Intelligent Information Management, 2024, 16, 21-34
https://www.scirp.org/journal/iim
ISSN Online: 2160-5920
ISSN Print: 2160-5912

Discrete Choice Analysis of Temporal Factors on Social Network Growth

Kwok-Wai Cheung, Yuk Tai Siu

School of Communication, The Hang Seng University of Hong Kong, Hong Kong, China
Email: keithcheung@hsu.edu.hk, ytsiu@hsu.edu.hk

How to cite this paper: Cheung, K.-W. and Siu, Y.T. (2024) Discrete Choice Analysis of Temporal Factors on Social Network Growth. *Intelligent Information Management*, **16**, 21-34.
https://doi.org/10.4236/iim.2024.161003

Received: November 18, 2023
Accepted: January 12, 2024
Published: January 15, 2024

Abstract

Social networks like Facebook, X (Twitter), and LinkedIn provide an interaction and communication environment for users to generate and share content, allowing for the observation of social behaviours in the digital world. These networks can be viewed as a collection of nodes and edges, where users and their interactions are represented as nodes and the connections between them as edges. Understanding the factors that contribute to the formation of these edges is important for studying network structure and processes. This knowledge can be applied to various areas such as identifying communities, recommending friends, and targeting online advertisements. Several factors, including node popularity and friends-of-friends relationships, influence edge formation and network growth. This research focuses on the temporal activity of nodes and its impact on edge formation. Specifically, the study examines how the minimum age of friends-of-friends edges and the average age of all edges connected to potential target nodes influence the formation of network edges. Discrete choice analysis is used to analyse the combined effect of these temporal factors and other well-known attributes like node degree (*i.e.*, the number of connections a node has) and network distance between nodes. The findings reveal that temporal properties have a similar impact as network proximity in predicting the creation of links. By incorporating temporal features into the models, the accuracy of link prediction can be further improved.

Keywords

Discrete Choice Models, Temporal Factors, Social Network, Link Prediction, Network Growth

1. Introduction

Nowadays, there is a wealth of user-generated content that can be easily shared

and spread via social networks like Facebook, X (Twitter), and LinkedIn. These platforms continuously create digital traces of social behaviours, providing researchers in social science and communication science with rich data for analysis. By employing computational methods [1], researchers can effectively study social networks to extract valuable information, including user interests, opinions, interactions, and online events.

The analysis of social network data has various practical applications. For instance, it can be used for node classification involving the assignment of labels or categories to nodes of a graph structure representing a social network [2]; for community detection that helps identify groups of nodes [3]; and for link prediction that anticipates the likelihood or presence of connections between nodes [4]. Recommendation systems also utilise social network data to suggest items or content to users based on their preferences, interests, or past behaviours [5], and online advertising systems leverage the data to effectively promote products, services, or brands across diverse online platforms [6]. In social healthcare analysis, the focus is on capturing the interaction between users within healthcare communities [7], whereas social influence analysis aims to understand how individuals or groups affect the thoughts, opinions, behaviours, or actions of others within a social network [8]. Furthermore, academic network analysis looks into the relationships, collaborations, and interactions within the academic and scientific community [9].

The aim of our study is to examine the impact of temporal factors on the growth of social networks using discrete choice analysis [10]. Network growth is a dynamic process that changes over time, and discrete choice framework enables us to analyse this process by viewing the formation of edges as nodes' decisions to connect with other nodes. Through considering temporal factors alongside other common variables, such as node degree and network proximity between nodes, we can gain a comprehensive understanding of how networks form and the underlying processes involved. Investigating the relevant factors contributing to edge formation in networks is, therefore, of great theoretical and practical importance that helps us understand the formation and structure of networks, as well as the processes that occur within them. This knowledge is valuable for identifying network hubs, such as opinion leaders, and for simulating and evaluating systems of different sizes.

Our study focuses on two main aspects as follows: 1) identifying temporal activities that influence the formation of network edges, and 2) comparing the effects of temporal activities with well-known factors that contribute to network growth.

- To investigate the first aspect, we analyze a publicly available network dataset that includes timestamps. Specifically, we examine the minimum age of friends-of-friends edges and the average age of all edges of potential target nodes. These measures provide insights into the recency of activity associated with the potential target node. We use conditional logit models to estimate

the probability of edge formation.

- To delve into the second aspect, which relates to the complex process of network growth influenced by mechanisms like preferential attachment and triadic closure, we employ mixed logit models that enable us to study and compare the effects and contributions of temporal factors with other factors. By doing so, we can gain a better understanding of how temporal factors contribute to the edge formation process driving the growth of online social networks.

