}) was most likely the candidates for misclassification by the linear classifier in which ensemble classifier performed better. This depicted the underlying performance variations demonstrated by different classifiers.

However, from these plots, the spreads of positively-labelled links that appeared in the test phase of the link prediction, were not observed. Therefore, binned distributions of these three dynamic features were presented for those links that appeared in the test phase of the corresponding datasets in Figure (8.5). From the distributions of positively-labelled links in three datasets by considering highest performing dynamic features, it was observed that in G_{FF}, the lower the feature value the higher the likelihood of forming links between actor-pairs. However, in the other two datasets, the feature values of the true positive links spread over range of values, large and small. Although a pattern of lower feature values of the high performing feature existed for the emerging links in G_{FF}, however, this phenomenon was not

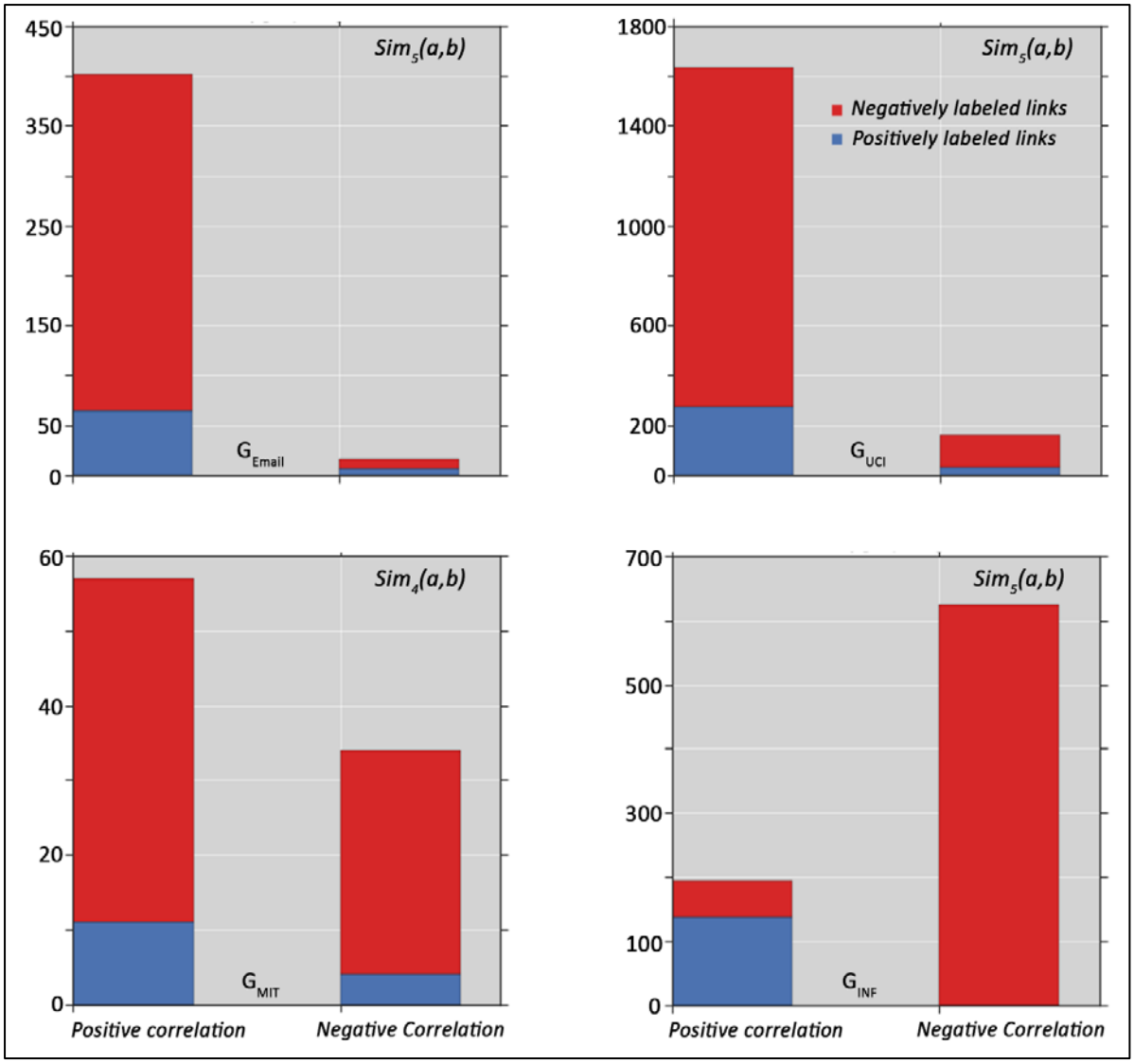


Figure 8.6: The four best performing correlation-based features in four datasets (i.e., G_{Email}, G_{UCI} , G_{MIT} and G_{INF}). These features measure the similarity between actor pairs by computing correlation between actor-level evolutionary information. $Sim_4(a,b)$ denotes the correlation between temporal dynamicity values of actor pairs, whereas $Sim_5(a,b)$ denotes the correlation between actor-level neighborhood dynamicity values. Considering these two features, the number of positive and negatively-correlated actors-pairs regarding their structural dynamicity $Sim_4(a,b)$ and neighborhood dynamicity $Sim_5(a,b)$ values are presented in four datasets in which the actor-pairs were either positively or negatively labelled in the corresponding classification datasets. The number of positive and negatively labelled links is denoted by the blue and red colour respectively. From the positively labelled links is observed that true links emerge among the actors those have positive correlation in regards to their evolution. It is observable from the figure that positive correlation between actor-level evolutions contributes more in emerging links.

applicable for the other two networks. It is noteworthy that only the high performing ones were considered to investigate the distribution of feature values. However, in a future study, the spread of distribution for the other feature values can be explored, despite their lower importance in the corresponding classification task. This justifies the selection of supervised link prediction over unsupervised strategy. If the ranking of feature values was performed in decreasing order by using unsupervised learning, and only the top-K values of the corresponding features were considered as probable links in future, then the result would be insignificant results. This was due to the phenomena described in Figure 8.5, in which true positive links not only emerged from higher values of top-performing features but also from the lower values.

Considering the top performing features, it was also observed that, correlation based features performed exclusively well across datasets. For example, the dynamic similarity metric $Sim_5(a, b)$that computed similarity between actor-pairs by considering the correlation between their neighborhood dynamicity values were the most prominent feature in three datasets (G_{Email}, $G_{UCI}$and G_{INF}). Therefore, to investigate what type of correlation existed between the dynamicity values of actor-pairs to form emerging links, correlation-based feature values were examined in four networks. These included $Sim_5(a, b)$ in G_{Email} , G_{UCI} and G_{INF}, and $Sim_4(a, b)$ in G_{MIT} . $Sim_4(a, b)$ measured the correlation between structural dynamicity values of actor-pairs. In Figure (8.6), the correlation types of both positively and negatively labelled actor-pairs (i.e., links) are presented in these four datasets. In each plot, the left bar represented the fraction of actor-pairs showing a positive correlation between their dynamicity values from both positive and negative classes. In addition, the right bar represented a negative correlation between dynamicity values of links in both classes. The red-coloured fraction denoted negatively-labelled actor pairs, whereas the blue-coloured fraction denoted positively-labelled links in each dataset. This figure shows that

most emerging links appear from the positive correlation of their corresponding dynamicity values incident to actor-pairs.

8.6 Concluding Remarks

Considering the problem of dynamic link prediction, this research attempted to develop evolutionary features by considering actor-oriented evolutionary information in dynamic networks. These features were constructed to act as input in supervised dynamic link prediction and subject to measure the similarity/proximity between actor-pairs with regard to different types of temporal changes they experience in evolving networks. Therefore, as the first step of defining dynamic similarity metrics, known as dynamic features, three different actor-level evolutionary aspects were identified: (i) network structural, (ii) neighbourhood, and (ii) cliquishness or community participation. Considering this evolutionary information associated with an individual actor, three different types of actor-oriented dynamicity measures were defined, known as structural, neighbourhood, and community dynamicity. Since these dynamicities include different temporal evolutions, experienced by actors, it is noteworthy that one of the important aspects of dynamic network analysis was to define the optimum time scale to sample the network and generate time series of network snapshots (i.e., SIN). For this purpose, the method described by chapter 3 was applied to find the optimal or best time scale for each SIN (i.e., link aggregation duration) for six real-life undirected social networks of different size and domains. With the optimal temporal window size defined for each SIN, ranging from half an hour to a month, time series of SINs were generated for each dynamic network datasets to compute the above mentioned three dynamicity values. This was followed by development of nine different dynamic features by using evolutionary features, denoting similarity between a pair of actors in evolutionary perspective, using dynamicity values.

To develop the dynamic features in this study, firstly, actor-level dynamicity values were leveraged in a dynamic programming-based temporal similarity method (i.e., DTW) and the Pearson correlation measures to develop the first six dynamic features. The seventh dynamic feature was constructed by measuring abundance of dynamicity using a similarity metric widely used in ecology, known as the Bray-Curtis similarity measure. In this measure, the normalized abundance of actor-level dynamicity values, incident to actor-pairs, were quantified by considering each SIN as a sample site (i.e., sampling zones in ecology) in temporal networks. Finally, based on two different existing community detection algorithms (i.e., Louvain and hierarchical agglomerative clustering), the eight and the last dynamic feature was developed by integrating evolutionary community-aware topologies, the actor's evolutionary community participation in conjunction with both an inter and intra-community network structure and associated neighbourhood changes.

In a supervised link prediction setup, two ensemble-based classifiers and one linear classifier were exploited to measure the performance of the aforementioned dynamic features. Considering the performance metrics, we observed that these features were only be indulged for a dynamic link prediction purpose can but also effectively support modelling of the network growth. Moreover, these features are so supportive in supervised dynamic link prediction that a simple linear classifier can perform well in classifying both positive and negatively-labelled links. The performances of dynamic features were also compared with a traditional topological metric (i.e., 'ResourceAllocation') widely used in link prediction purpose in cross-sectional networks (i.e., static network) and a time series forecasting-based dynamic link prediction strategy. In both cases, in this study, it was observed that dynamic features, constructed by leveraging the evolutionary aspect of actors, not only performed as good as the existing ones but also, surprisingly in most cases, outweighed them regarding prediction performance. Furthermore, in six dynamic network datasets, the ranks of dynamic

features were identified to determine the best performing feature(s), respectively. It was observed that correlation-based dynamic features measuring the level of correlation between actor-level dynamicity values performed well across most dynamic networks. These were followed by the eight metric, which was developed by exploiting evolutionary community-aware network structural information. Moreover, the spread and distribution of top-performing dynamic feature values were examined to investigate whether actor-pairs with either greater or smaller feature values showed a high likelihood. Although, no definite information could be extracted from the distribution of other dynamic features with regard to the range of their value in forging emerging links; however, considering the Pearson correlation-based features, it was evident that positively correlated actor-pairs, regarding their actor-level dynamicity values, were good indicators of forming future links in dynamic networks.