2. Background

Online social networks are complex networks (networks of complex topology), consisting of vertices representing the users of the system and edges depicting the social interaction between them. With the availability of topological data on large online networks due to the computerization of data acquisition, the study of the dynamic and topology of these networks become possible. Different growth mechanisms are observed in online social networks. Some common ones are:

Preferential Attachment: New links are attached preferentially to the users with high connectivity or degree. There is a positive correlation between the number of links users have and the probability of having new links attached to them [11] [12] [13] [14] [15]. Such a correlation is found responsible for the scale-free power-law distribution (scale invariant distribution) of many large networks in nature with their degree distribution following a power law, $P(k) \sim k^{-\gamma}$, where k is the degree of a vertex [11]. Networks will continuously grow by adding new vertices attached preferentially to vertices with higher connectivity. Examples following the power law degree distribution include collaboration graph of movie actors (actor collaboration graph), the Web, Power grid, and citation network.

Reciprocation: Online social networks are usually directed networks in which users choose to follow other users. The creation of a link between two users will very probably invoke the establishment of a reverse or reciprocal link within a short time [16].

Network Proximity: Users are more likely to link to the nearby users. The distance between users is measured with shortest path hop distance. This is a local mechanism of preferential attachment. New links are more likely created between users with friend-of-friend relationship, *i.e.* with two hops distance before new link is created [16].

These growth mechanisms are independently applied in the study of social networks. The growth of social network involves the formation of new nodes and edges. To study edge formation, assume a social network as a time-evolving simple directed graphs $G_t = (V_t, E_t)$, $t \in \mathbb{Z}$, with time-dependent feature vector x_j for node $j \in V_t$, and an edge $(i, j) \in E_t$ is formed by connecting node i to node j. The key question is to find out the probability of node i connecting to

node *j*.

For preferential attachment [11], the probability of a new node connecting to $|V_t|$ distinct existing nodes *j* is proportional to the power of the nodes' degree, which can be expressed as

$$P(j,V_t)=\frac{d_j^{\alpha}}{\sum_{n\in V_t} d_n^{\alpha}} \tag{1}$$

where d_j is the degree of nodes *j*.

For uniform attachment [17], an edge is formed by randomly sampling a neighbour node from all nodes. The probability is simplified to:

$$P(j,V_t)=\frac{1}{|V_t|} \tag{2}$$

For triadic closure [18], a variant of uniform attachment, an edge is formed by randomly sampling a neighbour node from the set of their friends-of-fiends, $FoF(i,j)_t$, instead of all nodes. The probability becomes:

$$P(j,V_t)=\frac{1}{|FoF(i,j)_t|} \tag{3}$$

However, all these mechanisms alone are not able to satisfactorily explain the edge formation and network growth as the process is influenced by multiple mechanisms and a number of node attributes including: 1) nodes' degree, 2) reciprocity, 3) friends-of-friends, and 4) network proximity.

Different models or frameworks such as discrete choice analysis framework [10] [19] and generalized triadic closure [20] were proposed to study network properties with power law degree distribution, community clustering, proximity, etc. In [20], a generalized configuration model with random triadic closure (GCTC) is presented. The model has five fundamental properties: large clustering coefficient, power law degree distribution, short path length, non-zero Pearson degree correlation, and existence of community structures. As the social links are not all equal, [21] estimates the different importance of social relationships to enhance feature-extraction link prediction algorithms. Most machine learning approaches only support static graphs, which train on snapshots of the system at a specific time [22] [23] [24]. By extracting a local subgraph around each target link, Zhang & Chen [25] [26] introduced a graph neural network (GNN) that extracts a local subgraph around each target link. This GNN can learn a function that maps subgraph patterns to link existence, effectively learning a "heuristic" suitable for the network. Furthermore, there is a growing interest in considering the temporal dynamics of networks [27]. To handle continuous time-dynamics graphs, Temporal Graph Attention (TGA) [28] [29] is employed.

Understanding the formation of network edges is an essential component in comprehending how networks organise themselves and the processes that take place within them. Given the prominence of online social network, it is useful to

gain the understanding of both their global and local characteristics, and to construct structural and growth models for network analysis. Applications of such understanding include searching for network hubs to identify opinion leaders as well as simulating and evaluating system of various sizes.

In this paper, the effect of temporal activity on network edge formation is investigated. In addition to the global and local mechanisms such as the well-known preferential attachment and triadic closure, temporal proximity is also considered as being a significant factor in link formation process. The temporal aspects driving the growth of online social network are further studied using discrete choice analysis to examine their effect and contribution in the process of edge formation.

3. Discrete Choice Analysis

Using discrete choice analysis, various network edge formation mechanisms can be investigated in a unified manner [19]. In discrete choice framework, $\mathcal{D}=\{(j_k,C_k)\,|\,k=1,\cdots,N\}$ denotes a dataset of N different choices. Each choice (j,C) consists of a set of mutually exclusive alternatives, called the choice set C, and a chosen alternative $j\in C$. Each alternative j is associated with a feature vector x_j representing its attributes.

In random utility-based discrete choice models [10], the decision maker i obtains utility U_{ij} from choosing alternative j. Under the assumption of utility-maximizing behaviour of utility theory, the decision maker will choose the alternative that gives the highest utility. With conditional logit, the utility obtained by decision maker i for alternative j depends on the attributes of the alternative, and is given by

$$U_{ij}=\theta^{\mathrm{T}}x_j+\varepsilon_{ij} \tag{4}$$

where θ is a fixed parameter vector for decision maker and ε_{ij} is iid (independent and identically distributed) extreme value. The probability of individual i choosing alternative j is given as:

$$P_i(j,C)=\frac{\exp\left(\theta^{\mathrm{T}}x_j\right)}{\sum_{n\in C}\exp\left(\theta^{\mathrm{T}}x_n\right)} \tag{5}$$

Under discrete choice framework, the choice dataset $\mathcal{D}$ of a growing network G_t consists of a chosen node j_t, a node set C_t, and a degree set $\{d_{j,t}\mid j\in C_t\}$ at each time-step t. By defining feature vector as $x_j=\log d_j$, the edge formation probability can be expressed as

$$P(j,C_t)=\frac{\exp\left(\alpha\log d_j\right)}{\sum_{n\in V_t}\exp\left(\alpha\log d_n\right)} \tag{6}$$

which is equivalent to that of preferential attachment. The power of degree at the time of the choice becomes the only choice model parameter $\theta=\alpha$. The log-likelihood for the parameter α for the choice dataset $\mathcal{D}$ is

$$l(\alpha;\mathcal{D}) = \sum_{(j,C)\in\mathcal{D}} \log \frac{\exp(\alpha \log d_j)}{\sum_{n\in C}\exp(\alpha \log d_n)} = \sum_{(j,C)\in\mathcal{D}} \left(\alpha \log d_j - \log \sum_{n\in C}\exp(\alpha \log d_n)\right). \tag{7}$$

For uniform attachment, an edge is formed by randomly sampling a neighbour node from all nodes, the likelihood is simplified to

$$l(\mathcal{D}) = \sum_{(j,C)\in\mathcal{D}} \log \frac{\exp(1)}{\sum_{n\in C}\exp(1)} = \sum_{(j,C)\in\mathcal{D}} -\log|C|. \tag{8}$$

For triadic closure, the neighbour node *j* is sampled uniformly at random from friends-of-friends of the chooser node *i*. Similar to uniform attachment except that the choice set is restricted, the likelihood is

$$l(\mathcal{D}) = \sum_{(j,C_{ij})\in\mathcal{D}} -\log|C_{ij}| \tag{9}$$

where C_{ij} is the friends-of-friends of *i* and *j*.

One advantage of using discrete choice analysis is that different mechanisms can be easily unified using mixed logit in handling choosers with different preferences. For discrete mixtures of logits, the edge formation probability is

$$P_i(j,C) = \sum_{m=1}^{M} \pi_m \frac{\exp(\theta_m^{\mathrm{T}} x_j)}{\sum_{n\in C}\exp(\theta_m^{\mathrm{T}} x_n)} \tag{10}$$

where $\pi_1,\cdots,\pi_m$ are class probabilities (weights of different modes) and $\sum_{m=1}^{M}\pi_m = 1$. Expectation maximization (EM) is used to optimize the likelihood of mixed models to estimate the parameters θ_m and class probabilities π_m.

The effect of temporal activities of target nodes on edge formation can be investigated together with other factors under this discrete choice framework. The temporal activities include the minimum age of the friends-of-friends edges of the potential target node, and the mean of the ages of all edges of potential target node. These measures indicate how recent the activity the potential target node is. Thus, the node attributes to be studied include nodes' degree, reciprocity, friends-of-friends, network proximity, and temporal activities of nodes.