Chapter 9

Discussion and Conclusion

9.1 Discussion

Time-varying systems have a complex underlying network structure in which entities and their relations or interactions change temporally. Due to the evolutionary nature of its constituents, efficiently performing link inference in a dynamic network is extremely challenging. Numerous methodologies have been attempted to address the problem of predicting dynamic links by acquiring knowledge from the static version of the problem. In several studies, time components were incorporated into prediction strategies because time information, which is associated with links in this type of network, is crucial for accurate prediction [219]. Temporal link prediction has attracted considerable research interest in various domains, including sociology, anthropology, information, and computer science [227], while in several domains, especially in biology and medical care research, temporal link prediction can support the prediction of future interactions between entities that are hard and expensive to understand physically [311-313]. The link prediction problem in complex networks has received considerable interest because, in addition to its diverse application scenarios, it can be leveraged to understand the underlying rationale behind network growth and evolution. Although divergent prediction strategies, metrics and methodologies have emerged to solve this problem in static networks, the ineptness of these strategies in accommodating the associated dynamicity and evolutionary information results in their inadequacy in dynamic link prediction.

In previous studies, the integration of temporal information has been complied considering both time series analyses and evolutionary aspects (for example, temporal link decay and the duration of link activeness). However, existing dynamic link prediction strategies are not free from hindrances. Their inherent shortcomings are twofold: firstly, there is a lack of a standard framework to identify the correct, appropriate or optimal choice of

aggregation granularity to perform binning on any stream of time-stamped links to discern meaningful information and to understand the rate of dynamics demonstrated by these networks; and, secondly, the disregarding of evolutionary information incident to actors, the principal constituents of the network, in the prediction task. Therefore, it is imperative that temporal components need to be integrated as a parameter to the actor-level evolutionary aspects that can support the link prediction problem in dynamic networks.

Link formations in dynamic networks have an inherent rhythm and often occur over a range of time scales. Moreover, temporal streams of links are commonly aggregated into dynamic networks for temporal analysis and the resolution, or window size, at which the original data is aggregated, has a great impact on the results extracted from this analysis. Any discrepancy between the inherent temporal window of the underlying process and the window size at which the analysis is performed can either obscure important insights of the dynamic data or lead to erroneous conclusions [94]. Furthermore, the level of aggregation of the temporal stream of dynamic networks greatly impacts the patterns observed, the inference strategies employed in the corresponding network, and its processes [258,314], including the identification of noisy, local and critical temporal orders. Furthermore, the identification of an appropriate temporal window length strongly impacts structural analyses, the efficacy of network mining and the dynamics demonstrated by the network and its actors [258,246,107]. Having too coarse or too fine temporal granularity may conceal or fail to unravel critical information about network dynamics and thereby impair the understanding of the evolutionary structure of underlying interactions. Therefore, for any stream of time-stamped links that form a dynamic network, it is imperative to make the right choice of aggregation granularity that is used to bin dynamic data [315].

Considering the aforementioned concerns, this thesis predominantly undertook two major research objectives: determining the optimal sampling duration to discretise dynamic

networks and integrating actor-level evolutionary aspects in dynamic link prediction. In addressing the former, this study proposed a novel algorithm based on actor-level network positional variations over time, including some validation measures to endorse the optimality of the window size. To address the later, this research developed dynamic similarity metrics that contribute as features in supervised link prediction tasks in dynamic networks by mining different actor-level evolutions in optimally sampled dynamic networks. In the section below, I summarize my research contribution in this thesis in regard to the research questions defined in the first chapter.

9.2 Research Contribution

As described in Chapter 2, the amount of information on exploiting the actor-level dynamicity values to detect the optimal temporal scale and predict emerging links in dynamic networks is limited. Often this task is left to the arbitrary choices of scholars depending on the experimental contexts or the requirements of the corresponding studies. In other studies, it is left to the data-collection process, which is impractical. Several studies attempted to exploit network-level structural properties across temporal network snapshots to identify the appropriate window length. Furthermore, in Chapter 2, in which existing dynamic link prediction strategies are described, it was evident that little or no attention was paid to the sampling or discretization issue of the dynamic networks, with a common tendency to randomly select a temporal window size to generate network snapshots. Considering these two research issues, a list of research questions were addressed by this thesis. The following sections look further at these questions and their answers.

9.2.1 Optimal Sampling of Dynamic Network

9.2.1.1 Research Question

- How can actor-level measures be used to determine the optimal sampling interval to discretise a dynamic network?

9.2.1.2 Research Contribution

The foremost motivation behind using actor-level measures is because they are the principal constituents of dynamic networks. In dynamic networks, actors may appear or disappear and change their link structures continuously over time, thus contributing to the network dynamics. This fact may trigger multiple events. First, some actors may demonstrate higher-level network activities than others while network activities by some actors are under-represented. Second, an actor may demonstrate a high rate of network activities at the very beginning of a particular time window or another actor may create all its new ties at the end of the immediate window. Finally, in relation to a given window size, an actor may reveal all its network activities in only one window of the sampled dynamic network while another might engage in the same activities in comparably in more windows. This would significantly affect the involvement of dynamic actors contributing to the evolution of the underlying dynamic network. Consequently, the analysis of a given network could produce different results for the actor-level social network measures (for example, network centrality) when considering different time scale sizes. However, using an appropriate or optimal time scale should reduce the differences in network activities demonstrated by the group of dynamic actors. Therefore, choosing an actor-level measure would create symmetry in the distribution of actor-level network activities over time.

To determine the optimal window length, this study proposed a novel algorithm by considering the distribution of actor-level positional dynamicity values measured by using

popular centrality measures. In Chapter 3, the algorithm was described and the appropriate rationales behind using the corresponding measurements were explained. In this algorithm, the variances of actor dynamicity values are compared by using different time-scale durations to sample a dynamic network. The window length with the minimum variance in actor dynamicity distributions defines the appropriate sampling window to analyse the dynamic network because the minimum variance will ensure that the suggested window size will be neither too large to reveal the high rates of network activities for some actors to exhibit a large volume of network activities nor be too small for other actors who reveal slow rates of network activities to exhibit a minimum number of network activities.

In Chapter 7, this research also demonstrated the experimental results of applying the algorithm over six real-life dynamic networks in order to identify the optimal and second-best optimal temporal window sizes and to discretise them. It is noteworthy that the proposed algorithm determines the optimal sampling time scale/window from a list of candidate time windows where the candidate time windows are network and context dependent and can be of any duration (for example, second, minute, hour, day or month). For example, if streaming links are collected or aggregated every second in a dynamic network, then choices of candidate windows in multiple of day(s) would be inappropriate. Similarly, if links are aggregated in a dynamic network using the temporal unit of single day, then selecting a candidate window of seconds or minutes would produce inaccurate results. However, selecting a large number of candidate windows would also be unreasonable. For example, if links appear at a rate of one per minute in a dynamic network, then considering candidate windows as multiples of microseconds, milliseconds or even seconds would result with most network snapshots having no links at all. Similarly, in a system where links are triggered once per day, considering ample candidate windows in multiples of seconds, minutes or hours would be unsuitable.

9.2.1.3 Research Question

- How can the optimality of the sampling resolution be validated?

9.2.1.4 Research Contribution

In Chapter 3, this thesis proposed three different evaluation criteria to validate the optimality of the candidate window(s). These validation measures are based on best-fit ARIMA model, time-series anomalies and an unsupervised clustering model known as *k*-means clustering. Chapter 3 describes these validation methods in detail, including the rationales behind using them to validate the optimality. In Chapter 7, these validation measures were applied over the optimal window resolution(s) of real-life dynamic networks identified by the proposed algorithm that considers the variances of actor-level positional dynamicity values. In all network datasets, this study observed that the identified optimal window(s) were valid and effective, despite a few exceptions found in case of the second-best optimal windows. In two network datasets, it was observed that multiple candidate windows, including the originally identified ones by the algorithm, became contenders to be the second-best optimal windows.

9.2.2 Actor-level Dynamicity

9.2.2.1 Research Question

- What kinds of evolutions or dynamicities are demonstrated by actors in dynamic networks?

9.2.2.2 Research Contribution

In Chapter 4, this study developed three types of actor-level dynamicities demonstrated by actors in dynamic networks. These dynamicities were developed based on the evolutionary changes experienced by actors in regard to their network structures, neighbourhoods and clustering tendency.

9.2.2.3 Research Question

- How can the actor-level dynamicities be quantified?

9.2.2.4 Research Contribution

In Chapter 4, this thesis defined the mathematical quantification of different actor-level dynamicities. The first one is the structural dynamicity, which was measured by using three prominent centrality measures used in social network. These are degree, closeness and betweenness centrality. Chapter 4 also provided necessary definitions of three centrality measures and why they were used to quantify this dynamicity. The second actor-level dynamicity was quantified by considering the actor's neighbourhood retention and gaining rate over time. This was called neighbourhood dynamicity. If an actor has both a high gaining rate in conjunction with a high retention rate of neighbours in a dynamic network then the corresponding actor has high neighbourhood dynamicity. The final actor-level dynamicity was defined by measuring the temporal changes of actors' clustering tendency or community participation. In all cases, the differences between actor-oriented centrality measures, neighbourhood counts and clustering coefficients in consecutive network snapshots were compared against the same values computed in an aggregated network that consisted of two individual snapshots in two consecutive timestamps.

9.2.3 Dynamic Similarity Metrics

9.2.3.1 Research Question

- How can the evolution similarity between actor-pairs be calculated by considering different actor-level dynamicities?

9.2.3.2 Research Contribution

In Chapter 5, this thesis proposed three different methods to compute the evolution similarity between actor-pairs: temporal similarity measures based on the dynamic time warping

(DTW) method, cross-correlation and using an ecological similarity measure known as the Bray-Curtis similarity measure. This chapter also provided detailed descriptions of the functionalities of these methods. The last similarity measure was developed by considering an actor's community participation pattern and community-aware structural evolutions over time in dynamic networks. By considering three different actor-level dynamicities in the first three similarity measures (DTW, cross-correlation and Bray-Curtis), this thesis proposed seven evolutionary similarity measures. In addition to this, by considering two different community detection algorithms (Louvain and hierarchical agglomerative clustering) in conjunction with the community-aware temporal changes in dynamic networks, this research also developed another two evolution similarity measures. The dominant rationales are that the community structure effectively manifests the information about actors with similar behaviour that can be conducive in predicting their future interaction [316], and that the high and low condensation of links among actors can be an effective prediction of emerging links [317]. Furthermore, incorporating community and structural information drastically improved the accuracy of link prediction [80]. These nine evolution similarity measures were called dynamic similarity metrics/dynamic features in this research and were used in supervised link prediction.

9.2.3.3 Research Question

- What impact do the evolutionary similarities between actor-pairs have in dynamic link prediction?

9.2.3.4 Research Contribution

In Chapter 8, this thesis applied nine different dynamic similarity metrics/dynamic features, theoretically constructed in Chapter 5, in a supervised link prediction setup by considering six dynamic networks. Two ensemble-based classifiers and one linear classifier were used to

measure the performance of the dynamic features. Considering the performance metrics, it was observed that these features are not only supportive for dynamic link prediction purposes but also effectively support modelling the network growth. Moreover, considering the classification performances demonstrated by both linear and ensemble classifiers, it was evident that dynamic features were competitive in supervised link prediction setup where even a mere linear classifier can successfully predict emerging links by classifying both positively and negatively labelled links. This thesis also observed that community-aware dynamic similarity metrics performed better in emerging link prediction in dynamic networks.

9.2.3.5 Research Question

- What is the impact of an optimal sampling window interval on dynamic link prediction?

9.2.3.6 Research Contribution

In Chapter 8, it was observed that evolutionary similarity-based features performed better where different actor-level evolutions were computed in an optimally sampled dynamic network. In all cases, this research observed that optimal sampling to discretise dynamic networks is imperative for improved performance in predicting emerging links.

9.2.3.7 Research Question

- What kind of actors participate in emerging links of a dynamic network in regard to their evolutionary similarity (i.e., similar/closer or dissimilar/distant)?

9.2.3.8 Research Contribution

The empirical results from Chapter 8 demonstrate that although actors with both similar and dissimilar evolutionary similarity participate in emerging links in dynamic networks, there

exists a positive correlation between their dynamicity values. More concretely, positively correlated actor-level dynamicities denote a higher likelihood of future link formations.

9.2.3.9 Research Question

- What are the performance enhancements of evolutionary similarity-based features over traditional neighbourhood-based prediction or time series-based link prediction in dynamic networks?

9.2.3.10 Research Contribution

The performances of dynamic features were compared with a traditional topological metrics (i.e., ResourceAllocation) that are widely used in link prediction purposes in cross-sectional (i.e., static) networks and a time series-based dynamic link prediction strategy developed by Soares and Prudêncio [195]. In both cases, it was observed that dynamic features, constructed by leveraging different evolutionary aspects of actors, not only perform as well as the existing dynamic features but also, surprisingly in most cases, outweigh them regarding prediction performance.

9.3 Conclusion

The literary discussions in Chapter 2 showed that the two most important aspects of dynamic link prediction framework have been overlooked by researchers in network science. These are the concrete abstraction of the dynamic or temporal networks and understanding the different types of evolutions experienced by the actors in these networks. In general, as discussed in Chapters 2 and 3, a given dynamic network is understood by a time series of smaller network snapshots, known as a short interval network (SIN). Instead of arbitrarily or randomly establishing the appropriate duration of this SIN over time for the entire duration of a given dynamic network, it is imperative to conjecture a standard framework to derive the

optimal temporal length of the SINs in a series. Failure to do so will lead to a huge performance differences in different link prediction strategies over the same dynamic network. Further, only looking at the construction of dynamic features by considering the actor-level network information computed in SIN over time will provide inadequate results unless an optimal sampling strategy exists to discretise temporal links. Furthermore, in dynamic network analysis, while capturing the rate of actor-level evolutions measured by different social network metrics depends on the underlying network structure of these SINs, the aggregation window in binning temporal links determines the underlying structure of these network snapshots. Thus, due to this inter-dependency, for the purpose of dynamic link prediction, the first prerequisite is to establish the optimum temporal duration of SINs in a dynamic network.

Therefore, in this study, the two most important issues in dynamic link prediction were the optimal temporal scale of dynamic networks and developing a supervised dynamic link prediction model by using actor-level evolutionary features to understand dynamic network growth. For the first issue, the proposed approach used in this thesis was different from previous approaches in terms of its simplicity, computational efficiency and applicability. It is free from relying on either maintaining the ground truth (i.e., more sampling is better) and it is not dependent on parametric distribution. Unlike other methods, as described in Chapter 3, it uses metrics related to the network structural evolutions of actors, the central constituents of a network, instead of network or graph metrics.

This study proposed that actor-level evolutionary measures be used to predict future links in dynamic networks without actor attributes. The common practice in dynamic link prediction is either to leverage the topological structure of networks, applicable to pair of actors, or to use a computation intensive probabilistic or parametric approach. In previous studies, the integration of temporal components into the prediction process was attempted.

However, in these studies actor-level temporal network evolutionary aspects were either completely or partially ignored. Therefore, in this research, I attempted to construct dynamic features that compute the similarity/proximity between actor-pairs by considering their dynamicity information. Instead of exploiting dyadic information, this study leveraged an individual actor's evolution over time for the feature of construction. Moreover, different, unique methods were considered to compute the similarity between dynamic actors which have not been explored before.

The approach used in this study to construct dynamic features can be further extended in several ways. For example, instead of using centrality measures or the clustering tendency of actors to predict the future dynamicity values of actors, other network structures or topology (for example, assortativity) can be exploited, including time series forecasting methods (for example, ARIMA). Moreover, other similarity measures (for example, Euclidean, Manhattan) can be employed instead of dynamic time warping to measure the similarity between temporal information. In case of the final dynamic similarity metric, other community detection algorithms (for example, edge betweenness) can be used to enhance prediction performance. Finally, like many other applications of link prediction problems, this study may be valuable to help define novel dynamic similarity metrics for dynamic link predictions in networks that inherently evolve over time, including terrorist networks, online social networks (for example, Twitter), scholarly and knowledge networks (for example, keyword networks) and collaborative filtering to model consumers' buying behaviour.

References

1. Yao L, Wang L, Pan L, Yao K (2016) Link Prediction Based on Common-Neighbors for Dynamic Social Network. Procedia Computer Science 83:82-89. doi:https://doi.org/10.1016/j.procs.2016.04.102

2. Lichtenwalter RN, Lussier JT, Chawla NV New perspectives and methods in link prediction. In: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, 2010. ACM, pp 243-252

3. Ozcan A, Oguducu SG (2018) Link prediction in evolving heterogeneous networks using the NARX neural networks. Knowledge and Information Systems 55 (2):333-360

4. Li Y, Luo P, Fan Z-p, Chen K, Liu J (2017) A utility-based link prediction method in social networks. European Journal of Operational Research 260 (2):693-705. doi:https://doi.org/10.1016/j.ejor.2016.12.041

5. Tsiotas D, Charakopoulos A (2018) Visibility in the topology of complex networks. Physica A: Statistical Mechanics and its Applications 505:280-292. doi:https://doi.org/10.1016/j.physa.2018.03.055

6. De Menezes MA, Barabási A-L (2004) Fluctuations in network dynamics. Physical review letters 92 (2):028701

7. Garlaschelli D, Caldarelli G, Pietronero L (2003) Universal scaling relations in food webs. Nature 423:165. doi:10.1038/nature01604

8. Newman ME (2003) The structure and function of complex networks. SIAM review 45 (2):167-256

9. Fortunato S (2010) Community detection in graphs. Physics reports 486 (3-5):75-174

10. Girvan M, Newman ME (2002) Community structure in social and biological networks. Proceedings of the national academy of sciences 99 (12):7821-7826

11. Yang J, Leskovec J (2015) Defining and evaluating network communities based on ground-truth. Knowledge and Information Systems 42 (1):181-213

12. Lu Z, Savas B, Tang W, Dhillon IS Supervised link prediction using multiple sources. In: Data Mining (ICDM), 2010 IEEE 10th International Conference on, 2010. IEEE, pp 923-928

13. Dong Y, Tang J, Wu S, Tian J, Chawla NV, Rao J, Cao H Link prediction and recommendation across heterogeneous social networks. In: 2012 IEEE 12th International Conference on Data Mining, 2012. IEEE, pp 181-190

14. Liu Q, Tang S, Zhang X, Zhao X, Zhao BY, Zheng H Network growth and link prediction through an empirical lens. In: Proceedings of the 2016 Internet Measurement Conference, 2016. ACM, pp 1-15

15. Wang P, Xu B, Wu Y, Zhou X (2015) Link prediction in social networks: the state-of-the-art. Science China Information Sciences 58 (1):1-38

16. Shang K-k, Small M, Yan W-s (2017) Link direction for link prediction. Physica A: Statistical Mechanics and its Applications 469:767-776. doi:https://doi.org/10.1016/j.physa.2016.11.129

17. Lü L, Zhou T (2011) Link prediction in complex networks: A survey. Physica A: Statistical Mechanics and its Applications 390 (6):1150-1170. doi:https://doi.org/10.1016/j.physa.2010.11.027

18. Sett N, Basu S, Nandi S, Singh SR (2018) Temporal link prediction in multi-relational network. World Wide Web 21 (2):395-419

19. Bilgic M, Namata GM, Getoor L Combining collective classification and link prediction. In: Data Mining Workshops, 2007. ICDM Workshops 2007. Seventh IEEE International Conference on, 2007. IEEE, pp 381-386

20. Almansoori W, Gao S, Jarada TN, Elsheikh AM, Murshed AN, Jida J, Alhajj R, Rokne J (2012) Link prediction and classification in social networks and its application in healthcare and systems biology. Network Modeling Analysis in Health Informatics and Bioinformatics 1 (1-2):27-36

21. Zhang L, Hu K, Tang Y (2010) Predicting disease-related genes by topological similarity in human protein-protein interaction network. Open Physics 8 (4):672-682

22. He X-S, Zhou M-Y, Zhuo Z, Fu Z-Q, Liu J-G (2015) Predicting online ratings based on the opinion spreading process. Physica A: Statistical Mechanics and its Applications 436:658-664

23. Liu R, Ouyang Y, Rong W, Song X, Tang C, Xiong Z Rating prediction based job recommendation service for college students. In: International conference on computational science and its applications, 2016. Springer, pp 453-467

24. Nilashi M, Ibrahim OB, Ithnin N, Zakaria R (2015) A multi-criteria recommendation system using dimensionality reduction and Neuro-Fuzzy techniques. Soft Computing 19 (11):3173-3207

25. Mori J, Kajikawa Y, Kashima H, Sakata I (2012) Machine learning approach for finding business partners and building reciprocal relationships. Expert Systems with Applications 39 (12):10402-10407

26. Wu S, Sun J, Tang J Patent partner recommendation in enterprise social networks. In: Proceedings of the sixth ACM international conference on Web search and data mining, 2013. ACM, pp 43-52

27. Vidmer A, Zeng A, Medo M, Zhang Y-C (2015) Prediction in complex systems: The case of the international trade network. Physica A: Statistical Mechanics and its Applications 436:188-199

28. Xie F, Chen Z, Shang J, Feng X, Li J (2015) A link prediction approach for item recommendation with complex number. Knowledge-Based Systems 81:148-158

29. Raymond R, Kashima H (2010) Fast and scalable algorithms for semi-supervised link prediction on static and dynamic graphs. Machine Learning and Knowledge Discovery in Databases:131-147

30. Huang Z, Zeng DD A link prediction approach to anomalous email detection. In: Systems, Man and Cybernetics, 2006. SMC'06. IEEE International Conference on, 2006. IEEE, pp 1131-1136

31. Berlusconi G, Calderoni F, Parolini N, Verani M, Piccardi C (2016) Link Prediction in Criminal Networks: A Tool for Criminal Intelligence Analysis. PLOS ONE 11 (4):e0154244. doi:10.1371/journal.pone.0154244

32. Rattigan MJ, Jensen D (2005) The case for anomalous link discovery. ACM SIGKDD Explorations Newsletter 7 (2):41-47

33. Qayyum S, Mansoor S, Khalid A, Khushbakht, Halim Z, Baig AR Fraudulent call detection for mobile networks. In: 2010 International Conference on Information and Emerging Technologies, 14-16 June 2010 2010. pp 1-5. doi:10.1109/ICIET.2010.5625718

34. Al-Oufi S, Kim H-N, El Saddik A Controlling privacy with trust-aware link prediction in online social networks. In: Proceedings of the Third International Conference on Internet Multimedia Computing and Service, 2011. ACM, pp 86-89

35. Yu H, Braun P, Yıldırım MA, Lemmens I, Venkatesan K, Sahalie J, Hirozane-Kishikawa T, Gebreab F, Li N, Simonis N (2008) High-quality binary protein interaction map of the yeast interactome network. Science 322 (5898):104-110

36. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabási A-L (2000) The large-scale organization of metabolic networks. Nature 407 (6804):651-654

37. Gül S, Kaya M, Kaya B Predicting links in weighted disease networks. In: Computer and Information Sciences (ICCOINS), 2016 3rd International Conference on, 2016. IEEE, pp 77-81