4. Experimental Results and Findings

Our study utilized a dataset from the Flickr social network, spanning from November 2006 to May 2007 [16] [30]. This dataset consists of 3.2 million nodes and 33.1 million edges, providing valuable insights into real-world network formation. In Flickr, users have the option to follow other users, with these "following" connections being publicly visible, while "followed by" connections are not. The data was collected using a breadth-first search crawl, focusing on the connected components reachable from the initial seed. Since a full crawl was conducted daily, the data enables the identification of the timing of new connections on a daily basis.

The analysis of the dataset reveals several common growth mechanisms in network formation. Reciprocation is observed, indicating that when two users

establish an initial link, there is a high likelihood of quickly establishing a reciprocal link. Preferential attachment is identified as a global mechanism, where users with a higher number of existing links have a greater probability of receiving new links. Proximity, on the other hand, is a local mechanism, where the shortest-path hop distance between two users significantly influences the likelihood of a link being created between them. Notably, it is observed that a majority of new links are formed between users who are two hops away from each other, indicating a friend-of-friend relationship.

To study the process of network formation, a set of conditional logit models was employed using a sample of 22,000 edge formation events that occurred around the same date. Out of these 22,000 events, 20,000 were used to fit the model, while the remaining 2000 were reserved for testing the model's performance. For link prediction study, a number of non-chosen/negative target nodes are sampled from network growth dataset due to the computational challenges associated with calculating the gradients of log-likelihood for choice sets with a large number of nodes. Negative sampling technique was applied to expedite the computation process. For each edge formation event, 24 non-chosen alternatives were negatively sampled, resulting in a smaller choice set for each choice analysis.

Four experiments were conducted using the Flickr dataset, and the results are as presented in Tables 1-4. It is important to note that since the data samples are

Table 1. Comparison of temporal proximity of FoF target with other common features. Conditional logit model fits for Flickr data. Standard errors of the estimates are given in parentheses. (Note: *p < 0.01, **p < 0.05)

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Log in-degree	1.113* (0.006)	1.132* (0.007)		0.696* (0.009)	0.522* (0.009)	0.674* (0.009)
Has degree	−1.111* (0.133)	−0.645* (0.187)		−0.687* (0.181)	−1.687* (0.229)	−0.673* (0.180)
Reciprocal		8.614* (0.244)	8.488* (0.241)	8.489* (0.266)	8.485* (0.281)	8.438* (0.293)
FoF			6.089* (0.044)	3.959** (0.049)		
2 hops					6.240* (0.184)	
3 hops					2.811* (0.180)	
4 hops					0.527* (0.184)	
5 hops					−0.458* (0.209)	
≥6 hops					−1.280* (0.275)	
Same day						6.772* (0.117)
1 day						6.259* (0.139)
2 days						5.742* (0.156)
3 days						5.041* (0.146)
4 days						4.614* (0.159)
5 days						4.359* (0.153)
≥6 days						3.279* (0.053)
Log-likelihood	−21,661	−16,770	−14,767	−10,940	−9999	−9734
Test accuracy	0.683	0.7486	0.754	0.8485	0.851	0.8735

Table 2. Comparison of temporal proximity with connected target feature with other common features. Conditional logit model fits for Flickr data. Standard errors of the estimates are given in parentheses. (Note: *p < 0.01, **p < 0.05).

	Model 4	Model 5	Model 7	Model 8
Log in-degree	0.694* (0.009)	0.511* (0.009)	0.577* (0.009)	0.295* (0.011)
Has degree	−0.847* (0.173)	−2.000* (0.225)	0.447** (0.178)	−1.866* (0.220)
Reciprocal	8.725* (0.285)	8.467* (0.292)	8.405* (0.287)	8.418* (0.327)
FoF	4.107* (0.051)		4.082* (0.054)	
2 hops		6.556* (0.195)		9.424* (0.240)
3 hops		2.979* (0.191)		5.984* (0.237)
4 hops		0.869* (0.194)		3.882* (0.240)
5 hops		−0.172 (0.216)		2.800* (0.261)
≥6 hops		−1.123* (0.279)		1.783* (0.321)
Last hop time diff			−0.076* (0.002)	−0.116* (0.003)
Log-likelihood	−10,802	−9886	−10,078	−8867
Test accuracy	0.8475	0.851	0.865	0.8735

Table 3. Comparison of temporal activities level of target feature with other common features. Conditional logit model fits for Flickr data. Standard errors of the estimates are given in parentheses. (Note: *p < 0.01)