38. Singh-Blom UM, Natarajan N, Tewari A, Woods JO, Dhillon IS, Marcotte EM (2013) Prediction and validation of gene-disease associations using methods inspired by social network analyses. PloS one 8 (5):e58977

39. Folino F, Pizzuti C Link prediction approaches for disease networks. In: International Conference on Information Technology in Bio-and Medical Informatics, 2012. Springer, pp 99-108

40. Guimerà R, Sales-Pardo M (2009) Missing and spurious interactions and the reconstruction of complex networks. Proceedings of the National Academy of Sciences 106 (52):22073-22078

41. Stanfield Z, Coşkun M, Koyutürk M (2017) Drug Response Prediction as a Link Prediction Problem. Scientific Reports 7:40321. doi:10.1038/srep40321

https://www.nature.com/articles/srep40321#supplementary-information

42. Turki T, Wei Z (2017) A link prediction approach to cancer drug sensitivity prediction. BMC Systems Biology 11 (Suppl 5):94. doi:10.1186/s12918-017-0463-8

43. Kaya B, Poyraz M (2016) Unsupervised link prediction in evolving abnormal medical parameter networks. International Journal of Machine Learning and Cybernetics 7 (1):145-155. doi:10.1007/s13042-015-0405-y

44. Benchettara N, Kanawati R, Rouveirol C A supervised machine learning link prediction approach for academic collaboration recommendation. In: Proceedings of the fourth ACM conference on Recommender systems, 2010. ACM, pp 253-256

45. Klimek P, Jovanovic AS, Egloff R, Schneider R (2016) Successful fish go with the flow: citation impact prediction based on centrality measures for term–document networks. Scientometrics 107 (3):1265-1282

46. Sun Y, Barber R, Gupta M, Aggarwal CC, Han J Co-author relationship prediction in heterogeneous bibliographic networks. In: Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on, 2011. IEEE, pp 121-128

47. Fazel-Zarandi M, Devlin HJ, Huang Y, Contractor N Expert recommendation based on social drivers, social network analysis, and semantic data representation. In: Proceedings of the 2nd international workshop on information heterogeneity and fusion in recommender systems, 2011. ACM, pp 41-48

48. Kc M, Chau R, Hagenbuchner M, Tsoi AC, Lee V A machine learning approach to link prediction for interlinked documents. In: International Workshop of the Initiative for the Evaluation of XML Retrieval, 2009. Springer, pp 342-354

49. Li Y, Wen A, Lin Q, Li R, Lu Z (2014) Name disambiguation in scientific cooperation network by exploiting user feedback. Artificial Intelligence Review 41 (4):563-578

50. Yadav A, Singh YN, Singh R (2015) Improving routing performance in AODV with link prediction in mobile Adhoc Networks. Wireless Personal Communications 83 (1):603-618

51. Han Q, Bai Y, Gong L, Wu W (2011) Link availability prediction-based reliable routing for mobile ad hoc networks. IET communications 5 (16):2291-2300

52. Yan G, Zhou T, Hu B, Fu Z-Q, Wang B-H (2006) Efficient routing on complex networks. Physical Review E 73 (4):046108

53. Jia X, Xin F, Chuan WR (2013) Adaptive spray routing for opportunistic networks. International Journal on Smart Sensing and Intelligent Systems 6 (1):95-119

54. Liu Z, Ma J, Zeng Y (2013) Secrecy transfer for sensor networks: From random graphs to secure random geometric graphs. International Journal on Smart Sensing and Intelligent Systems 6 (1):77-94

55. Guo L, Ma J, Chen Z, Zhong H (2015) Learning to recommend with social contextual information from implicit feedback. Soft Computing 19 (5):1351-1362

56. Wang M, Ma J (2016) A novel recommendation approach based on users' weighted trust relations and the rating similarities. Soft Computing 20 (10):3981-3990

57. Almohammadi K, Hagras H, Yao B, Alzahrani A, Alghazzawi D, Aldabbagh G (2017) A type-2 fuzzy logic recommendation system for adaptive teaching. Soft Computing 21 (4):965-979

58. Fournet J, Barrat A (2014) Contact patterns among high school students. PloS one 9 (9):e107878

59. Buccafurri F, Lax G, Nocera A, Ursino D (2015) Discovering missing me edges across social networks. Information Sciences 319:18-37. doi:https://doi.org/10.1016/j.ins.2015.05.014

60. Nguyen T, Phung DQ, Adams B, Venkatesh S Towards Discovery of Influence and Personality Traits through Social Link Prediction. In: ICWSM, 2011. pp 566-569

61. Yang Y, Tang J, Leung CW-k, Sun Y, Chen Q, Li J, Yang Q RAIN: Social Role-Aware Information Diffusion. In: AAAI, 2015. pp 367-373

62. Kim M, Leskovec J The network completion problem: Inferring missing nodes and edges in networks. In: Proceedings of the 2011 SIAM International Conference on Data Mining, 2011. SIAM, pp 47-58

63. Marchette DJ, Priebe CE (2008) Predicting unobserved links in incompletely observed networks. Computational Statistics & Data Analysis 52 (3):1373-1386

64. Xia P, Ribeiro B, Chen C, Liu B, Towsley D A study of user behavior on an online dating site. In: Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2013. ACM, pp 243-247

65. Albert R, Barabási A-L (2002) Statistical mechanics of complex networks. Reviews of modern physics 74 (1):47

66. Newman M (2010) Networks: an introduction. Oxford university press,

67. Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world'networks. nature 393 (6684):440-442

68. Li A, Cornelius SP, Liu Y-Y, Wang L, Barabási A-L (2017) The fundamental advantages of temporal networks. Science 358 (6366):1042-1046

69. Holme P (2015) Modern temporal network theory: a colloquium. The European Physical Journal B 88 (9):234

70. Masuda N, Klemm K, Eguíluz VM (2013) Temporal networks: slowing down diffusion by long lasting interactions. Physical Review Letters 111 (18):188701

71. Lentz HH, Selhorst T, Sokolov IM (2013) Unfolding accessibility provides a macroscopic approach to temporal networks. Physical review letters 110 (11):118701

72. Li X, Du N, Li H, Li K, Gao J, Zhang A A Deep Learning Approach to Link Prediction in Dynamic Networks. In: SIAM International Conference on Data Mining, Philadelphia, USA, 2014. Society of Industrial & Applied Mathematics, pp 289-297

73. Aggarwal C, Subbian K (2014) Evolutionary network analysis: A survey. ACM Computing Surveys (CSUR) 47 (1):10

74. Yu W, Aggarwal CC, Wang W Temporally Factorized Network Modeling for Evolutionary Network Analysis. In: Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, 2017. ACM, pp 455-464

75. Chen B, Chen L (2014) A link prediction algorithm based on ant colony optimization. Applied Intelligence 41 (3):694-708

76. Opsahl T, Hogan B (2011) Growth mechanisms in continuously-observed networks: Communication in a facebook-like community. arXiv:10102141

77. Liben-Nowell D, Kleinberg J (2007) The link-prediction problem for social networks. Journal of the American society for information science and technology 58 (7):1019-1031

78. Adrian K, Chocron P, Confalonieri R, Ferrer X, Giraldez-cru J Link Prediction in Evolutionary Graphs. In: Artificial Intelligence Research and Development: Proceedings of the 19th International Conference of the Catalan Association for Artificial Intelligence, Barcelona, Catalonia, Spain, October 19-21, 2016, 2016. IOS Press, p 187

79. Tabourier L, Libert A-S, Lambiotte R (2016) Predicting links in ego-networks using temporal information. EPJ Data Science 5 (1):1. doi:10.1140/epjds/s13688-015-0062-0

80. Feng X, Zhao J, Xu K (2012) Link prediction in complex networks: a clustering perspective. The European Physical Journal B 85 (1):3

81. Wang T, He X-S, Zhou M-Y, Fu Z-Q (2017) Link Prediction in Evolving Networks Based on Popularity of Nodes. Scientific reports 7 (1):7147

82. Nguyen CH, Mamitsuka H (2012) Latent feature kernels for link prediction on sparse graphs. IEEE transactions on neural networks and learning systems 23 (11):1793-1804

83. Lorrain F, White HC (1971) Structural equivalence of individuals in social networks. The Journal of mathematical sociology 1 (1):49-80

84. Liu Y, Zhao C, Wang X, Huang Q, Zhang X, Yi D (2016) The degree-related clustering coefficient and its application to link prediction. Physica A: Statistical Mechanics and its Applications 454:24-33

85. Barabási A-L, Albert R (1999) Emergence of scaling in random networks. science 286 (5439):509-512

86. Al Hasan M, Zaki MJ (2011) A survey of link prediction in social networks. In: Social network data analytics. Springer, pp 243-275

87. Chen Y, Chen K-J, Li Y A Link Prediction Method That Can Learn from Network Dynamics. In: 2014 IEEE International Conference on Data Mining Workshop, 2014. IEEE, pp 549-553

88. Li T, Wang J, Tu M, Zhang Y, Yan Y Enhancing Link Prediction Using Gradient Boosting Features. In: International Conference on Intelligent Computing, 2016. Springer, pp 81-92

89. Tylenda T, Angelova R, Bedathur S Towards time-aware link prediction in evolving social networks. In: Proceedings of the 3rd workshop on social network mining and analysis, 2009. ACM, p 9

90. Sarkar P, Chakrabarti D, Jordan MI Nonparametric Link Prediction in Dynamic Networks. In: International Conference on Machine Learning, Edinburgh, Scotland, 26th June-1st July 2012. pp 1687-1694

91. Li X, Du N, Li H, Li K, Gao J, Zhang A A deep learning approach to link prediction in dynamic networks. In: Proceedings of the 2014 SIAM International Conference on Data Mining, 2014. SIAM, pp 289-297

92. Choudhury N, Uddin S (2018) Evolutionary Community Mining for Link Prediction in Dynamic Networks. In: Cherifi C, Cherifi H, Karsai M, Musolesi M (eds) Complex Networks & Their Applications VI: Proceedings of Complex Networks 2017 (The Sixth International Conference on Complex Networks and Their Applications). Springer International Publishing, Cham, pp 127-138. doi:10.1007/978-3-319-72150-7_11

93. Ibrahim NMA, Chen L (2015) Link prediction in dynamic social networks by integrating different types of information. Applied Intelligence 42 (4):738-750

94. Caceres RS, Berger-Wolf T (2013) Temporal scale of dynamic networks. In: Temporal Networks. Springer, pp 65-94

95. Timmons AC, Preacher KJ (2015) The importance of temporal design: How do measurement intervals affect the accuracy and efficiency of parameter estimates in longitudinal research? Multivariate behavioral research 50 (1):41-55

96. Uddin S, Choudhury N, Farhad SM, Rahman MT (2017) The optimal window size for analysing longitudinal networks. Scientific Reports 7 (1):13389. doi:10.1038/s41598-017-13640-5

97. Sulo R, Berger-Wolf T, Grossman R Meaningful selection of temporal resolution for dynamic networks. In: Proceedings of the Eighth Workshop on Mining and Learning with Graphs, 2010. ACM, pp 127-136

98. Uddin S, Khan A, Piraveenan M (2016) A set of measures to quantify the dynamicity of longitudinal social networks. Complexity 21 (6):309-320

99. Caceres RS, Fish B (2017) A supervised approach to windowing detection on dynamic networks. MIT Lincoln Laboratory Lexington United States,

100. Holme P, Saramäki J (2012) Temporal networks. Physics Reports 519 (3):97-125. doi:https://doi.org/10.1016/j.physrep.2012.03.001

101. Sarr I, Missaoui R (2014) Temporal Analysis on Static and Dynamic Social Networks Topologies. In: Encyclopedia of Social Network Analysis and Mining. Springer, pp 2111-2119

102. Bliss CA, Frank MR, Danforth CM, Dodds PS (2014) An evolutionary algorithm approach to link prediction in dynamic social networks. Journal of Computational Science 5 (5):750-764. doi:https://doi.org/10.1016/j.jocs.2014.01.003

103. Vu DQ, Hunter D, Smyth P, Asuncion AU Continuous-time regression models for longitudinal networks. In: Advances in Neural Information Processing Systems, 2011. pp 2492-2500

104. Morris M, Kretzschmar M (1995) Concurrent partnerships and transmission dynamics in networks. Social Networks 17 (3-4):299-318

105. Park H-J, Friston K (2013) Structural and functional brain networks: from connections to cognition. Science 342 (6158):1238411

106. Steglich C, Snijders TA, Pearson M (2010) Dynamic networks and behavior: Separating selection from influence. Sociological methodology 40 (1):329-393

107. Ribeiro B, Perra N, Baronchelli A (2013) Quantifying the effect of temporal resolution on time-varying networks. Scientific reports 3:3006

108. Hinde RA (1976) Interactions, relationships and social structure. Man:1-17

109. Fish B, Caceres RS (2017) A supervised approach to time scale detection in dynamic networks. arXiv preprint arXiv:170207752

110. Moody J, McFarland D, Bender-deMoll S (2005) Dynamic network visualization. American journal of sociology 110 (4):1206-1241

111. Shannon CE (2001) A mathematical theory of communication. ACM SIGMOBILE Mobile Computing and Communications Review 5 (1):3-55

112. Feldmann A, Gilbert AC, Willinger W, Kurtz TG (1998) The changing nature of network traffic: Scaling phenomena. ACM SIGCOMM Computer Communication Review 28 (2):5-29

113. Pesaran H, Timmermann A (2007) Selection of estimation window in the presence of breaks. ournal of Econometrics 137 (1):134-161

114. Keogh E, Chu S, Hart D, Pazzani M An online algorithm for segmenting time series. In: Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on, 2001. IEEE, pp 289-296

115. Barron A, Rissanen J, Yu B (1998) The minimum description length principle in coding and modeling. IEEE Transactions on Information Theory 44 (6):2743-2760

116. Holme P (2015) Modern temporal network theory: a colloquium. The European Physical Journal B 88 (9):1-30

117. Kivelä M, Porter MA (2015) Estimating interevent time distributions from finite observation periods in communication networks. Physical Review E 92 (5):052813

118. Collins LM, Graham JW (2002) The effect of the timing and spacing of observations in longitudinal studies of tobacco and other drug use: Temporal design considerations. Drug and Alcohol Dependence 68:85-96

119. Siegler R (2006) Handbook of Child Psychology, Vol 2: Cognition, Perception and Language. John Wiley & Sons, Inc, New Jersey

120. Winkens B (2005) Optimal design and analysis of clinical trials with repeated measures. Maastricht University,

121. Krivitsky PN, Handcock MS (2014) A separable model for dynamic networks. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76 (1):29-46

122. Soundarajan S, Tamersoy A, Khalil EB, Eliassi-Rad T, Chau DH, Gallagher B, Roundy K Generating graph snapshots from streaming edge data. In: Proceedings of the 25th International Conference Companion on World Wide Web, 2016. International World Wide Web Conferences Steering Committee, pp 109-110

123. Jaccard P (1901) Étude comparative de la distribution florale dans une portion des Alpes et des Jura. Bull Soc Vaudoise Sci Nat 37:547-579

124. Darst RK, Granell C, Arenas A, Gómez S, Saramäki J, Fortunato S (2016) Detection of timescales in evolving complex systems. Scientific reports 6:39713

125. Fish B, Caceres RS Handling oversampling in dynamic networks using link prediction. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 2015. Springer, pp 671-686

126. Holme P (2003) Network dynamics of ongoing social relationships. EPL (Europhysics Letters) 64 (3):427

127. Onody RN, de Castro PA (2003) Optimization and self-organized criticality in a magnetic system. Physica A: Statistical Mechanics and its Applications 322:247-255

128. Vázquez A, Oliveira JG, Barabási A-L (2005) Inhomogeneous evolution of subgraphs and cycles in complex networks. Physical Review E 71 (2):025103

129. Eagle NN (2005) Machine perception and learning of complex social systems. Massachusetts Institute of Technology,

130. Sun J, Faloutsos C, Papadimitriou S, Yu PS Graphscope: parameter-free mining of large time-evolving graphs. In: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, 2007. ACM, pp 687-696

131. Getoor L, Sahami M Using probabilistic relational models for collaborative filtering. In: Workshop on Web Usage Analysis and User Profiling (WEBKDD'99), 1999.

132. Getoor L, Friedman N, Koller D, Taskar B (2002) Learning probabilistic models of link structure. Journal of Machine Learning Research 3 (Dec):679-707

133. Domingos P, Richardson M Mining the network value of customers. In: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, 2001. ACM, pp 57-66

134. Popescul A, Ungar LH Statistical relational learning for link prediction. In: IJCAI workshop on learning statistical models from relational data, 2003.

135. Yu K, Chu W, Yu S, Tresp V, Xu Z Stochastic relational models for discriminative link prediction. In: Advances in neural information processing systems, 2007. pp 1553-1560

136. Freeman LC (1989) Social networks and the structure experiment. Research methods in social network analysis:11-40

137. Wasserman S, Faust K (1994) Social network analysis: Methods and applications, vol 8. Cambridge university press,

138. Freeman LC (1982) Centered graphs and the structure of ego networks. Mathematical Social Sciences 3 (3):291-304

139. Everett M, Borgatti SP (2005) Ego network betweenness. Social networks 27 (1):31-38

140. Al Hasan M, Chaoji V, Salem S, Zaki M Link prediction using supervised learning. In: SDM06: workshop on link analysis, counter-terrorism and security, 2006.

141. Haghani S, Keyvanpour MR (2017) A systemic analysis of link prediction in social network. Artificial Intelligence Review:1-35

142. Srinivas V, Mitra P (2016) Link Prediction in Social Networks: Role of Power Law Distribution. Springer,

143. Zhu L, Guo D, Yin J, Ver Steeg G, Galstyan A (2016) Scalable temporal latent space inference for link prediction in dynamic social networks. IEEE Transactions on Knowledge and Data Engineering 28 (10):2765-2777

144. Berry MW, Browne M, Langville AN, Pauca VP, Plemmons RJ (2007) Algorithms and applications for approximate nonnegative matrix factorization. Computational statistics & data analysis 52 (1):155-173

145. Dunlavy DM, Kolda TG, Acar E (2011) Temporal Link Prediction Using Matrix and Tensor Factorizations. ACM Trans Knowl Discov Data 5 (2):1-27. doi:10.1145/1921632.1921636

146. Acar E, Dunlavy DM, Kolda TG Link prediction on evolving data using matrix and tensor factorizations. In: Data Mining Workshops, 2009. ICDMW'09. IEEE International Conference on, 2009. IEEE, pp 262-269

147. Katz L (1953) A new status index derived from sociometric analysis. Psychometrika 18 (1):39-43

148. Carroll JD, Chang J-J (1970) Analysis of individual differences in multidimensional scaling via an N-way generalization of “Eckart-Young” decomposition. Psychometrika 35 (3):283-319

149. Yu W, Cheng W, Aggarwal CC, Chen H, Wang W Link Prediction with Spatial and Temporal Consistency in Dynamic Networks. In: 26th International Joint Conferences on Artifical Intelligence (IJCAI), Melbourne, Australia, 2017. International Joint Conferences on Artificial Intelligence, pp 3343-3349

150. Ma X, Sun P, Wang Y (2018) Graph regularized nonnegative matrix factorization for temporal link prediction in dynamic networks. Physica A: Statistical Mechanics and its Applications 496:121-136. doi:https://doi.org/10.1016/j.physa.2017.12.092

151. Ma X, Sun P, Qin G (2017) Nonnegative matrix factorization algorithms for link prediction in temporal networks using graph communicability. Pattern Recognition 71:361-374

152. Gao S, Denoyer L, Gallinari P Temporal link prediction by integrating content and structure information. In: Proceedings of the 20th ACM international conference on Information and knowledge management, 2011. ACM, pp 1169-1174

153. Junuthula RR, Xu KS, Devabhaktuni VK Evaluating Link Prediction Accuracy in Dynamic Networks with Added and Removed Edges. In: Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom)(BDCloud-SocialCom-SustainCom), 2016 IEEE International Conferences on, 2016. IEEE, pp 377-384

154. Foulds J, DuBois C, Asuncion A, Butts C, Smyth P A dynamic relational infinite feature model for longitudinal social networks. In: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, 2011. pp 287-295

155. Leskovec J, Kleinberg J, Faloutsos C (2007) Graph evolution: Densification and shrinking diameters. ACM Transactions on Knowledge Discovery from Data (TKDD) 1 (1):2

156. Heaukulani C, Ghahramani Z Dynamic probabilistic models for latent feature propagation in social networks. In: International Conference on Machine Learning, 2013. pp 275-283

157. Xu K Stochastic block transition models for dynamic networks. In: Artificial Intelligence and Statistics, 2015. pp 1079-1087

158. Xu KS, Hero AO (2014) Dynamic stochastic blockmodels for time-evolving social networks. IEEE Journal of Selected Topics in Signal Processing 8 (4):552-562

159. Yang T, Chi Y, Zhu S, Gong Y, Jin R (2011) Detecting communities and their evolutions in dynamic social networks—a Bayesian approach. Machine learning 82 (2):157-189

160. Barbieri N, Bonchi F, Manco G Who to follow and why: link prediction with explanations. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, 2014. ACM, pp 1266-1275

161. Robins G, Pattison P, Kalish Y, Lusher D (2007) An introduction to exponential random graph (p*) models for social networks. Social networks 29 (2):173-191

162. Snijders TA, Pattison PE, Robins GL, Handcock MS (2006) New specifications for exponential random graph models. Sociological methodology 36 (1):99-153

163. Hanneke S, Xing EP (2007) Discrete temporal models of social networks. Lecture Notes in Computer Science 4503:115

164. Potgieter A, April KA, Cooke RJ, Osunmakinde IO (2009) Temporality in link prediction: Understanding social complexity. Emergence: Complexity and Organization 11 (1):69

165. Potgieter A, April K, Cooke R, Lockett M (2006) Adaptive Bayesian agents: Enabling distributed social networks. South African Journal of Business Management 37 (1):41-55

166. Zhu J, Hong J, Hughes JG (2002) Using markov chains for link prediction in adaptive web sites. In: Soft-Ware 2002: Computing in an Imperfect World. Springer, pp 60-73

167. Chen M-S, Park JS, Yu PS Data mining for path traversal patterns in a web environment. In: Distributed Computing Systems, 1996., Proceedings of the 16th International Conference on, 1996. IEEE, pp 385-392

168. Friedman N, Getoor L, Koller D, Pfeffer A Learning probabilistic relational models. In: IJCAI, 1999. pp 1300-1309

169. Sanghai S, Domingos P, Weld D Learning statistical models of time-varying relational data. In: Proceedings of the Workshop on Statistical Relational Learning, 18th International Joint Conference on Artificial Intelligence, 2003.