	Model 4	Model 5	Model 9	Model 10
Log in-degree	0.686* (0.009)	0.510* (0.009)	0.724* (0.009)	0.550* (0.010)
Has degree	−0.561* (0.193)	−1.650* (0.238)	9.612* (0.367)	7.766* (0.460)
Reciprocal	8.789* (0.301)	8.588* (0.311)	12.250* (0.555)	12.106* (0.579)
FoF	3.973* (0.049)		3.794* (0.052)	
2 hops		6.420* (0.196)		6.550* (0.257)
3 hops		2.966* (0.192)		3.246* (0.253)
4 hops		0.801* (0.195)		1.085* (0.255)
5 hops		−0.295 (0.221)		−0.067 (0.276)
≥6 hops		−1.030* (0.278)		−0.878* (0.328)
Link recency			−0.171* (0.004)	−0.173* (0.004)
Log-likelihood	−10,847	−9926	−9525	−8676
Test accuracy	0.8405	0.843	0.867	0.8755

Table 4. Temporal features. Conditional logit model fits for Flickr data. Standard errors of the estimates are given in parentheses. (Note: *p < 0.01, **p < 0.05)

	Model 5	Model 7	Model 8	Model 9	Model 10	Model 11
Log in-degree	0.541* (0.010)	0.712* (0.009)	0.509* (0.010)	0.771* (0.010)	0.590* (0.010)	0.555* (0.011)
Has degree	−1.367* (0.275)	−0.146 (0.233)	−1.346* (0.272)	7.694* (0.349)	5.754* (0.432)	5.662* (0.436)
Reciprocal	7.831* (0.263)	8.074* (0.261)	7.783* (0.267)	9.329* (0.398)	8.918* (0.411)	9.059* (0.422)

Continued

FoF		5.063* (0.066)		3.767* (0.053)		
2 hops	6.517* (0.214)		7.730* (0.218)		6.526* (0.259)	7.366* (0.262)
3 hops	3.103* (0.210)		3.184* (0.209)		3.219* (0.255)	3.307* (0.255)
4 hops	0.953* (0.213)		1.000* (0.212)		1.100* (0.257)	1.161* (0.257)
5 hops	−0.162 (0.235)		−0.147 (0.233)		−0.072 (0.277)	−0.023 (0.276)
≥6 hops	−0.781* (0.276)		−0.794* (0.274)		−0.812** (0.320)	−0.766** (0.318)
Last hop time diff		−0.069* (0.002)	−0.066* (0.002)			−0.046* (0.002)
Link recency				−0.164* (0.003)	−0.163* (0.003)	−0.138* (0.004)
Log-likelihood	−9,886	−10,035	−9114	−9327	−8531	−8275
Test accuracy	0.848	0.8635	0.868	0.869	0.8795	0.8845

taken independently in each experiment, the estimated parameter values for the same model number would differ across these experiments. It also shows that negative sampling could provide a practical trade-off between computation loading and estimation performance.

4.1. Experiment 1—Temporal Proximity of FoF's Connection to Target

The edge formation process is modelled as a node's decision to connect to a target node among a set of alternatives. Several conditional logit models with different parameters are estimated using the mlogit package [31]. **Table 1** presents the estimates of these conditional logit models for the Flickr data. Models 1 to 5 incorporate well-known mechanisms such as preferential attachment, reciprocity, and network proximity, as described in [19]. Model 6, on the other hand, considers the temporal proximity of a friend-of-friend's (FoF) connection to the target node instead of network proximity.

The "Log in-degree" feature represents the number of followers and incorporates the preferential attachment mechanism for nodes with a non-zero in-degree. The "Has degree" feature is used to distinguish nodes with positive and zero degrees. The "Reciprocal" feature indicates if there is already a link from the target node to the choosing node. This feature consistently shows a significant contribution to link creation, possibly due to the notification system of Flickr that encourages reciprocal links [16]. The "FoF" feature captures the effect of the target node being a friend-of-friend of the choosing node. A friend-of-friend is a user followed by another user who is already followed by the choosing user. There is a strong correlation indicating that the choosing user is likely to follow that friend-of-friend user. Including both the "Log in-degree" and "FoF" features in Model 4 leads to a significant decrease in their parameter estimates compared to Model 2 and Model 3, suggesting that these global and local features are not independent and highly correlated. Model 5 measures network proximity by counting the number of hops between the choosing user node and the target node.

Table 1 demonstrates that the "FoF" feature has a notable impact on the choice of link creation. In Model 6, instead of network proximity, the temporal proximity of the FoF's links to the target node is examined. The time difference between the creation of the link from the FoF node to the target node and the current time is included as a feature in the model. For a FoF target node, where the shortest path length between the choosing node and the target node is 2 hops, temporal proximity is defined as the minimum difference (in terms of the number of days) between the current date and the creation date of the FoF edge. Larger estimated values for closer temporal proximity imply that a choosing user is more likely to follow another user who has recently been followed by their friend. This temporal factor has a similar impact to network proximity. Model 5 and Model 6 perform the best, yielding comparable likelihood and test accuracy values. Model 6 demonstrates an improved prediction accuracy of 2.25% with the inclusion of this temporal factor compared to Model 5. Similar results are also observed for edge formation data on another day, indicating that temporal proximity is an important factor in link formation, comparable to or even better than network proximity.

4.2. Experiment 2—Temporal Proximity with Target that Already Has Connection Paths

In Model 6 of Table 1, only the temporal proximity of friend-of-friend (FoF) links is considered. The FoF set can be seen as a small local group or community where there is strong mutual influence and a higher likelihood of link creation among its users. In this experiment, the community can be extended to include users who are already connected to the choosing user when estimating the probability of link creation with a target node. The temporal information of the paths is taken into account.

In addition to the FoF target, the temporal factor for target nodes that have at least one path (not just a 2-hops path) from the choosing node is also investigated. For each of these target nodes, the creation times of the last hops in the paths from the choosing node to the target node are identified. The temporal activities factor, referred to as "Last hop time diff," measures the time difference between the creation time of the most recently created last hop to the target and the current time of choice. This factor indicates how recent there have been activities with the target node within this group. A smaller value implies that the target node has had activities more recently, making it more likely for the choosing node to connect to it.

The effect of including the "Last hop time diff" feature is shown in Table 2. Model 8 demonstrates an improved prediction accuracy of 2.25% with the inclusion of this temporal factor compared to Model 5.

4.3. Experiment 3—Temporal Activities Level of the Target Node

The "Log in-degree" parameter, which represents preferential attachment, is a

global mechanism for measuring the activity level and popularity of a node. It is a cumulative measure that takes into account the node's past activity up to the present time. However, in online social networks, the popularity of a node in the past does not necessarily imply the same popularity at the current time. Instead, there is a higher tendency for a user to follow a target node with more recent incoming links, such as a current opinion leader.

In this experiment, to measure the overall temporal activity level of a target node, all of its incoming connections are considered, rather than just the paths from the choosing node. The target node's "link recency" is calculated as the average of the time differences between the current time and the creation time of all its incoming connections. This metric captures the overall temporal activity level of the target node.

The effect of including the "link recency" feature is shown in Table 3. When comparing Model 5 and Model 9, it is evident that the temporal factor can provide prediction performance comparable to the network proximity factor. Furthermore, combining the temporal and network proximity factors can further enhance the prediction accuracy. Model 10 demonstrates an improved prediction accuracy of 3.25% with the inclusion of this temporal factor compared to Model 5. This target node feature serves as a valuable complement to the number of incoming links feature in link prediction.

4.4. Experiment 4—Combining the Effects of Temporal Proximity with and Temporal Activities Level of the Target Node

Table 4 presents the results of combining the temporal proximity factor from experiment 2 and the temporal activities level of the target node from experiment 3. The inclusion of the temporal activities level of target node (Models 9 and 10) improves the prediction ability more significantly compared to the temporal proximity factor (Models 7 and 8). When compared to Model 5, there is an improvement of 3.15% and 2.1%, respectively. This suggests that the recent activities level is a more crucial factor in influencing the creation of links to the target node. However, when both temporal factors are used together in Model 11, the prediction improvement is only slightly further raised to 3.65%. This is understandable as these two features are related to the incoming links of target nodes and are somewhat correlated. Hence, the improvement is not additive.

5. Discussion and Conclusions

The study of network structures has both theoretical and practical importance. In the past, various mechanisms such as preferential attachment and triadic closure have been proposed and studied separately, as they have a significant impact on the growth of social networks. However, it is valuable to examine these mechanisms in a unified manner, as they may be interrelated.

By employing a discrete choice framework, this study aims to analyse the growth of social networks and unify different network growth mechanisms. In

addition to well-known mechanisms like reciprocation, preferential attachment, and proximity, the study also investigates the influence of temporal factors on network growth. To conduct these experiments, a public Flickr dataset is used to analyse the prediction of network growth using temporal information.

In experiment 1, the study found that considering the recency of a friend's link to the FoF target in triadic closure is compatible with network proximity in predicting network growth. This suggests that temporal information, specifically the temporal proximity factor, is an important factor in understanding network growth.