170. Milch BC, Russell SJ (2006) Probabilistic models with unknown objects. University of California, Berkeley,

171. Sharan U, Neville J Exploiting time-varying relationships in statistical relational models. In: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis, 2007. ACM, pp 9-15

172. Hayashi K, Hirayama J-I, Ishii S (2009) Dynamic exponential family matrix factorization. Advances in Knowledge Discovery and Data Mining:452-462

173. McCullagh P, Nelder JA (1989) Generalized Linear Models, no. 37 in Monograph on Statistics and Applied Probability. Chapman & Hall/CRC.,

174. Sarkar P, Moore AW Dynamic social network analysis using latent space models. In: Advances in Neural Information Processing Systems, 2006. pp 1145-1152

175. Pujari M, Kanawati R Supervised rank aggregation approach for link prediction in complex networks. In: Proceedings of the 21st International Conference on World Wide Web, 2012. ACM, pp 1189-1196

176. Aalen O, Borgan O, Gjessing H (2008) Survival and event history analysis: a process point of view. Springer Science & Business Media,

177. Zeng Z, Chen K-J, Zhang S, Zhang H A link prediction approach using semi-supervised learning in dynamic networks. In: Advanced Computational Intelligence (ICACI), 2013 Sixth International Conference on, 2013. IEEE, pp 276-280

178. He Y, Liu JN, Hu Y-x, Wang X-z (2015) OWA operator based link prediction ensemble for social network. Expert Systems with Applications 42 (1):21-50

179. Bao Z, Zeng Y, Tay Y sonLP: social network link prediction by principal component regression. In: Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on, 2013. IEEE, pp 364-371

180. O'Madadhain J, Hutchins J, Smyth P (2005) Prediction and ranking algorithms for event-based network data. SIGKDD Explor Newsl 7 (2):23-30. doi:10.1145/1117454.1117458

181. Bringmann B, Berlingerio M, Bonchi F, Gionis A (2010) Learning and predicting the evolution of social networks. IEEE Intelligent Systems 25 (4):26-35

182. Berlingerio M, Bonchi F, Bringmann B, Gionis A (2009) Mining graph evolution rules. Machine learning and knowledge discovery in databases:115-130

183. Zhang Z, Wen J, Sun L, Deng Q, Su S, Yao P (2017) Efficient incremental dynamic link prediction algorithms in social network. Knowledge-Based Systems 132:226-235. doi:https://doi.org/10.1016/j.knosys.2017.06.035

184. Tseng P, Yun S (2009) A coordinate gradient descent method for nonsmooth separable minimization. Mathematical Programming 117 (1-2):387-423

185. Dorigo M, Birattari M, Stutzle T (2006) Ant colony optimization. IEEE computational intelligence magazine 1 (4):28-39

186. Sherkat E, Rahgozar M, Asadpour M (2015) Structural link prediction based on ant colony approach in social networks. Physica A: Statistical Mechanics and its Applications 419:80-94. doi:https://doi.org/10.1016/j.physa.2014.10.011

187. Eagle N, Pentland AS, Lazer D (2009) Inferring friendship network structure by using mobile phone data. Proceedings of the national academy of sciences 106 (36):15274-15278

188. Box GE, Jenkins GM (1976) Time series analysis: forecasting and control, revised ed. Holden-Day, San Francisco, CA

189. Huang Z, Lin DK (2009) The Time-Series Link Prediction Problem with Applications in Communication Surveillance. INFORMS Journal on Computing 21 (2):286-303

190. Holme P, Park SM, Kim BJ, Edling CR (2007) Korean university life in a network perspective: Dynamics of a large affiliation network. Physica A: Statistical Mechanics and its Applications 373:821-830

191. Moradabadi B, Meybodi MR (2017) A novel time series link prediction method: Learning automata approach. Physica A: Statistical Mechanics and its Applications 482:422-432. doi:https://doi.org/10.1016/j.physa.2017.04.019

192. Thathachar MA, Sastry PS (2002) Varieties of learning automata: an overview. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 32 (6):711-722

193. Sarkar P, Chakrabarti D, Jordan M Nonparametric link prediction in dynamic networks. In: International Conference on Machine Learning, Edinburgh, Scotland, 26 June-01 July 2012.

194. Murata T, Moriyasu S (2008) Link prediction based on structural properties of online social networks. New Generation Computing 26 (3):245-257

195. Soares PRdS, Prudêncio RBC Time series based link prediction. In: Neural Networks (IJCNN), The 2012 International Joint Conference on, 2012. IEEE, pp 1-7

196. Güneş İ, Gündüz-Öğüdücü Ş, Çataltepe Z (2016) Link prediction using time series of neighborhood-based node similarity scores. Data Mining and Knowledge Discovery 30 (1):147-180

197. Bütün E, Kaya M, Alhajj R A new topological metric for link prediction in directed, weighted and temporal networks. In: 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 18-21 Aug. 2016 2016. pp 954-959. doi:10.1109/ASONAM.2016.7752355

198. Schall D (2014) Link prediction in directed social networks. Social Network Analysis and Mining 4 (1):157

199. Romero DM, Kleinberg JM The directed closure process in hybrid social-information networks, with an analysis of link formation on Twitter. In: ICWSM, 2010.

200. Kim M, Han J (2009) A particle-and-density based evolutionary clustering method for dynamic networks. Proceedings of the VLDB Endowment 2 (1):622-633

201. Kunegis J, Lommatzsch A Learning spectral graph transformations for link prediction. In: Proceedings of the 26th Annual International Conference on Machine Learning, 2009. ACM, pp 561-568

202. Aggarwal CC, Xie Y, Yu PS (2014) A framework for dynamic link prediction in heterogeneous networks. Statistical Analysis and Data Mining 7 (1):14-33. doi:10.1002/sam.11198

203. Rossetti G, Guidotti R, Miliou I, Pedreschi D, Giannotti F (2016) A supervised approach for intra-/inter-community interaction prediction in dynamic social networks. Social Network Analysis and Mining 6 (1):86

204. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment 2008 (10):P10008

205. Rosvall M, Bergstrom CT (2011) Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems. PloS one 6 (4):e18209

206. Coscia M, Rossetti G, Giannotti F, Pedreschi D Demon: a local-first discovery method for overlapping communities. In: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, 2012. ACM, pp 615-623

207. Yuan H, Ma Y, Zhang F, Liu M, Shen W A Distributed Link Prediction Algorithm Based on Clustering in Dynamic Social Networks. In: Systems, Man, and Cybernetics (SMC), 2015 IEEE International Conference on, 2015. IEEE, pp 1341-1345

208. White T (2012) Hadoop: The definitive guide. " O'Reilly Media, Inc.",

209. De Meo P, Ferrara E, Fiumara G, Provetti A Generalized louvain method for community detection in large networks. In: Intelligent Systems Design and Applications (ISDA), 2011 11th International Conference on, 2011. IEEE, pp 88-93

210. Castaneda O (2011) Link Prediction and the Evolution of Communities on Twitter. Delft University of Technology, the Netherlands

211. Munasinghe L, Ichise R Time aware index for link prediction in social networks. In: International Conference on Data Warehousing and Knowledge Discovery, 2011. Springer, pp 342-353

212. Munasinghe L (2013) Time-aware methods for link prediction in social networks. Ph. D. thesis, The Graduate University for Advanced Studies,

213. Choudhary P, Mishra N, Sharma S, Patel R Link score: A novel method for time aware link prediction in social network. In: Ding W (ed) ICDMW Google Scholar, Dallas, Texas, 2013. IEEE Computer Society,

214. Merritt S, Jacobs A, Mason W, Clauset A Detecting Friendship Within Dynamic Online Interaction Networks. In: International AAAI Conference on Weblogs and Social Media, Boston, USA, 8-10 July 2013 2013. AAAI press,

215. Soares PR, PrudêNcio RB (2013) Proximity measures for link prediction based on temporal events. Expert Systems with Applications 40 (16):6652-6660

216. Salem Narasimhan J (2015) Link Prediction in Dynamic Networks. Doctoral, WASHINGTON STATE UNIVERSITY, Washington ,USA

217. Liu J, Deng G (2009) Link prediction in a user–object network based on time-weighted resource allocation. Physica A: Statistical Mechanics and its Applications 388 (17):3643-3650

218. Wang C, Han J, Jia Y, Tang J, Zhang D, Yu Y, Guo J Mining advisor-advisee relationships from research publication networks. In: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, 2010. ACM, pp 203-212

219. Lakshmi TJ, Bhavani SD (2017) Temporal probabilistic measure for link prediction in collaborative networks. Applied Intelligence 47 (1):83-95. doi:10.1007/s10489-016-0883-y

220. Wang C, Satuluri V, Parthasarathy S Local probabilistic models for link prediction. In: Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on, 2007. IEEE, pp 322-331

221. Ahmed NM, Chen L (2016) An efficient algorithm for link prediction in temporal uncertain social networks. Information Sciences 331:120-136. doi:https://doi.org/10.1016/j.ins.2015.10.036

222. Jeh G, Widom J (2002) SimRank: a measure of structural-context similarity. In: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, Edmonton, Alberta, Canada, 2002. Association for Computing Machinery, pp 538-543

223. Sajadmanesh S, Zhang J, Rabiee HR (2017) NPGLM: A Non-Parametric Method for Temporal Link Prediction. arXiv preprint arXiv:170606783

224. Adrian K, Chocron P, Confalonieri R, Ferrer X, Giraldez-cru J (eds) (2016) Link Prediction in Evolutionary Graphs, vol 288. Artificial Intelligence Research and Development. IOS Press, Amsterdam

225. Asur S, Parthasarathy S, Ucar D (2009) An event-based framework for characterizing the evolutionary behavior of interaction graphs. ACM Transactions on Knowledge Discovery from Data (TKDD) 3 (4):16

226. Kashima H, Kato T, Yamanishi Y, Sugiyama M, Tsuda K Link propagation: A fast semi-supervised learning algorithm for link prediction. In: Proceedings of the 2009 SIAM international conference on data mining, 2009. SIAM, pp 1100-1111

227. Ahmed NM, Chen L, Wang Y, Li B, Li Y, Liu W (2016) Sampling-based algorithm for link prediction in temporal networks. Information Sciences 374:1-14. doi:https://doi.org/10.1016/j.ins.2016.09.029

228. Lahiri M, Berger-Wolf TY Mining periodic behavior in dynamic social networks. In: Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on, 2008. IEEE, pp 373-382

229. Han J, Cheng H, Xin D, Yan X (2007) Frequent pattern mining: current status and future directions. Data mining and knowledge discovery 15 (1):55-86

230. Huang K-Y, Chang C-H (2005) SMCA: a general model for mining asynchronous periodic patterns in temporal databases. IEEE Transactions on Knowledge and Data Engineering 17 (6):774-785

231. Rahman M, Hasan MA Link Prediction in Dynamic Networks Using Graphlet. In, Cham, 2016. Machine Learning and Knowledge Discovery in Databases. Springer International Publishing, pp 394-409

232. Juszczyszyn K, Kazienko P, Musial K, Gabrys B (2008) Temporal Changes in Connection Patterns of an Email-Based Social Network. Paper presented at the Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 03,

233. Juszczyszyn K, Musial K, Budka M Link prediction based on subgraph evolution in dynamic social networks. In: Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on, 2011. IEEE, pp 27-34

234. Aggarwal C, Xie Y, Yu PS On dynamic link inference in heterogeneous networks. In: Proceedings of the 2012 SIAM International Conference on Data Mining, 2012. SIAM, pp 415-426