In experiment 2, a more general setting is investigated where the choosing user already has path(s) to the target node. This temporal feature is not limited to FoF links but can be applied to other scenarios as well. The results show that considering the recency of the path(s) estimated by the last hop creation time can provide a similar improvement compared to considering the recency of the friend's link to the FoF target. This finding is consistent with the expectation that FoF or triadic closure is a more important local mechanism in link creation, as the target nodes that already have connection paths are a superset of FoF nodes.

The inclusion of temporal features provides a different perspective on network growth. It is natural to expect that the shortest path network distance between a choosing user and nodes that already have connection paths will shorten over time as the number of links increases within a local group or community.

In addition to the in-degree measure, the average recency of a node's incoming connections serves as an indicator of a target's temporal activity level. Experiment 3 demonstrates that incorporating this link recency factor can enhance the predictive performance of a model using preferential attachment. Furthermore, experiment 4 shows that combining both the temporal proximity between nodes and the temporal activity level of the target can further improve the model's performance.

Since networks are composed of interconnected discrete entities and network growth involves the creation of links among nodes with different attributes, employing discrete choice analysis provides a unified modeling approach to examine and combine the effects of various mechanisms in network formation.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Atteveldt, W. and Peng, T.Q. (2018) When Communication Meets Computation: Opportunities, Challenges, and Pitfalls in Computational Communication Science. *Communication Methods and Measures*, **12**, 81-92. https://doi.org/10.1080/19312458.2018.1458084

[2] Bhagat, S., Cormode, G. and Muthukrishnan, S. (2011) Node Classification in Social Networks. In: Aggarwal, C., Ed., *Social Network Data Analytics*, Springer, Boston, 115-148. https://doi.org/10.1007/978-1-4419-8462-3_5

[3] Fortunato, S. (2010) Community Detection in Graphs. Physics Reports, **486**, 75-174. https://doi.org/10.1016/j.physrep.2009.11.002

[4] Liben-Nowell, D. and Kleinberg, J. (2007) The Link-Prediction Problem for Social Networks. *Journal of the American Society for Information Science and Technology*, **58**, 1019-1031. https://doi.org/10.1002/asi.20591

[5] Zhou, T., Ren, J., Medo, M. and Zhang, Y.C. (2007) Bipartite Network Projection and Personal Recommendation. *Physical Review E*, **76**, Article ID: 046115. https://doi.org/10.1103/PhysRevE.76.046115

[6] Li, Y., Zhang, D. and Tan, K.L. (2015) Real-Time Targeted Influence Maximization for Online Advertisements. *Proceedings of the VLDB Endowment*, **8**, 1070-1081. https://doi.org/10.14778/2794367.2794376

[7] Tang, X. and Yang, C.C. (2012) Ranking User Influence in Healthcare Social Media. *ACM Transactions on Intelligent Systems and Technology*, **3**, Article No. 73. https://doi.org/10.1145/2337542.2337558

[8] Peng, S., Wang, G. and Xie, D. (2017) Social Influence Analysis in Social Networking Big Data: Opportunities and Challenges. *IEEE Network*, **31**, 11-17. https://doi.org/10.1109/MNET.2016.1500104NM

[9] Guo, Z., Zhang, Z.M., Zhu, S., Chi, Y. and Gong, Y. (2014) A Two-Level Topic Model towards Knowledge Discovery from Citation Networks. *IEEE Transactions on Knowledge and Data Engineering*, **26**, 780-794. https://doi.org/10.1109/TKDE.2013.56

[10] Train, K.E. (2009) Discrete Choice Methods with Simulation. 2nd Edition, Cambridge University Press, Cambridge.

[11] Barabási, A.L. and Albert, R. (1999) Emergence of Scaling in Random Networks. *Science*, **286**, 509-512. https://doi.org/10.1126/science.286.5439.509

[12] Bagrow, J.P. and Dirk Brockmann, D. (2013) Natural Emergence of Clusters and Bursts in Network Evolution. *Physical Review X*, **3**, Article ID: 021016. https://doi.org/10.1103/PhysRevX.3.021016

[13] Caldarelli, G., Capocci, A., Rios, P.D.L. and Munoz, M.A. (2002) Scale-Free Networks from Varying Vertex Intrinsic Fitness. *Physical Review Letters*, **89**, Article ID: 258702. https://doi.org/10.1103/PhysRevLett.89.258702

[14] Holme, P. and Kim, B.J. (2002) Growing Scale-Free Networks with Tunable Clustering. *Physical Review E*, **65**, Article ID: 026107. https://doi.org/10.1103/PhysRevE.65.026107