235. Sun Y, Han J, Yan X, Yu PS, Wu T (2011) Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. Proceedings of the VLDB Endowment 4 (11):992-1003

236. Li J, Ge B, Yang K, Chen Y, Tan Y (2017) Meta-path based heterogeneous combat network link prediction. Physica A: Statistical Mechanics and its Applications 482:507-523

237. Yang Y, Chawla N, Sun Y, Hani J Predicting links in multi-relational and heterogeneous networks. In: Data Mining (ICDM), 2012 IEEE 12th International Conference on, 2012. IEEE, pp 755-764

238. Yang Y, Tang J, Keomany J, Zhao Y, Li J, Ding Y, Li T, Wang L Mining competitive relationships by learning across heterogeneous networks. In: Proceedings of the 21st ACM international conference on Information and knowledge management, 2012. ACM, pp 1432-1441

239. Tang J, Lou T, Kleinberg J Inferring social ties across heterogenous networks. In: Proceedings of the fifth ACM international conference on Web search and data mining, 2012. ACM, pp 743-752

240. Caceres RS, Berger-Wolf T, Grossman R Temporal scale of processes in dynamic networks. In: Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on, 2011. IEEE, pp 925-932

241. Lahiri M, Berger-Wolf TY Structure prediction in temporal networks using frequent subgraphs. In: Computational Intelligence and Data Mining, 2007. CIDM 2007. IEEE Symposium on, 2007. IEEE, pp 35-42

242. Chaiton M, Cohen J, O'Loughlin J, Rehm J (2010) Use of cigarettes to improve affect and depressive symptoms in a longitudinal study of adolescents. Addictive behaviors 35 (12):1054-1060

243. Uddin S, Choudhury N, Farhad SM, Rahman MT (2017) The optimal window size for analysing longitudinal networks. Scientific Reports 7

244. Clauset A, Eagle N Persistence and periodicity in a dynamic proximity network. In: DIMACS Workshop on Computational Methods for Dynamic Interaction Networks., Piscataway, 24-26 September 2007. DIMACS,

245. Kivelä M, Pan RK, Kaski K, Kertész J, Saramäki J, Karsai M (2012) Multiscale analysis of spreading in a large communication network. Journal of Statistical Mechanics: Theory and Experiment 2012 (03):P03005

246. Krings G, Karsai M, Bernhardsson S, Blondel VD, Saramäki J (2012) Effects of time window size and placement on the structure of an aggregated communication network. EPJ Data Science 1 (1):4

247. Uddin S, Choudhury N, Piraveenan M, Chung KSK (2017) Exploring Actor-Level Dynamics in Longitudinal Networks: The State of the Art. In: Alhajj R, Rokne J (eds) Encyclopedia of Social Network Analysis and Mining. Springer New York, New York, NY, pp 1-17. doi:10.1007/978-1-4614-7163-9_110155-1

248. Uddin S, Piraveenan M, Chung KSK, Hossain L Topological analysis of longitudinal networks. In: System Sciences (HICSS), 2013 46th Hawaii International Conference on, 2013. IEEE, pp 3931-3940

249. Uddin S, Khan A, Hossain L, Piraveenan M, Carlsson S (2015) A topological framework to explore longitudinal social networks. Computational and Mathematical Organization Theory 21 (1):48-68

250. Uddin S, Hossain L, Murshed ST, Crawford JW (2011) Static versus dynamic topology of complex communications network during organizational crisis. Complexity 16 (5):27-36

251. Uddin S, Khan A, Piraveenan M (2016) A set of measures to quantify the dynamicity of longitudinal social networks. Complexity 21 (6):309-320

252. Choudhury N, Uddin S Evolution Similarity for Dynamic Link Prediction in Longitudinal Networks. In: Workshop on Complex Networks CompleNet, 2017. Springer, pp 109-118

253. Field A (2009) Discovering statistics using SPSS. Sage Publications Ltd,

254. Makridakis S, Wheelwright SC, Hyndman RJ (2008) Forecasting methods and applications. John Wiley & Sons,

255. Vallis O, Hochenbaum J, Kejariwal A A Novel Technique for Long-Term Anomaly Detection in the Cloud. In: 6th USENIX conference on File and Storage Technologies, San Jose, California, February 26-29 2014.

256. Rosner B (1983) Percentage points for a generalized ESD many-outlier procedure. Technometrics 25 (2):165-172

257. Wang H, Song M (2011) Ckmeans. 1d. dp: optimal k-means clustering in one dimension by dynamic programming. The R journal 3 (2):29

258. Clauset A, Eagle N (2012) Persistence and periodicity in a dynamic proximity network. arXiv preprint arXiv:12117343

259. Papadopoulos S, Kompatsiaris Y, Vakali A, Spyridonos P (2012) Community detection in social media. Data Mining and Knowledge Discovery 24 (3):515-554

260. Choudhury N, Uddin S (2017) Mining Actor-level Structural and Neighborhood Evolution for Link Prediction in Dynamic Networks. Paper presented at the Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017, Sydney, Australia,

261. Strogatz SH (2001) Exploring complex networks. nature 410 (6825):268

262. Chen K, Chen Y, Li Y, Han J (2016) A supervised link prediction method for dynamic networks. Journal of Intelligent & Fuzzy Systems 31 (1):291-299

263. Liu Q, Tang S, Zhang X, Zhao X, Zhao BY, Zheng H Network Growth and Link Prediction Through an Empirical Lens. In: Proceedings of the 2016 ACM on Internet Measurement Conference, 2016. ACM, pp 1-15

264. Hanneman RA, Riddle M (2005) Introduction to Social Network Methods

265. Porter MD, Smith R Network neighborhood analysis. In: Intelligence and Security Informatics (ISI), 2010 IEEE International Conference on, 2010. IEEE, pp 31-36

266. Sørensen T (1948) A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. Biol Skr 5:1-34

267. Vintsyuk TK (1968) Speech discrimination by dynamic programming. Cybernetics and Systems Analysis 4 (1):52-57

268. Salvador S, Chan P (2007) Toward accurate dynamic time warping in linear time and space. Intelligent Data Analysis 11 (5):561-580

269. Müller M (2007) Dynamic time warping. Information retrieval for music and motion:69-84

270. Tse CK, Liu J, Lau FCM (2010) A network perspective of the stock market. Journal of Empirical Finance 17 (4):659-667. doi:http://dx.doi.org/10.1016/j.jempfin.2010.04.008

271. Kose F, Weckwerth W, Linke T, Fiehn O (2001) Visualizing plant metabolomic correlation networks using clique–metabolite matrices. Bioinformatics 17 (12):1198-1208

272. Toubiana D, Fernie AR, Nikoloski Z, Fait A (2013) Network analysis: tackling complex data to study plant metabolism. Trends in biotechnology 31 (1):29-36

273. Yu D, Kim M, Xiao G, Hwang TH (2013) Review of biological network data and its applications. Genomics & informatics 11 (4):200-210

274. Batushansky A, Toubiana D, Fait A (2016) Correlation-Based Network Generation, Visualization, and Analysis as a Powerful Tool in Biological Studies: A Case Study in Cancer Cell Metabolism. BioMed Research International 2016:9. doi:10.1155/2016/8313272

275. Namaki A, Shirazi AH, Raei R, Jafari GR (2011) Network analysis of a financial market based on genuine correlation and threshold method. Physica A: Statistical Mechanics and its Applications 390 (21):3835-3841. doi:https://doi.org/10.1016/j.physa.2011.06.033

276. Bray JR, Curtis JT (1957) An ordination of the upland forest communities of southern Wisconsin. Ecological monographs 27 (4):325-349

277. Ricotta C, Podani J (2017) On some properties of the Bray-Curtis dissimilarity and their ecological meaning. Ecological Complexity 31:201-205

278. Yoshioka PM (2008) Misidentification of the Bray-Curtis similarity index. Marine Ecology Progress Series 368:309-310

279. Legendre P, Legendre LF (2012) Numerical ecology, vol 24. Developments in Environmental Modelling. Elsevier, Amsterdam, NL

280. Rosvall M, Bergstrom CT (2008) Maps of random walks on complex networks reveal community structure. Proceedings of the National Academy of Sciences 105 (4):1118-1123

281. Soundarajan S, Hopcroft J Using community information to improve the precision of link prediction methods. In: Proceedings of the 21st International Conference on World Wide Web, 2012. ACM, pp 607-608

282. Valverde-Rebaza JC, de Andrade Lopes A (2012) Link prediction in complex networks based on cluster information. In: Advances in Artificial Intelligence-SBIA 2012. Springer, pp 92-101

283. Raghavan UN, Albert R, Kumara S (2007) Near linear time algorithm to detect community structures in large-scale networks. Physical review E 76 (3):036106

284. Newman ME (2004) Fast algorithm for detecting community structure in networks. Physical review E 69 (6):066133

285. Xu HH, Zhang LJ Application of Link Prediction in Temporal Networks. In: Advanced Materials Research, 2013. Trans Tech Publication, pp 2231-2236

286. Kunegis J Konect: the koblenz network collection. In: Proceedings of the 22nd International Conference on World Wide Web, 2013. ACM, pp 1343-1350

287. Rossi RA, Ahmed NK Networkrepository: A graph data repository with visual interactive analytics. In: 29th AAAI Conference on Artificial Intelligence, Austin, Texas, USA, 25-30 January 2015. Association for the Advancement of Artificial Intelligence, pp 4292-4293

288. Isella L, Stehlé J, Barrat A, Cattuto C, Pinton J-F, Van den Broeck W (2011) What's in a crowd? Analysis of face-to-face behavioral networks. Journal of theoretical biology 271 (1):166-180

289. De Sá HR, Prudêncio RB Supervised link prediction in weighted networks. In: Neural Networks (IJCNN), The 2011 International Joint Conference on, 2011. IEEE, pp 2281-2288

290. Al Hasan M, Chaoji V, Salem S, Zaki M (2006) Link prediction using supervised learning. In: 6th SDM' Workshop on Link Analysis, Counter-terrorism and Security, Bethesda, Maryland, 22-24 April 2006. Society for Industrial and Applied Mathematics,

291. Alejo R, García V, Sotoca JM, Mollineda RA, Sánchez JS Improving the performance of the RBF neural networks trained with imbalanced samples. In: International Work-Conference on Artificial Neural Networks, 2007. Springer, pp 162-169

292. Fu X, Wang L, Chua KS, Chu F Training RBF neural networks on unbalanced data. In: Neural Information Processing, 2002. ICONIP'02. Proceedings of the 9th International Conference on, 2002. IEEE, pp 1016-1020

293. Murphey YL, Wang H, Ou G, Feldkamp LA OAHO: an effective algorithm for multi-class learning from imbalanced data. In: Neural Networks, 2007. IJCNN 2007. International Joint Conference on, 2007. IEEE, pp 406-411

294. Nguyen GH, Bouzerdoum A, Phung SL A supervised learning approach for imbalanced data sets. In: Pattern Recognition, 2008. ICPR 2008. 19th International Conference on, 2008. IEEE, pp 1-4

295. Zhou Z-H, Liu X-Y (2006) Training cost-sensitive neural networks with methods addressing the class imbalance problem. IEEE Transactions on Knowledge and Data Engineering 18 (1):63-77

296. Berardi VL, Zhang GP (1999) The effect of misclassification costs on neural network classifiers. Decision Sciences 30 (3):659-682