[15] Jackson, M.O. and Rogers, B.W. (2007) Meeting Strangers and Friends of Friends: How Random Are Social Networks? *American Economic Review*, **97**, 890-915. https://doi.org/10.1257/aer.97.3.890

[16] Mislove, A., Koppula, H.S., Gummadi, K.P., Druschel, P. and Bhattacharjee, B. (2008) Growth of the Flickr Social Network. *WOSN*'08: *Proceedings of the First Workshop on Online Social Networks*, Seattle, 18 August 2008, 25-30. https://doi.org/10.1145/1397735.1397742

[17] Callaway, D.S., Hopcroft, J.E., Kleinberg, J.M., Newman, M.E.J. and Strogatz, S.H. (2001) Are Randomly Grown Graphs Really Random? Physical Review E, **64**, Article ID: 041902. https://doi.org/10.1103/PhysRevE.64.041902

[18] Rapoport, A. (1953) Spread of Information through a Population with Socio-Structural Bias: I. Assumption of Transitivity. *Bulletin of Mathematical Biophysics*, **15**, 523-533. https://doi.org/10.1007/BF02476440

[19] Overgoor, J., Benson, A.R. and Ugander, J. (2019) Choosing to Grow a Graph: Modeling Network Formation as Discrete Choice. *Proceedings of the Web Conference* 2019 (*WWW*'19), San Francisco, 13-17 May 2019, 1409-1420. https://doi.org/10.1145/3308558.3313662

[20] Zhang, R., Lee, D.S. and Chang, C.S. (2022) A Generalized Configuration Model with Triadic Closure. https://arxiv.org/pdf/2105.11688.pdf

[21] Toprak, M., Boldrini, C., Passarella, A. and Conti, M. (2022) Harnessing the Power of Ego Network Layers for Link Prediction in Online Social Networks. *IEEE Transactions on Computational Social Systems*, **10**, 48-60. https://doi.org/10.1109/TCSS.2022.3155946

[22] Martínez, V., Berzal, F. and Cubero, J.C. (2016) A Survey of Link Prediction in Complex Networks. *ACM Computing Surveys*, **49**, 1-33. https://doi.org/10.1145/3012704

[23] Bronstein, M.M., Bruna, J., LeCun, Y., Szlam, A. and Vandergheynst, P. (2017) Geometric Deep Learning: Going beyond Euclidean Data. *IEEE Signal Processing Magazine*, **34**, 18-42. https://doi.org/10.1109/MSP.2017.2693418

[24] Hamilton, W.L. (2020) Graph Representation Learning. Springer, Cham. https://doi.org/10.2200/S01045ED1V01Y202009AIM046

[25] Zhang, M. and Chen, Y. (2018) Link Prediction Based on Graph Neural Networks. 32*nd Conference on Neural Information Processing Systems* (*NeurIPS* 2018), Montréal, 3-8 December 2018, 5171-5181.

[26] Zhang, M. (2022) Graph Neural Networks: Link Prediction. In: Wu, L., Cui, P., Pei, J. and Zhao, L., Eds., *Graph Neural Networks*: *Foundations*, *Frontiers*, *and Applications*, Springer, Singapore, 195-223. https://doi.org/10.1007/978-981-16-6054-2_10

[27] Xu, D., Ruan, C., Korpeoglu, E., Kumar, S. and Achan, K. (2020) Inductive Representation Learning on Temporal Graphs. *Proceedings of the International Conference on Learning Representations*, Addis Ababa, 30 April 2020. https://openreview.net/forum?id=rJeW1yHYwH

[28] Rossi, E., Chamberlain, B., Frasca, F., Eynard, D., Monti, F. and Bronstein, M. (2020) Temporal Graph Networks for Deep Learning on Dynamic Graphs. arXiv: 2006.10637.

[29] Carchiolo, V., Cavallo, C., Grassia, M., Malgeri, M. and Mangioni, G. (2022) Link Prediction in Time Varying Social Networks. *Information*, **13**, Article 123. https://doi.org/10.3390/info13030123

[30] Mislove, A., Marcon, M., Gummadi, K.P., Druschel, P. and Bhattacharjee, B. (2007) Measurement and Analysis of Online Social Networks. *Proceedings of the* 7*th SIGCOMM Conference on Internet Measurement*, San Diego, 24-26 October 2007, 29-42. https://doi.org/10.1145/1298306.1298311

[31] Croissant, Y. (2020) Estimation of Random Utility Models in R: The Mlogit Package. *Journal of Statistical Software*, **95**, 1-41. https://doi.org/10.18637/jss.v095.i11