297. Yoon K, Kwek S (2007) A data reduction approach for resolving the imbalanced data issue in functional genomics. Neural Computing and Applications 16 (3):295-306

298. Choudhury N, Uddin S (2016) Time-aware link prediction to explore network effects on temporal knowledge evolution. Scientometrics 108 (2):745-776. doi:10.1007/s11192-016-2003-5

299. Choudhury N, Uddin S (2017) Evolution Similarity for Dynamic Link Prediction in Longitudinal Networks. In: Gonçalves B, Menezes R, Sinatra R, Zlatic V (eds) Complex Networks VIII: Proceedings of the 8th Conference on Complex Networks CompleNet 2017. Springer International Publishing, Cham, pp 109-118. doi:10.1007/978-3-319-54241-6_9

300. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research 16:321-357

301. Yang Y, Lichtenwalter RN, Chawla NV (2015) Evaluating link prediction methods. Knowledge and Information Systems 45 (3):751-782

302. Zhou T, Lü L, Zhang Y-C (2009) Predicting missing links via local information. The European Physical Journal B-Condensed Matter and Complex Systems 71 (4):623-630

303. da Silva Soares PR, Prudêncio RBC Time series based link prediction. In: Neural Networks (IJCNN), The 2012 International Joint Conference on, 2012. IEEE, pp 1-7

304. Hyndman RJ, Khandakar Y (2007) Automatic time series for forecasting: the forecast package for R. vol 6/07. Monash University, Department of Econometrics and Business Statistics,

305. Breiman L (1996) Bagging predictors. Machine learning 24 (2):123-140

306. Liaw A, Wiener M (2002) Classification and regression by randomForest. R news 2 (3):18-22

307. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) The WEKA data mining software: an update. ACM SIGKDD explorations newsletter 11 (1):10-18

308. Zhou T, Lü L, Zhang Y-C (2009) Predicting missing links via local information. The European Physical Journal B 71 (4):623-630

309. Adamic LA, Adar E (2003) Friends and neighbors on the Web. Social Networks 3 (25):211-230

310. Davis J, Goadrich M The relationship between Precision-Recall and ROC curves. In: Proceedings of the 23rd international conference on Machine learning, 2006. ACM, pp 233-240

311. Kaya B, Poyraz M (2014) Supervised link prediction in symptom networks with evolving case. Measurement 56:231-238

312. Symeonidis P, Iakovidou N, Mantas N, Manolopoulos Y (2013) From biological to social networks: Link prediction based on multi-way spectral clustering. Data & Knowledge Engineering 87:226-242

313. Neelamegam P, Jamaludeen A, Rajendran A (2011) Prediction of calcium concentration in human blood serum using an artificial neural network. Measurement 44 (2):312-319

314. Yosef N, Regev A (2011) Impulse control: temporal dynamics in gene transcription. Cell 144 (6):886-896

315. A supervised approach to time scale detection in dynamic networks (2017) Cornell University Library. https://arxiv.org/abs/1702.07752.

316. Fortunato S (2010) Community detection in graphs. Physics reports 486 (3):75-174

317. Valverde-Rebaza J, de Andrade Lopes A (2013) Exploiting behaviors of communities of twitter users for link prediction. Social Network Analysis and Mining 3 (4):1063-1074

Appendix A

AdamicAdar

AdamicAdar is a topological similarity index that is widely used in link prediction in static networks. This index refines the simple counting of common features (e.g., number of common neighbours) by weighting intermittent features heavily. The similarity index $S_{AA}(a, b)$ between two actors a and b is calculated by adding weights to the nodes which are connected to both nodes a and b:

$$S_{AA}(a, b) = \sum_{z \in \Gamma(a) \cap \Gamma(b)} \frac{1}{\log K_z}$$

Where $\Gamma(a)$ denotes the neighbourhood of actor a, z is the common neighbour to both a and b, and K denotes the degree of actor z

ResourceAllocation

ResourceAllocation is an index used in link prediction models over static networks. The index is a topological similarity index that calculates similarity between two actors in a network based on the intermittent actors connecting these two actors. The ResourceAllocation similarity index $S_{RA}(a, b)$ between actor a and b is computed as the amount of resource actor a receives from actor b through indirect links where each intermediate link contributes a unit of resource:

$$S_{RA}(a, b) = \sum_{z \in \Gamma(a) \cap \Gamma(b)} \frac{1}{K_z}$$

Where $\Gamma(a)$ denotes the neighbourhood of actor a and K denotes the degree of actor z

CommonNeighbours

It is one of the simplest methods of link prediction that captures the notion that two strangers who have a common friend may be introduced by that friend. This method introduces the triangle closure mechanism in graph topology and denotes the fact that two actors in a network are likely to form a link if they have many common neighbours and/or friends. CommonNeighbours similarity index $S_{CN}(a, b)$ between two actors a and b is determined by:

$$S_{CN}(a, b) = |\Gamma(a) \cap \Gamma(b)|$$

Jaccard Coefficient

This statistic was proposed to compare similarity and diversity of sample sets. It denotes the ratio of common neighbours of actors a and b to the all neighbors nodes of a and b. Jaccard coefficient prevents higher degree actors to have high similarity score with other actors. The Jaccard similarity measure $S_{JC}(a, b)$ between two actors a and b is denoted by:

$$S_{JC}(a, b) = \frac{|\Gamma(a) \cap \Gamma(b)|}{|\Gamma(a) \cup \Gamma(b)|}$$

Clustering Coefficient

In network theory, Clustering Coefficient measures the degree to which actors in a network tend to cluster together. In most real-world social network, actors are inclined to create tightly knit groups characterised by a relatively high density of ties. Clustering coefficient is generally considered as local (i.e., actor-level) measure. The clustering coefficient CC of an actor a is calculated as

$$CC_a = \frac{2L_a}{K_a(K_a - 1)}$$

Where K_a is the degree of actor a and L_a denotes the number of edges between the K_a neighbours of actor a. Alternatively, it is defined in regards to triadic closure which, in social network principle, denotes the fact that if two people in a social network have a friend in common, then there is an increased likelihood that they will form a relationship in future. It is also a measure to quantify how complete the neighbourhood of an actor is. The term clustering coefficient can also be defined in regards to triadic closure. The later denotes the tendency for people who share connections ia social network to become connected. Therefore, considering triangles (i.e., two actors sharing one common friend), the local clustering coefficient is defined as the fraction of pairs of actor's friends are friends with each other.

$$CC_a = \frac{number\ of\ triangles\ connected\ to\ actor\ a}{Number\ of\ triangles\ centered\ around\ a}$$

Cliquishness

In graph theory, a clique is a subset of vertices of an undirected graph such that every two distinct vertices in the clique are adjacent. The term cliquishness represents the tendency to associate with only select groups. In social network's perspective, cliquishness signifies a group of individuals interacting with one another or share similar interests.

Short Interval Network

A longitudinal network consists of a time series of network snapshots observed at different points in time to collect network data for analysis. These observed networks are short interval networks

Aggregated Network

Accumulation of a series of short interval networks into a bigger network is known as the aggregated network.

Positional Dynamicity

The positional dynamicity represents the changes of network positions of actors in different short interval networks relative to their positions in the aggregated network. In two consecutive short interval networks, an actor can change its neighbourhood connectivity in many different ways. This will ultimately change its network position between these two network snapshots.

The network position of individual actors could be quantified using any actor-level social network measure (e.g. degree centrality or closeness centrality).

Assuming that a given longitudinal network has been observed at $t_1, t_2, t_3 \ldots t_m$ different times where $t_m > t_{m-1} > t_{m-2} > \ldots > t_2 > t_1$ with m short interval networks and one aggregated network for this longitudinal network. These m short interval networks have $n_1, n_2, n_3 \ldots n_m$ actors or nodes whereas the aggregated network consists of N actors. An actor may appear in more than one short interval network. The sets of actors present in these m short-interval networks are $S_1, S_2, S_3 \ldots S_m$. Thus,

$$|S_1|=n_1;\ |S_2|=n_2 \ldots |S_{m-1}|=n_{m-1};\ |S_m|=n_m$$

$$|S_1\ U\ S_2\ U\ S_3\ U \ldots\ U\ S_{m-1}\ U\ S_m|=N$$

The two dimensional matrix $M\ (N\ X\ m)$ represents the presence and absence details of N actors in m short interval networks. This matrix contains only binary values, either 0 or 1. For example, M (2, 3) = 1 denotes that the second actor is present in the third short interval

network. Based on these assumptions, positional dynamicity can be calculated for an actor by the following equation:

$$PoD_i = \frac{\sum_{t=1}^{m}\left[\frac{|NP_{AN}^{i} - NP_{SIN(t)}^{i}|}{|NP_{AN}^{i} + NP_{SIN(t)}^{i}|} * M(i,t)\right]}{m} * 100\%$$

Where, PoD_i denotes the positional dynamicity demonstrated by actor i^{th} actor; NP_{AN}^{i} indicates the network position measure which is calculated by any actor-level social network measure such as closeness centrality for the ith actor in the aggregated network; $NP_{SIN(t)}^{i}$ denotes the network position measure based on the same social network measure (i.e. closeness centrality) in the tth short interval network for the ith actor; $M(i,t)$ represents the participation details of actors in all short interval networks; and m indicates the number of short interval networks in the longitudinal social network. The denominator m is used as a normalizing factor, allowing the above equation to compare the positional dynamicity of different actors in different longitudinal social networks.

Participation Dynamicity

This dynamicity component exemplifies the changing network participation of actors in any two consecutive short interval networks. In a given longitudinal social network that follows the similar assumption as in positional dynamicity, an actor may not be present in all short-interval networks. For example, an actor may participate in the $(t\text{-}1)^{th}$ short interval network but become absent in the tth short interval network, or alternatively, it may choose to participate in the tth short interval network and remain absent in the subsequent $(t+1)^{th}$ short interval network. These types of actor participatory transitions in consecutive short interval network also contribute to the dynamicity shown by the longitudinal network. Since an actor's presence in the current underlying short interval network can ensure its contribution

towards network dynamicity there exists two possible ways for an actor to show the participation dynamicity. Firstly, the actor is present in both the tth and $(t\text{-}1)^{th}$ short interval network. In this case, the participation dynamicity for that actor in the tth short interval network can be calculated with the following equation:

$$PaD^{i}_{SIN(t)} = \left[1 - \frac{n_t}{N} * \frac{n_{t-1}}{N}\right] \text{ ……. (5)}$$

Here, $PaD^{i}_{SIN(t)}$ represents the participation dynamicity of actor *i* in the tth short interval network. $\frac{n_t}{N}$ and $\frac{n_{t-1}}{N}$ indicate the probabilities that the actor will be present in the tth and $(t\text{-}1)^{th}$ short interval network, respectively. Hence, $\frac{n_t}{N} * \frac{n_{t-1}}{N}$ indicates the probability that the actor will be present in both the tth and $(t\text{-}1)^{th}$ short interval network. This value (i.e. $\frac{n_t}{N} * \frac{n_{t-1}}{N}$) represents how likely it is that that actor will be found in both the tth and $(t\text{-}1)^{th}$ short interval networks. Thus, the complement of this value is the participation dynamicity shown by the actor in the tth short interval network. As the maximum value for $\frac{n_t}{N} * \frac{n_{t-1}}{N}$ is 1, the complement of $\frac{n_t}{N} * \frac{n_{t-1}}{N}$ is $[1 - \frac{n_t}{N} * \frac{n_{t-1}}{N}]$.

.